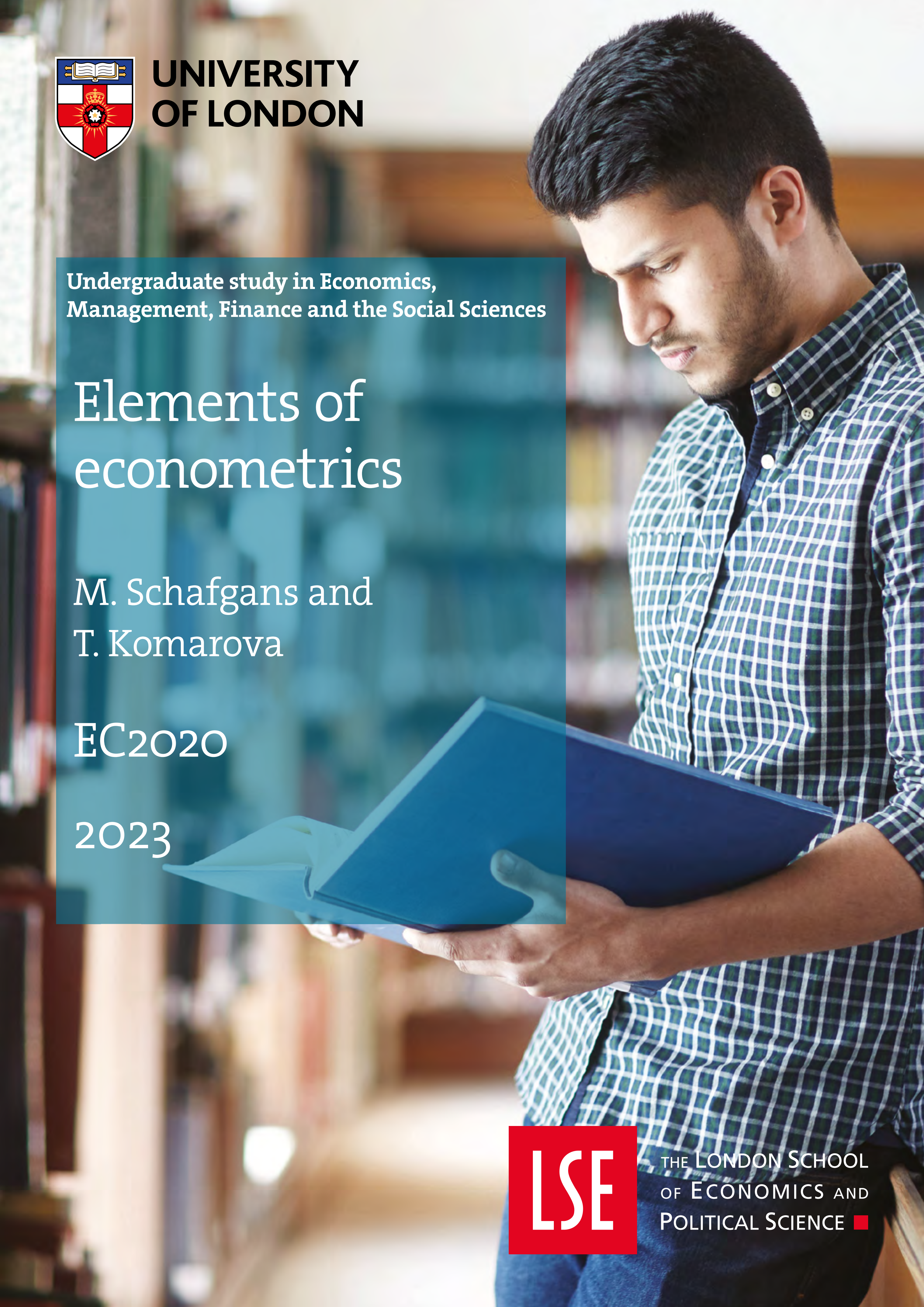


UNIVERSITY
OF LONDON

Undergraduate study in Economics, Management, Finance and the Social Sciences

Elements of econometrics

M. Schafgans and
T. Komarova

EC2020

2023

LSE
THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

UNIVERSITY
OF LONDON

Elements of econometrics

M. Schafgans and T. Komarova

EC2020

2023

Undergraduate study in
**Economics, Management,
Finance and the Social Sciences**

This subject guide is for a 200 course offered as part of the University of London undergraduate study in Economics, Management, Finance and the Social Sciences. This is equivalent to Level 5 within the Framework for Higher Education Qualifications in England, Wales and Northern Ireland (FHEQ).

For more information, see: london.ac.uk

LSE
THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

This guide was prepared for the University of London by:

M. Schafgans, Associate Professor of Economics

T. Komarova, Lecturer in Economics

Department of Economics, The London School of Economics and Political Science

With typesetting and proof-reading provided by:

James S. Abdey, BA (Hons), MSc, PGCertHE, PhD, Department of Statistics, London School of Economics and Political Science.

This is one of a series of subject guides published by the University. We regret that due to pressure of work the authors are unable to enter into any correspondence relating to, or arising from, the guide. If you have any comments on this subject guide, please communicate these through the discussion forum on the virtual learning environment.

University of London
Publications Office
Stewart House
32 Russell Square
London WC1B 5DN
United Kingdom

london.ac.uk

Published by: University of London

Contents

Chapter 0
Preface

0.1 What is econometrics, and why study it?

Econometrics is the application of statistical methods to the quantification and critical assessment of hypothetical economic relationships using data. It is with the aid of econometrics that we discriminate between competing economic theories and put numerical clothing onto the successful ones. Econometric analysis may be motivated by a simple desire to improve our understanding of how the economy works, at either the microeconomic or macroeconomic level, but more often it is undertaken with a specific objective in mind. In the private sector, the financial benefits that accrue from a sophisticated understanding of relevant markets and an ability to predict change may be the driving factor. In the public sector, the impetus may come from an awareness that evidence-based policy initiatives are likely to have the greatest impact.

It is now generally recognised that nearly all professional economists, not just those actually working with data, should have a basic understanding of econometrics. There are two major benefits. One is that it facilitates communication between econometricians and the users of their work. The other is the development of the ability to obtain a perspective on econometric work and to make a critical evaluation of it. Econometric work is more robust in some contexts than in others. Experience with the practice of econometrics and a knowledge of the potential problems that can arise are essential for developing an instinct for judging how much confidence should be placed on the findings of a particular study.

Such is the importance of econometrics that, in common with intermediate macroeconomics and microeconomics, an introductory course forms part of the core of any serious undergraduate degree in economics and is a prerequisite for admission to a serious Master's level course in economics or finance.

0.2 Aims and objectives

The aims of this course are to:

- understand the nature and scope of econometrics as a social science
- develop an understanding of the use of regression analysis and related techniques for quantifying economic relationships and testing economic theories
- provide a good foundation for further study in econometrics and statistical techniques
- equip students to read and evaluate empirical papers in professional journals

- provide students with practical experience of using mainstream regression programmes to fit economic models.

0.3 Learning outcomes

At the end of this course, and having completed the Essential reading and activities, you should be able to:

- describe and apply the classical regression model and its application to cross-section data
- describe and apply the:
  - Gauss–Markov conditions and other assumptions required in the application of the classical regression model
  - reasons for expecting violations of these assumptions in certain circumstances
  - tests for violations
  - potential remedial measures, including, where appropriate, the use of instrumental variables
- recognise and apply the advantages of logit, probit and similar models over regression analysis when fitting binary choice models
- competently use regression, logit and probit analysis to quantify economic relationships using R
- describe and explain the principles underlying the use of maximum likelihood estimation
- apply regression analysis to time-series models using stationary time series, with awareness of some of the econometric problems specific to time series applications (for example, autocorrelation) and remedial measures
- recognise the difficulties that arise in the application of regression analysis to nonstationary time series, know how to test for unit roots, and know what is meant by cointegration.

0.4 Overview of the learning resources

0.4.1 Essential reading

The Essential reading for the course is:

- Wooldridge, J. *Introductory Econometrics: A Modern Approach.* (Cengage, 2020) seventh edition [ISBN 9781337558860].

Detailed reading references in this subject guide refer to the edition of the set textbook listed above. A new edition of this textbook may have been published by the time you study this course. You can, however, use a more recent edition; use the detailed chapter and section headings to identify relevant readings. Also check the virtual learning environment (VLE) regularly for updated guidance on readings. You *must* have a copy of the textbook to be able to study this course.

0.4.2 Further reading

Please note that as long as you read the Essential reading you are then free to read around the subject area in any text, paper or online resource. You will need to support your learning by reading as widely as possible and by thinking about how these principles apply in the real world. To help you read extensively, you have free access to the virtual learning environment (VLE) and University of London Online Library (see below).

Other useful texts for this course include:

- Dougherty, C. *Introduction to econometrics.* (Oxford: Oxford University Press, 2016) 5th edition [ISBN 9780199676828].
- Gujarati, D.N. and D.C. Porter *Basic econometrics.* (McGraw–Hill, 2009, International edition) 5th edition [ISBN 9780071276252].
- Maddala, G.S. and K. Lahiri *Introduction to econometrics.* (Wiley, 2009) 4th edition [ISBN 9780470015124].
- Stock, J.H. and M.W. Watson *Introduction to econometrics.* (Pearson, 2012) 3rd edition [ISBN 9781408264331].

0.4.3 The subject guide

The subject guide is intended for you to use in conjunction with the main textbook. It will act to highlight important aspects of the content in the chapters, provide further clarifications where deemed useful and also to extend it in certain places. For the final examination, you will need to be familiar with the material in both the textbook and the subject guide.

The subject guide provides a range of activities that will enable you to test your understanding of the basic ideas and concepts. We want to encourage you to try the exercises you encounter throughout the material before working through the solutions provided on the virtual learning environment (VLE). With econometrics the motto has to be ‘practise, practise, practise…’. It is the best way to learn the material and prepare for examinations. The course is rigorous and demanding, but the skills you will be developing will be rewarding and well recognised by future employers.

Each chapter in the subject guide begins with an overview of the topic covered and a checklist of learning outcomes anticipated as a result of studying the associated, relevant, material in the textbook. The readings from the textbook (by section headings) are supplemented by a discussion and various activities that should be attempted along the way. This includes multiple choice questions (MCQs) that are available on the VLE, exercises from the textbook, and any additional exercises. Worked out solutions for the exercises can be found on the VLE (downloadable pdf files). At the end of each chapter you will find an overview of the material covered, a list of key terms and concepts, and a reminder of the learning outcomes. It concludes with additional examination based exercises that will enable you to test your knowledge and understanding with worked out solutions.

To support students with the pre-requisites for **EC2020 Elements of econometrics**, the first chapter provides a mathematics and statistics refresher. Students are encouraged to review this material and undertake the quizzes on the VLE to assess their preparation before taking the course. Even though we will refer back to this material throughout the course, students will benefit from reviewing this material beforehand. On the VLE you will find accompanying MCQs to test the degree of familiarity as required for the course.

A suggested approach for students studying **EC2020 Elements of econometrics**, in particular independent learners, is to split the material in 10 two-week blocks.

1. Chapter 2 (Nature of econometrics and economic data) and Chapter 3 (Simple regression model).

2. Chapter 4 (Multiple regression analysis: Estimation).

3. Chapter 5 (Multiple regression analysis: Inference).

4. Chapter 6 (Multiple regression analysis: Further issues and use of qualitative information).

5. Chapter 7 (Multiple regression analysis: Asymptotics), and Chapter 8 (Heteroskedasticity).

6. Chapter 9 (Instrumental variable estimation and two-stage least squares).

7. Chapter 10 (Measurement errors) and Chapter 11 (Simultaneous equation models).

8. Chapter 12 (Binary response models and maximum likelihood estimation).

9. Chapter 13 (Regression analysis with time series data) and Chapter 14 (Autocorrelation in time series regression).

10. Chapter 15 (Trends, seasonality, and highly persistent time series in regression analysis).

Throughout the course we discuss many empirical examples that help to motivate each econometric method and should make the course more engaging. To provide you with practical experience we have accompanied the material with an **introduction to R in RStudio** and provide R applications that show how to implement the approaches alongside. The subject guide will refer to this material where appropriate. Although not examinable, it provides you, in our opinion, with valuable transferable skills.

0.4.4 Online study resources

In addition to the subject guide and the Essential reading, it is crucial that you take advantage of the study resources that are available online for this course, including the VLE and the Online Library.

A free online supplement accompanying the textbook is available at http://www.cengagebrain.com. It provides students access to a solutions manual which includes detailed steps and solutions to odd-numbered problems and computer exercises in the text.

You can access the VLE, the Online Library and your University of London email account via the Student Portal at: http://mylondon.ac.uk

You should have received your login details for the Student Portal with your official offer, which was emailed to the address that you gave on your application form. You have probably already logged into the Student Portal in order to register! As soon as you registered, you will automatically have been granted access to the VLE, Online Library and your fully functional University of London email account.

If you have forgotten these login details, please click on the 'Forgotten your password' link on the login page.

0.4.5 The VLE

The VLE, which complements this subject guide, has been designed to enhance your learning experience, providing additional support and a sense of community. It forms an important part of your study experience with the University of London and you should access it regularly.

The VLE provides a range of resources for EMFSS courses:

- **Course materials:** Subject guides and other course materials available for download. In some courses, the content of the subject guide is transferred into the VLE and additional resources and activities are integrated with the text.
- **Readings:** Direct links, wherever possible, to essential readings in the Online Library, including journal articles and ebooks.
- **Video content:** Including introductions to courses and topics within courses, interviews, lessons and debates.
- **Screencasts:** Videos of PowerPoint presentations, animated podcasts and on-screen worked examples.
- **External material:** Links out to carefully selected third-party resources.
- **Self-test activities:** Multiple-choice, numerical and algebraic quizzes to check your understanding.
- **Collaborative activities:** Work with fellow students to build a body of knowledge.

- **Discussion forums:** A space where you can share your thoughts and questions with fellow students. Many forums will be supported by a 'course moderator', a subject expert employed by LSE to facilitate the discussion and clarify difficult topics.
- **Past examination papers:** We provide up to three years of past examinations alongside *Examiners' commentaries* that provide guidance on how to approach the questions.
- **Study skills:** Expert advice on getting started with your studies, preparing for examinations and developing your digital literacy skills.

Note: Students registered for Laws courses also receive access to the dedicated Laws VLE.

Some of these resources are available for certain courses only, but we are expanding our provision all the time and you should check the VLE regularly for updates.

0.4.6 Making use of the Online Library

The Online Library (http://onlinelibrary.london.ac.uk) contains a huge array of journal articles and other resources to help you read widely and extensively.

To access the majority of resources via the Online Library you will either need to use your University of London Student Portal login details, or you will be required to register and use an Athens login.

The easiest way to locate relevant content and journal articles in the Online Library is to use the **Summon** search engine.

If you are having trouble finding an article listed in a reading list, try removing any punctuation from the title, such as single quotation marks, question marks and colons.

For further advice, please use the online help pages (http://onlinelibrary.london.ac.uk/resources/summon) or contact the Online Library team: onlinelibrary@shl.london.ac.uk

Chapter 1
Mathematics and statistics refresher

1.1 Introduction

The material in this chapter permits you to revise some basic mathematical tools crucial for understanding regression analysis. It also provides a refresher of the basic statistical theory that is used throughout the course and is central to econometric analysis.

It should be noted that the material does not replace the prerequisite courses in mathematics, probability and statistics. By studying the essential readings listed, you will revise the material using notation that will be used throughout the course. At the end of each block you will be asked to complete associated quizzes on the VLE that will test your understanding.

1.1.1 Essential reading

Readings: Wooldridge Appendices A, B and C.

1.1.2 Synopsis of this chapter

The refresher is comprised of three modules and covers the key terms and concepts indicated.

A Basic mathematical tools required for understanding regression analysis.

Key terms and concepts:

- Summation Operator
- Average and Median
- Linear Function; Intercept and Slope
- Marginal effect
- Percentage Change, Percentage Point Change, and Proportionate Change
- Nonlinear Function: Quadratic, Exponential and Natural Logarithm
- Diminishing Marginal Effect
- Elasticity and Semi-Elasticity
- Derivative and Partial Derivative.

B Fundamentals of probability required in econometrics.

Key terms and concepts:

- Random Variable (discrete and continuous)
- Probability Density Function (pdf) and Cumulative Distribution Function (cdf)
- Joint Distribution and Independence
- Conditional Distribution
- Expected Value, Median, Variance and Standard Deviation
- Covariance and Correlation
- Conditional Expectation and Law of Iterated Expectation
- Variance of Sums of Random Variables
- Normal Distribution, Standard Normal Distribution, t Distribution, F Distribution and Chi-Squared distribution
- Degrees of Freedom.

C Fundamentals of mathematical statistics required in econometrics.

Key terms and concepts:

- Population versus Sample
- Random Sampling
- Estimator and Estimate
- Unbiasedness of an Estimator and Bias
- Sampling Variance of an Estimator (Efficiency)
- Mean Squared Error (MSE)
- Consistency of an Estimator
- Sufficient conditions of Consistency
- Law of Large Numbers (LLN)
- Asymptotic Normality of an Estimator
- Central Limit Theorem (CLT)
- Estimators: Ordinary Least Squares, Maximum Likelihood Estimation
- Confidence Interval (Interval Estimation)
- Fundamentals of Hypothesis Testing
  - Null and Alternative Hypothesis
  - One-sided and Two-sided Alternative
  - Type I and Type II Error
  - Significance Level
  - Power of a Test

  - Test Statistic
  - Critical Value
- t Test for the Mean of a Normal Population
- p-value of a Test.

1.2 Content of chapter

1.2.1 Basic mathematical tools

Read: Wooldridge Appendix A.

The **summation operator** is used extensively in this course, it provides a convenient shorthand for manipulating expressions involving the sums of many numbers. With $\{x_i : i = 1, 2, \dots, n\}$ denoting a sequence of n numbers:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

It is important to clearly indicate the index, in this case i (which may relate to a particular individual), and the values the index can take, in this case $1, 2, \dots, n$, where n denotes the number of observations (which may be the sample size). We can use the summation notation to describe various descriptive statistics of $\{x_i : i = 1, 2, \dots, n\}$, such as the sample mean:

$$\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

and the sample variance:

$$\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

As we will be using the summation operator a lot, it is important you are comfortable using the summation operator and are aware about its properties. (See Wooldridge property Sum.1–property Sum.3.)

The use of **proportions and percentages** is commonplace in applied econometrics. Mathematically, proportions are the decimal equivalent of a percentage, e.g. when 75% (75 percent) of the population is employed, we say that the proportion of employed individuals in the population equals 0.75. In other words, we need to divide a percentage by 100 to yield the proportion. *When describing changes in variables that are measured in percentages (or proportions), it is important to distinguish clearly between a percentage point change and a percentage change.* For instance if the percentage of employed individuals rises to 80%, this implies we have a **5 percentage point increase** in employment which is not the same as a 5 percent increase! A **5 percent increase** in employment would have yielded $(1 + .05) * 75 = 78.75$, which equals 78.75%. Using percentages (or proportions) to describe changes in variables, such as income, has the advantage that it is free of the units with which the variable is measured (in dollars or pounds). When we do not report changes in this way, we have to clearly indicate the units when describing changes.

We will mostly be working with linear equations (and you should understand their properties). Your ability to help understand modern economic research requires you to be familiar with various nonlinear functions as well. The use of nonlinear functions, such as quadratic forms, is useful as it will permit us to represent **diminishing marginal effects**. If:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

where $\beta_1 > 0$ and $\beta_2 < 0$, then the positive effect that x has on y ($\beta_1 > 0$) is stronger when x is smaller and diminishes thereafter ($\beta_2 < 0$). *The signs of the parameters β_1 and β_2 are crucial in describing this diminishing marginal effect.* Two other nonlinear functions we will encounter are the natural logarithm (or simply log function) and the exponential function. You should remember their properties and recall that, e.g. $\log(\exp(x)) = x$. When we encounter logarithms of variables in applied work, we need to interpret our parameters as elasticities and/or semi-elasticities. We will discuss this in more detail later.

Lastly, the block reviews a bit of **differential calculus**. The marginal effects and partial marginal effects are defined as the derivative of a function with respect to a particular argument, where for the partial marginal effect we need to hold all other arguments constant. Derivatives also play an important role in providing the first-order conditions that define the minimum (when we consider the ordinary least squares estimator) or the maximum (when we consider the maximum likelihood estimator) of a given objective function. *You will be expected to take derivatives of simple functions*, where it will be useful to remember the chain rule. With $f(x)$ and $g(x)$ denoting two functions of the variable x, the chain rule allows us to express:

$$\frac{\mathrm{d}}{\mathrm{d}x} f(g(x)) = f'(g(x))g'(x)$$

where $f'(x) = \mathrm{d}f/\mathrm{d}x$ and $g'(x) = \mathrm{d}g/\mathrm{d}x$.

> **Example 1.1** Let us use the chain rule to derive $\mathrm{d}y/\mathrm{d}x$ when $y = \log(1+2x)$. We first take the derivative of $\log(1+2x)$ with respect to its argument $1+2x$, and then multiply this by the derivative of $1+2x$ with respect to x:
>
> $$\frac{\mathrm{d}y}{\mathrm{d}x} = \frac{\mathrm{d}\log(1+2x)}{\mathrm{d}(1+2x)} \times \frac{\mathrm{d}(1+2x)}{\mathrm{d}x} = \frac{1}{1+2x} \times 2 = \frac{2}{1+2x}.$$

> **Activity 1.1** Take the ‘Quiz on Basic Mathematical tools’ on the VLE to test your understanding.

Fundamentals of probability

Read: Wooldridge Appendix B (you may want to skip Sections B.4e to B.4g during your refresher and return to this at a later stage).

Random variables play an important role in econometrics and it is important to distinguish random variables from realisations of random variables. Random variables take on numerical values that map a particular experiment to its possible outcomes; their outcome is unknown until we have conducted the experiment and may change each time we perform our experiment. Depending on whether the random variable can only take a finite (or countably infinite) number of values or not, we call our random variable a **discrete or continuous random variable**, where for a discrete random variable the probability function provides the probability of each of the possible outcomes. With continuous random variables, the probability of a particular outcome is known to be zero.

Much of our focus in this course will be on explaining continuous random variables, and it is important to recall the importance of the probability density function (pdf) and cumulative distribution function (cdf) to describe probabilities associated with the possible outcomes of continuous random variables. Figure B.2 in Wooldridge provides an example of the pdf of the random variable X, typically denoted $f(x)$, where the area underneath the pdf highlighted indicates the probability that X lies between a and b. As the probability that the random variable X takes on any value equals 1, the area underneath $f(x)$, also denoted $\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x$, is one.

As we will be interested in looking at relations between different random variables, we will want to consider the joint density and conditional density functions. When X and Y are both discrete random variables, the joint density describes the probability of all possible realisations (x, y) when we do an experiment where x and y are obtained at the same time (say we throw a dice and X denotes the outcome and Y is 1 when the outcome is 'even' and 0 otherwise). The random variables X and Y are said to be independent if and only if we can write the joint density, which we may denote as $f_{X,Y}(x, y)$, as the product of the marginals, i.e. $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. The use of subscripts clearly indicates what random variable(s) the probability density function refers to. Convince yourselves that in the example X and Y are clearly not independent.

The conditional distribution, denoted $f_{Y|X}(y \,|\, x)$, plays an important role in allowing us to describe the probability density of the random variable Y for a given value of the random variable X. When Y and X are dependent, the conditional density function will change dependent on what realisation x we want to consider. The usefulness of considering the conditional distribution of Y given X, is that it allows us to ignore the randomness of the random variable X, and treat its outcome as being fixed.

The expectation and variance of a random variable are two important features of the probability density function, describing the measures of central tendency and variability, respectively. The expectation of the random variable Y is denoted $E(Y)$, and:

$$E(Y) = \int_{-\infty}^{\infty} y f(y)\,\mathrm{d}y.$$

The expectation has important properties you need to be familiar with, given in Property E.1–Property E.3 in Wooldridge. We use the expectation operator often in derivations and it is important to recognise that typically (Jensen's inequality):

$$E(g(Y)) \neq g(E(Y))$$

where g is a given function. E.g. $E(Y^2) \neq E(Y)^2$.

The variance describes the spread of the distribution (indicating whether most realisations are tightly centred about the mean or not). The variance of the random variable Y is denoted $Var(Y)$, where

$$Var(Y) = E((Y - E(Y))^2) = E(Y^2) - E(Y)^2.$$

The variance has important properties you need to be familiar with, given in Property VAR.1–Property VAR.4 in Wooldridge, where the latter two describe the variance of sums of random variables. When defining the variance of sums of random variables it is important to know whether the random variables are independent or exhibit covariance/correlation. In Property VAR.3 we permit the two random variables to be dependent, in Property VAR.4 we require them to be independent. In general, when we do permit dependence we can write the variance of a sum of random variables as:

$$Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var(Y_i) + \sum_{i\neq j}^{n} Cov(Y_i, Y_j)$$

where $\sum_{i\neq j}^{n} Cov(Y_i, Y_j)$ is a shorthand for summing the covariances for all i, j pairs, where $i \neq j$. Under independence, this component is zero. Then we get the much simpler expression given in Property VAR.4:

$$Var\left(\sum_{i=1}^{n} Y_i\right) = \sum_{i=1}^{n} Var(Y_i).$$

When using the latter property it is important to clearly indicate that you make use of the assumption that all covariances are zero.

The covariance and correlation are two related measures that reveal the association between two random variables, with:

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)\, Var(Y)}}.$$

Their properties are summarised in Property COV.1–Property COV.3, and Property CORR.1–Property CORR.2 in Wooldridge. As the correlation coefficient is scale invariant, the magnitude of the correlation coefficient is easier to interpret, with perfect linear relationships giving rise to a correlation coefficient equalling 1 or -1.

Important to note: independence guarantees zero covariance and correlation but zero covariance (correlation) does not necessarily imply independence.

The conditional expectation of a random variable, $E(Y \mid X)$, and the conditional variance, $Var(Y \mid X)$ play an important role in our course and we will be discussing them in more detail. Both can formally be defined with the help of the conditional distribution that we defined earlier, but we will restrict our attention to their interpretation instead. As suggested, you may want to return later to look at the material in Sections B.4e to B.4g. The main properties of the conditional expectation are given in Property CE.1–Property CE.5 in Wooldridge. The main gist of its properties are as follows:

CE.1: By conditioning on X you can treat $c(X)$ as being constant.

CE.2: By conditioning on X you can treat $a(X)$ and $b(X)$ as being constant and use the linearity of the expectation operator.

CE.3: Under independence conditioning on X is irrelevant. (The values of Y do not depend on the realisation of X under independence.)

CE.4 (and CE.4′): These are examples of the important property called the 'law of iterated expectation'. As we will see often it is easy to formulate $E(Y \mid X)$. This is a function that will depend on the value of the random variable X. By accounting for all possible values that the random variable X can take, that is by evaluating the expectation of $E(Y \mid X)$, which we write as $E(E(Y \mid X))$, we obtain $E(Y)$.

CE.5 If the mean of Y does not depend on X, then X and Y must be uncorrelated.

You are expected to be familiar with the normal distribution, which has a nice bell-shaped curve that is centred around the mean and has a spread indicated by its variance (see Figure B.7 in Wooldridge). To obtain the standard normal distribution we need to subtract the mean and divide by the standard deviation (Property Normal.1). We typically use the notation $\phi(z)$ to denote the pdf of a $Normal(0, 1)$ distribution and denote its cumulative distribution function (cdf) as $\Phi(z)$. Statistical tables can be used to obtain values of $\Phi(z)$ for a given value of z, see Table G.1 in Wooldridge, whereby $\Phi(0) = 0.5$ due to the symmetry of the normal distribution. Other facts about the normal distribution are summarised in Property Normal.2–Property Normal.4 in Wooldridge.

We will use various distributions derived from the normal distribution in the course, the t distribution, F distribution and the χ^2 distribution. The details are important in this course, but it is good to look at the shape of these distributions. The shape of the t distribution is similar to that of the $Normal(0, 1)$, in comparison it tends to have 'thicker tails' (probability of obtaining realisations that are large (in either tail) is bigger) unless the degrees of freedom are large. Both the F and the χ^2 distributions are examples of skewed (not symmetric) distributions that only take non-negative values. These distributions, as we will see, will play an important role when we conduct hypothesis tests or provide confidence intervals.

> **Activity 1.2** Take the 'Quiz on Fundamentals of probability' on the VLE to test your understanding.

1.2.2 Fundamentals of mathematical statistics

Read: Wooldridge Appendix C (you may want to skip Section C.4 during your refresher and return to this at a later stage).

The goal of an empirical economist is to use a subset of observations (a **sample**) to learn something about the **population** from which the sample was drawn. For instance, we may want to learn what the mean family income is in the UK. Let θ denote the mean family income, and describe the pdf of family income in the UK with pdf $f(y; \theta)$

where Y denotes the random variable 'family income'. θ is an unknown fixed quantity in the population.

The easiest way to come up with a good estimator of θ is to draw independent observations (i.e. with replacement) from this population $f(y;\theta)$ and use a sample analogue of the population mean, the sample average, as its estimator:

$$\widehat{\theta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

where $Y_1, Y_2, \ldots, Y_n$ are n independent random variables drawn from the population that can be represented by the pdf $f(y;\theta)$. We call this a random sample, and we also say $Y_1, Y_2, \ldots, Y_n$ are i.i.d. (short for independent and identically distributed). This estimator $\widehat{\theta}$ is a random variable and we refer to its distribution as the **sampling distribution**. Depending on your particular sample, the estimator will take different values and the sampling distribution reveals information about the likely outcomes of the estimator for different samples. Typically, we only have one sample to our avail, and we will call the realisation of our estimator for this particular sample our estimate.

In the following table we see other examples of population quantities that we may be interested in, and their sample analogues as potential estimators.

Population moments		**Possible estimator**
Mean	$\mu_X = E(X_i)$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
Variance	$\sigma_X^2 = E[(X_i - \mu_X)^2]$ $= Var(X_i)$	$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ $= Sample\,Var(X_i)$
Covariance	$\sigma_{XY} = E[(X_i - \mu_x)(Y_i - \mu_Y)]$ $= Cov(X_i, Y_i)$	$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ $= SampleCov(X_i, Y_i)$
Fixed, unknown quantities		*Random variables*

If $(Y_1, X_1), (Y_2, X_2), \ldots, (Y_n, X_n)$ form a random sample drawn from the joint density $f(y, x;\theta)$, then our descriptive statistics (sample average, sample variance, and sample covariance) indeed form good estimators of their population analogues (mean, variance, and covariance). Associated realisation (estimates) are $\bar{x}$, s_X^2, and s_{XY}.

Common practice: When the parameter of interest is denoted as β_1, we typically will denote its estimator by $\widehat{\beta}_1$. The use of a 'hat' when referring to estimators is commonplace.

Useful finite sample properties of our estimators are that they are **unbiased**, $E(\widehat{\theta}) = \theta$, have a **small sampling variance**, $Var(\widehat{\theta})$, and are **efficient**. It is important to recognise that these are properties that describe the behaviour of our estimator *under repeated sampling*. The unbiasedness property ensures that there is no systematic error when estimating the parameter of interest, and the low sampling variability ensures that there is a high probability that we will obtain a realisation that is close to the parameter of interest. Clearly when choosing between two unbiased estimators, we should select the estimator with the smaller sampling variance (efficiency). When choosing between two estimators that are not necessarily unbiased, we may want to use the mean squared error (MSE) (which offers a trade-off between bias and variance) to compare the efficiency of estimators.

Check! You should be able to show that the sample average of a random sample $Y_1, Y_2, \ldots, Y_n$ from a distribution with mean μ and variance σ^2, is an unbiased estimator of μ, $E(\bar{Y}) = \mu$, whose sampling variance equals $Var(\bar{Y}) = \sigma^2/n$. Clearly, the larger the sample, the smaller the sample variability of our estimator. For efficiency it is never a good idea to throw away information. An estimator of μ that uses only part of the sample has a larger variance.

Useful asymptotic sample properties of our estimators are that they are **consistent** and **asymptotic normal**. We will say that an estimator is consistent if:

$$\text{plim}\ \widehat{\theta} = \theta.$$

To show that an estimator is consistent we can make use of the *sufficient conditions* of consistency ($\lim_{n\to\infty} E(\widehat{\theta}) = \theta$ and $\lim_{n\to\infty} Var(\widehat{\theta}) = 0$). Alternatively, we can rely on *laws of large numbers*. Laws of large numbers (LLNs) give us a convenient asymptotic device that shows that under suitable regularity conditions, sample averages will converge to their population analogues. E.g. plim $\bar{Y} = E(Y_i)$, hence guaranteeing that in the above example $\bar{Y}$ is a consistent estimator of μ as $E(Y_i) = \mu$. In the course we will show how we can use LLNs when considering the consistency of estimators making use of Property PLIM.1–Property PLIM.2 given in Wooldridge. It will be convenient to observe that, under suitable regularity conditions, all estimators have a distribution that can be well-approximated by a normal distribution when our sample is large. This is a powerful result we will rely on if we do not know the exact sampling distribution of an estimator. The asymptotic device that is used to establish this result is the *Central Limit Theorem* (CLT).

Instead of reporting point estimates, we may want to report **confidence intervals**. You want to recall how you can obtain the confidence interval for the mean from a normally distributed population using a random sample. Depending on whether the variance σ^2 is known or not, the form of the $100(1-\alpha)\%$ confidence interval is given by:

$$\left[\bar{y} - z_{\alpha/2}\sqrt{\sigma^2/n}, \bar{y} + z_{\alpha/2}\sqrt{\sigma^2/n}\right]$$

or:

$$\left[\bar{y} - t_{n-1,\alpha/2}\sqrt{s^2/n}, \bar{y} + t_{n-1,\alpha/2}\sqrt{s^2/n}\right]$$

where $z_{\alpha/2} = 1.96$ when $\alpha = 5\%$ (the 97.5th percentile in the $Normal(0,1)$ distribution) and $t_{n-1,\alpha/2}$ when $\alpha = 5\%$ is given by the 97.5th percentile of the t distribution. The limits are given by our estimate plus and minus a measure related to the precision of our estimator (standard deviation / standard error). Typically, the confidence interval is wider when the variance is unknown to account for the additional imprecision. It is important to recognise that these intervals are sample specific. When we obtain a different sample under identical conditions the interval will change, and when we do this repeatedly the interval will contain the true population parameter with probability given by the confidence level, $100(1-\alpha)\%$.

You should revisit the fundamentals of hypothesis testing. To conduct a test we need to specify the null hypothesis (H_0) and the alternative hypothesis (H_1) which are both statements about the true parameter. When we consider a single hypothesis, the alternative hypothesis can be one-sided or two-sided, for multiple hypotheses we only consider two-sided alternatives. You should be familiar with the concepts of a Type I error and a Type II error; these are two types of mistakes that we can make when conducting hypothesis tests due to the fact that the test is based on random variables. Related to these errors we can define the significance level (typically denoted by α) and the power of a test. Typically we want to choose our test in such a way that for a given level of significance we maximise the power of our test. To test our hypothesis we will need to define a suitable test statistic (a random variable that has no unknown quantities **and** has a nice (asymptotic) distribution under the null). This distribution plays an important role as it determines the critical value that defines, for a given level of significance, when we should start rejecting the null. The p-value tells us what the probability is that we will observe a more extreme realisation (averse to the null) of our test statistic. We will apply these ideas regularly in this course, and you should make sure that you are comfortable with the test about the mean in a normal population presented in the book.

> **Activity 1.3** Take the 'Quiz on Fundamentals of mathematical statistics' on the VLE to test your understanding.

Chapter 2

The nature of econometrics and economic data

2.1 Introduction

2.1.1 Aims of the chapter

This is an introductory chapter and it discusses several important general themes underlying any econometric analysis. Particularly important are topics of *ceteris paribus* and counterfactual reasoning.

2.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand steps for empirical analysis
- give an example of how economic theory can lead to an empirical econometric model
- know different types of economic data
- discuss *ceteris paribus*
- discuss counterfactual reasoning.

2.1.3 Essential reading

Readings: Wooldridge Sections 1.1–1.4.

2.1.4 Synopsis of this chapter

In this chapter we discuss the purpose of econometric analysis. Section 1.1 in Wooldridge discusses the scope of econometrics and, in particular, discusses that econometrics is used in all applied economics fields to test economic theories, to conduct empirical policy analysis to inform policymakers and to forecast economic variables in time series. Section 1.2 in Wooldridge discusses and gives examples of how econometric models can be derived from economic theory. Section 1.3 in Wooldridge talks about various types of economic data. Section 1.4 in Wooldridge discusses that finding associations (correlations) between variables may not be the same as finding causal relations. It also talks about the importance of the *ceteris paribus* notion in causal interpretations.

2.2 Content of chapter

2.2.1 What is econometrics?

Read: Wooldridge Section 1.1.

The main issues to learn in this segment are the scope of econometrics, the reliance of econometrics on observational data rather than experimental data, and the reliance on the statistical apparatus.

2.2.2 Steps in empirical economic analysis

Read: Wooldridge Section 1.2.

In this segment it is important to use examples in the textbook to understand and be able to discuss the following general four steps in the empirical analysis:

- Step 1: Carefully pose a question of interest.
- Step 2: Specify an economic or a conceptual model.
- Step 3: Turn the economic model into an econometric model.
- Step 4: Collect data on variables and use statistical methods to estimate parameters, construct confidence intervals, and test hypotheses.

> **Activity 2.1** Take the MCQs related to this section on the VLE to test your understanding.

2.2.3 The structure of economic data

Read: Wooldridge Section 1.3.

Economic data come in several different forms. This segment discusses these forms.

For cross-sectional data, pay particular attention to the random sampling. You also need to be able to discuss situations when this assumption is violated. For time series data the order of observations is very important and it cannot be assumed that outcomes are independent across time. You should be able to give examples of both cross-sectional and time series data.

You can skip Sections 1.3c and 1.3d as in this course we do not deal with pooled cross sections or panel data.

> **Activity 2.2** Take the MCQs related to this section on the VLE to test your understanding.

2.2.4 Causality, *ceteris paribus* and counterfactual reasoning

Read: Wooldridge Section 1.4.

This is the most important segment of this chapter.

The concept of **causality** is key in econometrics. How can we know that more spending causes better student performance (on average)? How can we know another year of education causes an increase in wages (on average)? Finding correlations in data might be suggestive but is rarely conclusive.

The definition of the causal effect of x on y

'How does variable y change if variable x is changed but all other relevant factors are held constant?'

Most of the economic questions that interest policy are **causal** questions. For instance, does an increase in minimum wage change poverty levels and, if yes, to what extent? Crucial to establishing causality is the notion of ***ceteris paribus***: 'all other (relevant) factors being equal. If we succeed – via statistical methods – in holding 'fixed' other relevant factors, then sometimes we establish that changes in one variable (say, education) in fact 'cause' changes in another variable (say, wage).

Most economic questions are *ceteris paribus* questions. You need to understand how the notion of *ceteris paribus* can be described through counterfactual reasoning and give the definition of counterfactual outcomes.

> **Activity 2.3** Exercise 2.2 from Wooldridge.

> **Activity 2.4** The data in ECONMATH were obtained on students from a large university course in introductory microeconomics. For this problem, we are interested in two variables: *score*, which is the final course score, and *econhs*, which is a binary variable indicating whether a student took an economics course in high school.
>
> There are 856 students in the sample. Of these, 317 students report having taken an economics course in high school.
>
> (i) From the data we can find that for those who took an economics course in high school, the average score is 72.08. For those who did not take an economics course in high school, the average score is 72.91. Do these findings necessarily

tell you anything about the causal effect of taking high school economics on college course performance? Explain.

(ii) If you wanted to obtain a good causal estimate of the effect of taking a high school economics course using the difference in averages, what experiment would you run?

2.3 Answers to activities

Solution to Activity 2.3

(i) Here is one way to pose the question: If two firms, say A and B, are identical in all respects except that firm A supplies job training for one hour per worker more than firm B, by how much would firm A's output differ from firm B's?

(ii) Firms are likely to choose job training depending on the characteristics of workers. Some observed characteristics are years of schooling, years in the workforce, and experience in a particular job. Firms might even discriminate based on age, gender, or race. Perhaps firms choose to offer training to more or less able workers, where 'ability' might be difficult to quantify but where a manager has some idea about the relative abilities of different employees. Moreover, different kinds of workers might be attracted to firms that offer more job training on average, and this might not be evident to employers.

(iii) The amount of capital and technology available to workers would also affect output. So, two firms with exactly the same kinds of employees would generally have different outputs if they use different amounts of capital or technology. The quality of managers would also have an effect.

(iv) If you find a positive correlation between *output* and *training*, would you have convincingly established that job training makes workers more productive? Explain.

Solution to Activity 2.4

(i) Comparing averages tells us very little about the causal effect of having taken an economics course in high school on performance in university-level economics. There are many factors that influence a student's performance in an economics course and some of these factors are likely correlated with whether or not they took economics in high school (and subsequently chose to take economics in university). Without controlling for these factors, we cannot really say anything about the causal effect of high school economics exposure on university-level economics performance. Though there is a small (negative) correlation between the two variables, there is little evidence of causation from simply comparing performance between these two groups.

(ii) Ideally we would have an experiment in which we could observe the same individuals in two different scenarios: one in which they took high school economics and one in which they did not. As this is not feasible, we would like to randomly select students into two groups: a control group that did not take economics in high school and a treatment group that did take economics in high school. Even this would be challenging as it would require selection into treatment and control when the students were in high school. Thus, the most feasible experiment would be to compare the difference in scores for students that are as similar as possible in every way except for their exposure to high school economics. Effectively, we would need to control for all of the other factors that could influence both performance in university-level economics and exposure to high school economics.

2.4 Overview of chapter

This chapter discussed the purpose of econometric analysis and how econometric models are derived from a formal economic model. The most common types of data structures used in applied econometrics are cross-sectional, time series, and panel data. In this course we will mostly use cross-sectional data. The definition of cross-sectional data is based on the random sampling concept. We also talked about the problems inherent in drawing causal inferences in the social sciences. We did this mostly through the agricultural yield and return to education examples. These examples also contrasted experimental and nonexperimental (observational) data. We discussed the notions of causality, *ceteris paribus*, and counterfactuals. In most cases, hypotheses in the social sciences are *ceteris paribus* in nature: all other relevant factors must be fixed when studying the relationship between two variables. Because of the nonexperimental nature of most data collected in the social sciences, uncovering causal relationships is very challenging.

2.4.1 Key terms and concepts

- Econometric model
- Economic model
- Empirical analysis
- Cross-sectional data
- Random sampling
- Time series data
- Panel data
- Experimental data
- Nonexperimental data
- Observational data
- Causal effect

- *Ceteris paribus*
- Counterfactual outcomes
- Counterfactual reasoning

2.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand steps for empirical analysis
- give an example of how economic theory can lead to an empirical econometric model
- know different types of economic data
- discuss *ceteris paribus*
- discuss counterfactual reasoning.

Chapter 3
Simple regression model

3.1 Introduction

3.1.1 Aims of the chapter

The purpose of this chapter is to introduce the simple linear regression model and discuss its interpretation and estimation.

3.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand three issues when formulating a simple regression model and discuss how they are addressed
- know the terminology for different components of the regression model
- discuss the key assumptions about how the regressor and error term are related
- derive and plot the population regression function
- derive and interpret OLS estimates in the simple regression model
- define OLS fitted values and residuals and derive their algebraic properties in OLS estimation
- discuss the effect of changing the units of measurement on the parameter estimates and the R^2
- interpret the OLS estimates when the dependent and/or independent variables appear in logarithmic form
- formulate and discuss assumptions of the linear regression model
- establish unbiasedness of the OLS estimator and understand the role of the assumptions in this result
- derive the variance of the OLS estimator under the homoskedasticity assumption and discuss how it is affected by different components.

3.1.3 Essential reading

Readings: Wooldridge Sections 2.1–2.5 and Appendices 2A, A.2, A.3, A.4, B.4a, B.4b, B.4c, B.4e, B.4f.

3.1.4 Synopsis of this chapter

In Section 2.1 in Wooldridge we discuss the simple linear regression model, which is a method that enables a researcher to quantify the relationship between a dependent variable that is assumed to be determined by one explanatory variable. In Sections 2.2 and 2.3 in Wooldridge we show how the method of ordinary least squares (OLS) is used to estimate the parameters in the simple regression model, discuss the interpretation of parameters, statistical properties of the ordinary least squares estimator, and learn how to evaluate the model fit. In Section 2.4 in Wooldridge we address the following two important issues in empirical research. First: how does changing the units of measurement of the dependent and independent variable affect the OLS parameter estimates? Second: how does changing the functional form (in particular, the use of variables in their logarithmic form) affect the interpretation of our parameter estimates? In Section 2.5 in Wooldridge we establish the unbiasedness of the OLS estimators under a series of assumptions and analyse the form of the variance of OLS estimators.

3.2 Content of chapter

3.2.1 Definition of the simple regression model

Read: Wooldridge Section 2.1 and Appendix A.2.

The simple linear regression model studies 'how y varies with changes in x' by writing down a simple equation relating y to x in the population of interest:

$$y = \beta_0 + \beta_1 x + u.$$

The first important thing to note is that in the simple linear regression model the variables y and x are not treated symmetrically. Indeed, we want to **explain y in terms of x** (not explain x in terms of y). The variable y is called the **dependent variable** or **response variable** (among other names). The variable x is called the **regressor** or **explanatory variable** (among other names). The variable u is called the **error term** or **disturbance** (among other names). It is treated as unobserved as it collects all the factors other than x that affect y. We call β_0 the **intercept parameter** and β_1 the **slope parameter**. These describe a population, and our ultimate goal is to estimate them.

Let Δ denote 'change'. Then holding u fixed means $\Delta u = 0$. So:

$$\begin{aligned}\Delta y &= \beta_1 \Delta x + \Delta u \\ &= \beta_1 \Delta x \quad \text{when } \Delta u = 0.\end{aligned}$$

This equation effectively defines β_1 as a slope, with the only difference being the restriction $\Delta u = 0$. Thus, β_1 measures the effect of x on y, holding all other factors (in u) fixed. In other words, the simple linear regression captures a *ceteris paribus* relationship of x on y.

It is important to understand that the simple linear regression (SLR) model is a population model. When it comes to **estimating** β_1 (and β_0) using a random sample of data and, thus, estimating the *ceteris paribus* effect of x on y, we must restrict how u and x are related to each other.

The first step towards formulating any restrictions on how the unobservable u is related to the explanatory variable x is to realise that x and u are viewed as **random variables** having distributions in the population. For example, if $x = cigs$ is the average number of cigarettes the mother smokes per day during pregnancy, then we could figure out its distribution in the population of women who have ever been pregnant. See the histogram of *cigs* obtained from BWGHT data in Figure 3.1. Suppose u includes the overall health of the mother and quality of prenatal care (imagine we want to analyse the causal effect of smoking in pregnancy on a child's birthweight). There is a range of values for these variables in the population and, hence, u also has a distribution in the population.

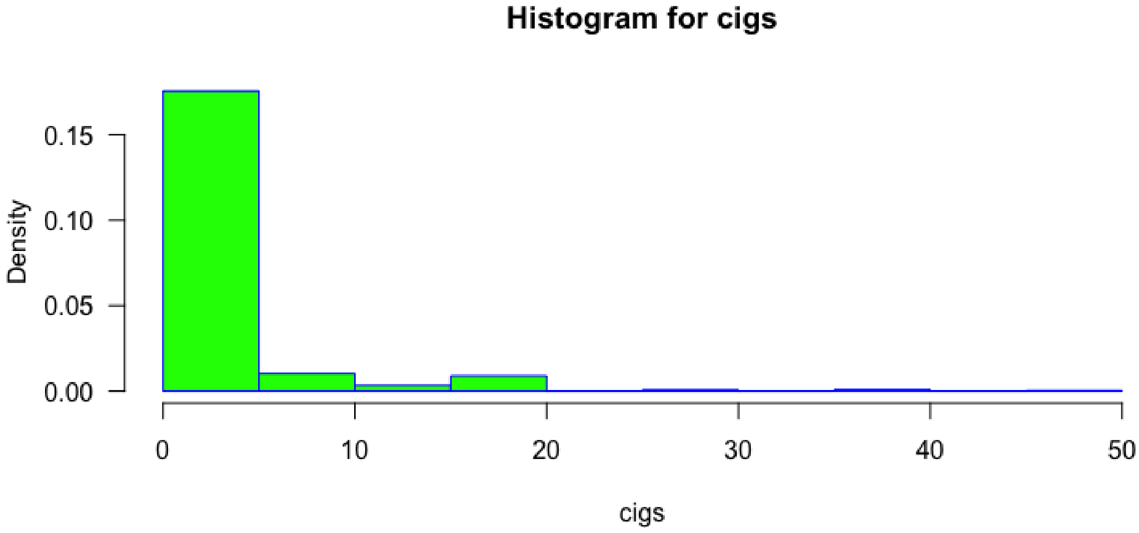


Figure 3.1: Histogram of *cigs* obtained from BWGHT data.

Our first key assumption is that the expected value of u is zero in the population:

$$E(u) = 0$$

where $E(\cdot)$ means the expected value. As long as the intercept β_0 is included in the equation, this assumption is an **innocuous normalisation** and can be imposed without a loss of generality. For example, normalising the mean 'land quality' in the regression of yield on the fertiliser amount or normalising the mean 'ability' in the regression of wage on education to be zero in the population should be harmless. It is. Let us show mathematically that this assumption is an innocuous normalisation.

Suppose $\alpha_0 = E(u) \neq 0$, then we can write:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u \\ &= (\beta_0 + \alpha_0) + \beta_1 x + (u - \alpha_0) \\ &= \beta_0^* + \beta_1 x + u^* \end{aligned}$$

where the new error term $u^* = u - \alpha_0$ satisfies $E(u^*) = 0$:

$$\begin{aligned} E(u^*) &= E(u - \alpha_0) \\ &= E(u) - E(\alpha_0) \\ &= \alpha_0 - \alpha_0 \\ &= 0. \end{aligned}$$

The new intercept is $\beta_0^* = \beta_0 + \alpha_0$, and, importantly, the **slope $\boldsymbol{\beta_1}$ does not change**. Thus, if the average of u is different from zero, we just adjust the intercept, leaving the slope the same. The presence of β_0 in the equation allows us to do that.

The second key assumption is on the expected value of u given x or, in other words, on the mean of the error term for each slice of the population determined by values of x, that is:

$$E(u \,|\, x) = E(u) \quad \text{for all values of } x$$

where $E(u \,|\, x)$ means 'the expected value of u given x'. In other words, under this assumption the error term has the same expectation given any value of the explanatory variable. We say u is **mean independent** of x. Suppose u is 'ability' and x is years of education. For the mean independence assumption to hold, we need, for example:

$$E(\mathit{ability} \,|\, x = 8) = E(\mathit{ability} \,|\, x = 12) = E(\mathit{ability} \,|\, x = 16)$$

so that the average ability is the same in the different portions of the population with an 8th grade education, a 12th grade education, and a four-year college education. Because people choose education levels partly based on ability, this assumption is almost certainly false. To give another example, suppose u is 'land quality' and x is fertiliser amount. Then $E(u \,|\, x) = E(u)$ if fertiliser amounts are chosen independently of quality. This assumption is reasonable but assumes fertiliser amounts are assigned at random.

Combining $E(u \,|\, x) = E(u)$ (the substantive assumption) with $E(u) = 0$ (normalisation) gives:

$$E(u \,|\, x) = 0 \quad \text{for all values of } x.$$

This is called the **zero conditional mean assumption**. It says that the expected value of the error term, u, is zero, regardless of what the value of the explanatory variable, x, is.

By properties of the conditional expectation, $E(u \,|\, x) = 0$ implies:

$$\begin{aligned} E(y \,|\, x) &= E(\beta_0 + \beta_1 x + u \,|\, x) \\ &= \beta_0 + \beta_1 x + E(u \,|\, x) \\ &= \beta_0 + \beta_1 x \end{aligned}$$

which shows the **population regression function**, $E(y \,|\, x)$, is a linear function of x. Thus, regression analysis is essentially about explaining effects of explanatory variables on average outcomes of y.

> **Activity 3.1** Take the MCQs related to this section on the VLE to test your understanding.

3.2.2 Deriving the Ordinary Least Squares estimates

Read: Wooldridge Section 2.2, Appendix 2A and Appendices B.4a, B.4b, B.4c.

Suppose one has a sample $\{(x_i, y_i) : i = 1, 2, \ldots, n\}$ of size n (the number of observations) from the population. Think of this as a random sample. We must next decide how to use the data to obtain estimates of the intercept and slope in the population regression of y on x.

The estimation procedure we are going to propose is based on two conditions (in the population) for the parameters β_0 and β_1. To obtain these two estimating equations for β_0 and β_1, we use restrictions:

$$E(u) = 0 \tag{3.2.1}$$

$$Cov(x, u) = 0. \tag{3.2.2}$$

Remember, condition (3.2.1) is innocuous. Condition (3.2.2) stated in terms of the covariance, means that x and u are **uncorrelated** (indeed, it can be equivalently written as $Corr(x, u) = 0$). With $E(u) = 0$, $Cov(x, u) = 0$ is the same as $E(xu) = 0$ because:

$$Cov(x, u) = E(xu) - E(x)\,E(u) = E(xu) - E(x) \cdot 0 = E(xu).$$

Both (3.2.1) and (3.2.2) are implied by the zero conditional mean assumption:

$$E(u \,|\, x) = 0.$$

Next we plug in u (recall from the regression equation that $u = y - \beta_0 - \beta_1 x$) into the restrictions (3.2.1) and (3.2.2):

$$E(y - \beta_0 - \beta_1 x) = 0$$

$$E[x(y - \beta_0 - \beta_1 x)] = 0.$$

These are the two conditions in the population that determine β_0 and β_1. Remember, however, that we do **not** observe everyone in the population. This means we do **not** know the population distribution of (y, x) and cannot compute expectations $E(\cdot)$ in the two conditions. What we have is a random sample from the population (which is just a subset). Therefore, we use the **sample analogs** of the expectations to estimate β_0 and β_1, which is a **method of moments** approach to estimation.

The main idea of the method of moments in general is to estimate the expected value '$E(\cdot)$' by the respective sample average '$\frac{1}{n}\sum_{i=1}^{n}\cdot$' (often called the sample analog). Therefore, in order to determine the estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we use the sample analogs:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) = 0$$

which result in two linear equations for two unknowns $\widehat{\beta}_0$ and $\widehat{\beta}_1$.

Solving these two equations with respect to $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we obtain:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sum_{i=1}^{n} x_i(x_i - \bar{x})} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\text{Sample Covariance}(x_i, y_i)}{\text{Sample Variance}(x_i)}$$

and:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

if $\sum_{i=1}^{n} x_i(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$ (that is, if the sample variance of the x_is is not zero, which only rules out the case where each x_i is the same value).

$\widehat{\beta}_0$ and $\widehat{\beta}_1$ are called the **ordinary least squares (OLS)** estimates.

The **fitted value** (or **predicted value**) for each data point i is:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

The difference between the observed and the fitted value is the **residual**:

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i.$$

Residuals show how well our model explains the real observations.

The name 'ordinary least squares' for $\widehat{\beta}_0$ and $\widehat{\beta}_1$ comes from the fact that $\widehat{\beta}_0$ and $\widehat{\beta}_1$ solve the optimisation problem:

$$\min_{b_0, b_1} \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2.$$

The objective function in this optimisation problem is called the **sum of squared residuals** (imagine we consider some generic b_0 and b_1 as parameters in calculating fitted values and residuals) and it is implied that we measure the prediction quality, for each i, by squaring the residual. Thus, the OLS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ minimise the sum of squared residuals. You need to study Appendix 2A for more technical details.

For OLS estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$, the **OLS regression line** as a function of x is:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

The OLS regression line allows us to predict y for any (sensible) value of x. It is also called the **sample regression function**. The slope, $\widehat{\beta}_1$, allows us to predict changes in y for any (reasonable) change in x, that is:

$$\Delta\widehat{y} = \widehat{\beta}_1 \Delta x.$$

With the help of various empirical examples you learn how to interpret the results of OLS estimation. In other words, you should be able to interpret findings given by the OLS regression line.

> **Activity 3.2** Take the MCQs related to this section on the VLE to test your understanding.

3.2.3 Properties of OLS on any sample of data

Read: Wooldridge Section 2.3.

You need to know how one can calculate the OLS fitted values and OLS residuals from OLS estimates and the data. It is important to remember that the *residual* $\widehat{u}_i$ is **different** from the *error* $u_i = y_i - \beta_0 - \beta_1 x_i$ (it is a common mistake by students to confuse these two).

It is useful to know and be able to explain and interpret the following algebraic properties of OLS statistics:

- OLS residuals always add up to zero:
$$\sum_{i=1}^{n} \widehat{u}_i = 0.$$
- Sample covariance (and thus sample correlation) between regressor and residual is always zero:
$$\sum_{i=1}^{n} x_i \widehat{u}_i = 0.$$
- Sample average of actual y_i is the same as sample average of the fitted values $\widehat{y}_i$:
$$\bar{y} = \bar{\widehat{y}}.$$
- Sample covariance of fitted values and residuals is zero:
$$\sum_{i=1}^{n} \widehat{y}_i \widehat{u}_i = 0.$$
- Point $(\bar{x}, \bar{y})$ is on the OLS regression line:
$$\bar{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}.$$

The analysis of residuals is useful in evaluating whether x is a good predictor of y. This segment introduces a formal common measure that measures how well x predicts y based on the OLS estimation. This measure is called the ***R*-squared** of the regression and is denoted as R^2:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

(assuming $SST > 0$), where:

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2$$

$$SSR = \sum_{i=1}^{n} \widehat{u}_i^2.$$

You should be able to show that:

$$SST = SSE + SSR.$$

Thus, R^2 is the fraction of the total variation in y_i that is explained by x_i (or the OLS regression line). You need to know and understand the following properties of R-squared:

- By construction it always holds that $0 \leq R^2 \leq 1$.
- $R^2 = 0$ means no linear relationship between y_i and x_i.
- $R^2 = 1$ means a perfect linear relationship.
- As R^2 increases, the y_is are closer and closer to falling on the OLS regression line.

However, you need to keep in mind that we do not want to fixate on R^2. It is a useful summary measure of the predictive quality of the OLS regression line but tells us nothing about causality. Having a 'high' R-squared is neither necessary nor sufficient to infer causality.

> **Activity 3.3** Take the MCQs related to this section on the VLE to test your understanding.

3.2.4 Units of measurement and functional form

Read: Wooldridge Section 2.4 and Appendices A.3–A.4.

With the help of various empirical examples you learn how both units of measurement and the functional form affect the interpretation of your parameters in the linear regression model. The fact that units of measurement affect the parameters should not be surprising when we consider a setting where we are interested in explaining the cost of travel as a function of distance travelled:

$$travelcost = \beta_0 + \beta_1 distance + u.$$

If we measure distance travelled in miles and cost of travel in UK pounds, we should interpret β_1 as the additional cost in pounds for an additional mile travelled, whereas if we measure distance travelled in kilometres and the cost of travel in euros, we should interpret β_1 as the additional cost in euros for an additional kilometre travelled. This difference in their interpretation explains why β_1 and our estimates thereof will be affected by the units of measurement. To understand the impact on β_0 and its estimate we simply consider its interpretation in the setting where the distance travelled equals zero.

The mathematical details of the impact on $\widehat{\beta}_0$ and $\widehat{\beta}_1$, when the units of measurement of y and x change should be understood as well.

You should understand how the interpretation of parameters is affected when our dependent and/or independent variables appear in logarithmic form. In particular, the interpretation of parameters in a log-log model have the nice interpretation of representing an elasticity, a well-recognised concept in economics. The following table summarises the interpretation of the slope parameter in different specifications involving the use of logarithms. You should study this table carefully as we will regularly make use of it when interpreting parameters in empirical research and will do so throughout this course.

Model	Dependent Variable	Regressor	Interpretation of β_1	
Level-Level	y	x	$\beta_1 = \dfrac{\Delta y}{\Delta x}$	If you change x by 1 unit, we expect y to change by β_1 units
Level-Log	y	$\log(x)$	$\dfrac{\beta_1}{100} = \frac{\Delta y}{\%\Delta x}$	If you change x by 1%, we expect y to change by $\beta_1/100$ units
Log-Level	$\log(y)$	x	$100\beta_1 = \dfrac{\%\Delta y}{\Delta x}$	If you change x by 1 unit, we expect y to change by $100\beta_1\%$
Log-Log	$\log(y)$	$\log(x)$	$\beta_1 = \dfrac{\%\Delta y}{\%\Delta x}$	If you change x by 1%, we expect y to change by $\beta_1\%$

Activity 3.4 Take the MCQs related to this section on the VLE to test your understanding.

Activity 3.5 Exercise 2.6 from Wooldridge.

Activity 3.6 Exercise 2.9 (i) from Wooldridge. You should recall that:

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x} \quad \text{and} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Activity 3.7 Exercise 2.9 (iv) from Wooldridge.

3.2.5 Expected values and variances of the OLS estimators

Read: Wooldridge Section 2.5.

The estimated regression coefficients are random variables because they are calculated from a random sample. This segment studies statistical properties of the OLS estimator, referring to a population model and assuming random sampling.

This segment starts with answering the following question posed by mathematical statistics: How do our estimators behave across different samples of data? On average, would we get the right answer if we could repeatedly sample? (Usually we only have one particular sample, so this is a hypothetical question.) This question is answered by finding the expected value of the OLS estimators. As we know, if the average outcome of an estimator across all possible random samples gives us the truth, this means that the estimator is unbiased. This segment establishes the unbiasedness of OLS estimators under a simple set of assumptions (SLR.1–SLR.4). You need to understand these assumptions and be able to discuss their role in the proof of unbiasedness. Assumption SLR.4 (zero conditional mean) is the key assumption for showing that OLS is unbiased. Note, however, that we can compute the OLS estimates whether or not this assumption holds, or even if there is an underlying population model. However, SLR.4 will guarantee good properties of OLS estimation. If this assumption fails, then OLS will be biased for β_1. For example, if an omitted factor contained in u is correlated with x, then SLR.4 typically fails.

It is also important to understand that unbiasedness is a property of an **estimator** (not an estimate)! After estimation like:

$$\widehat{wage} = -5.12 + 1.43educ$$

it may be tempting to say 1.43 is an 'unbiased estimate' of the education impact on wages, but technically this statement is incorrect. It is the **estimator (or the rule or recipe) used to get $\boldsymbol{\widehat{\beta}_1 = 1.43}$ that is unbiased** (under assumptions SLR.1–4).

The proof of OLS unbiasedness relies on the mathematical tool called the **law of iterated expectation**.

If we consider two random variables Z and X, then the law of iterated expectation says that:

$$E[E(Z \mid X)] = E(Z).$$

On the left-hand side, we take expectation twice. First we obtain $E(Z \mid X) = m(X)$ as a function of X. Then we take the expectation $E[m(X)]$. Sometimes to emphasise that the second expectation is taken with respect to the distribution of X we can alternatively write that:

$$E_X[E(Z \mid X)] = E(Z).$$

The proof of OLS unbiasedness first establishes that for any $\mathbf{X} = \{x_1, \ldots, x_n\}$, we get the same conditional expectation:

$$E(\widehat{\beta}_1 \mid \mathbf{X}) = \beta_1$$

and then use the law of iterated expectation to obtain that:

$$E(\widehat{\beta}_1) = E(E(\widehat{\beta}_1 \,|\, \mathbf{X})) = \beta_1.$$

In the proof of unbiasedness you should be able to explain every small step – whether it follows from one of the Assumptions SLR.1–SLR.4 (and indicate which one) or whether it follows from the properties of the conditional expectation.

Remember that unbiasedness is a finite sample property of the estimator. In other words, it is a feature of the sampling distributions of $\widehat{\beta}_1$ and $\widehat{\beta}_0$ and it does not say anything about how close $\widehat{\beta}_1$ and $\widehat{\beta}_0$ are to β_1 and β_0, respectively, in a given sample. We hope that, if the sample we obtain is somehow 'typical', then our estimate should be 'near' the population value. Unfortunately, it is always possible that we could obtain an unlucky sample that would give us a point estimate $\widehat{\beta}_1$ far from β_1, and relying on the unbiasedness property alone we can never know for sure whether this is the case. Let us illustrate this in simulations.

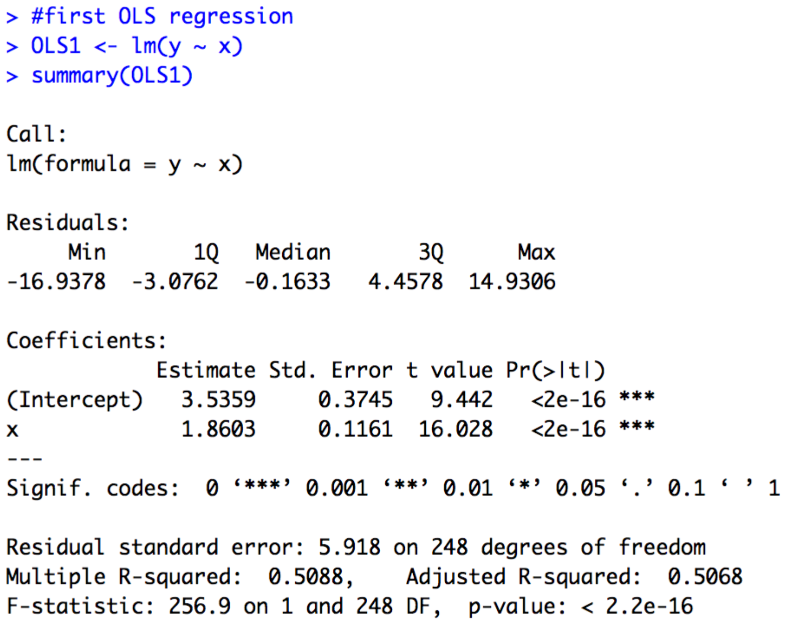
First, we draw 250 observations of x and u independently of each other and according to the distributions:

$$x \sim Normal(0, 9) \quad \text{and} \quad u \sim Normal(0, 36).$$

Then, we calculate y as:

$$y = 3 + 2x + u.$$

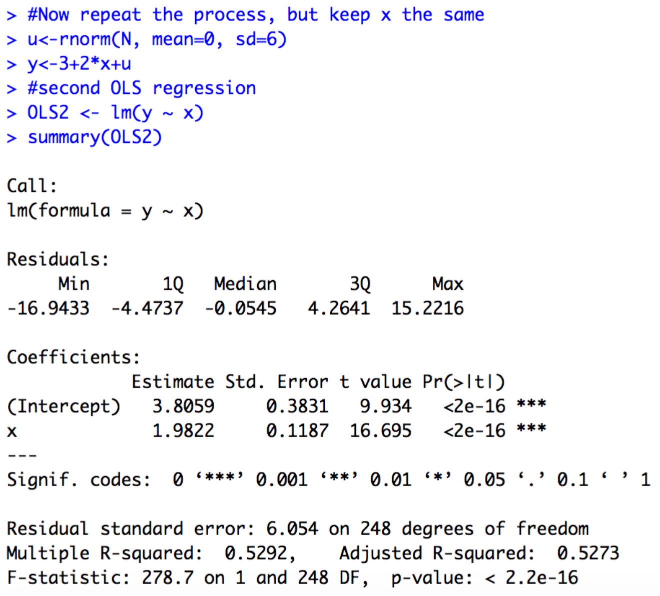
OLS estimation results in four different samples, given in Figures 3.2 to 3.5.



```
> #first OLS regression
> OLS1 <- lm(y ~ x)
> summary(OLS1)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.9378  -3.0762  -0.1633   4.4578  14.9306 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.5359     0.3745   9.442   <2e-16 ***
x             1.8603     0.1161  16.028   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.918 on 248 degrees of freedom
Multiple R-squared:  0.5088,	Adjusted R-squared:  0.5068 
F-statistic: 256.9 on 1 and 248 DF,  p-value: < 2.2e-16
```

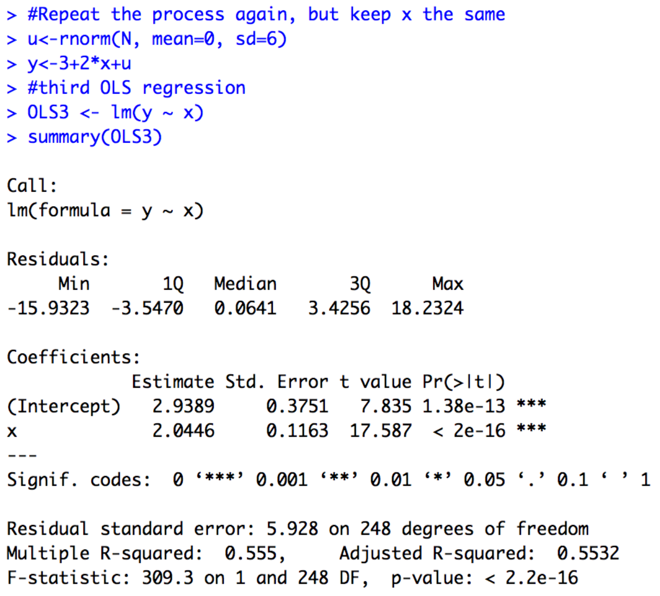
Figure 3.2: OLS estimation results from the first generated data set.

The second and fourth generated data sets give us $\widehat{\beta}_1 \approx 1.98$, very close to $\beta_1 = 2$. The first data set gives us $\widehat{\beta}_1 \approx 1.86$, which is a bit far. If we repeat the experiment again and again and then average the outcomes of $\widehat{\beta}_1$, we would get very close to 2. The problem is that we do not know which kind of sample we have. We never know whether we are close to the population value. Generally, we hope that our sample is 'typical' and produces a slope estimate close to β_1, but we never know.



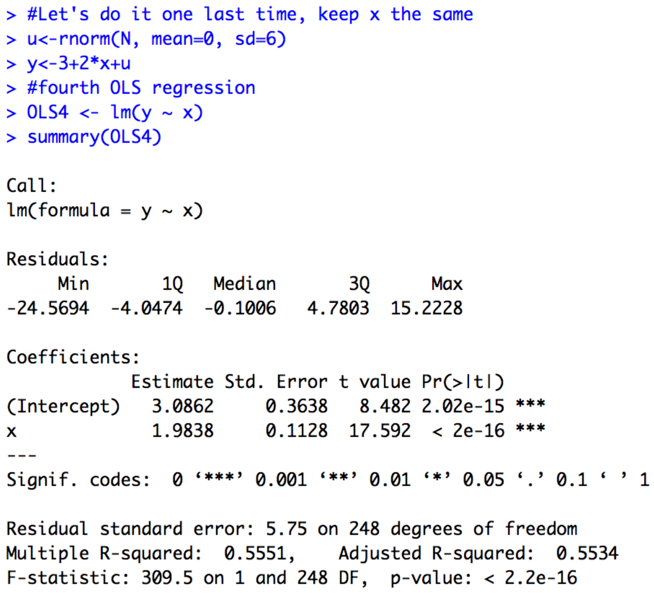
```
> #Now repeat the process, but keep x the same
> u<-rnorm(N, mean=0, sd=6)
> y<-3+2*x+u
> #second OLS regression
> OLS2 <- lm(y ~ x)
> summary(OLS2)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-16.9433  -4.4737  -0.0545   4.2641  15.2216 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.8059     0.3831   9.934   <2e-16 ***
x             1.9822     0.1187  16.695   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.054 on 248 degrees of freedom
Multiple R-squared:  0.5292,    Adjusted R-squared:  0.5273 
F-statistic: 278.7 on 1 and 248 DF,  p-value: < 2.2e-16
```

Figure 3.3: OLS estimation results from the second generated data set.



```
> #Repeat the process again, but keep x the same
> u<-rnorm(N, mean=0, sd=6)
> y<-3+2*x+u
> #third OLS regression
> OLS3 <- lm(y ~ x)
> summary(OLS3)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.9323  -3.5470   0.0641   3.4256  18.2324 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.9389     0.3751   7.835 1.38e-13 ***
x             2.0446     0.1163  17.587  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.928 on 248 degrees of freedom
Multiple R-squared:  0.555,     Adjusted R-squared:  0.5532 
F-statistic: 309.3 on 1 and 248 DF,  p-value: < 2.2e-16
```

Figure 3.4: OLS estimation results from the third generated data set.



```
> #Let's do it one last time, keep x the same
> u<-rnorm(N, mean=0, sd=6)
> y<-3+2*x+u
> #fourth OLS regression
> OLS4 <- lm(y ~ x)
> summary(OLS4)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max 
-24.5694  -4.0474  -0.1006   4.7803  15.2228 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   3.0862     0.3638   8.482 2.02e-15 ***
x             1.9838     0.1128  17.592  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.75 on 248 degrees of freedom
Multiple R-squared:  0.5551,	Adjusted R-squared:  0.5534 
F-statistic: 309.5 on 1 and 248 DF,  p-value: < 2.2e-16
```

Figure 3.5: OLS estimation results from the fourth generated data set.

Even though under SLR.1 to SLR.4, the OLS estimators are unbiased, depending on the sample, the estimates will be nearer or farther away from the true population values. How far we can expect our estimates to be away from the true population values depends on the variance (and, ultimately, standard deviation) of the estimators. This segment characterises the variance of the OLS estimators under SLR.1 to SLR.4 and under an additional Assumption SLR.5 (homoskedasticity) which simplifies the calculations.

Homoskedasticity requires $Var(u \,|\, x) = \sigma^2$ and implies that:

$$Var(y \,|\, x) = \sigma^2$$

that is, that the variance of y does not change with x. It is important to keep in mind that the constant variance assumption may not be realistic; it must be determined on a case-by-case basis (it will be discussed later in the course how to use the data to test whether SLR.5 holds and how to proceed without it). Consider the following two examples of when Assumption SLR.5 is violated.

> **Example 3.1** Suppose $y = sav$, $x = inc$ and we think:
>
> $$E(sav \,|\, inc) = \beta_0 + \beta_1 inc$$
>
> with $\beta_1 > 0$. This means average family saving increases with income. If we impose SLR.5, then:
>
> $$Var(sav \,|\, inc) = \sigma^2$$
>
> which means the variability in saving does not change with income. There are reasons to think saving would be more variable as income increases.

> **Example 3.2** Suppose y is an average score, at the school level, and x is school enrolment. The variance of an average is inversely related to the number of units being averaged. In fact:
>
> $$Var(y \,|\, x) = \frac{\eta^2}{x}$$
>
> often holds, where η^2 is the variance in individual scores.

Under Assumptions SLR.1–SLR.5 we have:

$$Var(\widehat{\beta}_1 \,|\, \mathbf{X}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$Var(\widehat{\beta}_0 \,|\, \mathbf{X}) = \frac{\sigma^2 \left(n^{-1} \sum_{i=1}^n x_i^2\right)}{SST_x}.$$

The textbook by Wooldridge denotes $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ as '$Var(\widehat{\beta}_1)$ conditional on $\{x_1, \ldots, x_n\}$'. These are the 'standard' formulae for the variance of the OLS estimators. They are **not** valid if Assumption SLR.5 does not hold.

It is also important to understand that the homoskedasticity assumption was **not** used to show unbiasedness of the OLS estimators. That requires only SLR.1 to SLR.4.

You should be able to derive the above formula for $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ and explain every step in the derivation by either referring to Assumptions SLR.1–SLR.5 or to the properties of the variance. You are also expected to be able to interpret the formula for $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$.

- As the error variance increases, that is, as σ^2 increases, so does $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$. The more 'noise' in the relationship between y and x – that is, the larger the variability in u – the harder it is to learn about β_1.
- By contrast, more variation in $\{x_i\}$ is a **good** thing:

$$SST_x \uparrow \text{ implies } Var(\widehat{\beta}_1 \,|\, \mathbf{X}) \downarrow .$$

Notice that SST_x/n is the sample variance in x. We can think of this as getting close to the population variance of x, σ_x^2, as n gets large. This means:

$$SST_x \approx n\sigma_x^2$$

which means that, as n grows, $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ shrinks at the rate $1/n$. This is why more data is a good thing: **more data shrinks the sampling variance of our estimators**.

The **standard deviation** of $\widehat{\beta}_1$ is the square root of the variance. So:

$$sd(\widehat{\beta}_1) = \frac{\sigma}{\sqrt{SST_x}}.$$

Since σ^2 is not known, $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ (or $sd(\widehat{\beta}_1)$) cannot be computed. However, we can find a way to *estimate* σ^2. The estimator of σ^2 used universally is:

$$\widehat{\sigma}^2 = \frac{SSR}{n-2} = (n-2)^{-1} \sum_{i=1}^n \widehat{u}_i^2.$$

This estimator is unbiased under Assumptions SLR.1 to SLR.5. In regression outputs in various regression packages, it is:

$$\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} = \sqrt{\frac{SSR}{n-2}}$$

that is usually reported. This is an estimator of $sd(u)$, the standard deviation of the population error. One small glitch is that $\widehat{\sigma}$ is not unbiased for σ. This will not matter for our purposes as it is very close to σ for a large enough sample size (and there is no unbiased estimator of σ). $\widehat{\sigma}$ is called the **standard error of the regression**, which means it is an estimate of the standard deviation of the error in the regression.

Given $\widehat{\sigma}$, we can now estimate the standard deviations $sd(\widehat{\beta}_1)$ and $sd(\widehat{\beta}_0)$. The estimates of these are called the **standard errors** of $\widehat{\beta}_1$ and $\widehat{\beta}_0$, respectively. You will use these a lot. In order to find these estimates, we just plug $\widehat{\sigma}$ in for σ. E.g. for $\widehat{\beta}_1$:

$$se(\widehat{\beta}_1) = \frac{\widehat{\sigma}}{\sqrt{SST_x}}$$

where both the numerator and denominator are easily computed from the data.

Activity 3.8 Show that under Assumptions SLR.1 through SLR.4 we have:

$$E(\widehat{\beta}_0) = \beta_0.$$

Instructions. You may find it helpful to proceed using the following steps.

1. Write down the formula for $\widehat{\beta}_0$.
2. Obtain the expression for $\bar{y}$ from the regression equation $y_i = \beta_0 + \beta_1 x_i + u_i$, for $i = 1, \ldots, n$. Plug that expression for $\bar{y}$ into the formula for $\widehat{\beta}_0$.
3. Find the conditional expectation $E(\widehat{\beta}_0 \,|\, \mathbf{X})$. You can use the fact that it is already known that under Assumptions SLR.1–SLR.4:

$$E(\widehat{\beta}_1 \,|\, \mathbf{X}) = \beta_1.$$

4. Apply the law of iterated expectation.

Activity 3.9 Take the MCQs related to this section on the VLE to test your understanding.

3.2.6 Introduction to R in RStudio

Complete ‘R Task – Activity 1’ on the VLE.

Here we will provide instructions on how to install R and RStudio on your computer. We will be discussing how to implement econometrics using real data with R throughout this course. It is a great transferable skill to have. You will not be tested on your ability to use R in the examination. You should be able to read regression output.

3.3 Answers to activities

Solution to Activity 3.5

(i) As this is a log-log model, we need to interpret the parameter as the elasticity of *price* with respect to *dist*. By increasing the house's distance to a recently built garbage incinerator by 1%, its price can be expected to increase by .312%. We clearly would expect this sign, as living closer to an incinerator depresses housing prices.

(ii) For the *ceteris paribus* condition to hold we need to ensure that by changing the location of the incinerator all other factors that affect the price remain unchanged. Clearly this is not true if cities decide to locate the incinerator in areas that are not close to expensive neighbourhoods. We should therefore be concerned that SLR.4 is not satisfied.

(iii) There are many factors that could affect the price of houses, such as the size of the house, the number of bathrooms, age of the home, and the quality of the neighbourhood (including school quality). These variables could easily be correlated with *dist* and log(*dist*).

Solution to Activity 3.6

We start by specifying the two models we want to consider. The model with the variables in their original units of measurement:

$$y = \beta_0 + \beta_1 x + u$$

and the model where both the dependent and independent variables have different units of measurement:

$$y^* = \beta_0 + \beta_1 x^* + u \quad \text{where} \quad y_i^* = c_1 y_i \quad \text{and} \quad x_i^* = c_2 x_i$$

where c_1 and c_2, with $c_2 \neq 0$, are known constants.

- **We start by looking at the slope.** With the new measurement, the estimator of the slope is:
$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^n (x_i^* - \bar{x}^*)(y_i^* - \bar{y}^*)}{\sum_{i=1}^n (x_i^* - \bar{x}^*)^2}.$$
We need to replace $y_i^* = c_1 y_i$ and $x_i^* = c_2 x_i$.
Notice:
$$\bar{y}^* = \frac{1}{n}\sum_{i=1}^n y_i^* = \frac{1}{n}\sum_{i=1}^n (c_1 y_i) = c_1 \frac{1}{n}\sum_{i=1}^n y_i = c_1 \bar{y}.$$
Therefore $\bar{y}^* = c_1 \bar{y}$ and similarly $\bar{x}^* = c_2 \bar{x}$.
Using $y_i^* = c_1 y_i$, $x_i^* = c_2 x_i$, $\bar{y}^* = c_1 \bar{y}$, and $\bar{x}^* = c_2 \bar{x}$, we obtain:
$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^n (c_2 x_i - c_2 \bar{x})(c_1 y_i - c_1 \bar{y})}{\sum_{i=1}^n (c_2 x_i - c_2 \bar{x})^2} = \frac{\sum_{i=1}^n c_2 (x_i - \bar{x}) c_1 (y_i - \bar{y})}{\sum_{i=1}^n c_2^2 (x_i - \bar{x})^2}.$$

We can take c_1 and c_2 again outside the summation expressions to yield:

$$\widetilde{\beta}_1 = \frac{c_2 c_1 \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{c_2^2 \sum_{i=1}^n (x_i - \bar{x})^2}$$

which can be simplified to yield $\widetilde{\beta}_1 = \dfrac{c_1}{c_2}\widehat{\beta}_1$ as required.

- **Next we look at the intercept.** With the new measurement, the estimator of the intercept is:

  $$\widetilde{\beta}_0 = \bar{y}^* - \widetilde{\beta}_1 \bar{x}^*.$$

  Let us replace $y_i^* = c_1 y_i$ and $x_i^* = c_2 x_i$ again.

  This gives us:

  $$\widetilde{\beta}_0 = c_1 \bar{y} - \widetilde{\beta}_1 (c_2 \bar{x}) = c_1 \bar{y} - c_2 \widetilde{\beta}_1 \bar{x}.$$

  To show our result, we use our previous result $\widetilde{\beta}_1 = (c_1/c_2)\widehat{\beta}_1$:

  $$\widetilde{\beta}_0 = c_1 \bar{y} - \frac{c_1}{c_2}\widehat{\beta}_1 c_2 \bar{x} = c_1(\bar{y} - \widehat{\beta}_1 \bar{x}) = c_1 \widehat{\beta}_0.$$

Solution to Activity 3.7

In this problem, we are asked to compare the parameter estimates of the following two log-linear models: the model where y is in its original unit of measurement:

$$\log y = \beta_0 + \beta_1 x + u$$

where $y > 0$, and the model where we use a different unit of measurement:

$$\log y^* = \beta_0 + \beta_1 x + u \quad \text{with} \quad y_i^* = c_1 y_i$$

where $c_1 > 0$ is a known constant.

- **We start by looking at the slope.**

  With the new measurement, the estimator of the slope is:

  $$\widetilde{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(\log(y_i^*) - \overline{\log(y^*)})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

  where $\overline{\log(y^*)} = \dfrac{1}{n}\sum_{i=1}^n \log(y_i^*)$.

  Let us replace $y_i^* = c_1 y_i$.

  Using the properties of logarithms:

  - $\log(y_i^*) = \log(c_1 y_i) = \log(c_1) + \log(y_i)$
  - $\overline{\log(y^*)} = \frac{1}{n}\sum_{i=1}^n log(c_1 y_i) = \frac{1}{n}\sum_{i=1}^n [\log(c_1) + \log(y_i)] = \log(c_1) + \overline{\log(y)}$.

  Hence $\log(y_i^*) - \overline{\log(y^*)} = \log(y_i) - \overline{\log(y)}$, and we obtain $\widetilde{\beta}_1 = \widehat{\beta}_1$.

- **Next we look at the intercept.**

  With the new measurement, the estimator of the intercept is:

  $$\widetilde{\beta}_0 = \overline{\log(y^*)} - \widetilde{\beta}_1 \bar{x}.$$

  Let us replace $y_i^* = c_1 y_i$, and recall $\overline{\log(y^*)} = \log(c_1) + \overline{\log(y)}$:

  $$\widetilde{\beta}_0 = \log(c_1) + \overline{\log(y)} - \widetilde{\beta}_1 \bar{x}.$$

  Next, we replace our previous result that showed $\widetilde{\beta}_1 = \widehat{\beta}_1$.

  Thus, we obtain:

  $$\begin{aligned}\widetilde{\beta}_0 &= \log(c_1) + \overline{\log(y)} - \widehat{\beta}_1 \bar{x} \\ &= \log(c_1) + \widehat{\beta}_0.\end{aligned}$$

Solution to Activity 3.8

Let us follow the steps in the instructions.

1. We know that:

   $$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}.$$

2. From the regression equation $y_i = \beta_0 + \beta_1 x_i + u_i$, for $i = 1, \ldots, n$, we can obtain that:

   $$\bar{y} = \beta_0 + \beta_1 \bar{x} + \bar{u}.$$

   Plugging this expression for $\bar{y}$ into the formula for $\widehat{\beta}_0$, we get:

   $$\widehat{\beta}_0 = \beta_0 + (\beta_1 - \widehat{\beta}_1)\bar{x} + \bar{u}.$$

3. Let us find the conditional expectation $E(\widehat{\beta}_0 \,|\, \mathbf{X})$:

   $$\begin{aligned}E(\widehat{\beta}_0 \,|\, \mathbf{X}) &= E(\beta_0 + (\beta_1 - \widehat{\beta}_1)\bar{x} + \bar{u} \,|\, \mathbf{X}) \\ &= \beta_0 + E((\beta_1 - \widehat{\beta}_1)\bar{x} \,|\, \mathbf{X}) + E(\bar{u} \,|\, \mathbf{X}) \\ &= \beta_0 + E(\beta_1 - \widehat{\beta}_1 \,|\, \mathbf{X})\bar{x} + \frac{1}{n}\sum_{i=1}^{n} E(u_i \,|\, \mathbf{X}) \\ &= \beta_0 + (\beta_1 - E(\widehat{\beta}_1 \,|\, \mathbf{X}))\bar{x} + \frac{1}{n}\sum_{i=1}^{n} E(u_i \,|\, x_i) \\ &= \beta_0 + (\beta_1 - \beta_1)\bar{x} + 0 \\ &= \beta_0.\end{aligned}$$

4. We can now apply the law of iterated expectation to conclude that:

   $$E(\widehat{\beta}_0) = E(E(\widehat{\beta}_0 \,|\, \mathbf{X})) = E(\beta_0) = \beta_0.$$

3.4 Overview of chapter

In this chapter we introduce a simple linear regression model in which a dependent variable y is a linear function of a regressor x plus an unobserved error u. The intercept and the slope parameters of the linear function are the parameters of interest.

The chapter discusses the Ordinary Least Squares (OLS) estimation of the intercept and slope parameters and the ideas behind this estimation. It analyses the effects of changing the units of measurement on OLS estimates. It also introduces the goodness-of-fit measure R^2 which characterises the predictive power of the regressor for the dependent variable.

In the chapter it is established that under Assumptions SLR.1–SLR.4 with Assumption SLR.4 being the key, the OLS estimator is unbiased. Adding the homoskedasticity Assumption SLR.5, the chapter derives the formula for the variance of the OLS estimator and analyses how it is affected by different components.

3.4.1 Key terms and concepts

- Simple linear regression model
- Explanatory variable
- Dependent variable
- Error term
- Slope coefficient
- Intercept
- Ordinary Least Squares (OLS)
- Residual
- Fitted value
- Population regression function
- Sample regression function
- Explained Sum of Squares (SSE)
- Residual Sum of Squares (SSR)
- Total Sum of Squares (SST)
- R-squared
- Ordinary Least Squares (OLS)
- Homoskedasticity
- Heteroskedasticity

3.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand three issues when formulating a simple regression model and discuss how they are addressed
- know the terminology for different components of the regression model
- discuss the key assumptions about how the regressor and error term are related
- derive and plot the population regression function
- derive and interpret OLS estimates in the simple regression model
- define OLS fitted values and residuals and derive their algebraic properties in OLS estimation
- discuss the effect of changing the units of measurement on the parameter estimates and the R^2
- interpret the OLS estimates when the dependent and/or independent variables appear in logarithmic form
- formulate and discuss assumptions of the linear regression model
- establish unbiasedness of the OLS estimator and understand the role of the assumptions in this result
- derive the variance of the OLS estimator under the homoskedasticity assumption and discuss how it is affected by different components.

3.5 Test your knowledge and understanding

3.5.1 Additional examination based exercises

3E.1 Let *price* denote the selling price (in thousands of dollars) of a house in a community, and let *bdr* denote the number of bedrooms. A simple model relating the selling price to the number of bedrooms is:

$$price = \beta_0 + \beta_1 bdr + u \tag{3.5.1}$$

where u is the unobserved error.

(a) List at least one factor contained in u. Is this likely to be correlated with the number of bedrooms?

(b) What does it mean to say that 'the OLS estimator is unbiased'?

(c) Do you think that the OLS estimator in (3.5.1) provides an unbiased estimator of the *ceteris paribus* effect of *bdr* on *price*? (You are not expected to present a formal proof, just give your intuition.)

(d) Define the R^2 in general terms. Suppose $R^2 = 0.17$ in the regression in (3.5.1). Interpret this result.

3E.2 The dataset BWGHT.RAW contains data on births to women in the United States. Two variables of interest are the dependent variable, infant birthweight in ounces (*bwght*), and an explanatory variable, the average number of cigarettes the mother smoked per day during pregnancy (*cigs*). The following simple regression was estimated using data on $n = 1{,}388$ births:

$$\widehat{bwght} = 119.77 - 0.514 cigs.$$

(a) What is the predicted birth weight when $cigs = 0$? What about when $cigs = 20$ (one pack per day)? Comment on the difference.

(b) Does this simple regression necessarily capture a causal relationship between the child's birthweight and the mother's smoking habits? Explain.

(c) To predict a birthweight of 125 ounces, what would *cigs* have to be? Comment.

(d) The proportion of women in the sample who do not smoke while pregnant is about 0.85. Does this help reconcile your finding from part (c)?

3E.3 Consider running the regression on a constant only and, thus, only estimating an intercept.

(a) Given a sample $\{y_i : i = 1, 2, \ldots, n\}$, let $\widetilde{\beta}_0$ be the solution to:

$$\min_{b_0} \sum_{i=1}^{n} (y_i - b_0)^2.$$

Show that $\widetilde{\beta}_0 = \bar{y}$, that is, the sample average minimises the sum of squared residuals.

(b) Define residuals $\widetilde{u}_i = y_i - \bar{y}$. Argue that these residuals always sum to zero.

3E.4 Consider the bivariate regression model without intercept:

$$y_i = \beta x_i + u_i$$

for $i = 1, \ldots, n$. We impose the following assumptions:

- **SLR.1** The population model is $y = \beta x + u$.
- **SLR.2** We have a random sample of size n, $\{(y_i, x_i) : i = 1, \ldots, n\}$, following the population model in SLR.1.
- **SLR.3** The sample outcomes on $\{x_i : i = 1, \ldots, n\}$ are not all the same value.
- **SLR.4** The error term u satisfies $E(u \,|\, x) = 0$ for any value of x.
- **SLR.5** The error term u satisfies $Var(u \,|\, x) = \sigma^2$ for any value of x (homoskedasticity).

(a) The OLS estimator $\widehat{\beta}$ is defined as a solution of:

$$\min_b \sum_{i=1}^{n} (y_i - bx_i)^2.$$

Show that $\widehat{\beta}$ is written as:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(b) Show that the above $\widehat{\beta}$ can be written as:

$$\widehat{\beta} = \beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}.$$

(c) Show that $\widehat{\beta}$ is an unbiased estimator of β.

(d) Let $\mathbf{X} = (x_1, \ldots, x_n)$. Compute the conditional variance $Var(\widehat{\beta} \,|\, \mathbf{X})$ for the above OLS estimator.

3E.5 To investigate the relationship between the price of wine and consumption of wine, an economist estimates the following regression using a sample of 32 individuals for one week in 2013:

$$\widehat{\log(wine)} = \underset{(0.8911)}{4.2514} - \underset{(0.0031)}{0.8328} \log(price)$$
$$n = 32, \; R^2 = 0.89.$$

wine denotes the amount of wine consumed per week in millilitres (a medium glass contains 175 ml), and *price* denotes the average price of a medium glass of wine of a selection of wines during the week in GBP (£). The numbers in parentheses are the standard errors.

(a) Discuss what would happen to the parameter estimate of the slope coefficient if we had measured the amount of wine consumed per week in number of medium glasses instead of millilitres. Explain your answer.

(b) Discuss what would happen to the parameter estimate of the intercept if we had measured the amount of wine consumed per week in number of medium glasses instead of millilitres. Explain your answer.

3.5.2 Solutions to examination based exercises

3E.1 (a) The total size of the house (in sq m), the number of bathrooms, the size of a garden, and the condition of the house are a few possibilities. It seems that at least two of these (except for the garden size and the condition of the house maybe) could be correlated with the number of bedrooms. (The total size and the number of bedrooms are probably positively correlated; the number of bathrooms and the number of bedrooms are probably positively correlated.)

(b) Unbiasedness is a feature of the sampling distribution of the estimator. It means the expected value of the OLS estimator coincides with the true parameter value in the population.

(c) It is very unlikely that the estimator for the slope coefficient is unbiased because there are variables contained in the unobservable term that are correlated with the number of bedrooms. For example, the number of bedrooms and the total size of the house are probably positively correlated.

(d) The R^2 is a measure of goodness-of-fit and defined as:

$$R^2 = \frac{SSE}{SST}$$

where $SSE = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2$ is the explained sum of squares and $SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the total sum of squares.

From $R^2 = 0.17$ in the regression we can see that *bdr* explains 17% of the variation in *price*.

3E.2 (a) When $cigs = 0$, the predicted birthweight is 119.77 ounces. When $cigs = 20$, $bwght = 109.49$. This is about an 8.6% drop.

(b) Not necessarily. There are many other factors that can affect birthweight, particularly education. Education could be correlated with both cigarette smoking during pregnancies and knowledge about prenatal care.

(c) If we want a predicted *bwght* of 125, then:

$$cigs = \frac{125 - 119.77}{-.524} \approx -10.18$$

or about -10 cigarettes! This is nonsense, of course, and it shows what happens when we are trying to predict something as complicated as birthweight with only a single explanatory variable. The largest predicted birthweight is necessarily 119.77. Yet almost 700 of the births in the sample had a birthweight higher than 119.77.

(d) 1,176 out of 1,388 women did not smoke while pregnant, or about 84.7%. Because we are using only *cigs* to explain birthweight, we have only one predicted birthweight at $cigs = 0$. The predicted birthweight 119.77 is necessarily roughly in the middle of the observed birthweights at $cigs = 0$, and therefore a large number of birthweights are above the intercept 119.77. To predict such higher birthweights, the value of *cigs* has to be negative.

3E.3 (a) The first-order condition for $\widetilde{\beta}_0$ is:

$$(-2)\sum_{i=1}^{n}(y_i - \widetilde{\beta}_0) = 0$$

which can be rewritten as:

$$\sum_{i=1}^{n} y_i - n\widetilde{\beta}_0 = 0.$$

By solving this for $\widetilde{\beta}_0$, we obtain $\widetilde{\beta}_0 = \bar{y}$.

(b) From (a) we have:

$$\widetilde{u}_i = y_i - \widetilde{\beta}_0 = y_i - \bar{y}.$$

Thus, from the first-order condition:

$$\sum_{i=1}^{n} \widetilde{u}_i = \sum_{i=1}^{n}(y_i - \widetilde{\beta}_0) = \sum_{i=1}^{n} y_i - n\widetilde{\beta}_0 = 0.$$

3E.4 (a) By taking the derivative of $\sum_{i=1}^{n}(y_i - bx_i)^2$ with respect to b, the first-order condition of the above minimisation problem is:

$$(-2)\sum_{i=1}^{n} x_i(y_i - \widehat{\beta}x_i) = 0$$

or equivalently:

$$0 = \sum_{i=1}^{n} x_i(y_i - \widehat{\beta}x_i) = \sum_{i=1}^{n} x_i y_i - \widehat{\beta}\sum_{i=1}^{n} x_i^2.$$

By SLR.2, $\{x_i : i = 1, \ldots, n\}$ cannot be all zero. Thus, $\sum_{i=1}^{n} x_i^2 \neq 0$. So, by solving for $\widehat{\beta}$, we obtain the conclusion:

$$\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}.$$

(b) By SLR.1–2, we have $y_i = \beta x_i + u_i$ for all $i = 1, \ldots, n$. By plugging this into the expression $\widehat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ obtained in (a), we have:

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

i.e. the conclusion is obtained.

(c) From (b) and conditional on $\mathbf{X} = (x_1, \ldots, x_n)$ (or treat $\mathbf{X}$ as non-random), we have:

$$\begin{aligned} E(\widehat{\beta}) &= E\left(\beta + \frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right) \\ &= \beta + E\left(\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right) \\ &= \beta + \frac{\sum_{i=1}^{n} x_i E(u_i)}{\sum_{i=1}^{n} x_i^2} \\ &= \beta \end{aligned}$$

where the first equality follows from (b), the second equality follows from the property of $E(\cdot)$, the third equality follows from the fact that we treat $\mathbf{X}$ as non-random and property of $E(\cdot)$, and the last equality follows from SLR.4.

(d) Note that:

$$\begin{aligned}
Var(\widehat{\beta}\,|\,\mathbf{X}) &= Var\left(\beta + \left.\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right| \mathbf{X}\right) \\
&= Var\left(\left.\frac{\sum_{i=1}^{n} x_i u_i}{\sum_{i=1}^{n} x_i^2}\right| \mathbf{X}\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} Var\left(\left.\sum_{i=1}^{n} x_i u_i\right| \mathbf{X}\right) \\
&= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} Var(x_i u_i\,|\,\mathbf{X}) \\
&= \frac{1}{\left(\sum_{i=1}^{n} x_i^2\right)^2} \sum_{i=1}^{n} x_i^2\, Var(u_i\,|\,\mathbf{X}) \\
&= \frac{\sigma^2}{\left(\sum_{i=1}^{n} x_i^2\right)}
\end{aligned}$$

where the first equality follows from the definition of $\widehat{\beta}$, the second equality follows from the property of the variance, the third equality follows from the property of the variance and the fact that $\mathbf{X}$ is treated as non-random, the fourth equality follows from SLR.2, the fifth equality follows from the property of the variance and the fact that $\mathbf{X}$ is treated as non-random, and the last equality follows from SLR.5.

3E.5 (a) The parameter estimate of the slope coefficient will stay the same if we measure the amount of wine consumed per week in medium glasses instead of millilitres. The coefficient should be interpreted as the elasticity of wine consumption with respect to price. A 1.0% increase in price, will result in a 0.8328% decline in the amount of wine consumed. This elasticity does not vary with the way we measure the wine consumption (or prices for that matter). Formally, we note (subscript indicating units of measurement):

$$\log(wine_{ml}) = \log(wine_{glasses} \times 175) \approx \log(wine_{glasses}) + 5.2$$

where we make use of the properties of the log operator, that ensures $\log(x \cdot y) = \log(x) + \log(y)$, for $x, y > 0$. With our original specification given by:

$$\log(wine_{ml}) = \alpha + \beta \log(price) + u$$

we obtain, upon substitution:

$$\log(wine_{glasses}) = \underbrace{\alpha - \log(175)}_{\text{new intercept}} + \underbrace{\beta}_{\text{same slope}} \log(price) + u.$$

(b) As shown above, the intercept will go down by log(175).

Chapter 4
Multiple regression analysis: Estimation

4.1 Introduction

4.1.1 Aims of the chapter

In this chapter we extend the simple linear regression model to the multiple linear regression model. You will learn what the benefits are of using multiple linear regression models, and you will learn how to properly interpret parameters in these models. We extend the Gauss–Markov assumptions of the linear regression model, which now will include an additional assumption: the absence of perfect multicollinearity. We extend the unbiasedness result of the OLS estimators to the multiple linear regression model under the Gauss–Markov assumptions and provide an expression for the sampling variance of the OLS slope estimators. You will learn the consequences associated with two forms of misspecification: the consequences of including irrelevant variables in our multiple linear regression model and the (more serious) consequences of excluding relevant variables. By discussing the Gauss–Markov theorem, we aim to justify the use of the OLS estimator rather than another competing estimator.

4.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- discuss the advantage of the multiple regression model over the simple regression model
- understand the principles behind the derivation of the OLS estimator of parameters in the multiple regression model
- interpret regression coefficients including the importance of 'holding other factors fixed' (*ceteris paribus*)
- discuss the 'partialling-out' approach to estimating parameters (Frisch–Waugh–Lovell)
- understand the effect on the R^2 of adding more explanatory variables
- demonstrate knowledge of the assumptions that guarantee the unbiasedness of OLS estimators (MLR.1–MLR.4) and the assumption (MLR.5) used to derive the variance of the regression parameters

- discuss the consequence of including irrelevant variables
- discuss the consequence of excluding relevant variables: Omitted Variable Bias (OVB)
- discuss the likely direction of the OVB in empirical settings
- knowledge of the expression for $Var(\widehat{\beta}_j \,|\, X)$ under MLR.1–MLR.5, and have a clear understanding of what factors improve the precision of the OLS estimators
- understand the concept of multicollinearity and discuss the consequences of near multicollinearity
- understand how the standard errors of the regression parameters are obtained
- discuss the Gauss–Markov Theorem and efficiency of the OLS estimator
- use statistical software to apply the OLS estimation technique to analyse the relationship between real-world economic variables.

4.1.3 Essential reading

Readings: Wooldridge, Sections 3.1–3.5 (skip 3.2i) and Appendix 3A.1.

4.1.4 Synopsis of this chapter

After motivating the use of the multiple linear regression model in Section 3.1 in Wooldridge, the chapter shows how to extend the ordinary least squares approach to the multiple regression setting in Section 3.2 in Wooldridge. In this section, the importance of interpreting the parameters as providing partial effects, effects of a particular independent variable on the dependent variable 'controlling', that is holding fixed, all other factors is highlighted. Rather than explicitly controlling for everything else, it argues that we can also initially 'partial-out' the effect of all other factors (the Frisch–Waugh–Lovell Theorem). Terminology introduced in the simple linear regression model, such as fitted values, residuals and the goodness-of-fit R^2 easily generalise to the multiple regression model. Section 3.3 in Wooldridge discusses the conditions that guarantee the OLS estimator is unbiased and reveals that while including irrelevant variables will not change this result, excluding relevant variables will lead to the Omitted Variable Bias (OVB) result. In Section 3.4 in Wooldridge, the sampling variance of the OLS slope parameters, and its components, is discussed under the additional assumption of homoskedasticity. We learn how to estimate this sampling variance using data and obtain standard errors of our estimates. In Section 3.5 in Wooldridge, the Gauss–Markov Theorem is discussed which ensures that the OLS is the best linear unbiased estimator (BLUE), which presents an efficient result.

4.2 Content of chapter

4.2.1 Motivation for multiple regression

Read: Wooldridge Section 3.1.

It is important to understand the motivation for the multiple linear regression model. It provides a framework that will, more readily, allow us to provide a causal interpretation to our parameters. By introducing more independent variables (controls, also referred to as confounders), we can isolate the effect we are interested in: the effect of a particular independent variable 'holding everything else constant'.

4.2.2 Mechanics and interpretation of Ordinary Least Squares (OLS)

Read: Wooldridge Section 3.2 (skip 3.2i) and Appendix 3A.1.

You should be able to derive the first order conditions that define the OLS estimates in the multiple linear regression setting (see also Activity 4.1). You are not expected to solve for the parameter estimates in the *multiple* linear regression setting. It is important that you interpret the OLS estimates as **partial effects**, effects that control (hold constant) all other factors (also called confounders) and apply this correctly in empirical examples.

It is useful to understand the important result that shows that, instead of explicitly controlling for everything by including additional regressors (controls), we may alternatively 'partial-out' the influence of the other explanatory variables beforehand. Specifically, one can show that the estimated coefficient of an explanatory variable, x_j, in a multiple regression can be obtained in two steps:

1. Regress the explanatory variable x_j on an intercept and all other explanatory variables and obtain the residuals:

$$\widehat{r}_j = x_j - \widehat{x}_j$$

   where $\widehat{x}_j$ denotes the fitted values.

2. Regress y on the residuals $\widehat{r}_j$ from this regression (no intercept):

$$\widehat{\beta}_j = \frac{\sum_{i=1}^{n} \widehat{r}_{ij} y_i}{\sum_{i=1}^{n} \widehat{r}_{ij}^2}.$$

Why does this procedure work?

- The residuals from the first regression is the part of the explanatory variable that is uncorrelated with the other explanatory variables. The residuals explicitly control for the other explanatory variables by removing any correlation they have with them.

- The slope coefficient of the second regression therefore represents the isolated effect of the explanatory variable on the dependent variable.

You should be familiar with the terminology and properties of OLS fitted values, residuals (described in 3.2e) and the R^2 (described in 3.2h).

Activity 4.1 Consider the multiple regression model:

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u.$$

Provide the three FOCs that define the OLS parameter estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$.

Activity 4.2 Exercise 3.2 from Wooldridge.

Activity 4.3 Exercise 3.4 from Wooldridge.

Activity 4.4 Exercise 3.16 from Wooldridge.

Activity 4.5 Take the MCQs related to this section on the VLE to test your understanding.

4.2.3 The expected value of the OLS estimators

Read: Wooldridge Sections 3.3 and 3.2g.

The section starts by providing the conditions that ensure that the OLS estimators in the multiple linear regression model are unbiased, see Theorem 3.1 in Wooldridge. It is important to understand these conditions MLR.1–MLR.4 well. The zero conditional mean assumption (also called a conditional independence assumption) is the most complicated one to interpret. It ensures that we can assume that the expected value of the unobservables, u, will not change if we change an independent variable holding all other controls fixed, so that:

$$\frac{\partial E(y \,|\, x_1, \ldots, x_k)}{\partial x_j} = \beta_j, \quad \text{for } j = 1, \ldots, k.$$

It is crucial therefore in enabling us to give our parameters a causal interpretation. By adding additional regressors, we have removed components from the errors that may have given rise to correlation between the errors and regressors beforehand.

You should be able to explain why this unbiasedness result remains when we include irrelevant variables in our multiple linear regression model, and be able to derive the omitted variable bias (see details below). You should feel comfortable discussing the likely direction of this OVB in empirical settings. You need to clearly indicate the fact that the omitted variable is relevant (non-zero coefficient) and that the omitted variable is correlated with the included regressor(s).

Let us look here in detail at the consequences of omitting relevant variables adding some notation that you may want to adopt. We consider the setting where the true population model has two explanatory variables and satisfies MLR.1 to MLR.4:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u. \tag{*}$$

This regression will give unbiased OLS estimators which we denote as $\widehat{\beta}_{j,long}$, for $j = 0, 1, 2$. Let $\widetilde{\beta}_{1,short}$ denote the estimator of β_1 when we exclude the relevant variable x_2, that is when we estimate:

$$y = \beta_0 + \beta_1 x_1 + v.$$

The subscripts on our parameter estimates clearly denote whether we are considering the multiple regression model (long) or the simple regression model (short).

In Section 3.2g in Wooldridge it is pointed out that it can be shown that:

$$\widetilde{\beta}_{1,short} = \widehat{\beta}_{1,long} + \widehat{\beta}_{2,long}\widehat{\delta}_1$$

where $\widehat{\delta}_1$ is the OLS estimator of the slope when regressing x_2 on x_1. This expression indicates that when we exclude a relevant variable (i.e. consider the short model), that we are not only estimating the effect of interest ($\widehat{\beta}_{1,long}$ is an unbiased estimator of β_1), but are estimating an additional effect $\widehat{\beta}_{2,long}\widehat{\delta}_1$ which represents the effect x_1 has on y through its correlation with x_2.

Whereas deriving this result is beyond the scope of this course, it is good to understand why this is true. Let us describe the relation between x_2 and x_1 as:

$$x_2 = \delta_0 + \delta_1 x_1 + e \tag{**}$$

where e is an error which is uncorrelated with x_1. If we substitute this expression in (*), we obtain:

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2(\delta_0 + \delta_1 x_1) + e) + u \\ &= \underbrace{(\beta_0 + \beta_2\delta_0)}_{intercept} + \underbrace{(\beta_1 + \beta_2\delta_1)}_{slope} x_1 + error. \end{aligned}$$

In the textbook you can find a derivation of the OVB that makes use of this relation. Let us look at an alternative derivation, which you may find more appealing.

We are interested in deriving $E(\widetilde{\beta}_{1,short} \,|\, \mathbf{X})$. The conditioning on $\mathbf{X}$ allows us to treat the regressors $\{(x_{1i}, x_{2i}), i = 1, \ldots, n)\}$ as fixed constants (not stochastic).

- The estimator $\widetilde{\beta}_{1,short}$ (simple regression) is defined as:

$$\widetilde{\beta}_{1,short} = \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(y_i - \bar{y})}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}.$$

- When we plug in the true (long) model (**), this allows us to write:

$$\widetilde{\beta}_{1,short} = \beta_1 + \beta_2 \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} + \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(u_i - \bar{u})}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}.$$

- Observe that using the true model:

$$\begin{aligned} y_i - \bar{y} &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i - (\beta_0 + \beta_1 \bar{x}_1 + \beta_2 \bar{x}_2 + \bar{u}) \\ &= \beta_1(x_{1i} - \bar{x}_1) + \beta_2(x_{2i} - \bar{x}_2) + (u_i - \bar{u}). \end{aligned}$$

► If we now take the conditional expectation, we obtain:

$$E(\widetilde{\beta}_{1,short} \,|\, \mathbf{X}) = \beta_1 + \underbrace{\beta_2 \frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}}_{OVB} \neq \beta_1$$

where $\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}$ is the OLS formula for $\widehat{\delta}_1$ (assumed to be non-zero).

- Observe, as we condition on the independent variables, $\mathbf{X}$, the first two terms should be obvious. The conditioning on $\mathbf{X}$ allows us to treat the regressors $\{(x_{1i}, x_{2i}), i = 1, \ldots, n)\}$ as fixed constants (not stochastic).
- Let us look at why the third term vanishes when we take expectation. First, realise that:

$$E\left(\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)(u_i - \bar{u})}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \,\middle|\, \mathbf{X} \right) \equiv E\left(\frac{\sum_{i=1}^n (x_{1i} - \bar{x}_1)u_i)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2} \,\middle|\, \mathbf{X} \right)$$

  as $\sum_{i=1}^n (x_{1i} - \bar{x}_1)\bar{u} = \bar{u} \sum_{i=1}^n (x_{1i} - \bar{x}_1) = 0$.

  We can simplify the resulting expression as:

$$E\left(\sum_{i=1}^n d_i u_i \,\middle|\, \mathbf{X} \right), \quad \text{where } d_i = \frac{(x_{1i} - \bar{x}_1)}{\sum_{i=1}^n (x_{1i} - \bar{x}_1)^2}.$$

- Next, let us take the constants d_i outside of the expectation operator:

$$E\left(\sum_{i=1}^n d_i u_i \,\middle|\, \mathbf{X} \right) = \sum_{i=1}^n d_i E(u_i \,|\, \mathbf{X}).$$

- This is zero. By random sampling, MLR.2, $E(u_i \,|\, \mathbf{X}) = E(u_i \,|\, x_i)$, and MLR.4 ensures $E(u_i \,|\, x_i) = 0$.

We conclude therefore, when we do not control for x_2, our estimator of β_1 does not have a causal interpretation as it incorporates the effect x_2 has on y as well (a confounder). We should, however, recognise that there are two cases where excluding variables does not lead to bias:

1. $\beta_2 = 0$, which means that x_2 is irrelevant, and bivariate (short) regression is fine.
2. $\widehat{\delta}_1 = 0$, which means that x_1 and x_2 are uncorrelated in the sample.

> **Activity 4.6** Take the MCQs related to this section on the VLE to test your understanding.

4.2.4 The variance of the OLS estimators

Read: Wooldridge Section 3.4.

In this section we introduce a simplifying assumption that the errors have constant variance (homoskedasticity), given in MLR.5. The *advantages of adding the assumption of homoskedasticity* are twofold.

1. The formula for the sample variance of our slope estimators will be simple and permit us to discuss the factors that will help us obtain more precise parameter estimates.

2. Under this additional assumption, our OLS estimator will have an important *efficiency property* (Gauss–Markov Theorem) (see next section).

You should be familiar with the expression of the conditional variance of the OLS slope parameters that can be derived under Assumptions MLR.1–MLR.5, specifically:

$$Var(\widehat{\beta}_j \,|\, \mathbf{X}) = \frac{\sigma^2}{SST_j(1-R_j^2)} = \frac{\sigma^2}{(n-1)Sample\,Var(x_j)(1-R_j^2)}, \quad \text{for } j = 1, \ldots, k$$

where n is the sample size, SST_j the total sample variation in x_j, and R_j^2 the goodness-of-fit from regressing x_j on all other regressors. $\mathbf{X}$ represents the observable independent variables, $\{(x_{1i}, \ldots, x_{ki}), i = 1, \ldots, n\}$. The conditioning simply allows us to ignore the random nature of the independent variables. (As in the simple regression model, the variance depends on the explanatory variables, and by conditioning on $\mathbf{X}$ we can treat the independent variables as fixed constants.)

You are not expected to be able to derive the above expression in the multiple regression setting, but they are expected to discuss its components and relate this discussion to the presence of near multicollinearity. The discussion about the variance inflation factor (VIF) presented in the textbook is not examinable.

You should know how to obtain an estimate of the variance under the Gauss–Markov assumptions (we need to replace σ^2 by an unbiased estimator thereof) and how we obtain the standard errors using:

$$se(\widehat{\beta}_j) = \sqrt{\widehat{Var}(\widehat{\beta}_j \,|\, \mathbf{X})}.$$

It is critical to realise that the formula for the sampling variation of $\widehat{\beta}_j$ displayed above and its estimate, requires all Gauss–Markov assumptions to be satisfied. Any violations of the Gauss–Markov assumptions will render this formula invalid, and standard errors based on it will be incorrect. (We will later discuss how we can make the standard errors robust to violations such as heteroskedasticity.)

Section 3.4b in Wooldridge finally highlights that including irrelevant variables, while leaving the OLS estimator unbiased, is not without its cost as it will lead to a higher variance. Nevertheless, the validity of the standard errors remains in the presence of irrelevant variables. Excluding relevant variables, on the other hand, will not only lead to biased parameters but it will also lead to invalid standard errors.

> **Activity 4.7** Exercise 3.6 from Wooldridge.

> **Activity 4.8** Exercise 3.10 from Wooldridge.
>
> You may want to introduce the following easier-to-interpret notation when answering this question.
>
> - $\widetilde{\beta}_{1,short}$ is the estimate of β_1 when estimating the simple regression of y on x_1
> - $\widehat{\beta}_{1,long}$ is the estimate of β_1 when we are estimating the multiple regression of y on x_1, x_2, x_3.

> **Activity 4.9** Take the MCQs related to this section on the VLE to test your understanding.

4.2.5 Efficiency of OLS: The Gauss–Markov Theorem

Read: Wooldridge Section 3.5.

In this last section, we discuss the Gauss–Markov Theorem. It shows that under the assumptions MLR.1–MLR.5 (also called our Gauss–Markov assumptions) the OLS estimator has the desirable property that there is no other linear, unbiased estimator that has a lower variance, in other words the OLS estimator is 'Best'. The proof is not examinable. It is good to point out that this efficiency result is limited in the sense that it does not guarantee that there might be a nonlinear and/or biased estimator that 'beats' the OLS estimator.

It should be clear also that any violation of the Gauss–Markov assumptions (for instance, due to the presence of heteroskedasticity) will render our OLS estimator $\widehat{\beta}_j$ no longer efficient (BLUE). In such settings, clearly, we may want to discuss how we can regain an efficiency result in its presence.

> **Activity 4.10** Take the MCQs related to this section on the VLE to test your understanding.

4.2.6 R Application for multiple regression analysis: Estimation

Complete 'R Task – Activity 2' on VLE.

Here we show how to load a relevant dataset in R. We also look at how to perform OLS on a bivariate and multivariate model using the command 'lm'. We also briefly consider the issue of multicollinearity.

4.3 Answers to activities

Solution to Activity 4.1

- In this case we have $k = 2$ (two slopes) and have used $x_{i1} = inc_i$ and $x_{i2} = inc_i^2$.

  The three first-order conditions required for OLS are:

$$\frac{\partial SSR}{\partial b_0} : -2\sum_{i=1}^{n}(cons_i - \widehat{\beta}_0 - \widehat{\beta}_1 inc_i - \widehat{\beta}_2 inc_i^2) = 0$$
$$\frac{\partial SSR}{\partial b_1} : -2\sum_{i=1}^{n}(cons_i - \widehat{\beta}_0 - \widehat{\beta}_1 inc_i - \widehat{\beta}_2 inc_i^2)inc_i = 0$$
$$\frac{\partial SSR}{\partial b_2} : -2\sum_{i=1}^{n}(cons_i - \widehat{\beta}_0 - \widehat{\beta}_1 inc_i - \widehat{\beta}_2 inc_i^2)inc_i^2 = 0$$

  or equivalently, denoting $\widehat{u}_i = cons_i - \widehat{\beta}_0 - \widehat{\beta}_1 inc_i - \widehat{\beta}_2 inc_i^2$, we have:

$$\frac{\partial SSR}{\partial b_0} : -2\sum_{i=1}^{n}\widehat{u}_i = 0$$
$$\frac{\partial SSR}{\partial b_1} : -2\sum_{i=1}^{n}\widehat{u}_i inc_i = 0$$
$$\frac{\partial SSR}{\partial b_2} : -2\sum_{i=1}^{n}\widehat{u}_i inc_i^2 = 0.$$

Solution to Activity 4.2

(Exercise 3.2 from Wooldridge.)

(i) Yes. Because of budget constraints, it makes sense that, the more siblings there are in a family, the less education any one child in the family has. To find the increase in the number of siblings that reduces predicted education by one year, we solve:

$$\Delta\widehat{educ} = -1 = -.094(\Delta sibs)$$

so $\Delta sibs = 1/.094 \approx 10.6$.

(ii) Holding *sibs* and *feduc* fixed, one more year of mother's education implies .131 years more of predicted education.

If a mother has four more years of education, her son is predicted to have about half a year ($4 \times .131 = .524$) more education.

(iii) Since the number of siblings is the same, but *meduc* and *feduc* are both different, the coefficients on *meduc* and *feduc* both need to be accounted for.

The predicted difference in education between B and A is:

$$.131 \times (16 - 12) + .210 \times (16 - 12) = 1.364.$$

Solution to Activity 4.3

(Exercise 3.3 from Wooldridge.)

(i) A larger rank for a law school means that the school has less prestige; this lowers starting salaries. For example, a rank of 100 means there are 99 schools thought to be better.

(ii) We expect both β_1 and β_2 to be positive since both LSAT and GPA are measures of the quality of the entering class. No matter where better students attend law school, we expect them to earn more, on average.

We also expect β_3 and β_4 to be positive since the number of volumes in the law library and the tuition cost are both measures of the school quality. (Cost is less obvious than library volumes, but should reflect quality of the faculty, facilities, and so on.)

(iii) This is a semi-elasticity. One additional GPA unit, increases the salary by 24.8%, *ceteris paribus*. This is an example of the **log-linear specification** interpretation.

(iv) This is an elasticity. A one percent increase in library volumes implies a 0.095% increase in predicted median starting salary, other things being equal. This is an example of the **log-log specification** interpretation.

(v) It is definitely better to attend a law school with a lower rank.

If law school A has a ranking 20 less than law school B, the predicted difference in starting salary is $100(.0033 \times 20) = 6.6\%$ higher for law school A. This is another example of the **log-linear specification** interpretation.

Solution to Activity 4.4

(Exercise 3.16 from Wooldridge.)

An important fact about R^2 is that it never decreases, and it usually increases, when another independent variable is added to a regression (using the same dependent variable) and the same set of observations is used for both regressions.

Here the sets of observations used are different! If two regressions use different sets of observations (or we compare models with different dependent variables) we cannot compare the R-squareds.

Solution to Activity 4.7

(Exercise 3.6 from Wooldridge.)

(i) Using the properties of the expectation operator (Property E.3), we have:

$$E(\widehat{\theta}) \equiv E(\widehat{\beta}_1 + \widehat{\beta}_2) = E(\widehat{\beta}_1) + E(\widehat{\beta}_2).$$

By unbiasednesss of the OLS estimators under Assumptions MLR.1–MLR.4, this gives us:

$$E(\widehat{\theta}) = \beta_1 + \beta_2 \equiv \theta$$

which reveals the unbiasedness of $\widehat{\theta}$ as an estimator of θ.

(ii) Using the properties of the variance operator (Property VAR.4), we have:

$$Var(\widehat{\theta}) \equiv Var(\widehat{\beta}_1 + \widehat{\beta}_2) = Var(\widehat{\beta}_1) + Var(\widehat{\beta}_2) + 2Cov(\widehat{\beta}_1, \widehat{\beta}_2).$$

By definition of $Corr(\widehat{\beta}_1, \widehat{\beta}_2)$, we obtain:

$$\begin{aligned} Var(\widehat{\theta}) &= Var(\widehat{\beta}_1) + Var(\widehat{\beta}_2) + 2\,\frac{Corr(\widehat{\beta}_1, \widehat{\beta}_2)}{\sqrt{Var(\widehat{\beta}_1)\,Var(\widehat{\beta}_2)}} \\ &= Var(\widehat{\beta}_1) + Var(\widehat{\beta}_2) + 2\,\frac{Corr(\widehat{\beta}_1, \widehat{\beta}_2)}{sd(\widehat{\beta}_1)sd(\widehat{\beta}_2)}. \end{aligned}$$

Solution to Activity 4.8

(Exercise 3.10 from Wooldridge.)

Let $\widetilde{\beta}_{1,short}$ denote the estimate of β_1 when we estimate the simple regression of y on x_1 and let $\widehat{\beta}_{1,long}$ denote the estimate of β_1 when we estimate the multiple regression of y on x_1, x_2, x_3.

(i) The fact the x_2 and x_3 have large partial effects on y ensures they are both relevant variables.

The high correlation of x_1 with x_2 and x_3 indicates that we will get omitted variable bias (OVB) (we only get OVB when the omitted variables are correlated with x_1). This means that $\widetilde{\beta}_{1,short}$ will then capture their effect on y (large) as well. It is therefore likely that the estimates will be very different.

(ii) Here we would expect $\widetilde{\beta}_{1,short}$ and $\widehat{\beta}_{1,long}$ to be similar (subject to what we mean by 'almost uncorrelated'). Leaving out variables that are not correlated with included regressors does not result in OVB. The amount of correlation between x_2 and x_3 does not directly affect the multiple regression estimate on x_1.

(iii) In this case we are (unnecessarily) introducing multicollinearity into the regression: x_2 and x_3 have small partial effects on y (relevant?) that are highly correlated with x_1.

This multicollinearity is likely to result in an increase of the standard error of the coefficient on x_1; that is $se(\widehat{\beta}_{1,long})$ is likely to be much larger than $se(\widetilde{\beta}_{1,short})$.

(iv) In this case, adding x_2 and x_3 will decrease the residual variance, σ^2, without causing much collinearity (because x_1 is almost uncorrelated with x_2 and x_3), so we should see $se(\widehat{\beta}_{1,long})$ smaller than $se(\widetilde{\beta}_{1,short})$. The amount of correlation between x_2 and x_3 does not directly affect $se(\widehat{\beta}_{1,long})$.

4.4 Overview of chapter

The multiple regression model provides a framework that will, more readily, allow us to provide a causal interpretation to our parameters. We discussed how to extend the ordinary least squares approach to the multiple regression setting, where the parameters provide partial effects, effects that control for all other factors (confounders).

The consequences of including irrelevant variables and excluding relevant variables were discussed. We showed that excluding relevant variables is more severe, and is likely to result in an omitted variable bias, whereby parameter estimates do not only signal the causal effect of interest but incorporate an indirect effect that the excluded factors have on the outcome of interest. It is important that you understand the likely direction of the bias associated with omitting relevant variables.

We discussed what factors will permit us to obtain more precise parameter estimators and discussed the problem of multicollinearity.

4.4.1 Key terms and concepts

- *Ceteris paribus*
- Partial effect
- Confounders
- Frisch–Waugh–Lovell Theorem (FWL)
- Excluding a relevant variable
- Omitted variable bias
- Inclusion of an irrelevant variable
- Perfect collinearity
- Multicollinearity
- Standard deviation of $\widehat{\beta}_j$
- Standard error of $\widehat{\beta}_j$
- Best linear unbiased estimator (BLUE)
- Gauss–Markov Theorem

4.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- discuss the advantage of the multiple regression model over the simple regression model

- understand the principles behind the derivation of the OLS estimator of parameters in the multiple regression model
- interpret regression coefficients including the importance of 'holding other factors fixed' (*ceteris paribus*)
- discuss the 'partialling-out' approach to estimating parameters (Frisch–Waugh–Lovell)
- understand the effect on the R^2 of adding more explanatory variables
- demonstrate knowledge of the assumptions that guarantee the unbiasedness of OLS estimators (MLR.1–MLR.4) and the assumption (MLR.5) used to derive the variance of the regression parameters
- discuss the consequence of including irrelevant variables
- discuss the consequence of excluding relevant variables: Omitted Variable Bias (OVB)
- discuss the likely direction of the OVB in empirical settings
- knowledge of the expression for $Var(\widehat{\beta}_j \,|\, X)$ under MLR.1–MLR.5, and have a clear understanding of what factors improve the precision of the OLS estimators
- understand the concept of multicollinearity and discuss the consequences of near multicollinearity
- understand how the standard errors of the regression parameters are obtained
- discuss the Gauss–Markov Theorem and efficiency of the OLS estimator
- use statistical software to apply the OLS estimation technique to analyse the relationship between real-world economic variables.

4.5 Test your knowledge and understanding

4.5.1 Additional examination based exercises

4.1E Suppose that average worker productivity at manufacturing firms (*avgprod*) depends on two factors, average hours of training (*avgtrain*) and average worker ability (*avgabil*):

$$avgprod = \beta_0 + \beta_1 avgtrain + \beta_2 avgabil + u.$$

Assume that this equation satisfies the Gauss–Markov assumptions.

(a) Provide a clear interpretation of the parameter β_1.

(b) Suppose firms whose workers have less than average ability receive grants to provide training, so that *avgtrain* and *avgabil* are negatively correlated. Discuss the effect of omitting *avgabil* on the OLS estimator of β_1. Will the bias be upward or downward? Explain your answer.

(c) You are told that you can obtain the estimator of β_1 by running the regression:

$$avgprod_i = \beta_1 avgtrain_i^* + e_i, \quad \text{for } i = 1, \ldots, n$$

where $avgtrain_i^*$ is obtained by running a regression of *avgtrain* on an intercept and *avgabil*. Discuss this statement.

4.2E In this question we are interested to see whether fast-food restaurants charge higher prices in areas with a larger concentration of ethnic minorities. ZIP code-level data on prices for various items along with characteristics of the ZIP code population in New Jersey and Pennsylvania are used.

Let us consider a model to explain the price of soda, *psoda*, in terms of the proportion of the population from ethnic minorities, *prpem*, and median income, *income*. Price and income are measured in US$. We obtain the following result:

$$\widehat{psoda} = .956 + .115 prpem + .0000016 income. \tag{2.1}$$

(a) Interpret the parameter on *prpem*.

(b) What would happen to the parameter estimates if we use the percentage of the population that is from ethnic minorities (*pctem*) instead of the decimal equivalent *prpem*? What would happen to the parameter estimates if we measured income in \$10,000s?

(c) When adding a further measure of income to (2.1), the proportion of the population that is in poverty (*prppov*), the coefficient on *prpem* falls to .089. Discuss the following statement 'Because log(*income*) and *prppov* are highly correlated, it is inappropriate to include both'.

(d) A model with constant price elasticity with respect to income gives:

$$\widehat{\log(psoda)} = -.793 + .122 prpem + .077 \log(income). \tag{2.2}$$

Interpret the parameter estimate on *prpem*, and discuss what would happen to the parameter estimates if we use *pctem* instead of *prpem*.

4.3E We are interested in investigating the factors governing the precision of regression coefficients. Consider the model:

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

with OLS parameter estimates $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$. Under the Gauss–Markov assumptions, we have:

$$Var(\widehat{\beta}_2 \,|\, X) = \frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_{2i} - \bar{X}_2)^2} \times \frac{1}{1 - r_{X_2X_3}^2}$$

where σ_ε^2 is the variance of ε and $r_{X_2X_3}$ is the sample correlation between X_2 and X_3.

(a) Provide at least four factors that help us obtain more precise parameter estimates of $\widehat{\beta}_2$.

(b) In light of your answer to (a), discuss the concept of near multicollinearity. Provide a real-life example where this problem is likely to occur.

(c) Are the following statements true or false? Give an explanation.

i. In multiple regression, multicollinearity implies that the least squares estimators of the coefficients are biased and standard errors invalid.

ii. If the coefficient estimates in an equation have high standard errors, this is evidence of high multicollinearity.

4.5.2 Solutions to additional examination based exercises

4.1E (a) This indicates the effect that an additional hour of training has on the average worker productivity at manufacturing firms holding the average worker ability and other factors (in errors) constant.

(b) Omitting a relevant variable ($\beta_2 \neq 0$) gives rise to omitted variable bias (OVB) here as the omitted variable is correlated with the included regressor (*avgabil* and *avgtrain* are negatively correlated). The omitted variable bias is given by:

$$\text{Bias}(\widehat{\beta}_{1,short}) = \beta_2 \widehat{\delta}_1$$

where $\widehat{\delta}_1$, obtained by regressing:

$$avgtrain = \delta_0 + \delta_1 avgabil + v$$

is expected to be negative as *avgabil* and *avgtrain* are negatively correlated.

Since by definition $\beta_2 > 0$, this means that the OVB is negative, that is $E(\widehat{\beta}_{1,short}) < \beta_1$. This means that, on average, across different random samples, the simple regression estimator underestimates the effect of the training programme. It is even possible that $E(\widehat{\beta}_{1,short})$ is negative even though $\beta_1 > 0$.

Notice:

$$\begin{aligned}
\widehat{\beta}_{1,short} &= \frac{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})(avgprod_i - \overline{avgprod})}{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})^2} \\
&= \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})(avgabil_i - \overline{avgabil})}{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})^2} + \frac{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})(u_i - \bar{u})}{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})^2} \\
&= \beta_1 + \beta_2 \frac{SampleCov(avgtrain_i, avgabil_i)}{SampleVar(avgtrain_i)} + \frac{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})u_i}{\sum_{i=1}^{n}(avgtrain_i - \overline{avgtrain})^2}.
\end{aligned}$$

As $E(u \mid avgtrain, avgabil) = 0$ by our Gauss–Markov assumptions, we obtain:

$$E(\widehat{\beta}_{1,short} \mid avgtrain,\ avgabil) = \beta_1 + \underbrace{\beta_2 \frac{SampleCov(avgtrain_i, avgabil_i)}{SampleVar(avgtrain_i)}}_{OVB}.$$

(c) This relates to the partialling out interpretation of our slope parameters. Hereby, we first remove all correlation that *avgprod* and *avgtrain* have with *avgabil* (given by the residuals of the OLS regression of these variables on an intercept and *avgabil*) before estimating the effect of *avgprod* on *avgtrain*. Both yield the same estimates of the effect training has on productivity that control for ability, *ceteris paribus*.

4.2E (a) If, say, *prpem* increases by .10 (ten percentage points), the price of soda is estimated to increase by .0115 dollars, or about 1.2 cents. While this does not seem large, there are communities with no ethnic minorities and others that are almost all ethnic minorities, in which case the difference in *psoda* is estimated to be almost 11.5 cents.

(b) If we use *pctem* instead of *prpem*, its coefficient would equal .0015 to ensure it has the same interpretation, everything else remains the same. If we measure *income* in \$10,000s, its coefficient becomes .016, everything else remains the same.

(c) The discrimination effect is likely to be larger when we control for income. This is due to the fact that we would expect a negative omitted variable bias when excluding income (relevant variable). The bias is negative as income has a positive coefficient in the multiple regression and we anticipate a negative correlation between *prpblck* and *income*.

(d) There is no argument that they are highly correlated, but we are using them simply as controls to determine the price discrimination against ethnic minorities. In order to isolate the pure discrimination effect, controlling for multiple measures of income (even if highly correlated) is a good idea.

(e) If *prpem* increases by .10 (or 10 percentage points), log(*psoda*) is estimated to increase by .10(.122) = .0122, or about 1.22%. If we use *pctem* instead of *prpem*, its coefficient would equal .00122 to ensure it has the same interpretation, everything else remains the same.

(f) If we use *pctem* instead of *prpem*, its coefficient would equal .00122 to ensure it has the same interpretation, everything else remains the same.

4.3E (a) You should note that the expression of the variance can be rewritten as:

$$Var(\widehat{\beta}_2) = \frac{\sigma_\varepsilon^2}{n \times Sample\,Var(X_2) \times (1 - r^2_{X_2X_3})}.$$

The four factors that can improve the precision of our parameter estimates therefore are: (i) larger sample size, n, (ii) larger sample variability of the X_2 regressor, (iii) smaller (in absolute value) correlation among the X_2 and X_3 regressors, $r^2_{X_2X_3}$, and (iv) smaller error variance, σ^2_ε.

(b) Perfect multicollinearity means that some of the regressors can be written as a linear combination of other regressors. For example, $X_i = \sum_{j\neq i}^n \lambda_j X_j$, where X are regressors. Multicollinearity (often referred to as near multicollinearity) indicates that there is a close linear relation between the regressors.

A real-life example of multicollinearity is in a setting where we try to explain the determinants of educational attainment and we use various measures of cognitive skills, which are highly correlated.

(c) Are the following statements true or false? Give an explanation.

i. Presence of near multicollinearity is not a violation of the Gauss–Markov assumptions. In the presence of multicollinearity, the OLS estimators remain unbiased and their standard errors remain valid as long as all the Gauss–Markov assumptions are satisfied. The statement is false.

ii. It is true that presence of near multicollinearity gives rise to high standard errors, but having high standard errors could also be due to the fact that we have a very small sample or high variance of the error term. The statement is false.

Chapter 5
Multiple regression analysis: Inference

5.1 Introduction

5.1.1 Aims of the chapter

In this chapter we discuss how to test single and multiple linear restrictions in the classical linear regression model. We discuss the t test for single linear restrictions and the F test for multiple linear restrictions. You also learn how you can use confidence intervals or p-values to decide whether, at a given level of significance, a null hypothesis should be rejected.

5.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the meaning of the classical linear model (CLM)
- contrast the sampling distribution of $(\widehat{\beta}_j - \beta_j)/se(\widehat{\beta}_j)$ under MLR.1 to MLR.6 with the sampling distribution of $(\widehat{\beta}_j - \beta_j)/sd(\widehat{\beta}_j)$
- perform the t test for testing a hypothesis about a single population parameter
- understand the benefit of using one-sided t test over two-sided t test
- obtain p-values for t tests (one-sided and two-sided)
- construct confidence interval (CI) for population parameters and explain its interpretation
- use confidence interval (CI) for hypothesis testing against two-sided alternatives
- explain the meaning of linear restrictions
- perform the t test for testing a hypothesis about a single linear restriction of the parameters and understand how to obtain the standard error of linear combinations of the estimators
- formulate the F test in two equivalent forms: (i) based on comparing the SSR of the restricted and unrestricted models or (ii) based on a comparison of the R^2 of the restricted and unrestricted models. Interpret the F test as a test for the loss in goodness of fit

- perform the F test for testing multiple (joint) linear restrictions of the population parameters, and perform the F test for the overall significance of a regression

- obtain p-values for F tests

- explain the relation between the F test and the t test when testing a single linear restriction

- use statistical software to implement the t test and the F test when analysing the relationship between real-world economic variables.

5.1.3 Essential reading

Readings: Wooldridge, Sections 4.1–4.5 and Appendices B.5, C.5a, C.5b, C.6a, C.6b, C.6d, C.6e.

5.1.4 Synopsis of this chapter

In order to conduct hypothesis tests and construct confidence intervals, we need to obtain the sampling distribution of the OLS estimator $\widehat{\beta}_j$. Section 4.1 in Wooldridge discusses this distribution (conditional on the regressors) under the Gauss–Markov assumption when the unobserved error is normally distributed and an important standardisation of this sampling distribution is provided. With the help of various empirical examples, Section 4.2 in Wooldridge then proceeds to discuss the t test for a single population parameter; you are shown how to compute p-values for the t test and see how you can use p-values when conducting hypothesis tests. In Section 4.3 in Wooldridge, you learn how to use confidence intervals and how these intervals can be used when testing hypotheses. In Section 4.4 in Wooldridge, you will see how these results can be extended to the setting where we are interested in testing a hypothesis about a single linear combination of the parameters (say whether two parameters are identical). Finally, Section 4.5 in Wooldridge discusses the F test for testing multiple linear restrictions, and briefly discusses the relationship between the F and t statistics.

5.2 Content of chapter

5.2.1 Sampling distribution of the OLS estimators

Read: Wooldridge Section 4.1, and Appendices B.5a, B.5b, B.5c.

It is important to realise that in this chapter we assume that **all Gauss–Markov assumptions hold** and, in addition, that the errors are independent of the regressors and have a **normal distribution** (independence is really stronger than needed). These convenient assumptions, also labelled the classical regression model assumptions (CLM), enable us to conduct statistical inference as it permits us to obtain the following standardised sampling distribution of $\widehat{\beta}_j$:

$$\frac{\widehat{\beta}_j - \beta_j}{sd(\widehat{\beta}_j)} \sim Normal(0,1) \tag{*}$$

where $sd(\widehat{\beta}_j)$ denotes the standard deviation (square root of the variance) of $\widehat{\beta}_j$.

The normal distribution is very convenient. Moreover, all tests we will construct based on MLR.1 to MLR.6 will turn out to be valid in **large samples** even when the normality assumption, MLR.6, is not made explicit.

You are expected to be familiar with properties of the normal distribution and be able to use the statistical table of the $Normal(0,1)$ distribution (Table G.1 in Wooldridge).

> **Activity 5.1** Take the MCQs related to this section on the VLE to test your understanding.

5.2.2 Testing hypotheses about a single population parameter: The *t* test

Read: Wooldridge Section 4.2 and Appendix B.5e.

You should make sure they understand the general principles underlying testing the significance of a particular population parameter first (in other words, is the partial effect of x_j on y zero: i.e. is $\beta_j = 0$?).

Intuitively: We should reject $H_0 : \beta_j = 0$ when the realisation of our OLS estimator $\widehat{\beta}_j$ is not close to zero.

- To determine whether this is the case, we need to **look at the sampling distribution of $\boldsymbol{\widehat{\beta}_j}$ under $\boldsymbol{H_0}$**, as this distribution reflects the likely values that our estimator will take when the null is true under repeated sampling from the same population.
- Under H_0, this distribution (bell-shaped) will be centred around the true value (zero in this case). Realisations in the tails of this distribution (far from the centre) suggest we should reject H_0.

Formally: To test this (and other hypotheses), we will look for a **test statistic** that satisfies the following two criteria under H_0.

1. The test statistic has a *nice distribution.* (Allows us to use statistical tables to determine *critical values.*)
2. The test statistic *does not contain any unknown quantities*, such as σ^2, the variance of the error. (Allows us to compute it given the data.)

This latter statement is the reason why it is useful to recognise that under MLR.1 to MLR.6, by replacing $sd(\widehat{\beta}_j)$ with $se(\widehat{\beta}_j)$ in (*) above, we obtain a t distribution for the standardised estimators:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \sim t_{n-k-1} = t_{df}. \qquad (**)$$

The t distribution is fully characterised by the degrees of freedom, which is given here by $df = n - k - 1$ (sample size minus the number of parameters – intercept and k slopes).

To test the hypothesis $H_0 : \beta_j = 0$, we should use the ***t* statistic**:

$$t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)}.$$

Our test statistic $t_{\widehat{\beta}_j}$ satisfies the two requirements: we can obtain a realisation of the t test for a given sample as $se(\widehat{\beta}_j)$ can be computed using the data unlike $sd(\widehat{\beta}_j)$, and the distribution under the null is nice. Using (**) given above, we should conclude that under the null, when $\beta_j = 0$, this implies that:

$$t_{\widehat{\beta}_j} \sim t_{n-k-1}.$$

For testing $H_0 : \beta_j = 0$, we therefore rely on the following result:

$$\text{Under } H_0 : t_{\widehat{\beta}_j} = \frac{\widehat{\beta}_j}{se(\widehat{\beta}_j)} \sim t_{n-k-1}.$$

Our test should now reject when our realisation of $\widehat{\beta}_j$ is far away from zero relative to its standard error, $se(\widehat{\beta}_j)$. The decision rule (whether we should reject or not) will depend on:

- the alternative hypothesis we want to consider: one-sided or two-sided

- the significance level of our test. This denotes the probability of rejecting H_0 when it is true, or equivalently, denotes our willingness to commit a Type I error. Typical value of the significance level, denoted α, is 5% or 1%.

With the help of empirical examples, we see how to conduct both one-sided and two-sided tests and the graphical discussion of the rejection rule should be well understood.

When conducting hypothesis tests, it is important to include all ingredients of the following testing recipe:

1. Provide H_0 and H_1 (denoted in terms of the true population parameters!).
2. The test statistic and its distribution under H_0.
3. The significance level and rejection rule (critical value).
4. The assumptions underlying the test (MLR.1 to MLR.6).
5. Economic interpretation of our decision.

Violations of MLR.1 to MLR.6, such as omission of relevant variables, or the presence of heteroskedasticity, **will invalidate the test**, meaning we cannot trust the conclusion of our testing procedure.

In Section 4–2c the above discussion is generalised to the setting where we want to test another hypothesis about β_j, say $H_0 : \beta_j = 1$. Under CLM assumptions, we can use the following t statistic:

$$t = \frac{\widehat{\beta}_j - 1}{se(\widehat{\beta}_j)} \sim t_{n-k-1} \text{ under } H_0.$$

We should reject when the realisation of $\widehat{\beta}_j$ is far from 1 relative to $se(\widehat{\beta}_j)$, which is the same as rejecting when t is far away from zero.

It is good to point out here that there is a benefit associated with using a one-sided test instead of a two-sided test when the direction of violations to the null are obvious (e.g. when testing whether the returns to education are zero). When we compare the rejection rules for one-sided and two-sided alternatives, we should recognise that for the same level of significance, the critical value is in absolute value smaller when using a one-sided alternative compared to a two-sided alternative. This reveals that for the same level of significance (and knowing the direction where violations from the null appear) we will start rejecting earlier using one-sided alternatives than two-sided alternatives. As this ensures that we will also be rejecting earlier when the null is wrong, this shows that one-sided hypothesis tests have the benefit of being more powerful (recall: the power of a test is our ability to reject the null when the null is false).

Instead of conducting a hypothesis test for a given level of significance, it may be more informative to answer the question: 'What is the lowest level of significance at which you would reject the null given the realisation of your test statistic?'. This level is known as the ***p*-value**. It signals the probability of observing a more extreme relation of our test statistic (favouring the alternative hypothesis) than the realisation of our test statistic. If the p-value is lower than α (the significance level), then we should reject the null. Again, a graphical discussion of how to obtain the p-values (one-sided and two-sided) is most illustrative and should be well understood. It is good to recognise that the p-value of a one-sided test is half the p-value of a two-sided test.

Activity 5.2 Take the MCQs related to this section on the VLE to test your understanding.

Activity 5.3 Exercise 4.2 from Wooldridge and also answer the following.

(iv) Show graphically how you can obtain the p-value of the test $H_0 : \beta_3 = 0$ against $H_1 : \beta_3 > 0$.

Activity 5.4 Exercise 4.4 from Wooldridge and also answer the following.

(v) Test the hypothesis that $H_0 : \beta_2 = 0.4$ against $H_1 : \beta_2 \neq 0.4$ at the 5% level. Interpret your result.

5.2.3 Confidence intervals

Read: Wooldridge Section 4.3 and Appendices C.5a, C.5b.

In this section we are shown that under CLM assumptions, the 95% confidence interval for the population parameter β_j, is given by:

$$\left[\widehat{\beta}_j - t_{df,2.5\%} \times se(\widehat{\beta}_j), \widehat{\beta}_j + t_{df,2.5\%} \times se(\widehat{\beta}_j)\right]$$

where $df = n - k - 1$. Properly interpreting confidence intervals is tricky. We should recognise that the limits of the interval will depend on the particular sample we happen to observe as both $\widehat{\beta}_j$ and $se(\widehat{\beta}_j)$ will change from sample to sample.

- Our 95% confidence interval tells us therefore that under repeated sampling from the same population, in 95% of the samples the interval we compute will contain the true value β_j.
- For any particular sample, we do not know whether β_j lies in the interval or not.

We should understand how to use confidence intervals for hypothesis testing purposes, and appreciate the important link between the confidence level and the significance level of our test. If a_j does not lie in the **95% confidence interval**, then we would reject the null that $\beta_j = a_j$ at the **5% level of significance**.

Activity 5.5 Take the MCQs related to this section on the VLE to test your understanding.

Activity 5.6 As an extension of Activity 5.4, use Exercise 4.4 from Wooldridge to answer the following.

(vi) Provide the 99% confidence interval for β_3 (parameter on *pctstu*).

(vii) Provide the 95% confidence interval for β_2 (parameter on $\log(avginc)$).

5.2.4 Testing hypotheses about a single linear combination of the parameters

Read: Wooldridge Section 4.4.

It is important that you study this section carefully. It reveals that under the CLM assumptions we can use the t test to test **any** single linear restriction on our parameters. Examples of this are whether **the effect of two variables is identical** (see discussion in textbook), or the example considered here whether **the sum of two parameters equals one** (Cobb–Douglas production function):

$$\log(output) = \beta_0 + \beta_1 \log(capital) + \beta_2 \log(labour) + u \qquad (*)$$

where $H_0 : \beta_1 + \beta_2 = 1$ suggests the presence of 'constant returns to scale' in an industry.

The main difficulty of the implementation of t tests in these settings is the computation of the standard errors that appear in the denominator of our test statistic. To test $H_0 : \beta_1 + \beta_2 = 1$, for instance, we should use:

$$t = \frac{\widehat{\beta}_1 + \widehat{\beta}_2 - 1}{se(\widehat{\beta}_1 + \widehat{\beta}_2 - 1)} \equiv \frac{\widehat{\beta}_1 + \widehat{\beta}_2 - 1}{se(\widehat{\beta}_1 + \widehat{\beta}_2)} \sim t_{df} \text{ under } H_0$$

where $df = n - 3$ (we estimate here three parameters). Depending on the alternative (one-sided or two-sided) and the significance level, this distribution allows us to specify the critical values that determine whether we should reject the null or not. **To obtain the standard error is a bit more involved** as knowing $se(\widehat{\beta}_1)$ and $se(\widehat{\beta}_2)$ is not enough, we need to recognise here:

$$se(\widehat{\beta}_1 + \widehat{\beta}_2) = \sqrt{[se(\widehat{\beta}_1)]^2 + [se(\widehat{\beta}_2)]^2 + 2s_{12}}$$

where s_{12} is the estimate of $Cov(\widehat{\beta}_1, \widehat{\beta}_2)$.

An alternative method we could use instead, requires us to **reformulate (reparameterise) our model** (*). We will want to introduce a new parameter $\theta = \beta_1 + \beta_2$, as this will permit us to test $H_0 : \theta = 1$, using:

$$t = \frac{\widehat{\theta} - 1}{se(\widehat{\theta})} \sim t_{df} \text{ under } H_0.$$

To see what regression we need to run to obtain an estimate of θ and its standard error, let us rewrite $\theta = \beta_1 + \beta_2$, as:

$$\beta_1 = \theta - \beta_2.$$

Plugging this expression into our original model yields:

$$\log(output) = \beta_0 + (\theta - \beta_2)\log(capital) + \beta_2 \log(labour) + u.$$

Upon rearranging, we can write this as:

$$\log(output) = \beta_0 + \theta \log(capital) + \beta_2(\log(labour) - \log(capital)) + u.$$

Using the properties of the log operator this equals:

$$\log(output) = \beta_0 + \theta \log(capital) + \beta_2 \log(labour/capital) + u.$$

Once we run a regression of $\log(output)$ on an intercept, $\log(capital)$ and $\log(labour/capital)$ we obtain $\widehat{\theta}$ and $se(\widehat{\theta})$ that allow us to obtain our test statistic.

> **Activity 5.7** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 5.8** Exercise 4.8 from Wooldridge.

5.2.5 Testing multiple linear restrictions: The F test

Read: Wooldridge Section 4.5 and Appendix B.5f.

Often we may want to test multiple hypotheses about the underlying parameters simultaneously. Specific examples discussed in this section include (i) testing the significance of the regression, which involves testing the hypothesis that all slope parameters are zero, and (ii) testing whether, after controlling for some dependent variables, other independent variables have no effect on the dependent variable.

The F test we discuss for this purpose involves comparing the 'fit' of the original model with the 'fit' of the model where the restrictions are imposed. To obtain the latter (also called the restricted model) we simply need to plug in the restrictions in our original model (also called the unrestricted model). The test wants to ascertain whether the loss of fit by imposing restrictions is statistically significant.

Important: The F test discussed assumes all Gauss–Markov assumptions and the normality of the errors (i.e. the CLM assumptions MLR.1 to MLR.6) are satisfied.

A nice formulation of the F statistic is given by:

$$F = \frac{(SSR_r - SSR_{ur})/\text{number of restrictions}}{SSR_{ur}/df \text{ of unrestricted model}}$$

where SSR_r denotes the residual sum of squares of the restricted model and SSR_{ur} the residual sum of squares of the unrestricted model. Under the CLM assumptions this test statistic has an F distribution under the null which is fully characterised by the degrees of freedom of the numerator (number of restrictions) and the degrees of freedom of the denominator (df of the unrestricted model):

$$\text{Under } H_0 : F \sim F_{\text{number of restrictions}, df \text{ of unrestricted model}}.$$

Note: Our test statistic satisfies both requirements: (i) we can compute it given the data, and (ii) it has a nice distribution under the null.

Intuitively, we should reject the null if the loss in fit is statistically significant. Formally, we should reject the null at a particular level of significance α if the realisation of our test statistic exceeds the critical value. Denoting the critical value (see Table G.3 in Wooldridge) by $F_{q,\,df_{ur},\alpha}$, with q denoting the number of restrictions, and df_{ur} the degrees of freedom of the unrestricted model, that is:

$$\text{Reject } H_0 : F > F_{q,df_{ur},\alpha}$$

$$\text{Not reject } H_0 : F \leq F_{q,\,df_{ur},\alpha}.$$

As with the t test, you should know how to obtain the p-value and how they can be used for inference purposes.

An alternative formulation of the F test is based on the R^2 of the restricted and unrestricted models instead of their residual sum of squares. Recall, our goodness of fit measure $R^2 = 1 - SSR/SST$. Therefore, the F test can also be written as:

$$F = \frac{(R^2_{ur} - R^2_r)/\text{number of restrictions}}{(1 - R^2_{ur})/df \text{ unrestricted model}}.$$

Since the R^2 is commonly reported for regression results, unlike the SSR, this is a convenient form.

Important: Care should be taken when using this formation of the F test, that the R^2 of both the restricted and unrestricted models provide the proportion of the variance of the **same** dependent variable that can be explained by the model. We can only compare the R^2 of two models when the dependent variable is identical. Furthermore, as $R^2_{ur} \geq R^2_r$, $F \geq 0$.

Two further results discussed are noteworthy:

1. The test of the significance of the regression (null that all slope parameters are zero) can be written in a form that shows a close relation to the goodness of fit of the original model (k slopes and 1 intercept):

$$F = \frac{R^2/k}{(1 - R^2)/(n - k - 1)}.$$

   Clearly when our R^2 is large our independent variables do a good job in explaining the variation in the dependent variable and hence we should reject the null that all slopes are zero!

2. We can also use the F test to test a **single** linear restriction. In fact the F statistic equals the square of the t statistic. The advantage of using the t statistic instead (not the square of the test statistic) is that we can use a one-sided test.

> **Activity 5.9** Take the MCQs related to this section on the VLE to test your understanding.

Activity 5.10 Exercise 4.6 from Wooldridge and also answer the following.

(v) Show in a graph how you can obtain the p-value of the test conducted in (iii).

Activity 5.11 Exercise 4.10 from Wooldridge.

Activity 5.12 Exercise 4.12 from Wooldridge.

5.2.6 R Application for multiple regression analysis: Inference

Complete 'R Task – Activity 3' on VLE.

Here we show how to test single and multiple linear hypothesis tests in R.

5.3 Answers to activities

Solution to Activity 5.3

(Exercise 4.2 from Wooldridge and extension.)

(i) $H_0 : \beta_3 = 0$ and $H_1 : \beta_3 > 0$.

(ii) The proportionate effect on *salary* is $.00024 \times 50 = .012$.

To obtain the percentage effect, we multiply this by 100. The 50% point *ceteris paribus* increase in *ros*, therefore, is predicted to increase *salary* by only 1.2%. Practically speaking, this is a very small effect for such a large change in *ros*.

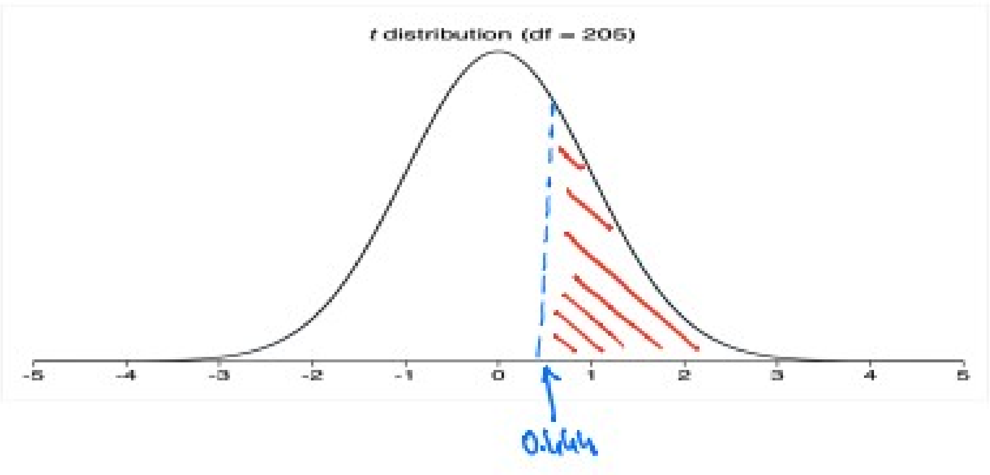
(iii) To test the hypothesis we use the t statistic $\widehat{\beta}_3/se(\widehat{\beta}_3)$, which under H_0 is distributed as t_{df} with $df = 209 - 4 = 205$. The realisation of our test statistic equals $.0024/.00054 \approx 0.444$.

At the 10% level of significance, the one-sided critical value equals $t_{df,10\%} = 1.282$. As the realisation of our t statistic lies well below this critical value, we conclude that we find no evidence for a statistically significant positive effect.

The test assumes MLR.1 to MLR.6 are satisfied.

(iv) Based on this sample, the estimated *ros* coefficient appears to be different from zero only because of sampling variation. Still, including irrelevant variables may not be causing any harm.

(v) Under H_0 the test statistic is distributed as t_{205} and the realisation of our test statistic was 0.444. We need both to indicate the p-value graphically:



Solution to Activity 5.4

(Exercise 4.4 from Wooldridge and extension.)

(i) $H_0 : \beta_3 = 0$ and $H_1 : \beta_3 \neq 0$.

(ii) Other things being equal, a larger population increases the demand for rental housing, which should increase rents; $\beta_1(+)$.

The demand for overall housing is higher when average income is higher, pushing up the cost of housing, including rental rates; $\beta_2(+)$.

Both are elasticities!

(iii) The coefficient on $\log(pop)$ is an elasticity. The correct statement is that 'a 10% increase in population increases rent by 0.66% (= .066 × 10) *ceteris paribus*'.

(iv) A 1-unit increase in *pctstu* (that is a 1% point increase, as it is measured in percentages) results in a .56% increase in rent, *ceteris paribus.*

(iv) The t statistic $\widehat{\beta}_3/se(\widehat{\beta}_3) = .0056/.0017$ is about 3.29.

Given our CLM assumptions (MLR.1 to MLR.6), we know:

$$\text{under } H_0 : t \sim t_{df} \quad \text{where } df = 64 - 4 = 60.$$

At the 1% level of significance, the two-sided critical value equals $t_{df,0.5\%} = 2.660$.

As our t statistic is well above the critical value, we conclude that $\widehat{\beta}_3$ is statistically different from zero at the 1% level. The student population as a percentage of total population affects rent rates.

(v) The t statistic for this hypothesis equals $(\widehat{\beta}_2 - 0.4)/se(\widehat{\beta}_2)$ which is about 1.32.

Given our CLM assumptions (MLR.1 to MLR.6), we know:

$$\text{under } H_0 : t \sim t_{df} \quad \text{where } df = 64 - 4 = 60.$$

Under H_0 our test statistic is $t \sim t_{df}$ with $df = 64 - 4 = 60$. At the 5% level of significance, the two-sided critical value equals $t_{df,2.5\%} = 2.00$.

As our t statistic lies below this critical value, we conclude that $\widehat{\beta}_2$ is not statistically different from 0.4 at the 5% level. That is, the elasticity is not significantly different from .4%, *ceteris paribus*.

Test assumes MLR.1 to MLR.6 are satisfied.

Solution to Activity 5.6

(Extension Exercise 4.4 from Wooldridge.)

(vi) Using the same critical value as in (iv), $t_{60,\,0.5\%} = 2.660$ we obtain the 99% confidence interval as:

$$\left[\widehat{\beta}_3 - 2.66 \times se(\widehat{\beta}_3), \widehat{\beta}_3 + 2.66 \times se(\widehat{\beta}_3)\right]$$

$$\Rightarrow \quad [.0056 - 2.66 \times .0017, .0056 + 2.66 \times .0017] = [.0011, .010].$$

Zero does not lie in this interval confirming the result in (d): We should reject H_0 at the 1% level of significance.

(vii) Using the same critical value as in (v), $t_{60,\,2.5\%} = 2.00$ we obtain the 95% confidence interval as:

$$\left[\widehat{\beta}_2 - 2.00 \times se(\widehat{\beta}_2), \widehat{\beta}_2 + 2.00 \times se(\widehat{\beta}_2)\right]$$

$$\Rightarrow \quad [.507 - 2.00 \times .081, .507 + 2.00 \times .081] = [.345, .669].$$

As 0.4 lies in this interval, our result in (e) is confirmed: We should not reject H_0 at the 5% level of significance.

Solution to Activity 5.8

(Exercise 4.8 from Wooldridge.)

(i) Recall the property of the variance: If W and Z are random variables and a and b constants:

$$Var(aW + bZ) = a^2\,Var(W) + b^2\,Var(Z) + 2ab\,Cov(W, Z).$$

The same result holds for the conditional variance. So we obtain:

$$Var(\widehat{\beta}_1 - 3\widehat{\beta}_2 \,|\, \mathbf{X}) = Var(\widehat{\beta}_1 \,|\, \mathbf{X}) + 9\,Var(\widehat{\beta}_2 \,|\, \mathbf{X}) - 6\,Cov(\widehat{\beta}_1, \widehat{\beta}_2 \,|\, \mathbf{X}).$$

Since the standard error equals the square root of the estimated variance (obtained by replacing the error variance σ^2 with its estimator s^2):

$$se(\widehat{\beta}_1 - 3\widehat{\beta}_2) = \sqrt{se(\widehat{\beta}_1)^2 + 9se(\widehat{\beta}_2)^2 - 6s_{12}}$$

where s_{12} is an estimate of $Cov(\widehat{\beta}_1, \widehat{\beta}_2 \,|\, \mathbf{X})$.

(ii) The t statistic to test $H_0 : \beta_1 - 3\beta_2 = 1$ is:

$$t = \frac{\widehat{\beta}_1 - 3\widehat{\beta}_2 - 1}{se(\widehat{\beta}_1 - 3\widehat{\beta}_2)}.$$

Note: $se(\widehat{\beta}_1 - 3\widehat{\beta}_2 - 1) = se(\widehat{\beta}_1 - 3\widehat{\beta}_2)$.

In order to compute the test statistic we need the standard error of $\widehat{\beta}_1 - 3\widehat{\beta}_2$, which, in addition to knowing $se(\widehat{\beta}_1)$ and $se(\widehat{\beta}_2)$, requires s_{12}.

(iii) We can rewrite $\theta_1 = \beta_1 - 3\beta_2$, as:

$$\beta_1 = \theta_1 + 3\beta_2.$$

Plugging this into the population model gives:

$$y = \beta_0 + (\theta_1 + 3\beta_2)x_1 + \beta_2 x_2 + \beta_3 x_3 + u.$$

Rewriting gives:

$$y = \beta_0 + \theta_1 x_1 + \beta_2(3x_1 + x_2) + \beta_3 x_3 + u.$$

By running a regression of y on x_1, $3x_1 + x_2$, and x_3 we can directly obtain $\widehat{\theta}$ and its standard error, and $(\widehat{\theta}_1 - 1)/se(\widehat{\theta}_1)$ gives the same test statistic as in (ii).

Solution to Activity 5.10

(Exercise 4.6 from Wooldridge and extension.)

(i) For both hypotheses we should use the t test, and under H_0 $t \sim t_{df}$ with $df = n - 2 = 86$ (MLR.1 to MLR.6).

The critical value at the 5% level is approximately $t_{90,\,2.5\%} = 1.987$ (two-sided).

- First hypothesis: $t = \widehat{\beta}_0/se(\widehat{\beta}_0) = -14.47/16.27 \approx -.89$. We fail to reject $H_0 : \beta_0 = 0$.
- Second hypothesis: $t = (\widehat{\beta}_1 - 1)/se(\widehat{\beta}_1) = (.976 - 1)/.049 \approx -.49$. Fail to reject $H_0 : \beta_1 = 1$.

Remember, we reject H_0 in favour of H_1 only if $|t| > 1.987$.

(ii) To test the joint hypothesis that $\beta_0 = 0$ and $\beta_1 = 1$, we need the SSR in the restricted model. This turns out to yield $SSR = 209{,}448.99$. Carry out the F test for the joint hypothesis.

■ Here we should use the SSR form of the F statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}.$$

The restricted model is given by $price = 0 + 1 \times assess + u$, which does not contain any parameters (no need to estimate): $SSR_r = \sum_{i=1}^{n}(price_i - assess_i)^2$.

- Here $q = 2$ (number of restrictions) and $n - k - 1 = df_{ur} = 88 - 2 = 86$.

- We are given $SSR_r = 209{,}448.99$ and $SSR_{ur} = 165{,}644.51$.
- Our test statistic, therefore, equals:

$$F = \frac{(209{,}448.99 - 165{,}644.51)/2}{165{,}644.51/86} \approx 11.37.$$

- Given MLR.1 to MLR.6, under $H_0 : F \sim F_{2,86}$.
- At the 1% significance level, our critical value $c = F_{2,90,1\%} = 4.85$. Hence we have found strong evidence (1% level) against the assumption of rational assessment as 11.37 exceeds the critical value.

(iii) Here we should use the R^2 form of the F statistic:

$$F = \frac{(R^2_{ur} - R^2_r)/q}{(1 - R^2_{ur})/(n-k-1)}.$$

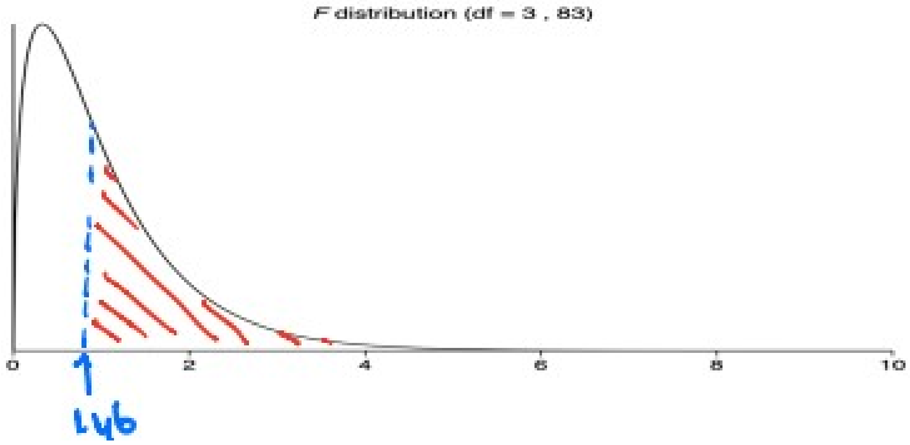
The restricted model is the simple regression model, the unrestricted model has 3 more regressors.

- We are testing $q = 3$ restrictions $n - k - 1 = df_{ur} = 88 - 5 = 83$.
- We are given $R^2_r = .820$ and $R^2_{ur} = .829$.
- Our test statistic, therefore, equals:

$$F = \frac{(.829 - .820)/3}{1 - .829/83} \approx 1.46.$$

- Given MLR.1 to MLR.6, under $H_0 : F \sim F_{3,83}$.
- At the 10% level of significance, our critical value is approximately $F_{3,90,10\%} = 2.15$.
- We fail to reject H_0 even at the 10% level. After controlling for *assess* the additional regressors are jointly not significant.

(iv) Recall: Under H_0 the test statistic F is distributed as $F_{3,83}$ and the realisation of our test statistic was 1.46. We need both to indicate the p-value graphically.



Solution to Activity 5.11

(Exercise 4.10 from Wooldridge.)

(i) We need to compute the F statistic for the overall significance of the regression with $n = 142$ and $k = 4$: $F = \dfrac{.0395/4}{(1{-}.0395)/137} = 1.41$.

The 5% critical value with 4 for the numerator df and using 120 for the denominator df, is 2.45, which is well above the value of F. Therefore, we fail to reject $H_0 : \beta_{dkr} = \beta_{eps} = \beta_{nentinc} = \beta_{salary} = 0$ at the 5% level. With the t test we conclude that no explanatory variable is individually significant at the 5% level. The largest absolute t statistic is on *dkr*, $t_{dkr} = 1.60$, which is not significant at the 5% level against a two-sided alternative.

(ii) The F statistic (with the same df) is now $F = \dfrac{.0330/4}{(1{-}.0330)/137} = 1.17$, which is even lower than in part (i). None of the t statistics is significant at a reasonable level.

(iii) Both proposals are problematic as the use of logarithms necessitates that the original variables are positive only. It will generate missing values and these firms will be dropped, which will be a problem as you will no longer have a random sample.

(iv) It seems very weak. There are no significant t statistics at the 5% level (against a two-sided alternative), and the F statistics are insignificant in both cases. Plus, less than 4% of the variation in returns is explained by the independent variables.

Solution to Activity 5.12

(Exercise 4.12 from Wooldridge.)

(i) A 10% increase in spending is associated with a 1.1 unit increase in the mathematics pass rate, that is a 1.1 percentage point increase (as *math10* is measured in percentages).

The estimated intercept is the predicted value of the dependent variable when all regressors are set to zero. As the minimum value of *lexpend* is 8.11, this does not seem to be reasonable; a prediction of a pass rate equalling -69 indeed makes no sense.

(ii) The low R^2 means that *lexpend* alone does not help explain a lot of the variation in *math10* (but it also does not suggest that there is no correlation between the two). There are indeed likely to be many omitted factors that will help explain the variation in *math10* (see examples in (iii)).

If expenditure levels were randomly assigned to schools it is unlikely that the amount of variation in *math10* explained by spending could be more than a few percent of the total variation.

(iii) The coefficient on *lexpend* falls to 7.75, but its t statistic is 2.55, which is statistically significant even at the 1% level. Recognising that *lexpend* and *lnchprg* are likely correlated suggests that the simple regression estimates are likely to exhibit OVB.

(iv) Adding *lenroll* and *lnchprg* (essentially the poverty rate) allows us to explain notably more variation in *math10*, but it is still only about 19%. One might improve the explanatory power by including variables such as average family income and measures of parental education levels. Or, features of the teachers' qualifications.

5.4 Overview of chapter

In this chapter we discussed how to test linear hypotheses about the parameters in the population regression model. For this purpose, we added the assumption that the unobserved error is normally distributed while continuing to rely on the Gauss–Markov assumptions. We called this our *Classical Linear Model.*

We discussed the t test for testing single linear restrictions and discussed the F test for testing multiple linear restrictions.

For a given level of significance (our willingness to commit a Type I error), we need to obtain the critical value to determine whether we should reject the null hypothesis or not. When conducting hypothesis testing in an examination setting, you are advised to follow the testing recipe provided that clearly indicates the ingredients required.

We showed how to obtain p-values and discussed their relevance for inference. We also discussed how to construct confidence intervals for our population parameters and discussed how they can be used for hypothesis testing.

5.4.1 Key terms and concepts

- Normality assumption
- Classical linear model (CLM) and Classical linear model assumptions
- Hypothesis test
- Type I error
- Type II error
- Null hypothesis
- Alternative hypothesis
- One-sided and two-sided alternative
- Test statistic
- Significance level

- Critical value
- Rejection region
- Confidence interval (interval estimator)
- t statistic
- Unrestricted and restricted model
- Joint hypothesis test
- F statistic
- Test for the significance of the regression
- p-value
- Power of a test

5.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to

- understand the meaning of the classical linear model (CLM)
- contrast the sampling distribution of $(\widehat{\beta}_j - \beta_j)/se(\widehat{\beta}_j)$ under MLR.1 to MLR.6 with the sampling distribution of $(\widehat{\beta}_j - \beta_j)/sd(\widehat{\beta}_j)$
- perform the t test for testing a hypothesis about a single population parameter
- understand the benefit of using one-sided t test over two-sided t test
- obtain p-values for t tests (one-sided and two-sided)
- construct confidence interval (CI) for population parameters and explain its interpretation
- use confidence interval (CI) for hypothesis testing against two-sided alternatives
- explain the meaning of linear restrictions
- perform the t test for testing a hypothesis about a single linear restriction of the parameters and understand how to obtain the standard error of linear combinations of the estimators
- formulate the F test in two equivalent forms: (i) based on comparing the SSR of the restricted and unrestricted models or (ii) based on a comparison of the R^2 of the restricted and unrestricted models. Interpret the F test as a test for the loss in goodness of fit
- perform the F test for testing multiple (joint) linear restrictions of the population parameters, and perform the F test for the overall significance of a regression

- obtain p-values for F tests
- explain the relation between the F test and the t test when testing a single linear restriction
- use statistical software to implement the t test and the F test when analysing the relationship between real-world economic variables.

5.5 Test your knowledge and understanding

5.5.1 Additional examination based exercises

5.1E Let us consider a model for the sale price of Monet paintings. The data we have contain the sale prices, widths, and heights of 430 Monet paintings, which sold at auction for prices ranging from \$10,000 to \$33 million. A linear regression provided the following result:

$$\widehat{\log \text{Price}}_i = \underset{(0.612)}{-8.427} + \underset{(0.091)}{1.334} \log \text{Area}_i - \underset{(0.128)}{0.165} \text{Aspect Ratio}_i$$
$$n = 430, \;\; R^2 = 0.336$$

where Area = Width × Height and Aspect Ratio = Height/Width. The standard errors are given in parentheses. We will assume all Gauss–Markov assumptions are satisfied.

(a) Interpret the parameters and test the significance of each. What additional assumption do you need to make?

(b) Explain whether the p-value for the test of significance of the Aspect Ratio is larger or smaller than 0.05. Sketch a graph to indicate how the p-value is obtained.

(c) Provide the 95% confidence interval for $\beta_{\log \text{Area}}$. Discuss how you can use this interval to test the significance of log Area.

(d) Test the joint significance of the regression. Discuss its relation to the goodness of fit measure: R^2.

(e) You want to test $H_0 : \beta_{\log \text{Area}} = 1$ against $H_1 : \beta_{\log \text{Area}} > 1$. Perform this test using the information provided and provide an economic interpretation of your result. In view of your answer, indicate whether the p-value for the test is bigger or smaller than 5% and provide a sketch that indicates how the p-value is obtained.

5.2E Let us consider the estimation of a hedonic price function for houses. The hedonic price refers to the implicit price of a house given certain attributes (e.g. the number of bedrooms). The data contains the sale price of 546 houses sold in the summer of 1987 in Canada along with their important features. The following characteristics are available: the lot size of the property in square feet (*lotsize*), the numbers of bedrooms (*bedrooms*), the number of full bathrooms (*bathrooms*), and a variable that indicates the presence of an air conditioner (*airco)* (1 = yes, 0 = no).

Consider the following ordinary least squares result:

$$\widehat{\log(price)}_i = \underset{(.232)}{7.094} + \underset{(.028)}{0.400} \log(lotsize)_i + \underset{(.015)}{0.078} bedrooms_i \tag{2.1}$$

$$+ \underset{(.023)}{0.216} bathrooms_i + \underset{(.024)}{0.212} airco_i$$

$$n = 546, \ SSR = 32.622.$$

The standard errors are in parentheses, and SSR measures the residual sum of squares.

(a) Interpret the parameter estimates on $\log(lotsize)$ and *bedrooms*.

(b) Test the significance of $\log(lotsize)$ and *bedrooms*.

(c) Suppose that the lot size of the property was measured in square metres rather than square feet. How would this affect the parameter estimates of the slopes and intercept? How would this affect the fitted values? *Note:* the conversion (approximate) 1 $\text{m}^2 = 10$ ft^2.

(d) We are interested in testing the hypothesis $H_0 : \beta_{bedrooms} = \beta_{bathrooms}$ against the alternative $H_1 : \beta_{bedrooms} \neq \beta_{bathrooms}$.

i. Give an interpretation to the hypothesis of interest.

ii. Provide the t statistic we can use to test this hypothesis and indicate what we would need to estimate to obtain this test statistic.

iii. Discuss the following statement: 'Standard regression output of the following regression:

$$\log(price) = \alpha_0 + \alpha_1 \log(lotsize) + \alpha_2 bbrooms + \alpha_3 bedrooms + \alpha_4 airco + u \tag{2.2}$$

where $bbrooms = bedrooms + bathrooms$, will automatically provide us with the test statistic for our hypothesis of interest.' What is the interpretation of α_3?

iv. Discuss how you can test the hypothesis making use of the following, so-called restricted, regression result:

$$\widehat{\log(price)}_i = \underset{(.234)}{6.994} + \underset{(.282)}{0.408} \log(lotsize)_i + \underset{(.011)}{0.127} bbrooms_i + \underset{(.024)}{0.215} airco_i,$$

$$n = 546, \ SSR = 33.758.$$

Clarify how the regression incorporates the restriction.

(e) Discuss how you would test the joint significance of *bedrooms*, *bathrooms*, and *airco*. Clearly indicate what additional result you would need to be able to conduct this test and show graphically how to obtain its p-value.

5.5.2 Solutions to additional examination based exercises

5.1E (a) The parameter on log Area denotes an elasticity, a 1% increase in the *Area* of painting results in a 1.334% increase in the price of the painting, *ceteris paribus.*

To test its significance, we test $H_0 : \beta_{\log \text{Area}} = 0$ against $H_1 : \beta_{\log \text{Area}} \neq 0$ using the t test. Under CLM assumptions (MLR.1 to MRL.6):

$$\frac{\widehat{\beta}_{\log \text{Area}}}{se(\widehat{\beta}_{\log \text{Area}})} \sim t_{n-3} \quad \text{under } H_0.$$

The realisation of our test statistic is very large $1.334/.091 = 14.66$ and exceeds the two-sided critical value even at the 1% level of significance (2.576) revealing its statistical significance. Conclude: *Area* matters in explaining *Price* after controlling for *Aspect Ratio*.

The parameter on *Aspect Ratio* suggests that a one-unit increase in the Aspect Ratio results in a 16.5% reduction in the price of the Monet painting, *ceteris paribus*.

Using the t test here does not reveal a statistically significant effect. The t statistic equals $-.165/.128 = -1.29$, which is even below the two-sided critical value at the 10% level of significance. Conclude: After controlling for $\log(Area)$, *Aspect Ratio* does not have a statistically significant effect on *Price*.

(b) The p-value for the test of significance is larger than .05, as we cannot reject the null at the 5% level of significance as discussed in (a). Graph of the p-value should clearly show area in both tails of a Student's t distribution with $n - 3$ degrees of freedom; test statistic taking realisations in absolute value in excess of 1.29.

(c) Under the CLM assumptions (MLR.1 to MLR.6), the 95% confidence interval for $\beta_{\log \text{Area}}$ is given by:

$$\left[\widehat{\beta}_{\log \text{Area}} - 1.96 \times se(\widehat{\beta}_{\log \text{Area}}), \widehat{\beta}_{\log \text{Area}} + 1.96 \times se(\widehat{\beta}_{\log \text{Area}})\right] = [1.16, 1.51]$$

where 1.96 is obtained from the Student's t distribution with large degrees of freedom (same as $N(0, 1)$ therefore). As zero does not lie in this 95% confidence interval, we can conclude that we should reject the null $H_0 : \beta_{\log \text{Area}} = 0$ against $H_1 : \beta_{\log \text{Area}} \neq 0$ at the 5% level of significance.

(d) Here we are asked to test the joint hypothesis $H_0 : \beta_{\log \text{Area}} = 0$ and $\beta_{\text{Aspect Ratio}} = 0$ against H_1 : at least one non-zero. Under the CLM assumptions (MLR.1 to MLR.6) we should use the F statistic:

$$F = \frac{(SSR_r - SSR_{ur})/2}{SSR_{ur}/(n-3)} = \frac{R^2/2}{(1-R^2)/(n-3)} \sim F_{2,n-3} \quad \text{under } H_0.$$

We divide the numerator by 2 as there are two restrictions and the denominator by $n - 3 = 427$, the degrees of freedom of the unrestricted model (3 parameters). The realisation of the test statistic equals $\frac{.336/2}{.664/427} = 108.04$ which is large, reflective of their joint significance. At the 1% level of significance the critical value equals 4.61.

A large goodness of fit measure R^2 ensures that the F statistic is large as well, both resulting in evidence that the regressors help to explain the dependent variables.

(e) Under the CLM assumptions (MLR.1 to MLR.6), we will use the t statistic, here:

$$\frac{\widehat{\beta}_{\text{log Area}} - 1}{se(\widehat{\beta}_{\text{log Area}})} \sim t_{427} \text{ under } H_0.$$

The realisation of the test statistic equals $(1.334 - 1)/0.091 = 3.6703$. At the 5% level we need to reject when it exceeds 1.645 (one sided), which it does. We therefore find evidence that auction prices are elastic. As we reject our hypothesis at the 5% level, the p-value is smaller than .05.

5.2E (a) On average, holding the remaining variables in the regression constant, a 1% increase in lot size is associated with a 0.4% increase in house price, and each extra bedroom is associated with a 7.8% increase in house price.

(b) Under the CLM assumptions (MLR.1 to MLR.6) we will use the t test to test these hypotheses (as in Question 1). For log(*lotsize*), we should test: $H_0 : \beta_{\log(lotsize)} = 0$ against $H_1 : \beta_{\log(lotsize)} \neq 0$ using:

$$\frac{\widehat{\beta}_{\text{log lotsize}}}{se(\widehat{\beta}_{\text{log lotsize}})} \sim t_{n-5} \quad \text{under } H_0.$$

The realisation of our test statistic is very large $.400/.028 = 14.28$ and exceeds the two-sided critical value even at the 1% level of significance (2.576). *bedrooms* is also statistically significant at the 1% level of significance with a realisation of the test statistic equal to $.075/.015 = 5$. Both log(*lotsize*) and *bedrooms* are therefore individually significant.

(c) Let $lotsize_i$ be the lot size in square feet and $\widetilde{lotsize_i}$ be the lot size in square metres. We have that $\widetilde{lotsize_i} = 10^{-1} lotsize_i$ and:

$$\log(\widetilde{lotsize_i}) = \log(10^{-1}) + \log(lotsize_i).$$

Since this is an additive transformation of one of the explanatory variables we have that: (i) the regression slopes will not be affected (elasticity), (ii) the intercept will change to $7.094 - \log(10^{-1}) \times 0.4$, and (iii) the fitted values will also not be affected.

(d) i. We want to test the hypothesis that the effect of an additional bedroom has the same effect as an additional bathroom on the percentage increase in house prices, *ceteris paribus.*

ii. Under the CLM assumptions (MLR.1 to MLR.6) we can use the t statistic to test this hypothesis. The test statistic is given by:

$$\frac{\widehat{\beta}_{bedrooms} - \widehat{\beta}_{bathrooms}}{se(\widehat{\beta}_{bedrooms} - \widehat{\beta}_{bathrooms})} \sim t_{n-5} \quad \text{under } H_0.$$

In order to obtain this test statistic we would need to obtain an estimate of $Cov(\widehat{\beta}_{bedrooms}, \widehat{\beta}_{bathrooms})$, which we shall call s_{23}. This is needed to obtain the standard error of the difference, with:

$$se(\widehat{\beta}_{bedrooms} - \widehat{\beta}_{bathrooms}) = \sqrt{se(\widehat{\beta}_{bedrooms})^2 + se(\widehat{\beta}_{bathrooms})^2 - 2s_{23}}.$$

iii. Let us substitute *bbrooms* = *bedrooms* + *bathrooms* in (2.2), this reveals that:

$$\log(price) = \alpha_0 + \alpha_1 \log(lotsize) + \alpha_2 bathrooms + (\alpha_2 + \alpha_3) bedrooms + \alpha_4 airco + u.$$

This shows that the effect of *bathrooms* and *bedrooms* is equal when $\alpha_3 = 0$. Our test therefore equals a test of $H_0 : \alpha_3 = 0$, which forms part of standard regression output when we estimate (2.2). This reveals the advantage of reparameterisation for hypothesis testing.

We should interpret α_3 as the additional (when positive) effect *bedrooms* have over *bathrooms* in affecting the percentage increase in house prices.

Comment: It is important to recognise the usefulness of parameterisation for testing and interpretation purposes.

iv. The regression result is obtained by considering our model (2.2) and imposing the restriction that $\alpha_3 = 0$. This is a restricted model as the effect of *bedrooms* and *bathrooms* is assumed to be identical.

Under our CLM assumptions (MLR.1 to MLR.6) we can now use the F statistic to test our hypothesis against the two-sided alternative $H_1 : \alpha_3 \neq 0$. Our F statistic is given by:

$$F = \frac{(SSR_r - SSR_{ur})/1}{SSR_{ur}/(n-5)} \sim F_{1,n-5} \quad \text{under } H_0$$

where $SSR_r = 33.758$ is the residual sum of squares of the restricted model and $SSR_{ur} = 32.622$ is the residual sum of squares of the unrestricted model. We divide the numerator by 1 (single restriction) and the denominator by the degrees of freedom of the unrestricted model $(n - 5)$.

The realisation of our test statistic equals:

$$\frac{33.758 - 32.622}{32.622/541} = 18.84.$$

At the 1% level of significance the critical value equals 6.63, so we will reject the null. The p-value will be smaller than 1%. Graph of the p-value should clearly show area in the right tail of an F distribution with $1, n-5$ degrees of freedom; test statistic taking realisations in excess of 18.84.

(e) Here we would like to test the joint hypothesis $H_0 : \beta_{\text{bedrooms}} = 0, \beta_{\text{bathrooms}} = 0$ and $\beta_{\text{airco}} = 0$ against the alternative that at least one is non-zero. As in the previous part we would perform an F test that determines whether the loss in fit is significant. We will be testing 3 restrictions, and the restricted model we need to estimate is given by:

$$\log(price) = \alpha_0 + \alpha_1 \log(lotsize) + u. \tag{2.3}$$

Using the residual sum of squares of this model, SSR_r, together with the residual sum of squares from (2.1), we compute the realisation of our test statistic:

$$F = \frac{(SSR_r - SSR_{ur})/3}{SSR_{ur}/(n-5)} \sim F_{3,n-5} \quad \text{under } H_0$$

and we reject at the 5% level of significance if the realisation:

$$\frac{(SSR_r - 32.622)/3}{32.622/541}$$

is in excess of 2.60.

Graph should display the p-value as the area where a random variable with distribution $F_{3,\,n-5}$ is more extreme than 2.60.

Chapter 6

Multiple regression analysis: Further issues and use of qualitative information

6.1 Introduction

6.1.1 Aims of the chapter

This chapter is focused on the discussion of various specifications of multiple regression models widely used by applied economists, social scientists and policymakers. In particular, it covers models with quadratics, interaction terms and dummy variables.

6.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to

- understand how to formulate and interpret models with quadratic and/or interaction terms
- be able to conduct hypothesis testing using t tests and F tests and interpret the results of such testing
- motivate the use of the adjusted R-squared
- discuss limitations of pursuing high values of R^2 or adjusted R-squared
- estimate and interpret models with quadratic and/or interaction terms in practice
- define a dummy variable and describe an impact it has on the regression specification and the interpretation of regression coefficients when it is used as a regressor
- describe how multiple dummy variables can be used to create further more refined categories in practice
- understand how ordinal information, such as ratings, can be captured by using dummy variables
- know how the choice of the base category affects the interpretation of the regression results (including the interpretation of parameters and t tests)

- explain the dummy variable trap and how it can be avoided
- motivate and describe a setting where a dummy variable is interacted with a continuous regressor
- describe and perform the Chow test in two equivalent ways (and also two variations of the Chow test – allowing or not intercepts to be different)
- estimate and interpret models with dummy variables in practice.

6.1.3 Essential reading

Reading: Wooldridge, Sections 6.2b, 6.2c, 6.3a, 6.3c, 7.1–7.4 and Appendices A.4a, A.5.

6.1.4 Synopsis of this chapter

In Section 6.2b in Wooldridge we discuss how to design, estimate, and interpret regression models that allow for decreasing or increasing partial effects or allow the regression function to have a turning point. Such models build on quadratic functions and enrich empirical specifications a researcher can use in practice. Section 6.2c in Wooldridge introduces models with interaction terms in which partial effects can depend on other regressors. We can then apply t and F tests in models with quadratics and interaction terms and draw conclusions from such hypothesis testing. Section 6.3a in Wooldridge delves into a more detailed analysis of the goodness of fit measure, discussing how the R^2 changes when more regressors are added to the model and motivates the use of the adjusted R-squared as the measure that 'penalises' a researcher for adding the extra predictor variables that do not improve the existing model. Section 6.3c in Wooldridge discusses the dangers of the pursuit of high R^2 or adjusted R-squared.

Section 7.1 in Wooldridge discusses regressors that describe qualitative information or, in other words, information about categories. Traditionally such information is captured by dummy variables. Section 7.2 discusses models with a single dummy variable. It defines the base category and teaches how to interpret regression parameters in such a setting. There is a discussion of the dangers of the dummy variable trap and instructions on how it can be avoided. Section 7.3 illustrates how in practice one can use multiple dummy variables to create further more refined categories or to capture ordinal information (such as ratings) and learn the interpretation of parameters in these more advanced settings. In particular, the chapter analyses the impact of the choice of the base category on the interpretation of t tests. Section 7.4 in Wooldridge introduces regression settings where a researcher may employ not only interactions among different dummy variables but also interactions of dummy variables with continuous regressors and discusses the meaning of such settings. The variety of regression specifications achieved by using dummy variables is illustrated with empirical applications. Finally, the knowledge of dummy variables and the F test to design the so-called Chow test that allows testing for differences among regression functions for different categories or groups.

6.2 Content of chapter

6.2.1 Functional form: Models with quadratics

Read: Wooldridge, Section 6.2b and Appendices A.4a, A.5.

It is recommended that you start by reviewing the properties of quadratic functions by reading Appendix A.4a and differentiation in Appendix A.5.

Models with quadratics are used quite often in applied economics and eonometrics to capture nonlinearities and, in particular, decreasing or increasing marginal effects. They allow us to test whether the relationship between the regressor and the dependent variable has a constant or changing marginal effect. Intuitively, we can imagine that in a public sector company wages tend to directly follow a formula based on tenure where the marginal effects would be constant, or that in a private sector company the rate of wage increases is steeper at lower tenure levels and slower later on, which would lead to decreasing marginal effects. Second, such models also allow for a turning point, which is sometimes of interest. (For example, what is the optimal number of students at a high school for high school performance?)

Consider the case when y depends on a single observed factor x, but it does so in a quadratic way:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u.$$

This is a model with two regressors, $x_1 = x$ and $x_2 = x^2$. In this model β_1 no longer measures the change in y with respect to x as it makes no sense to hold x^2 fixed while changing x. Using differentiation, we can easily deduce that the slope of y with respect to x depends on β_1 and β_2, and the value of x, i.e. we have:

$$\frac{\mathrm{d}y}{\mathrm{d}x} = \beta_1 + 2\beta_2 x \quad \text{holding } u \text{ fixed.}$$

Such a model is easy to estimate by running OLS of y on $x_1 = x$ and $x_2 = x^2$.

After estimating the model, we can write:

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$$

and conclude that the estimated slope of y with respect to x is $\widehat{\beta}_1 + 2\widehat{\beta}_2 x$. We can write it in the form of an approximation for the change in y with respect to x:

$$\frac{\Delta\widehat{y}}{\Delta x} \approx \widehat{\beta}_1 + 2\widehat{\beta}_2 x.$$

In order to provide richer interpretation of models with quadratics, there are several general facts about the quadratic function $\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$ that we need to keep in mind (let's consider the case $\widehat{\beta}_2 \neq 0$, so the function is certainly not a linear one). They can be summarised as follows:

- If $\widehat{\beta}_2 > 0$, the function is convex and has one minimum point.
  - Its graph is a parabola that opens upwards.

- If $\widehat{\beta}_2 < 0$, the function is concave and has one maximum point.
  - Its graph is a parabola that opens downwards.
- The location of the optimum point can be found from the first-order condition:

$$\widehat{\beta}_1 + 2\widehat{\beta}_2 x^* = 0 \quad \Rightarrow \quad x^* = -\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}$$

(sometimes x^* is called the turning point).

Figures 6.1–6.4 show the graphs of the quadratic function $\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2$ in four different cases determined by the signs of $\widehat{\beta}_1$ and $\widehat{\beta}_2$.

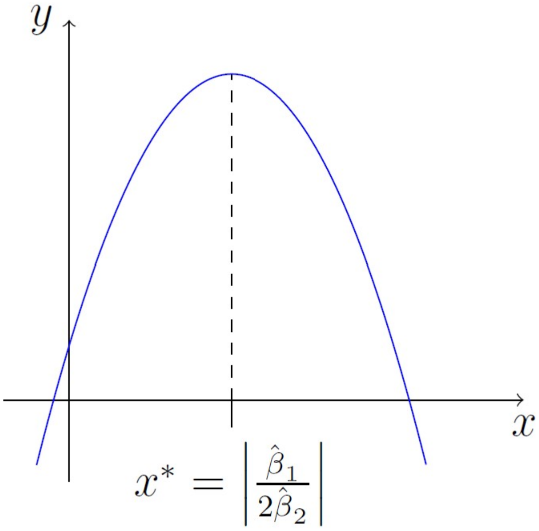


Figure 6.1: Case 1: $\widehat{\beta}_1 > 0$ and $\widehat{\beta}_2 < 0$.

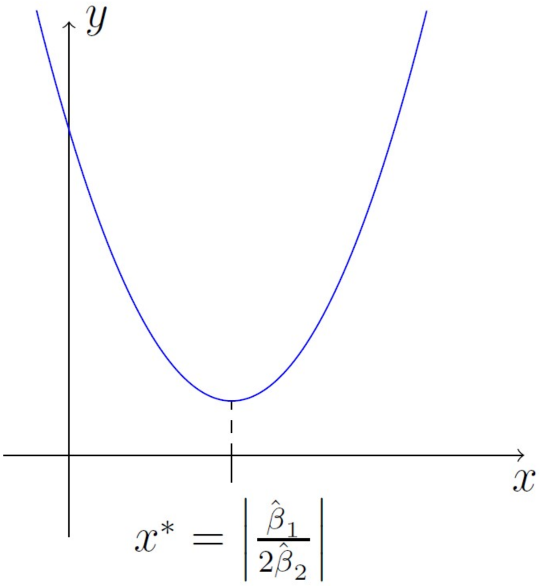


Figure 6.2: Case 2: $\widehat{\beta}_1 < 0$ and $\widehat{\beta}_2 > 0$.

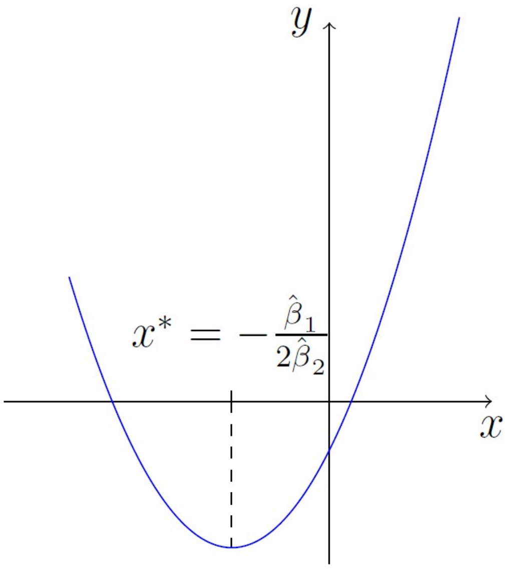


Figure 6.3: Case 3: $\widehat{\beta}_1 > 0$ and $\widehat{\beta}_2 > 0$.

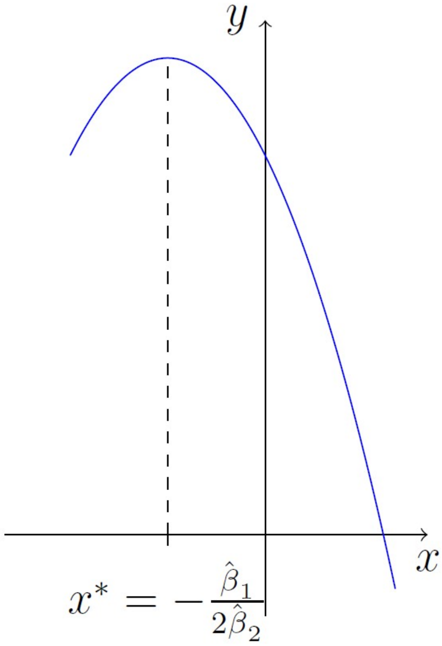


Figure 6.4: Case 4: $\widehat{\beta}_1 < 0$ and $\widehat{\beta}_2 < 0$.

For instance, in Case 1 the slope is initially positive, but it decreases as x increases. The graph of the quadratic function has a hump shape and, as we know, turns at the value:

$$x^* = \left|\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}\right|$$

with the function reaching its maximum at x^*. The slope is negative at the values of the regressor greater than x^*.

Similarly, in Case 2 the slope is initially negative, but it increases (gets less negative) as x increases. The graph of the quadratic function has a U shape and, as we know, turns at:

$$x^* = \left|\frac{\widehat{\beta}_1}{2\widehat{\beta}_2}\right|.$$

However, now the function reaches the minimum at x^* and for the values of the regressor greater than x^*, the slope is positive.

Cases 2 and 3 can be analysed analogously. In both these cases the turning point occurs at a negative value of x. In Case 3 the slope $(\widehat{\beta}_1 + 2\widehat{\beta}_2 x)$ increases with x and for $x > 0$ it is always positive. In Case 4 the slope $(\widehat{\beta}_1 + 2\widehat{\beta}_2 x)$ is always negative for $x > 0$. For instance, using WAGE1 data we can estimate the following wage equation with a quadratic in experience:

$$\widehat{\log(wage)} = \underset{(.1059)}{.1280} + \underset{(.0075)}{.0904}\, educ + \underset{(.0052)}{.0410}\, exper - \underset{(.000116)}{.000714}\, exper^2$$

$$n = 526,\ R^2 = .3003.$$

We can find that the quadratic turns at about $.041/[2(.000714)] \approx 28.7$ years of experience. In the sample, there are 121 people with at least 29 years of experience. This is a fairly sizable fraction of the sample. Thus, there is a change in the slope occurring in our model within the range of *exper* in our data.

In models with quadratics, we can conduct various significance tests. If we want to test:

$$H_0 : exper \text{ has constant effect on } \log(wage)$$
$$H_1 : exper \text{ effect on } \log(wage) \text{ is not constant}$$

this can be written as:

$$H_0 : \beta_{exper} = 0$$
$$H_1 : H_0 \text{ is not true.}$$

This can be conducted by a t test (under MLR.1 to MLR.6) using the t statistic on $exper^2$.

However, if we wanted to test:

$$H_0 : exper \text{ has no effect on } \log(wage)$$
$$H_1 : exper \text{ does have an effect on } \log(wage)$$

this would be written as:

$$H_0 : \beta_{exper} = 0,\ \beta_{exper^2} = 0$$
$$H_1 : H_0 \text{ is not true.}$$

This can be tested by an F test (under MLR.1 to MLR.6).

Activity 6.1 Take the MCQs related to this section on the VLE to test your understanding.

Activity 6.2 The following equation is estimated by OLS using WAGE1 data:

$$\widehat{\log(wage)} = \underset{(.1059)}{.1280} + \underset{(.0075)}{.0904}\, educ + \underset{(.0052)}{.0410}\, exper - \underset{(.000116)}{.000714}\, exper^2$$

$$n = 526,\ R^2 = .3003.$$

The estimated return to education is 9.04%. This model assumes this is the same for all years of education.

(i) Is $exper^2$ statistically significant at the 1% level?

(ii) Using the approximation:

$$\%\Delta\widehat{wage} = 100(\widehat{\beta}_2 + 2\widehat{\beta}_3 exper)\Delta exper$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

Activity 6.3 This activity is based on data in HPRICE3 but only for the year 1981. The data are for houses that sold during 1981 in North Andover, MA. 1981 was the year construction began on a local garbage incinerator and our objective is to study the effects of the incinerator location on housing price.

i. Consider the simple regression model:

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u$$

where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β_1 if the presence of the incinerator depresses housing prices?

ii. The estimated equation from part (i) is:

$$\widehat{\log(price)} = \underset{(0.65)}{8.05} + \underset{(.066)}{.365} \log(dist)$$

$$n = 142,\ R^2 = .180.$$

Interpret the results.

iii. To the simple regression model in part (ii), we add the variables $\log(intst)$, $\log(area)$, $\log(land)$, *rooms*, *baths*, and *age*, where *intst* is distance from the home to the interstate, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years.

The estimated equation is:

$$\widehat{\log(price)} = \underset{(.642)}{7.592} + \underset{(.058)}{.055} \log(dist) - \underset{(.052)}{.039} \log(intst) + \underset{(.077)}{.319} \log(area)$$

$$+ \underset{(.040)}{.077} \log(land) + \underset{(.028)}{.043}\, rooms + \underset{(.042)}{.167}\, baths - \underset{(.001)}{.004}\, age$$

$$n = 142,\ R^2 = .748.$$

Now, what do you conclude about the effects of the incinerator? Explain why (ii) and (iii) give conflicting results.

(iv) Adding $[\log(inst)]^2$ to the model from part (iii), we obtain:

$$\widehat{\log(price)} = \underset{(2.646)}{-3.317} + \underset{(.062)}{.185}\log(dist) + \underset{(.501)}{2.073}\log(intst) - \underset{(.028)}{.119}[\log(intst)]^2$$

$$+ \underset{(.073)}{.359}\log(area) + \underset{(.037)}{0.091}\log(land) + \underset{(.027)}{.038}\,rooms + \underset{(.040)}{.150}\,baths - \underset{(.001)}{.003}\,age$$

$$n = 142,\ R^2 = .778.$$

Now what happens? What do you conclude about the importance of functional form?

Functional form: Models with interaction terms

Read: Wooldridge, Section 6.2c and Appendix A.5.

This subchapter analyses regressions with interactions which model situations when a partial effect of one regressor depends on the value of another regressor. For example, the partial effect of education could depend on the level of intelligence.

For instance, consider the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u \tag{6.2.1}$$

which contains the **interaction term** x_1x_2. Holding x_2 (and u) fixed, the partial effect of x_1 on y is:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

so that effect of x_1 depends on x_2 unless $\beta_3 = 0$. Similarly, holding x_1 (and u) fixed, the partial effect of x_2 on y in the model with the interaction term is:

$$\frac{\Delta y}{\Delta x_2} = \beta_2 + \beta_3 x_1.$$

In the model above, the null hypothesis $H_0 : \beta_3 = 0$ is the same as saying the partial effects are constant. This hypothesis can be tested by using a t test. If it cannot be rejected at a small significance level (smaller than 5%), it is not worth complicating the model.

It is important to realise that it is easy to get confused in models with interactions. From:

$$\frac{\Delta y}{\Delta x_1} = \beta_1 + \beta_3 x_2$$

we see that β_1 is now the partial effect of x_1 on y when $x_2 = 0$. But $x_2 = 0$ may be very far from a legitimate, or at least interesting, part of the population. A similar comment holds for β_2.

An important point for models with interactions (as well as quadratics), is that we need to evaluate the partial effect at interesting values of the explanatory variables. Often, zero is not an interesting value for an explanatory variable and is well outside the range in the sample.

Often, two interesting parameters are the partial effects evaluated at the mean of the other variable:

$$\delta_1 = \beta_1 + \beta_3\mu_2$$
$$\delta_2 = \beta_2 + \beta_3\mu_1.$$

We can estimate these as:

$$\widehat{\delta}_1 = \widehat{\beta}_1 + \widehat{\beta}_3\bar{x}_2$$
$$\widehat{\delta}_2 = \widehat{\beta}_2 + \widehat{\beta}_3\bar{x}_1$$

where $\bar{x}_2$ and $\bar{x}_1$ are the sample averages and the $\widehat{\beta}_j$s are the OLS estimates. We can also obtain confidence intervals for δ_1 and δ_2. However, it is often easier to let a regression compute $\widehat{\delta}_1$ and $\widehat{\delta}_2$, and to get appropriate standard errors. To achieve that, we can **reparametrise** interaction effects and instead of the original model (6.2.1) consider:

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3(x_1 - \mu_1)(x_2 - \mu_2) + u$$

where $\alpha_0 = \beta_0 - \beta_3\mu_1\mu_2$ (it is a nice exercise to obtain this alternative representation of the regression model with the interaction term). In other words, the means are subtracted **before** creating the interactions. The regression we run is:

$$y_i \quad \text{on} \quad x_{i1},\ x_{i2},\ (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2), \quad i = 1, \ldots, n.$$

The intercept changes, too, but this is not important. The coefficient on the interaction does not change.

The advantages of such a reparametrisation can be summarised as follows:

- It gives an easy interpretation of all parameters.
- Standard errors for partial effects at the mean values are readily available in the regression output.
- If necessary, interaction may be centred at other interesting values.

To give an example, consider the following wage equation estimated with the interaction *educ* · *IQ* using WAGE2 data:

$$\widehat{\log(wage)} = \underset{(.568)}{5.838} + \underset{(.044)}{.025}\,educ + \underset{(.005)}{.004}\,IQ + \underset{(.0004)}{.0001}\,educ \cdot IQ$$
$$n = 935,\ R^2 = .1298.$$

Here the interaction term is not significant. Not only is it insignificant but it also renders *educ* and *IQ* insignificant. Indeed, here is the estimated model without the interaction term:

$$\widehat{\log(wage)} = \underset{(.096)}{5.658} + \underset{(.007)}{.039}\, educ + \underset{(.001)}{.006}\, IQ$$

$$n = 935,\ R^2 = .1297.$$

This has happened due to high correlation of the interaction term with its components, which is supported by the following correlation table:

	educ	*IQ*	*educ* · *IQ*
educ	1	0.515	0.888
IQ	0.515	1	0.845
educ · *IQ*	0.888	0.845	1

This collinearity can be reduced by defining the interaction in terms of demeaned variables:

	educ	*IQ*	$(educ - \overline{educ})(IQ - \overline{IQ})$
educ	1	0.515	0.186
IQ	0.515	1	-0.013
$(educ - \overline{educ})(IQ - \overline{IQ})$	0.186	-0.013	1

We can estimate the wage equation with the reparameterised interaction term – that is, with the term $(educ - \overline{educ})(IQ - \overline{IQ}) = (educ - 13.46845)(IQ - 101.2824)$:

$$\widehat{\log(wage)} = \underset{(.096)}{5.659} + \underset{(.007)}{.038}\, educ + \underset{(.001)}{.006}\, IQ + \underset{(.0004)}{.0001}(educ - \overline{educ})(IQ - \overline{IQ})$$

$$n = 935,\ R^2 = .1298.$$

We can see that once the reparameterisation is done, the estimates are much closer to the regression without the interaction. For example, for the average *IQ* (about 101), the return to education is about 3.8%. However, the interaction term is not at all statistically significant, which implies that there is no evidence that variability in IQ as such affects the return to education.

> **Activity 6.4** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 6.5** Exercise 6.4 from Wooldridge.

More on goodness-of-fit and selection of regressors

Read: Wooldridge, beginning of Sections 6.3 (before 6.3a), 6.3a, 6.3c.

The beginning of Section 6.3 (before 6.3a) gives several general recommendations about using R^2.

Section 6.3a introduces the adjusted R-squared as another goodness of fit measure. To motivate this new measure, we need to note how the R-squared changes when more explanatory variables are added. It can be shown that R^2 **can never go down**, and usually increases, when one or more variables are added to a regression. This is because additional variables almost always help to explain more variation in the dependent variable.

To give this a more mathematical justification, recall that $R^2 = 1 - \frac{SSR}{SST}$. Imagine a model in which we start with one regressor x_1 and then add another regressor x_2. Then SST does not change as it is calculated from the observations of the dependent variable, which remains unchanged. If OLS happens to choose the coefficient on x_2 to be exactly zero, then the SSR is the same whether or not x_2 is included in the regression. However, if OLS chooses any value other than zero, then it must be that this value reduced the SSR relative to the regression that excludes x_2. In practice, an estimated coefficient will most likely be not zero, so in general the SSR will decrease when a new regressor is added. This means that the $R^2 = 1 - \frac{SSR}{SST}$ generally increases (can never decrease) when a new regressor is added.

The **adjusted *R*-squared**, also called 'R-bar-squared', is defined as:

$$\bar{R}^2 = 1 - \frac{[SSR/(n-k-1)]}{[SST/(n-1)]} = 1 - (1-R^2)\frac{(n-1)}{(n-k-1)}$$

where $\widehat{\sigma}^2$ is the usual estimator of the variance parameter.

$\bar{R}^2$ can increase or decrease when more regressors are added. Indeed, as more regressors are added, SSR falls, but so does $n-k-1$. $\bar{R}^2$ penalises adding regressors that have weak explanatory power for the dependent variable (in this case we can expect $\bar{R}^2$ to decrease). For $k \geq 1$, $\bar{R}^2 < R^2$ unless $SSR = 0$ (not an interesting case). In addition, it is possible that $\bar{R}^2 < 0$, especially if $n-k-1$ is small (recall that $R^2 \geq 0$ always).

Section 6.3c discusses the dangers of overemphasising goodness of fit. Fixating on making R^2 or $\bar{R}^2$ as large as possible can lead to silly mistakes and can lead to a model that has no interesting *ceteris paribus* interpretation. It is very important to always remember the *ceteris paribus* interpretation of a regression. Sometimes it does not make sense to hold other factors fixed when studying the effect of a particular variable.

To give an illustration of the importance of keeping in mind the *ceteris paribus* interpretation of the regression, consider the effects of spending on mathematics pass rates (using MEAP93 data). The following linear regression model of *math10* (percentage of students passing the MEAP mathematics test) on *lexpend* (logarithm of expenditure per student, in \$) and *lnchprg* (percentage of students eligible for free school lunch programme – this is a measure of poverty) can help us study the effects of total spending per student in a school on the percentage of students passing a standardised mathematics test:

$$\widehat{math10} = \underset{(25.0729)}{-20.3608} + \underset{(2.9726)}{6.2297}\, lexpend - \underset{(.0354)}{.3046}\, lnchprg$$

$$n = 408,\ R^2 = .1799.$$

Since spending is in logs:

$$\Delta\widehat{math10} = (6.23/100)\%\Delta expend \approx .0623(\%\Delta expend).$$

So, if spending increases by 10%, holding *lnchprg* fixed, $\widehat{math}10$ increases by about 0.62 percentage points. Now suppose we add *staff* (staff per 1,000 students), and the log of the average teacher salary, *lsalary* (in $):

$$\widehat{math10} = \underset{(44.0357)}{-98.8195} - \underset{(11.6959)}{20.0524}\, lexpend - \underset{(.0384)}{.2802}\, lnchprg + \underset{(.1140)}{.2510}\, staff + \underset{(11.2251)}{26.3234}\, lsalary$$

$$n = 408,\ R^2 = .1910.$$

As we can see, the coefficient on *lexpend* is actually negative (but not statistically different from zero). Does it mean that spending is not important? No, we cannot draw this conclusion because a part of spending goes towards lowering student–teacher ratios and increasing salaries. Once we hold those fixed, the role for spending is obviously limited. What we seem to be finding in the second regression is that spending other than to lower *staff* or increase *lsalary* has no effect on performance. But this does **not** mean that total spending has no effect!

To give an illustration of a model that has no interesting *ceteris paribus* interpretation, consider the study of the effects of spending on mathematics pass rates using MEAP01 data. We start with the model:

$$\widehat{math4} = \underset{(13.1284)}{138.2842} - \underset{(.9095)}{3.3771}\, lexpend - \underset{(.0142)}{.4483}\, lunch$$
$$n = 1823,\ \bar{R}^2 = .3603$$

where $math44$ is the percentage of students getting a passing fourth-grade mathematics score, *lexpend* is the logarithm of expenditure per student (in $), *lunch* is the percentage of students eligible for free or reduced lunch. We then add $read4$ (percentage of students getting a passing fourth-grade reading score):

$$\widehat{math4} = \underset{(8.9047)}{40.8296} - \underset{(.6032)}{.8786} lexpend - \underset{(.0118)}{.0986}\, lunch + \underset{(.0164)}{.7932}\, read4$$

$$n = 1{,}823,\ \bar{R}^2 = .7207.$$

$\bar{R}^2$ goes from .3603 to .7207, which is a huge increase and, thus, we may be tempted to use the second regression. But why should we hold $read4$ fixed when looking at the effects of spending on $math4$? We want to allow spending to improve reading pass rates, too. Thus, $read4$ can be considered to be a *bad control* – it is itself an outcome of *lexpend* and, thus, $read4$ can be a dependent variable in its own regression.

The last regression example with $read4$ as a regressor is what can be called **over-controlling** for factors in multiple regression. It may be tempting to **over-control** because often the goodness-of-fit statistics increase substantially or sometimes this happens because researchers are afraid of omitting important explanatory variables. It is important to avoid including any bad controls by remembering the *ceteris paribus* nature of multiple regression.

Activity 6.6 Exercise 6.8 from Wooldridge.

6.2.2 Describing qualitative information

Read: Wooldridge, Section 7.1.

The main takeaway here is that qualitative factors often come in the form of binary information: a worker belongs to a union or does not, a firm offers a 401(k) pension plan or it does not. Other examples of qualitative information include race, industry, region, rating grade. In these examples the relevant information can be captured by defining a **binary variable** (or **dummy variable**). Sometimes it is also called a **zero-one** variable to emphasise the two values it takes on.

In defining a dummy variable, it is very important to decide which outcome is coded as a zero and which is coded as a one. It is also recommended to choose the variable name to be descriptive.

It is also useful to know that dummy variables can be used to create more refined categorisations. For example, if we start with two pieces of qualitative information – say, gender and marital status, as captured by $female$ and *married* dummy variables – we can define more than two categories. These would be married male (*marrmale*), married female (*marrfem*), single male (*singmale*), and single female (*singfem*).

6.2.3 A single dummy independent variable

Read: Wooldridge, Section 7.2.

In models with a single dummy independent variable it is important to know the interpretation of the regression coefficient corresponding to the dummy variable.

Start by considering a model where the dummy variable is the only regressor. For instance:

$$\mathit{wage} = \beta_0 + \delta_0 \mathit{union} + u \tag{6.2.2}$$

where *union* is a binary variable that takes the value of 1 if a worker belongs to a union and takes the value of 0 otherwise. Then β_0 is the expected wage of non-union workers and $\beta_0 + \delta_0$ is the expected wage of union workers, and δ_0 is the gap between these two expected wages. When a model has other regressors on the right-hand side – for example:

$$\mathit{wage} = \beta_0 + \delta_0 \mathit{union} + \beta_1 \mathit{tenure} + u \tag{6.2.3}$$

then even though β_0 and $\beta_0 + \delta_0$ no longer have such interpretations, δ_0 can still be interpreted as the difference between expected wages for union and non-union workers *conditional on the other regressors*. In example (6.2.3) above:

$$\delta_0 = E(\mathit{wage} \mid \mathit{union} = 1, \mathit{tenure}) - E(\mathit{wage} \mid \mathit{union} = 0, \mathit{tenure}).$$

Graphically, δ_0 in (6.2.3) is the gap between two regression lines – one for union workers and the other for non-union ones. Usually δ_0 is referred to as the difference in the intercept between the the **base group** and the other group.

In the example above the base group is non-union workers and this is why β_0 is the intercept for non-union workers, and δ_0 is the difference in intercepts between union and non-union workers.

It is important to understand how the interpretation of coefficients changes if we decide to choose a different base group. We could choose union workers as the base group by writing the model as:

$$wage = \alpha_0 + \gamma_0 \textit{non-union} + u$$

where the intercept for union workers is α_0 and the intercept for non-union workers is $\alpha_0 + \gamma_0$. Because $\textit{non-union} = 1 - \textit{union}$, this implies that $\alpha_0 = \beta_0 + \delta_0$ and $\alpha_0 + \gamma_0 = \beta_0$ (which also gives $\gamma_0 = -\delta_0$). Thus, we essentially get the same answer if we set union workers as the base group. The coefficient on the dummy variable changes sign but it must remain the same magnitude. The intercept changes because now the base group is union workers. It is important to keep track of which group is the base group.

In this topic you get your first exposure to the **dummy variable trap** problem. In our example with wage and union/non-union variables, putting both *union* and *non-union* in the model and also including the constant term would result in the dummy variable trap. The dummy variable trap is a situation of perfect collinearity (and, thus, a violation of assumption MLR.3). Indeed, in this case because of the presence of the constant term in the regression equation, we would have the following linear relationship between regressors 1, *union* and *non-union*:

$$\textit{union} + \textit{non-union} = 1.$$

Later subchapters consider situations when there are more than two dummy variables representing multiple categories. If we have, say, c categories represented by dummy variables, then to avoid the dummy variable trap we can only include $(c - 1)$ dummy variables in the regression alongside the intercept.

Finally, with $\log(y)$ as the dependent variable, δ_0 is given a percentage change interpretation.

Activity 6.7 Take the MCQs related to this section on the VLE to test your understanding.

Activity 6.8 Suppose we have a dataset where everyone identifies as either male or female. The model we consider is:

$$wage = \beta_0 + \delta_0 \textit{female} + u$$

where *female* takes the value of 1 for women and takes the value of 0 otherwise. Suppose we have the random sample $\{(wage_i, female_i)\}_{i=1}^n$ of size n.

Derive the following formulas for the OLS estimators in this model:

$$\widehat{\delta}_0 = \overline{wage}_f - \overline{wage}_m$$

$$\widehat{\beta}_0 = \overline{wage}_m.$$

Activity 6.9 The Bechdel test is a test (more of a minimum threshold) for whether a film portrays women in a 'normal/real life' way. A film passes the Bechdel test if there are at least two named female characters in the film who have a conversation together that doesn't pertain to a male character. (Whether this constitutes a good 'test' or not is definitely up for debate.)

(i) Consider a regression of the logarithm of international grossing (adjusted for inflation in 2013 dollars) on the dummy variable *pass_test* whether the film passes the Bechdel test:

$$\log(\mathit{intgross\_13}) = \beta_0 + \delta_0 \mathit{pass\_test} + \varepsilon.$$

Give the interpretation of the coefficient δ_0.

(ii) The estimation of the equation from part (i) gives the following results:

$$\log(\widehat{\mathit{intgross}}\mathit{\_13}) = \underset{(.0559)}{4.4844} - \underset{(.0836)}{.3503}\,\mathit{pass\_test}$$

$$n = 1776,\ R^2 = .0098.$$

Discuss what the estimation results indicate about audiences' preferences for films.

(iii) Let us also include the logarithm of the budget of the film (adjusted for inflation in 2013 dollars), in particular taking this to be the $\log(\mathit{budget\_13})$:

$$\log(\widehat{\mathit{intgross}}\mathit{\_13}) = \underset{(.0907)}{1.4187} - \underset{(.0625)}{.0662}\,\mathit{pass\_test} + \underset{(.0226)}{0.8603}\log(\mathit{budget\_13})$$

$$n = 1{,}776,\ R^2 = .4545.$$

How do you interpret the coefficient on the budget? Discuss what happens to the coefficient on *pass_test* in comparison to part (ii). In particular, what does this mean for sexism in the film industry?

(iv) In parts (ii) and (iii) we estimated two equations:

$$\log(\widetilde{\mathit{intgross}}\mathit{\_13}) = \widetilde{\beta}_0 + \widetilde{\delta}_0\,\mathit{pass\_test}$$

and

$$\log(\widehat{\mathit{intgross}}\mathit{\_13}) = \widehat{\beta}_0 + \widehat{\delta}_0\,\mathit{pass\_test} + \widehat{\beta}_1\log(\mathit{budget\_13}).$$

Describe the algebraic relationship between $\widetilde{\delta}_0$ and $\widehat{\delta}_0$. Use this algebraic relationship and the fact that $\log(\mathit{budget\_13})$ and *pass_test* are negatively correlated to explain differences in findings in parts (ii) and (iii).

6.2.4 Using dummy variables for multiple categories

Read: Wooldridge, Section 7.3.

Here you will learn how to allow for more than two groups. This can be accomplished by defining a set of dummy variables.

Suppose in a dataset we have observations of wage and everyone identifies as either male or female. In addition, there is a marital status variable which takes only two values – married or single. We can define four exhaustive and mutually exclusive groups. These are married males (*marrmale*), married females (*marrfem*), single males (*singmale*), and single females (*singfem*).

We can allow each of the four groups to have a different intercept by choosing a base group and then including dummies for the other three groups. If, for instance, we choose single males as the base group, we include *marrmale*, *marrfem*, and *singfem* in the regression. The coefficients on these variables are relative to single men. With log(*wage*) as the dependent variable, intercepts will have a percentage change interpretation.

Consider, for instance:

$$\log(\mathit{wage}) = \beta_0 + \delta_1 \mathit{marrmale} + \delta_2 \mathit{marrfem} + \delta_3 \mathit{singfem} + \beta_1 \mathit{educ} + u.$$

Then δ_1 is the 'marriage premium' for men (holding education fixed). We can easily test $H_0 : \delta_1 = 0$ vs. $H_1 : \delta_1 \neq 0$ (or $H_1 : \delta_1 > 0$) directly from the regression output in a statistical package by using the t statistic associated with *marrmale*. The 'marriage premium' for women is $\delta_2 + \beta_0 - (\delta_3 + \beta_0) = \delta_2 - \delta_3$. We cannot test $H_0 : \delta_2 = \delta_3$ directly from the regression output. We can usually run a special command in a statistical package to test it. An alternative approach would be to choose, say, married females as the base group and re-estimate the model (including the dummies *marrmale*, *singmale*, and *singfem*) and use the t statistic for *singfem*.

In this example with four groups, no matter which group we choose as the base group, we include only three of four dummy variables. If we include all four we fall into the **dummy variable trap**.

You also need to know how dummy variables can be used in capturing information given in the form of categories with a clear ordering of the categories.

Let us describe this in an example. Suppose that we would like to estimate the effect of a person's physical attractiveness on performance in the labour market. The BEAUTY dataset includes rankings of physical attractiveness of each man or woman, on a scale of 1 to 5 with 5 being 'strikingly beautiful or handsome'. The 'looks' variable is an **ordinal variable** (other examples of ordinal variables include credit ratings, or ratings of individual happiness): we know that order of outcomes conveys information (5 is better than 4, and 2 is better than 1) but we do not know that the difference between 5 and 4 is same as 2 and 1. Including the ordinal 'looks' variable as we would include any other explanatory variable would imply that as we move up the scale from 1 to 5, a one-unit increase means the same amount of 'beauty'. It might not make sense to assume that a one-unit increase in 'looks' has a constant effect on the dependent variable.

An alternative approach is to define dummy variables. This approach is usually preferable to including the ordinal variable itself as a regressor. For example, we can create three dummy variables corresponding to the following categories:

- below average (*belavg*, where $\mathit{looks} \leq 2$)
- average ($\mathit{looks} = 3$)

- above average (*abvavg*, where *looks* ≥ 4).

This is an illustration of how we can create multiple dummy variables to capture the ordinal information.

> **Activity 6.10** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 6.11** Exercise 7.4 from Wooldridge.

6.2.5 Interactions involving dummy variables

Read: Wooldridge, Section 7.4.

Dummy variables can be interacted with each other and can be interacted with continuous variables. It is important to understand that these two cases lead to different regression specifications with different interpretations.

When dummy variables are interacted with each other (Wooldridge, Section 7.4a), this creates a more refined categorisation of the population. For example, suppose we have dummy variables *female* and *union*. Consider a regression model that contains both these dummy variables as well as an interaction term between *female* and *union*. This allows the union memebership premium to depend on gender. For concreteness, take:

$$\log(wage) = \beta_0 + \beta_1 female + \beta_2 union + \beta_3 female \cdot union + \beta_4 educ + u.$$

Then β_0 is the intercept for males who are not union members (plug in $female = 0$ and $union = 0$), $\beta_0 + \beta_1$ is the intercept for females who are not union members (plug in $female = 1$ and $union = 0$), $\beta_0 + \beta_2$ is the intercept for males who are union members (plug in $female = 0$ and $union = 1$), and, finally, $\beta_0 + \beta_1 + \beta_2$ is the intercept for females who are union members (plug in $female = 1$ and $union = 1$). Equivalently, this model can be recast by defining dummy variables for three out of four groups and using specifications discussed in Wooldridge, Section 7.3.

Situations of interacting dummy variables with continuous explanatory variables (Wooldridge, Section 7.4b) will lead to regression models with **different slopes**, as well as different intercepts. As an example, consider the model:

$$\log(wage) = \beta_0 + \delta_0 union + \beta_1 exper \ + \delta_1 union \cdot exper + u \qquad (6.2.4)$$

where we added the interaction term $union \cdot exper$ to the model where *union* and *exper* appear separately. Then the intercept for non-union members is β_0 and that for union members is $\beta_0 + \delta_0$. The slope for non-union members is β_1 and that for union members is $\beta_1 + \delta_1$.

The summary of parameter interpretations for non-union and union members is in the following table:

	Intercept	Slope
non-union	β_0	β_1
union	$\beta_0 + \delta_0$	$\beta_1 + \delta_1$
Difference (*union* – *non-union*)	δ_0	δ_1

It is important to be careful when interpreting the coefficient on *union* when $union \cdot exper$ is in the equation. To see why, at any level of experience, the predicted difference in log(*wage*) between union and non-union members is:

$$\delta_0 + \delta_1 exper.$$

Therefore, the coefficient on *union*, δ_0, is the predicted difference in log(*wage*) between a union and non-union worker when $exper = 0$. This is not an especially interesting subset of the population.

It is also worth mentioning that if we are including an interaction effect, we need to include the levels of the two variables. Indeed, it would be hard to justify omitting the level of a variable but including an interaction involving that variable. For example, suppose we drop *exper*, but include $union \cdot exper$. Such a model imposes zero return to experience for non-union members (hard to justify) and the coefficient on $union \cdot exper$ is now a direct estimate of the return to experience for union members, rather than being the difference in returns between union and non-union members.

In the context of model (6.2.4), we necessarily get the same estimated intercepts and slopes if we estimate regressions **separately** for union and non-union employees. The null hypothesis that there is no difference in log(*wage*) between union and non-union employees at the same levels of experience is:

$$H_0 : \delta_0 = 0, \ \delta_1 = 0.$$

This H_0 can be tested using the F test. This is a version of the **Chow test** (Wooldridge, Section 7.4c) for the equality of regression functions across two groups. We test joint significance of the dummy variable defining the groups as well as the interaction terms.

In a general model with k explanatory variables and an intercept, suppose we have two groups. If we would like to test whether the intercept and all slopes are the same across the two groups, we can define a dummy variable, say w, indicating one of the two groups. Then in the model:

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\ + \delta_0 w + \delta_1 w \cdot x_1 + \delta_2 w \cdot x_2 + \cdots + \delta_k w \cdot x_k + u \end{aligned}$$

we want to test:

$$H_0 : \delta_0 = 0, \ \delta_1 = 0, \ \delta_2 = 0, \ \ldots, \ \delta_k = 0$$

for $k+1$ restrictions. It can be done using a standard F test of the $k+1$ exclusion restrictions (that is, $F_{k+1,\, n-2\,(k+1)}$ distribution):

$$F = \frac{(SSR_r - SSR_{ur})/(k+1)}{SSR_{ur}/(n - 2(k+1))}.$$

It is important to know another variation of the Chow test. Namely, it is often of interest to allow $\delta_0 \neq 0$ and just test equality of the slopes:

$$H_0 : \delta_1 = 0,\ \delta_2 = 0,\ \ldots,\ \delta_k = 0.$$

Now we are testing k parameters under H_0. We, once again, use a standard F test (but this time with the $F_{k,\,n-2(k+1)}$ distribution):

$$F = \frac{(SSR_r - SSR_{ur})/k}{SSR_{ur}/(n-2(k+1))}$$

(the restricted model, of course, is now different from the previous case).

You also need to know the following equivalent way of conducting the Chow test (it produces an identical statistic). This approach relies on using the sum of squared residuals version of the F statistic. This approach does not require us to construct all of the interactions. Instead, we can compute three SSRs:

1. Pool the data and estimate a single regression. This is the restricted model, and produces the **restricted SSR**. Call this the **pooled** SSR, SSR_P.

   If we want to test to the equality of the intercepts and all the slopes, then SSR_P comes from the estimation of the following model:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

   This step depends on the null hypothesis H_0 we are testing as the restricted model we estimate in this step depends on our H_0.

2. Estimate the regressions on the two groups (say, 1 and 2) separately in the usual way:

   $$y = \beta_{10} + \beta_{11} x_1 + \beta_{12} x_2 + \cdots + \beta_{1k} x_k + u_1$$

   for group 1 and:

   $$y = \beta_{20} + \beta_{21} x_1 + \beta_{22} x_2 + \cdots + \beta_{2k} x_k + u_2$$

   for group 2. Get the SSRs, SSR_1 and SSR_2. The **unrestricted SSR** is $SSR_1 + SSR_2$ (and this is the same as the regression that includes the full set of interactions).

   This step does not depend on the null hypothesis H_0 we are testing as its outcome is the unrestricted SSR.

Let n denote the sample size in the pooled regression where we use both groups. The F statistic for the null hypothesis of the equality of the intercepts and all the slopes is

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]/(k+1)}{(SSR_1 + SSR_2)/[n - 2(k+1)]}$$

and has the $F_{k+1,\,n-2(k+1)}$ distribution under H_0.

If we leave the intercepts unrestricted under H_0, then SSR_P is obtained from the pooled regression but with the dummy variable added (SSR_1 and SSR_2 obtained from regressions for each group remain the same as those regressions do not change):

$$y = \beta_0 + \delta_0 w + beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u.$$

The $k+1$ in the numerator becomes k, and we use the $F_{k,\,n-2(k+1)}$ distribution.

Activity 6.12 Take the MCQs related to this section on the VLE to test your understanding.

Activity 6.13 Exercise 7.8 from Wooldridge.

Activity 6.14 This exercise is based on SLEEP75 data. The equation of interest is:

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u$$

where:

- *sleep* is minutes of sleep at night, per week
- *totwrk* is minutes worked per week
- *yngkid* is the dummy variable (= 1 if children under the age of 3 are present).

(i) Estimation of the regression equation separately for men and women gives the following results:

- for men:

$$\widehat{sleep)} = \underset{((310.0)}{3{,}648.2} - \underset{(.024)}{.182}\, totwrk - \underset{(7.41)}{13.05}\, educ + \underset{(14.32)}{7.16}\, age - \underset{(.1684)}{.0448}\, age^2 + \underset{(9.02)}{60.38}\, yngkid$$

$$n = 400,\ R^2 = .156$$

- for women:

$$\widehat{sleep)} = \underset{((384.9)}{4{,}238.7} - \underset{(.028)}{.140}\, totwrk - \underset{(9.59)}{10.21}\, educ - \underset{(18.53)}{30.36}\, age - \underset{(.223)}{.368}\, age^2 - \underset{(93.19)}{118.28}\, yngkid$$

$$n = 306,\ R^2 = .098.$$

Are there notable differences in the two estimated equations?

(ii) We want to compute the Chow test for equality of the parameters in the sleep equation for men and women. In the model:

$$\begin{aligned} sleep = {} & \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ & + \delta_0 male + \delta_1 totwrk \cdot male + \delta_2 educ \cdot male + \delta_3 age \cdot male + \delta_4 age^2 \cdot male \\ & + \delta_5\, yngkid \cdot male + u \end{aligned}$$

that uses the full set of observations we test:

$$H_0 : \delta_0 = \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$$

against the alternative:

$$H_1 :\ \text{at least one of } \delta_j, j = 0, \ldots, 5,\ \text{is not } 0.$$

The F statistic for this test is about 2.12.

What are the relevant *df* for the test? Should you reject the null at the 5% level?

(iii) Now we allow for different intercepts for males and females and determine whether the interaction terms involving *male* are jointly significant. In other words, in the model:

$$\begin{aligned} sleep &= \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid \\ &\quad + \delta_0 male + \delta_1 totwrk \cdot male + \delta_2 educ \cdot male + \delta_3 age \cdot male + \delta_4 age^2 \cdot male \\ &\quad + \delta_5 yngkid \cdot male + u \end{aligned}$$

that uses the full set of observations we test:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$$

against the alternative:

$$H_1 : \text{ at least one of } \delta_j, j = 1, \ldots, 5, \text{ is not } 0.$$

The F statistic for this test is about 1.26

What are the relevant *df* for the test? Should you reject the null at the 5% level?

(iv) Given the results from parts (ii) and (iii), what would be your final model?

6.3 Answers to activities

Solution to Activity 6.2

(i) Our null and alternative hypotheses are:

$$H_0 : \beta_{exper^2} = 0$$
$$H_1 : \beta_{exper^2} \neq 0.$$

(Essentially we are testing the null that the equation is linear in *exper* against the alternative that it is quadratic.) To answer the question affirmatively, we want to reject H_0 at the 1% level.

The t statistic on $exper^2$ is:

$$t_{\widehat{\beta}_{exper^2}} = \frac{\widehat{\beta}_{exper^2}}{se(\widehat{\beta}_{exper^2})} = \frac{-.000714}{.000116} \approx -6.16.$$

Given our CLM assumptions (MLR.1 to MLR.6), we know:

$$\text{under } H_0 : t_{\widehat{\beta}_{exper^2}} \sim t_{df}, \text{ where } df = 526 - 4 = 522.$$

Using the 1% level of significance, our critical value equals $t_{df,.005} = 2.576$ (taking into account that the alternative is two-sided). The absolute value of our test statistic exceeds $t_{df,.01}$, so we reject the null hypothesis and conclude that $exper^2$ is statistically significant at the 1% level.

(ii) To estimate the return to the fifth year of experience, we start at $exper = 4$ and increase $exper$ by one, so $\Delta exper = 1$:

$$\%\Delta\widehat{wage} = \%100(.04102 - 2(.000714)4) \approx 3.53\%.$$

Similarly, for the 20th year of experience:

$$\%\Delta\widehat{wage} = \%100(.04102 - 2(.000714)19) \approx 1.39\%.$$

Solution to Activity 6.3

(i) We expect $\beta_1 \geq 0$: all other relevant factors equal, it is better to have a home farther away from the incinerator.

(ii) A 1% increase in distance from the incinerator is associated with a predicted price that is about .37% higher.

(iii) The coefficient on $\log(dist)$ becomes about .055 ($se \approx .058$). The effect is much smaller now, and statistically insignificant. This is because we have explicitly controlled for several other factors that determine the quality of a home (such as its size and number of baths) and its location (distance to the interstate). This is consistent with the hypothesis that the incinerator was located near less desirable homes to begin with.

(iv) The coefficient on $\log(dist)$ is now very statistically significant, with a t statistic of about three. The coefficients on $\log(intst)$ and $[\log(intst)]^2$ are both very statistically significant, each with t statistics above four in absolute value. Just adding $[\log(intst)]^2$ has had a very big effect on the coefficient important for policy purposes. This means that distance from the incinerator and distance from the interstate are correlated in some nonlinear way that also affects housing price.

We can find the value of $\log(intst)$ where the effect on $\log(price)$ actually becomes negative: $2.073/[2(.1193)] = 8.69$. When we exponentiate this we obtain about 5,943 feet from the interstate. Therefore, it is best to have your home away from the interstate for distances less than just over a mile.

Solution to Activity 6.5

(i) Holding all other factors fixed we have:

$$\Delta\log(wage) = \beta_1\Delta educ + \beta_2\Delta educ \cdot educ \cdot pareduc.$$

Dividing both sides by $\Delta educ$ gives the result. The sign of β_2 is not obvious, although $\beta_2 > 0$ if we think a child gets more out of another year of education the more highly educated are the child's parents.

(ii) We use the values $pareduc = 32$ and $pareduc = 24$ to interpret the coefficient on $educ \cdot pareduc$. The difference in the estimated return to education is $.00078(32 - 24) = .0062$, or about .62 percentage points.

(iii) When we add *pareduc* by itself, the coefficient on the interaction term is negative. The t statistic on $educ \cdot pareduc$ is about -1.33, which is not significant at the 10% level against a two-sided alternative ($t_{716,.05} = 1.645$). Note that the coefficient on *pareduc* is significant at the 5% level against a two-sided alternative ($t_{716,.025} = 1.96$). This provides a good example of how omitting a level effect (*pareduc* in this case) can lead to biased estimation of the interaction effect. In the hypotheses testing we assume MLR.1–MLR.6.

Solution to Activity 6.6

(i) The answer is not entirely obvious, but one must properly interpret the coefficient on *alcohol* in either case. If we include *attend*, then we are measuring the effect of alcohol consumption on college GPA, holding attendance fixed. Because attendance is likely to be an important mechanism through which drinking affects performance, we probably do not want to hold it fixed in the analysis. In other words, *attend* can be considered a bad control as it's a result of *alcohol*.

If we do include *attend*, then we interpret the estimate of $\beta_{alcohol}$ as measuring those effects on *colGPA* that are not due to attending class. (For example, we could be measuring the effects that drinking alcohol has on study time.) To get a total effect of alcohol consumption, we would leave *attend* out.

(ii) We would want to include *SAT* and *hsGPA* as controls, as these measure student abilities and motivation. Drinking behaviour in college could be correlated with one's performance in high school and on standardised tests. Other factors, such as family background, would also be good controls.

Solution to Activity 6.8

Using the formula for the OLS slope estimate:

$$\begin{aligned}\widehat{\delta}_0 &= \frac{\sum_{i=1}^n (female_i - \overline{female}) \cdot (wage_i - \overline{wage})}{\sum_{i=1}^n (female_i - \overline{female})^2} \\ &= \frac{\sum_{i=1}^n female_i \cdot wage_i - n\overline{female} \cdot \overline{wage}}{\sum_{i=1}^n female_i^2 - n(\overline{female})^2}.\end{aligned}$$

Let n_f denote the number of women in the sample. That is:

$$n_f = \sum_{i=1}^n female_i = \sum_{i=1}^n female_i^2$$

(using that $female_i^2 = female_i$ due to the binary nature of this variable).

Then:

$$\begin{aligned}\overline{female} &= \frac{n_f}{n} \\ \overline{wage}_f &= \frac{1}{n_f}\sum_{i=1}^{n} female_i \cdot wage_i \\ \overline{wage}_m &= \frac{1}{n-n_f}\sum_{i=1}^{n}(1-female_i)\cdot wage_i\end{aligned}$$

where $\overline{wage}_f$ is the average wage of women in the sample and $\overline{wage}_m$ is the average wage of men in the sample. Also note that:

$$\overline{wage} = \frac{n_f}{n}\overline{wage}_f + \left(1-\frac{n_f}{n}\right)\overline{wage}_m.$$

From the formula for $\widehat{\delta}_0$ above we then obtain

$$\begin{aligned}\widehat{\delta}_0 = \frac{n_f\overline{wage}_f - n\cdot\frac{n_f}{n}\overline{wage}}{n_f - n\cdot(n_f/n)^2} &= \frac{\overline{wage}_f - \overline{wage}}{1-\frac{n_f}{n}} \\ &= \frac{\left(1-\frac{n_f}{n}\right)\overline{wage}_f - \left(1-\frac{n_f}{n}\right)\overline{wage}_m}{1-\frac{n_f}{n}} \\ &= \overline{wage}_f - \overline{wage}_m.\end{aligned}$$

Using the formula for the OLS intercept estimate:

$$\begin{aligned}\widehat{\beta}_0 &= \overline{wage} - \widehat{\delta}_0\cdot\overline{female} \\ &= \frac{n_f}{n}\overline{wage}_f + \left(1-\frac{n_f}{n}\right)\overline{wage}_m - \left(\overline{wage}_f - \overline{wage}_m\right)\cdot\frac{n_f}{n} \\ &= \overline{wage}_m.\end{aligned}$$

Solution to Activity 6.9

(i) $100\cdot\delta_0$ is the approximate expected percentage difference in *intgross_13* between those films that pass the Bechdel test and those that do not.

(ii) The estimation results indicate that films which pass the Bechdel test have an international gross that is approximately 35% lower than films which don't pass it! This seems to suggest that audiences only want to watch films where women have no meaningful roles.

(iii) The coefficient on budget indicates that increasing your budget by 1% increases your international gross by 0.86%. So there are slightly decreasing returns to investment into the film.

The coefficient on passing the test is now insignificant. So, actually audiences don't seem to care about how women are portrayed in films.

Sexism seems to enter the film industry through the budget of the film. Producers appear to put more money into films which do not pass the test. These films are then of a better quality (because of a greater budget) and audiences prefer these films. So it is not the absence of women which causes audiences to enjoy a film; it is the amount of money put into the film.

(iv) By the algebraic relationship:

$$\widetilde{\delta}_0 = \widehat{\delta}_0 + \widehat{\beta}_1 \widetilde{\gamma}_1$$

where $\widetilde{\gamma}_1$ is the slope from simple regression of $\log(\mathit{budget\_13})$ on *pass_test*. Since $\log(\mathit{budget\_13})$ and *pass_test* are negatively correlated, $\widetilde{\gamma}_1$ is negative. It is reasonable to expect that $\widehat{\beta}_1$ would be positive.

Since $\widehat{\beta}_1 > 0$, we have:

$$\widetilde{\delta}_0 = \widehat{\delta}_0 + (+)(-) \leq \widehat{\delta}_0$$

i.e. if we omit $\log(\mathit{budget\_13})$, the slope estimate of *pass_test* tends to be smaller (more negative).

Solution to Activity 6.11

(i) The approximate difference is just the coefficient on utility times 100, or -28.3%. The t statistic is $t = -.283/.099 \approx -2.86$, which is statistically significant at the 1% level since $|t| > t_{203,\,0.005} = 2.576$. (Implicitly we use the two-sided alternative and, of course, assume MLR.1 to MLR.6.)

(ii) $100 \cdot [\exp(-.283) - 1] \approx -24.7\%$, and so the estimate is somewhat smaller in magnitude.

(iii) The proportionate difference is $.181 - .158 = .023$, or about 2.3%. One equation that can be estimated to obtain the standard error of this difference is:

$$\log(\mathit{salary}) = \beta_0 + \beta_1 \log(\mathit{sales}) + \beta_2 \mathit{roe} + \delta_1 \mathit{consprod} + \delta_2 \mathit{utility} + \delta_3 \mathit{trans} + u$$

where *trans* is a dummy variable for the transportation industry. Now, the base group is *finance*, and so the coefficient δ_1 directly measures the difference between the consumer products and finance industries, and we can use the t statistic on *consprod*.

Solution to Activity 6.13

(i) We want to have a constant semi-elasticity model, so a standard wage equation with marijuana usage included would be:

$$\log(\mathit{wage}) = \beta_0 + \beta_1 \mathit{usage} + \beta_2 \mathit{educ} + \beta_3 \mathit{exper} + \beta_4 \mathit{exper}^2 + \beta_5 \mathit{female} + u.$$

Then $100 \cdot \beta_1$ is the approximate percentage change in wage when marijuana usage increases once per month.

(ii) We would add an interaction term in female and usage:

$$\log(wage) = \beta_0+\beta_1 usage+\beta_2 educ+\beta_3 exper+\beta_4 exper^2+\beta_5 female+\beta_6 female\cdot usage+u.$$

The null hypothesis that the effect of marijuana usage does not differ by gender is $H_0 : \beta_6 = 0$.

(iii) We take the base group to be nonuser. Then we need dummy variables for the other three groups: *lghtuser*, *moduser*, and *hvyuser*. Assuming no interactive effect with gender, the model would be:

$$\begin{aligned}\log(wage) = \beta_0 &+ \delta_1 lghtuser + \delta_2 moduser + \delta_3 hvyuser\\ &+ \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + \beta_5 female + u.\end{aligned}$$

(iv) The null hypothesis is $H_0 : \delta_1 = 0, \delta_2 = 0, \delta_3 = 0$, for a total of $q = 3$ restrictions. We have to estimate restricted and unrestricted models and compute sums of squared residuals in these two models: denote them as SSR_r and SSR_{ur}, respectively.

In the F test the numerator *df* is $q = 3$. If n is the sample size, the denominator *df* in the F distribution is $n - 8$. The F statistic is computed using the formula:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-8)}.$$

We obtain the critical value from the $F_{q,\,n-8}$ distribution and compare it with our F statistic. In this testing procedure, we assume that MLR.1–MLR.6 hold.

(v) The error term could contain factors, such as family background (including parental history of drug abuse), that could directly affect wages and also be correlated with marijuana usage. We are interested in the effects of a person's drug usage on his or her wage, so we would like to hold other confounding factors fixed. We could try to collect data on relevant background information.

Solution to Activity 6.14

(i) There are certainly notable differences in the point estimates. For example, having a young child in the household leads to less sleep for women (about two hours per week) while men are estimated to sleep about an hour more. The quadratic in age is a hump shape for men: however, the turning point is $age^* \approx 79.92$, so the regression line for men is increasing in age for all of the sample values. The quadratic in age is a U-shape for women: the turning point is $age^* \approx 41.25$, so the regression line for women is decreasing up to that point and is increasing after it.

The intercepts for men and women are also notably different.

(ii) The *df* are 6 (numerator) and 694 (denominator).

The 5% critical value of the $F_{6,\,694}$ distribution is 2.10. Since $F = 2.12 > 2.10$, under MLR.1 to MLR.6 we reject the null that the sleep equations are the same at the 5% level (but barely so).

(iii) The *df* are 5 (numerator) and 694 (denominator).

The 5% critical value of the $F_{5,694}$ distribution is 2.21. Since, $F = 1.26 < 2.21$, under MLR.1–MLR.6 we do not reject the null that the sleep equations are the same at the 5% level.

(iv) The outcome of the test in part (iii) shows that, once an intercept difference is allowed, there is not strong evidence of slope differences between men and women. This is one of those cases where the practically important differences in estimates for women and men in part (i) do not translate into statistically significant differences. We apparently need a larger sample size to determine whether there are differences in slopes. For the purposes of studying the sleep–work trade-off, the original model with male added as an explanatory variable seems sufficient.

6.4 Overview of chapter

This chapter begins with a discussion of models with quadratics which provide flexible regression specifications and can deliver increasing or decreasing partial effects. In addition, these models allow for a constant partial effect as a special case, which can be easily tested. Quadratics also permit a researcher to model U-shaped (or inverse U-shaped) relationships between a regressor and the dependent variable, which could be of interest in some applications. The estimation and interpretation of such models is illustrated in applications and the analysis is supplemented with hypothesis testing that relies on t or F tests.

The chapter then discusses how to formulate and interpret models with interaction terms which allow a partial effect of a regressor to depend on another regressor. These models allow for a constant partial effect as a special case and it is discussed how one could test for this. An important point for models with interactions (as well as quadratics) is that partial effects need to be evaluated at interesting values of the explanatory variables. In general, this would require centering variables that we want to interact around interesting values before constructing the interaction term, which typically leads to an equation that is more appealing.

The chapter also reviews the main properties of R^2 and discusses advantages and limitations of its use in empirical work. It is noted that R^2 will always increase when adding an additional regressor even if it is irrelevant. This motivates the use of the adjusted R-squared which penalises the number of regressors and can drop when an independent variable is added. The chapter further discusses the dangers of pursuing high values of R^2 (or the adjusted R-squared) in practice.

This chapter then considers dummy/binary variables and discusses an example of a regression setting with a single dummy variable. It defines the base (or reference) category and gives the interpretation of the regression parameters in the presence of a dummy variable as a regressor. It then discusses the changes in the regression interpretation if the baseline category is chosen in a different way. The issue of the dummy variable trap is introduced in this simple setting and it is discussed how it can be avoided.

It is shown how to construct dummy variables for multiple categories and discussed how the choice of the benchmark category affects the interpretation of regression parameters and t tests. It is analysed and illustrated how ordinal information in data can be captured by means of dummy variables.

Finally, the chapter considers models where the set of regressors includes dummy variables interacted either with each other or continuous variables. There are substantial differences between these two situations and it is explained why the latter case creates situations when we allow for a difference in slopes among different groups. Finally, it is described how to test the null hypothesis that two populations or groups follow the same regression function, against the alternative that one or more of the slopes differ across the groups (Chow test).

6.4.1 Key terms and concepts

- Quadratic function
- Turning point
- Interaction term
- Interaction effect
- Centering of variables
- Adjusted R-squared
- Dummy variable
- Base group
- Benchmark group
- Difference in intercepts
- Dummy variable trap
- Ordinal variable
- Difference in slopes
- Difference in both intercepts and slopes
- Chow test

6.4.2 A reminder of your learning outcomes

After completing this segment, and having completed the accompanying reading and activities, you should be able to:

- understand how to formulate and interpret models with quadratic and/or interaction terms
- be able to conduct hypothesis testing using t tests and F tests and interpret the results of such testing

- motivate the use of the adjusted R-squared
- discuss limitations of pursuing high values of R^2 or adjusted R-squared
- estimate and interpret models with quadratic and/or interaction terms in practice
- define a dummy variable and describe an impact it has on the regression specification and the interpretation of regression coefficients when it is used as a regressor
- describe how multiple dummy variables can be used to create further more refined categories in practice
- understand how ordinal information, such as ratings, can be captured by using dummy variables
- know how the choice of the base category affects the interpretation of the regression results (including the interpretation of parameters and t tests)
- explain the dummy variable trap and how it can be avoided
- motivate and describe a setting where a dummy variable is interacted with a continuous regressor
- describe and perform the Chow test in two equivalent ways (and also two variations of the Chow test – allowing or not intercepts to be different)
- estimate and interpret models with dummy variables in practice.

6.5 Test your knowledge and understanding

6.5.1 Additional examination based exercises

6.1E Consider the human capital earnings function given by:

$$earnings_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i, \quad i = 1, \ldots, n$$

where *earnings* denotes the hourly earnings of an individual and *educ* and *exper* denote the years of schooling and experience, respectively. We assume we have obtained a random sample $\{(earnings_i, educ_i, exper_i)\}_{i=1}^n$ from the population. The errors $\{u_i\}_{i=1}^n$ are i.i.d. normal random variables with zero mean and variance σ^2. We assume independence between the errors and regressors.

(a) Discuss the rationale for including both *exper* and *exper*^{2} in this model. In your answer be explicit about the expected signs for β_2 and β_3.

(b) Provide a clear interpretation of the parameter β_1.

6.2E Let us consider the following cross-sectional model for household consumption:

$$C_i = \beta_0 + \beta_1 Y_i + \beta_2 Y_i^2 + u_i$$

where Y_i denotes household income, $E(u_i) = 0$ and $E(u_i^2) = \sigma^2$. We will assume that the error is independent of Y_i (and therefore Y_i^2 as well). You are provided with a random sample $\{(C_i, Y_i)\}_{i=1}^n$.

(a) What will happen to the OLS parameter estimates when we measure consumption and income in £1,000s instead of £? Explain your answer.

(b) Let the above model represent the true relationship between consumption and income. Discuss the effect of ignoring the quadratic term on the OLS estimator of β_1. Will the bias be upward or downward? Explain your answer.

6.3E This question looks at the determinants of extramarital affairs using the AFFAIRS dataset. It has 601 observations and 19 variables. The outcome we will work with is the variable '*naffairs*', which lists the number of affairs the individual had in the past year. While the dataset lists many possible determinants (regressors of interests) of this outcome, we will here focus on three, namely how individuals rate their marriage ('*ratemarr*'), how religious they are ('*relig*'), and how many years they have been married ('*yrsmarr*'). '*ratemarr*' is a variable from 1 to 5, where 5 = very happy, 4 = somewhat happy, 3 = average, 2 = somewhat unhappy, and 1 = very unhappy. '*relig*' is a variable from 1 to 5 with 5 = very religious, 4 = somewhat religious, 3 = slightly religious, 2 = not at all religious, and 1 = anti religion. Every individual in the dataaset identifies either as a male or a female (please see the comment on the use of gender variables in the Course Overview).

(a) Using only observations for women, we estimate the following regression equation:

$$\widehat{naffairs} = \underset{(.878)}{4.987} - \underset{(.161)}{.678}\, ratemarr - \underset{(.158)}{.519}\, relig + \underset{(.034)}{.089}\, yrsmarr$$

$$n = 315, \quad R^2 = .1204, \quad \bar{R}^2 = .1119.$$

Interpret the results of this estimation. Do you think any effects you found here can be interpreted causally?

(b) Do you think it makes sense to include the square of *yrsmarr* as an additional regressor of interest? Discuss. If so, what sign would you expect for the quadratic term of *yrsmarr*?

(c) Including the square of *yrsmarr* into the regression in (a), we obtain the following results:

$$\widehat{naffairs} = \underset{(.936)}{4.950} - \underset{(.162)}{.678}\, ratemarr - \underset{(.159)}{.519}\, relig + \underset{(.144)}{.105}\, yrsmarr - \underset{(.008)}{.000978}\, yrsmarrsq$$

$$n = 315, \quad R^2 = .1205, \quad \bar{R}^2 = .1091.$$

What conclusion do you draw based on this output? Additionally, what do you make of the fact that the usual R^2 increased but the adjusted R-squared decreased compared to (a)?

(d) Now we return to the regression in (a) and consider including an interaction between *yrsmarr* and *ratemarr*. Discuss how we would interpret such an interaction term, were we to include it in our model in (a).

(e) Estimation of the model in (d) gives the following results:

$$\widehat{naffairs} = \underset{(1.394)}{2.057} - \underset{(.304)}{.018}\, ratemarr - \underset{(.157)}{.515}\, relig + \underset{(.117)}{.392}\, yrsmarr - \underset{(.028)}{.076}\, yrsmarr \times ratemarr$$

$$n = 315, \quad R^2 = .1404, \quad \bar{R}^2 = .1294.$$

Interpret these results.

(f) Estimation of models in (a) and (d) but using observations for men only gives the following results:

$$\widehat{naffairs} = \underset{(.8783)}{4.9631} - \underset{(.1774)}{.7431}\, ratemarr - \underset{(.1578)}{.4746}\, relig + \underset{(.0349)}{.1116}\, yrsmarr$$

$$n = 286, \quad R^2 = .1260, \quad \bar{R}^2 = .1136$$

and:

$$\widehat{naffairs} = \underset{(1.451)}{4.575} - \underset{(.341)}{.065}\, ratemarr - \underset{(.158)}{.478}\, relig + \underset{(.132)}{.154}\, yrsmarr - \underset{(.032)}{.011}\, yrsmarr \times ratemarr$$

$$n = 286, \quad R^2 = .1256, \quad \bar{R}^2 = .1163.$$

What do you make of these results? Are your conclusions for men similar to those for women? Discuss.

6.4E Suppose a researcher has wage data on randomly selected workers and each worker in the dataset identifies either as a male or a female. The researcher uses the data to estimate the OLS regression:

$$\widehat{wage} = \underset{(.16)}{10.73} + \underset{(.29)}{1.78}\, male$$

$$n = 440, \quad R^2 = .09$$

where *wage* is measured in dollars per hour and *male* is a binary variable that is equal to 1 if the person is a male and 0 if the person is a female. Define the wage–gender gap as the difference in mean earnings between men and women.

(a) In the sample, what is the mean wage of women? What is the mean wage of men?

(b) What is the estimated gender gap?

(c) Is the estimated gender gap significantly different from 0 at the 5% significance level?

(d) Construct a 95% confidence interval for the gender gap.

(e) Another researcher suggests including $male^2$ along with the dummy variable *male*. Explain whether this suggestion makes sense.

(f) A different researcher suggests including both dummy variables *female* (a variable that is equal to 1 if the person is female and 0 if the person is male) and *male* in the regression equation. Explain whether this suggestion makes sense.

(g) Instead of the model above you decide to estimate the model:

$$wage = \beta_0 + \beta_1 sinfemale + \beta_2 marrfemale + \beta_3 marrmale + u$$

where *sinfemale* is the dummy variable for single women, *marrfemale* is the dummy variable for married women, and *marrmale* is the dummy variable for married men. What is the interpretation of the coefficient β_1 in this model? What is the interpretation of $\beta_2 - \beta_1$ in this model?

(h) Another researcher uses these same data but regresses *wage* on *female*. What are the regression estimates calculated from this regression? In other words, indicate $\widehat{\alpha}_0$, $\widehat{\alpha}_1$ and R^2 in the fitted regression:

$$\widehat{wage} = \widehat{\alpha}_0 + \widehat{\alpha}_1 female$$
$$n = 440, \quad R^2 = \ldots$$

6.5E The OLS output in Table 6.1 is produced using a dataset on teaching evaluations for university professors. *fem_age* is the interaction between the variables *female* and *age*, *beauty*2 to *beauty*4 are dummy variables for the second, third and fourth quartile in the distribution of beauty ratings, and *onecredit* refers to one credit (short) courses. The total number of observations is 463, of which 195 are women. Suppose that each professor in the dataset identifies either as a male or a female.

female	.5949 (.2731)
age	.0042 (.0034)
fem_age	-.0169 (.0057)
*beauty*2	.2320 (.0738)
*beauty*3	.1831 (.0748)
*beauty*4	.3407 (.0737)
onecredit	.5096 (.1087)
constant	3.6421 (.1964)
R^2	.1361

Table 6.1: The dependent variable is *course_eval* (course evaluations). Standard errors are in parentheses.

(a) Is the coefficient on *beauty*2 significantly different from 0 at the 1% level?

(b) What is the interpretation of the coefficient on *age*?

(c) By how much do we expect the course evaluation of a female professor to differ if she is one year older? (Give a number.)

(d) What would happen if you added the interaction between variables *male* and *age* to the regression (where $male = 1 - female$)?

(e) How would you test whether the effect of age for females is significantly different from 0? Can you conduct this test using the information in the output in Table 6.1?

(f) Test whether the effect of age is different for males and females.

(g) Suppose you would like to test whether the intercept and all slopes are the same across men and women. To do that we estimate the model:

$$course\_eval = \beta_0 + \beta_1 age + \beta_2 beauty2 + \beta_3 beauty3 + \beta_4 beauty4 + \beta_5 onecredit + u$$

three times. First, we estimate it on the whole data set and obtain $SSR = 129.3731$. Then, we estimate it on the subset of women only and obtain $SSR = 51.3907$. Finally, we estimate it on the subset of men only and obtain $SSR = 66.8034$.

Formulate the null hypothesis for your test, calculate the test statistic for your test and discuss whether or not you reject your null hypothesis at the 1% level.

6.5.2 Solutions to additional examination based exercises

6.1E (a) With β_2 positive and β_3 negative, this permits diminishing returns to experience. While earnings increase with an additional year of experience its effect is smaller for individuals for more experience (as long as $\beta_2 + 2\beta_3 exper > 0$, after which additional years of experience reduce earnings). For other values of the β_js a different explanation should be given.

(b) β_1 denotes the effect one additional year of schooling has on the average hourly earnings of an individual holding everything else constant.

6.2E (a) The parameter β_1 remains unchanged, but β_0 and β_2 need to be rescaled. Let $\widetilde{C} = C/1000$ and $\widetilde{Y} = Y/1000$ denote our new variables, let's substitute them:

$$1000\widetilde{C} = \beta_0 + \beta_1 1000\widetilde{Y} + \beta_2(1000\widetilde{Y})^2 + u$$

$$\widetilde{C} = \frac{\beta_0}{1000} + \beta_1\widetilde{Y} + 1000\beta_2\widetilde{Y}^2 + u.$$

The intercept will be divided by 1,000 whereas β_2 will be multiplied by 1,000.

(b) If we estimate:

$$C = \beta_0 + \beta_1 Y + e \text{ (short equation)}$$

we will get the classical OVB problem. We need to consider the estimator given by:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(C_i - \bar{C})}{\sum_{i=1}^n (Y_i - \bar{Y})^2}.$$

Plugging in the true model (long equation), we can write:

$$\widehat{\beta}_1 = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i^2 - \overline{Y^2})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} + \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(u_i - \bar{u})}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \text{ where } \overline{Y^2} = \tfrac{1}{n}\sum Y_i^2$$

$$= \beta_1 + \beta_2 \frac{SampleCov(Y_i, Y_i^2)}{SampleVar(Y_i)} + \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})u_i}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}.$$

As $E(u_i \,|\, Y_1, \ldots, Y_n) = E(u_i)$ by independence, and $E(u_i) = 0$, this yields:

$$E(\widehat{\beta}_1 \,|\, Y_1, \ldots, Y_n) = \beta_1 + \underbrace{\beta_2 \frac{SampleCov(Y_i, Y_i^2)}{SampleVar(Y_i)}}_{OVB}.$$

As we expect that $SampleCov(Y_i, Y_i^2)$> 0 and $\beta_2 < 0$, the bias will be negative.

6.3E (a) The coefficient on *ratemarr* tells us that women who rate their marriage as being happier tend to have fewer affairs. Specifically, a one-unit increase in *ratemarr* is associated with a decrease of .678 affairs. The coefficient is statistically significantly different from zero at the 1% level (under MLR.1 to MLR.6).

The coefficient on *relig* tells us that more religious women tend to have fewer affairs. Specifically, a one-unit increase in *relig* is associated with a decrease of .519 affairs. The coefficient is statistically significantly different from zero at the 1% level (under MLR.1 to MLR.6).

The coefficient on *yrsmarr* tells us that women who are married for longer tend to have more affairs compared to women who are married less long. Specifically, being married for one more year is associated with an increase of .089 affairs. The coefficient is statistically significant at the 1% level (under MLR.1 to MLR.6).

(b) If we were to include the square of *yrsmarr*, we would presume that the coefficient on this new regressor is negative. Why is that? Well, (a) above tells us that there is a positive correlation between the length of the marriage and the number of affairs a woman has. However, note that a marriage that holds for a long time also concerns people who tend to 'be good' together (in that they otherwise may have gotten divorced). We may think that such couples are less likely to cheat on one another. Thus, we would expect a negative sign on the square of *yrsmarr*: the longer a couple is married, we expect a decreasing positive effect on the number of affairs someone has. (Indeed, it seems possible that at some point the effect even becomes negative.) It thus may indeed make sense to include this square in the regression.

(c) Notice that the intercept and the coefficients on *ratemarr* and *relig* are basically unchanged, in terms of magnitude, sign, and significance. The interpretation of these is therefore analogous to (a).

yrsmarrsq is indeed negative, as hypothesised in (b). However, notice that both *yrsmarr* and *yrsmarrsq* are both no longer significant. Also notice that the effect of *yrsmarrsq* is incredibly small. Recall what this means: the coefficient on *yrsmarrsq* is the 'amount' by which the positive effect of *yrsmarr* decreases over time. Since this decrease is basically zero, the data is telling us that the hypothesised effect in (b) may be irrelevant, at least in this dataset. Hence, we may be better off dropping it.

R-squared cannot decrease when we add an additional regressor, so its increase is expected. However, the fact that the adjusted R-squared decreased means that *yrsmarrsq* does not do much in terms of explaining variation in the dependent variable.

(d) When including an interaction between *yrsmarr* and *ratemarr* we would expect a negative sign in front of this coefficient. Why? We may think that an individual in a longer marriage who is also happier may cheat less often on their spouse because they are happier. Hence, this interaction should have a negative effect in that it should reduce the number of times an individual cheats on his or her spouse.

(e) Note first that the coefficient on *relig* is still pretty much unchanged in terms of sign, magnitude, and significance. Thus, the interpretation here is analogous to the above.

The interaction term of *yrsmarr* and *ratemarr* does indeed have a negative coefficient and is statistically significant at the 1% level (under MLR.1 to MLR.6). The coefficient on *yrsmarr* is still positive (but much larger) and also statistically significant at the 1% level (under MLR.1 to MLR.6). The coefficient on *ratemarr* is, while still negative, no longer significantly different from zero. What does this tell us?

This tells us that for women (recall, we are only running this regression for women at the moment), how happy they are in their marriage is only a determinant for the number of affairs they have when interacted with the length of the marriage, but not by itself. More intuitively, a women who has been in a marriage longer (compared to a woman who has been in a marriage less long) tends to have more affairs. This (positive) effect, however, is diminished for women who are happy in their marriage. Another way of putting this is that women who are unhappy in their marriage only tend to have more affairs as time goes by.

(f) For the output using only the subsample of men, we see that all the coefficients have the same sign as for women. However, with the exception of *relig* (and the intercept) all of them are statistically insignificant (under MLR.1 to MLR.6). This tells us that the effect of the interaction term on women we hypothesised about may not be prevalent in our data for men. Indeed, if we run the regression on men without the interaction (as reported), we find coefficients on all three right-hand side variables that are (highly) statistically significant (under MLR.1 to MLR.6). What does this tell us? It suggests that the nuanced effect we found for women (i.e. that how happy they are in their marriage only matters when interacted with *yrsmarr*) doesn't exist for men. By running a regression without the interaction, we see that men who are happy in their marriage or who are highly religious tend to cheat less and men who have been in their marriage longer tend to cheat more. But, at least judging from this one interaction term we included, the data for men does not exhibit the heterogeneity that the female sample did. To put it differently, a man in an unhappy marriage tends to cheat more often on his partner (no interaction term!), whereas an unhappy woman tends to cheat more often on her partner only as time goes by (interaction term!).

6.4E (a) The average wage for women is \$10.73/hour. The average wage for men is \$12.51/hour.

(b) The estimated gender gap equals \$1.78/hour.

(c) The hypothesis testing for the gender gap is $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. With a t statistic of:

$$t = \frac{1.78}{0.29} \approx 6.14$$

we can reject the null hypothesis that there is no gender gap at the 5% significance level.

(d) The 95% confidence interval for the gender gap β_1 is $1.78 - 1.96 \times 0.29, 1.78 + 1.96 \times 0.29]$, that is $[1.2116, 2.3484]$.

(e) $male^2$ coincides with *male* and, thus, adding $male^2$ creates the situation of perfect collinearity, so it is not a reasonable suggestion.

(f) It creates the dummy variable trap and the situation of perfect collinearity, so it is not a reasonable suggestion.

(g) β_1 is the gender premium/discrimination for single people. β_1 is the gap in the average wages between single men and single women. $\beta_2 - \beta_1$ is the marriage premium for women – the difference in the average wages for married women and single women.

(h) The binary variable regression model relating wages to gender can be written as either:

$$wage = \beta_0 + \beta_1 male + u$$

or:

$$wage = \alpha_0 + \alpha_1 female + v.$$

In the first regression equation, *male* equals 1 for men and 0 for women; β_0 is the population mean of wages for women and $\beta_0 + \beta_1$ is the population mean of wages for men. In the second regression equation, *female* equals 1 for women and 0 for men; α_0 is the population mean of wages for men and $\alpha_0 + \alpha_1$ is the population mean of wages for women. We have the following relationship for the coefficients in the two regression equations:

$$\alpha_0 = \beta_0 + \beta_1$$
$$\beta_0 = \alpha_0 + \alpha_1.$$

Given the OLS estimated $\widehat{\beta}_0$ and $\widehat{\beta}_1$, we have:

$$\widehat{\alpha}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 = 12.51$$
$$\widehat{\alpha}_1 = \widehat{\beta}_0 - \widehat{\alpha}_0 = -\widehat{\beta}_1 = -1.78.$$

Due to the relationship among coefficient estimates, for each individual observation, the OLS residual is the same under the two regression equations: $\widehat{u}_i = \widehat{v}_i$. Thus the sum of squared residuals, $SSR = \sum_{i=1}^{n} \widehat{u}_i^2$ is the same under the two regressions. This implies that $R^2 = 1 - \frac{SSR}{SST}$ is unchanged. In summary, in regressing *wage* on *male*, we will get:

$$\widehat{wage} = 12.5 - 1.78 male, \quad R^2 = .09.$$

6.5E (a) The t statistic for *beauty*2 is about 3.143. Since $|3.143| > 2.576$, we conclude that *beauty*2 is significantly different from 0 at the 1% level (under MLR.1–MLR.6).

(b) The coefficient on *age* denotes the average difference in the course evaluation for a male professor if he is one year older, holding everything else constant.

(c) We expect the course evaluation to be lower by 0.012696:

$$.004242 - .016938 = -.012696.$$

(d) We would find ourselves in the situation of perfect multicollinearity as:

$$age = female \cdot age + male \cdot age.$$

(e) We would want to test the null hypothesis:

$$H_0 : \beta_{age} + \beta_{fem\_age} = 0$$

against, say, $H_1 : \beta_{age} + \beta_{fem\_age} \neq 0$. The t statistic to test this H_0 is written as:

$$t = \frac{\widehat{\beta}_{age} + \widehat{\beta}_{fem\_age}}{\sqrt{se(\widehat{\beta}_{age})^2 + se(\widehat{\beta}_{fem\_age})^2 + 2s_{12}}}$$

where s_{12} is the estimator of $Cov(\widehat{\beta}_{age}, \widehat{\beta}_{fem\_age})$. We cannot compute the t statistic just from the output in Table 6.1 as we need s_{12} to compute the t statistic.

(f) We want to test the null hypothesis:

$$H_0 : \beta_{fem\_age} = 0$$

against $H_1 : \beta_{fem\_age} \neq 0$. This can be done by the t test. Under CLM assumptions (MLR.1 to MLR.6):

$$\frac{\beta_{fem\_age}}{se(\beta_{fem\_age})} \sim t_{455} \quad \text{under } H_0.$$

The realisation of our test statistic is $-.016938/.005713 = -2.9648$. The absolute value of this realisation exceeds the two-sided critical value even at the 1% level of significance (2.576) revealing its statistical significance. Thus, the effect of age is different for males and females at the 1% significance level. The fact that we use all CLM assumptions (MLR.1 to MLR.6) means that, in addition to our Gauss–Markov assumptions, we need to assume normality of the errors.

(g) In the regression model:

$$\begin{aligned} course\_eval = {} & \beta_0 + \beta_1 age + \beta_2 beauty2 + \beta_3 beauty3 + \beta_4 beauty4 + \beta_5 onecredit \\ & + \delta_0 female + \delta_1 female \cdot age + \delta_2 female \cdot beauty2 + \delta_3 female \cdot beauty3 \\ & + \delta_4 female \cdot beauty4 + \delta_5 female \cdot onecredit + u \end{aligned}$$

we want to test:

$$H_0 : \delta_0 = 0,\ \delta_1 = 0,\ \delta_2 = 0,\ \ldots,\ \delta_5 = 0$$

for 6 restrictions against the alternative which is the negation of the null:

$$H_1 : \text{at least one of } \delta_j, j = 0, \ldots, 5, \text{ is not 0.}$$

Under CLM assumptions (MLR.1 to MLR.6):

$$F = \frac{[SSR_P - (SSR_f + SSR_m)]/6}{(SSR_f + SSR_m)/451} \sim F_{6,\,451} \quad \text{under } H_0$$

where SSR_P is obtained on the whole data set, SSR_f is obtained on the subset of women only, and SSR_m is obtained on the subset of men only.

The realisation of our Chow test statistic is:

$$\frac{[129.3731 - (51.3907 + 66.8034)]/6}{(51.3907 + 66.8034)/451} = 7.1094$$

and it exceeds the critical value even at 1% level of significance (2.80) leading us to reject the null hypothesis at the 1% level of significance.

Chapter 7
Multiple regression analysis: Asymptotics

7.1 Introduction

7.1.1 Aims of the chapter

In this chapter we expand our knowledge of the properties of OLS estimators. We focus on the so-called large sample properties, which describe the behaviour of OLS estimators as the sample size increases to infinity.

7.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the implications of relaxing the assumption of the normal distribution of regression errors for the finite-sample inference
- explain the consequences of endogeneity for OLS estimation
- have an intuition of what the consistency property means and be able to establish the consistency of the OLS estimator
- understand which conditions guarantee the consistency property of the OLS estimator
- explain what it means for the standardised OLS estimator to be asymptotic normal
- understand the implications of using large sample results for statistical inference
- explain in which and under what assumptions the OLS estimator is asymptotically efficient.

7.1.3 Essential reading

Reading: Wooldridge, Sections 5.1–5.3 and Appendices C.3a and C3.b.

7.1.4 Synopsis of this chapter

Section 5.1 in Wooldridge establishes conditions under which the OLS estimator is consistent – that is, converges to the true parameter (with probability approaching 1) as the sample size increases. Consistency is especially useful in cases where the finite properties like unbiasedness do not hold, and we do discuss situations when OLS estimators are consistent but biased.

Section 5.2 in Wooldridge then presents the asymptotic distribution of the normalised OLS estimator. This asymptotic distribution is standard normal, which allows us to conclude that in large samples a researcher is able to conduct approximate inference using confidence intervals and t and F statistics even when removing the assumption of the normal distribution of regression errors (or equivalently, removing the assumption of the normal distribution of the dependent variable conditional on the explanatory variables).

Section 5.3 in Wooldridge discusses under what conditions OLS is asymptotically efficient.

7.2 Content of chapter

7.2.1 Consistency

Read: Wooldridge, Section 5.1 and Appendix C.3a.

It is recommended that you start by reviewing Appendix C.3a about the notion of the consistency of an estimator. In a nutshell, consistency means that the probability that the estimate is arbitrarily close to the true population value can be made arbitrarily close to 1 by increasing the sample size. It can be said that consistency is a minimum requirement for sensible estimators. It is important for us to know that consistency is especially useful in cases where the finite properties like unbiasedness do not hold. What does it mean if neither unbiasedness nor consistency holds? It means the estimator is biased even in large samples and, hence, cannot tell us much about the population parameter.

This section analyses whether OLS estimators are consistent or, more precisely, what assumptions are required to guarantee the consistency of OLS estimators. It establishes that under Assumptions MLR.1 through MLR.4, the OLS estimator $\widehat{\beta}_j$ is consistent for β_j , for all $j = 0, \ldots, k$, as $n \to \infty$. Thus, as we get more data, we can expect the sampling distribution of $\widehat{\beta}_j$ to become more tightly centred around β_j. Intuitively, this result also says that *the sampling variance* $\text{Var}(\widehat{\beta}_j \,|\, \mathbf{X})$ *shrinks to zero as* $n \to \infty$.

We can illustrate the consistency property of the OLS estimators in simulations. We simulate data from the model:

$$y = 5 - 3x_1 + 2x_2 + u$$

where u has mean 0 and is independent of x_1 and x_2 (there are some distributions used for x_1, x_2, and u with these properties). We consider sample sizes $n = 100$, $n = 1{,}000$, and $n = 10{,}000$. The results of simulations are presented in Figures 7.1 to 7.9. As we can see from these figures, the distribution of $\widehat{\beta}_j$ becomes more and more concentrated around β_j, for $j = 1, 2, 3$, as n increases. This is exactly what we expect from the consistency property.

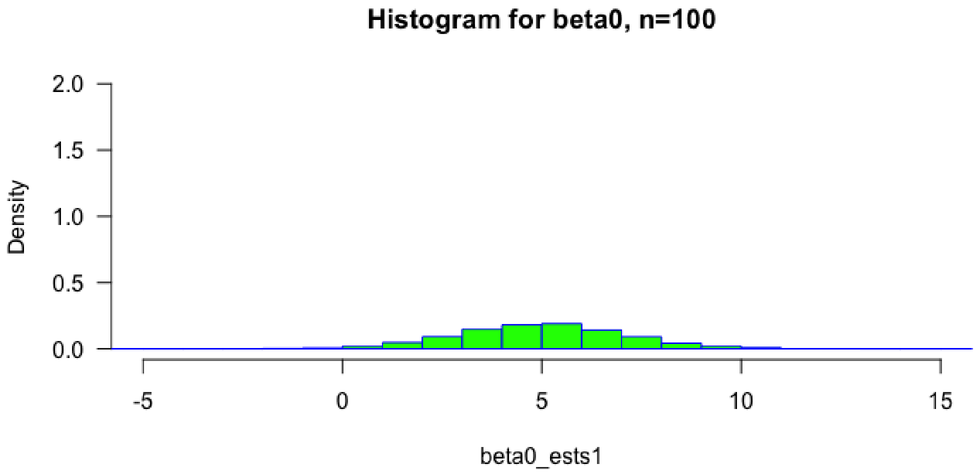


Figure 7.1: Approximate distribution of $\widehat{\beta}_0$ for $n = 100$.

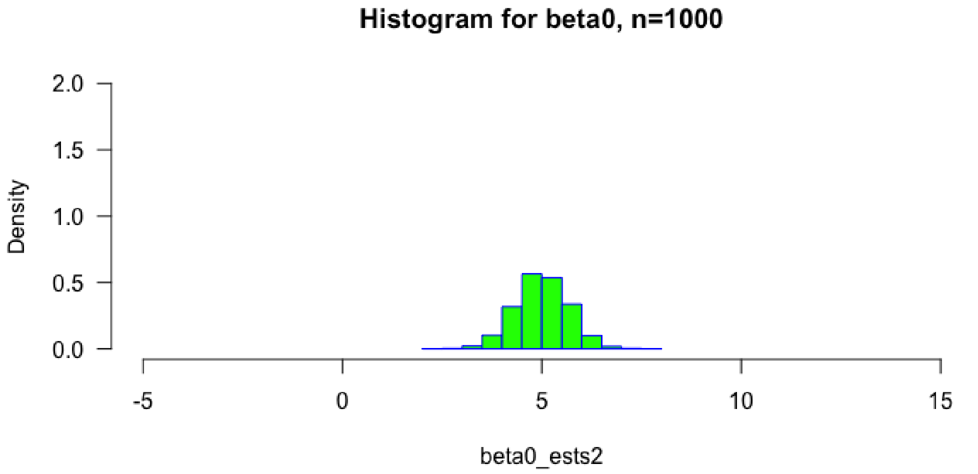


Figure 7.2: Approximate distribution of $\widehat{\beta}_0$ for $n = 1{,}000$.

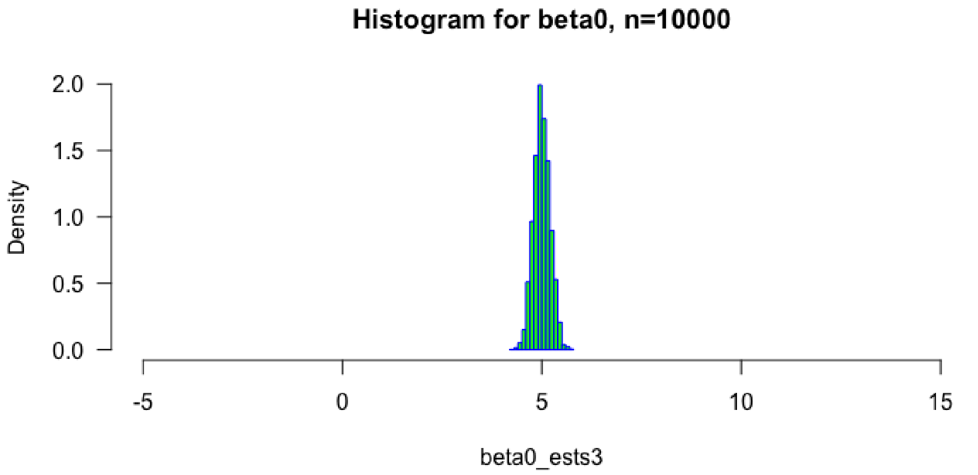


Figure 7.3: Approximate distribution of $\widehat{\beta}_0$ for $n = 10{,}000$.

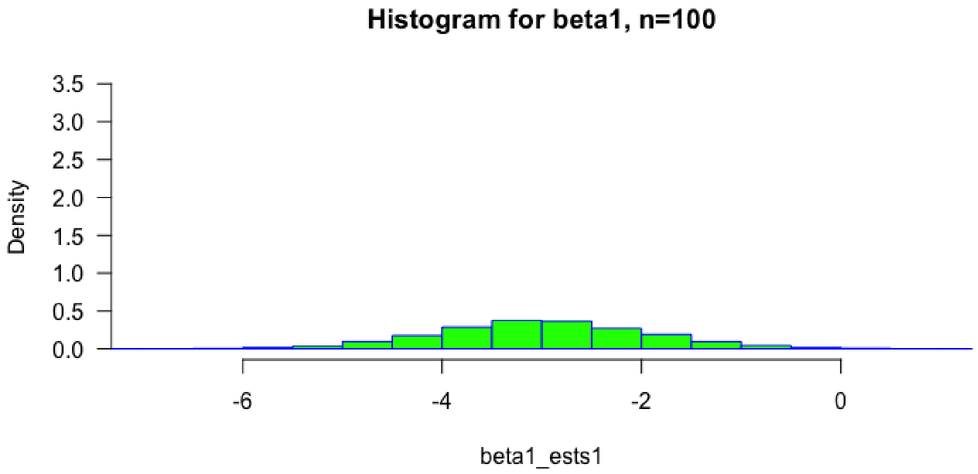


Figure 7.4: Approximate distribution of $\widehat{\beta}_1$ for $n = 100$.

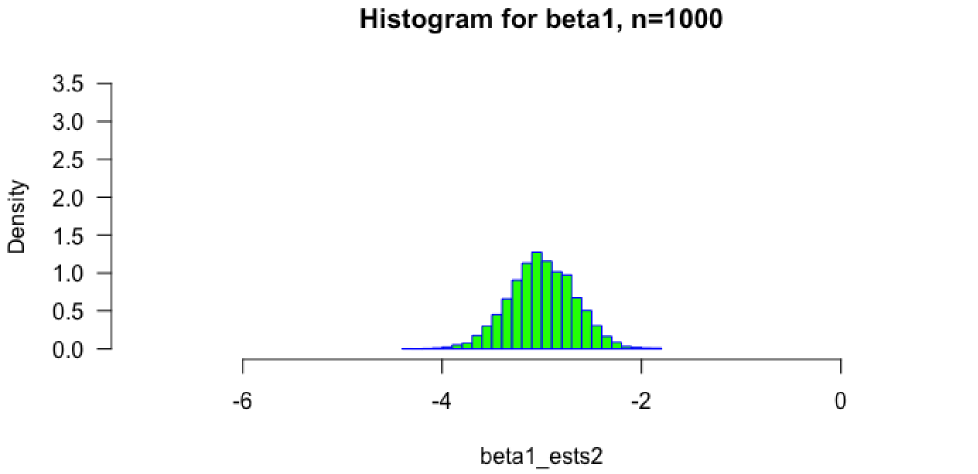


Figure 7.5: Approximate distribution of $\widehat{\beta}_1$ for $n = 1{,}000$.

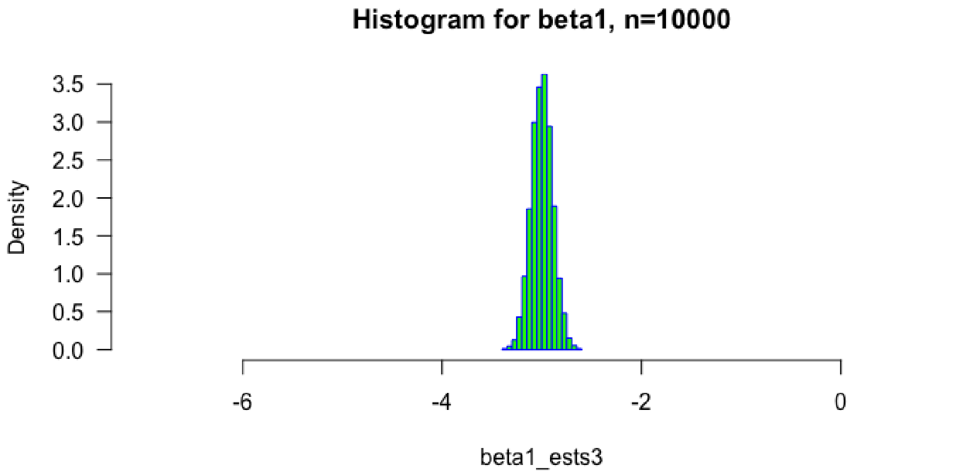


Figure 7.6: Approximate distribution of $\widehat{\beta}_1$ for $n = 10{,}000$.

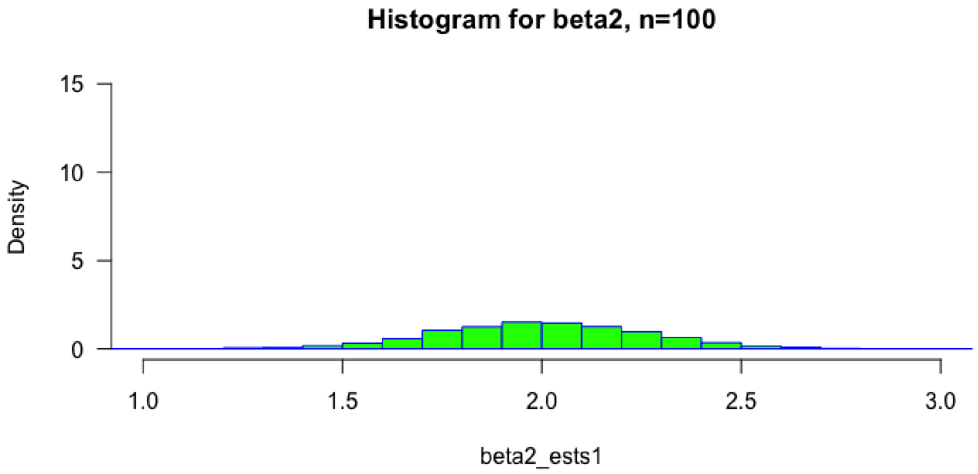


Figure 7.7: Approximate distribution of $\widehat{\beta}_2$ for $n = 100$.

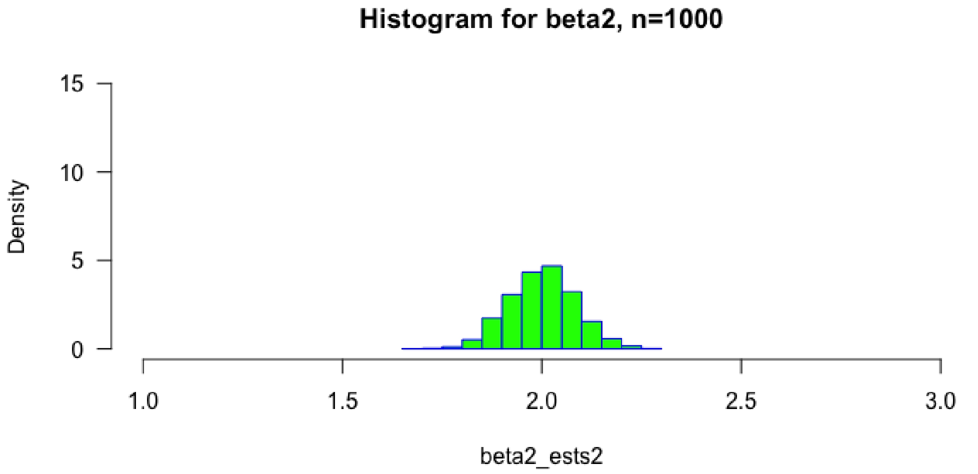


Figure 7.8: Approximate distribution of $\widehat{\beta}_2$ for $n = 1{,}000$.

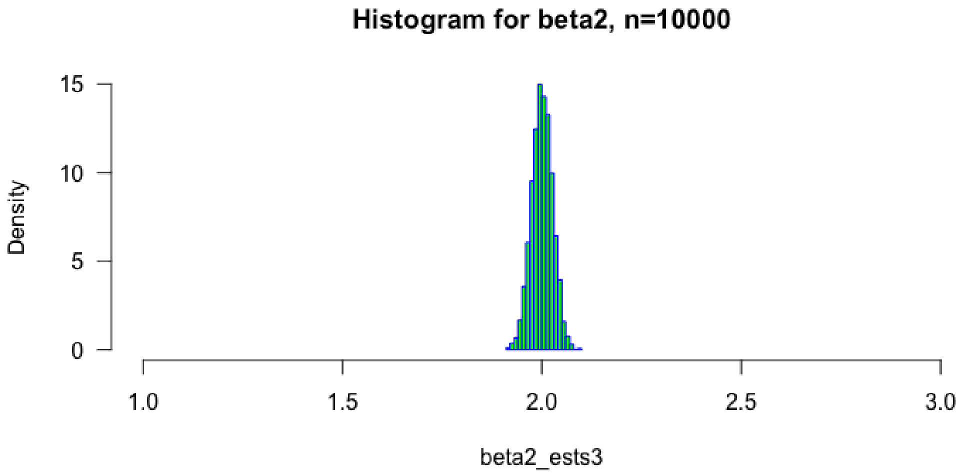


Figure 7.9: Approximate distribution of $\widehat{\beta}_2$ for $n = 10{,}000$.

The proof of consistency of the OLS slope estimator in the bivariate model presented in Wooldridge, Section 5.1 uses LLN (a general case would use LLN as well). The proof also makes it clear that in the bivariate model consistency of the slope estimator would hold if we required just uncorrelatedness between the regressor x and the error u. This extends to a general case and it turns out that the OLS estimator is consistent under assumptions MLR.1 through MLR.3 and MLR.4$'$, where Assumption MLR.4$'$ is the following:

Assumption MLR.4$'$ (Zero Mean and Zero Correlation)

Assumption MLR.4$'$ is:

$$E(u) = 0 \quad \text{and} \quad Cov(x_j, u) = 0, \quad j = 1, 2, \ldots, k.$$

Importantly, Assumption MLR.4$'$ is a much weaker requirement than Assumption MLR.4. This is because under MLR.4 $E(u \,|\, x_1, x_2, \ldots, x_k) = 0$, the error u should be uncorrelated with any function of $(x_1, \ldots, x_k)$, and MLR.4$'$ only requires u to be uncorrelated with each x_j, which is a particular function of $(x_1, \ldots, x_k)$.

We can also conclude that there are situations when OLS estimators are biased but consistent. These are the situations when MLR.1 to MLR.3 hold and Assumption MLR.4$'$ holds but MLR.4 does not hold.

One implication of the consistency result MLR.1 through MLR.3 and MLR.4$'$ is that correlation between u and any of $x_1, x_2, \ldots, x_k$ generally causes **all** of the OLS estimators to be inconsistent. This simple but important observation is often summarised as: **if the error is correlated with any of the regressors, then OLS is biased and inconsistent** (note that if MLR.4$'$ is violated, then MLR.4 is violated as well).

Activity 7.1 Take the MCQs related to this section on the VLE to test your understanding.

Activity 7.2 Exercise 5.2 from Wooldridge.

7.2.2 Asymptotic normality and large sample inference

Read: Wooldridge, Section 5.2 and Appendix C.3b.

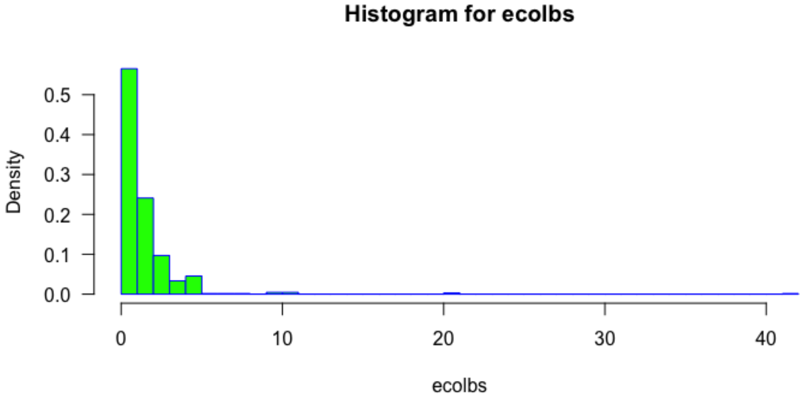
It is recommended that you start by reviewing Appendix C.3b about the notion of the asymptotic normality of an estimator.

This section discusses that the exact t and F distributions for the t and F statistics are derived under MLR.6, along with the other Classical Linear Model assumptions. Therefore, it is of importance to analyse whether Assumption MLR.6 is plausible. Since under MLR.1 ($y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$), Assumption MLR.6 can be translated into the condition of the normal distribution of the dependent variable: *conditional on* *the regressors, our dependent variable y is normally distributed*, that is:

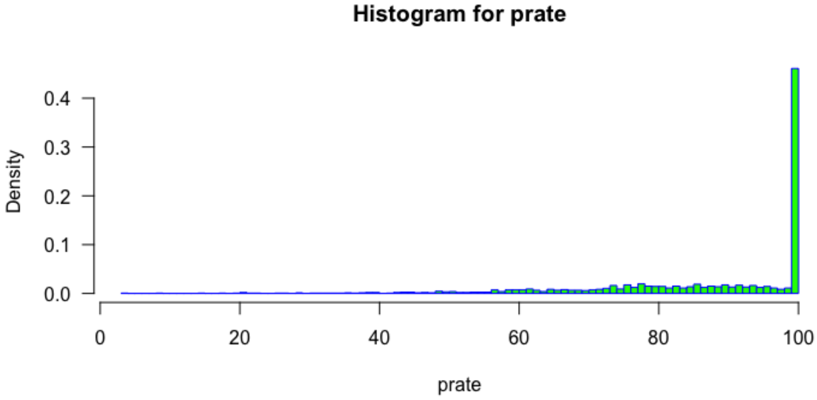
$$y \,|\, x_1, \ldots, x_k \sim Normal(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \sigma^2)$$

then whether assumption MLR.6 holds or not can be studied from the perspective of the distribution of the dependent variable.

It is argued in this section that in practice, the normality assumption MLR.6 is often questionable. As an example, take the APPLE dataset, in which a dependent variable *ecolbs* (pounds of apples demanded) does not have a conditional normal distribution. By definition, *ecolbs* cannot be negative. More importantly, it equals zero more than one-third of the time. This is confirmed by the histogram (or bar chart) for *ecolbs* below.



Many dependent variables are the same way. Another example is the 401K dataset. The interesting dependent variable there is *prate*, the percentage of employees at a firm participating in a 401(k) pension plan. By definition, $0 \leq prate \leq 100$. Plus, more than 44% of the observations are at 100. The histogram for *prate* below shows its distribution is nothing like the continuous bell shape of the normal.



This section discusses that in cases such as described above, fortunately, we do not have to abandon statistical inference if our sample size is even moderately large. (The sample size in the APPLE dataset is 660 and that in the 401K dataset is 1,534, and these turn out to be plenty large enough.) Namely, even when the observations of the dependent variable are not from a normal distribution, we can use the central limit theorem to conclude that the OLS estimators satisfy **asymptotic normality**. This is formulated in the following theorem.

Theorem (Asymptotic normality of OLS)

Under Assumptions MLR.1 to MLR.5, the standardised version of $\widehat{\beta}_j$ is asymptotically standard normal:

$$\frac{\widehat{\beta}_j - \beta_j}{sd(\widehat{\beta}_j)} \overset{a}{\sim} N(0,1).$$

The same convergence holds when we replace $sd(\widehat{\beta}_j)$ with $se(\widehat{\beta}_j)$:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \overset{a}{\sim} N(0,1).$$

The proof of this theorem is not given as it is somewhat involved (it relies largely on the CLT and also the LLN and some other probabilistic laws). The theorem essentially states that what was exactly true with MLR.6 remains approximately true, in ‘large’ samples, even without MLR.6.

Because t_{n-k-1} tends to $Normal(0,1)$ as n tends to ∞, it is also legitimate to write:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)} \overset{a}{\sim} t_{n-k-1}.$$

In practice, we just treat the quantity $\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)}$ as having an approximate t_{n-k-1} distribution. This means that *everything* we discussed for inference – t testing, confidence intervals, and even F testing – can be applied without the normality assumption. The caveat is that the testing is only **approximate**.

Sometimes when we know MLR.6 fails – as in the APPLE dataset example – we emphasise that our p-values are ‘asymptotic p-values’ or ‘approximate p-values’ that are based on ‘asymptotic standard errors’. But often this is left implicit.

We can illustrate the asymptotic normality property of the OLS estimators in simulations. We will simulate data from the model:

$$y = 5 - 3x_1 + 2x_2 + u$$

where x_1 is drawn from the binomial distribution with mean 0.3, x_2 is drawn from the standard normal distributions, and u is a random variable that takes values -1 and 4 with probabilities 0.8 and 0.2, respectively.

The distribution visuals in Figures 7.10 to 7.12 are for the ratio:

$$\frac{\widehat{\beta}_j - \beta_j}{se(\widehat{\beta}_j)}.$$

In Figures 7.10 to 7.12 we also draw the density of the standard normal distribution for comparison. As the distribution of $\frac{\widehat{\beta}_j-\beta_j}{se(\widehat{\beta}_j)}$ is very close to the normal distribution even for $n = 500$.

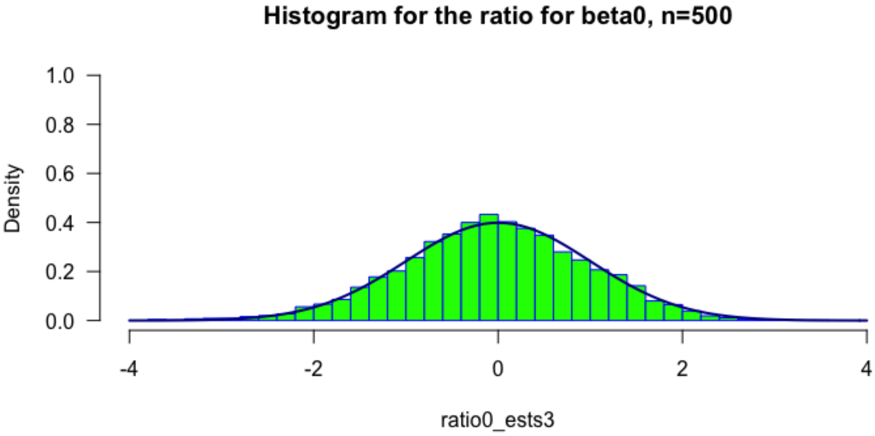


Figure 7.10: Approximate distribution of $\widehat{\beta}_0$, $n = 500$.

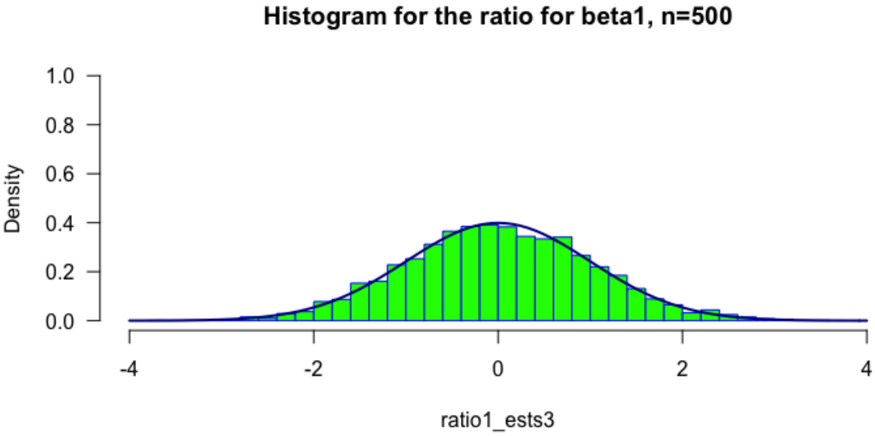


Figure 7.11: Approximate distribution of $\widehat{\beta}_1$, $n = 500$.

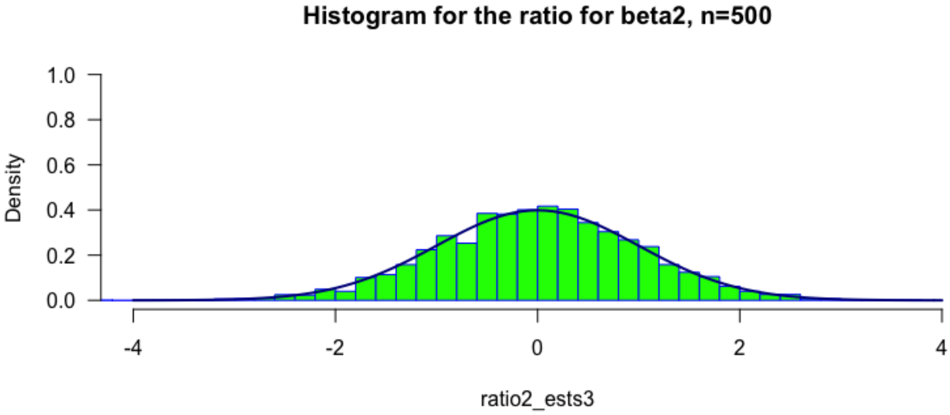


Figure 7.12: Approximate distribution of $\widehat{\beta}_2$, $n = 500$.

Activity 7.3 Take the MCQs related to this section on the VLE to test your understanding.

Activity 7.4 After estimating the equation:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u$$

using WAGE1 data, a researcher plots the histogram of OLS residuals and also plots the normal distribution that provides the best fit to the histogram in Figure 7.13:

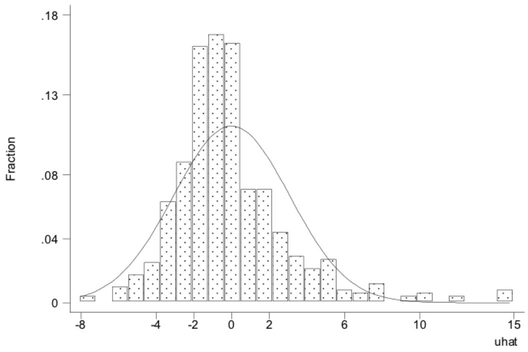


Figure 7.13: Histogram of OLS residuals.

Repeating above, but with log($wage$) as the dependent variable, the researcher obtains the histogram in Figure 7.14 for the residuals with the best-fitting normal distribution overlaid:

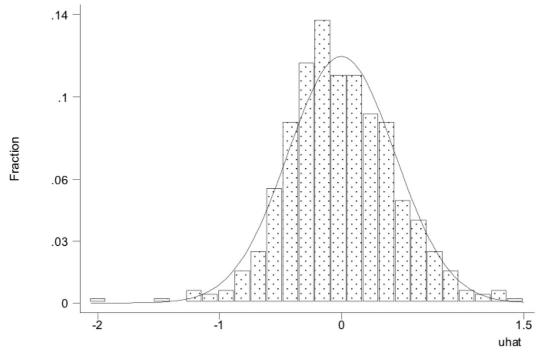


Figure 7.14: Histogram of OLS residuals.

Would you say that Assumption MLR.6 is closer to being satisfied for the level-level model or the log-level model?

Activity 7.5 Several statistics are commonly used to detect nonnormality in underlying population distributions. Here we will study one that measures the amount of skewness in a distribution. Recall that any normally distributed random variable is symmetric about its mean; therefore, if we standardise a symmetrically distributed random variable, say $z = (y - \mu_y)/\sigma_y$, where $\mu_y = E(y)$ and $\sigma_y = sd(y)$, then z has mean zero, variance one, and $E(z^3) = 0$. Given a sample of data $\{y_i : i = 1, \ldots, n\}$, we can standardise y_i in the sample by using $z_i = (y_i - \widehat{\mu}_y)/\widehat{\sigma}_y$, where $\widehat{\mu}_y$ is the sample mean and $\widehat{\sigma}_y$ is the sample standard deviation. (We ignore the fact that these are estimates based on the sample.) A sample statistic that measures skewness is $\frac{1}{n}\sum_{i=1}^{n} z_i^3$, or where n is replaced with $(n-1)$ as a degrees-of-freedom adjustment. If y has a normal distribution in the population, the skewness measure in the sample for the standardised values should not differ significantly from zero.

(i) First, using the data set 401KSUBS and keeping only observations with *fsize* $= 1$, we find that the measure of skewness for *inc* is about 1.86 and when we use log(*inc*), the skewness measure is about .360. Which variable has more skewness and therefore seems less likely to be normally distributed?

(ii) Next, using BWGHT2 we find that the skewness for *bwght* is about $-.60$. When we use log(*bwght*), the skewness measure is about -2.95. What do you conclude?

(iii) Evaluate the following statement: 'The logarithmic transformation always makes a positive variable look more normally distributed.'

(iv) If we are interested in the normality assumption in the context of regression, should we be evaluating the unconditional distributions of y and $\log(y)$? Explain.

7.2.3 Asymptotic efficiency of OLS

Read: Wooldridge, Section 5.3.

We know that the OLS estimator is best linear unbiased under the Gauss–Markov assumptions. It turns out that this same set of assumptions implies an asymptotic efficiency property of OLS, too. The discussion of this is in Section 5.3 of Wooldridge but the details are not important for our purposes. We just need to remember that OLS is the best we can do, using finite sample and asymptotic arguments, under Assumptions MLR.1 to MLR.5.

7.3 Answers to activities

Solution to Activity 7.2

What is this question about? In the regression we are asked to run – of *pctstck* on *funds* – regressor *funds* is correlated with the error because of the omitted variable *risktol*. We will obtain the **inconsistency** in $\widetilde{\beta}_1$ (sometimes loosely called the **asymptotic bias**). In other words, we will derive the asymptotic analogue of the omitted variable bias.

Let's consider a more general setting. Suppose the true model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + v$$

satisfies the assumptions MLR.1–MLR.4. Then v has a zero mean and is uncorrelated with x_1 and x_2.

If we omit x_2 from the regression and do the simple regression of y on x_1, then $u = \beta_2 x_2 + v$.

Let $\widetilde{\beta}_1$ denote the simple regression slope estimator. We can derive the probability limit of $\widetilde{\beta}_1$:

$$\begin{aligned}\text{plim } \widetilde{\beta}_1 = \beta_1 + \frac{Cov(x_1, u)}{Var(x_1)} &= \beta_1 + \frac{Cov(x_1, \beta_2 x_2 + v)}{Var(x_1)} \\ &= \beta_1 + \frac{Cov(x_1, \beta_2 x_2 + v)}{Var(x_1)} \\ &= \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)} + \frac{Cov(x_1, v)}{Var(x_1)} \\ &= \beta_1 + \beta_2 \frac{Cov(x_1, x_2)}{Var(x_1)}\end{aligned}$$

as $Cov(x_1, v) = 0$. Thus, the asymptotic bias is $\beta_2 \frac{Cov(x_1,x_2)}{Var(x_1)}$.

Now, let's go back to our question. A higher tolerance of risk means more willingness to invest in the stock market, so $\beta_2 > 0$. By assumption, *funds* and *risktol* are positively correlated. By the formula we derived above, plim $\widetilde{\beta}_1 > \beta_1$, so $\widetilde{\beta}_1$ has a positive inconsistency (asymptotic bias). This makes sense; if we omit *risktol* from the regression and it is positively correlated with *funds*, some of the estimated effect of *funds* is actually due to the effect of *risktol*.

The main takeaway from this activity is that the effect of the omitted variable bias persists in the limit and leads to inconsistency of OLS estimators. Thus, even if we have a large sample, we still have to worry about the omitted variable bias.

Solution to Activity 7.4

The residuals from the log(*wage*) regression appear to be more normally distributed. Certainly the histogram of the residuals from the log(*wage*) equation fits under its comparable normal density better than the histogram of the residuals from the *wage* equation, and the histogram for the *wage* residuals is notably skewed to the left.

In the *wage* regression, there are some very large residuals (roughly equal to 15) that lie almost five estimated standard deviations ($\widehat{\sigma} = 3.085$) from the mean of the residuals, which is identically zero, of course. Residuals far from zero do not appear to be nearly as much of a problem in the log(*wage*) regression.

Solution to Activity 7.5

(i) There is much less skewness in log of income, which means *inc* is less likely to be normally distributed. (In fact, the skewness in income distributions is a well-documented fact across many countries and time periods.)

(ii) In this case, there is much more skewness after taking the natural log.

(iii) The example in part (ii) clearly shows that this statement cannot hold generally. It is possible to introduce skewness by taking the natural log. As an empirical matter, for many economic variables, particularly dollar values, taking the log often does help to reduce or eliminate skewness. But it does not **have** to.

(iv) For the purposes of regression analysis, we should be studying the **conditional** distributions: that is, the distributions of y and $\log(y)$ conditional on the explanatory variables. If we think the mean is linear, as in Assumptions MLR.1 and MLR.3, then this is equivalent to studying the distribution of the population error, u. In fact, the skewness measure studied in this question often is applied to the residuals from OLS regression.

7.4 Overview of chapter

In this chapter the focus is on the large sample properties of OLS estimators. We learn that given MLR.1–MLR.3, the OLS estimators are consistent if MLR.4$'$ also holds. Assumption MLR.4$'$ imposes weaker restrictions than MLR.4 and only requires the error to be uncorrelated with each regressor. Given MLR.1–MLR.3, the OLS estimators are, of course, consistent if the more restrictive assumption MLR.4 holds.

Furthermore, it is discussed that all of the methods of testing and constructing confidence intervals that we learned earlier remain approximately valid without assuming that the errors are drawn from a normal distribution (equivalently, the distribution of y given the explanatory variables is not normal). This means that we can apply OLS and use associated testing methods for applications where the dependent variable is not even approximately normally distributed.

Finally, we learn that OLS is asymptotically efficient under Gauss–Markov assumptions.

7.4.1 Key terms and concepts

- Large sample properties
- Asymptotic properties
- Consistency
- Inconsistency
- Asymptotic distribution
- Asymptotic normality
- Asymptotic efficiency

7.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the implications of relaxing the assumption of the normal distribution of regression errors for the finite-sample inference
- explain the consequences of endogeneity for OLS estimation
- have an intuition of what the consistency property means and be able to establish the consistency of the OLS estimator
- understand which conditions guarantee the consistency property of the OLS estimator
- explain what it means for the standardised OLS estimator to be asymptotic normal
- understand the implications of using large sample results for statistical inference
- explain in which and under what assumptions the OLS estimator is asymptotically efficient.

7.5 Test your knowledge and understanding

7.5.1 Additional examination based questions

7.1E Consider the bivariate regression $y_i = \beta_0 + \beta_1 x_i + u_i$. Suppose a sample $\{(x_i, y_i\}_{i=1}^n$ is available and Assumptions MLR.1 to MLR.4 hold true. Assume that $\beta_0 = 0$. Consider the OLS estimator without intercept:

$$\widetilde{\beta}_1 = \arg\min_{b_1} \sum_{i=1}^{n} (y_i - b_1 x_i)^2.$$

Is $\widetilde{\beta}_1$ consistent for β_1? If so, prove it. If not, explain the reason.

7.2E (i) Consider the linear model:

$$score = \beta_0 + \beta_1 colgpa + \beta_2 actmth + \beta_3 acteng + u$$

where *score* is the final score in the course measured as a percentage, *colgpa* is the college GPA, *actmth* is the mathematics ACT score, *acteng* is the English ACT score.

Suppose the distribution of *score* is skewed to left tail. Why cannot Assumption MLR.6 hold for the error term u? What consequences does this have for using the usual t statistic to test $H_0 : \beta_3 = 0$?

(ii) Using the ECONMATH data, we can estimate the model from part (i) and obtain that the t value for *acteng* is -2.48 and the corresponding p-value is 0.013.

How would you defend your findings to someone who makes the following statement: 'You cannot trust that p-value because clearly the error term in the equation cannot have a normal distribution?'.

7.5.2 Solutions to additional examination based questions

7.1E We can derive that $\widetilde{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum x_i^2}$. Rewrite it as follows:

$$\widetilde{\beta}_1 = \frac{\sum_{i=1}^n x_i(\beta_1 x_i + u_i)}{\sum x_i^2} = \beta_1 + \frac{\sum_{i=1}^n x_i u_i}{\sum x_i^2} = \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^n x_i u_i}{\frac{1}{n}\sum x_i^2}.$$

By LLN:

$$\text{plim } \frac{1}{n}\sum_{i=1}^n x_i u_i = E[x_i u_i], \quad \text{plim } \frac{1}{n}\sum_{i=1}^n x_i^2 = E[x_i^2].$$

Then using the properties of the plim operator:

$$\text{plim } \widetilde{\beta}_1 = \beta_1 + \frac{E[x_i u_i]}{E[x_i^2]} = \beta_1 + \frac{0}{E[x_i^2]} = \beta_1$$

where we used $E[x_i u_i] = 0$ (implied by MLR.4) and $E[x_i^2] \neq 0$ (implied by MLR.3). Thus, $\widetilde{\beta}_1$ is consistent for β_1.

7.2E (i) The distribution of *score* being skewed to the left tail violates the normality assumption even conditional on the explanatory variables. Therefore, the assumption MLR.6 does not hold for the error term u and *score* will not be normally distributed, which means that the t statistics will not have t distributions and the F statistics will not have F distributions. This is a potentially serious problem because our inference hinges on being able to obtain critical values or p-values from the t or F distributions.

(ii) If the sample size is large, then the t distribution can be approximated to the distribution of the t statistics when the error term is not normally distributed.

Chapter 8
Heteroskedasticity

8.1 Introduction

8.1.1 Aims of the chapter

In this chapter we examine in detail the situation of heteroskedasticity in linear regression models. Heteroskedasticity occurs when the assumption of homoskedasticity is violated or, in other words, when the variance of the regression error, conditional on the explanatory variables, is not constant.

8.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the definitions of homoskedasticity and heteroskedasticity in linear regression models
- explain the consequences of heteroskedasticity for OLS estimators, their standard errors, confidence intervals, t and F tests
- describe the principles of heteroskedasticity-robust inference for the OLS estimation
- discuss different approaches to detecting and testing for heteroskedasticity
- implement different approaches to detecting and testing for heteroskedasticity in practice
- explain what a weighted least squares estimator is and describe its properties.

8.1.3 Essential reading

Reading: Wooldridge, Sections 8.1–8.4.

8.1.4 Synopsis of this chapter

Section 8.1 in Wooldridge discusses the impact of heteroskedasticity on the properties of OLS estimators and notes that OLS estimators are still unbiased and consistent (given that the first four Gauss–Markov assumptions hold), but the formulas for standard errors obtained under homoskedasticity are no longer valid. This leads us to employing heteroskedasticity-robust inference for the OLS estimation which uses the OLS estimators and focuses on modifying their standard errors and also test statistics in the presence of heteroskedasticity. This is discussed in Section 8.2 in Wooldridge. In Section 8.3 in Wooldridge we turn to the discussion of testing for heteroskedasticity and discuss various approaches: graphical analysis, Breusch–Pagan test, and the White test. In Section 8.4 in Wooldridge we look at dealing with the situation of heteroskedasticity from a different angle – instead of using OLS estimation and applying heteroskedasticity-robust inference, we develop and implement a weighted least squares method which delivers a more efficient estimator than OLS under correctly specified heteroskedasticity.

8.2 Content of chapter

8.2.1 Consequences of heteroskedasticity for OLS

Read: Wooldridge, Section 8.1.

This section analyses what happens with the OLS inference if we drop MLR.5 and act as if we know nothing about the conditional variance:

$$Var(u\,|\,x_1,\ldots,x_k) = Var(u\,|\,\mathbf{x}).$$

We know that OLS is still **unbiased and consistent** under MLR.1 to MLR.4. (We did not use MLR.5 to obtain either of these properties.) This is an important conclusion: *Heteroskedasticity does not cause bias or inconsistency in the OLS estimators* $\widehat{\beta}_j$ *as long as MLR.1 to MLR.4 continue to hold.*

However, if $Var(u\,|\,\mathbf{x})$ depends on $\mathbf{x}$ – that is, **heteroskedasticity** is present – then OLS is no longer BLUE as it is no longer 'best' (that is, it no longer has the lowest variance among linear and unbiased estimators). In principle, it is possible to find unbiased estimators that have smaller variances than the OLS estimators. A similar comment holds for asymptotic efficiency.

Importantly, the usual standard errors are **no longer valid**, which means the t statistics and confidence intervals that use these standard errors cannot be trusted. This is true even in large samples. Similarly, joint hypothesis tests using the usual F statistic are **no longer valid** in the presence of heteroskedasticity. In fact, the usual sum of squared residuals form of the F statistic is not valid under heteroskedasticity.

Both R^2 and $\bar{R}^2$ remain consistent estimators of the population R-squared. You may ask why the F statistic is no longer valid under heteroskedasticity even though R^2 remains valid (recall that we were able to write our formula for the F statistic in terms of R^2). This is because the usual sum of squared residuals form of the F statistic is not valid under heteroskedasticity and, consequently, the usual expression for the F statistic in terms of R^2 is no longer valid either.

Activity 8.1 Take the MCQs related to this section on the VLE to test your understanding.

8.2.2 Heteroskedasticity-robust inference after OLS estimation

Read: Wooldridge, Section 8.2.

Without MLR.5, there are still good reasons to use OLS as OLS estimators are unbiased and consistent, but we need to modify test statistics to make them valid under heteroskedasticity. Fortunately, it is known how to modify standard errors and t, F, and other statistics to be valid in the presence of **heteroskedasticity of unknown form**. This is very convenient because it means we can report new inferential statistics that are correct regardless of the kind of heteroskedasticity present in the population. This also includes the possibility of homoskedasticity (that is, MLR.5 actually holds). So we can compute confidence intervals and conduct statistical inference without worrying about whether MLR.5 holds.

Most regression packages include an option with the OLS estimation that computes **heteroskedasticity-robust standard errors**, which then produces **heteroskedasticity-robust *t* and *F* statistics** and **heteroskedasticity-robust confidence intervals**.

To give an illustration of the technical basis for heteroskedasticity-robust inference, let's consider the simple regression model:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

and assume throughout that MLR.1 to MLR.4 hold. We will use the notation:

$$Var(u_i \,|\, x_i) = \sigma_i^2$$

where an i subscript on σ_i^2 indicates that the variance of the error depends upon the particular value of x_i.

Recall that the OLS estimator can be written as:

$$\widehat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} w_i u_i$$

where $w_i = \frac{x_i - \bar{x}}{SST_x}$. Denote $\mathbf{X} = \{x_1, \ldots, x_n\}$. Under MLR.1 to MLR.4 (that is, without the homoskedasticity assumption), we can show that:

$$\begin{aligned} Var(\widehat{\beta}_1 \,|\, \mathbf{X}) = Var\left(\sum_{i=1}^{n} w_i u_i \,|\, \mathbf{X}\right) &= \sum_{i=1}^{n} Var(w_i u_i \,|\, \mathbf{X}) \\ &= \sum_{i=1}^{n} w_i^2 \, Var(u_i \,|\, \mathbf{X}) \\ &= \sum_{i=1}^{n} w_i^2 \, Var(u_i \,|\, x_i) \\ &= \sum_{i=1}^{n} w_i^2 \sigma_i^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}. \end{aligned}$$

For comparison: in the case of homoskedasticity, the variance of $\widehat{\beta}_1$ would be $\frac{Var(u)}{SST_x}$ (a special case of the formula above if you take $\sigma_i^2 = Var(u)$ for all i). How can we estimate this $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$? If $\widehat{u}_i$ denote the **OLS residuals** from the regression of y on x, then a valid estimator of $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ for **heteroskedasticity of any form (including homoskedasticity)**, is:

$$\widehat{Var}(\widehat{\beta}_1 \,|\, \mathbf{X}) = \frac{\sum_{i=1}^n (x_i - \bar{x})^2 \widehat{u}_i^2}{SST_x^2}.$$

The square root of $\widehat{Var}(\widehat{\beta}_1 \,|\, \mathbf{X})$ is called the **heteroskedasticity-robust standard error** for $\widehat{\beta}_1$.

There is an analogous formula for the variance of OLS estimators in the multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i.$$

Under assumptions MLR.1 to MLR.4, a valid estimator of $Var(\widehat{\beta}_j \,|\, \mathbf{X})$ is:

$$\widehat{Var}(\widehat{\beta}_j \,|\, \mathbf{X}) = \frac{\sum_{i=1}^n \widehat{r}_{ij}^2 \widehat{u}_i^2}{SSR_j^2}$$

where $\widehat{r}_{ij}$ is the ith residual from regressing x_j on all other regressors, and SSR_j is the sum of squared residuals from the regression of x_j on all other regressors.

The heteroskedasticity-robust t statistic is calculated as:

$$t = \frac{\text{estimate} - \text{hypothesised value}}{\text{standard error}}$$

where we use the OLS estimate in the numerator and the heteroskedasticity-robust standard error in the denominator.

It is sometimes incorrectly claimed that heteroskedasticity-robust standard errors for OLS are always larger than the usual OLS standard errors.

The heteroskedasticity-robust F statistic (or a simple transformation of it) is also called a **heteroskedasticity-robust Wald statistic**. The robust Wald statistic has an approximate chi-squared distribution in large samples, but it is easy to turn it into an approximate F distribution (this is what software packages do).

> **Activity 8.2** Take the MCQ related to this section on the VLE to test your understanding.

> **Activity 8.3** Exercise 8.4 in Wooldridge.

> **Activity 8.4** Answer the following.
>
> (i) Using the data in HPRICE1, we estimate the following regression equation and obtain both usual and heteroskedasticity-robust standard errors:
>
> $$\begin{array}{rlclcl} \widehat{price} = - & 21.77 & + & .00207 \; lotsize + & .123 \; sqrft + & 13.85 \; bdrms \\ & (29.48) & & (.00064) & (.013) & (9.01) \\ & [36.28] & & [.00122] & [.017] & [8.28] \end{array}$$
>
> $$n = 88, \quad R^2 = .672.$$

Discuss any important differences with the usual standard errors.

(ii) Using the same data, we estimate the regression equation with log(*price*) instead of *price*, and log(*lotsize*) and log(*sqrft*) instead of *lotsize* and *sqrft*, respectively:

$$\begin{array}{lrcrlrlrl}
\widehat{\log(price)} = & -\ 1.30 & + & .168 & \log(lotsize) + & .700 & \log(sqrft) + & .037 & bdrms \\
 & (.65) & & (.038) & & (.093) & & (.028) & \\
 & [.76] & & [.041] & & [.101] & & [.030] &
\end{array}$$

$$n = 88, \quad R^2 = .643.$$

Discuss any important differences with the usual standard errors.

(iii) What does this example suggest about heteroskedasticity and the transformation used for the dependent variable?

8.2.3 Testing for heteroskedasticity

Read: Wooldridge, Section 8.3.

This section begins with a discussion of why one might want to test for heteroskedasticity and focuses on some of the modern tests. There is a general framework we can apply to all these tests.

Before discussing these tests, it is worth mentioning that an informal way to detect heteroskedasticity is by means of a visual inspection of OLS residuals (or squared residuals) plotted against a fitted value or a regressor. To illustrate this, Figures 8.1 and 8.2 depict situations of homoskedasticity, whereas Figure 8.3 describes a situation of heteroskedasticity.

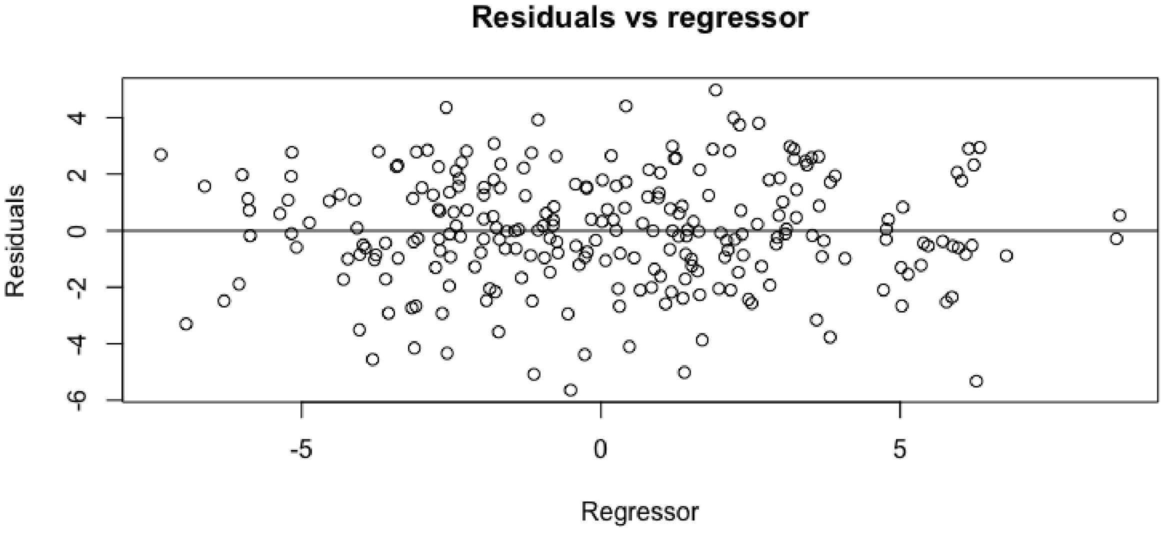


Figure 8.1: Example of homoskedasticity.

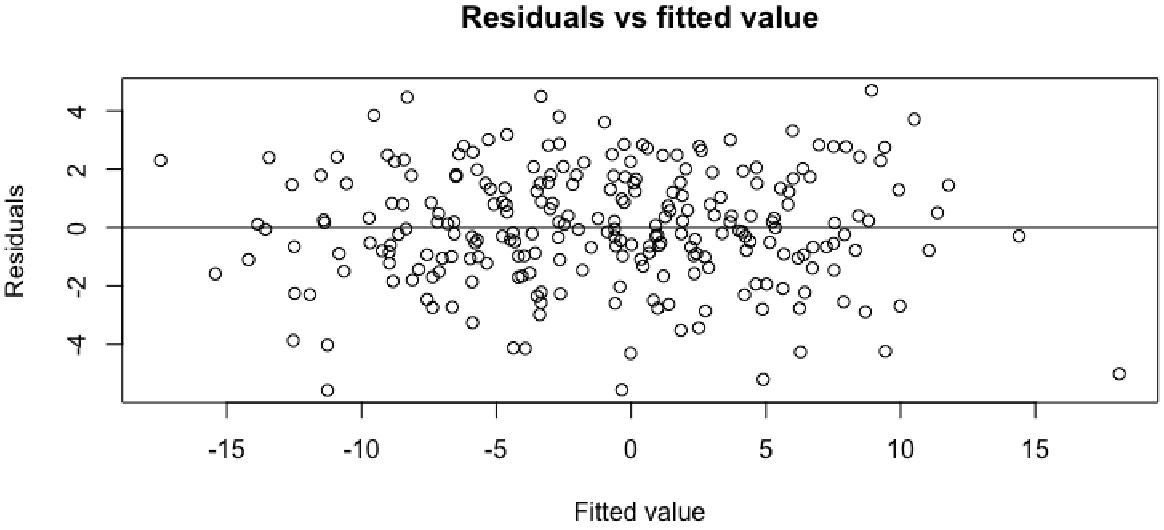


Figure 8.2: Example of homoskedasticity.

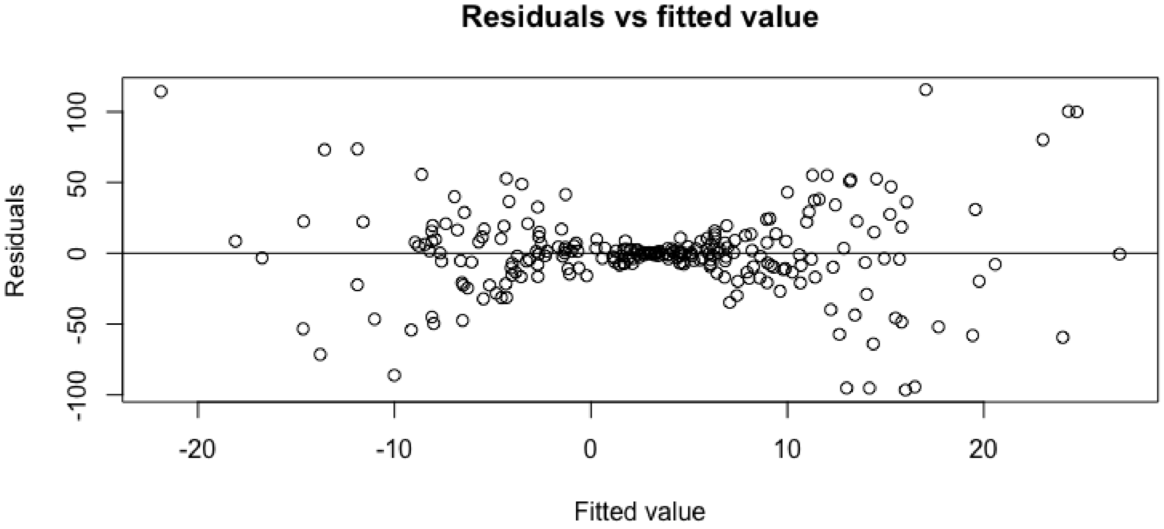


Figure 8.3: Example of heteroskedasticity.

In Figure 8.3 the residuals exhibit a systematic pattern as they appear to have a greater variance for large absolute values of fitted values.

Examination of residuals may also help detect a wrong functional form. For example, in Figure 8.4 positive residuals occur at moderate fitted values and negative residuals for high or low fitted values. This can be corrected by considering an appropriate non-linear model. Figure 8.5 illustrates another violation of this type.

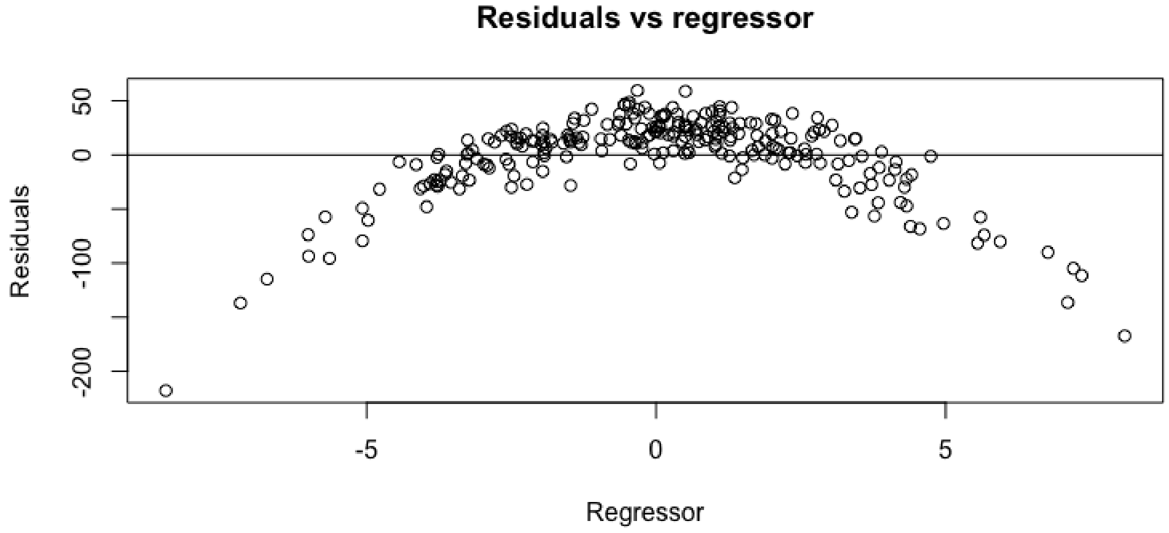


Figure 8.4: Examination of residuals.

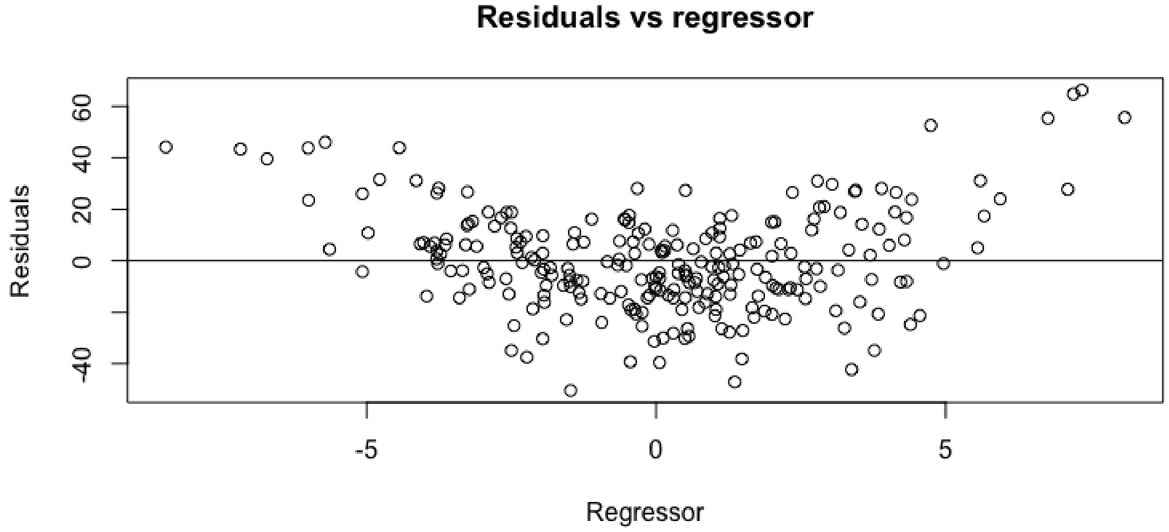


Figure 8.5: Examination of residuals.

For statistical tests for heteroskedasticity, we maintain:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$
$$E(u \,|\, \mathbf{x}) = 0$$

which are MLR.1 and MLR.4, respectively. As before we assume random sampling (MLR.2) and, of course, we rule out perfect collinearity (MLR.3).

Let $\mathbf{x} = (x_1, \ldots, x_k)$. Given MLR.3, Assumption MLR.5 can be written:

$$E(u^2 \,|\, \mathbf{x}) = \sigma^2 = E(u^2)$$

and so we want to test the null hypothesis:

$$H_0 : E(u^2 \,|\, \mathbf{x}) = \sigma^2 \text{ (constant)}.$$

If we cannot reject H_0 at a sufficiently small significance level, we usually conclude that heteroskedasticity is not a problem.

In order to test H_0 we need an alternative. The **Breusch–Pagan** test considers $E(u^2 \,|\, \mathbf{x})$ to be a linear function of explanatory variables:

$$E(u^2 \,|\, \mathbf{x}) = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k.$$

Therefore, the null hypothesis of homoskedasticity is:

$$H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$$

whereas the alternative hypothesis is:

$$H_1 : \text{at least one } \delta_j \text{ is not } 0.$$

Since we can write the linear model for $E(u^2 \,|\, x_1, \ldots, x_k)$ as the following model with an error term:

$$\begin{aligned} u^2 &= \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v \\ E(v \,|\, \mathbf{x}) &= 0 \end{aligned}$$

and since under $H_0 : \delta_1 = \delta_2 = \cdots = \delta_k = 0$ it makes sense to assume that v is independent of the x_js, in which case the equation:

$$u^2 = \delta_0 + \delta_1 x_1 + \cdots + \delta_k x_k + v$$

satisfies MLR.1 through MLR.5, then H_0 can be tested using the usual F test for joint significance of all explanatory variables. Note that we are testing the null hypothesis that the **squared** error is uncorrelated with each of the explanatory variables. Clearly, $u^2 \geq 0$ cannot be normally distributed. Consequently, any test using u^2 as the dependent variable must rely on large-sample approximations. Let us write the equation for a random draw i as:

$$u_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik} + v_i.$$

Now, we do not observe u_i^2 because these are the squared errors in the regression of y on $x_1, \ldots, x_k$! (Remember, we observe the y_i and x_{ij} for each draw i, but not the error u_i.) The solution is to **replace the errors with the OLS residuals**. In other words, the equation we estimate looks like:

$$\widehat{u}_i^2 = \delta_0 + \delta_1 x_{i1} + \cdots + \delta_k x_{ik} + \mathit{error}_i$$

where error_i is complicated because it depends on the $\widehat{\beta}_j$s as well as on v_i. Fortunately, it can be shown (not easily) that replacing u_i^2 with $\widehat{u}_i^2$ does **not** affect the asymptotic distribution of the F statistic. So we regress the **squared** OLS residuals on all of the explanatory variables and test their joint significance.

It is important to clearly understand the steps in computing the Breusch–Pagan test. First, the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$ is estimated by OLS and the OLS residuals, $\widehat{u}_i$, are saved. Then the squared residuals, $\widehat{u}_i^2$, are computed and regressed on all explanatory variables. For this regression, the usual F test of joint significance of the explanatory variables is computed. Then the p-value of the F test in step 2 is analysed. If it is sufficiently small, then the null of homoskedasticity is rejected and it is concluded that Assumption MLR.5 fails (throughout this testing procedure we maintain assumptions MLR.1–MLR.4).

The White test for heteroskedasticity supposes that $E(u^2 \mid \mathbf{x})$ is a linear function of all the explanatory variables, as with the Breusch–Pagan test, but also of all the nonredundant squares and interactions of all explanatory variables.

The first step of the White test for heteroskedasticity is exactly the same as in the Breusch–Pagan test – we run OLS of y on $x_1, \ldots, x_k$ and save the OLS residuals and calculate their squared values. The second step of the White test uses the squared OLS residuals from the first step as the dependent variable and regresses them on all the explanatory variables, all the nonredundant squares and interactions of all explanatory variables. For instance, for the regression:

$$\log(\mathit{wage}) = \beta_0 + \beta_1 \mathit{educ}_i + \beta_2 \mathit{exper}_i + u_i$$

the second step of the White test for heteroskedasticity would be the regression:

$$\widehat{u}_i^2 = \delta_0 + \delta_1 \mathit{educ}_i + \delta_2 \mathit{exper}_i + \delta_3 \mathit{educ}_i^2 + \delta_4 \mathit{exper}_i^2 + \delta_5 \mathit{exper}_i \cdot \mathit{educ}_i + \mathit{error}_i$$

and the null hypothesis would be:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = \delta_5 = 0$$

and the test statistic would, of course, be the joint F test for all explanatory variables. The third step of the White test is comparing the F statistic with a suitable critical value or just using its p-value. In conducting this test, we maintain assumptions MLR.1–MLR.4.

The White test has the advantage of detecting more general departures from heteroskedasticity than the Breusch–Pagan test. But the White test also has the disadvantage of using lots of degrees of freedom when there are lots of regressors in the original model. As we can see above, in the case of just two regressors *educ* and *exper* it leads to 6 parameters δ_j that need to be estimated. In the case of 6 regressors it would lead to 28 parameters δ_j that would need to be estimated.

There is a modification of the White test that is easier to implement and always has two degrees of freedom. The idea is to use the OLS fitted values in a test for heteroskedasticity. In this modification, in the first step we run OLS of y on $x_1, \ldots, x_k$ and save both the OLS residuals and fitted values. In the second and third steps we test for heteroskedasticity by estimating the equation:

$$\widehat{u}_i^2 = \delta_0 + \delta_1 \widehat{y}_i + \delta_2 \widehat{y}_i^2 + \mathit{error}_i$$

where $\widehat{y}_i$ stands for the fitted value for observation i, and then test $H_0 : \delta_1 = \delta_2 = 0$ using an F test (under MLR.1–MLR.4). It is important not to confuse $\widehat{y}_i$ and y_i in this equation.

On a general note, if we find heteroskedasticity after OLS estimation, at a minimum, we should compute the heteroskedasticity-robust standard errors, along with t statistics and confidence intervals. Joint hypothesis tests should also be made heteroskedasticity-robust. In modern practice, this is the most common response (if heteroskedasticity is even tested for in the first place).

Activity 8.5 Take the MCQs related to this section on the VLE to test your understanding.

Activity 8.6 Exercise 8.6 from Wooldridge.

Activity 8.7 This exercise uses data VOTE1 which has 173 observations.

(i) After estimating a model with *voteA* (percent vote for A) as the dependent variable and *prtystrA* (percent vote for president), *democA* (dummy variable equal to 1 if A is democrat), log(*expendA*), and log(*expendB*) as regressors, we obtain the fitted values, $\widehat{y}_i$, and the OLS residuals, $\widehat{u}_i$, and regress the OLS residuals on all of the regressors. In this latter regression we obtain $R^2 = 6.145 \times 10^{-32}$ (practically zero).

Explain why you obtain $R^2 \approx 0$.

(ii) The regression of $\widehat{u}_i^2$ on *prtystrA*, *democA*, log(*expendA*) and log(*expendB*) gives $R^2 = .05256$. Compute the F statistic for the Breusch–Pagan test for heteroskedasticity and discuss whether you reject homoskedasticity at the 5% level and whether you reject it at the 10% level.

(iii) The regression of $\widehat{u}_i^2$ on $\widehat{y}_i$, $\widehat{y}_i^2$ gives $R^2 = .03173$. Compute the F statistic for the special case of the White test for heteroskedasticity. How strong is the evidence for heteroskedasticity now? Discuss whether you reject homoskedasticity at the 5% level and whether you reject it at the 10% level.

8.2.4 Weighted least squares

Read: Wooldridge, Section 8.4.

Under heteroskedasticity (that is, the failure of MLR.5), the OLS estimator is no longer BLUE in general. There is a more efficient (i.e. a smaller variance) estimator than OLS under heteroskedasticity. This estimator is the so-called **weighted least squares (WLS)** estimator.

The construction of the WLS estimator requires us to know exactly the structure of heteroskedasticity, which was not the case before with heteroskedasticity-robust inference for OLS. If we have correctly specified the form of the variance as a function of explanatory variables, then weighted least squares (WLS) is more efficient than OLS.

Suppose that:

$$Var(u \,|\, x_1, \ldots, x_k) = \sigma^2 h(x_1, \ldots, x_k)$$

where σ^2 is unknown and the function $h(x_1, \ldots, x_k) > 0$ is known. From the multiple regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

we get a transformed model:

$$y_i^* = \beta_0 x_{i0}^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + u_i^*$$

where $y_i^* = \frac{y_i}{\sqrt{h(x_{i1},\ldots,x_{ik})}}$, $x_{i0}^* = \frac{1}{\sqrt{h(x_{i1},\ldots,x_{ik})}}$, $x_{ij}^* = \frac{x_{ij}}{\sqrt{h(x_{i1},\ldots,x_{ik})}}$, $j = 1, \ldots, k$, and $u_i^* = \frac{u_i}{\sqrt{h(x_{i1},\ldots,x_{ik})}}$. (Notice how the intercept also gets multiplied, it is no longer an intercept.)

Because $h(x_{i1}, \ldots, x_{ik})$ is just a function of $x_{i1}, \ldots, x_{ik}$, the transformed error $\frac{u_i}{\sqrt{h(x_{i1},\ldots,x_{ik})}}$ has a zero expected value conditional on $x_{i1}, \ldots, x_{ik}$:

$$E(u_i^* \,|\, x_{i1}, \ldots, x_{ik}) = E\left(\frac{u_i}{\sqrt{h(x_{i1}, \ldots, x_{ik})}} \middle|\, x_{i1}, \ldots, x_{ik} \right) = \frac{E(u_i \,|\, x_{i1}, \ldots, x_{ik})}{\sqrt{h(x_{i1}, \ldots, x_{ik})}} = 0$$

(here we use MLR.4 for u_i). This implies that MLR.4 holds in the transformed model:

$$E(u_i^* \,|\, x_{i0}^*, x_{i1}^*, \ldots, x_{ik}^*) = E(u_i^* \,|\, x_{i1}, \ldots, x_{ik}) = 0$$

(the first equality uses the fact that $x_{i0}^*, x_{i1}^*, \ldots, x_{ik}^*$ are functions of $x_{i1}, \ldots, x_{ik}$).

Since the conditional variance satisfies:

$$Var(a(x)u \,|\, x) = a(x)^2 \, Var(u \,|\, x)$$

for any function $a(x)$, we obtain:

$$\begin{aligned} Var(u_i^* \,|\, x_{i1}, \ldots, x_{ik}) &= Var\left(\frac{u_i}{\sqrt{h(x_{i1}, \ldots, x_{ik})}} \middle|\, x_{i1}, \ldots, x_{ik} \right) \\ &= \frac{Var(u_i \,|\, x_{i1}, \ldots, x_{ik})}{h(x_{i1}, \ldots, x_{ik})} = \frac{\sigma^2 h(x_1, \ldots, x_k)}{h(x_{i1}, \ldots, x_{ik})} = \sigma^2 \end{aligned}$$

so the transformed error u_i^* is **homoskedastic**.

Thus, if the original equation satisfies the first four Gauss–Markov assumptions, then the transformed equation satisfies all five Gauss–Markov assumptions. Hence, the OLS estimators in the transformed model will be BLUE. These new estimators are **Weighted Least Squares (WLS) estimators**.

Activity 8.8 Take the MCQs related to this section on the VLE to test your understanding.

Activity 8.9 This question uses the data in 401KSUBS, restricting the sample to $fsize = 1$.

(i) We estimated the equation that explains net total financial wealth (*nettfa*, measured in \$1,000s) in terms of income (*inc*, also measured in \$1,000s) and some other variables, including *age*, *gender*, an indicator $e401k$ for whether the person is eligible for a 401(k) pension plan, and also the interaction term $e401k \cdot inc$. We estimated the equation by OLS and obtained the usual and robust standard errors (the usual OLS standard errors are in $(\cdot)$ and the heteroskedasticity-robust standard errors are in $[\cdot]$):

$$\begin{aligned}\widehat{nettfa} = &- \underset{\substack{(2.82)\\ [3.23]}}{17.20} + \underset{\substack{(.080)\\ [.098]}}{.628}\ inc + \underset{\substack{(.0026)\\ [.0044]}}{.0251}\ (age-25)^2 + \underset{\substack{(2.04)\\ [2.06]}}{2.54}\ male \\ &- \underset{\substack{(4.40)\\ [6.25]}}{3.83}\ e401k + \underset{\substack{(.124)\\ [.220]}}{.343}\ e401k \cdot inc \\ &n = 2{,}107, \quad R^2 = .131.\end{aligned}$$

What do you conclude about the statistical significance of the interaction term?

(ii) Now we estimated the more general model by WLS obtained under the assumption $Var(u_i \,|\, inc_i) = \sigma^2 inc_i$. The results below show the usual and robust standard errors for the WLS estimator:

$$\begin{aligned}\widehat{nettfa} = &- \underset{\substack{(2.27)\\ [2.53]}}{14.09} + \underset{\substack{(.084)\\ [.091]}}{.619}\ inc + \underset{\substack{(.0019)\\ [.0026]}}{.0175}\ (age-25)^2 + \underset{\substack{(1.56)\\ [1.31]}}{1.78}\ male \\ &- \underset{\substack{(3.66)\\ [3.51]}}{2.17}\ e401k + \underset{\substack{(.130)\\ [.160]}}{.295}\ e401k \cdot inc.\end{aligned}$$

Note that even after applying WLS estimation, we may want to use heteroskedasticity-robust standard errors if we have doubts that we used the correct specification of heteroskedascitity or in order to convince ourselves the usual and heteroskedasticity-robust standard errors are similar.

Is the interaction term statistically significant using the robust standard error?

(iii) Discuss the WLS coefficient on $e401k$ in the more general model. Is it of much interest by itself? Explain.

(iv) If we reestimate the model by WLS but replace $e401k \cdot inc$ with the interaction term $e401k \cdot (inc - 30)$ (the average income in the sample is about 29.44), the coefficient on $e401k$ becomes 6.68 (robust $t = 3.20$).

Now interpret the coefficient on $e401k$.

8.2.5 R applications for heteroskedasticity

Activity: Complete ‘R Task – Activity 4’ on the VLE.

Here we show how to obtain White robust standard errors and test for heteroskedasticity in R.

8.3 Answers to activities

Solution to Activity 8.3

(i) The coefficients corresponding to *crsgpa*, *cumgpa*, and *tothrs* have the anticipated signs. If a student takes courses where grades are, on average, higher – as reflected by higher *crsgpa* – then their grades will be higher, other things being equal. The better the student has been in the past – as measured by *cumgpa* – the better the student does (on average) in the current semester, other things being equal. Finally, *tothrs* is a measure of effort, and its coefficient indicates an increasing return to effort, other things being equal.

The t statistic for *crsgpa* is very large, over five using the usual standard error (which is the largest of the two). Using the robust standard error for *cumgpa*, its t statistic is about 2.61, which is also significant at the 5% level (under MLR.1 to MLR.4). The t statistic for *tothrs* is only about 1.17 using either standard error, so it is not significant at the 5% level (under MLR.1 to MLR.4). Using robust standard errors makes the difference, e.g. for our conclusion regarding the significance of the effect of *season* (under MLR.1 to MLR.4).

(ii) If *crsgpa* was the only explanatory variable, $H_0 : \beta_{crsgpa} = 1$ means that, without any information about the student, the best predictor of term GPA is the average GPA in the students' courses; this holds essentially by definition. (The intercept would be zero in this case.)

With additional explanatory variables, it is not necessarily true that $H_0 : \beta_{crsgpa} = 1$ because $crsgpa$ could be correlated with characteristics of the student. For example, perhaps the courses students take are influenced by ability – as measured by test scores – and past college performance. But it is still interesting to test this hypothesis.

The t statistic using the usual standard error is $t = \frac{.900-1}{.175} \approx -.57$; using the heteroskedasticity-robust standard error gives $t \approx -.60$. Under MLR.1 to MLR.4, we fail to reject $H_0 : \beta_{crsgpa} = 1$ at any reasonable significance level, certainly including 5% (in either case).

(iii) The in-season effect is given by the coefficient on season, which implies that, other things equal, an athlete's GPA is about .16 points lower when their sport is competing. The t statistic using the usual standard error is about -1.60, while that using the robust standard error is about -1.96.

Under MLR.1 to MLR.4, against a two-sided alternative, the t statistic using the robust standard error is just significant at the 5% level (cv ≈ 1.96), while using the usual standard error, the t statistic is not quite significant at the 10% level (cv ≈ 1.65).

So the standard error used makes a difference in this case.

Solution to Activity 8.4

(i) $n = 88$ is still considered large enough for us to use robust standard errors. The robust standard error on *lotsize* is almost twice as large as the usual standard error, making *lotsize* much less significant (under MLR.1 to MLR.4) – the t statistic falls from about 3.23 to 1.70.

The t statistic on *sqrft* also falls, but it is still very significant (under MLR.1 to MLR.4). The variable *bdrms* actually becomes somewhat more significant, but it is still barely significant. The most important change is in the significance of *lotsize*.

(ii) For the log-log model, the heteroskedasticity-robust standard error is always slightly greater than the corresponding usual standard error, but the differences are relatively small. In particular, $\log(\textit{lotsize})$ and $\log(\textit{sqrft})$ still have very large t statistics.

The t statistic on *bdrms* is not significant at the 5% level (under MLR.1 to MLR.4) against a one-sided alternative using either standard error.

(iii) Using the logarithmic transformation of the dependent variable often mitigates, if not entirely eliminates, heteroskedasticity. This is certainly the case here, as no important conclusions in the model for $\log(\textit{price})$ depend on the choice of standard error. (We have also transformed two of the independent variables to make the model of the constant elasticity variety in *lotsize* and *sqrft*.)

Solution to Activity 8.6

(i) The proposed test is a hybrid of the Breusch–Pagan and White tests. There are $k + 1$ regressors, each original explanatory variable, and the squared fitted values. So, the number of restrictions tested is $k + 1$ and this is the numerator *df*. The denominator *df* is $n - (k + 2) = n - k - 2$.

(ii) In the Breusch–Pagan test, we regress:

$$\widehat{u}_i^2 \quad \text{on} \quad x_{i1}, x_{i2}, \ldots, x_{ik}, \quad i = 1, \ldots, n.$$

In the hybrid test, we regress:

$$\widehat{u}_i^2 \quad \text{on} \quad x_{i1}, x_{i2}, \ldots, x_{ik}, \widehat{y}_i^2, \quad i = 1, \ldots, n.$$

The only difference is that in the hybrid test we have an extra regressor, and so the R-squared will be no less for the hybrid test than that for the Breusch–Pagan test (we know that R-squared cannot decrease when we add a regressor).

For the special case of the White test, the argument is a bit more subtle. In the special case of the White test, we regress:

$$\widehat{u}_i^2 \quad \text{on} \quad \widehat{y}_i, \widehat{y}_i^2, \quad i = 1, \ldots, n$$

whereas in the hybrid test we regress:

$$\widehat{u}_i^2 \quad \text{on} \quad x_{i1}, x_{i2}, \ldots, x_{ik}, \widehat{y}_i^2, \quad i = 1, \ldots, n.$$

Thus, it comes down to whether it is $\widehat{y}_i$ or $x_{i1}, x_{i2}, \ldots, x_{ik}$ that reduce SSR more. Since the fitted values are a linear function of the regressors:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}$$

then when we run the regression:

$$\widehat{u}_i^2 \quad \text{on} \quad \widehat{y}_i, \widehat{y}_i^2, \quad i = 1, \ldots, n$$

we are putting a restriction on how the original explanatory variables appear on the right-hand side.

This means that the regression for the special case of the White test is a restricted form of the regression in the hybrid test. This means that the R-squared from the special case of the White test regression will be no greater than the R-squared from the hybrid regression.

(iii) No. The F statistic for the joint significance of the regressors in the second step regression depends on $R^2_{\widehat{u}^2}/(1 - R^2_{\widehat{u}^2})$, and it is true that this ratio increases as $R^2_{\widehat{u}^2}$ increases.

But, the F statistic also depends on the *df*, and the *df* are different among all three tests: the BP test, the special case of the White test, and the hybrid test. So we do not know which test will deliver the smallest p-value.

(iv) As discussed in part (ii), the OLS fitted values are a linear combination of the original regressors. Because those regressors appear in the hybrid test, adding the OLS fitted values is redundant; perfect multicollinearity would result.

Solution to Activity 8.7

(i) Recall that OLS works by choosing the estimates, $\widehat{\beta}_j$, such that the residuals are uncorrelated in the sample with each regressor (and the residuals have a zero sample average, too). Therefore, if we regress $\widehat{u}_i$ on the regressors, the R-squared will be exactly zero!

(ii) The F statistic for joint significance is about $F = \frac{.05256/4}{(1-.05256)/168} = 2.33$ (with 4 numerator *df* and 168 denominator *df*). Under:

$$H_0 : \delta_{prtystrA} = \delta_{democA} = \delta_{\log(expendA)} = \delta_{\log(expendB)}$$

and MLR.1 to MLR.4:

$$F \sim F_{4,\,168}.$$

The 5% critical value of $F_{4,\,168}$ is 2.37 and the 10% critical value of $F_{4,\,168}$ is 1.94, so there is evidence of heteroskedasticity at the 10%, but not quite at the 5% level (the p-value is $\approx .058$).

(iii) The F statistic for joint significance is about $F = \frac{.03173/2}{(1-.03173)/170} \approx 2.79$ (with 2 numerator *df* and 170 denominator *df*). Under $H_0 : \delta_{\widehat{y}} = \delta_{\widehat{y}^2}$ and MLR.1 to MLR.4:

$$F \sim F_{2,\,170}.$$

The 5% critical value of $F_{2,170}$ is 3 and the 10% critical value is 2.3, so there is evidence of heteroskedasticity at the 10%, but not quite at the 5% level.

In fact, the p-value is $\approx .065$. There is slightly less evidence of heteroskedasticity than provided by the BP test, but the conclusion is similar.

Solution to Activity 8.9

(i) Although the usual OLS t statistic on the interaction term is about 2.8, the heteroskedasticity-robust t statistic is just under 1.6. Therefore, using OLS, we must conclude the interaction term is only marginally significant (under MLR.1 to MLR.4).

But the coefficient is nontrivial; it implies a much more sensitive relationship between financial wealth and income for those eligible for a 401(k) plan.

(ii) Usual and robust standard errors are now much closer (which we can take as an indication that our heteroskedasticity specification works reasonably well and, thus, WLS dealt well with heteroskedasticity).

The robust t statistic is about 1.84, and so the interaction term is marginally significant (two-sided p-value is about .066) under MLR.1 to MLR.4.

(iii) The coefficient on $e401k$ literally gives the estimated difference in financial wealth at $inc = 0$, which obviously is not interesting. It is not surprising that it is not statistically different from zero; we obviously cannot hope to estimate the difference at $inc = 0$, nor do we care to.

(iv) When we replace $e401k \cdot inc$ with $e401k \cdot (inc - 30)$, the coefficient on $e401k$ is the estimated difference in $nettfa$ between those with and without 401(k) eligibility at roughly the average income, \$30,000. Naturally, we can estimate this much more precisely, and its magnitude (\$6,680) makes sense.

8.4 Overview of chapter

In this chapter we discussed the properties of ordinary least squares in the presence of heteroskedasticity. Since homoskedasticity played no role in showing that OLS is unbiased or consistent for the parameters in the regression equation, heteroskedasticity does not cause bias or inconsistency in the OLS estimators. However, the usual standard errors and test statistics are no longer valid. We illustrated how to compute heteroskedasticity-robust standard errors and t and F statistics. We also discussed that there is nothing wrong with the R-squared or adjusted R-squared as goodness-of-fit measures in the presence of heteroskedasticity.

We also covered two common ways to test for heteroskedasticity: the Breusch–Pagan test and the White test (including a special case of the White test). Breusch–Pagan involves regressing the squared OLS residuals on the regressors. The special case of the White test involves regressing the squared OLS residuals on the fitted and squared fitted values. A simple F test is asymptotically valid.

Finally, we discussed that OLS is no longer the best linear unbiased estimator in the presence of heteroskedasticity. This leads us to weighted least squares as a means of obtaining the BLUE estimator. To construct WLS, we needed to have the correct model of heteroskedasticity.

8.4.1 Key terms and concepts

- Heteroskedasticity
- Homoskedasticity
- Robust
- Wald statistic
- Breusch–Pagan
- White test
- Weighted least squares

8.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the definitions of homoskedasticity and heteroskedasticity in linear regression models
- explain the consequences of heteroskedasticity for OLS estimators, their standard errors, confidence intervals, t and F tests
- describe the principles of heteroskedasticity-robust inference for the OLS estimation
- discuss different approaches to detecting and testing for heteroskedasticity
- implement different approaches to detecting and testing for heteroskedasticity in practice
- explain what a weighted least squares estimator is and describe its properties.

8.5 Test your knowledge and understanding

8.5.1 Additional examination based questions

8.1E (a) A researcher investigating economies of scale in the production of electricity has the following data for the year 1955 for 115 plants: OUT, annual output, measured in trillion kilowatt hours, and $COST$, annual cost, measured in US $ million. She defines a new variable, $OUTSQ$, as the square of OUT, and regresses $COST$ on OUT and $OUTSQ$. The output is shown in Table 8.1.

OUT	3.43 (.47)
OUTSQ	.26 (.04)
constant	2.78 (.90)
R^2	.93

Table 8.1: The dependent variable is COST. Standard errors are in parentheses.

She plots the data and fitted line in a scatter diagram, shown in Figure 8.6.

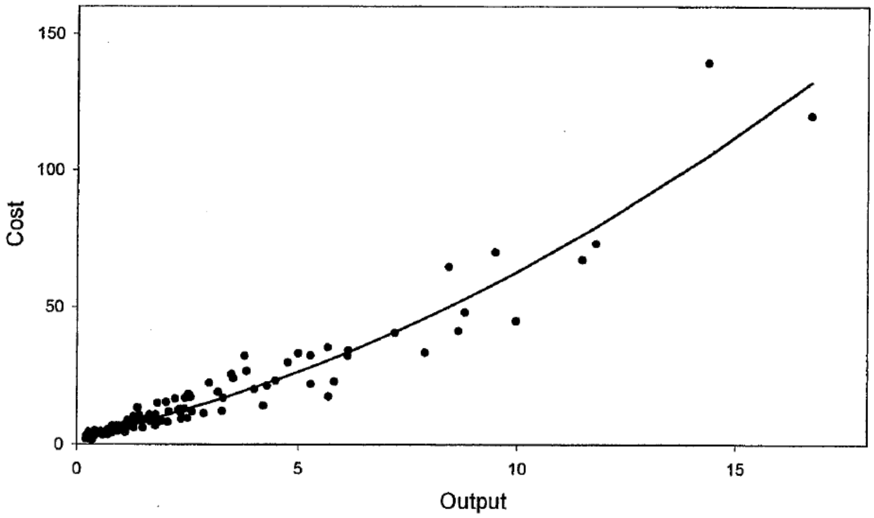


Figure 8.6: Scatter diagram of the data.

i. Explain why Figure 8.6 suggests a problem of heteroskedasticity.

ii. Describe the consequences of heteroskedasticity.

iii. Describe how you would perform the Breusch–Pagan test for this model if you had a copy of the dataset, stating any necessary assumptions, writing the testing equation(s) and explaining how the test statistic is constructed.

iv. Describe how you would perform the White test for this model if you had a copy of the dataset, stating any necessary assumptions, writing the testing equation(s) and explaining how the test statistic is constructed. The researcher performs a White test. The test statistic is 58.2. State the conclusion in this case.

(b) The researcher defines *LCOST* and *LOUT* as the natural logarithms of *COST* and *OUT* and regresses *LCOST* on *LOUT*. Breusch–Pagan and White tests for this specification do not reject the null hypothesis of homoskedasticity. Explain this finding, given the answer in part (a).

8.2E Consider the following result of the OLS estimation:

$$\widehat{wage} = \underset{(.21)}{7.10} - \underset{(.30)}{2.51}\,female \tag{8.5.1}$$

$$n = 526, \quad R^2 = .116$$

where *wage* is average hourly earnings in dollars and *female* is a binary variable that is equal to 1 if the person is female and 0 otherwise.

(a) In the model:

$$wage = \beta_0 + \beta_1 female + u$$

suppose that the variances are also different for female and non-female: that is:

$$Var(wage\,|\,female = 0) \neq Var(wage\,|\,female = 1)$$

(or equivalently $Var(u\,|\,female = 0) \neq Var(u\,|\,female = 1)$). Is the result in (1) still reliable (make sure you discuss the reliability of $\widehat{\beta}_0$, $\widehat{\beta}_1$ and their standard errors)? If not, how would you modify the result?

(b) In the model:

$$wage = \beta_0 + \beta_1 female + u$$

suppose:

$$Var(u\,|\,female) = \sigma^2(1 + 0.2 \cdot female)$$

for some unknown σ^2. Suggest an estimator different from the OLS that may have better properties.

8.3E We use a dataset on hospital finance for 299 hospitals in California in 2003. The data include many variables related to hospital finance and hospital utilisation.

(a) We estimate the following model by OLS:

$$\begin{array}{rccccc}
\widehat{patient\_revenue} = & -9.189\cdot 10^7 & + & 988.5 \; ER\_visits & + & 6890 \; surgeries \\
 & (1.399\cdot 10^7) & & (339.5) & & (1163) \\
 & [1.291\cdot 10^7] & & [665.47] & & [2769]
\end{array}$$

$$\begin{array}{cccc}
+ & 547600 \; beds & + & 5.918\cdot 10^7 \; occupancy \\
 & (43680) & & (2.126\cdot 10^7) \\
 & [85763] & & [1.836\cdot 10^7]
\end{array}$$

$$n = 299,\; R^2 = .7597,\; \bar{R}^2 = .7564$$

Here *patient_revenue* is the patient revenue (in \$), *ER_visits* is the number of visits to the hospital's emergency room, *surgeries* is the number of surgeries conducted in the hospital, *beds* is the number of available hospital beds, and *occupancy* is the average occupancy rate.

The usual OLS standard errors are in parentheses, (·), below the corresponding OLS estimate, and the heteroskedasticity-robust standard errors are in brackets, [·].

Do the regressors have the expected estimated signs? How do heteroskedasticity-robust standard errors compare to usual standard errors? Which of the regressors are statistically significant at the 5% level? Does it matter which standard errors are used?

(b) A researcher performs the Breusch–Pagan test for the model in (a). The BP test statistic is 34.52. State the conclusion in this case.

(c) The researcher changes the model and now considers *patient_revenue*, *ER_visits*, *surgeries* and *beds* in logarithms:

$$\begin{aligned}\log(\widehat{patient\_revenue}) = {} & \underset{\substack{(.3362)\\ [.3536]}}{8.9055} + \underset{\substack{(.0534)\\ [.0568]}}{.3841} \log(ER\_visits) + \underset{\substack{(.0304)\\ [.0323]}}{.2356} \log(surgeries) \\ & + \underset{\substack{(.0390)\\ [.0469]}}{.5962} \log(beds) + \underset{\substack{(.1096)\\ [.1178]}}{.6424}\ occupancy\end{aligned}$$

$$n = 299,\ R^2 = .8823,\ \bar{R}^2 = .8807.$$

How do heteroskedasticity-robust standard errors and usual standard errors compare now? Which of the regressors are statistically significant at the 1% level? Does it matter which standard errors are used?

(d) The researcher performs the Breusch–Pagan test for the model in (c). The BP test statistic is 0.5117. The researcher also performs the special case of the White test and finds the test statistic to be 8.118. State and discuss your conclusions.

8.5.2 Solutions to additional examination based questions

8.1E (a) i. Figure 8.6 shows that the spread of $COST$ is greater for greater values of $OUTPUT$ suggesting a problem of heteroskedasticity.

ii. If MLR.1–MLR.4 hold, then the OLS estimators remain unbiased and consistent. However, the usual standard errors and the usual t and F statistics are no longer valid.

iii. Let u denote the error in the regression of $COST$ on OUT and $OUTSQ$:

$$COST = \beta_0 + \beta_1 OUT + \beta_2 OUTSQ + u.$$

The Breusch–Pagan test assumes that $E[u^2 \mid OUT, OUTSQ]$ is a linear function of regressors:

$$E[u^2 \mid OUT, OUTSQ] = \delta_0 + \delta_1 OUT + \delta_2 OUTSQ.$$

The null hypothesis of homoskedasticity is:

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

(the alternative is the negation of the null). This null hypothesis is tested using the F test of the joint significance of OUT and $OUTSQ$. Under H_0 and MLR.1–MLR.4, $F \sim F_{2,112}$.

Here is how we would perform the Breusch–Pagan test if we had a copy of the dataset (throughout this testing procedure we maintain assumptions MLR.1–MLR.4):

- After estimating the equation $COST = \beta_0 + \beta_1 OUT + \beta_2 OUTSQ + u$ by OLS (estimates given) we would save the OLS residuals, $\widehat{u}_i$, and compute the squared residuals, $\widehat{u}_i^2$.
- We then would regress $\widehat{u}_i^2$ on OUT and $OUTSQ$ and compute the usual F test of joint significance of OUT and $OUTSQ$.
- If the p-value in the previous test is sufficiently small, we would reject the null of homoskedasticity and conclude Assumption MLR.5 fails.

iv. The White test in our case assumes that:

$$E[u^2 \mid OUT, OUTSQ] = \delta_0 + \delta_1 OUT + \delta_2 OUTSQ + \delta_3 OUT^3 + \delta_4 OUT^4$$

(OUT^4 is the square of the regressor $OUTSQ$ and OUT^3 is the interaction of two regressors; the square of OUT just gives $OUTSQ$). The null hypothesis of homoskedasticity in our case is:

$$H_0 : \delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$$

(the alternative is the negation of the null). This null hypothesis is tested using the F test of the joint significance of OUT, $OUTSQ$, OUT^3, OUT^4. Under H_0 and MLR.1–MLR.4, $F \sim F_{4,110}$.

Here is how we would perform the White test if we had a copy of the dataset (throughout this testing procedure we maintain assumptions MLR.1–MLR.4):

- After estimating the equation $COST = \beta_0 + \beta_1 OUT + \beta_2 OUTSQ + u$ by OLS (estimates given) we would save the OLS residuals, $\widehat{u}_i$, and compute the squared residuals, $\widehat{u}_i^2$.
- We then would regress $\widehat{u}_i^2$ on OUT, $OUTSQ$, OUT^3, OUT^4 and compute the usual F statistic of joint significance of these four explanatory variables.
- If the p-value of the previous test is sufficiently small, we would reject the null of homoskedasticity and conclude Assumption MLR.5 fails.

We are given that the realisation of the F statistic is 58.2. The approximate 1% critical value for $F_{4,110}$ is 4.01. Since 58.2 > 4.01, we reject the null of homoskedasticity at the 1% significance level.

(b) It is known that models with the logarithm of a dependent variable such as cost, wage, income, GDP, etc. often suffer less from heteroskedasticity. In our case this finding is consistent with the homoskedastic error u entering through the following multiplicative form:

$$COST = \exp^{\beta_0 + u} OUT^{\beta_1}$$

(takes logs to obtain $LCOST = \beta_0 + \beta_1 LOUT + u$).

8.2E (a) If MLR.1–MLR.4 hold, OLS estimators $\widehat{\beta}_0$, $\widehat{\beta}_1$ are still unbiased and consistent. However, their usual standard errors are no longer valid due to heteroskedasticity (violation of MLR.5). We would have to compute heteroskedasticity-robust standard errors.

(b) WLS estimators are unbiased and consistent under MLR.1–MLR.4 and have a smaller variance than OLS estimators. We can obtain them as OLS estimators in the model:

$$\frac{wage}{\sqrt{1+0.2\cdot female}} = \beta_0\frac{1}{\sqrt{1+0.2\cdot female}}+\beta_1\frac{female}{\sqrt{1+0.2\cdot female}}+\frac{u}{\sqrt{1+0.2\cdot female}}.$$

8.3E (a) Yes, regressors have the expected signs. The demand for a hospital's services (*ER_visits*, *surgeries*, *occupancy*) has a positive effect on patient revenue as well as the hospital capacity (*beds*). For *ER_visits*, *surgeries* and *beds* the heteroskedasticity-robust standard errors are larger than the usual ones whereas for *occupancy* the heteroskedasticity-robust standard error is lower. When using the usual standard errors, all the regressors are statistically significant at the 5% level. When using heteroskedasticity-robust standard errors, *ER_visits* becomes statistically insignificant at the 5% level (and even at the 1% level).

(b) The Breusch–Pagan test statistic is the F statistic in the auxiliary model. Under MLR.1 to MLR.4 and under the null of homoskedasticity, $F \sim F_{4,294}$. The 1% critical value for the $F_{4,294}$ distribution is 3.32. Since $34.52 > 3.32$, we reject the null of homoskedasticity at the 1% level.

(c) All heteroskedasticity-robust standard errors are larger than the usual ones. All the regressors are statistically significant at the 1% level and this conclusion holds when using either usual or heteroskedasticity-robust standard errors.

(d) The result of the BP test does not reject the null of homoskedasticity at the 5% significance level ($.5117 < 2.37$, where 2.37 is the 5% critical value for the $F_{4,294}$). In fact, the p-value associated with it is .7272. The special White test, however, rejects the null of homoskedasticity since $8.118 > 3$, where 3 is the 5% critical value for the $F_{2,296}$ distribution. These two findings suggest that there is still heteroskedasticity in the model in (c) but it is captured by quadratic and interaction terms of the regressors rather than in linear terms.

Chapter 9

Instrumental variable estimation and two-stage least squares

9.1 Introduction

9.1.1 Aims of the chapter

In this chapter we examine linear regression models in the presence of endogeneity. The situation of endogeneity occurs when at least one of the explanatory variables is correlated with the regression error. We outline some scenarios where endogeneity can arise. Since the OLS estimators in this case are inconsistent (as well as biased), we propose a different estimation approach – the so-called instrumental variables (IV) method. We first describe it in detail in the simple linear regression case and discuss its properties and performance. We then discuss the requirements on instrumental variables and the general ideas behind using them in a multiple regression model. We also introduce the situation of overidentification, which happens when the number of available instruments exceeds the number of endogenous regressors, and give a general approach to finding the estimator based on instrumental variables. This is the so-called two-stage least squares (2SLS) estimation approach. Finally, we describe a test for endogeneity of an explanatory variable that shows whether IV-based methods are even necessary.

9.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand what is meant by the presence of endogeneity in linear regression models
- explain the consequences of endogeneity for OLS estimation
- explain the concept of instrumental variables in both simple and multiple regression models
- describe the principles underlying the use of instrumental variables in both simple and multiple regression models
- understand the properties of instrumental variables estimators and major challenges associated with the IV approach

- explain what is meant by overidentification and describe principles underlying the use of two-stage least squares estimation
- describe a test for endogeneity of an explanatory variable
- discuss whether endogeneity may be present in a given empirical model
- discuss choices and properties of instrumental variables in practice and conduct IV/2SLS estimation using statistical software.

9.1.3 Essential reading

Readings: Wooldridge, Sections 15.1, 15.3a, 15.3d, 15.5a.

9.1.4 Synopsis of this chapter

In this chapter we first give a brief overview of the concept of endogeneity, highlight the wide range of settings that give rise to endogeneity, provide intuition behind the undesirable consequences of endogeneity on the OLS estimator, and provide possible solutions to deal with endogeneity.

After motivating the endogeneity problem with the help of omitted variables, Section 15.1 in Wooldridge introduces the definition of an instrumental variable (IV) in the simple linear regression model and discusses the key idea of the IV method. In the simple regression model an instrument for the endogenous regressor needs to satisfy the instrument exogeneity (validity) and instrument relevance requirements and you learn whether (and how) these requirements can be tested. With the help of empirical examples we discuss these conditions. You will learn how these conditions can be used to identify and derive the IV estimator in the simple linear regression model. You will analyse its properties (consistency) and discuss the variance of the IV estimator which highlights the consequences of having poor instruments.

In Section 15.2 in Wooldridge, IV estimation is extended to the multiple regression model. This is quite an important extension to consider because sometimes a potential instrument is exogenous only when other factors are controlled for. In addition to the instrument exogeneity and relevance requirements, we will add here a third condition on our instrument: an exclusion restriction.

In Sections 15.3a and 15.3d in Wooldridge we consider the setting where we have more than one instrument for our endogenous regressor(s). The setting where we have more instruments than the number of endogenous regressors is called overidentification (contrast with exact (or just) identification where you have exactly as many instruments as endogenous regressor(s)). You learn that the two-stage least squares (2SLS) estimator will give an estimator which is better than the IV estimator based on a single instrument (in the sense of having a smaller variance).

Finally, in Section 15.5a you learn how to test for the endogeneity of an explanatory variable. As OLS is more efficient than IV (2SLS) when there is no endogeneity this is very useful.

9.2 Content of chapter

9.2.1 Overview of endogeneity

A key assumption we have used when studying the linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

was MLR.4: $E(u \,|\, x_1, \ldots, x_k) = 0$. This assumption, which ensured the unbiasedness of OLS, implied:

$$Cov(x_j, u) = 0, \quad \text{for } j = 1, \ldots, k$$

i.e. each regressor x_j and the error u should be uncorrelated. This implies that as soon as one of the explanatory variables x_j is correlated with u the OLS estimators of all β_js will be biased and inconsistent – this is undesirable.

Why is this happening? For intuition, consider a simple example:

$$\log(wage) = \beta_0 + \beta_1 educ + u.$$

- We know that unobserved u contains ability as there is no question ability affects wages.
- At the same time, we know people with higher ability tend to choose more education and, thus, *educ* and u are correlated.
- It turns out that because *educ* and u are correlated, the OLS approach cannot correctly separate the effect of *educ* on y from the effect of the error u on y (sometimes it is said that in this case OLS cannot identify the effect of *educ* on y).
- This will lead to bias and inconsistency in the OLS estimators of both β_0 and β_1.

By **endogeneity** we refer to any correlation between regressors and the error term – it is a key concept in econometrics and econometric applications. We will use the following terminology:

- Regressor x_j with $Cov(x_j, u) \neq 0$ is called an endogenous regressor.
- Regressor x_j with $Cov(x_j, u) = 0$ is called an exogenous regressor.

There may be various statistical or economic reasons why we might expect that the **errors and regressors are correlated**. While we focus our attention in this chapter on the first, we will discuss the latter three later in the course. In particular, they are:

- omitted variables (Chapter 15 in Wooldridge)
- measurement errors in the regressors (Section 9.4 in Wooldridge)
- simultaneity (Chapter 16 in Wooldridge)
- lagged dependent variables in the presence of autocorrelation in the error term (Chapter 12 in Wooldridge).

OLS properties under endogeneity

We shouldn't be surprised that our OLS estimators are bad (biased and inconsistent) if there is correlation between the errors and regressors. Indeed, in the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i$$

the first-order conditions (FOCs) that define our OLS estimator are:

$$\sum_{i=1}^{n} x_{ij}\widehat{u}_i = 0, \quad \text{for all } j = 0, \ldots, k.$$

- We now need to recognise that these FOCs only make sense if:

  $$E(x_{ij}u_i) = 0 \quad \text{for all } j = 0, \ldots, k$$

  or, equivalently:

  $$Cov(x_{ij}, u_i) = 0 \quad \text{for all } j = 0, \ldots, k.$$

- *Intuition*: $\frac{1}{n}\sum_{i=1}^{n} x_{ij}\widehat{u}_i = 0$ is the sample analogue of the requirement that $E(x_{ij}u_i) = 0$, and $Cov(x_{ij}u_i) = E(x_{ij}u_i)$ because $E(u_i) = 0$.

Solutions to endogeneity

If we think one or more explanatory variables in a model is endogenous, we have basically two choices.

1. Include good controls in the hope that the endogenous explanatory variable becomes exogenous.
   - Adding more controls allows us to control for confounders – needed for causal interpretation.
   - Adding additional lags (explanatory variables) in dynamic time series models may help remove any remaining dependence in the errors.
2. Find one or more **instrumental variables** for the endogenous explanatory variable.
   - The idea of the estimator is to replace the 'bad' FOCs of OLS with conditions that are reasonable.
   - By imposing that instruments **are** uncorrelated with the error $E(z_i u_i) = 0$ (validity, instrument exogeneity), we **can** use their sample analogues to define our new estimator:

     $$\sum_{i=1}^{n} z_i\widehat{u}_i = 0.$$

   - Like OLS, our IV estimator is a 'method of moments estimator'.

> **Activity 9.1** Take the MCQs related to this section on the VLE to test your understanding.

9.2.2 Motivation: Omitted variables in a simple regression model

Read: Wooldridge, Section 15.1 introduction (up until discussion of identification).

It is important to be able to explain (in empirical settings) why omitted variables give rise to the endogeneity problem whereby included explanatory variables are correlated with the regression error (which includes relevant variables that are correlated with the included regressor). This is directly related to our omitted variable bias (OVB) discussion and you should revisit it if required.

In the presence of endogeneity we need additional information to enable consistent estimation of our parameters, say in the simple linear regression model:

$$y = \beta_0 + \beta_1 x + u, \quad Cov(x, u) \neq 0.$$

The key idea of the instrumental variables method is that we need to find a new variable (say, z), called an **instrument**, that satisfies the following two conditions.

1. z is uncorrelated with u, i.e. we have:

$$Cov(z, u) = 0.$$

   This requirement is often referred to as instrument exogeneity or instrument validity.

2. z is correlated with x, i.e. we have:

$$Cov(z, x) \neq 0.$$

   This requirement is often referred to as instrument relevance.

In empirical settings you need to be able to argue intuitively whether it is reasonable that these requirements on your instrument are satisfied. The textbook and various activities will help you with the required reasoning. In addition, you need to understand whether (and how) these requirements can be tested. We learn that in this setting we cannot test the instrument exogeneity, but we can test the instrument relevance. The latter can easily be done by estimating a simple regression:

$$x = \pi_0 + \pi_1 z + v$$

and testing the hypothesis $H_0 : \pi_1 = 0$ against $H_1 : \pi_1 \neq 0$. For instrument relevance, we should be able to reject H_0. We must therefore find a convincing economic argument as to why the instrument exogeneity is not violated.

We can explain the need for these two requirements intuitively as follows. Let us think of the endogenous regressor x as having both 'good' and 'bad' variation.

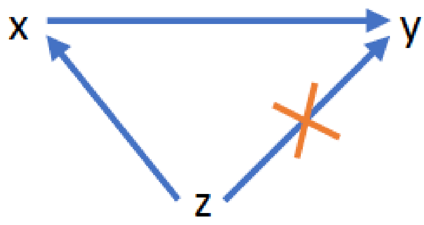
- 'Good' variation is **not** correlated with u.
- 'Bad' variation is correlated with u.

The instrumental variable z is a variable that explains the variation in x, but doesn't explain y, that is:

- z only explains the 'good' variation in x.

The instrument z can then be used to extract the 'good' variation from x and the instrumental variable method effectively replaces x with only that component.

The properties of an instrumental variable can be expressed using the following diagram, which we will use later to explain intuitively why the instrumental variable works in more detail. The diagram contains the dependent variable, $\mathbf{y}$; the endogenous regressor, $\mathbf{x}$, and the instrument variable, $\mathbf{z}$.



As the diagram shows: $\mathbf{x}$ affects $\mathbf{y}$ indicated by the directed arrow in accordance with our regression model. The instrument variable $\mathbf{z}$ is shown to affect the dependent variable only through its effect on the endogenous regressor $\mathbf{x}$. The instrument is not allowed to affect y directly. This is a nice summary of the property of the instrumental variables that we want to have.

> **Activity 9.2** Take the MCQs related to this section on the VLE to test your understanding.

9.2.3 IV estimation of the simple regression model

Read: Wooldridge Section 15.1 remainder, 15.1a–15.1c.

We should know how the IV estimator can be obtained in the simple linear regression model:

$$y = \beta_0 + \beta_1 x + u. \qquad (*)$$

IV estimator β_1 and identification

The instrument exogeneity and instrument relevance assumption enable us to estimate β_1 consistently as they guarantee that β_1 is identified. This means, that we can write β_1

in terms of population moments that can be estimated using a sample of data:

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)}.$$

(Notice the importance of instrument relevance in ensuring that the denominator is non-zero.)

Observe that plugging in (*) allows us to write $Cov(z, y) = Cov(z, \beta_0 + \beta_1 x + u)$. Using the properties of covariances, this yields $Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u)$. Using the instrument validity $(Cov(z, u) = 0)$ and instrument exogeneity $(Cov(z, x))$ we obtain the result.

Given a random sample $\{(y_i, x_i, z_i) : i = 1, \dots, n\}$, we can can then use the sample covariances to estimate the population covariances, which gives us the following **IV estimator** of β_1:

$$\widehat{\beta}_{1,IV} = \frac{Sample\ Cov(z, y)}{Sample\ Cov(z, x)} \equiv \frac{\widehat{Cov}(z, y)}{\widehat{Cov}(z, x)} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}.$$

IV estimator, a method of moments estimator

We can equivalently obtain the estimators of β_0 and β_1 by taking two moment conditions:

$$0 = E(u) = E(y - \beta_0 - \beta_1 x)$$
$$0 = Cov(z, u) \equiv E(zu) = E(z(y - \beta_0 - \beta_1 x))$$

and solving their sample analogues:

$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_{0,IV} - \widehat{\beta}_{1,IV} x_i) = 0 \quad \Leftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i = 0$$
$$\sum_{i=1}^{n}(y_i - \widehat{\beta}_{0,IV} - \widehat{\beta}_{1,IV} x_i) z_i = 0 \quad \Leftrightarrow \quad \frac{1}{n}\sum_{i=1}^{n}\widehat{u}_i z_i = 0.$$

Similar to how we derived the OLS estimators, we can rewrite the first equation to obtain the IV estimator of β_0:

$$\widehat{\beta}_{0,IV} = \bar{y} - \widehat{\beta}_{1,IV}\bar{x}$$

which looks just like the OLS intercept estimator except that the slope estimator, $\widehat{\beta}_{1,IV}$, is now the IV estimator. To derive the IV estimator of β_1 we plug this expression into the second equation:

$$\sum_{i=1}^{n}(y_i - (\bar{y} - \widehat{\beta}_{1,IV}\bar{x}) - \widehat{\beta}_{1,IV} x_i) z_i = 0$$

which we can rewrite as $\sum_{i=1}^{n}((y_i - \bar{y}) - \widehat{\beta}_{1,IV}(x_i - \bar{x})) z_i = 0$. Using the fact that, for example:

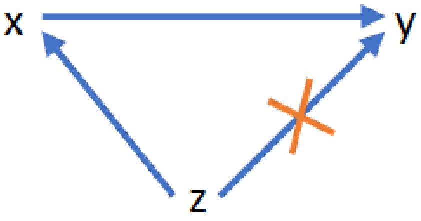
$$\sum_{i=1}^{n}(y_i - \bar{y})\bar{z} = \bar{z}\sum_{i=1}^{n}(y_i - \bar{y}) = 0$$

we then obtain:

$$\widehat{\beta}_{1,IV} = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \equiv \frac{\widehat{Cov}(z,y)}{\widehat{Cov}(z,x)}.$$

IV estimator in the simple regression model: Intuition

Recall the following picture which highlights what we need from our IV:



Based on the above graph, we can see the effect of instrument z on outcome y:

$$\text{Effect of } z \text{ on } y = (\text{Effect of } z \text{ on } x) \times (\text{Effect of } x \text{ on } y).$$

The thing we are interested in, the causal effect of x on y (the top arrow in the picture), is actually in the above equation. All we need to do is rearrange terms:

$$\text{Effect of } x \text{ on } y = \frac{\text{Effect of } z \text{ on } y}{\text{Effect of } z \text{ on } x}.$$

In the IV estimation, we calculate both of the effects on the right-hand side of this equation by OLS. To get the estimator of our causal effect we need to take their ratio:

$$\text{Effect of } z \text{ on } y = \frac{\widehat{Cov}(z,y)}{\widehat{Var}(z)}$$

$$\text{Effect of } z \text{ on } x = \frac{\widehat{Cov}(z,x)}{\widehat{Var}(z)}.$$

Properties of IV: Consistent, not unbiased, variance

- Given the instrument exogeneity and instrument validity, you should be able to prove the consistency of the IV estimator by application of the law of large numbers. Plugging in the true model $y_i = \beta_0 + \beta_1 x_i + u_i$ allows us to rewrite $\widehat{\beta}_{1,IV}$ as:

$$\begin{aligned}\widehat{\beta}_{1,IV} &= \beta_1 + \frac{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(u_i - \bar{u})}{\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \\ &= \beta_1 + \frac{\widehat{Cov}(z_i,u_i)}{\widehat{Cov}(z_i,x_i)}.\end{aligned}$$

By using properties of the plim operator, we then obtain:

$$\begin{aligned}\text{plim } \widehat{\beta}_{1,IV} &= \beta_1 + \frac{\text{plim}(\widehat{Cov}(z_i, u_i))}{\text{plim}(\widehat{Cov}(z_i, x_i))} \\ &= \beta_1 + \frac{Cov(z_i, u_i)}{Cov(z_i, x_i)} \\ &= \beta_1.\end{aligned}$$

- The IV estimator **cannot be unbiased** when x and u are correlated. Showing the finite-sample bias in the IV estimation is beyond the scope of the course, but can be quite substantial.

- You should be familiar with the expression of the variance of the IV estimator $\widehat{\beta}_{1,IV}$, which is approximately:

$$Var(\widehat{\beta}_{1,IV}) \approx \frac{\sigma_u^2}{n\sigma_x^2\rho_{x,z}^2}$$

where $\sigma_x^2 = Var(x)$ and $\rho_{x,z} = Corr(x,z)$. You should be able to use this formula to discuss (i) why we should simply stick to OLS if x is not endogenous ($\rho_{x,z}^2 < 1$ when $z \neq x$), and (ii) why IV estimates can have large standard errors, especially if x and z are only weakly correlated.

For the IV estimator, we can compute a heteroskedasticity-robust or nonrobust standard error and conduct large-sample inference using t statistics and confidence intervals. We impose no restrictions on the nature of x_i or z_i. For example, each variable could be binary (or, more generally, discrete), or just one of them.

In Section 15.1b we learn that a weak correlation between x and z can have more serious consequences than simply giving rise to large standard errors. Even under a mild violation of the instrument exogeneity assumption (so that z and u are only moderately correlated) the IV estimator may have large asymptotic bias. While important to be aware about when conducting empirical research, this will not be examinable. Section 15.1c briefly comments about the use of the R^2 after IV estimation.

Activity 9.3 Exercise 15.1 from Wooldridge.

Activity 9.4 Take the MCQs related to this section on the VLE to test your understanding.

Activity 9.5 This question investigates the effect of fertility on female labour supply and income and uses the dataset *same_sex_kids* (from the 1980 US census). We would like to answer the question: ‘How much does a woman’s hours worked and labour income fall when she has additional children?’

(i) Consider two simple regression models:

$$hoursm = \beta_0 + \beta_1 more\_kids + u$$

and:

$$income1m = \beta_0 + \beta_1 more\_kids + u$$

where *income1m* is mom's labour earnings in the year prior to the census, in 1995 dollars, *hoursm* is mom's average hours worked per week, and *more_kids* is the dummy variable equal to 1 if mom had more than 2 children, and is equal to 0 otherwise.

Explain why these OLS regressions are inappropriate for estimating the causal effect of fertility (*more_kids*) on labour supply (*hoursm*) or labour income (*income1m*).

(ii) The data set contains the variable *same_sex*, which is equal to 1 if the first two children are of the same sex (boy-boy or girl-girl) and equal to 0 otherwise. The OLS estimation of *more_kids* on *same_sex* (with heteroskedasticity-robust standard errors) gives:

$$\widehat{more\_kids} = \underset{(.0008)}{.3557} + \underset{(.0012)}{.0548}\, same\_sex$$

$$n = 655{,}169,\ R^2 = .0032.$$

Are couples whose first two children are of the same sex more likely to have a third child? Is the effect large? Is it statistically significant?

(iii) Explain why *same_sex* is a valid instrument for the instrumental variable regression of *hoursm* on *more_kids* and also for the instrumental variable regression of *income1m* on *more_kids*.

(iv) The estimation of the regression of *hoursm* on *more_kids* and of the regression of *income1m* on *more_kids* using *same_sex* as an instrument gives the following results (with heteroskedasticity-robust standard errors):

$$\widehat{hoursm} = \underset{(.3262)}{20.2929} - \underset{(.84486)}{3.5282}\, more\_kids$$

$$n = 655{,}169,\ R^2 = .0068$$

and:

$$\widehat{income1m} = \underset{(92.512)}{3823.213} - \underset{(240.784)}{766.881}\, more\_kids$$

$$n = 655{,}169,\ R^2 = .0077.$$

How large is the fertility effect on labour supply and how large is the fertility effect on mom's labour income?

Activity 9.6 Consider the model:

$$y = \beta_0 + \beta_1 x + u$$

with endogenous x and a *candidate* instrument z. Write down the plim of the OLS estimator of β_1 and the plim of the IV estimator of β_1 without assuming the required instrument exogeneity holds.

> Assume that $\sigma_u = \sigma_x$, so that the population standard deviation in the error term is the same as it is in x. Suppose that the instrumental variable, z, is slightly correlated with u: $Corr(z, u) = .1$. Suppose also that z and x have a somewhat stronger correlation: $Corr(z, x) = .2$.
>
> (i) What is the asymptotic bias in the IV estimator?
>
> (ii) How much correlation would have to exist between x and u before OLS has more asymptotic bias than the IV estimator?

9.2.4 IV estimation of the multiple regression model

Read: Wooldridge, Section 15.2.

In Section 15.2 in Wooldridge we learn how to extend the IV estimation procedure to the multiple regression model. The discussion focuses on the setting where there is a single endogenous explanatory variable in the model and one exogenous explanatory variable, but the results easily generalise to the presence of more exogenous explanatory variables.

In particular, we consider the linear model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where:

$$E(u) = 0, \quad E(x_1 u) = 0, \quad E(x_2 u) \neq 0$$

that is, when x_1 is exogenous but x_2 is endogenous.

It is clearly wrong to use OLS to estimate this model because OLS imposes (FOCs):

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} = 0$$

where:

$$\widehat{u}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \widehat{\beta}_2 x_{i2}$$

which are sample analogues of $E(u_i) = 0$, $E(u_i x_{i1}) = 0$ and $E(u_i x_{i2}) = 0$, respectively. Since $E(u_i x_{i2}) \neq 0$ due to the fact that x_2 is endogenous, we should not impose $\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} = 0$ when defining our estimator. In fact, if we estimate this model by OLS, **all of the estimators $\widehat{\boldsymbol{\beta}}_0$, $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2$ will be biased and inconsistent**, not just the estimator $\widehat{\beta}_2$ which corresponds to the endogenous regressor.

As only one of the conditions we used for the OLS estimator is incorrect here, it seems natural that only one instrument, which we shall call z, is required to deal with the endogeneity of x_2. Assuming **instrument exogeneity**, so that $Cov(z, u) = 0$, which in light of $E(u) = 0$ equivalently satisfies $E(zu) = 0$, we can use its sample analogue to replace the unreasonable requirement $\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i2} = 0$ with the reasonable requirement that $\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i z_i = 0$. The IV estimator imposes:

$$\frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i x_{i1} = 0, \quad \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i z_i = 0$$

where:

$$\widehat{u}_i = y_i - \widehat{\beta}_{0,IV} - \widehat{\beta}_{1,IV}x_{i1} - \widehat{\beta}_{2,IV}x_{i2}.$$

It is important to point out here that in order to solve for $(\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2)$ we need at least three conditions. This implies, for instance, that x_1 cannot be our instrument for x_2 as in this case the third moment condition in the IV estimation would be identical to the second moment condition, and, thus, we would effectively have the system of only two moment conditions to estimate three parameters and this system would have an infinity of solutions! This is an additional assumption, called the **exclusion restriction**.

We should recall that another assumption instruments need to satisfy is the **relevance** requirement. Indeed, we need the instrument z to be correlated with x_2, but the precise sense in which these two variables must be correlated is complicated by the presence of x_1 in the regression equation. Formulated somewhat informally, we require:

$$Partial\ Corr(z, x_2) \neq 0.$$

After partialling out x_1, we need z to be correlated with x_2. (The partial correlation effectively removes or nets out the effect of x_1.) More formally, to verify the relevance condition we consider the, so-called **reduced form for the endogenous variable** x_2:

$$x_2 = \pi_0 + \pi_1 x_1 + \pi_2 z + v$$

where $E(v) = 0$, $E(x_1 v) = 0$ and $E(zv) = 0$, and test whether $\pi_2 \neq 0$. The reduced form expresses the endogenous variable as a linear function of all exogenous variables and an error term. We can estimate the reduced form by OLS and use a t test to see if we can reject that the coefficient on z is equal to zero (usually making it robust to heteroskedasticity).

To summarise, the requirements on our instruments z for the model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where:

$$E(u) = 0, \quad E(x_1 u) = 0, \quad E(x_2 u) \neq 0$$

are the following.

1. **Instrument exogeneity (validity)**: Instrument z must satisfy $Cov(z, u) = 0$.
2. **Instrument relevance**: $Partial\ Corr(z, x_2) \neq 0$. After controlling for x_1 (same as partialling out), there is correlation between z and x_2. π_2 in the reduced form is non-zero.
3. **Exclusion restriction**. We cannot use the exogenous regressor that appears in the regression model as the IV. The IV should be another exogenous variable.

Together these three requirements give us the so-called **exact identification**.

They give us **identification** because we can **uniquely** define our true parameters in terms of the three population moment conditions:

$$E(u_i) = E(u_i x_{i1}) = E(u_i z_i) = 0.$$

Our IV estimator uses the sample analogues of the moment conditions. This identification is **exact** because we do not have any extra information (an extra moment condition) we can afford to discard. If we consider any two of the three moment conditions, we will no longer be able to identify parameters or, equivalently, uniquely recover them.

Activity 9.7 Exercise 15.8 from Wooldridge (you may skip 15.8d).

9.2.5 Two-stage least squares

Read: Wooldridge, Sections 15.3a–15.3d.

In this section we discuss how to use the instrumental variable estimator in settings where we have more instruments available than the number of endogenous regressors in the model. In this setting we have **overidentification**. The situation of having the same number of instruments as the number of endogenous regressors, considered thus far, is called **exact (or just)** identification.

Consider, for instance, the regression model with one endogenous and one exogenous regressor:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

where:

$$E(u) = 0, \quad E(x_1 u) = 0, \quad E(x_2 u) \neq 0$$

that is, where regressor x_1 is exogenous and regressor x_2 is endogenous.

Suppose now that we have *two* variables – let's call them z_1 and z_2 – that are different from x_1 (that is, satisfy the **exclusion restriction**) and satisfy the **exogeneity condition**:

$$E(z_1 u) = 0, \quad E(z_2 u) = 0.$$

If each of z_1 and z_2 is partially correlated with x_2, then we can take either of them (say, z_1) and implement the IV estimation we discussed previously. But this approach would discard additional data z_2 and the information from the exogeneity condition that $E(z_2 u) = 0$!

The 2SLS (**Two Stage Least Squares**) approach is a clever method that will use **both IVs** z_1 and z_2 and combine them in an optimal way to estimate β_1. It will give an estimator which is better than the IV estimator based on a single instrument (in the sense of having a smaller variance).

To gain intuition, let us consider a simple model with *one regressor x, which is endogenous*:

$$y = \beta_0 + \beta_1 x + u$$

where:

$$E(u) = 0, \quad E(xu) \neq 0$$

and assume we have *two instruments* for x: z_1 and z_2 that satisfy our usual requirements:

- $Cov(z_j, u) = E(z_j u) = 0$, for $j = 1, 2$ exogeneity (validity)
- $Cov(z_j, x) \neq 0$, for $j = 1, 2$ relevance.

In this overidentified setting, we should realise that the following IV estimators of the slope parameter β_1:

$$\widehat{\beta}_{1,IV}^{(1)} = \frac{Sample\ Cov(z_1, y)}{Sample\ Cov(z_1, x)} \quad \text{and} \quad \widehat{\beta}_{1,IV}^{(2)} = \frac{Sample\ Cov(z_2, y)}{Sample\ Cov(z_2, x)}$$

are both **consistent**. Both are the usual IV estimators in the simple regression model that select either z_1 ($\widehat{\beta}_{1,IV}^{(1)}$) or z_2 ($\widehat{\beta}_{2,IV}^{(1)}$) as instrument for x. **Question:** Which one should we use?

Since both are consistent, we need to look at the precision of IV estimators. Based on the expression for the variance, we should choose the estimator that uses the instrument that has the higher correlation with x. We have:

$$Var(\widehat{\beta}_{1,IV}^{(1)} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_1) = \frac{\sigma_u^2}{n\sigma_x^2 Corr(x, z_1)^2}$$

$$Var(\widehat{\beta}_{1,IV}^{(2)} \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_n, \mathbf{z}_2) = \frac{\sigma_u^2}{n\sigma_x^2 Corr(x, z_2)^2}.$$

Important: This approach discards additional information, which is not a good idea! We should use both! This is what the 2SLS approach will do! The basic idea is that since each of z_1 and z_2 is uncorrelated with u, any linear combination is also uncorrelated with u and, therefore, is a valid IV.

Optimality: To find the best IV, here, the 2SLS method will choose the linear combination that has the highest correlation with the endogenous x (and thereby will give the most precise IV estimator). Let us denote such a combination as z^{opt}. 2SLS would then use:

$$\widehat{\beta}_{1,IV}^{(opt)} = \frac{Sample\ Cov(z^{opt}, y)}{Sample\ Cov(z^{opt}, x)}.$$

The 2SLS estimator in this model will be obtained in **two stages**.

- **First stage:** Run OLS for:

  $$x = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v \qquad (*)$$

  and get the fitted values $\widehat{x} = \widehat{\pi}_0 + \widehat{\pi}_1 z_1 + \widehat{\pi}_2 z_2$.

  Equation $(*)$ is often called the **first stage** equation or, as we previously referred to analogous equations, the **reduced form** for x. It can be estimated by OLS.

  The purpose of this stage is to predict the endogenous x by a linear function of variables that are uncorrelated with u, to obtain the best instrument (i.e. which gives the highest correlation with the endogenous variable)!

  We recall that the first stage equation also allows us to formally give the **relevance condition** on the vector (z_1, z_2) of IVs in our model. The relevance condition on x being partially correlated with the vector (z_1, z_2) can be written as $\pi_1 \neq 0$ or $\pi_2 \neq 0$ which we can test using an F statistic, where $H_0 : \pi_1 = 0,\ \pi_2 = 0$ against $H_1 : \pi_1 \neq 0$ or $\pi_2 \neq 0$. If we reject this H_0, then this indicates that the relevance condition holds.

- **Second stage**: Run the OLS regression of y on $\widehat{x}$ obtained in the first stage:

  $$y = \beta_0 + \beta_1 \widehat{x} + e. \tag{**}$$

  We can estimate (**) by OLS, as we have replaced at this stage the endogenous x by its prediction from the first stage (exogenous!).

  The estimators $\widehat{\beta}_{0,2SLS}$ and $\widehat{\beta}_{1,2SLS}$ obtained in the second stage are the 2SLS estimators of β_0 and β_1, respectively.

To summarise the results of 2SLS for the multiple regression model in the presence of multiple endogenous explanatory variables, let us consider the following multiple regression setting with two endogenous variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u$$

where:

$$E(u) = 0, \quad E(x_1 u) = 0, \quad E(x_2 u) \neq 0, \quad E(x_3 u) \neq 0$$

that is, where regressor x_1 is exogenous and x_2 and x_3 are endogenous. We assume that we have *three* variables – let's call them z_1, z_2, z_3 – that are different from x_1 (that is, satisfy the **exclusion restriction**) and that satisfy the exogeneity and relevance conditions. Again, this is the setting of overidentification. The 2SLS estimator in this model will be obtained in **two stages**.

- **First stage:** Estimate by OLS the reduced form equations for both endogenous variables:

  $$x_2 = \pi_{2,0} + \pi_{2,1} x_1 + \pi_{2,2} z_1 + \pi_{2,3} z_2 + \pi_{2,4} z_3 + v_2$$

  reduced form for x_2 and:

  $$x_3 = \pi_{3,0} + \pi_{3,1} x_1 + \pi_{3,2} z_1 + \pi_{3,3} z_2 + \pi_{3,4} z_3 + v_2$$

  reduced form for x_3 and compute the fitted values $\widehat{x}_2$ and $\widehat{x}_3$.

  Each reduced form equation has all the instruments (z_1, z_2, z_3) and all the exogenous regressors (in our example it is just x_1) on the right-hand side. Including all the exogenous regressors on the right-hand side is very important. If this is not done, then you are not doing 2SLS and you are not getting consistent estimators.

- **Second stage**: Run the OLS regression of y on x_1, $\widehat{x_2}$, and $\widehat{x_3}$ to obtain $\widehat{\beta}_{0,2SLS}$, $\widehat{\beta}_{1,2SLS}$ and $\widehat{\beta}_{2,2SLS}$.

Practical advice

(a) Do not do two stages on your own but let the software do it. This is because the standard errors would be wrong if you tried to do 2SLS on your own due to the second stage using 'estimated' values that have their own estimation error. This error needs to be taken into account when calculating standard errors.

(b) All regressors that are not endogenous, need to be included in the first stage. If this is not done, then you are not doing 2SLS and you are not getting consistent estimators. This is yet another reason to let statistical software do the 2SLS estimation for you.

(c) Always report your first stage results (including R-squared). This helps to analyse/test the relevance condition. It can also help us determine whether there might be a **weak IV** problem (a weak instrument is an IV that does not explain very much of the variation in the endogenous regressor).

Final comment: When we have exact (just) identification we can also conduct 2SLS. In that case the 2SLS estimator will be identical to the IV estimator (intuition, when we have exact identification there is no need to look for the 'optimal' instruments).

> **Activity 9.8** Take the MCQs related to this section on the VLE to test your understanding.

9.2.6 Testing for endogeneity

Read: Wooldridge, Section 15.5a.

In this section we are shown how to test for the endogeneity of one or more of the explanatory variables. While Hausman (1987) suggested a test based on directly comparing the OLS and 2SLS estimates and testing whether the differences are statistically significant, the test discussed in Wooldridge applies a simpler regression-based test.

As shown in Section 15.5a this test involves running a regression where we include residuals of the reduced form for the potentially endogenous explanatory variables as additional regressors and tests whether their coefficients are statistically significant (using asymptotic t or F tests).

To illustrate, suppose we have the following model, where we have two suspected endogenous variables y_2 and y_3:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + u_1.$$

As there are two endogenous variables, we need at least two additional exogenous variables (instruments). Let us consider the exact identified setting, where we have two additional exogenous variables z_2 and z_3. The procedure is now as follows:

1. Estimate two reduced form equations (one for each endogenous variable):

$$y_2 = \pi_{20} + \pi_{21} z_1 + \pi_{22} z_2 \pi_{23} z_3 + v_2$$

(*reduced form for* y_2) and:

$$y_3 = \pi_{30} + \pi_{31} z_1 + \pi_{32} z_2 \pi_{33} z_3 + v_3$$

(*reduced form for* y_3) and save the residuals of these regressions, respectively $\widehat{v}_2$ and $\widehat{v}_3$.

2. Add the reduced form residuals to the model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 y_3 + \beta_3 z_1 + \delta_1 \widehat{v}_2 + \delta_2 \widehat{v}_3 + \mathit{error}$$

and estimate it by OLS. We should test $H_0 : \delta_1 = 0, \delta_2 = 0$ using an F statistic. If we reject H_0 at a small level of significance, we conclude that either y_2 or y_3 are endogenous.

Intuitively, when both $\delta_1 = 0$ and $\delta_2 = 0$ the result indicates that we can run the original regression without controlling for the potential endogeneity of y_2 and y_3 (we can exclude these controls). We simply should stick to OLS in this case.

Activity 9.9 (Based on Exercise 15.C2 from Wooldridge.)

The data in FERTIL2 include, for women in Botswana during 1988, information on number of children, years of education, age, and religious and economic status variables.

(i) We first estimate the model:

$$children = \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 age^2 + u$$

by OLS. The estimation result is:

$$\widehat{children} = \underset{(.2436)}{-4.1383} - \underset{(.0061)}{.0906}\, educ + \underset{(.0192)}{.3325}\, age - \underset{(.0004)}{.0026}\, age^2$$

$$n = 4{,}361,\ R^2 = 0.5687.$$

Interpret the estimates. In particular, holding age fixed, what is the estimated effect of another year of education on fertility? If 100 women receive another year of education, how many fewer children are they expected to have?

(ii) The variable *frsthalf* is a dummy variable equal to one if the woman was born during the first six months of the year. Assuming that *frsthalf* is uncorrelated with the error term from part (i), how would you show that *frsthalf* is a reasonable IV candidate for *educ*?

(iii) Estimating the model from part (i) by using *frsthalf* as an IV for *educ* gives:

$$\widehat{children} = \underset{(.5452)}{-3.3878} - \underset{(.0524)}{.1715}\, educ + \underset{(.0202)}{.3236}\, age - \underset{(.0004)}{.0027}\, age^2.$$

Compare the estimated effect of education with the OLS estimate from part (i).

(iv) When we add binary variables *electric*, *tv*, and *bicycle* to the equation and estimate it by OLS, we obtain:

$$\widehat{children} = \underset{(.2444)}{-4.3898} - \underset{(.0063)}{.0767}\, educ + \underset{(.0191)}{.3402}\, age - \underset{(.0004)}{.0027}\, age^2$$

$$- \underset{(.0744)}{.3028}\, electric - \underset{(.0827)}{.2531}\, tv + \underset{(.0490)}{.3179}\, bicycle$$

$$n = 4{,}361,\ R^2 = 0.5761.$$

The 2SLS (or IV) estimates are:

$$\widehat{children} = \underset{(.6394)}{-3.5913} - \underset{(.0644)}{.1640}\, educ + \underset{(.0213)}{.3282}\, age - \underset{(.0004)}{.0027}\, age^2$$

$$- \underset{(.1584)}{.1065}\, electric - \underset{(.2044)}{.0026}\, tv + \underset{(.0507)}{.3321}\, bicycle.$$

Compare the estimated coefficients on *educ*. Interpret the coefficient on *tv* and explain why television ownership has a negative effect on fertility.

Activity 9.10 (Based on Exercise 15.2 from Wooldridge.)

Suppose that you wish to estimate the effect of class attendance on student performance. A basic model is:

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + u$$

where *stndfnl* is the standardised outcome on a final examination, *atndrte* is the percentage of classes attended, *priGPA* is the prior college grade point average, and *ACT* is the ACT score.

(i) Let *dist* be the distance from the student's living quarters to the lecture hall. Assuming that *dist* and u are uncorrelated, what other assumptions must *dist* satisfy to be a reliable IV for *atndrte*?

(ii) Suppose we add the interaction term *priGPA* × *atndrte* giving the model:

$$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \times atndrte + u.$$

If *atndrte* is correlated with u, then, in general, so is *priGPA* × *atndrte*. If:

$$E(u \mid priGPA, ACT, dist) = 0$$

as happens when *priGPA*, *ACT*, and *dist* are all exogenous, what might be a good IV for *priGPA* × *atndrte*?

[*Hint*: If $E(u \mid priGPA, ACT, dist) = 0$, then any function of *priGPA* and *dist* is uncorrelated with u.]

(iii) Given your choice of IV for *priGPA* × *atndrte*, write down the first stage equation(s) you would estimate as part of 2SLS estimation. How many first stage equations do you have?

Are you in the situation of exact identification or in the situation of overidentification?

9.2.7 R Applications for instrumental variable estimation and two-stage least squares

Activity 9.11 Complete 'R Task – Activity 5' on VLE.

Here we show how to implement IV and 2SLS using the command `ivreg` in R. We also provide an implementation of the test for endogeneity.

9.3 Answers to activities

Solution to Activity 9.3

(Exercise 15.1 from Wooldridge.)

(i) It has been fairly well established that socioeconomic status affects student performance.

The error term u contains, among other things, family income, which has a positive effect on GPA and is also very likely to be correlated with PC ownership.

(ii) Families with higher incomes can afford to buy computers for their children. Therefore, family income certainly satisfies the relevance requirement for an instrumental variable: it is correlated with the endogenous explanatory variable (take $x = PC$ and $z = faminc$ in the relevance condition).

(iii) This is a natural experiment that affects whether or not some students own computers. Some students who buy computers when given the grant would not have without the grant. (Students who did not receive the grants might still own computers.)

Define a dummy variable, *grant*, equal to 1 if the student received a grant, and 0 otherwise. This is our candidate instrumental variable z (in our model $x = PC$ and $y = GPA$).

- Then, if $z = grant$ was randomly assigned, it is uncorrelated with u. In particular, it is uncorrelated with family income and other socioeconomic factors in u.
- Further, $z = grant$ should be correlated with $x = PC$: the probability of owning a PC should be significantly higher for a student receiving grants.
- Incidentally, if the university gave grant priority to low-income students, grant would be negatively correlated with u, and IV would be inconsistent.

Solution to Activity 9.5

(i) Both of the above OLS regression models suffer from a possible omitted variable bias.

- More educated women may both work more and be less likely to have an additional child than less educated women.
- More educated women may both earn more and be less likely to have an additional child than less educated women.

- This would imply that *more_kids* is positively correlated with the regression error in each of these regressions so that the OLS estimator for the parameter corresponding to *more_kids* is biased and inconsistent.

(ii) These estimates show that women in 1980 with same-sex children are estimated to be 5.48 percentage points more likely to have a third child (we give the percentage interpretation because the dependent variable is binary).

- Couples with two children of the same sex may want to have an additional child hoping for a different sex of the third child.
- The family's budget constraint is affected – children of the same sex can more easily share a room, clothes, and toys.
- The effect is statistically significant at any conventional significance level: t statistic $= 45.652$.

(iii) *same_sex* has no direct effect on either labour supply or labour earnings. At the same time, *same_sex* is random and so it is plausible that it is unrelated to any of the variables in the error term in the labour supply equation and labour earnings equation. Thus, the instrument is exogenous. There is no way to test whether *same_sex* is exogenous.

(iv) The estimates in the first equation suggest that women with more than 2 children on average work 3.5 fewer hours per week than women with 2 or fewer children.

The estimates in the second equation suggest that women with more than 2 children on average earn \$766 less per year than women with 2 or fewer children.

Solution to Activity 9.6

It is useful to recall:

$$\text{plim}(\widehat{\beta}_{1,OLS}) = \beta_1 + \frac{Cov(x,u)}{Var(x)} = \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot Corr(x,u)$$

$$\text{plim}(\widehat{\beta}_{1,IV}) = \beta_1 + \frac{Cov(z,u)}{Cov(z,x)} = \beta_1 + \frac{\sigma_u}{\sigma_x} \cdot \frac{Corr(z,u)}{Corr(z,x)}.$$

Taking into account that $\sigma_u = \sigma_x$, we can write:

$$\text{plim}(\widehat{\beta}_{1,OLS}) = \beta_1 + Corr(x,u)$$

$$\text{plim}(\widehat{\beta}_{1,IV}) = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)}.$$

The asymptotic bias is obtained by subtracting β_1 from the plim of these estimators.

(i) The asymptotic bias in the IV estimator is:

$$\text{plim}(\widehat{\beta}_{1,IV}) - \beta_1 = \frac{Corr(z,u)}{Corr(z,x)} = \frac{.1}{.2} = .5.$$

So the asymptotic bias is .5.

(ii) Because the asymptotic bias in the OLS estimator is:

$$\text{plim}(\widehat{\beta}_{1,OLS}) - \beta_1 = Corr(x, u)$$

we would have to have $Corr(x, u) > .5$ before the asymptotic bias in OLS exceeds that of IV.

This is a simple illustration of how a seemingly small correlation (.1 in this case) between the IV (z) and error (u) can still result in IV being more biased than OLS if the correlation between z and x is weak (.2).

Solution to Activity 9.7

(Exercise 15.8 from Wooldridge.)

(i) A few examples include family income and background variables, such as parents' education.

(ii) The population model is:

$$score = \beta_0 + \beta_1 girlhs + \beta_2 faminc + \beta_3 meduc + \beta_4 feduc + u$$

where the names of the variables are self-explanatory.

(iii) Parents who are supportive and motivated to have their daughters do well in school may also be more likely to enrol their daughters in a girls' high school. It seems likely that *girlhs* and u are correlated.

(iv) Let *numghs* be the number of girls' high schools within a 20-mile radius of a girl's home. To have reliable IV estimation, *numghs* must satisfy three requirements: it must be uncorrelated with u (instrument exogeneity), it must be partially correlated with *girlhs* (instrument relevance), and it must satisfy the exclusion restriction.

The exclusion restriction holds as *numghs* is different from exogenous regressors in the model. The second requirement (instrument relevance) probably holds and can be tested by estimating:

$$girlhs = \pi_0 + \pi_1 faminc + \pi_2 meduc + \pi_3 feduc + \pi_4 numghs + v$$

and testing $H_0 : \pi_4 = 0$ (testing *numghs* for statistical significance).

The first requirement (instrument exogeneity) is more problematic for the following reasons.

- Girls' high schools tend to locate in areas where there is a demand, which can reflect the seriousness with which people in the community view education.
- Some areas of a state have better students on average for reasons unrelated to family income and parents' education, and these reasons might be correlated with *numghs*.

One possibility to deal with these issues to achieve instrument exogeneity is to include in the regression models community-level variables that can control for differences across communities.

Solution to Activity 9.9

(Based on Exercise 15.C2 from Wooldridge.)

(i) Another year of education, holding age fixed, results in about .091 fewer children. In other words, for a group of 100 women, if each gets another year of education, they collectively are predicted to have about nine fewer children.

(ii) The way to show this is to estimate the reduced form for *educ*:

$$educ = \pi_0 + \pi_1 age + \pi_2 age^2 + \pi_3 frsthalf + v$$

and test $H_0 : \pi_3 = 0$ against $H_1 : \pi_3 \neq 0$ using a t test.

(iii) The estimated effect of education on fertility is now much larger. Naturally, the standard error for the IV estimate is also larger, about nine times larger. This produces a fairly wide 95% confidence interval for β_1.

(iv) Adding *electric*, *tv*, and *bicycle* to the model reduces the estimated effect of *educ* in both cases, but not by too much.

In the equation estimated by OLS, the coefficient on *tv* implies that, other factors fixed, four families that own a television will have about one fewer child than four families without a TV. A causal interpretation is that TV provides an alternative form of recreation.

Interestingly, the effect of TV ownership is practically and statistically insignificant in the equation estimated by IV. The coefficient on *electric* is also greatly reduced in magnitude in the IV estimation. The substantial drops in the magnitudes of these coefficients suggest that a linear model might not be the best functional form, which would not be surprising since children is a count variable.

Solution to Activity 9.10

(i) The variable *dist* satisfies the exclusion restriction as it is different from regressors in the model. Therefore, the only condition we need to check is that the variable *dist* is partially correlated with *atndrte* (instrument relevance). More precisely, in the reduced form equation for *atndrte*, which is:

$$atndrte = \pi_0 + \pi_1 priGPA + \pi_2 ACT + \pi_3 dist + v$$

we must have $\pi_3 \neq 0$. Given a sample of data, we can test $H_0 : \pi_3 = 0$ against $H_1 : \pi_3 \neq 0$ using a t test. If we reject H_0, then we consider it as evidence of instrument relevance.

(ii) This part says that in the new model the interaction term $priGPA \times atndrte$ is another endogenous regressor (in addition to *atndrte*) – indeed, it is the interaction of exogenous *priGPA* and endogenous *atndrte*, and it will generally be correlated with u.

Since we have *two* endogenous regressors (*atndrte* and $priGPA \times atndrte$) in the model, we will need at least *two* IVs.

- For *atndrte* we can use *dist* as an instrument. Since:

  $$E(u \,|\, priGPA, ACT, dist) = 0$$

  implies that u and *dist* are uncorrelated, at the very least *dist* satisfies instrument exogeneity.

- For $priGPA \times atndrte$ we can consider the instrument $priGPA \times dist$.

  Under the exogeneity assumption that $E(u \,|\, priGPA, ACT, dist) = 0$, any function of *priGPA*, *ACT*, and *dist* is uncorrelated with u (by the law of iterated expectation). In particular, we may argue that the interaction $priGPA \times dist$ is uncorrelated with u. Let us show this:

  $$E(priGPA \times dist \cdot u) = 0.$$

  First, you argue:

  $$\begin{aligned} &E(priGPA \times dist \cdot u \,|\, priGPA, ACT, dist) \\ &\quad = priGPA \times dist \cdot E(u \,|\, priGPA, ACT, dist)) = 0. \end{aligned}$$

  Next, you argue that by the law of iterated expectation:

  $$E(priGPA \times dist \cdot u) = E(E(priGPA \times dist \cdot u)) = E(0) = 0.$$

  If *dist* is partially correlated with *atndrte*, then $priGPA \times dist$ is also partially correlated with $priGPA \times atndrte$. So, we can estimate the equation:

  $$stndfnl = \beta_0 + \beta_1 atndrte + \beta_2 priGPA + \beta_3 ACT + \beta_4 priGPA \times atndrte + u$$

  by 2SLS using IVs *dist* and $priGPA \times dist$.

(iii) Since we now have **two** endogenous regressors, we have to estimate **two** first stage equations. Apart from including the two instruments in both, both equations include the included exogenous regressors in the original model *priGPA* and *ACT*.

First stage equation (or reduced form) for *atndrte*:

$$atndrte = \pi_{1,0} + \pi_{1,1} priGPA + \pi_{1,2} ACT + \pi_{1,3} dist + \pi_{1,4} priGPA \cdot dist + v_1.$$

First stage equation (or reduced form) for $priGPA \cdot atndrte$:

$$priGPA \cdot atndrte = \pi_{2,0} + \pi_{2,1} priGPA + \pi_{2,2} ACT + \pi_{2,3} dist + \pi_{2,4} priGPA \cdot dist + v_2.$$

Two endogenous regressors and **two** instruments ⇒ exact identification.

9.4 Overview of chapter

In this chapter we addressed a very important issue of endogeneity in empirical research. Endogeneity is a key concept in economics and econometrics as endogenous explanatory variables are present in many empirical models of interest. We discussed a classical approach to dealing with endogenous explanatory variables – the so-called *Instrumental Variables* method.

An instrumental variable must have the following properties: (1) it must be exogenous or valid, that is, uncorrelated with the error term of the original equation of interest; (2) it must be partially correlated with the endogenous explanatory variable; (3) it must satisfy the exclusion restriction. Finding a variable with these properties is usually challenging and is sometimes considered to be the most difficult part of the IV method.

We first discussed how to conduct the IV estimation when we have one endogenous regressor and one instrumental variable. We then discussed the method of two stage least squares (2SLS), which allows for more instrumental variables than we have endogenous explanatory variables and also allows for several endogenous explanatory variables. The 2SLS approach is used routinely in the empirical social sciences. When used properly, it can allow us to estimate *ceteris paribus* effects in the presence of endogenous explanatory variables.

As a part of the IV/2SLS method, it is important to always estimate the reduced form/first stage equation(s) and test whether the IVs are partially correlated with the endogenous explanatory variable(s).

Finally, we discussed that when we have valid instrumental variables, we can test whether an explanatory variable is endogenous and discussed details of how to implement such a test.

9.4.1 Key terms and concepts

- Endogenous explanatory variables
- Exogenous explanatory variables
- Instrument/instrumental variable
- Instrument exogeneity/instrument validity
- Instrument relevance
- Reduced form equation
- Exclusion restriction
- Instrumental variables (IV) estimator
- First stage equation
- Two Stage Least Squares (2SLS) estimator
- Exact identification
- Overidentification

9.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand what is meant by the presence of endogeneity in linear regression models
- explain the consequences of endogeneity for OLS estimation
- explain the concept of instrumental variables in both simple and multiple regression models
- describe the principles underlying the use of instrumental variables in both simple and multiple regression models
- understand the properties of instrumental variables estimators and major challenges associated with the IV approach
- explain what is meant by overidentification and describe principles underlying the use of two-stage least squares estimation
- describe a test for endogeneity of an explanatory variable
- discuss whether endogeneity may be present in a given empirical model
- discuss choices and properties of instrumental variables in practice and conduct IV/2SLS estimation using statistical software.

9.5 Test your knowledge and understanding

9.5.1 Additional examination based exercises

9.1E An article in the *Journal of Banking & Finance* studies the causal effect of board size on the performance of small- and medium-sized firms.

Suppose you are given a rich dataset of a large number of closely held corporations, including information on board size, several measures of firm performance, and a number of other firm characteristics that impact performance and are correlated with board size. Assume board size is binary (i.e. there are large board and small board companies).

The paper suggests using the number of children of the chief executive officer (CEO) of the firms as an instrument for board size. What is the first stage and what sign would you expect for the key coefficient in the first stage? What assumptions should a good instrument satisfy? Discuss whether this is a good instrument.

9.2E Let us consider some simulated data on the relation between number of police ($POLICE$) and crime ($CRIME$). Some of the data refer to election years ($ELECTION = 1$), the other data to non-election years ($ELECTION = 0$). We want to estimate the causal effect of the number of police on crime, β, using the model:

$$CRIME = \alpha + \beta POLICE + u.$$

The sample size is 300.

(a) We are worried that $POLICE$ is correlated with other factors in u. Discuss the properties of the OLS estimator $\widehat{\beta}$ in light of this problem. Prove any claim you make.

(b) It is argued that the election dummy $ELECTION$ could serve as an instrument. Discuss what conditions instruments should satisfy and whether they are reasonable in this case (contrast this with the use of a variable like 'Police Department Funding'). Can we test some of these requirements, and if so how? To help you answer this question, suppose you are told that the OLS estimation of $POLICE$ on $ELECTION$ gives the following results:

$$\widehat{POLICE} = \underset{(.1679)}{9.8475} + \underset{(.3358)}{1.0599}\, ELECTION.$$

(c) Derive the formula for our IV estimator of β in this case and discuss the properties of the estimator assuming $ELECTION$ is a valid instrument. Prove any claim you make.

(d) Interpret the following estimated IV results:

$$\widehat{CRIME} = \underset{(9.2951)}{18.2210} - \underset{(.9182)}{0.8670}\, POLICE.$$

(e) Say I had used another valid instrument instead of $ELECTION$. How should I judge which of the two IV estimators I should use? Is there a better estimator than either one of them?

(f) Discuss how you would proceed to test for endogeneity of $POLICE$.

9.3E The National Health Service (NHS) is trying to assess the effect of maternal smoking (S_i) during pregnancy on birthweight (BW_i) by estimating the model:

$$BW_i = \beta_0 + \beta_1 S_i + u_i.$$

Five years ago, the NHS mailed a brochure about the potential dangers of smoking during pregnancy to all the women in Bristol aged 20 to 35 whose birthdays were on even days of the year. Among women giving birth in Bristol, the NHS linked information on who received a brochure in the post ($R_i = 1$ indicates a woman who received the brochure and 0 otherwise) to their records on which of these women gave birth during the five-year period, and data on birthweight (let BW_i indicate the birthweight in grams) of their children for those who did. The NHS has no information on the smoking behaviour of mothers in Bristol.

(a) Suppose you run a regression of BW_i on R_i. Clearly explain what the interpretation of the coefficient on R_i is. What substantive question does this coefficient speak to?

(b) Which econometric method would you like to use to address the question the NHS wants to answer and why? Discuss the assumptions necessary for this method to identify a causal effect and whether they are likely to hold in this context. What prevents you from applying this method with the available data?

(c) Suppose the NHS mailed the same brochure to a similar set of women with even birthdays in Bath at the same time. Now the NHS in Bath is conducting a survey of women aged 25 to 40 asking them about their smoking behaviour during the past five years (resulting in a variable $S = 1$ if the woman smoked regularly during the past five years, 0 otherwise). The survey also asks whether the women had a birth during the past five years. However, the survey data are not linked to other NHS records and we don't know the birthweight of the children in this population.

i. What regression can you run with the data from Bath that is relevant to the relationship between smoking and birthweight? What is this regression in terms of the method you discussed in part (b)? What substantive question does this regression address and why is this interesting?

ii. How can you use the results you have from women in Bristol and from women in Bath to say something about the question the NHS is interested in? What do you need to assume for this to make sense?

9.4E An economist is interested in estimating the production function for widgets which is postulated to follow a Cobb–Douglas specification:

$$Y_i = \exp(\beta_0) L_i^{\beta_L} K_i^{\beta_K} \exp(u_i)$$

where Y_i is a measure of output for firm i, L_i is labour, K_i is capital stock and u_i is an unobserved term that captures technological or managerial efficiency and other external factors (for example, weather). The parameters to be estimated are $(\beta_0, \beta_L, \beta_K)$. Taking logs:

$$\ln Y_i = \beta_0 + \beta_L \ln L_i + \beta_K \ln K_i + u_i.$$

(a) Assume that you have a cross-section of independent firms and that more productive firms hire less workers (labour). Explain why OLS would not provide consistent estimates of $(\beta_0, \beta_L, \beta_K)$. Would it overestimate or underestimate β_L on average? Clearly explain your answer.

(b) Instead of applying OLS, the economist decides to use the average wage paid by firm i, W_i, as an instrument for the (log) quantity of labour employed by that firm, $\ln L_i$.

Describe in detail how you would estimate the parameters of the production function using Two Stage Least Squares (2SLS). What restrictions would be necessary for this researcher to successfully use this instrumental variable in the estimation of the parameters $(\beta_0, \beta_L, \beta_K)$ and what would you need to assume about capital stock?

(c) If average wages per firm do not vary much by firm (potentially because of unionisation or high mobility of the labour force), how would this affect the properties of the estimation procedure suggested in (b)? Explain your answer.

9.5.2 Solutions to additional examination based exercises

9.1E The first stage is a regression function of board size on the CEO family size (possibly also including other exogenous covariates). Clearly, as these are small- and medium-sized firms, the expectation is that the larger the family of the CEO, the larger the size of the board.

An instrumental variable should fulfil three requirements.

- The first requirement is the exogeneity of the instrument condition. It amounts to claiming that the size of the CEO's family is uncorrelated with the unobservables affecting the firm's performance. It amounts to claiming the only reason why the size of the CEO's family might impact firm performance is through its effect on the size of the board of directors. It is not hard to come up with stories why this might not be the case (for example, a CEO with a large family might have less time to dedicate to the firm).
- The second requirement is the instrument relevance. It amounts to requiring a strong first stage, which we can check by running the regression of board size on the CEO family size (possibly also including other exogenous covariates).
- Finally, the third requirement is the exclusion restriction which means that the CEO family size is not a regressor in the model (we can assume this is the case and that we don't include this variable on the right-hand side of the regression equation).

Thus, it is somewhat doubtful that it is a good instrument as the instrument exogeneity condition is doubtful.

9.2E (a) The problems associated with endogeneity are bias and inconsistency. You should be happy to prove these claims.

(b) We require two conditions from *ELECTION* in order for it to qualify as an instrumental variable:

- *ELECTION* is uncorrelated to the error, i.e. we require $Cov(u_i, ELECTION_i) = 0$ (Validity)
- $Cov(ELECTION_i, POLICE_i) \neq 0$ (Relevance).

The second condition states that *ELECTION* should have a direct effect on *POLICE*. It can be tested from the reduced form estimation given above. The realisation of the t statistic for the null hypothesis of the relevance violation is:

$$t = \frac{1.0599}{0.3358} = 3.1563.$$

This exceeds the two-sided 1% critical value. Therefore, we reject the null hypothesis of no relevance and conclude that the instrument is relevant. The validity condition cannot be tested.

Aside: Consider now a variable like *FUNDING* which is short for police department funding. Is this a good instrument? Clearly it satisfies relevance: we expect more police funding in general to lead to more police officers. What about validity? Now it doesn't seem very likely that funding directly affects crime, i.e. directly enters u_i. On the other hand, however, it might be correlated with other variables in u_i – for example, unemployment, city size etc. For instance places which have higher unemployment in general tend to be poorer which could be adversely affecting police funds; at the same time unemployment has a direct effect on crime. This would violate validity.

(c) For the ease of notation, let $y_i \equiv CRIME_i$, $x_i \equiv POLICE_i$ and $z_i \equiv ELECTION_i$. The IV estimators of the intercept ($\widehat{\alpha}^{IV}$) and the slope ($\widehat{\beta}^{IV}$) solve the following system of moment conditions:

$$\sum_{i=1}^{n}(y_i - \widehat{\alpha}^{IV} - \widehat{\beta}^{IV} x_i) = 0 \tag{2.1}$$

$$\sum_{i=1}^{n}(y_i - \widehat{\alpha}^{IV} - \widehat{\beta}^{IV} x_i) z_i = 0. \tag{2.2}$$

Start from equation 1 and solve for $\widehat{\alpha}^{IV}$ to obtain:

$$\widehat{\alpha}^{IV} = \bar{y} - \widehat{\beta}^{IV} \bar{x}. \tag{2.3}$$

Plug that 2.3 into equation 2 and solve for $\widehat{\beta}^{IV}$:

$$\sum_{i=1}^{n}(y_i - \widehat{\alpha}^{IV} - \widehat{\beta}^{IV} x_i) z_i = 0$$

$$\sum_{i=1}^{n}(y_i - (\bar{y} - \widehat{\beta}^{IV} \bar{x}) - \widehat{\beta}^{IV} x_i) z_i = 0$$

$$\sum_{i=1}^{n}(y_i - \bar{y}) z_i - \widehat{\beta}^{IV} \sum_{i=1}^{n}(x_i - \bar{x})) z_i = 0$$

giving:

$$\widehat{\beta}^{IV} = \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})(z_i - \bar{z})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}. \tag{2.4}$$

We shall now prove that if z_i is indeed a valid instrument for x_i ($Cov(z_i, x_i) \neq 0$ and $E(z_i\varepsilon_i) = 0$), then $\widehat{\beta}^{IV}$ is consistent for β. Notice that by properties of sums we can write equation 2.4 as:

$$\widehat{\beta}^{IV} = \frac{\sum\limits_{i=1}^{n}(z_i - \bar{z})y_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}.$$

Plug in the true model for y_i and rearrange to obtain:

$$\widehat{\beta}^{IV} = \frac{\sum\limits_{i=1}^{n}(z_i - \bar{z})(\alpha + \beta x_i + \varepsilon_i)}{\sum\limits_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})} = \beta + \frac{\sum\limits_{i=1}^{n}(z_i - \bar{z})\varepsilon_i}{\sum\limits_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}.$$

Divide the numerator and denominator of the last term by n and take probability limits:

$$\text{plim } \widehat{\beta}^{IV} = \beta + \frac{\text{plim } n^{-1}\sum\limits_{i=1}^{n}(z_i - \bar{z})\varepsilon_i}{\text{plim } n^{-1}\sum\limits_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}.$$

By the Weak Law of Large Numbers:

$$\text{plim } \widehat{\beta}^{IV} = \beta + \frac{E(z_i\varepsilon_i)}{Cov(z_i, x_i)}.$$

If $Cov(z_i, x_i) \neq 0$ and $E(z_i\varepsilon_i) = 0$, then:

$$\text{plim } \widehat{\beta}^{IV} = \beta.$$

Therefore, the IV estimator is *consistent* for β.

Notice, however, that the IV estimator is not *unbiased*. To get unbiasedness we would need to have $E(\varepsilon_i \,|\, x_1, \ldots, x_n, z_1, \ldots, z_n) = 0$, but the fact that $Cov(x_i, z_i) \neq 0$ (i.e. x_i is endogenous) implies $E(\varepsilon_i \,|\, x_1, \ldots, x_n, z_1, \ldots, z_n) \neq 0$, so:

$$E(\widehat{\beta}^{IV} \,|\, x_1, \ldots, x_n, z_1, \ldots, z_n) \neq \beta$$

and $E(\widehat{\beta}^{IV}) \neq \beta$. Therefore, IV is biased.

(d) The effect is negative as we expected – but it is estimated very imprecisely!

(e) When choosing either of the *valid* instruments we want to choose the one with the higher correlation with *POLICE* – this is due to the formula for the variance of the IV estimator of β (asymptotic formula):

$$AVar(\widehat{\beta}_{IV}) = \frac{\sigma_u^2}{n \times Sample\ \ Var(POLICE)} \frac{1}{(Corr(POLICE, z))^2}.$$

Better than using either one of them we should use both of them. Discuss the 2SLS approach.

(f) To test whether *POLICE* indeed is an endogenous variable (correlated with the error term), we want to test the following null against the alternative which is the negation of the null:

$$H_0 : E(POLICE_i \varepsilon_i) = 0 \text{ (not endogenous)}$$

$$H_1 : E(POLICE_i \varepsilon_i) \neq 0 \text{ (endogenous).}$$

The test for this discussed in Wooldridge is easy to implement because we are asked to simply add the residuals of our reduced form for the endogenous (*POLICE*) variable and test its coefficients' significance. So it proposes we run the following regression:

$$CRIME_i = \alpha + \beta POLICE_i + \delta \widehat{v}_i + \varepsilon_i$$

where $\widehat{v}_i$ are the residuals obtained from the reduced form estimation (in our case this estimation is OLS of *POLICE* on the instrument *ELECTION* and all exogenous regressors in the model, if we had any). After performing this regression we want to test $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$ and can use $\widehat{\delta}/se(\widehat{\delta})$ as usual! The inclusion of the reduced form residual 'controls' for the endogeneity of *POLICE*.

9.3E (a) The coefficient on R_i measures the difference in birthweight between children of mothers who did and did not receive the brochure. This coefficient speaks directly to the effect of an information campaign of mailing brochures on a health outcome of the children of the women who received the brochure. Since the brochures were sent in a quasi-random fashion (even birthdays are probably fairly uncorrelated with personal characteristics) the coefficient is causal.

(b) Instrumental variables. The NHS is interested in the effect of smoking but only manipulates receipt of the brochure. The brochure should affect smoking, which in turn might affect birthweight. So R_i is the instrument and smoking is the endogenous regressor. The assumptions are relevance (verified through a first stage), which seems plausible but we don't know whether the brochure affects smoking; and instrument validity – the brochure should not affect children's birthweight other than through smoking. This is plausible but pointing out the risks of smoking during pregnancy may also make women more aware of other health behaviours like drinking alcohol, coffee etc.

We can't do IV because we don't have observations for smoking.

(c) (i) We can run the regression of smoking on R_i, which is the reduced form or the first stage of the IV model in (b). This tells us whether receiving the brochure actually has any effect on smoking behaviour. This is the missing piece for the IV model in (b).

(ii) Recall in the discussion of the identification of the parameter β_1 using the IV approach we derived that:

$$\beta_1 = \frac{Cov(R, BW)}{Cov(R, S)}.$$

We can rewrite it as:

$$\beta_1 = \frac{Cov(R, BW)/\,Var(R)}{Cov(R, S)/\,Var(R)}.$$

Thus, we can obtain a consistent estimator of β_1 if we obtain consistent estimators of $Cov(R, BW)/\,Var(R)$ and $Cov(R, S)/\,Var(R)$, respectively.

- A consistent estimator of $Cov(R, S)/\,Var(R)$ is obtained as the slope coefficient in (a).
- A consistent estimator of $Cov(R, S)/\,Var(R)$ is obtained as the slope coefficient in the reduced form in (c) (i).
- Thus, we should divide the slope coefficient in (a) by the slope coefficient in (c) (i). We can do this with the Bristol and Bath results if we believe the women and children in the two cities are similar enough that the results would be the same in either place.

9.4E (a) The problem with using OLS on the above regression is that there will be correlation between the error u_i and the regressor $\ln L_i$ as more productive firms (higher u_i) are associated with hiring less workers (lower L_i). This correlation will make the parameter estimates inconsistent. This problem is a result of the omission of relevant variables which creates OVB. The parameter estimates for β_L will be underestimates as they capture the fact that firms with higher technological or managerial efficiency require less labour for the same output.

(b) Two-stage least squares (2SLS) proceeds in two steps. In the first stage, $\ln L_i$ is regressed on all other regressors and the instrument which provides the fitted values of $\ln L_i$ (giving the exogenous variation later):

$$\widehat{\ln L_i} = \widehat{\pi}_0 + \widehat{\pi}_1 w_i + \widehat{\pi}_2 \ln K_i.$$

In the second stage, the original regression is run where the predicted values from the first stage above $\widehat{\ln L_i}$ are used instead of the original regressor $\ln L_i$. We run:

$$\ln Y_i = \beta_0 + \beta_L \widehat{\ln L_i} + \beta_K \ln K_i + u_i.$$

This instrumental variable approach requires the instrument to be valid, relevant and to satisfy the exclusion restriction (the exclusion restriction here obviously holds). Instrument relevance means that the instrument must be partially correlated with the original regressor, or, in terms of the reduced form:

$$\ln L_i = \pi_0 + \pi_1 w_i + \pi_2 \ln K_i + error_i$$

relevance requires $\pi_1 \neq 0$. Instrument validity requires the instrument to be uncorrelated with the error term, that is, $Cov(w_i, u_i) = 0$.

For consistent parameter estimators, we require capital stock to also be exogenous, i.e. uncorrelated with the error term.

(c) If there is not much variation in wages, the instrument is likely not to be very partially correlated with $\ln L_i$. We face the weak instrument problem which means we will have a weak first stage ($\ln L_i$ is not predicted well by the instrument and other exogenous regressors), resulting in imprecise 2SLS estimators. The formula for the variance of the IV estimator clearly indicates that a low correlation between instrument and regressor increases the imprecision of the IV estimator.

Chapter 10
Measurement error

10.1 Introduction

10.1.1 Aims of the chapter

In this chapter we deal with a common phenomenon when working with economic data: measurement error. You will learn the consequences associated with measurement error in the dependent variable and measurement error in the independent variable(s). The consequences critically depend on the assumptions you are willing to make about the measurement error.

10.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the concept of classical errors-in-variables (CEV)
- discuss the impact of measurement error of the dependent variable on OLS
- discuss the impact of measurement error of the explanatory variable on OLS
- derive the consequences of CEV in the simple regression model
- discuss the IV solution to classical measurement error in explanatory variables.

10.1.3 Essential reading

Readings: Wooldridge, Sections 9.4a, 9.4b and 15.4.

10.1.4 Synopsis of this chapter

When using economic data it is important to recognise that many variables may not be measured accurately and contain what is called measurement error. We can think of many reasons for this, such as individuals misreporting values on purpose, recall bias, misreporting values due to lack of clarity of question asked (accidental), or simply because of erroneous data entry.

The consequences of the presence of measurement error in the dependent and/or independent variable on the properties of the OLS estimators are discussed in Section 9.4 in Wooldridge. In Section 9.4a you learn that as long as the measurement error in the dependent variable is not systematically related to any of the explanatory variables, OLS remains unbiased and consistent. Measurement error will render the estimator less precise (increase its variance). In Section 9.4b you learn that measurement error in the independent variable(s) will typically render OLS invalid (inconsistent) because it will give rise to the problem of endogeneity (correlation between the error and the regressor that is measured with error). All parameters will be affected by the presence of measurement error in one or more of the explanatory variables. For the regression model, it is shown how classical measurement error in the explanatory variable will give rise to attenuation bias (whereby the estimated OLS effect, on average, is closer to zero than the true effect). When considering the effect of measurement error on the properties of OLS, it is common to make the classical errors-in-variables (CEV) assumption whereby the measurement error is uncorrelated with the unobserved variable, the other explanatory variables, and the original error term.

In Section 15.4 in Wooldridge, we consider an implementation of the Instrumental Variable estimator to deal with the endogeneity problem associated with measurement error in the explanatory variables.

10.2 Content of chapter

10.2.1 Properties of OLS under measurement error

Read: Wooldridge, Sections 9.4a and 9.4b.

The presence of measurement error may give rise to the endogeneity problem we just discussed which would render the use of OLS undesirable (inconsistent). As discussed in Section 9.4 in Wooldridge, we can have measurement error in the dependent and/or independent variables.

The consequences for the OLS estimator will depend critically on the assumptions you are willing to make about the measurement error. It is common to impose the classical errors-in-variables (CEV) assumption whereby the measurement error is assumed to be uncorrelated with the unobserved variable, the (other) explanatory variables, and the original error term. These assumptions are not necessarily reasonable (and you should be able to explain this), but under this assumption we have the following clear implications associated with the presence of measurement errors.

- Under CEV of the dependent variable, OLS will remain consistent, but less precise (larger variance).
- Under CEV of a single independent variable, OLS will be inconsistent (all parameters). The parameter on the variable measured with error will exhibit **attenuation bias**.

You should be able to prove these properties in the simple regression model setting only. When more than one independent variable is measured with error (CEV), the OLS estimator is inconsistent and little can be said about the direction of the inconsistency.

Activity 10.1 Take the MCQs related to this section on the VLE to test your understanding.

Activity 10.2 Exercise 9.4 from Wooldridge.

10.2.2 IV solutions to errors-in-variable problems

Read: Wooldridge, Section 15.4.

As discussed in Section 15.4 in Wooldridge, one way to resolve the endogeneity problem associated with classical measurement error in an explanatory variable is to use a second measurement of the same explanatory variable as an instrument. It is important to recognise that the measurement errors associated with these two measured-with-error explanatory variables are uncorrelated.

Activity 10.3 Consider the following equation for students' GPA:

$$colGPA_i = \beta_0 + \beta_1 faminc_i^* + \beta_2 hsGPA_i + \beta_3 SAT_i + u_i.$$

Precise data on *colGPA*, *hsGPA*, *SAT* are easy to obtain (using students' transcripts). Students, however, report the family income as *famincs*, which provides a measurement of *faminc** containing measurement error. We assume:

$$famincs_i = faminc_i^* + e_{si}$$

where e_s is a classical measurement error. We obtain a random sample given by $\{(colGPA_i, hsGPA_i, SAT_i, famincs_i) : i = 1, \ldots, n\}$.

(i) Provide the estimable model and show that the error term in the new equation, say, v_i, is negatively correlated with *famincs* if $\beta_1 > 0$. What does this imply about the OLS estimator of β_1 from the above regression?

(ii) Say you are able to get the parents to provide another measure of family income, $fam incp$. Let us assume that:

$$famincp_i = faminc_i^* + e_{pi}$$

where the classical measurement error e_p is uncorrelated with e_s. Our random sample is given by $\{(colGPA_i, hsGPA_i, SAT_i, famincs_i, famincp_i) : i = 1, \ldots, n\}$.

Show that $E(famincp_i\ v_i) = 0$ where v_i is the error term in the model from part (i).

(iii) Are $famincs_i$ and $famincp_i$ likely to be correlated? Explain.

(iv) Use the results in (ii) and (iii) to propose an IV estimator for our β parameters.

10.3 Answers to activities

Solution to Activity 10.2

(Exercise 9.4 from Wooldridge.)

(i) For the CEV assumptions to hold, we must be able to write:

$$tvhours = tvhours^* + e$$

where the measurement error e has zero mean and is uncorrelated with $tvhours^*$ and each explanatory variable in the equation.

Note that for OLS to consistently estimate the parameters, we do not need e to be uncorrelated with $tvhours^*$.

(ii) No, the CEV assumptions are unlikely to hold in this example.

First: e and $tvhours^*$ are likely to be correlated.

- For children who do not watch TV at all, $tvhours^* = 0$, and it is very likely that reported TV hours is zero. So, if $tvhours^* = 0$, then $e = 0$ with high probability.
- For children who do watch TV, $tvhours^* > 0$. Here the measurement error can be positive or negative. However, since $tvhours \geq 0$, e must satisfy $e \geq -tvhours^*$.

Second: e is likely correlated with the explanatory variables.

- As $tvhours^*$ depends directly on the explanatory variables, e will likely be correlated with them too.
- More highly educated parents may tend to underreport how much television their children watch. This means that the measurement error e and the education variables are negatively correlated.

Solution to Activity 10.3

(i) Using $faminc_i^* = famincs_i - e_{si}$, the estimable model can be obtained as:

$$\begin{aligned} colGPA_i &= \beta_0 + \beta_1(famincs_i - e_{si}) + \beta_2 hsGPA_i + \beta_3 SAT_i + u_i \\ &= \beta_0 + \beta_1 famincs_i + \beta_2 hsGPA_i + \beta_3 SAT_i + \underbrace{u_i - \beta_1 e_{si}}_{v_i}. \end{aligned}$$

Consider $Cov(famincs_i, v_i) = Cov(faminc_i^* + e_{si}, u_i - \beta_1 e_{si})$.

- Using CEV this yields $Cov(famincs_i, v_i) = -\beta_1 Var(e_{si})$.
- If $\beta_1 > 0$ this reveals a negative correlation.

The OLS estimator of β_1 will suffer from attenuation bias. The parameter estimate on average will be closer to zero than β_1.

(ii) Realise: $E(famincp_i v_i) = Cov(famincp_i, v_i) = Cov(faminc_i^* + e_{pi}, u_i - \beta_1 e_{si})$.

$faminc_i^*$ is uncorrelated with u_i and e_{si}, and e_{pi} is uncorrelated with u_i and e_{si}. Therefore, $E(famincp_i v_i) = 0$.

(iii) $famincs_i$ and $famincp_i$ are correlated as both measure $faminc_i^*$. We have:

$$\begin{aligned} Cov(famincs_i, famincp_i) &= Cov(faminc_i^* + e_{si}, faminc_i^* + e_{pi}) \\ &= Var(faminc_i^*) \\ &\neq 0. \end{aligned}$$

(iv) Our results show that we can use $famincp$ as an instrument for $famincs$. In particular, $E(famincp_i\ v_i) = 0$ ensures the **validity** of our instrument, and we also have that $Cov(famincs_i, famincp_i) \neq 0$ ensures its **relevance**.

We can implement this estimator using 2SLS.

- Step 1: Obtain the fitted values from the following regression:

$$famincs_i = \pi_0 + \pi_1 famincp_i + \pi_2 hsGPA_i + \pi_3 SAT_i + \varepsilon_i.$$

- Step 2: Estimate the following regression:

$$colGPA_i = \beta_0 + \beta_1 \widehat{famincs}_i + \beta_2 hsGPA_i + \beta_3 SAT_i + \nu_i.$$

10.4 Overview of chapter

We discussed the consequences of measurement error in the dependent and independent variables on the OLS estimator.

As long as the measurement error in the dependent variable is not systematically related to one or more of the explanatory variables, OLS remains unbiased and consistent. Measurement error in the independent variables on the other hand will typically render OLS invalid (inconsistent) because it will give rise to the problem of endogeneity (correlation between the error and the regressor that is measured with error). All parameters will be affected by the presence of measurement error in one or more of the explanatory variables.

When considering the effect of measurement error on the properties of OLS, it is common to consider a classical error-in-variables (CEV) assumption whereby the measurement error is uncorrelated with the unobserved variable, the other explanatory variables in the model, and the original error term.

In the simple regression model, we showed that CEV in the explanatory variable will give rise to attenuation bias, where the estimated OLS effect, on average, is closer to zero than the true effect.

We concluded with a discussion of IV to deal with the problem of CEV in explanatory variables.

10.4.1 Key terms and concepts

- Measurement error
- Classical errors-in-variables (CEV)
- Attenuation bias

10.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the concept of classical errors-in-variables (CEV)
- discuss the impact of measurement error of the dependent variable on OLS
- discuss the impact of measurement error of the explanatory variable on OLS
- derive the consequences of CEV in the simple regression model
- discuss the IV solution to classical measurement error in explanatory variables.

10.5 Test your knowledge and understanding

10.5.1 Additional examination based exercises

10.1E Consider the following regression model:

$$y_i = \beta x_i^* + u_i, \quad i = 1, \ldots, n$$

which satisfies the Gauss–Markov assumptions. The variables are presented in deviations of their means. We do not observe x_i^*, instead we observe:

$$x_i = x_i^* + e_i$$

where e_i is a classical error-in-variables (CEV) error with zero mean and constant variance.

(a) Explain what it means to say that we have a CEV. How realistic is this assumption?

(b) We assume we can obtain a random sample $\{(y_i, x_i), i = 1, \ldots, n\}$. If $\widehat{\beta}$ is the OLS estimator of β from regressing y_i on x_i show that $\widehat{\beta}$ is inconsistent and exhibits attenuation bias.

(c) Say you are told now that the dependent variable exhibits measurement error as well. The observed dependent variable is given by:

$$y_i^* = y_i + w_i$$

where the measurement error w_i is assumed to be uncorrelated with e_i. Under what additional assumption will the OLS estimator of β that uses the random sample $\{(y_i^*, x_i), i = 1, \ldots, n\}$ give rise to the same attenuation bias result as discussed in (b)?

10.5.2 Solutions to additional examination based exercises

10.1E (a) The CEV assumptions require the measurement error to be uncorrelated with the *unobserved* explanatory variable, $Cov(e, x_1^*) = E(ex_1^*) = 0$, and to be uncorrelated with the error in the original equation, $Cov(e, u) = E(eu) = 0$.

The assumption that the measurement error is uncorrelated with the *unobserved* explanatory variable is very strong, and in many cases unlikely to be true. It assumes that the measurement error is purely random, such as with erroneous data entry.

(b) The equation we estimate is given by:

$$y_i = \beta x_i + (u_i - \beta e_i), \quad i = 1, \ldots, n \qquad (*)$$

where we denote the composite error term $v_i = u_i - \beta e_i$. We need to show that $\text{plim}(\widehat{\beta}) \neq \beta$, where:

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \beta + \frac{\sum_{i=1}^n x_i v_i}{\sum_{i=1}^n x_i^2}. \qquad \text{(plug in (*))}$$

Rewriting in terms of averages first yields:

$$\widehat{\beta} = \beta + \frac{\frac{1}{n}\sum_{i=1}^n x_i v_i}{\frac{1}{n}\sum_{i=1}^n x_i^2}.$$

By properties of the plim we obtain:

$$\text{plim}\ \widehat{\beta} = \beta + \frac{\text{plim}\ \frac{1}{n}\sum_{i=1}^n x_i v_i}{\text{plim}\ \frac{1}{n}\sum_{i=1}^n x_i^2} = \beta + \frac{E(x_i v_i)}{E(x_i^2)}$$

where the second equality applies the law of large numbers (which ensures that sample averages converge in probability to their population analogues).

Using the CEV and the Gauss–Markov assumption that $E(x_i^* u_i) = 0$, we can show:

$$E(x_i^2) = E[(x_i^* + e_i)^2] = E[(x_i^*)^2 + e_i^2 + 2x_i^* e_i)] = E[(x_i^*)^2] + Var(e_i) + 0 \neq 0$$

and:

$$E(x_i v_i) = E[(x_i^* + e_i)(u_i - \beta e_i)] = -\beta\, Var(e_i).$$

Combining results yields:

$$\text{plim}\ \widehat{\beta} = \beta\left(1 - \frac{Var(e_i)}{E[(x_i^*)^2] + Var(e_i)}\right).$$

As:

$$0 < \left(1 - \frac{Var(e_i)}{E[(x_i^*)^2] + Var(e_i)}\right) < 1$$

we establish the attenuation bias result.

(c) The equation we estimate is given by:

$$y_i^* = \beta x_i + (u_i - \beta e_i + w_i), \quad i = 1, \ldots, n \tag{**}$$

where we denote the composite error term $v_i = u_i - \beta e_i + w_i$. We need to consider:

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i^*}{\sum_{i=1}^n x_i^2} = \beta + \frac{\sum_{i=1}^n x_i v_i}{\sum_{i=1}^n x_i^2}. \qquad \text{(plug in (**))}$$

By properties of the plim and application of the law of large numbers we obtain:

$$\text{plim } \widehat{\beta} = \beta + \frac{\text{plim } \frac{1}{n}\sum_{i=1}^n x_i v_i}{\text{plim } \frac{1}{n}\sum_{i=1}^n x_i^2} = \beta + \frac{E(x_i v_i)}{E(x_i^2)}$$

where now:

$$E(x_i v_i) = E[x_i(u_i - \beta e_i + w_i)].$$

To obtain the same result as in (b), we will need to assume $E(x_i w_i) = 0$. As $x_i = x_i^* + e_i$ that is satisfied if the measurement error w_i is uncorrelated with the measurement error e_i (as is assumed) and w_i is uncorrelated with x_i^*.

Chapter 11
Simultaneous equation models

11.1 Introduction

11.1.1 Aims of the chapter

Until now we have considered the estimation of equations in isolation. In this chapter we argue that in economics many behavioural relations (such as the demand and supply equations) are really part of a system of simultaneous equations. If this is the case the application of OLS to a single behavioural relationship in isolation yields biased and inconsistent estimates arising from the presence of explanatory variables in the behavioural equation that are jointly determined with the dependent variable. We will show that these variables are correlated with the error, hence giving rise to the endogeneity problem. To solve the problem, we will use 2SLS (IV) which involves the search for instruments to deal with the simultaneity problem (endogeneity of explanatory variables that are jointly determined with the dependent variables). As we will see, simultaneous equation models (SEM) provide a setting where the discussion of suitable instruments to deal with endogeneity is quite natural.

11.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the concept of simultaneity in regression
- understand the need for supply (demand) shifters to identify the demand (supply) equation graphically
- explain the concepts of a structural equation and a reduced form equation in the simultaneous equations framework
- distinguish endogenous variables (determined within the SEM) and exogenous variables (determined outside the model) in the simultaneous equations framework
- explain the problem of simultaneity bias when OLS is applied to a single structural equation in isolation
- derive the correlation between the jointly determined explanatory variable and the structural equation error
- derive an expression for the large-sample simultaneity bias in the slope coefficients when OLS is used to fit a simple regression in a simultaneous equations model

- discuss the order and rank condition of identification in a two-equation system
- explain what is meant by a just identified, unidentified, and overidentified equation
- discuss 2SLS estimation of the structural form equations in SEM.

11.1.3 Essential reading

Readings: Wooldridge, Sections 16.1–16.3.

11.1.4 Synopsis of this chapter

In Section 16.1 in Wooldridge simultaneous equations models (SEMs) are introduced, which describe a system of behavioural relations that **interact** with each other.

A classic example is given by the demand and supply for a commodity (such as labour). A key aspect we should recognise is that observational data we can use to estimate these behavioural relations are given by the market-clearing equilibria of prices and quantities where demand equals supply. It is argued that the ability for us to estimate (identify) the demand and supply equations will require that we have supply and demand shifters (observable variables that enter only the supply or demand equation) as without these there is no way to distinguish the demand from the supply equation given the observed data. This issue is discussed in more detail in Section 16.3. When we consider relationships between economic variables, simultaneity is also present when we recognise that the causal relationship may go either way (describe behaviour of different agents). This is illustrated by specifying a SEM for murder rates and the size of the police force.

In Section 16.2 you learn how to show that SEMs give rise to the endogeneity problem, and hence explain the simultaneity bias associated with estimating single behaviour relationships (structural equation) in isolation. For this we will derive the reduced form for the endogenous variables, where reduced form equations express the endogenous variable in terms of all the exogenous variables and errors in the model.

In Section 16.3, the identification issue of SEMs is discussed in more detail. This centres around the question whether with observable data we are able to estimate the behavioural equations and is directly related to the availability of instruments that can be used to deal with the endogenous (jointly determined) regressor. We will distinguish between underidentification, exact (just) identification, and overidentification. Equations that can be identified will be estimated by 2SLS (IV), where the instruments to deal with endogeneity can be found in the SEM itself.

11.2 Content of chapter

11.2.1 The nature of simultaneous equations models

Read: Wooldridge, Section 16.1.

Simultaneity arises when some of the regressors are **jointly determined** with the dependent variable in the same economic model. A typical example is that of a market equilibrium, where we consider the simultaneous equations model (SEM) given by:

$$q_i = \alpha_1 p_i + \beta_1 z_{1i} + u_{1i} \quad \text{(the supply function)}$$

$$q_i = \alpha_2 p_i + \beta_2 z_{2i} + u_{2i} \quad \text{(the demand function).}$$

We observe a random sample of q_i (quantity), p_i (price), z_{1i} (observable supply shifter), z_{2i} (observable demand shifter), and (q_i, p_i) is the equilibrium outcome. Each equation has a *ceteris paribus*, causal interpretation, representing behavioural relations; we refer to them as the **structural equations** of a SEM. The error terms, u_{1i} and u_{2i}, are called the **structural errors**.

- Variables jointly determined in this model are called the **endogenous variables** (in this example q_i and p_i).
- Variables determined outside this model are called **exogenous variables** (in this example z_{1i} and z_{2i}).

Without the inclusion of z_{1i} and z_{2i} in this model, there is no way to tell which equation is supply and which is demand (called the **identification problem**). Similarly if $z_{1i} = z_{2i}$, then the two equations become identical and there is no hope to estimate either the supply or demand equation. This indicates that to be able to estimate supply or demand, they must have different exogenous variables. This is illustrated graphically later.

You should recognise when it is inappropriate to suggest the use of a SEM.

> **Activity 11.1** Take the MCQs related to this section on the VLE to test your understanding.

11.2.2 Simultaneity bias in OLS

Read: Wooldridge, Section 16.2.

Estimation of structural equations by OLS will typically give rise to bias and inconsistency. The reason for this is that the explanatory variable that is determined simultaneously with the dependent variable typically is correlated with the error term, that is endogenous. We refer to this as the **simultaneity bias** of OLS.

The key ingredient we use to prove this is the **reduced form**. The reduced form expresses the endogenous variables in terms of the exogenous variables in the model and the structural errors. You should be able to derive the reduced forms of the endogenous variables from a two-equation structural model. Let us here consider the market equilibrium SEM model, given by:

$$q = \alpha_1 p + \beta_1 z_1 + u_1 \quad \text{(the supply function)}$$

$$q = \alpha_2 p + \beta_2 z_2 + u_2 \quad \text{(the demand function).}$$

We can solve the SEM for p by equating demand and supply:

$$\begin{aligned}\alpha_1 p + \beta_1 z_1 + u_1 &= \alpha_2 p + \beta_2 z_2 + u_2\\ (\alpha_1 - \alpha_2)p &= -\beta_1 z_1 + \beta_2 z_2 + u_2 - u_1.\end{aligned}$$

Since $\alpha_1 \neq \alpha_2$ (supply and demand slope, respectively), we can divide this equation by $\alpha_1 - \alpha_2$ as follows:

$$p = \frac{-\beta_1}{\alpha_1 - \alpha_2} z_1 + \frac{\beta_2}{\alpha_1 - \alpha_2} z_2 + \frac{u_2 - u_1}{\alpha_1 - \alpha_2}.$$

The resulting equation is called the reduced form equation for p, which is:

$$p = \pi_{p1} z_1 + \pi_{p2} z_2 + v_p$$

where π_{p1} and π_{p2} are the reduced form parameters (nonlinear functions of the structural parameters) and v_p is the reduced form error, which is a linear function of u_1 and u_2.

To obtain the reduced form for q, we can plug this equation into either the supply or demand function. Alternatively, we can use the following approach. Let us multiply the supply equation by α_2 and the demand equation by α_1 and take their difference:

$$\begin{array}{cl} \alpha_2 q & = \alpha_2 \alpha_1 p + \alpha_2 \beta_1 z_1 + \alpha_2 u_1 \\ \alpha_1 q & = \alpha_1 \alpha_2 p + \alpha_1 \beta_2 z_2 + \alpha_1 u_2 \\ \hline (\alpha_2 - \alpha_1) q & = \alpha_2 \beta_1 z_1 - \alpha_1 \beta_2 z_2 + \alpha_1 u_2 - \alpha_2 u_1. \end{array} \quad -$$

we can divide this equation by $\alpha_2 - \alpha_1$ to obtain the result:

$$q = \underbrace{\frac{\alpha_2 \beta_1}{\alpha_2 - \alpha_1}}_{\pi_{q1}} z_1 + \underbrace{\frac{-\alpha_1 \beta_2}{\alpha_2 - \alpha_1}}_{\pi_{q2}} z_2 + \underbrace{\frac{\alpha_1 u_2 - \alpha_2 u_1}{\alpha_2 - \alpha_1}}_{v_q}.$$

To show that the supply equation, i.e. $q = \alpha_1 p + \beta_1 z_1 + u_1$, suffers from simultaneity bias and hence should not be estimated by OLS, we need to show that $Cov(p, u_1) \neq 0$. Using the reduced form, we then note:

$$Cov(p, u_1) = Cov(\pi_{p1} z_1 + \pi_{p2} z_2 + v_p, u_1) = Cov(v_p, u_1) \neq 0.$$

The second equality uses the property of the covariance operator and the exogeneity of z_1 and z_2. Indeed, using the definition of v_p we obtain (say u_1 and u_2 are uncorrelated):

$$\begin{aligned} Cov(v_p, u_1) = Cov\left(\frac{u_2 - u_1}{\alpha_1 - \alpha_2}, u_1\right) &= \frac{1}{\alpha_1 - \alpha_2} Cov(u_2 - u_1, u_1)\\ &= \frac{1}{\alpha_1 - \alpha_2}(Cov(u_2, u_1) - Cov(u_1, u_1))\\ &= -\frac{\sigma_1^2}{\alpha_1 - \alpha_2}\\ &\neq 0. \end{aligned}$$

The same reasoning reveals that the demand equation suffers from simultaneity bias as well.

Activity 11.2 Take the MCQs related to this section on the VLE to test your understanding.

Activity 11.3 Exercise 16.1 from Wooldridge.

Activity 11.4 Let us consider the simultaneous equation model (SEM) for murder rates and the size of the police force. Let the SEM be given by:

$$murdpc = \alpha_0 + \alpha_1 polc + u_1$$
$$polc = \beta_0 + \beta_1 murdpc + \beta_2 election + u_2$$

where *election* is an exogenous variable (determined outside the model) and *murdpc* and *polc* are endogenous. Let $Var(u_1) = \sigma_1^2$, $Var(u_2) = \sigma_2^2$, and assume that $Cov(u_1, u_2) = 0$.

(i) Derive the reduced form for *polc*.

(ii) Show that the endogenous variable *polc* is correlated with u_1 using your answer in (a).

(iii) Show that the OLS estimator of α_1 is inconsistent.

(iv) Discuss whether the simultaneity bias in the OLS estimator is upward or downward and interpret this result.

11.2.3 Identifying and estimating a structural equation

Read: Wooldridge, Section 16.3.

In Section 16.3a in Wooldridge we learn that it is important to establish whether a structural form equation is identified *prior* to its estimation. In the market equilibrium setting, the identification discussion establishes whether given observable data we can really distinguish between the demand and supply equations. As we learn, the identification of the supply equation is directly related to the availability of instruments (demand shifters) that enable us to deal with the endogeneity of prices.

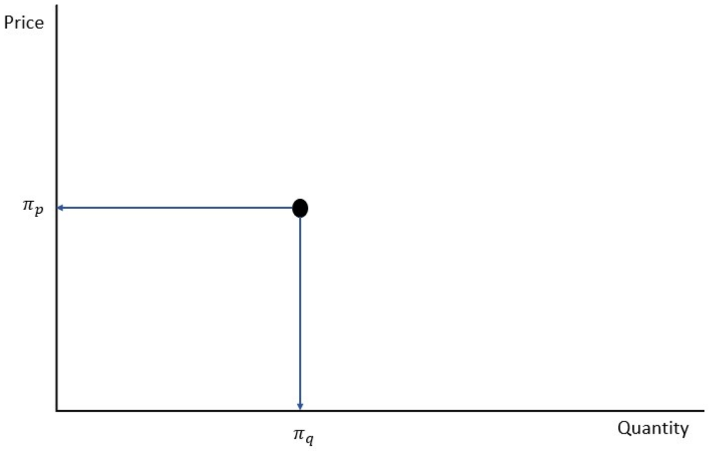
Explanation of the identification problem

Let us discuss here what the identification problem is when we consider a demand and supply model where there are no demand/supply shifters, i.e. consider:

$$q = \alpha_0 + \alpha_1 p + u_1 \quad \text{(supply)}$$
$$q = \beta_0 + \beta_1 p + u_2 \quad \text{(demand)}$$

where (u_1, u_2) are i.i.d. error terms with zero mean.

In this model the only observable variables are quantity (q) and prices (p), and we need to recognise that this data is only informative about the equilibrium price and quantity (π_p, π_q).



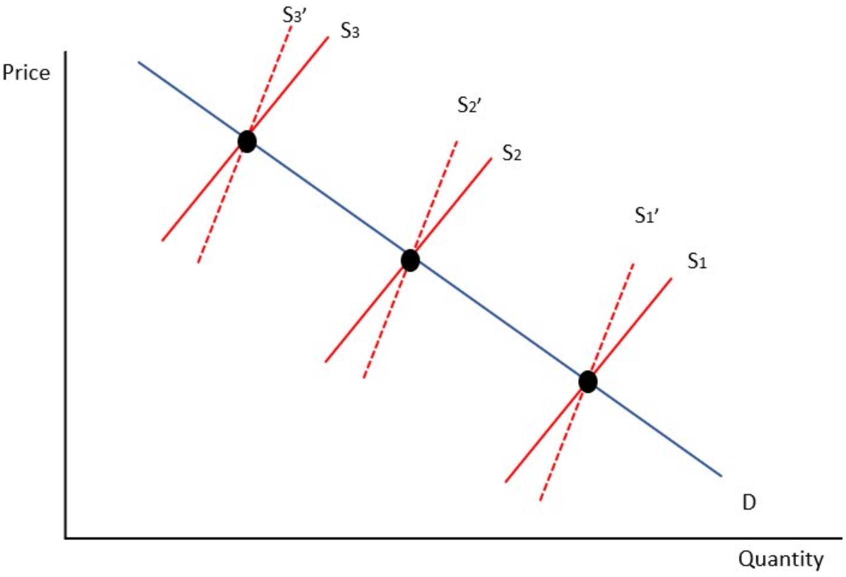
The data is not informative about the slope (intercept) of the demand and supply equation.

Next, let us discuss what the identification problem is when we consider a demand and supply model where we only have a supply shifter, i.e. consider:

$$q = \alpha_0 + \alpha_1 p + \alpha_2 z + u_1 \quad \text{(supply)}$$
$$q = \beta_0 + \beta_1 p + u_2 \quad \text{(demand)}$$

where (u_1, u_2) are i.i.d. error terms with zero mean and z is an exogenous supply shifter. Do we also have an **identification problem** here?

In this model, the observable variables are quantity (q), prices (p), and a supply shifter (z). Here the data is informative about the equilibrium price and quantity for different values for z. Since the demand function does not depend on the supply shifter, this permits us to identify (trace out) the demand equation.



We cannot identify the supply equation as there are many supply functions (say S and S') that we cannot distinguish given the observable data. *Note:* the slope of the supply function stays the same when z changes.

Identification condition of a structural equation

Let us consider a general SEM:

$$y_1 = \beta_{10} + \alpha_1 y_2 + \beta_{11} z_{11} + \cdots + \beta_{1k_1} z_{1k_1} + u_1$$

$$y_2 = \beta_{20} + \alpha_2 y_1 + \beta_{21} z_{21} + \cdots + \beta_{2k_2} z_{2k_2} + u_2$$

where (y_1, y_2) are endogenous variables determined in this model, $(z_{11}, \ldots, z_{1k_1})$ are exogenous variables appearing in the first equation, and $(z_{21}, \ldots, z_{2k_2})$ are exogenous variables appearing in the second equation. Typically there is an overlap between $(z_{11}, \ldots, z_{1k_1})$ and $(z_{21}, \ldots, z_{2k_2})$. In fact: *To ensure our equations are identified, the regressors cannot completely overlap.*

The first equation in the SEM is identified if and only if the *second* equation contains at least one exogenous variable with a non-zero coefficient that is excluded from the first equation. This is the **rank condition**. A necessary condition for this requirement is that the *first* equation excludes at least one exogenous variable. This is the **order condition**. [Equivalently, we need at least as many excluded exogenous variables in the first equation as there are included endogenous variables in the first equation, which in this case equals one (y_2) – See also the previous discussion of 2SLS.]

Recognising that the excluded exogenous variables are needed as instruments for our included endogenous variables, we use the following terminology:

- An equation is **overidentified** if we have more excluded exogenous variables than included endogenous variables.
- An equation is **just identified** if we have exactly as many excluded exogenous variables as included endogenous variables.
- An equation is **unidentified** if we have less excluded exogenous variables than included endogenous variables.

Estimation of a structural equation

Let us focus on estimation of the first equation. Once our identification condition is satisfied, we can then apply the 2SLS estimator.

- Identification ensures that we have enough instruments to allow us to deal with the endogeneity of y_2. Since we have one endogenous variable in the first equation, we need at least one instrument.

The 2SLS estimator of the first equation takes the following two steps.

1. *Run OLS of y_2 on $(\mathbf{z}_1, \mathbf{z}_2)$ and obtain the fitted values $\widehat{y}_2$.*

- Effectively, we are estimating the reduced form of the endogenous regressor y_2.

2. *Run OLS of y_1 on $\widehat{y}_2$ and $\mathbf{z_1}$.*

- By using $\widehat{y}_2$ in place of y_2, we can estimate the structural equation by OLS:

$$y_1 = \beta_{10} + \alpha_1 \widehat{y}_2 + \beta_{11} z_{11} + \cdots + \beta_{1k_1} z_{1k_1} + e.$$

In Wooldridge various empirical applications are discussed.

Activity 11.5 Take the MCQs related to this section on the VLE to test your understanding.

Activity 11.6 Exercise 16.4 from Wooldridge.

Activity 11.7 (Based on Exercise 16.C1 from Wooldridge.)

(i) A model to estimate the effects of smoking on annual income (perhaps through lost work days due to illness, or productivity effects) is:

$$\log(\mathit{income}) = \beta_0 + \beta_1 \mathit{cigs} + \beta_2 \mathit{educ} + \beta_3 \mathit{age} + \beta_4 \mathit{age}^2 + u_1$$

where *cigs* is the number of cigarettes smoked per day, on average. Interpret β_1.

(ii) To reflect the fact that cigarette consumption might be jointly determined with income, a demand for cigarettes equation is given:

$$\begin{aligned} \mathit{cigs} = \gamma_0 + \gamma_1 \log(\mathit{income}) + \gamma_2 \mathit{educ} + \gamma_3 \mathit{age} + \gamma_4 \mathit{age}^2 \\ + \gamma_5 \log(\mathit{cigpric}) + \gamma_6 \mathit{restaurn} + u_2 \end{aligned}$$

where *cigpric* is the price of a pack of cigarettes (in cents) and *restaurn* is a binary variable equal to 1 if the person lives in a state with restaurant smoking restrictions, 0 otherwise. We will assume these are exogenous to the individual.

Discuss under what assumption the income equation from part (i) is identified. What signs you would expect for γ_5 and γ_6?

(iii) The estimated reduced form for *cigs* is given by:

$$\widehat{\mathit{cigs}} = \underset{(23.70)}{1.58} - \underset{(.162)}{.450}\, \mathit{educ} + \underset{(.154)}{.823}\, \mathit{age} - \underset{(.0017)}{.0096}\, \mathit{age}^2 - \underset{(5.766)}{.351}\, \log(\mathit{cigpric}) - \underset{(1.110)}{2.736} \log(\mathit{restaurn})$$

$$n = 807,\ R^2 = .051.$$

The standard errors are reported in parentheses.

Briefly discuss how we can estimate the reduced form and discuss whether $\log(\mathit{cigpric})$ and *restaurn* are statistically significant. What can you conclude from this result?

(iv) Estimating the log(*income*) equation by 2SLS gives:

$$\widehat{\log(income)} = \underset{(.23)}{7.78} - \underset{(.026)}{.042}\, cigs + \underset{(.016)}{.040}\, educ + \underset{(.023)}{.094}\, age - \underset{(.00027)}{.00105}\, age^2.$$

By comparison, when estimating the log(*income*) equation by OLS we obtain:

$$\widehat{\log(income)} = \underset{(.17)}{7.80} + \underset{(.0017)}{.0017}\, cigs + \underset{(.008)}{.060}\, educ + \underset{(.008)}{.058}\, age - \underset{(.00008)}{.00063}\, age^2$$

$$n = 807,\ R^2 = .165.$$

Discuss the estimates of β_1.

(v) How would you test whether the OLS and 2SLS estimates in (iv) are statistically different (that is whether we should worry about the endogeneity of *cigs*)?

11.3 Answers to activities

Solution to Activity 11.3

(Exercise 16.1 from Wooldridge.)

(i) If $\alpha_1 = 0$, then $y_1 = \beta_1 z_1 + u_1$ is the reduced form for y_1. As required, the right-hand side depends only on the exogenous variable z_1 and the error term u_1.

If $\alpha_2 = 0$, the reduced form for y_1 is $y_1 = \beta_2 z_2 + u_2$.

- Note that having both α_1 and α_2 equal zero is not interesting as it implies the bizarre condition $u_2 - u_1 = \beta_1 z_1 - \beta_2 z_2$.

To derive the reduced form for y_2 when $\alpha_1 \neq 0$ and $\alpha_2 = 0$, we proceed as follows.

- We can plug $y_1 = \beta_2 z_2 + u_2$ into the first equation and solve for y_2:

$$\beta_2 z_2 + u_2 = \alpha_1 y_2 + \beta_1 z_1 + u_1.$$

- We can rearrange this as: $\alpha_1 y_2 = -\beta_1 z_1 + \beta_2 z_2 + u_2 - u_1$.
- Divide by α_1 (because $\alpha_1 \neq 0$):

$$y_2 = \underbrace{-(\beta_1/\alpha_1)}_{\pi_{21}} z_1 + \underbrace{(\beta_2/\alpha_1)}_{\pi_{22}} z_2 + \underbrace{(u_2 - u_1)/\alpha_1}_{v_2}.$$

(ii) If we multiply the first structural equation by α_2 and multiply the second equation by α_1 and subtract them from each other we can derive the reduced form for y_1.

Observe that the suggested subtraction allows y_2 to cancel out:

$$\begin{array}{rl} \alpha_2 y_1 = & \alpha_2\alpha_1 y_2 + \alpha_2\beta_1 z_1 + \alpha_2 u_1 \\ \alpha_1 y_1 = & \alpha_1\alpha_2 y_2 + \alpha_1\beta_2 z_2 + \alpha_1 u_2 \quad - \\ \hline (\alpha_2 - \alpha_1) y_1 = & \alpha_2\beta_1 z_1 - \alpha_1\beta_2 z_2 + \alpha_2 u_1 - \alpha_1 u_2 \end{array}$$

Because $\alpha_1 \neq \alpha_2$, we can divide the equation by $\alpha_2 - \alpha_1$ to obtain the reduced form for y_1.

The reduced form for y_1 therefore is:

$$y_1 = \pi_{11}z_1 + \pi_{12}z_2 + v_1$$

where $\pi_{11} = \frac{\beta_1\alpha_2}{\alpha_2-\alpha_1}$, $\pi_{12} = \frac{-\beta_2\alpha_1}{\alpha_2-\alpha_1}$, and $v_1 = \frac{\alpha_2}{\alpha_2-\alpha_1}u_1 - \frac{\alpha_1}{\alpha_2-\alpha_1}u_2$.

If $\alpha_1 \neq 0$ and $\alpha_2 \neq 0$, y_2 also has a reduced form. We can obtain it by subtracting the two structural equations from each other:

$$\begin{array}{rcl} y_1 = & \alpha_1 y_2 + \beta_1 z_1 + u_1 & \\ y_1 = & \alpha_2 y_2 + \beta_2 z_2 + u_2 & - \\ \hline 0 = & (\alpha_1 - \alpha_2)y_2 + \beta_1 z_1 - \beta_2 z_2 + u_1 - u_2 & \end{array}$$

We can rearrange this to yield:

$$(\alpha_1 - \alpha_2)y_2 = -\beta_1 z_1 + \beta_2 z_2 + u_2 - u_1.$$

Because $\alpha_1 \neq \alpha_2$, we can divide by $\alpha_1 - \alpha_2$ to obtain the reduced form:

$$y_2 = \pi_{21}z_1 + \pi_{22}z_2 + v_2$$

where $\pi_{21} = \frac{-\beta_1}{\alpha_1-\alpha_2}$, $\pi_{22} = \frac{\beta_2}{\alpha_1-\alpha_2}$, and $v_2 = \frac{1}{\alpha_1-\alpha_2}(u_2 - u_1)$.

(iii) In supply and demand examples, $\alpha_1 \neq \alpha_2$ is very reasonable.

If the first equation is the supply function, we generally expect $\alpha_1 > 0$, and if the second equation is the demand function, $\alpha_2 < 0$.

The reduced forms can exist even in cases where the supply function is not upward sloping and the demand function is not downward sloping, but we might question the usefulness of such models.

Solution to Activity 11.4

(i) Let us substitute the first structural equation in the second one:

$$polc = \beta_0 + \beta_1(\alpha_0 + \alpha_1 polc + u_1) + \beta_2 election + u_2.$$

We can rearrange this to yield:

$$(1 - \beta_1\alpha_1)polc = \beta_0 + \beta_1\alpha_0 + \beta_1 u_1 + \beta_2 election + u_2.$$

As $\beta_1\alpha_1 \neq 1$ we can divide this equation by $1 - \beta_1\alpha_1$ to yield the reduced form.

- $\alpha_1 < 0$ as more police disincentivises crime.
- $\beta_1 > 0$ as higher murder rates will increase the police force.

The reduced form for *polc*, therefore, is:

$$polc = \pi_0 + \pi_1 election + v$$

where $\pi_0 = \frac{\beta_0+\beta_1\alpha_0}{1-\beta_1\alpha_1}$, $\pi_1 = \frac{\beta_2}{1-\beta_1\alpha_1}$ and $v = \frac{\beta_1}{1-\beta_1\alpha_1}u_1 + \frac{1}{1-\beta_1\alpha_1}u_2$.

(ii) Let us plug in the reduced form and apply the rules of covariances:

$$\begin{aligned} Cov(\pi_0 + \pi_1 election + v, u_1) &= \pi_1 Cov(election, u_1) + Cov(v, u_1) \\ &= Cov(v, u_1) \text{ because } election \text{ is exogenous.} \end{aligned}$$

Given the definition of $v = \frac{\beta_1}{1-\beta_1\alpha_1}u_1 + \frac{1}{1-\beta_1\alpha_1}u_2$, we obtain:

$$Cov(v, u_1) = Cov\left(\frac{\beta_1}{1-\beta_1\alpha_1}u_1 + \frac{1}{1-\beta_1\alpha_1}u_2, u_1\right) = \frac{\beta_1}{1-\beta_1\alpha_1}\sigma_1^2$$

as we assumed $Cov(u_1, u_2) = 0$.

Therefore, $Cov(polc, u_1) = \frac{\beta_1}{1-\beta_1\alpha_1}\sigma_1^2 \neq 0$.

(iii) The OLS estimator $\widehat{\alpha}_1$ is given by:

$$\begin{aligned} \widehat{\alpha}_1 &= \frac{\sum_{i=1}^n (polc_i - \overline{polc})(murdpc_i - \overline{murdpc})}{\sum_{i=1}^n (polc_i - \overline{polc})^2} \\ &= \alpha_1 + \frac{\sum_{i=1}^n (polc_i - \overline{polc})(u_{i1} - \overline{u_1})}{\sum_{i=1}^n (polc_i - \overline{polc})^2} = \alpha_1 + \frac{Sample\ Cov(polc_i, u_{i1})}{Sample\ Var(polc_i)}. \end{aligned}$$

We now apply the plim operator, to reveal:

$$\text{plim}(\widehat{\alpha_1}) = \alpha_1 + \frac{\text{plim } Sample\ Cov(polc_i, u_{i1})}{\text{plim } Sample\ Var(polc_i)} = \alpha_1 + \frac{Cov(polc, u_1)}{Var(polc)}$$

where the second equality applies the law of large numbers.

As $Cov(polc, u_1) \neq 0$ we will get the inconsistency as plim $(\widehat{\alpha_1}) \neq \alpha_1$.

(iv) The simultaneity bias (inconsistency) is given by:

$$\frac{Cov(polc, u_1)}{Var(polc)}.$$

As $Var(polc) > 0$, the sign of this bias is given by the sign of $Cov(polc, u_1)$.

Using result in (ii) $Cov(polc, u_1) = \frac{\beta_1}{1-\beta_1\alpha_1}\sigma_1^2$, as $\beta_1\alpha_1 < 0$, and recognising that $\beta_1 > 0$, this is positive.

The upward bias will mean that OLS will underestimate the effectiveness of a larger police force.

Solution to Activity 11.6

Let us look at the second equation first.

Observe: There are no exogenous variables appearing in the first equation that are not also in the second equation.

We need a 'log(*earnings*) shifter' to identify the *alcohol* equation.

The rank condition for identifying the second equation therefore does not hold, and this equation is **not identified**.

The first equation is identified, provided $\gamma_3 \neq 0$ (and we would expect $\gamma_3 < 0$).

Here we do have an exogenous '*alcohol* shifter' we can use to identify the $\log(\mathit{earnings})$ equation.

The equation is **just identified**, as we have one exogenous variable, $\log(\mathit{price})$, we can use to deal with the endogenous regressor *alcohol*.

We can use 2SLS to estimate the β parameters (IV can be used too as the equation is just identified).

The two steps of 2SLS are as follows:

1. Estimate the reduced form for *alcohol* by OLS:

$$\mathit{alcohol} = \pi_0 + \pi_1 \mathit{educ} + \pi_2 \log(\mathit{price}) + v.$$

2. Use the fitted values of the estimated reduced form to estimate by OLS:

$$\log(\mathit{earnings}) = \beta_0 + \beta_1 \widehat{\mathit{alcohol}} + \beta_2 \mathit{educ} + e.$$

Solution to Activity 11.7

(i) Assuming the structural equation represents a causal relationship, $100 \cdot \beta_1$ is the approximate percentage change in income if a person smokes one more cigarette per day, *ceteris paribus* (semi-elasticity).

(ii) For identification we need at least one exogenous variable in the *cigs* equation that is not also in the $\log(\mathit{income})$ equation.

Therefore, we need $\gamma_5 \neq 0$ or $\gamma_6 \neq 0$ to be different from zero.

If both are non-zero the equation is overidentified, otherwise it is just identified.

- Since consumption and price are, *ceteris paribus*, negatively related: $\gamma_5 \leq 0$.
- Everything else equal, restaurant smoking restrictions should reduce cigarette smoking: $\gamma_6 \leq 0$.

(iii) As all regressors are exogenous, we can estimate the reduced form by OLS.

Under CLM assumptions we can use the t test to test the significance of $\log(\mathit{cigpric})$ and *restaurn*, with the distribution under H_0 being t_{807-6}.

- $\log(\mathit{cigpric})$ is highly insignificant: $\widehat{\pi}_{\log(\mathit{cigpric})}/se(\widehat{\pi}_{\log(\mathit{cigpric})}) \approx -0.06$.
- *restaurn* is significant: $\widehat{\pi}_{\mathit{restaurn}}/se(\widehat{\pi}_{\mathit{restaurn}}) \approx -2.47$; has sign as expected.
- It appears we have identification (can drop $\log(\mathit{cigpric})$).

(iv) Remember, **OLS ignores potential simultaneity** between income and cigarette smoking. OLS is expected to be biased/inconsistent.

The coefficient on *cigs* using OLS implies that cigarette smoking causes income to increase, although the coefficient is not statistically different from zero.

Remember, **2SLS accounts for the potential simultaneity** between income and cigarette smoking by using $\log(cigpric)$ and *restaurn* as instruments for *cigs* (consistent).

The coefficient on *cigs* is negative and almost significant at the 10% level against a two-sided alternative.

The estimated effect is very large: each additional cigarette someone smokes lowers predicted income by about 4.2%, *ceteris paribus*. Of course, the 95% confidence interval for β_{cigs} is very wide.

(v) We would obtain the residuals from the reduced form of *cigs*, which we shall call $\widehat{v}$. We then estimate the $\log(income)$ structural equation by OLS while controlling for $\widehat{v}$, that is:

$$\log(income) = \beta_0 + \beta_1 cigs + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \delta\widehat{v} + e.$$

Then we test $H_0 : \delta = 0$ against $H_1 : \delta \neq 0$ using an (asymptotic) t test.

If we reject the null we have found evidence that the difference between OLS and 2SLS is statistically significant and the OLS suffers from endogeneity.

11.4 Overview of chapter

In economics many behavioural relations (such as the demand and supply of labour) are part of a system of equations, where each equation in the system has a causal, *ceteris paribus*, interpretation of its own. Variables that are jointly determined in this system are endogenous variables, and the other variables, which are determined outside the system, are exogenous variables. These behavioural equations are also called our structural equations.

The simultaneity (joint determinacy) induces a correlation between the endogenous explanatory variable(s) and the error in these structural equations. This will give rise to inconsistent estimators when OLS is applied to these behavioural relations, a problem called simultaneity bias. We derived this result, which uses the reduced form equations for our endogenous explanatory variables.

To solve the problem of simultaneity bias, we used 2SLS (IV) which involved the search for instruments which necessitates exclusion restrictions on our structural equations. If we lack the required instruments for a particular equation, we call the equation unidentified. This would be the case, for example, when there are no observed demand shifters in which case the supply equation is not identified, as the data will not allow us to identify (estimate) the supply equation. When we have more instruments than we need, we call our equation overidentified. An equation is just identified when we have exactly as many instruments as we need.

We showed that simultaneous equation models provide a setting where the discussion of suitable instruments to deal with endogeneity is quite natural.

11.4.1 Key terms and concepts

- Simultaneous equations model (SEM)

- Endogenous and exogenous variables
- Structural equation
- Simultaneity
- Simultaneity bias
- Identified equation
- Exclusion restrictions
- Just identified equation
- Overidentified equation
- Unidentified equation
- Order condition
- Rank condition
- Reduced form equation

11.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the concept of simultaneity in regression
- understand the need for supply (demand) shifters to identify the demand (supply) equation graphically
- explain the concepts of a structural equation and a reduced form equation in the simultaneous equations framework
- distinguish endogenous variables (determined within the SEM) and exogenous variables (determined outside the model) in the simultaneous equations framework
- explain the problem of simultaneity bias when OLS is applied to a single structural equation in isolation
- derive the correlation between the jointly determined explanatory variable and the structural equation error
- derive an expression for the large-sample simultaneity bias in the slope coefficients when OLS is used to fit a simple regression in a simultaneous equations model
- discuss the order and rank condition of identification in a two-equation system
- explain what is meant by a just identified, unidentified, and overidentified equation
- discuss 2SLS estimation of the structural form equations in SEM.

11.5 Test your knowledge and understanding

11.5.1 Additional examination based exercises

11.1E The following model jointly determines monthly child support payments and monthly visitation rights for divorced couples (parent 1 and parent 2) with children:

$$\textit{support} = \alpha_1 + \alpha_2\ \textit{visits} + \alpha_3\ \textit{inc1} + \alpha_4\ \textit{remarr1} + \alpha_5 \textit{dist} + \varepsilon_1 \tag{1.1}$$

$$\textit{visits} = \beta_1 + \beta_2\ \textit{support} + \beta_3 \textit{remarr2} + \beta_4\ \textit{dist} + \varepsilon_2. \tag{1.2}$$

It is assumed that following the divorce, children live with parent 2 and parent 1 pays child support. The errors have zero mean and constant variance and $Cov(\varepsilon_1, \varepsilon_2) = \sigma_{12} \neq 0$.

(1.1) is the 'reaction function' of parent 1: it describes the amount of child support paid by parent 1 for any given level of visitation rights; *inc1* is the income of parent 1, *remarr1* indicates whether parent1 is remarried (1 = yes, 0 = no), and *dist* measures the distance in miles between the current residence of parent 1 and parent 2.

(1.2) is the 'reaction function' of parent 2: it describes visitation rights for a given amount of child support; *remarr2* indicates whether parent 2 is remarried (1 = yes, 0 = no).

We assume that *inc1*, *remarr1*, *remarr2*, and *dist* are exogenous explanatory variables.

(a) Explain the concepts of endogenous versus exogenous explanatory variables and show that *support* is an endogenous variable in (1.2). You are expected to derive the reduced form for *support* to fully answer this question.

(b) Examine for each structural (behavioural) equation whether the equation is over-identified, exactly identified or unidentified. Provide clear arguments for your answer.

(c) Your friend suggests you should implement the IV estimator to estimate the β parameters consistently. He tells you to use *inc1* as an instrument for *support.* Provide a critical discussion of this suggestion.

11.2E It is postulated that a reasonable demand–supply model for the wine industry in Australia, under the market-clearing assumption, would be given by:

$$Q_t = \alpha_0 + \alpha_1 P_t^w + \alpha_2 P_t^b + \alpha_3 Y_t + \alpha_4 A_t + u_{1t} \qquad \text{(demand)}$$

$$Q_t = \beta_0 + \beta_1 P_t^w + \beta_2 S_t + u_{2t} \qquad \text{(supply)}$$

where Q_t = real per capita consumption of wine, P_t^w = price of wine relative to CPI, P_t^b = price of beer relative to CPI, Y_t = real per capita disposable income, A_t = real per capita advertising expenditure, and S_t = storage cost at time t. CPI is the Consumer Price Index at time t. The endogenous variables in this model are Q and P^w, and the exogenous variables are P^b, Y, A and S. The variance of u_{1t} and u_{2t} are, respectively σ_1^2, and σ_2^2, and $Cov(u_{1t}, u_{2t}) = \sigma_{12} \neq 0$.

(a) Derive the reduced form for P_t^w.

(b) The OLS estimation of the demand function, based on annual data from 1955–1975 ($T = 20$), gave the following results (all variables are in logs and figures in parentheses are t ratios):

$$\widehat{Q}_t = \underset{(-6.04)}{-23.651} + \underset{(4.0)}{1.158}P_t^w \underset{(-0.45)}{-0.275}\, P_t^b \underset{(4.5)}{+3.212}\, Y_t \underset{(-1.3)}{-0.603}\, A_t.$$

All the coefficients except that of Y have the wrong signs. The coefficient of P^w (price elasticity of demand, α_1) not only has the wrong sign but also appears significant.

Explain why the OLS parameter estimator may give rise to these counterintuitive results. You are expected to use your results in part (a) to support your answer.

(c) The supply equation is overidentified. Clearly explain this terminology. What distinguishes overidentification from just identified and unidentified? Provide one set of assumptions that would render the supply equation exactly identified.

(d) Discuss how you should estimate the supply equation in light of the overidentification. What are the benefits of overidentification?

11.3E Let us consider the demand for fish. Using 97 daily price (*avgprc*) and quantity (*totqty*) observations on fish prices at the Fulton Fish Market in Manhattan, the following results were obtained by OLS:

$$\begin{aligned}\log(\widehat{totqty_t}) &= \underset{(.163)}{8.244} - \underset{(.176)}{.425}\log(avgprc_t) - \underset{(.226)}{.311} mon_t - \underset{(.223)}{.683}\, tues_t \\ &\quad - \underset{(.220)}{.533}\, wed_t + \underset{(.220)}{.067}\, thurs_t \\ &\quad R^2 = .217,\ n = 97. \qquad (3.1)\end{aligned}$$

The equation allows demand to differ across the days of the week, and Friday is the excluded dummy variable. The standard errors are in parentheses.

(a) Interpret the coefficient of log(*avgprc*) and discuss whether it is significant.

(b) It is commonly thought that prices are jointly determined with quantity in equilibrium where demand equals supply. What are the consequences of this simultaneity for the properties of the OLS estimator?

(c) The variables $wave2_t$ and $wave3_t$ are measures of ocean wave heights over the past several days. In view of your answer in part (b), what two assumptions do we need to make in order to use $wave2_t$ and $wave3_t$ as instruments for $\log(avgprc_t)$ in estimating the demand equation? Discuss whether these assumptions are reasonable.

(d) Below we report two sets of OLS regression results:

$$\widehat{\log(totprc_t)} = \underset{(.144)}{-1.022} + \underset{(.021)}{.094}\, wave2_t + \underset{(.020)}{.053}\, wave3_t - \underset{(.113)}{.012}\, mon_t - \underset{(.112)}{.009}\, tues_t$$

$$- \underset{(.112)}{.051}\, wed_t + \underset{(.111)}{.124}\, thurs_t$$

$$R^2 = .3041,\ SSR = 10.934,\ n = 97 \tag{3.2}$$

$$\widehat{\log(totprc_t)} = \underset{(.092)}{-.276} + \underset{(.134)}{.012}\, mon_t - \underset{(.132)}{.007}\, tues_t + \underset{(.130)}{.046}\, wed_t + \underset{(.130)}{.095}\, thurs_t$$

$$R^2 = .0088,\ SSR = 15.576,\ n = 97. \tag{3.3}$$

Use these results to test whether $wave2_t$ and $wave3_t$ are jointly significant. Clearly specify, the null and alternative hypotheses, the test statistic with its distribution under the null, and the rejection rule as well as any assumptions you make use of. How is your finding related to your answer in part (c)?

(e) The following IV results were obtained in R:

$$\widehat{\log(totqty_t)} = \underset{(.182)}{8.164} - \underset{(.327)}{.815}\log(avgprc_t) - \underset{(.229)}{.307}\, mon_t - \underset{(.226)}{.685}\, tues_t$$

$$- \underset{(.223)}{.521}\, wed_t + \underset{(.225)}{.095}\, thurs_t. \tag{3.4}$$

Discuss how these results can be obtained using two stage least squares (2SLS).

11.5.2 Solutions to additional examination based exercises

11.1E (a) We say that an explanatory variable is endogenous if it is correlated with the error term, $Cov(x_i, u_i) \neq 0$, and exogenous if it is not correlated with the error term, $Cov(x_i, u_i) = 0$. Note as $E(u_i) = 0$, we can replace $Cov(x_i, u_i)$ by $E(x_i u_i)$ in these statements as they are the same as:

$$Cov(x_i, u_i) = E(x_i u_i) - E(x_i)\, E(u_i).$$

A statement which says that explanatory variables that are jointly determined in a SEM are endogenous and explanatory variables that are determined outside the SEM model are exogenous is fine, but relating the issue of endogeneity to the correlatedness with the error is more to the point.

To show that *support* is an endogenous variable in (1.2), we need to obtain the reduced form for *support* (express *support* in terms of exogenous variables and the structural form errors only). Let us plug (1.2) into (1.1) first:

$$support = \alpha_1 + \alpha_2(\beta_1 + \beta_2 support + \beta_3 remarr2 + \beta_4 dist + \varepsilon_2)$$

$$+ \alpha_3 inc1 + \alpha_4 remarr1 + \alpha_5\ dist + \varepsilon_1.$$

We now need to rearrange, to yield:

$$(1 - \alpha_2\beta_2) support = \alpha_1 + \alpha_2\beta_1 + \alpha_2\beta_3 remarr2 + (\alpha_2\beta_4 + \alpha_4) dist + \alpha_2\varepsilon_2$$

$$+ \alpha_3 inc1 + \alpha_4 remarr1 + \varepsilon_1.$$

Assuming $\alpha_2\beta_2 \neq 1$, we can divide the equation to yield the reduced form:

$$\begin{aligned} \textit{support} = {} & \frac{\alpha_1 + \alpha_2\beta_1}{(1-\alpha_2\beta_2)} + \frac{\alpha_2\beta_3}{(1-\alpha_2\beta_2)}\textit{remarr2} + \frac{(\alpha_2\beta_4 + \alpha_4)}{(1-\alpha_2\beta_2)}\textit{dist} \\ & + \frac{\alpha_3}{(1-\alpha_2\beta_2)}\textit{inc1} + \frac{\alpha_4}{(1-\alpha_2\beta_2)}\textit{remarr1} + \frac{\varepsilon_1 + \alpha_2\varepsilon_2}{(1-\alpha_2\beta_2)} \end{aligned}$$

or:

$$\textit{support} = \pi_1 + \pi_2\ \textit{remarr2} + \pi_3\ \textit{dist} + \pi_4\ \textit{inc1} + \pi_5\ \textit{remarr1} + v_s$$

where $v_s = \frac{1}{1-\alpha_2\beta_2}(\varepsilon_1 + \alpha_2\varepsilon_2)$ is the reduced form error. Using this reduced form, we can now show that:

$$Cov(\textit{support}, \varepsilon_2) = Cov(v_s, \varepsilon_2) = \frac{1}{1-\alpha_2\beta_2}(\sigma_{12} + \alpha_2\sigma_2^2) \neq 0$$

where the first equality uses the exogeneity of the explanatory variables in the reduced form expression. Hence as $Cov(\textit{support}, \varepsilon_2) \neq 0$, *support* is endogenous.

(b) The first equation is *exactly identified* since there is one exogenous variable (*remarr2*) (excluded in (1.1)) available as an instrument for the endogenous regressor (*visits*) (rank condition is satisfied as long as $\beta_3 \neq 0$ in (1.2)).

The second equation is *overidentified* since there are two exogenous variables (*inc1* and *remarr1*) (excluded in (1.2)) available as instruments to the endogenous regressor (*support*) (if either α_3 or α_4 in (1.1) are zero, the equation is just identified, if both are zero, the equation is not identified).

(c) The IV approach suggested by the friend would yield consistent estimates for the β parameters since *inc1* is a suitable instrument (exogenous and correlated with support (assuming $\alpha_3 \neq 0$)). However, given this equation is overidentified, we could obtain more efficient estimates by using a two-stage least squares approach in which *both remarr1* and *inc1* are included as instruments.

11.2E (a) We need to combine the demand curve and supply curve and set quantity of demand equal to quantity of supply:

$$\alpha_0 + \alpha_1 P_t^w + \alpha_2 P_t^b + \alpha_3 Y_t + \alpha_4 A_t + u_t = \beta_0 + \beta_1 P_t^w + \beta_3 S_t + v_t.$$

We then need to rewrite this to get an expression of P_t^w in terms of the exogenous regressors and errors only:

$$\alpha_0 + \alpha_1 P_t^w + \alpha_2 P_t^b + \alpha_3 Y_t + \alpha_4 A_t + u_t = \beta_0 + \beta_1 P_t^w + \beta_3 S_t + v_t.$$

Rearranging yields:

$$(\alpha_1 - \beta_1)P_t^w = \beta_0 - \alpha_0 + \beta_3 S_t - \alpha_2 P_t^b - \alpha_3 Y_t - \alpha_4 A_t + v_t - u_t.$$

As $\alpha_1 \neq \beta_1$ (demand/supply slopes different) we can then obtain:

$$P_t^w = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \tfrac{\beta_3}{\alpha_1-\beta_1} S_t - \frac{\alpha_2}{\alpha_1 - \beta_1}P_t^b - \frac{\alpha_3}{\alpha_1 - \beta_1}Y_t - \frac{\alpha_4}{\alpha_1 - \beta_1}A_t + \frac{v_t - u_t}{\alpha_1 - \beta_1}.$$

The reduced form equation, therefore, equals:

$$P_t^w = \pi_0 + \pi_1 S_t + \pi_2 P_t^b + \pi_3 Y_t + \pi_4 A_t + V_t$$

where π_j are our reduced form parameters and $V_t = \frac{v_t - u_t}{\alpha_1 - \beta_1}$ is the reduced form error.

(b) The key reason that OLS gives rise to these counterintuitive results is due to the endogeneity of P_t^w which renders our OLS inconsistent and may result in parameter estimates of the wrong sign.

Q_t and P_t^w are jointly determined in this SEM, and this results in an endogeneity problem as $Cov(P_t^w, u_t) \neq 0$.

Using the result from (a), we have:

$$Cov(P_t^w, u_t) = Cov(\pi_0 + \pi_1 S_t + \pi_2 P_t^b + \pi_3 Y_t + \pi_4 A_t + V_t, u_t)$$

with $V_t = (v_t - u_t)/(\alpha_1 - \beta_1)$. As we are told that S_t, P_t^b, Y_t, and A_t are exogenous, we get:

$$Cov(P_t^w, u_t) = Cov\left(\frac{v_t - u_t}{\alpha_1 - \beta_1}, u_t\right) = \frac{1}{\alpha_1 - \beta_1}(\sigma_{vu} - \sigma_u^2) \neq 0.$$

(c) The supply equation is overidentified, since here we have three instrumental variables P_t^b, Y_t and A_t to deal with our endogenous variable P_t^w.

If, in the demand equation we have, say, $\alpha_4 \neq 0$, $\alpha_2 = 0$ and $\alpha_3 = 0$, then the supply equation becomes exactly identified. In that case we would only have a single instrument, A_t (advertising), for the endogenous variable P_t^w.

If we do not have any instrument our equation becomes unidentified and we would no longer be able to estimate the equation.

(d) As the supply equation is overidentified we should use 2SLS.

Step 1: Estimate the reduced form of P_t^w and get the fitted value $\widehat{P}_t^w$:

$$\widehat{P}_t^w = \widehat{\pi}_0 + \widehat{\pi}_1 S_t + \widehat{\pi}_2 P_t^b + \widehat{\pi}_3 Y_t + \widehat{\pi}_4 A_t.$$

Step 2: Use $\widehat{P}_t^w$ in place of P_t^w to estimate the supply equation by OLS, that is, estimate:

$$Q_t = \beta_0 + \beta_1 \widehat{P}_t^w + \beta_2 S_t + error_t.$$

The benefit of overidentification is that of improved efficiency of our IV estimator (2SLS chooses the optimal instrument).

11.3E (a) We want to interpret this parameter as the price elasticity of demand after controlling for days of the week differences. If the price increases by 1% then quantity demanded decreases by .425%.

We need to test $H_0 : \beta_{\log(avgprc)} = 0$ against $H_1 : \beta_{\log(avgprc)} \neq 0$. Under the CLM assumptions (MLR.1 to MLR.6), we can use the t statistic:

$$\frac{\widehat{\beta}_{\log(avgprc)}}{se(\widehat{\beta}_{\log(avgprc)})} \sim t_{n-6} \quad \text{under } H_0.$$

At the 5% level of significance the critical value is 1.96. Given the realisation of our test statistic is $-.425/.176 = -2.415$ we conclude that it is significant (i.e. we reject H_0).

(b) This is a problem of endogeneity. The simultaneity of quantity and prices will induce a correlation between prices ($\log(avgprc)$) and the error term in the demand (and supply) equation. This endogeneity will lead to inconsistent parameter estimates when OLS is used to estimate the demand equation. It is also called 'simultaneity bias'.

(c) To estimate the demand equation, we need at least one exogenous variable, not in the demand equation, that appears in the supply equation (supply shifter).

We need to assume that these variables can properly be excluded from the demand equation and are uncorrelated with the error in the demand equation. **(Exclusion and validity requirement.)** These assumptions may not be entirely reasonable; wave heights are determined partly by weather, and demand at a local fish market could depend on weather too.

The requirement that at least one of $wave2_t$ and $wave3_t$ appears in the supply equation ensures that we have a correlation of our instruments with $\log(avgrpc)$. **(Relevance requirement.)** There is indirect evidence of this in part (d), where we will show that they are jointly significant in the reduced form for $\log(avgprc_t)$.

(d) We test the joint hypothesis: $H_0 : \beta_{wave2} = \beta_{wave3} = 0$ against $H_1 : \beta_{wave2} \neq 0$ and/or $\beta_{wave3} \neq 0$.

The F test is obtained as:

$$F = \frac{(RSS_r - SSR_{ur})/2}{SSR_{ur}/(97-7)} = \frac{(15.576 - 10.934)/2}{10.934/90} = 19.0.$$

Equivalently, the following form could be used to obtain the the F statistic:

$$\frac{(R_u^2 - R_r^2)/2}{(1 - R_u^2)/90}.$$

Both tests evaluate the loss in fit of imposing the restrictions we are testing.

Under the CLM assumptions (MLR.1 to MLR.6) this gives us an $F_{2,\,90}$ random variable under the null. For any reasonable level of significance we will want to reject the null. We conclude that our test result ensures that our IVs indeed are relevant.

(e) Discussion of 2SLS:

- Step 1: Estimate the reduced form for $\log(avgprc)$:

$$\log(avgprc_t) = \delta_0 + \delta_1 mon_t + \cdots + \delta_4 thurs_t + \delta_5 wave2_t + \delta_6 wave3_t + e_t.$$

We obtain the fitted values of this regression: $\widehat{\log(avgprc_t)}$. All exogenous variables have to be included here!

- Step 2: Use the fitted values for $\log(avgprc)$ to estimate the demand equation:

$$\begin{aligned}\log(totqty_t) = \beta_0 + \beta_1 \log\widehat{(avgprc_t)} + \beta_2 mon_t + \beta_3 tues_t \\ + \beta_4 wed_t + \beta_5 thurs_t + u_t. \qquad (3.5)\end{aligned}$$

  The parameter estimates of this regression are our 2SLS estimators.

Chapter 12

Binary response models and maximum likelihood

12.1 Introduction

12.1.1 Aims of the chapter

In this chapter we discuss a special class of models where the dependent variable is a binary variable. These models are also called binary response, or binary choice models. These models have many empirical applications and can be used for instance when trying to explain the decision of a married woman to participate in the labour market (1 = yes) or not (0 = no).

We discuss the linear probability model (LPM) and the logit and probit models. The LPM permits us to estimate the parameters by OLS. To estimate the logit and probit model (examples of nonlinear models) we will use the maximum likelihood estimator (MLE). We will review the fundamentals of MLE and discuss the attractive properties ML estimators exhibit before describing the method for logit and probit models. We also discuss how to conduct hypothesis tests when using logit and probit models.

12.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- motivate the interest and use of binary response models
- describe the linear probability model (LPM) and discuss the interpretation of parameters
- discuss the benefits/disadvantages of the LPM
- discuss heteroskedasticity-robust inference for the LPM
- explain the benefits of using a nonlinear model for binary response models
- understand the principles of maximum likelihood estimation (MLE)
- discuss the properties of ML estimators
- derive the MLE from first principles in simple models
- specify the logit model and the log-likelihood function for its estimation by MLE

- specify the probit model and the log-likelihood function for its estimation by MLE
- understand regression output provided by logit/probit models
- understand how to obtain partial effects in logit and probit models
- conduct hypothesis tests using logit and probit models (in particular use the z test and the likelihood ratio (LR) test)
- discuss a goodness of fit measure for binary response models
- use statistical software to estimate binary choice models (linear and nonlinear) using real-world data.

12.1.3 Essential reading

Readings: Wooldridge, Sections 7.5, 8.5, 17.1, and Appendices 17A and C4.b.

12.1.4 Synopsis of this chapter

Section 7.5 in Wooldridge discusses the consequences of using ordinary least squares (OLS) in settings where the dependent variable, y, is a binary random variable (i.e. where y takes only the values zero and one). As the conditional expectation $E(y \,|\, x)$ in this case should be interpreted as a conditional probability $Pr(y = 1 \,|\, x)$, we call our multiple regression model also the linear probability model (LPM). The partial effects in the LPM describe the effects independent variables have on the probability of $y = 1$ holding all other factors constant.

While the LPM is easy to estimate and interpret, the linear probability model has various disadvantages. First, the LPM exhibits heteroskedasticity (a violation of our Gauss–Markov assumptions). As discussed in Section 8.5 in Wooldridge, this requires the use of heteroskedasticity-robust standard errors for statistical inference. Second, the LPM gives rise to the possibility that we obtain predicted values smaller than zero or bigger than one. In light of the fact that we need to interpret these predicted values as predicted probabilities such a result is undesirable. To deal with this drawback of the LPM, we argue for the benefit of using a nonlinear model for binary response models and consider the logit and probit models in Section 17.1 in Wooldridge.

After discussing the fundamentals of the maximum likelihood estimator (MLE) (see Appendix C4.b), we discuss the MLE for the logit and probit models (this includes a discussion of Appendix 17A). You will learn how to obtain predicted probabilities and, importantly, how to obtain partial effects for the logit and probit models. Finally, we discuss how to conduct hypothesis testing in this framework (MLE), which will include a discussion of the Likelihood Ratio test.

12.2 Content of chapter

In many empirical examples, we have incorporated binary variables as regressors (also known as dummy variables). For instance, we used gender and race dummies in wage regressions to test whether there are gender and ethnic differences in wages. We also incorporated interactions involving dummy variables – for example, we considered the interactions between gender and education in our wage regression to allow us to detect whether there are gender differences in the returns to education. In this chapter, we consider the use of **binary variables as our dependent variable**. This setting is referred to as the **binary response model**.

In binary response models, we attempt to explain a qualitative event. Individuals and/or firms are asked to choose between two alternatives and we denote one of the outcomes as $y = 1$ and the other as $y = 0$. There are many empirical applications of this nature in economics as discussed in Wooldridge.

12.2.1 A binary dependent variable: The linear regression model

Read: Wooldridge, Section 7.5.

Here, we consider the implication of using the multiple linear regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

when the dependent variable is a binary variable. We continue to assume $E(u \,|\, x) = 0$ (needed for causal interpretation of our parameters), so that:

$$E(y \,|\, \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

A key observation we need to make when the dependent variable y is a discrete random variable that only takes the values zero and one, is that we need to interpret the conditional expectation $E(y \,|\, x)$ as a probability. Using properties of discrete random variables (see Wooldridge, Appendix B.1a):

$$\begin{aligned} E(y \,|\, \mathbf{x}) &= 0 \times Pr(y = 0 \,|\, \mathbf{x}) + 1 \times Pr(y = 1 \,|\, \mathbf{x}) \\ &= Pr(y = 1 \,|\, \mathbf{x}). \end{aligned}$$

Hence, when we apply the multiple linear regression model to a binary dependent variable we obtain a **linear probability model (LPM)**:

$$Pr(y = 1 \,|\, \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

This terminology is due to the fact that the probability that $y = 1$ is assumed to be *linear in the parameters*. The parameters of the LPM have a nice (easy) causal interpretation: β_j represents the effect a unit change in x_j has on the probability that $y = 1$ holding everything else fixed.

We can easily estimate the parameters of the LPM by running an OLS regression of y on the explanatory variables, and predicted values from this regression need to be interpreted as predicted probabilities:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \cdots + \widehat{\beta}_k x_{ki} = Pr(\widehat{y_i = 1} \,|\, \mathbf{x}_i).$$

Empirical applications of this are provided in Wooldridge and you need to be able to discuss the results, clearly relating them to changes in the probability of the event that $y = 1$.

1. Let us highlight some points with reference to the classic example of labour market participation (using the data from Mroz, 1987), Wooldridge reports:

$$\begin{aligned}\widehat{inlf} = &\underset{(.154)}{.586} - \underset{(.0014)}{0.0034}\, nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper \\ &- \underset{(.00018)}{.00060}\, exper^2 - \underset{(.002)}{0.016}\, age - \underset{(.034)}{.262}\, kidslt6 + \underset{(.013)}{.013}\, kidsge6 \\ &n = 753,\ R^2 = .264\end{aligned}$$

where *inlf* is a binary dependent variable which indicates whether a married woman was in the labour force during 1975 (1 = yes) or not (1 = no).

   - A 35-year old woman with 16 years of education, 7 years of experience and 1 young child, with no other sources of income ($nwifeinc = 0$) has a **predicted probability of being in the labour force** of:

$$\widehat{inlf} = .586 + .038 \times 16 + .039 \times 7 - .00060 \times 49 - .016 \times 35 - .262 = .616$$

   where $\widehat{inlf} = \widehat{Pr}(inlf = 1 \,|\, age = 35, educ = 7, \ldots)$.

   - Let us focus here only on two parameters $\widehat{\beta}_{educ}$ and $\widehat{\beta}_{kidslt6}$ (see Wooldridge for a discussion of the parameters).

   $\widehat{\beta}_{educ}$: *Ceteris paribus*, each *additional year of education* increases the predicted probability of a woman being in the labour market by .038 units, or **3.8 percentage points**. For the 35-year old woman described above, it increases her predicted probability from 61.6% to 65.4%. (Careful: this is *not* the same as a 3.8 percentage increase!)

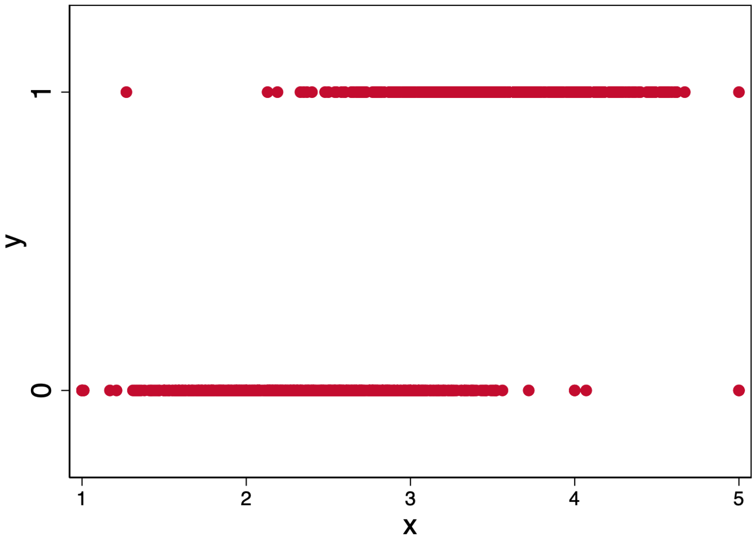
   $\widehat{\beta}_{kidslt6}$: *Ceteris paribus*, each *additional young child* (< 6 years) reduces the predicted probability of a woman being in the labour market by .262 units, that is a 26.2 percentage points drop.

Various limitations of the LPM are discussed in Wooldridge. Two of them are directly related to the fact that we need to interpret the fitted values of this regression as predicted probabilities. (1) The **fitted values for our LPM are not guaranteed to lie between zero and one**. Due to their interpretation this is clearly unsatisfactory. (2) The **estimated partial effects are constant** throughout the range of values the explanatory variables can take. In order to ensure that the predicted values represent a probability (and are restricted to lie between zero and one), *eventually they must have a diminishing effect.*

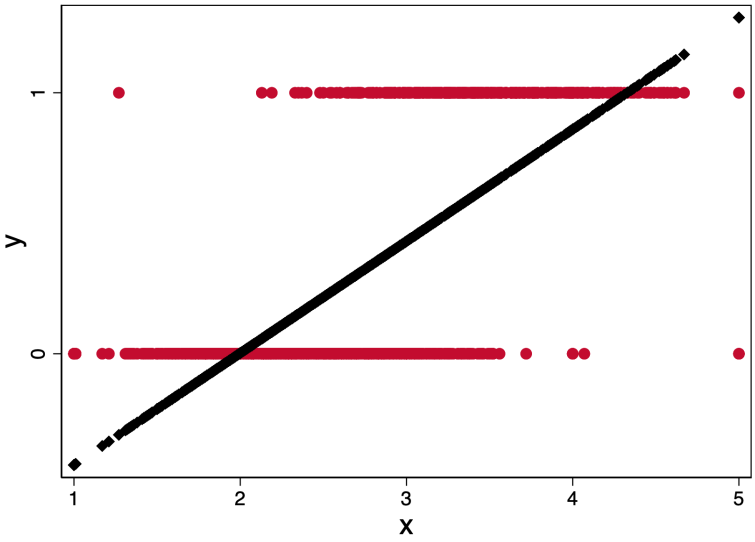
To give a graphic exposition of these points let us consider the simple bivariate LPM model setting. This discussion also reinforces the special nature of binary dependent variables. Let:

$$y = \beta_0 + \beta_1 x + u.$$

The following graph displays a scatter graph of observations. Observe that y only can take two values 0 and 1. The scatter graph suggests a positive relation between x and y.



The fitted regression line (LPM regression line) is a straight line in this graph that minimises the residual sum of squares.



For large values of x, the predicted value may exceed one, whereas for small values of x, the predicted value may lie below zero. This is a result of the fact that the LPM assumes the marginal effect of x on y is constant.

A further drawback of the LPM is that (3) the **LPM exhibits heteroskedasticity** (which invalidates the OLS standard errors) due to the fact that:

$$Var(y \,|\, \mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x}))$$

where:

$$p(\mathbf{x}) \equiv Pr(y = 1 \,|\, \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k.$$

To see this result, we recall $Var(y \mid \mathbf{x}) = E(y^2 \mid \mathbf{x}) - E(y \mid \mathbf{x})^2$. Using the properties of discrete random variables, we can obtain the result provided as:

$$E(y \mid \mathbf{x}) = 0 \times Pr(y = 0 \mid \mathbf{x}) + 1 \times Pr(y = 1 \mid \mathbf{x}) = p(\mathbf{x})$$

$$E(y^2 \mid \mathbf{x}) = 0^2 \times Pr(y = 0 \mid \mathbf{x}) + 1^2 \times Pr(y = 1 \mid \mathbf{x}) = p(\mathbf{x})$$

$$Var(y \mid \mathbf{x}) = p(\mathbf{x}) - p(\mathbf{x})^2 = p(\mathbf{x})(1 - p(\mathbf{x})).$$

Using the properties of the variance operator, with:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

we should note that the conditional variance of y given $\mathbf{x}$ is identical to the conditional variance of u:

$$Var(u \mid \mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x})).$$

Hence, unless all slopes β_j, for $j = 1, 2, \ldots, k$, are zero, the conditional variance of u will depend on $\mathbf{x}$ and exhibit heteroskedasticity. A violation of MLR.5!

Caution: The results provided for the empirical applications of the LPM report the usual standard errors alongside the parameter estimates. As the presence of heteroskedasticity invalidates the standard errors, any discussion of the significance of parameters based on them in this section should be read with caution.

Activity 12.1 Take the MCQs related to this section on the VLE to test your understanding.

Activity 12.2 Exercise 7.7 from Wooldridge.

12.2.2 The linear probability model revisited

Read: Wooldridge, Section 8.5.

We have discussed that the multiple regression model when y is a binary variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \quad E(u \mid \mathbf{x}) = 0$$

known as the LPM, suffers from heteroskedasticity as $Var(u \mid \mathbf{x})$ depends on $\mathbf{x}$. When using the LPM, therefore, we should recall the consequences of the presence of heteroskedasticity for the multiple regression model:

1. OLS will still give unbiased and consistent parameter estimates.
2. OLS will no longer be efficient (violation of Gauss–Markov theorem).
3. We will need to correct the standard errors when using OLS. Without it we cannot trust our usual t and F statistics, not even in large samples.

Heteroskedastic-robust estimation of LPM

The most common approach to deal with heteroskedasticity in the LPM is to use heteroskedasticity-robust standard errors. Using robust standard errors (which are easy to obtain using statistical packages) we can conduct inference and construct confidence intervals as usual.

Let us highlight some points with reference to the classic example of labour market participation (using the data from Mroz, 1987). Wooldridge reports:

$$\begin{aligned}\widehat{inlf} &= \underset{(.151)}{.586} - \underset{(.0015)}{0.0034}\, nwifeinc + \underset{(.007)}{.038}\, educ + \underset{(.006)}{.039}\, exper \\ &\quad - \underset{(.00019)}{.00060}\, exper^2 - \underset{(.002)}{0.016}\, age - \underset{(.032)}{.262}\, kidslt6 + \underset{(.013)}{.013}\, kidsge6 \\ &\quad n = 753,\ R^2 = .264\end{aligned}$$

where numbers in brackets are the heteroskedasticity-robust standard errors.

The parameter estimates are the same as before (we have used OLS again). The *only difference is that robust standard errors are reported.* They are not that different from the standard errors we had before.

The **robust** 95% confidence interval for $\beta_{nwifeinc}$ is given by:

$$\left[\widehat{\beta}_{nwifeinc} - 1.96se(\widehat{\beta}_{nwifeinc}), \widehat{\beta}_{nwifeinc} + 1.96se(\widehat{\beta}_{nwifeinc})\right] = [-.0063, -.0005].$$

The critical value 1.96 used is obtained from $Normal(0,1)$ (large sample). As zero does not lie in this interval, we reject the null that $H_0 : \beta_{nwifeinc} = 0$ against $H_1 : \beta_{nwifeinc} \neq 0$ at the 5% level of significance.

To test $H_0 : \beta_{educ} = 0$ against $H_1 : \beta_{educ} \neq 0$, we now use the **robust *t* statistic**:

$$t_{\widehat{\beta}_{educ}} = \frac{\widehat{\beta}_{educ}}{se(\widehat{\beta}_{educ})} = \frac{.038}{.007} = 5.426.$$

We get a clear strong rejection of H_0. As we use robust standard errors, we rely on asymptotic $Normal(0,1)$ critical values.

To test $H_0 : \beta_{exper} = 0$ and $\beta_{exper^2} = 0$, we now use the **robust *F* statistic**. You are not able to derive the statistic itself. Here $F = 67.17$, which gives strong evidence against H_0. We use critical values given by $F_{2,df}$ where $df = 753 - 8$.

Estimating the LPM by weighted least squares

In order to regain efficiency, we may want to consider using weighted least squares on our LPM:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u, \quad E(u \,|\, \mathbf{x}) = 0. \tag{*}$$

We recall, $Var(u_i \,|\, \mathbf{x_i}) = p(\mathbf{x}_i)(1 - p(\mathbf{x_i}))$, where $p(\mathbf{x}_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$. By multiplying our LPM (*) by $h_i \equiv 1/\sqrt{p(\mathbf{x}_i)(1 - p(\mathbf{x_i}))}$ for each observation i, we are effectively able to remove the problem of heteroskedasticity:

$$h_i y_i = \beta_0 h_i + \beta_1 x_{1i} h_i + \beta_2 x_{2i} h_i + \cdots + \beta_k x_{ki} h_i + u_i h_i \tag{**}$$

as

$$Var(u_i h_i \,|\, \mathbf{x_i}) = h_i^2\, Var(u_i \,|\, \mathbf{x_i}) = 1 \quad \text{i.e. homoskedastic.}$$

As this model (**) satisfies our Gauss–Markov assumptions, estimating this model by OLS (which defines our Weighted Least Squares) will be efficient.

In order to implement this regression, we will need to use estimates of h_i, where:

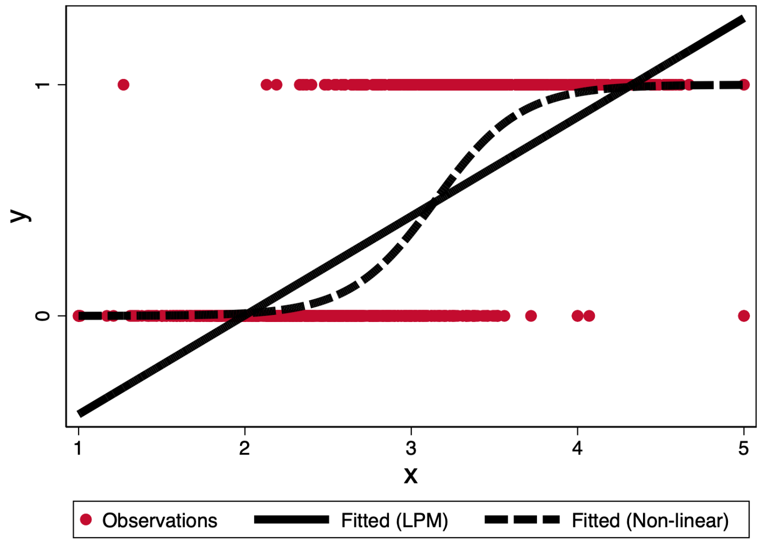
$$\widehat{h}_i = \frac{1}{\sqrt{\widehat{p}(\mathbf{x}_i)(1-\widehat{p}(\mathbf{x_i}))}}$$

instead. As fitted values of the LPM (which define $\widehat{p}_i$) may not lie inside the zero–one range, these weights may not be well defined. The use of weighted least squares for LPM is very uncommon.

12.2.3 Logit and probit models for binary response (part 1)

Read: Wooldridge, Section 17.1a – Specifying logit and probit models.

In binary response models, we have seen that the fitted dependent variable represents a predicted probability which, by its nature, needs to be restricted to lie in the interval $[0, 1]$. As failure to satisfy this requirement is one of the limitations of the LPM, we argue here in favour of using a non-linear model for binary response models.



In particular, we will consider a general model:

$$Pr(y = 1 \,|\, \mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

where the G function takes values on the interval $[0, 1]$ only. This specification ensures that $Pr(y = 1 \,|\, \mathbf{x})$ is always in $[0, 1]$.

An important consequence, we discuss in detail later, is that the β_j parameters do not have the same, easy, interpretation they had in the LPM. To obtain the *ceteris paribus* effect of explanatory variables (continuous) on the probability that $y = 1$, we should consider the partial derivative (using the chain-rule):

$$\frac{\partial Pr(y = 1 \,|\, \mathbf{x})}{\partial x_j} = g(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)\beta_j$$

where $g(z) = \mathrm{d}G(z)/\mathrm{d}z$.

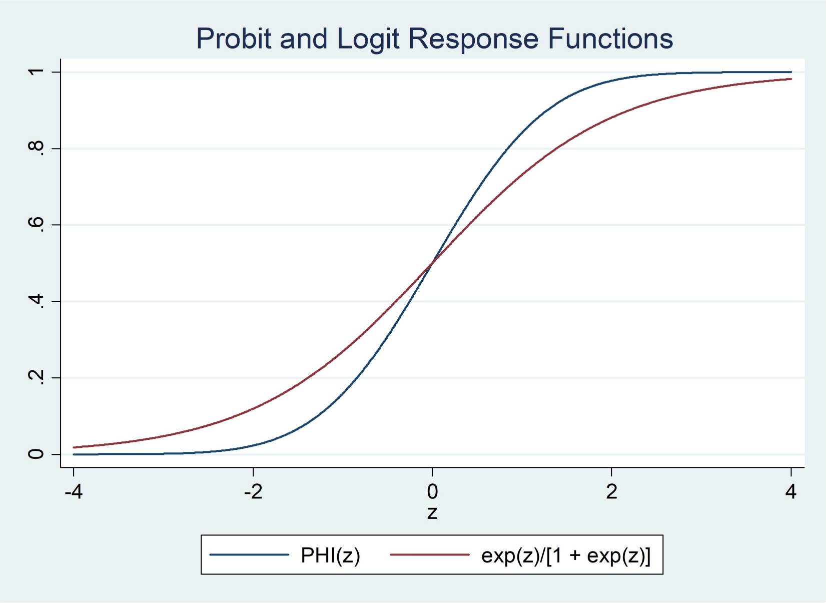
A natural choice for G is to use some **cumulative distribution function (cdf)**. We should recall that the cdf, $G(z)$, is defined as $Pr(Z \leq z)$. It is a non-decreasing function of z. For large values of z, $G(z)$ is close to 1, for small values of z, $G(z)$ is close to 0. Here $z = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$.

The two common choices for G, which define the logit and probit models are the *cdf of the logistic distribution* and the *cdf of the standard normal distribution*, given by:

$$G(z) = \Lambda(z) = \frac{\exp(z)}{1 + \exp(z)} \qquad \text{(logit model)}$$

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) \mathrm{d}u \qquad \text{(probit model)}$$

respectively. Both G functions have similar shapes but the logistic is more spread out, as we can see in the following graph.



Comment: The discussion in Wooldridge which shows that the logit and probit models can be derived from an underlying latent variable model is interesting, but is not examinable.

To estimate the β_j parameters in our non-linear binary response model:

$$Pr(y = 1 \,|\, \mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)$$

we will use the **maximum likelihood estimator** (MLE). The estimator makes use of the fact our binary random variable can be described by the **Bernoulli distribution**:

$$f(y) = p(\mathbf{x})^y (1 - p(\mathbf{x}))^{1-y}, \quad y = 0, 1$$

where $p(\mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$ is the response probability, $Pr(y = 1 \,|\, x)$. Before discussing further details, let us first review the fundamentals of MLEs and discuss the desirable (asymptotic) properties they have.

12.2.4 Maximum likelihood estimation

Read: Wooldridge, Appendix C.4b.

Comment: In this segment, we will use notation that allows us to clearly distinguish between random variables (denoted with capital letters) and their realisations (denoted with small letters). It is noted that this distinction is not made as explicit in the main text of Wooldridge (and the course in general).

Fundamentals of maximum likelihood estimation

Maximum likelihood is a general approach used when estimating parameters. It is an optimisation method which *assumes that we can fully describe the distribution of the data up to some unknown parameters of interest*, θ. The maximum likelihood estimator (MLE) uses a sample of n observations to determine the value of θ that best describes the distribution from which the data was drawn.

- **Intuition:** Say we have a sample of i.i.d. observations drawn from a normal distribution with mean μ and variance σ^2. We recall that the normal distribution has a nice bell-shaped probability density function, centred around the mean and dispersion determined by the variance. The maximum likelihood estimator (estimate) will look at the random sample (realisations) to determine the normal distribution that best represents the distribution from which the random sample was drawn. Hence, the central tendency of our random sample will determine the estimator of μ, while the dispersion of observations in the random sample will determine the estimator of σ^2.

The likelihood function we maximise to obtain the MLE is given by the joint density of the data as a function of the unknown parameters θ. In the simplest case we will consider the setting where we have a random sample $Y_1, Y_2, \ldots, Y_n$, where each Y_i is drawn independently from the same (identical) population whose density function is given by $f(y; \theta)$, that is Y_i is i.i.d. (independent, identically distributed). By independence, the joint density of $Y_1, \ldots, Y_n$ then equals the product of marginals:

$$f(y_1, y_2, \ldots, y_n; \theta) = f(y_1; \theta) f(y_2; \theta) \cdots f(y_n; \theta)$$

where y_i denotes all possible realisations the random variable Y_i can take. The likelihood function is defined in terms of the random variables $Y_1, Y_2, \ldots, Y_n$, which in this case yields:

$$L(\theta; Y_1, \ldots, Y_n) = f(Y_1; \theta) f(Y_2; \theta) \cdots f(Y_n; \theta).$$

The MLE of our parameter is given by that value of θ that will maximise this function, which we may write formally as:

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} L(\theta; Y_1, \ldots, Y_n).$$

The first-order conditions of this optimisation problem, which we may write formally as:

$$\left.\frac{\partial L(\theta; Y_1, \ldots, Y_n)}{\partial \theta}\right|_{\theta=\widehat{\theta}_{MLE}} = 0$$

will give our estimator $\widehat{\theta}_{MLE}$ in terms of the random variables $(Y_1, Y_2, \ldots, Y_n)$. We recall that an **estimator** is a random variable, and an **estimate** is a realisation of this random variable for a given sample $(y_1, \ldots, y_n)$.

Instead of maximising the likelihood function, we typically use the log-likelihood function. Since log is a monotone transformation, the estimator that maximises the log-likelihood function is the same as the estimator that maximises the likelihood function. Using the properties of the log operator, the log-likelihood can be written as the sum of logs:

$$\log L(\theta; Y_1, \ldots, Y_n) = \sum_{i=1}^{n} \log f(Y_i; \theta).$$

Desirable properties of maximum likelihood estimators

Under very general conditions (often labelled the MLE regularity conditions) the estimator has nice asymptotic properties as follows.

- **ML estimator is consistent**:

  $$\text{plim } \widehat{\theta}_{MLE} = \theta.$$

  ML estimators are not always unbiased. Hence, the consistency result is important.

- **ML estimator is asymptotically normal**

  This property is important as it allows us to perform inference: construct (asymptotic) confidence intervals and conduct (asymptotic) hypothesis tests. (ML estimators do not always have an explicit form and their finite sampling distribution may not be known.)

  For large samples the distribution of the ML estimator is well-approximated by a normal distribution centred around the population parameter of interest (consistent):

  $$\widehat{\theta}_{MLE} \overset{\text{a}}{\sim} \textit{Normal}(\theta, \textit{Var}(\widehat{\theta}_{MLE})).$$

- **ML estimator is asymptotically efficient**

  This property is important as it ensures the precision of the ML estimators. (It is related to the concept of the *Cramér–Rao lower bound*, which ensures that there is no other consistent, asymptotic normal estimator that has a lower variance.)

Maximum likelihood estimator: Mathematical fundamentals

You are expected to derive the MLE in simple models from first principles. Mathematically, this requires you to be comfortable with the following mathematical rules (see also Basic mathematical tools in the pre-course material).

Product function satisfies these rules:

- $a^b a^c = a^{b+c}$. For example:

$$\left(\frac{1}{2\pi}\right)^{1/2}\left(\frac{1}{2\pi}\right)^{1/2} = \left(\frac{1}{2\pi}\right).$$

- $(a^b)^c = a^{bc}$.

Natural logarithm (log) function satisfies the rules:

- $\log a^b = b \log a$.
- $\log(ab) = \log a + \log b$. In general:

$$\log(z_1 z_2 \cdots z_n) = \log\left(\prod_{i=1}^{n} z_i\right) = \sum_{i=1}^{n} \log(z_i).$$

- $\log(1/\sigma) = -\log\sigma$ because $\log(1) = 0$.
- $\log(e^a) \equiv \log(\exp(a)) = a$ because $\log(e) = 1$.
- $\mathrm{d}\log(x)/\mathrm{d}x = 1/x$.

Simple example of the maximum likelihood estimator (MLE)

Let us consider a random sample $Y_1, \ldots, Y_n$, where $Y_i = 1$ with probability θ and $Y_i = 0$ with probability $1 - \theta$. Y_i is a discrete random variable, whose probability density function (pdf) is given by the Bernoulli distribution:

$$f(y;\theta) = \theta^y (1-\theta)^{1-y} \quad \text{for } y = 0, 1.$$

Let us derive the MLE of θ from first principles.

Step 1: Construct the joint density.

- Our random sample ensures that $Y_1, \ldots, Y_n$ are independent, which permits us to write the joint density as the product of the marginals, i.e. we have:

$$f(y_1, \ldots, y_n; \theta) = \underbrace{\theta^{y_1}(1-\theta)^{1-y_1}}_{pdf(Y_1)} \cdot \underbrace{\theta^{y_2}(1-\theta)^{1-y_2}}_{pdf(Y_2)} \cdots \underbrace{\theta^{y_n}(1-\theta)^{1-y_n}}_{pdf(Y_n)}.$$

  (Advice: write this expression out in full before simplifying!)

  Using the product rules, we can simplify this as:

$$f(y_1, \ldots, y_n; \theta) = \theta^{\sum\limits_{i=1}^{n} y_i}(1-\theta)^{n-\sum\limits_{i=1}^{n} y_i}.$$

Step 2: Obtain the log-likelihood.

- The likelihood function is defined by the joint density:

$$L(\theta; Y_1, \ldots, Y_n) = \theta^{\sum_{i=1}^{n} Y_i}(1-\theta)^{n-\sum_{i=1}^{n} Y_i}.$$

  Using the rules of the log function, we obtain the log-likelihood function:

$$\begin{aligned}\log L(\theta; Y_1, \ldots, Y_n) &= \log\left(\theta^{\sum_{i=1}^{n} Y_i}\right) + \log\left((1-\theta)^{n-\sum_{i=1}^{n} Y_i}\right) \quad \log(ab) = \log(a) + \log(b) \\ &= \sum_{i=1}^{n} Y_i \log\theta + \left(n - \sum_{i=1}^{n} Y_i\right)\log(1-\theta) \quad \log(a^c) = c\log(a).\end{aligned}$$

Step 3: Obtain $\widehat{\theta}_{MLE}$ by solving FOC.

- We need to obtain the derivative $\mathrm{d}\log L/\mathrm{d}\theta$ for this purpose. Making use of the rules of the log function, and the chain rule, we can obtain:

$$\frac{\mathrm{d}\log L}{\mathrm{d}\theta} = \sum_{i=1}^{n} Y_i \frac{1}{\theta} + \left(n - \sum_{i=1}^{n} Y_i\right)\frac{1}{1-\theta}(-1) \quad \frac{\mathrm{d}\log(\theta)}{\mathrm{d}\theta} = \frac{1}{\theta}.$$

  Observe, by the chain rule we have:

$$\frac{\mathrm{d}\log(1-\theta)}{\mathrm{d}\theta} = \frac{\mathrm{d}\log(1-\theta)}{\mathrm{d}(1-\theta)}\frac{\mathrm{d}(1-\theta)}{\mathrm{d}\theta} = \frac{1}{1-\theta}(-1)$$

- The $\widehat{\theta}_{MLE}$ therefore needs to satisfy:

$$\sum_{i=1}^{n} Y_i \frac{1}{\widehat{\theta}_{MLE}} - \left(n - \sum_{i=1}^{n} Y_i\right)\frac{1}{1-\widehat{\theta}_{MLE}} = 0$$

  or:

$$\frac{\sum_{i=1}^{n} Y_i}{\widehat{\theta}_{MLE}} = \frac{n - \sum_{i=1}^{n} Y_i}{1-\widehat{\theta}_{MLE}}.$$

  To solve for $\widehat{\theta}_{MLE}$ we rewrite this as:

$$(1-\widehat{\theta}_{MLE})\sum_{i=1}^{n} Y_i = \widehat{\theta}_{MLE}\left(n - \sum_{i=1}^{n} Y_i\right).$$

  Hence $\widehat{\theta}_{MLE} = \bar{Y}$.

Activity 12.3 We have just discussed the MLE of θ when we have a random sample $Y_1, \ldots, Y_n$, from the Bernoulli distribution where $\theta = Pr(Y_i = 1)$, for $i = 1, \ldots, n$.

Since θ defines the probability that $Y_i = 1$ we should not be surprised that:

$$\widehat{\theta}_{MLE} = \bar{Y}$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ denotes the proportion of ones in the sample.

Properly supported, answer the following questions:

(i) Is $\widehat{\theta}_{MLE}$ an unbiased estimator of θ? Do all ML estimators satisfy this property?

(ii) Is $\widehat{\theta}_{MLE}$ a consistent estimator of θ?

(iii) Would $1/\widehat{\theta}_{MLE}$ be an unbiased estimator of $1/\theta$?

(iv) Would $1/\widehat{\theta}_{MLE}$ be a consistent estimator of $1/\theta$?

Activity 12.4 Suppose you are given a random sample $X_1, \ldots, X_n$ from the exponential distribution, whose probability density function (pdf) is defined by:

$$f(x;\beta) = \frac{1}{\beta}\exp\left(-\frac{x}{\beta}\right), \quad x > 0,\ \beta > 0.$$

According to this distribution $E(X_i) = \beta$ and $Var(X_i) = \beta^2$ for $i = 1, \ldots, n$.

(i) Show that the maximum likelihood estimator of β is $\bar{X}$, where $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$.

(ii) Is the estimator unbiased and/or consistent? Prove your claims.

12.2.5 Logit and probit models for binary response (part 2)

Wooldridge, Sections 17.1b, 17.1c and 17.1d.

Here we provide details of the MLE for the logit and probit models and discuss how to interpret logit and probit estimates. We also address how to conduct hypothesis testing for logit and probit models. In this section, we no longer explicitly use capital letters to denote random variables (as in the rest of the course).

Recall, the logit and probit model specifies:

$$Pr(y = 1 \mid \mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$$

where y is a binary random variable (taking values of zero and one only),

$\mathbf{x} = \{x_1, \ldots, x_k\}$ and:

$$G(z) = \Lambda(z) = \frac{\exp(z)}{1+\exp(z)} \qquad \text{(logit model)}$$

$$G(z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du. \qquad \text{(probit model)}$$

For notational simplicity, let $\mathbf{x}\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

Maximum likelihood estimation of the probit and logit models

Our ability to use MLE to estimate the β parameters for the probit and logit models lies in the realisation that the (conditional) density of a binary dependent variable y given $\mathbf{x}$ can be described by the Bernoulli distribution:

$$f(y \,|\, x) = G(\mathbf{x}\beta)^y (1 - G(\mathbf{x}\beta))^{1-y}, \quad y = 0, 1$$

where $G(\mathbf{x}\beta)$ denotes $Pr(y = 1 \,|\, \mathbf{x})$.

Assuming random sampling of n observations (y_i, x_i), the **joint (conditional) density** can then be obtained by using the product of the 'marginals':

$$f(y_1, \ldots, y_n \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^{n} f(y_i \,|\, \mathbf{x}_i) = \prod_{i=1}^{n} G(\mathbf{x}_i\beta)^{y_i} (1 - G(\mathbf{x}_i\beta))^{1-y_i}$$

where $\prod_{i=1}^{n} f(y_i \,|\, \mathbf{x}_i)$ is shorthand for $f(y_1 \,|\, \mathbf{x}_1) f(y_2 \,|\, \mathbf{x}_2) \cdots f(y_n \,|\, \mathbf{x}_n)$. This joint (conditional) density defines our likelihood function $L(\beta; y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n)$. Given a random sample, we want to determine the value of β that will maximise this likelihood function.

Instead of maximising the likelihood function, it is more common to use the log-likelihood function which has a very nice interpretable form here, given by:

$$\log L(\beta; y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \sum_{i=1}^{n} \left\{ y_i \log \underbrace{G(\mathbf{x}_i\beta)}_{Pr(y_i=1 \,|\, x_i)} + (1 - y_i) \log(\underbrace{1 - G(\mathbf{x}_i\beta)}_{Pr(y_i=0 \,|\, x_i)}) \right\}.$$

Interpretation: This function reveals that the contribution to the log-likelihood for an observation i for which $y_i = 1$ is given by $\log(Pr(y_i = 1 \,|\, x_i)$ (the other term vanishes), whereas the contribution to the log-likelihood for an observation i for which $y_i = 0$ is given by $\log(Pr(y_i = 0 \,|\, x_i)$. *Intuitively*, this reveals that the value of β that will maximise this log-likelihood function should ensure that $\widehat{Pr}(y_i = 1 \,|\, x_i)$ is large when $y_i = 1$, while $\widehat{Pr}(y_i = 0 \,|\, x_i)$ should be large when $y_i = 0$.

We can derive the log-likelihood from the joint density by using our mathematical fundamentals as shown below:

- As $\log(ab) = \log(a) + \log(b)$, we know:

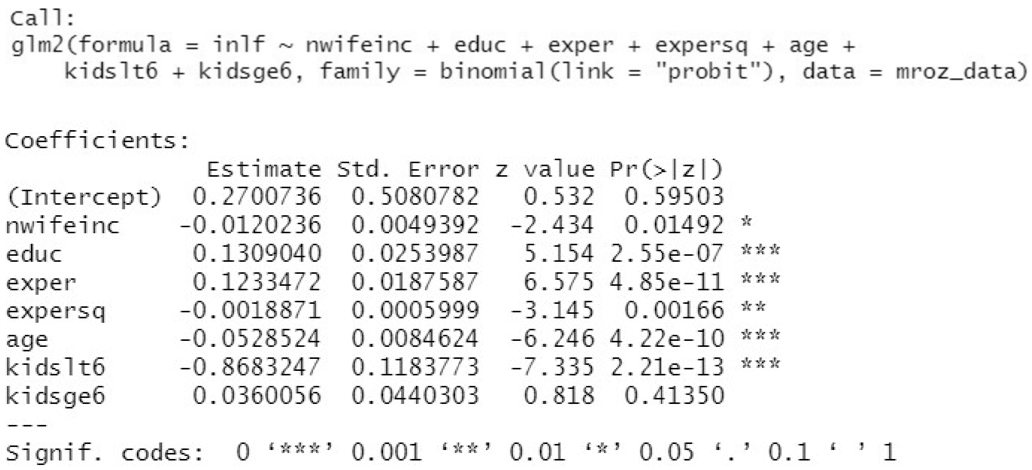
$$\log\left(\prod_{i=1}^{n} f(y_i \,|\, x_i)\right) = \sum_{i=1}^{n} \log(f(y_i \,|\, x_i)).$$

- Using the definition of $f(y_i \,|\, x_i)$, and recalling $\log(a^c) = c\log(a)$ we have:

$$\begin{aligned}\log(G(\mathbf{x}_i\beta)^{y_i}(1-G(\mathbf{x}_i\beta))^{1-y_i}) &= \log(G(\mathbf{x}_i\beta)^{y_i}) + \log((1-G(\mathbf{x}_i\beta)^{1-y_i}) \\ &= y_i\log(G(\mathbf{x}_i\beta)) + (1-y_i)\log(1-G(\mathbf{x}_i\beta)).\end{aligned}$$

Mathematically, the MLE involves solving the first-order condition of the log-likelihood function for $\widehat{\beta}_{MLE}$. This problem does not give an explicit formula for $\widehat{\beta}_{MLE}$. The estimator is obtained using numerical optimisation methods and is readily available in statistical packages such as R.

The regression output of MLE provides the parameter estimates, $\widehat{\beta}_{MLE}$ and associated (asymptotic) standard errors in a table similar to that of OLS. In particular, when we consider the empirical application of labour market participation of women using the Mroz data again, we obtain the following output when estimating the probit model in R:



```
Call:
glm2(formula = inlf ~ nwifeinc + educ + exper + expersq + age +
    kidslt6 + kidsge6, family = binomial(link = "probit"), data = mroz_data)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  0.2700736  0.5080782   0.532  0.59503    
nwifeinc    -0.0120236  0.0049392  -2.434  0.01492 *  
educ         0.1309040  0.0253987   5.154 2.55e-07 ***
exper        0.1233472  0.0187587   6.575 4.85e-11 ***
expersq     -0.0018871  0.0005999  -3.145  0.00166 ** 
age         -0.0528524  0.0084624  -6.246 4.22e-10 ***
kidslt6     -0.8683247  0.1183773  -7.335 2.21e-13 ***
kidsge6      0.0360056  0.0440303   0.818  0.41350    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

It is common to also report the value of the log-likelihood function when providing MLE regression results.

Interpreting results from probit and logit models

Fitted values: Using the probit and logit parameter estimates $\widehat{\beta}_j$, for $j = 0, \ldots, k$, we can obtain the predicted response probability for all individuals.

In probit regression models:

$$Pr(\widehat{y_i = 1} \,|\, \mathbf{x_i}) \equiv \widehat{y}_i = \Phi(\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_k x_{ki}).$$

In logit regression models:

$$Pr(\widehat{y_i = 1} \,|\, \mathbf{x_i}) \equiv \widehat{y}_i = \Lambda(\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \cdots + \widehat{\beta}_k x_{ki}).$$

In order to obtain the predicted probabilities it is important not to forget the cdf functions! With reference to the probit results reported above, for instance we can obtain the predicted probability of a woman with particular characteristics to be in the labour force by computing:

$$\begin{aligned}\widehat{inlf} &= \Phi(.270 - 0.012 nwifeinc + .131 educ + .123 exper \\ &\quad - .0019 exper^2 - 0.053 age - .868 kidslt6 + .036 kidsge6)\end{aligned}$$

where we need to use the cdf of a standard normal distribution (see Table G.1 in Wooldridge) as we used the probit model.

Parameter estimates: It is important to realise that the parameter estimates, $\widehat{\beta}_j$, for $j = 0, \dots, k$, from probit and logit cannot be interpreted as the partial effect that explains how each explanatory variable affects the predicted probability that $y = 1$ holding everything else constant. The reason for this is that the logit and probit models are specified using a *nonlinear function*, where:

$$Pr(y = 1 \,|\, \mathbf{x}) = G(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \equiv G(\mathbf{x}\beta).$$

Only when the relationship is *linear* (which defines the LPM), can we argue that $\Delta Pr(y = 1 \,|\, \mathbf{x})/\Delta x_j = \beta_j$. To get the partial effect for the logit and probit model, we need to use:

$$\frac{\partial Pr(y = 1 \,|\, \mathbf{x})}{\partial x_j} = \frac{\partial G(\mathbf{x}\beta)}{\partial x_j} = \beta_j g(\mathbf{x}\beta) \quad \text{where } g(z) = \frac{\mathrm{d}G(z)}{\mathrm{d}z}.$$

As the derivative of the cumulative density function (G) equals the probability density function (g), we use:

- $g(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2)$ (probit model)
- $g(z) = \exp(z)/[1 + \exp(z)]^2$ (logit model).

Conveniently, as $g(z) > 0$, this shows that the partial effect does have the same sign as $\widehat{\beta}_j$. If $\widehat{\beta}_j > 0$: *Ceteris paribus*, x_j affects the predicted probability that $y = 1$ positively. If $\widehat{\beta}_j < 0$: *Ceteris paribus*, x_j affects the predicted probability that $y = 1$ negatively.

Using the probit regression results:

$$\begin{aligned}\widehat{inlf} &= \Phi(.270 - 0.012 nwifeinc + .131 educ + .123 exper \\ &\quad - .0019 exper^2 - 0.053 age - .868 kidslt6 + .036 kidsge6)\end{aligned}$$

this allows us to state the following.

- As $\widehat{\beta}_{nwifeinc} < 0$: A higher non-wife income reduces the predicted probability of a woman being in the labour market, *ceteris paribus*.
- As $\widehat{\beta}_{kidslt6} < 0$: The presence of young children reduces the probability of a woman being in the labour market, *ceteris paribus*.
- Whether these effects (their directions) are statistically significant will depend on the precision with which these estimates are estimated (their standard errors). More about this later.

Marginal effects: The partial effects for the logit and probit models are not constant across the range of values the explanatory variables can take, unlike the LPM model. Depending on the values of $\mathbf{x}$ these partial effects will take different numerical values.

- For continuous explanatory variables x_j, the partial effect is given by:

$$\frac{\partial Pr(y=1\,|\,\mathbf{x})}{\partial x_j} = \frac{\partial G(\mathbf{x}\beta)}{\partial x_j} = \beta_j g(\mathbf{x}\beta) \quad \text{where } g(z) = \frac{\mathrm{d}G(z)}{\mathrm{d}z}.$$

- For explanatory variables that are dummy variables (discrete), the partial effect evaluates the difference in probability of participation when the dummy variable switches from zero to one. Say x_1 is a dummy variable, then:

$$\begin{aligned}\frac{\Delta Pr(y=1\,|\,\mathbf{x})}{\Delta x_1} &= G(\beta_0 + \beta_1 \times 1 + \beta_2 x_2 + \cdots + \beta_k x_k) \\ &\quad - G(\beta_0 + \beta_1 \times 0 + \beta_2 x_2 + \cdots + \beta_k x_k).\end{aligned}$$

Rather than reporting estimated partial effects for all individuals in the sample, the following approaches are popular. We report the average of the partial effect (APE) over all individuals, we report the partial effect for a particular individual, or we report the partial effect of an individual with characteristics given by the average person (PEA). That is, for continuous variables:

- *Estimate of the Average Partial Effect (APE) over all individuals*:

$$\widehat{APE}_j = \frac{1}{n}\sum_{i=1}^{n} g(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}) \cdot \widehat{\beta}_j.$$

- *Estimate of the partial effect for a particular individual*:

$$g(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}) \cdot \widehat{\beta}_j.$$

- *Estimate of the Partial Effect of the Average individual (PEA)*. Let $\bar{x}_j$ denote the average of the jth explanatory variable in the sample:

$$\widehat{PEA}_j = g(\widehat{\beta}_0 + \widehat{\beta}_1 \bar{x}_1 + \widehat{\beta}_2 \bar{x}_2 + \cdots + \widehat{\beta}_k \bar{x}_k) \cdot \widehat{\beta}_j.$$

The discrete analogues of these expressions can be found in Wooldridge.

The partial effects, which depend on the level of the characteristics, are characterised by diminishing effects (see the example which follows).

Using the probit regression results:

$$\begin{aligned}\widehat{inlf} &= \Phi(.270 - 0.012\,nwifeinc + .131\,educ + .123\,exper \\ &\quad - .0019\,exper^2 - 0.053\,age - .868\,kidslt6 + .036\,kidsge6)\end{aligned}$$

this allows us to state the following.

- The marginal effect of *nwifeinc* is equal to:

$$-.012 \times \phi(.270 - 0.012\,nwifeinc + .131\,educ + \cdots + .036\,kidsge6)$$

where $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)$.

- Rather than reporting this result for all individuals, we may report the average marginal (partial) effect over all individuals, $\widehat{APE}_{nwifeinc}$:

$$-.012 \times \frac{1}{753}\sum_{i=1}^{753}\phi(.270 - 0.012 nwifeinc_i + .131 educ_i + \cdots + .036 kidsge6_i).$$

- Our statistical package can help us to obtain $\widehat{APE}_{nwifeinc} = -.0036$ (we will discuss how to do this in R on the VLE). It reveals that \$10,000 additional other income (i.e. $\Delta nwifeinc = 10$) on average reduces the probability of women working in the labour market by .036, that is a 3.6 percentage point decrease, holding everything else constant. Our statistical package will also provide the associated standard error, $se(\widehat{APE}_{nwifeinc})$, which we can use to test whether this effect is statistically significant.

- Whereas the LPM assumes constant marginal effects, logit and probit imply diminishing magnitudes of the partial effect. To show this, let us consider the effect of an additional young child for women with $nwifeinc = 20.13$, $educ = 12.3$, $exper = 10.6$, $age = 42.5$ and $kidsge6 = 1$.

  Below we report the predicted probability of being in the labour market for this woman as a function of the number of young children:

$$\widehat{Pr}(inlf = 1 \,|\, kidslt6 = 0, nwifeinc = 20.13, educ = 12, \ldots) = .707$$

$$\widehat{Pr}(inlf = 1 \,|\, kidslt6 = 1, nwifeinc = 20.13, educ = 12, \ldots) = .373$$

$$\widehat{Pr}(inlf = 1 \,|\, kidslt6 = 2, nwifeinc = 20.13, educ = 12, \ldots) = .117.$$

  Comparing these probabilities, we note that the first child reduces her predicted probability by (.707 − .373) = .334 (a 33.4 percentage points drop). The second child reduces her predicted probability further but by a smaller amount: (.373 − .117) = .256 (a 25.6 percentage points drop).

Goodness of fit

A measure of goodness of fit that is commonly reported in binary response models (also for LPM) indicates the **percentage of observations that are correctly predicted.**

A common rule for this is to consider y as being correctly predicted if:

$$y_i = 1 \text{ and } Pr(\widehat{y_i = 1} \,|\, \mathbf{x}_i) \geq 0.5 \quad \text{or} \quad y_i = 0 \text{ and } Pr(\widehat{y_i = 1} \,|\, \mathbf{x}_i) < 0.5.$$

This rule has a drawback in that it does not mirror the quality of the prediction. Whether $Pr(\widehat{y_i = 1} \,|\, \mathbf{x}_i) = 0.5001$ or $Pr(\widehat{y_i = 1} \,|\, \mathbf{x}_i) = 0.9999$ is not reflected in this measure as in both cases we will indicate that the observation is predicted correctly if $y_i = 1$, and incorrectly if $y_i = 0$.

There are alternative measures, such as the pseudo R-squared (not examinable). Later we will discuss how we can formally test the joint significance of the slopes as a practical alternative.

Hypothesis testing

Single linear restrictions: As we have seen, the MLE regression output provides parameter estimates $(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k)$ and their standard errors. This will permit us to test hypotheses, such as:

$$H_0 : \beta_2 = 1 \quad \text{against} \quad H_1 : \beta_2 \neq 1$$

using an **asymptotic *t* test**. The procedure is the same as before, except that we will not use the Student's t distribution to obtain the critical values, but the standard normal distribution. The reason for this is that, in general, MLEs are typically only *asymptotically normal*. Specifically, we should use here the test statistic:

$$z = \frac{\widehat{\beta}_2 - 1}{se(\widehat{\beta}_2)} \overset{a}{\sim} N(0, 1) \text{ under } H_0.$$

At the 5% level of significance, we should reject H_0 if the realisation of our test statistic:

$$|z| > 1.96.$$

Comment: MLE regression output provides the z statistic for the hypotheses:

$$H_0 : \beta_j = 0$$

together with their associated two-sided p-values just as it did in OLS! (The third and fourth column of our MLE regression output.) If the p-value is smaller than 5%, we should reject the null that $\beta_j = 0$ at the 5% level of significance and we would state that $\widehat{\beta}_j$ is statistically significant.

Multiple linear restrictions: The MLE regression output typically also provides the value of the log-likelihood function ($\log L$). The log-likelihood value plays an important role when testing multiple linear restrictions such as

$$H_0 : \beta_2 = 0 \text{ and } \beta_3 = 0 \quad \text{against} \quad H_1 : \beta_2 \neq 0 \text{ and/or } \beta_3 \neq 0.$$

The test is based on detecting whether imposing restrictions leads to a significant loss of fit. The test requires us to compare the value of the likelihood function of the unrestricted model, L_{ur}, with the value of the likelihood function of the restricted model (where we impose the restrictions), L_r. The test is called the **likelihood ratio (LR) test**.

The likelihood ratio test statistic is given by:

$$LR = 2\log\left(\frac{L_{ur}}{L_r}\right) = 2(\log L_{ur} - \log L_r) \overset{a}{\sim} \chi_q \text{ under } H_0$$

where $q = df_{ur} - df_r$ equals the number of restrictions. This is an (asymptotic) χ^2 test. For a given level of significance, α, we should reject the null when the realisation of the likelihood ratio test:

$$LR > \chi^2_{q,\,\alpha}.$$

With $q = 2$ and $\alpha = 5\%$ this critical value equals 5.99. We should recognise that realisations of the LR test statistic that are small, signal that there is not a significant loss in fit and therefore should not lead to a rejection of the null.

Using the probit regression results:

$$\begin{aligned}\widehat{inlf} &= \Phi\Bigg(\underset{(.509)}{.270} - \underset{(.0048)}{0.012}\, nwifeinc + \underset{(.025)}{.131}\, educ + \underset{(.019)}{.123}\, exper \\ &\quad -\underset{(.0006)}{.0019}\, exper^2 - \underset{(.008)}{0.053}\, age - \underset{(.119)}{.868}\, kidslt6 + \underset{(.054)}{.036}\, kidsge6\Bigg) \\ n &= 753,\ \log L = -401.302\end{aligned}$$

this allows us to state the following.

- We can test the hypothesis:

  $$H_0 : \beta_{educ} = 0 \quad \text{against} \quad H_1 : \beta_{educ} \neq 0$$

  using the **asymptotic *t* test** (also called z test):

  $$t_{\widehat{\beta}_{educ}} = \frac{\widehat{\beta}_{educ}}{se(\widehat{\beta}_{educ})} = \frac{.131}{.025} \approx 5.18.$$

  At the 5% level of significance the critical value is 1.96. The p-value $< .001$, hence we conclude that education has a significant effect on being in the labour force. The test relies on MLE regularity conditions which ensures that:

  $$\text{Under } H_0 : \frac{\widehat{\beta}_{educ}}{se(\widehat{\beta}_{educ})} \overset{a}{\sim} Normal(0,1).$$

- We can test the joint hypothesis:

  $$H_0 : \beta_{exper} = 0,\ \beta_{exper^2} = 0 \quad \text{against} \quad H_1 : \beta_{exper} \neq 0 \text{ and/or } \beta_{exper^2} \neq 0$$

  using the likelihood ratio test:

  $$LR = 2(\log L_{ur} - \log L_r) \overset{a}{\sim} \chi^2_2 \text{ under } H_0.$$

  To obtain $\log L_{ur}$ we need to know the value of the log-likelihood of the model where these restrictions are not imposed. This regression is reported above. Hence $\log L_{ur} = -401.302$.

  To obtain $\log L_r$ we need to estimate the probit model where the two restrictions are imposed. Estimating the probit model where these two variables are left out, gives us:

  $$\begin{aligned}\widehat{inlf} &= \Phi\Bigg(\underset{(.473)}{.422} - \underset{(.0046)}{0.021}\, nwifeinc + \underset{(.024)}{.156}\, educ \\ &\quad -\underset{(.008)}{0.034}\, age - \underset{(.114)}{.892}\, kidslt6 + \underset{(.040)}{.037}\, kidsge6\Bigg) \\ n &= 753,\ \log L = -454.225 \longrightarrow \text{Hence } \log L_r = -454.223.\end{aligned}$$

  Our test statistic $LR = 2(-401.302 - -454.225) = 105.85$ is extremely large. Experience and experience squared are jointly significant with p-value $< .001$. At the 5% level of significance our critical value $\chi^2_{2,5\%} = 5.99$.

- We can also use the LR test to test the joint significance of all explanatory variables:

$$H_0 : \beta_{nwifeinc} = 0, \beta_{educ} = 0, \ldots, \text{ and } \beta_{kidsge6} = 0.$$

This would require us to test **seven restrictions**, with:

$$LR = 2(\log L_{ur} - \log L_r) \overset{a}{\sim} \chi^2_7 \text{ under } H_0.$$

To obtain $\log L_r$ we need to run a probit regression that includes an intercept only (*all explanatory variables should be excluded*). This gives $\log L_r = -514.873$.

The $\log L_{ur}$ is the same as before, $\log L_{ur} = -401.302$. Our test statistic $LR = 2(-401.302 - -514.873) = 227.14$ shows that our explanatory variables are jointly highly significant (p-value $< .001$).

Frequently MLE regression output reports this test statistic and its associated p-value as it provides a **goodness of fit measure**.

Activity 12.5 Take the MCQs related to this section on the VLE to test your understanding.

Activity 12.6 Exercise 17.2 from Wooldridge.

Activity 12.7 (Based on Exercise 17.C9 from Wooldridge.)

Let us try to explain the willingness of households to buy ecologically produced apples. We use data where each family was presented with a description of ecologically friendly apples, along with prices (in $) of regular apples (*regprc*) and prices of the hypothetical eco-labelled apple (*ecoprc*).

The variable we want to explain is the dummy variable *ecobuy*, which equals 1 if the household wants to buy ecologically friendly apples, and 0 otherwise. Using a sample of 660 households, the following results were obtained:

	OLS	probit	logit
constant	0.890 (.068)	1.088 (.206)	1.760 (.341)
regprc	0.735 (.132)	2.029 (.378)	3.283 (.623)
ecoprc	−0.845 (.106)	−2.344 (.318)	−3.792 (.527)
R^2	.086		
$\log L$		−407.60	−407.89

The numbers in parentheses are the heteroskedasticity-robust standard errors for OLS and the (asymptotic) standard errors for probit and logit.

The average price of regular apples in the sample is $0.90 and the average price of ecologically produced apples in the sample is $1.10. At these average prices:

(i) Discuss how you can estimate the probability (i.e. willingness) to buy ecologically produced apples using the Linear Probability Model (LPM).

(ii) Discuss how you can estimate the probability to buy ecologically produced apples using the probit results.

(iii) Discuss the effect of a \$.10 reduction of the price of ecologically produced apples using the LPM.

(iv) Discuss the effect of a \$.10 reduction of the price of ecologically produced apples using the logit model.

(v) How do your answers to (c) and (d) change for a further \$.10 reduction of the price of ecologically produced apples?

Activity 12.8 (Based on Exercise 17.C9 from Wooldridge.)

We return to our problem which tries to explain the willingness of households to buy ecologically produced apples. We use data where each family was presented with a description of ecologically friendly apples, along with prices (in \$) of regular apples (*regprc*) and prices of the hypothetical eco-labelled apples (*ecoprc*).

The variable we want to explain is the dummy variable *ecobuy*, which equals 1 if the household wants to buy ecologically friendly apples, and 0 otherwise. Additional household variables we have are family income in \$1000s, *faminc*, household size, *hhsize*, years of schooling, *educ*, and *age*.

Using a sample of 660 households, the following results were obtained:

	OLS A	OLS B	probit A	probit B
constant	0.890	0.424	1.088	−0.244
	(.068)	(.168)	(.206)	(.474)
regprc	0.735	0.719	2.029	2.030
	(.132)	(.131)	(.378)	(.381)
ecoprc	−0.845	−0.803	−2.344	−2.267
	(.106)	(.106)	(.318)	(.321)
faminc		0.0006		0.0014
		(.0005)		(.0015)
hhsize		0.024		0.069
		(.012)		(.037)
educ		0.024		0.071
		(.008)		(.024)
age		−0.001		−0.001
		(.001)		(.004)
R^2	.086	.110		
$\log L$			−407.60	−399.04

The numbers in parentheses are the heteroskedasticity-robust standard errors for OLS and the (asymptotic) standard errors for probit.

(i) Test the hypothesis $H_0 : \beta_{ecoprc} = 0$ against $H_1 : \beta_{ecoprc} < 0$ using the OLS and probit results that include the additional non-price controls. What interpretation can you give this parameter in the OLS and probit setting?

We want to test whether, after controlling for prices, the nonprice variables do not affect the willingness to buy ecofriendly apples.

(ii) Clearly indicate the null and alternative hypotheses.

(iii) Using the LPM (OLS) the robust F statistic for this hypothesis is given by 4.24 with a p-value $< .001$. Interpret the finding, and explain the importance of using a robust F statistic.

(iv) Test this hypothesis using the probit regression results using the LR test. Clearly indicate the (asymptotic) distribution and the rejection rule.

12.2.6 R Applications for binary response models

Activity 12.9 Complete 'R Task – Activity 6' on VLE.

Here we provide details of how to estimate binary response models, including the linear probability model (LPM) and nonlinear models such as probit and logit.

12.3 Answers to activities

Solution to Activity 12.2

(Exercise 7.7 from Wooldridge.)

$$inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + u. \qquad (*)$$

(i) Following the suggestion, let us plug in $inlf = 1 - outlf$:

$$1 - outlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \beta_3 exper + \beta_4 exper^2 + u.$$

Rearranging this yields:

$$outlf = \underbrace{(1-\beta_0)}_{\alpha_0} \underbrace{-\beta_1}_{\alpha_1} nwifeinc \underbrace{-\beta_2}_{\alpha_2} educ \underbrace{-\beta_3}_{\alpha_3} exper \underbrace{-\beta_4}_{\alpha_4} exper^2 \underbrace{-u}_{e}. \qquad (**)$$

If we regress $outlf$ on all independent variables:

$$\widehat{\alpha}_0 = 1 - \widehat{\beta}_0 \quad \text{and} \quad \widehat{\alpha}_j = -\widehat{\beta}_j,\ j = 1,2,3,4.$$

- The slope parameters have the opposite sign when using *outlf* instead of *inlf* as the dependent variable.
- **Intuition of result**: Recall, the slope parameters in the linear probability model explain how regressors affect the probability that $inlf = 1$, *ceteris paribus.*
- When we use *outlf* as the dependent variable, the slope parameters will explain how regressors affect the probability that $outlf = 1$, *ceteris paribus.*

$$Pr(outlf = 1 \mid x) = 1 - Pr(outlf = 0 \mid x) = 1 - Pr(inlf = 1 \mid x).$$

(ii) The standard errors will not change. Recall $se(\widehat{\beta}_j) = \sqrt{\widehat{Var(\widehat{\beta}_j \mid \mathbf{X})}}$.

Slopes: Changing the signs of the estimators does not change their variances:

$$Var(-\widehat{\beta}_j \,|\, \mathbf{X}) = Var(\widehat{\beta}_j \,|\, \mathbf{X})$$

and therefore the standard errors are unchanged (but the t statistics change sign).

Intercept: As:

$$Var(1 - \widehat{\beta}_0 \,|\, \mathbf{X}) = Var(\widehat{\beta}_0 \,|\, \mathbf{X})$$

the standard error of the intercept is also the same as before.

(iii) Recall, by definition $R^2 = 1 - SSR/SST$.

Observations regarding R^2:

- Changing the units of measurement of independent variables, or entering qualitative information using different sets of dummy variables, does not change the R^2.
- Changing the dependent variable typically will affect the R^2. However, due to the special nature of the dependent variables ($inlf$ and $outlf$), the R-squareds from the regressions will be the same.

Let us show that the SSRs are the same when we estimate (*) or (**).

- Recall (*): $e = -u$. For each i, the error in the equation for $outlf_i$ is just the negative of the error in the other equation for $inlf_i$. The same is true of the residuals.

Let us show that the SSTs are the same when we estimate (*) or (**).

- The SST for (**) equals:

$$SST = \sum_{i=1}^{n} (outlf_i - \overline{outlf})^2.$$

- Let us plug in $outlf = 1 - inlf$:

$$SST = \sum_{i=1}^{n} [(1 - inlf_i) - (1 - \overline{inlf})]^2.$$

- By rearranging we obtain:

$$SST = \sum_{i=1}^{n} [-inlf_i + \overline{inlf})]^2 = \sum_{i=1}^{n} [inlf_i - \overline{inlf})]^2$$

which is the SST for (*).

Hence, the R-squareds are the same in the two regressions.

Solution to Activity 12.3

We are asked to consider the properties of $\widehat{\theta}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

(i) For unbiasedness we need to show $E(\widehat{\theta}_{MLE}) = \theta$.

$$E(\widehat{\theta}_{MLE}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}\sum_{i=1}^{n} \theta = \theta.$$

It should be noted that **not all ML estimators are unbiased**.

(ii) For consistency we need to show $\text{plim}(\widehat{\theta}_{MLE}) = \theta$.

We can directly refer to the **law of large numbers** to establish:

$$\text{plim}(\widehat{\theta}_{MLE}) \equiv \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = E(Y_i).$$

Since $E(Y_i) = \theta$ consistency is established.

Recall that the law of large numbers ensures that the sample average $\frac{1}{n}\sum_{i=1}^{n} Y_i$ converges to the population mean $E(Y_i)$.

(iii) Since $E(1/\widehat{\theta}_{MLE}) \neq 1/E(\widehat{\theta}_{MLE})$, we do not obtain an unbiased estimator of $1/\theta$. This should remind you of the limitation of the expectation operator.

(iv) Using the properties of the plim operator, we note:

$$\text{plim}\left(\frac{1}{\widehat{\theta}_{MLE}}\right) = \frac{\text{plim}(1)}{\text{plim}(\widehat{\theta}_{MLE})} = \frac{1}{\theta}.$$

Hence we get a consistent estimator of $1/\theta$. This should remind you of the nice features of the plim operator.

Solution to Activity 12.4

(i) Let us follow the steps required to derive the MLE.

Step 1: Construct the joint density.

- By independence:

$$f(x_1,\ldots,x_n;\beta) = \prod_{i=1}^{n} f(x_i;\beta) = \frac{1}{\beta}\exp\left(-\frac{x_1}{\beta}\right)\frac{1}{\beta}\exp\left(-\frac{x_2}{\beta}\right)\cdots\frac{1}{\beta}\exp\left(-\frac{x_n}{\beta}\right).$$

Using the product rule we obtain:

$$f(x_1,\ldots,x_n;\beta) = \frac{1}{\beta^n}\exp\left(-\frac{\sum_{i=1}^{n} x_i}{\beta}\right).$$

Step 2: Obtain the log-likelihood function.

- The likelihood function is defined by the joint density:

$$L(\beta; X_1, \ldots, X_n) = \frac{1}{\beta^n} \exp\left(-\frac{\sum_{i=1}^n X_i}{\beta}\right).$$

- Using the rules of the log function:

$$\log L(\beta; X_1, \ldots, X_n) = -n\log(\beta) - \frac{\sum_{i=1}^n X_i}{\beta}.$$

Step 3: Obtain $\widehat{\beta}_{MLE}$ by solving the FOC.

- The derivative is:

$$\frac{\mathrm{d}\log L}{\mathrm{d}\beta} = -\frac{n}{\beta} + \frac{\sum_{i=1}^n X_i}{\beta^2}.$$

- $\widehat{\beta}_{MLE}$ should solve:

$$-\frac{n}{\widehat{\beta}_{MLE}} + \frac{\sum_{i=1}^n X_i}{\widehat{\beta}^2_{MLE}} = 0.$$

- Multiply equation by $\widehat{\beta}^2_{MLE}$ gives: $-n\widehat{\beta}_{MLE} + \sum\limits_{i=1}^n X_i = 0.$
- Hence $\widehat{\beta}_{MLE} = \bar{X}$ which confirms the answer provided in question.

(ii) The ML estimators are known to have nice properties: consistent, asymptotic normal, and asymptotic efficient.

Recall, we are told that $E(X_i) = \beta$.

Since $E(\bar{X}) = \frac{1}{n}\sum E(X_i) = \beta$ we establish that the MLE is unbiased as well.

Using the law of large numbers we know $\text{plim}(\bar{X}) = E(X_i)$. Since $E(X_i) = \beta$ this establishes that the MLE is consistent.

Solution to Activity 12.6

(Exercise 7.2 from Wooldridge.)

Compute for a student athlete with *hsGPA* = 3.0, *SAT* = 1,200, *study* = 10 the probability of graduation within five years:

1. Compute z using the student characteristics and parameters as:

$$z = -1.17 + .24(3.0) + .00058(1{,}200) + .073(10) = .976.$$

2. Apply the logit function to z, where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$:

$$\frac{\exp(.976)}{1 + \exp(.976)} \approx .726$$

which is 72.6%!

Similarly, compute the estimated probability of graduating within five years for a student athlete ($hsGPA = 3.0,\ SAT = 1,200,\ study = 5$):

1. Recompute z using the new characteristics and parameters as:

$$z = -1.17 + .24(3.0) + .00058(1{,}200) + .073(5) = .611.$$

2. We apply the logit function to z:

$$\frac{\exp(.611)}{1+\exp(.611)} \approx .648$$

which is 64.8%!

From this we conclude that the difference in estimated probabilities is 72.6% − 64.8% = 7.8%. That is, there is a 7.8 percentage point difference! Having spent more time in an organised study hall increased the probability of a student with an $hsGPA = 3$ and $SAT = 1{,}200$ graduating in five years. [Note how far off the calculation would be if we simply used the coefficient on study to conclude that the difference in probabilities is $.073(10-5) = .365$, or 36.5%.]

Solution to Activity 12.7

(Based on Exercise 17.C9 from Wooldridge.)

(i) The **OLS (LPM)** results give the following regression line:

$$\widehat{Pr(y=1\,|\,\mathbf{x})} \equiv \widehat{y} = .890 + .735 regprc - .845 ecoprc.$$

Evaluating at average prices ($.90 for regular and $1.10 for ecologically produced apples) this yields a predicted probability equal to:

$$0.890 + .735 \times .90 - 0.845 \times 1.10 = 0.622.$$

(ii) The **probit** results give the following (nonlinear) regression line:

$$\widehat{Pr(y=1\,|\,\mathbf{x})} = \Phi(1.088 + 2.029 regprc - 2.344 ecoprc).$$

Evaluating at average prices this yields a predicted probability equal to:

$$\Phi(1.088 + 2.029 \times .90 - 2.344 \times 1.10) = \Phi(0.336) = 0.633$$

using Table G.1.

Similarly, the **logit** results give the following (nonlinear) regression line:

$$\widehat{Pr(y=1\,|\,\mathbf{x})} = \Lambda(1.760 + 3.283 regprc - 3.792 ecoprc)$$

where $\Lambda(z) = \exp(z)/[1+\exp(z)]$.

Evaluating at average prices this yields a predicted probability equal to:

$$\Lambda(1.760 + 3.283 \times .90 - 3.792 \times 1.10) = \Lambda(0.5435)$$
$$= \frac{\exp(.5435)}{1+\exp(.5435)} = .633$$

the same (or typically very similar) to the probit result.

(iii) Using the OLS (LPM) results, the effect of reducing *ecoprc* by \$.10 equals:

$$-.845 \times -.10 = .0845.$$

This reveals an 8.45 percentage point increase in the probability of buying ecofriendly apples.

(iv) To obtain the effect using the logit model we can use:

$$\begin{aligned}&\Lambda(1.760 + 3.283 \times .90 - 3.792 \times \mathbf{1.00})\\&\quad - \Lambda(1.760 + 3.283 \times .90 - 3.792 \times \mathbf{1.10})\\&= \Lambda(.9227) - \Lambda(.5435) = .7156 - .633 = .083\end{aligned}$$

or a 8.3 percentage points increase.

- Alternatively, we can use the result that the partial effect of a **unit change** equals:
$$g(\mathbf{x}\widehat{\beta})\widehat{\beta}_{ecoprc} \quad \text{where} \quad g(z) = \exp(z)/(1 + \exp(z))^2.$$

- Evaluated at the average prices (PEA) this yields:

$$\begin{aligned}&g(1.760 + 3.283 \times .90 - 3.792 \times 1.10) \times -3.792\\&= \exp(.5435)/(1 + \exp(.5435))^2 \times -3.792 = -.8813.\end{aligned}$$

(Had we evaluated with *ecoprc* $= 1.00$ it would have equalled $-.772$.)

- To obtain the effect of a \$.10 drop in prices, we need to multiply this by -0.10.

(v) For the OLS (LPM) results the effect remains the same. The effect is constant. The effect of reducing *ecoprc* with a further \$.10 equals:

$$-.845 \times -.10 = .0845.$$

For logit/probit the effect will not the be same. We should expect a diminishing effect. Easiest is to compare the predicted probabilities:

$$\begin{aligned}&\Lambda(1.760 + 3.283 \times .90 - 3.792 \times \mathbf{.90})\\&\quad - \Lambda(1.760 + 3.283 \times .90 - 3.792 \times \mathbf{1.00})\\&= \Lambda(1.3019) - \Lambda(.9227) = .786 - .715 = .071.\end{aligned}$$

A further \$.10 decrease, increases the probability by 7.1 percentage points.

Solution to Activity 12.8

(Based on Exercise 17.C9 from Wooldridge.)

(i) We are asked to test $H_0 : \beta_{ecoprc} = 0$ against $H_1 : \beta_{ecoprc} < 0$ using the OLS B and probit B results.

For inference in the LPM setting (OLS) we have to use robust standard errors as there is presence of heteroskedasticity (violation of MLR.5). Specifically, the LPM, which assumes:

$$p(x) = \beta_0 + \beta_{regprc} regprc + \beta_{ecoprc} ecoprc + \cdots$$

exhibits heteroskedasticity as:

$$Var(ecobuy \,|\, \mathbf{x}) = p(\mathbf{x})(1 - p(\mathbf{x})) \neq \text{constant}.$$

For inference in the probit setting, we rely on the asymptotic properties of our ML estimators and use the reported (asymptotic) standard errors.

- In both cases we use the asymptotic t test statistic (also called z statistic):

  $$z = \frac{\widehat{\beta}_{ecoprc}}{se(\widehat{\beta}_{ecoprc})} \overset{a}{\sim} Normal(0, 1) \text{ under } H_0.$$

  Using a 5% level of significance we should reject if $z < -1.645$.

- For LPM, we obtain $-.803/.106 = -7.58$; for probit, we obtain $-2.267/.321 = -7.06$. In both settings we find strong evidence that $\beta_{ecoprc} < 0$.

- Interpretation estimates:
  - LPM [$\widehat{\beta}_{ecoprc} = -.803$]: A \$.10 increase in *ecoprc* will reduce the probability of buying an ecofriendly apple by .0803, i.e. an 8.03 percentage point drop.
  - Probit [$\widehat{\beta}_{ecoprc} = -2.267$]: The parameter does not have a direct (easy) interpretation. All we can say is that an increase in *ecoprc* will reduce the probability of buying an ecofriendly apple.

(ii) We are asked to test:

$$H_0 : \beta_{faminc} = 0, \beta_{hhsize} = 0, \beta_{educ} = 0, \text{ and } \beta_{age} = 0 \text{ against}$$

$$H_1 : \text{ at least one non-zero.}$$

(iii) For testing multiple linear restrictions for the LPM we use the robust F test (cannot be obtained using the results provided) – **heteroskedasticity!**

$F = 4.24$. Under H_0: $F \sim F_{4,653}$; the degrees of freedom of the numerator is 4 (number of restrictions) and the denominator $df - k - 1 = 653$.

As the p-value is small ($< .001$), we should reject the null hypothesis, finding evidence of the joint significance of the nonprice effects on the willingness to buy ecologically produced apples after controlling for prices.

(iv) For testing multiple linear restrictions for the probit model we use the LR test.

The test statistic can easily be obtained given the results:

$$LR = 2(\log L_{ur} - \log L_r) = 2(-316.60 - -407.60) = 182.0.$$

Under H_0 its (asymptotic) distribution is χ^2_4 (4 restrictions).

Comparing it to the critical value of the χ^2_4 (5% critical value 9.49) we reject the null that nonprice variables have no impact on the willingness to buy ecologically produced apples after controlling for prices.

12.4 Overview of chapter

In this chapter we discussed the binary response model, a model where the dependent variable is a binary variable. The use of OLS in this setting assumes the use of a linear probability model (LPM), where the probability of our choice is assumed to be linear. The partial effects in the LPM describe the effects the independent variables have on the probability of a particular choice holding all other factors constant.

After raising various drawbacks of the LPM, we introduced the logit and probit models (assumes the probability of our choice is nonlinear). We discussed the use of MLE for these models which exhibit various attractive properties (consistency, asymptotic efficiency and asymptotic normality). We discussed how to obtain predicted probabilities and partial effects for the LPM and the logit/probit model. Whereas the partial effects are constant in the LPM, the partial effects are not constant in the logit/probit models and exhibit diminishing effects.

We also discussed hypothesis testing for binary choice models. When we use the LPM, inference needs to account for the presence of heteroskedasticity (use robust standard errors). Hypothesis testing when relying on the logit/probit models (MLE) makes use of the z test (asymptotic t test) for single linear restrictions, or the likelihood ratio LR test for joint linear restrictions.

12.4.1 Key terms and concepts

- Binary response models
- Linear probability model (LPM)
- Response probability
- Maximum likelihood estimator (MLE)
- Log-likelihood function
- Asymptotic normality
- Asymptotic efficiency
- Logit and probit model
- Average partial effect (APE)
- Partial effect of the average individual (PEA)
- z test (asymptotic t test)
- Likelihood ratio (LR) test

12.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- motivate the interest and use of binary response models
- describe the linear probability model (LPM) and discuss the interpretation of parameters
- discuss the benefits/disadvantages of the LPM
- discuss heteroskedasticity-robust inference for the LPM
- explain the benefits of using a nonlinear model for binary response models
- understand the principles of maximum likelihood estimation (MLE)
- discuss the properties of ML estimators
- derive the MLE from first principles in simple models
- specify the logit model and the log-likelihood function for its estimation by MLE
- specify the probit model and the log-likelihood function for its estimation by MLE
- understand regression output provided by logit/probit models
- understand how to obtain partial effects in logit and probit models
- conduct hypothesis tests using logit and probit models (in particular use the z test and the likelihood ratio (LR) test)
- discuss a goodness of fit measure for binary response models
- use statistical software to estimate binary choice models (linear and nonlinear) using real-world data.

12.5 Test your knowledge and understanding

12.5.1 Additional examination based exercises

12.1E Suppose you are given a random sample $X_1, \ldots, X_n$ from a distribution whose probability density function (pdf) is defined by:

$$f(x; \lambda) = \lambda \exp(-\lambda x), \quad x > 0, \ \lambda > 0.$$

According to this distribution $E(X_i) = 1/\lambda$ and $Var(X_i) = 1/\lambda^2$, for $i = 1, \ldots, n$.

(a) Show that the maximum likelihood estimator of λ is $1/\bar{X}$, where $\bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$.

(b) Is the estimator unbiased and/or consistent? Prove your claims.

12.2E Let us consider an analysis on recidivism (probability of arrest) using a random sample of young men in California who have at least one arrest prior to 1986. The dependent variable, *arr86*, is equal to unity if the man was arrested at least once during 1986, and zero otherwise.

	OLS A	OLS B	logit A	logit B	logit B Marginal Effect
pcnv	$-.152$	$-.162$	$-.880$	$-.901$	$-.176$
	(.021)	(.021)	(.122)	(.120)	(.023)
avgsen	.005	.006	.027	.031	.006
	(.006)	(.006)	(.035)	(.034)	(.007)
tottime	$-.003$	$-.002$	$-.014$	$-.010$	$-.002$
	(.005)	(.005)	(.028)	(.027)	(.005)
ptime86	$-.023$	$-.022$	$-.140$	$-.127$	$-.025$
	(.005)	(.005)	(.031)	(.031)	(.006)
qemp86	$-.038$	$-.043$	$-.199$	$-.216$	$-.042$
	(.005)	(.005)	(.028)	(.028)	(.005)
black	.170	–	.823	–	–
	(.024)		(.117)		
hispan	.096	–	.522	–	–
	(.021)		(.109)		
constant	.380	.380	$-.464$	$-.169$	
	(.019)	(.019)	(.095)	(.084)	
R^2	.068	.047			
$\log L$			-1512.35	-1541.24	

pcnv is the proportion of prior arrests that led to a conviction, *avgsen* is the average sentence served from prior convictions, *tottime* is the months spent in prison since age 18 prior to 1986, *ptime86* is months spent in prison in 1986, *qemp86* is the number of quarters the man was legally employed in 1986, while *black* and *hispan* are two race dummies (*white* the excluded dummy). In parentheses, the robust standard errors are reported for OLS and the (asymptotic) standard errors for MLE.

(a) When estimating the parameters by OLS, we are using the Linear Probability Model (LPM).

i. Explain why using OLS when the dependent variable is binary implies we use the Linear Probability Model.

ii. Explain why we should report heteroskedasticity-robust standard errors.

iii. Using OLS A: Interpret the coefficient on the variable *black* and discuss whether it is significant. Clearly indicate the null and alternative hypothesis, the test statistic and the rejection rule.

iv. Using OLS B: What is the estimated effect of an increase in *pcnv* from 0.25 to 0.75, holding everything else constant, on the probability of arrest?

(b) It is argued that the linear probability model may not be appropriate for explaining the binary variable *arr86* and a logit regression model has been estimated.

i. Explain how the logit estimates are obtained.

Hint: You may recall, that for the logit model A, we will specify:

$$Pr(arr86_i = 1 \mid X_i) = \Lambda(\beta_0 + \beta_1 pcnv_i + \beta_2 avgsen_i + \cdots + \beta_6 black_i + \beta_7 hispan_i)$$

where $\Lambda(z) = \frac{1}{1+\exp(-z)}$ and X_i denote the explanatory variables.

ii. Using logit A: What interpretation can you give to the coefficient on the variable *black* if any? Discuss whether it is significant? Clearly indicate the null and alternative hypothesis, the test statistic and the rejection rule.

iii. Discuss whether *black* and *hispan* are jointly significant using the logit model results. Clearly indicate the null and alternative hypothesis, the test statistic and the rejection rule.

iv. Using logit B: Discuss how the marginal effects (reported in the last column), evaluated at the mean values of the explanatory variables, are obtained. Give a brief comment as to how they compare to the marginal effects of the associated LPM.

v. Using logit B: What is the estimated effect of an increase in *pcnv* from 0.25 to 0.75 for a *white* man, with characteristics *avgsen=1, tottime=1, ptime86=0 and qemp86=2* on the probability of arrest? A clear explanation of what calculations are required is sufficient.

12.5.2 Solutions to additional examination based exercises

12.1E **Step 1**: Construct the joint density.

- By independence:

$$f(x_1,\ldots,x_n;\beta)=\prod_{i=1}^{n}f(x_i;\beta)=\lambda\exp(-\lambda x_1)\lambda\exp(-\lambda x_2)\cdots\lambda\exp(-\lambda x_n).$$

- Using the product rule we obtain:

$$f(x_1,\ldots,x_n;\beta)=\lambda^n\exp\left(-\lambda\sum_{i=1}^{n}x_i\right).$$

Step 2: Obtain the log-likelihood function.

- The likelihood function is defined by the joint density:

$$L(\lambda;X_1,\ldots,X_n)=\lambda^n\exp\left(-\lambda\sum_{i=1}^{n}X_i\right).$$

- Using the rules of the log function:

$$\log L(\lambda;X_1,\ldots,X_n)=n\log(\lambda)-\lambda\sum_{i=1}^{n}X_i.$$

Step 3: Obtain $\widehat{\lambda}_{MLE}$ by solving the FOC:

- The derivative is:

$$\frac{\mathrm{d}\log L}{\mathrm{d}\lambda}=\frac{n}{\lambda}-\sum_{i=1}^{n}X_i.$$

- $\widehat{\lambda}_{MLE}$ should solve: $\frac{n}{\widehat{\lambda}_{MLE}}-\sum_{i=1}^{n}X_i=0.$
- Hence $\widehat{\lambda}_{MLE}=1/\bar{X}$ as required.

Comment: Make sure you are comfortable with the mathematics fundamentals, such as:

$$\begin{aligned}\log(AB) &= \log(A) + \log(B)\\ \log(A^c) &= c\log(A)\\ \log(\exp(A)) &= A\\ \frac{\mathrm{d}\log(x)}{\mathrm{d}x} &= \frac{1}{x}.\end{aligned}$$

The estimator is **not unbiased**, that is $E(\widehat{\lambda}_{MLE}) \neq \lambda$.

As $E(X_i) = 1/\lambda$, we know that $E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = 1/\lambda$.

In general:

$$E\left(\frac{1}{\bar{X}}\right) \neq \frac{1}{E(\bar{X})} \quad \text{(Jensen's inequality).}$$

Hence we need to conclude that the ML estimator in general is not unbiased as the right-hand side equals λ.

The estimator is **consistent**, that is $\text{plim}(\widehat{\lambda}_{MLE}) = \lambda$.

By the law of large numbers we know $\text{plim}(\bar{X}) = E(X_i) = 1/\lambda$. By properties of the plim operator moreover:

$$\text{plim}\left(\frac{1}{\bar{X}}\right) = \frac{\text{plim }(1)}{\text{plim}(\bar{X})} = \frac{1}{1/\lambda} = \lambda.$$

Comment: You may want to recall **Jensen's inequality** which reveals:

$$E(g(X)) \neq g(E(X))$$

in general. By definition (for continuous random variables):

$$E(g(X)) = \int g(x)f(x)\,\mathrm{d}x$$

where $f(x)$ is the probability density function of the random variable X. Forgetting this is a mistake often made in examinations.

12.2E (a) i. The reason for this is that with a binary dependent variable Y, we need to interpret its conditional expectation $E(Y \mid X)$ as $Pr(Y = 1 \mid X)$, where X denotes the explanatory variables. As OLS assumes that this probability is linear in the parameters, i.e. we have:

$$Pr(\mathit{arr86} = 1 \mid \mathit{pcnv}, \ldots, \mathit{hispan}) = \beta_0 + \beta_1 \mathit{pcnv} + \cdots + \beta_7 \mathit{hispan}$$

it is called a linear probability model.

Comment (not asked): In the LPM the marginal effects are constant – for example, each additional quarter spent being legally employed in 1986 reduces the probability of arrest by the same amount (constant marginal effects).

ii. We need to report heteroskedasticity-robust standard errors as the LPM:

$$arr86_i = \beta_0 + \beta_1 pcnv_i + \cdots + \beta_7 hispan_i + e_i \equiv p(X_i) + e_i$$

inherently suffers from heteroskedasticity unless all slope parameters are zero. In particular, we have:

$$Var(e_i \,|\, X_i) = p(X_i)(1 - p(X_i)) \neq \text{constant.}$$

As this implies heteroskedasticity, we need to recognise that the usual standard errors are no longer valid (and inference based thereon incorrect) as they assume homoskedasticity (and no autocorrelation).

Comment (not asked): The OLS estimator is unbiased and consistent!

iii. The coefficient on the variable *black* reveals that black men have a .170 (that is 17 percentage points) higher probability of arrest compared to white men, *ceteris paribus*. Important to make reference to the base category!

To test for its significance, we need to test $H_0 : \beta_{\text{black}} = 0$ against $H_1 : \beta_{\text{black}} \neq 0$ using an asymptotic t test as we should use robust standard errors (as reported). The test statistic is $\widehat{\beta}_{\text{black}}/se(\widehat{\beta}_{\text{black}})$. At the 5% level of significance our critical value is 1.96. Given the realisation of our test statistic .170/.024 = 7.08, we conclude that it is significant (i.e. we reject H_0).

iv. Since:

$$\widehat{arr86}_i = \widehat{\beta}_0 + \widehat{\beta}_1 pcnv_i + \widehat{\beta}_2 avgsen_i + \cdots + \widehat{\beta}_5 qemp86$$

denotes the predicted probability of arrest in the LPM, the estimated effect is:

$$\widehat{\beta}_1(0.75 - 0.25) = -.162 \times .5 = -.081.$$

So the probability of arrest decreases by 0.081, or 8.1 percentage points.

(b) i. You are expected to indicate that we use maximum likelihood to estimate the β parameters.

The log-likelihood that will be maximised is given by:

$$\begin{aligned}\log L(\beta) &= \sum_{i=1}^{n} arr86_i * \log(Pr(arr86_i = 1 \,|\, X_i)) \\ &\quad + (1 - arr86_i) * \log(Pr(arr86_i = 0 \,|\, X_i)) \\ &= \sum_{i=1}^{n} arr86_i * \log(\Lambda(z_i)) + (1 - arr86_i) * \log(1 - \Lambda(z_i))\end{aligned}$$

where:

$$z_i = \beta_0 + \beta_1 pcnv_i + \beta_2 avgsen_i + \cdots + \beta_6 black_i + \beta_7 hispan_i.$$

Solving the FOC will give the MLE parameter estimates. Intuitively, we choose the βs so as to ensure that for individuals who are arrested their predicted probability of arrest is high.

ii. Unlike the LPM, the β parameters have no easy interpretation. Only the sign of the β parameters is interpretable, where a positive sign indicates that, *ceteris paribus*, hispanics have a higher probability of arrest than whites.

We use the z statistic to test the significance of β_{black}, as our ML estimator is only asymptotically normal. Here:

$$z = \frac{\widehat{\beta}_{\text{black}}}{se(\widehat{\beta}_{\text{black}})} \sim Normal(0,1) \quad \text{under } H_0 \text{ asymptotically.}$$

The realisation of our test statistic equals $.823/.117 = 7.03$, which indicates that it is statistically significant at the 1% level (critical value equal to 2.576).

iii. To test the joint hypothesis $H_0 : \beta_{\text{black}} = 0$ and $\beta_{\text{hispan}} = 0$ against the alternative H_1 that at least one is non-zero, we will use the likelihood ratio (LR) test. The test statistic is:

$$LR = 2(\log L_{ur} - \log L_r) \overset{a}{\sim} \chi^2_2 \text{ under } H_0$$

where $\log L_{ur}$ is the value of the log-likelihood of the unrestricted model (Model A), and $\log L_r$ is the value of the log-likelihood of the restricted model (Model B).

The realisation of our test statistic equals:

$$2(-1512.35 - -1541.24) = 57.78$$

which indicates at any reasonable level of significance we want to reject the null (i.e. they are jointly significant). At the 1% level of significance the critical value equals 9.21.

iv. The marginal effects of interest describe how $Pr(arr86_i = 1 \,|\, X_i)$ changes as a result of the explanatory variables. For continuous variables that means, for example:

$$\frac{\partial Pr(arr86_i = 1 \,|\, X_i)}{\partial pcnv} = f(z_i) * \beta_1 \text{ with } f(z) = \frac{\mathrm{d}\Lambda(z)}{\mathrm{d}z} = \mathrm{e}^{-z}/((1+\mathrm{e}^{-z}))^2$$

and:

$$z_i = \beta_0 + \beta_1 pcnv_i + \beta_2 avgsen_i + \cdots + \beta_6 black_i + \beta_7 hispan_i.$$

The marginal effects therefore depend on the characteristics of the individual. The marginal effects reported in the table are for an individual with average characteristics:

$$\bar{z} = \beta_0 + \beta_1 \overline{pcnv} + \beta_2 \overline{avgsen} + \cdots + \beta_6 \overline{black} + \beta_7 \overline{hispan.}$$

Whether this individual exists is another matter. Indeed, it may be more interesting to report the average of these marginal effects over all individuals:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial Pr(arr86_i = 1 \,|\, X_i)}{\partial pcnv} = \frac{1}{n}\sum_{i=1}^{n} f(z_i) * \beta_1.$$

The marginal effects for the average person are quite comparable to those obtained by the LPM.

v. We want to compare the predicted probability of arrest for this individual if *pcnv* goes from 0.25 to 0.75, that is:

$$\begin{aligned}
&\widehat{Pr}(\mathit{arr86} = 1 \,|\; \mathit{pcnv} = .75, \mathit{avgsen} = 1, \mathit{tottime} = 1, \mathit{ptime86} = 0, \\
&\qquad\qquad \mathit{qemp86} = 2, \mathit{black} = 0, \mathit{hispan} = 0) \\
&\quad - \widehat{Pr}(\mathit{arr86} = 1 \,|\; \mathit{pcnv} = .25, \mathit{avgsen} = 1, \mathit{tottime} = 1, \mathit{ptime86} = 0, \\
&\qquad\qquad \mathit{qemp86} = 2, \mathit{black} = 0, \mathit{hispan} = 0) \\
&= \Lambda(-1.69 - 0.901 * .75 + .031 - .010 - .216 * 2) \\
&\quad - \Lambda(-1.69 - 0.901 * .25 + .031 - .010 - .216 * 2) \\
&= \Lambda(-1.256) - \Lambda(-.805) \\
&= .222 - .309 \\
&= -.087
\end{aligned}$$

with:

$$\Lambda(z) = \frac{1}{1 + \exp(-z)}.$$

This indicates a decrease in the probability of arrest of .087, that is a 8.7% points reduction.

Comment (not asked): The marginal effect is different for different individuals unlike in the LPM setting where the estimated effect of an 8.1% point reduction (see (a) (iv) is the same for all individuals).

Chapter 13
Regression analysis with time series data

13.1 Introduction

13.1.1 Aims of the chapter

In this chapter, we extend our regression model to time series data. We will learn what the substantial differences are between time series data and cross-sectional data (our focus thus far) and will discuss three different types of time series models (i) the static model, (ii) a finite distributed lag model, and (iii) models where we include lagged values of the dependent variable as regressors.

We will learn how to adapt the Gauss–Markov assumptions to time series data. Importantly, with time series data we will no longer be able to rely on the random sampling assumption (MLR.2 will not hold). We replace this assumption by a new assumption which states that, conditional on the regressors, there is no correlation between the errors in different time periods (TS.5). Under the classical linear model assumptions TS.1 to TS.6, we then obtain exactly the same finite sample properties as before (unbiasedness, variance, efficiency) and statistical inference is standard.

We learn that the key assumption required for unbiasedness, TS.3 (strict exogeneity), is very restrictive here. It imposes limitations on the economic environment that are likely not satisfied and we conclude that, typically, we do not expect our OLS estimator to remain unbiased. We will discuss a weaker, more reasonable, exogeneity assumption (TS.3$'$) that will enable us to establish the consistency of the OLS estimator.

In order to establish large sample properties in the time series setting, we need to impose two assumptions: (1) weak dependence (need to restrict the dependence inherent in time series processes) and (2) stationarity (need to ensure that the statistical properties of these processes do not change over time). We will discuss these two properties and discuss commonly used processes, such as autoregressive (AR) and moving average (MA) processes of order 1, and analyse their dependence and stationarity properties.

Finally, we establish the assumptions (TS.1$'$ to TS.5$'$) that ensure the OLS estimator is approximately normally distributed which guarantees that the usual t tests, F tests and confidence intervals are asymptotically valid.

13.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand essential differences between time series data and cross-sectional data
- discuss limitations of static models in a time series setting
- describe a general form of finite distributed lag (FDL or just DL) models, interpret their coefficients and explain how to express immediate and cumulative effects on the response variable
- discuss the role of multicollinearity in FDL models and its consequences for the precision of estimates
- formulate and interpret models with lagged dependent variables as regressors
- explain which conditions guarantee the unbiasedness of OLS in regressions with time series data, explain the strict exogeneity assumption, and describe what economic situations it eliminates
- explain which conditions guarantee that the usual OLS variance formulas are valid in models with regressors correlated across time
- describe conditions that guarantee that the OLS estimator is BLUE
- estimate and interpret static and FDL in practice
- understand notions of stationarity, covariance stationarity and weak dependence
- analyse covariance stationarity and weak dependence properties of an AR(1) process
- analyse covariance stationarity and weak dependence properties of an MA(1) process
- explain which conditions guarantee the consistency of OLS in regressions with time series data, formulate the contemporaneous exogeneity assumption, and describe how it is different from the strict exogeneity assumption
- explain which conditions guarantee that the OLS estimator is approximately normally distributed and the usual t tests, F tests and confidence intervals are all asymptotically valid.

13.1.3 Essential reading

Readings: Wooldridge, Sections 10.1–10.3 and 11.1–11.2.

13.1.4 Synopsis of this chapter

In Section 10.1 in Wooldridge, the substantial differences between time series data and cross-sectional data are discussed. Various time series models are introduced in Section 10.2. First, we consider the so-called static models and discuss their limitations. We then introduce a class of finite distributed lag (FDL or just DL) models which allows one variable to depend on contemporaneous **and** lagged values of other variables. We discuss how the regression coefficients in FDL models capture immediate and cumulative impacts of regressors on the response variable. We also introduce and interpret models which include lagged values of the dependent variable as regressors at this point.

In Section 10.3 in Wooldridge, the finite sample properties of the OLS estimator in models where regressors typically are correlated across time are examined. Assumptions TS.1–TS.3 are formulated under which the OLS estimator remains unbiased and we discuss the limitations on the economic environment imposed by the strict exogeneity assumption TS.3. We formulate additional assumptions TS.4 and TS.5 which together with TS.1–TS.3 guarantee the validity of the formula for the variance of the OLS estimator and also guarantee that the OLS estimator is BLUE. Finally, we discuss that under the additional assumption on the normal distribution of the errors, we can conduct finite sample (or exact) statistical inference.

In Chapter 11 in Wooldridge, the large sample properties of the OLS estimator with time series data are discussed. As argued in Section 11.1, we require the stochastic processes to be stationary and weakly dependent. We first discuss these two concepts and give examples of some commonly used time series processes exhibiting dependence, such as the autoregressive process (AR) and moving average (MA) processes of order 1. We analyse their dependence and stationarity properties in Section 11.1b.

In Section 11.2 we learn that the consistency of the OLS estimator does not require us to assume strict exogeneity (deemed unreasonable in time series models). Instead, we can use a weaker, more reasonable, contemporaneous exogeneity assumption (TS.3′). Under this more reasonable assumption of contemporanous exogeneity, given a form of homoskedasticity (TS.4′) and no serial correlation (TS.5′), we can continue to rely on the usual t tests, F tests and confidence intervals (asymptotically valid).

13.2 Content of chapter

13.2.1 The nature of time series data

Read: Wooldridge, Section 10.1.

In Section 10.1 we are introduced to the use of time series data, where special attention needs to be given to the substantial differences between time series data and cross-sectional data.

The first key aspect that distinguishes time series data from cross-sectional data is that there is a clear temporal ordering. The temporal order indicates that observations should not be arbitrarily reordered as for many purposes we must know that the data for the time period t immediately precedes the data for the time period $t + 1$. When analysing time series data in the social sciences, we also must recognise that the past can affect the future, but not vice versa.

The second key aspect is that time series data are almost always correlated across time (that is they exhibit dependence) sometimes even strongly. This means that we can no longer rely on the assumption of random sampling (MLR.2) commonly made when using cross-sectional data.

As we will discuss in the final chapter of this course, particular features of time series data require special attention. First, when we collect time series at monthly or quarterly frequencies they can exhibit **seasonality**. Examples of time series data with seasonality include (and are not limited to) Christmas effect on expenditures, effectiveness of fertiliser on production, and ice cream sales. Second, many time series processes exhibit **trends**. Examples of time series with trends include (and are not limited to) output per labour hour, and nominal imports (see Wooldridge). These features (nonstationarity) will be relegated to the final chapter of the subject guide. For ease of exposition, we will ignore these issues for now.

Summary of some terminology used in the time series context

The observation on the time series variable y made at date t is denoted y_t. The interval between observations, that is, the period of time between observation t and observation $t + 1$, is some unit of time such as weeks, months, quarters, years etc. For instance, if the unit is years and $t = 2010$, then $t + 1 = 2011$, $t + 2 = 2012$, $\ldots$, $t + 25 = 2035$, and so on.

Special terminology and notation are used to indicate future and past values of y. The value of y in the previous period is called its first lagged value or its **first lag**, and is denoted y_{t-1}. Its jth lagged value (or its **jth lag**) is its value j periods ago, which is y_{t-j}. Similarly, y_{t+1} denotes the value of y one period into the future.

> **Activity 13.1** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 13.2** In this activity, we will use unemployment data for the period 1992–2001 given in the table below:
>
> | year | *unem* |
> |---|---|
> | 1992 | 6.9 |
> | 1993 | 7.2 |
> | 1994 | 7.4 |
> | 1995 | 7.1 |
> | 1996 | 6.3 |
> | 1997 | 6.2 |
> | 1998 | 6.0 |
> | 1999 | 5.9 |
> | 2000 | 5.8 |
> | 2001 | 6.1 |
>
> (i) What are the values of $unem_t$ for $t = 1992$, 1993, 1999?
>
> (ii) What are the values of $unem_{t-1}$ for $t = 1992$, 1993, 1999?

(iii) What are the values of $unem_{t-2}$ for $t = 1992, 1993, 1999$?

(iv) What are the values of $unem_{t+1}$ for $t = 1992, 1993, 1999$?

13.2.2 Examples of time series regression models

Read: Wooldridge, Section 10.2 and Example 10.4.

In Section 10.2 in Wooldridge two time series models are discussed that can easily be estimated by OLS: the static model and the distributed lag (DL) model. Here, we will briefly introduce another common model (relegated to Section 11.2 by Wooldridge), where we allow lagged dependent variables to affect the dependent variable as well, the so-called autoregressive distributed lag (ADL) model. You need to be able to interpret the results from such regression in empirical examples. In particular, we will need to distinguish between short-run and long-run effects.

Static model

The simplest model relates the outcome on the dependent variable at time t, y_t, to one or more explanatory variables **dated at the same time**. A static model is used to estimate a **contemporaneous relation** (hence, the name 'static model').

With just one explanatory variable z_t, we have:

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

such as the *static Phillips curve* example in Wooldridge. It cannot capture effects that take place with a lag. The effect is instantaneous (contemporaneous).

Distributed Lag Model (DL)

In finite distributed lag models, the explanatory variables are allowed to influence the dependent variable with a time lag (see also Example 10.4 in Wooldridge where the effect of personal exemption on fertility rates is considered). Allowing for a two-year effect on the general fertility rate (gfr):

$$\underbrace{gfr_t}_{\text{Children born per 1,000 women in year t}} = \alpha_0 + \delta_0 \underbrace{pe_t}_{\text{Tax exemption in year t}} + \delta_1 \underbrace{pe_{t-1}}_{\text{Tax exemption in year t - 1}} + \delta_2 \underbrace{pe_{t-2}}_{\text{Tax exemption in year t - 2}} + u_t$$

The fertility rate example is a special case of a **finite distributed lag model** with two lags (in addition to the contemporaneous variable):

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

In such a model, we think that a change in a variable, z, today, can affect y up to two periods into the future, denoted DL(2). To see this, consider the following.

- We can write this equation from the perspective of period $t+1$ (one period ahead):

$$y_{t+1} = \alpha_0 + \delta_0 z_{t+1} + \delta_1 z_t + \delta_2 z_{t-1} + u_{t+1}.$$

- From the perspective of period $t+2$ (two periods ahead) this equation is:

$$y_{t+2} = \alpha_0 + \delta_0 z_{t+2} + \delta_1 z_{t+1} + \delta_2 z_t + u_{t+2}.$$

Such models are good for estimating lagged effects of z (say, a particular policy).

For example, we may be interested in the effect of minimum wage on employment. Assuming that we have monthly data on employment and the minimum wage, then we may expect that the effect of a change in the minimum wage will take several months to have its full effect on employment. We may want to specify an FDL model here with more than two lags.

In general, we can use an FDL of order q, denoted DL(q) (see Wooldridge). Whereas for a practical matter, the choice of q can be hard, its choice is often dictated by the frequency of data. With annual data, q is usually small. With monthly data, q is often chosen as 12 or 24 or even higher, depending on how many months of data we have. Under some assumptions we can use an F test to see if additional lags are jointly significant.

The slope parameters in our DL model are called the distributed lag parameters which you need to be able to interpret. In particular, for our DL(2) model:

$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

- The coefficient on the contemporaneous variable z, δ_0, provides the immediate change in y due to the one-unit increase in z at time t. It is called:

$$\text{Impact Propensity} = \delta_0.$$

- The sum of all distributed lag coefficients gives us the answer to the following thought experiment. Suppose the level of z increases **permanently** today (for example, the minimum wage increases by \$1.00 per hour and stays there). The sum provides the (*ceteris paribus*) change in y after the change in z has passed through all q time periods. It is called the:

$$\text{Long Run Propensity} = \delta_0 + \delta_1 + \delta_2.$$

  You are advised to read the discussion in Example 10.4 in Wooldridge carefully. It provides details to show how you can obtain the standard error of the estimated long run propensity: $\widehat{\delta}_0 + \widehat{\delta}_1 + \widehat{\delta}_2$ using a simple reparameterisation of the model.

- We typically can obtain very precise (low variance) parameter estimates of the long run propensity (LRP). Estimates of the distributed lag parameters individually are typically not estimated very precisely due to the problem of near multicollinearity: if $\{z_t\}$ is slowly moving over time, then z_t, z_{t-1}, and z_{t-2} can be highly correlated.

A few remarks

- Notice that if z increases by one unit today, but then falls back to its original level in the next period (that is, we are dealing with a **transitory** shock), the lag distribution tells us how y changes in each future period. Eventually y falls back to its original level with a temporary change in z.
- With a **permanent change** in z, y changes to a new level, and the change from the old to the new level is the LRP.
- We can, of course, have more than one variable appear with multiple lags. For example, a simple equation to explain how the Federal Reserve Bank in the U.S. changes the Federal Funds Rate is:

  $$\begin{aligned} \mathit{ffrate}_t = \alpha_0 + \delta_0 \mathit{inf}_t + \delta_1 \mathit{inf}_{t-1} + \delta_2 \mathit{inf}_{t-2} \\ + \gamma_0 \mathit{gdpgap}_t + \gamma_1 \mathit{gdpgap}_{t-1} + \gamma_2 \mathit{gdpgap}_{t-2} + u_t \end{aligned}$$

  where *inf* is inflation rate and *gdpgap* is the GDP gap (measured as a percentage).
- FDLs are often more realistic than static models, and they can do better forecasting because they account for some dynamic behaviour. Nevertheless, they are not usually the most preferred for forecasting because they do not allow lagged outcomes on y to directly affect current outcomes.

Autoregressive Distributed Lag model (ADL)

With time series data, there is also the possibility of allowing past outcomes on y to affect current y. The simplest model is:

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

which is a simple regression model for time series data where the explanatory variable at time t is y_{t-1} (see also Section 11.2 in Wooldridge). It is called an **autoregressive model of order 1**, or **AR(1)**. This simple model typically does not have much economic or policy interest because we are just using lagged y to explain current y. Nevertheless, autoregressive models can be remarkably good at forecasting, even compared with complicated economic models. It is easy to add other explanatory variables along with a lag. For example, we can consider:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_t + u_t.$$

This is an example of the so-called **Autoregressive Distributed Lag (ADL)** model, that combines autoregressive and finite distributed lag models. In this particular example, it is an ADL$(1, 0)$ model: the dependent variable y depends on one lag of itself; y also depends on the current value of an explanatory variable z (0 lags).

- Here β_2 measures the effect of changing z_t on y_t, holding fixed y_{t-1}. It is a kind of *short-run effect* of z on y.

  Controlling for y_{t-1} while estimating the effect of z_t can be effective for estimating the causal effect of z_t on y_t: it recognises that the policy variable (z_t) may be correlated with y_{t-1}. Since y_{t-1} is likely a relevant regressor, its omission will cause OVB if z_t is related to y_{t-1}.

- The *long-run effect* of z on y is given by $\beta_2/(1-\beta_1)$. In order to see this, assume that $|\beta_1| < 1$. Let z^* denote the equilibrium (or, long-run) value of z_t and let y^* be the equilibrium (or, long-run) value of y_t, such that:

  $$y^* = \beta_0 + \beta_1 y^* + \beta_2 z^*$$

  (the error term is taken to be its mean value zero in the long-run). Rewriting this relation yields:

  $$y^* = \frac{\beta_0}{1-\beta_1} + \frac{\beta_2}{1-\beta_1} z^*.$$

Final remark

- Statistically, models with lagged dependent variables are more difficult to study. For one, the OLS estimators are no longer unbiased under any assumption, so large-sample analysis is very important. (As we will discuss, large-sample analysis is critical for static and FDL models too, but at least a finite-sample analysis makes sense sometimes.)

Activity 13.3 Take the MCQs related to this section on the VLE to test your understanding.

Activity 13.4 Consider the following model with a lagged dependent variable:

$$inf_t = \beta_0 + \beta_1 inf_{t-1} + \beta_2 unem_{t-1} + u_t$$

where $inft_t$ is the inflation rate in period t and $unem_t$ is unemployment in period t. Let $t = 1$ denote the first period for which the full set of regressors is available (with this convention, inf_0 and $unem_0$ are observed). Assume that u_t is independent of all lagged inf_{t-s}, $s \geq 1$, and all lagged $unem_{t-s}$, $s \geq 1$.

Discuss whether in this model it is possible for the following condition to be satisfied:

$$E(u_t \mid inf_0, \ldots, inf_t, \ldots, inf_n, unem_0, \ldots, unem_t, \ldots, unem_n) = 0$$

for any period t. (If this conditions holds, it implies that unobservable u_t in period t is uncorrelated with all current, past and future values of the regressors.)

Activity 13.5 Exercise 10.6 from Wooldridge.

13.2.3 Finite sample properties of OLS under classical assumptions

Read: Wooldridge, Section 10.3.

In Section 10.3, the assumptions required to establish the same kind of finite-sample properties (unbiasedness, variance calculations, normality) in the time-series setting as we had with the cross-sectional setting are discussed.

Let the time series process: $\{(y_t, x_{t1}, \ldots, x_{tk}) : t = 1, \ldots, T\}$ satisfy the following model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + u_t, \quad t = 1, \ldots, T.$$

The assumptions that will ensure the OLS estimators for time series regression are unbiased are given by Assumptions TS.1 to TS.3 (see Wooldridge).

Assumption TS.3 (strict exogeneity) is very restrictive

As we can no longer rely on random sampling (MLR.2), the assumption that guarantees uncorrelatedness between the errors and regressors now *explicitly needs to rule out any correlation between the error in period t, u_t, and the regressors x_{sj}, for $j = 1, \ldots, k$ in any period s even when the time periods s and t do not match up* (s could coincide with t but could also be in the past or in the future of period t). This assumption is called the assumption of **strict exogeneity**.

$$\text{TS.3} \qquad E(u_t \,|\, x_{s1}, \ldots, x_{sk}) = 0 \text{ for all } s, t.$$

In the cross-sectional setting, we made a similar assumption, whereby:

$$\text{TS.3}' \qquad E(u_t \,|\, x_{t1}, \ldots, x_{tk}) = 0.$$

This assumption, we will label (TS.3′) *only rules out correlation between the error in period t and the regressors x_{sj}, for $j = 1, \ldots, k$ when $s = t$.* The random sampling in the cross-sectional setting, automatically rules out correlations between u_t and x_{sj}, for $j = 1, \ldots, k$ with $s \neq t$.

Assumption TS.3 (strict exogeneity) rules out some practically important situations

1. TS.3 can **never** be truly satisfied when a lagged dependent variable, y_{t-1}, is included among the regressors, x_{tj}.

   To see this: let $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$. By lagging this equation one period, we see that y_{t-1} is a function of u_{t-1}. Hence u_{t-1} cannot be mean independent of y_{t-1}, and equivalently u_t cannot be mean independent of y_t (can use any period). The requirement $E(u_t \,|\, y_1, \ldots, y_n) = 0$ therefore fails necessarily.

2. It imposes important restrictions in static and FDL models as well. Consider:

   $$y_t = \alpha_0 + \delta_1 z_t + \delta_2 z_{t-1} + \delta_2 z_{t-2} + u_t.$$

   The Assumption TS.3 requires that u_t is not only uncorrelated with z_t, z_{t-1} and z_{t-2}, but also needs to be uncorrelated with all past **and** future outcomes on z.

   Uncorrelatedness of u_t with future values of z requires that z_{t+1} does not react to changes in u_t (that is, to changes in the unobserved part of y_t). Whether this may hold or not, depends on the reason why z_t changes over time. For example, say we are interested in estimating the effect of a change in minimum wages on employment. It may be reasonable to expect that minimum wage changes (future) are a response to shocks in employment (past).

If we add assumptions of homoskedasticity (TS.4) and no serial correlation (TS.5), OLS is again BLUE (efficient) and the usual OLS variance formula holds, that is:

$$Var(\widehat{\beta}_j \,|\, x_1, \ldots, x_n) = \frac{\sigma^2}{SST_j(1 - R_j^2)} \quad \text{for } j = 1, \ldots, k.$$

To permit us to perform exact inference (use t and F tests) we need to add a normality assumption again. For inference purposes, it is important to recognise that the presence of heteroskedasticity or serial correlation will invalidate the simple formula for the variance and associated standard error and we should use robust (HAC – heteroskedasticity and autocorrelation consistent) standard errors if we are concerned about the validity of TS.4 or TS.5 (we will discuss this in more detail later).

It is important to distinguish three different kinds of correlations that may arise in time series regression:

1. The explanatory variables, x_{tj}, might be correlated over time.
   - This is almost always true. For example, in the dataset PHILLIPS the correlation between $unem_t$ and $unem_{t-1}$ is 0.752. TS.2 only rules out **perfect** correlation in the regressors.
2. Correlation between x_{sj} and u_t (s and t are any two time periods).
   - If the errors and any of the regressors are correlated, TS.3 is violated and OLS is biased. So, we want to rule out this kind of correlation.
3. Correlation between u_s and u_t (s and t are any two time periods).
   - This is the problem of serial correlation ruled out in TS.5. The presence of serial correlation itself does **not** cause bias in the OLS estimators but it is imposed to obtain the familiar variance formula.

Summary of finite-sample analysis of OLS for time series data

1. Under TS.1–TS.3, the OLS estimator is unbiased. The strict exogeneity assumption for the explanatory variables (TS.3) is key but it rules out some interesting cases.
2. When we add homoskedasticity (TS.4) and no serial correlation (TS.5) assumptions, OLS is BLUE and the usual OLS variance formulae hold. Unfortunately, in models where Assumption TS.3 (strict exogeneity) has a chance of holding – namely, static and finite distributed lag models – serial correlation is often a problem.
3. If we add normality (TS.6), then exact inference is possible. The classical linear model assumptions for time series data are much more restrictive than those for cross-sectional data – in particular, strict exogeneity (TS.3) and no serial correlation (TS.5) assumptions can be unrealistic. Nevertheless, the CLM framework is a good starting point for many applications.

> **Activity 13.6** Take the MCQs related to this section on the VLE to test your understanding.

13.2.4 Stationary and weakly dependent time series

Read: Wooldridge, Section 11.1.

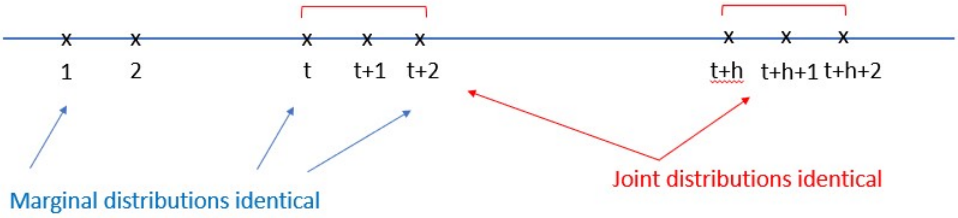
As we discussed, the assumptions used to derive the finite sample properties of the OLS estimator in the time series context are rather restrictive. Violation of strict exogeneity (TS.3) and serial correlation (TS.5) are commonplace in time series models. We also relied on the normality assumption (TS.6) for statistical inference.

As in the cross-sectional setting, we can expect that weaker assumptions are needed if the sample size is large. In the cross-sectional case, OLS inference is **approximately valid** even without a normality assumption, and we also know how to adjust our statistics for the presence of heteroskedasticity of unknown form (please recall the use of robust standard errors). In the cross-sectional setting, we relied on random sampling to justify these results. As we can no longer rely on random sampling when using time series data, we must impose alternative assumptions on the underlying time series processes ($\{\mathbf{x}_t\}$ and $\{u_t\}$).

Key requirements for large sample analysis of time series

1. **Weak dependence.** While it is unreasonable to assume independence in TS, we will need to limit this dependence (replacing *'independent'*).

2. **Stationarity.** We will require that the joint distribution of $(x_{t_1}, \ldots, x_{t_m})$ is identical to that of $(x_{t_1+h}, \ldots, x_{t_m+h})$ for any h (replacing '*identically distributed*').
   - Stationarity requires the future to be like the past in a probabilistic sense. The assumption is much stronger than identically distributed (corresponds to the case where $m = 1$) as it includes the dependence structure over time.
   - Schematically, this assumption can be visualised as requiring:


1
2
t
t+1
t+2
t+h
t+h+1
t+h+2
Marginal distributions identical
Joint distributions identical

These concepts are discussed in Section 11.1 in Wooldridge. The importance of these concepts is that with weak dependence and stationarity, we can again apply the law of large numbers and central limit theorem to justify the OLS approximate inference, here:

- Law of Large Numbers (LLN):

$$\text{plim } \bar{Y}_T \equiv \text{plim } \frac{1}{T}\sum_{t=1}^{T} Y_t = \mu, \quad \text{where } \mu = E(Y_t).$$

- Central Limit Theorem (CLT):

$$\frac{\bar{Y}_T - \mu}{\sigma/\sqrt{T}} \overset{a}{\sim} N(0,1), \quad \text{where } \mu = E(Y_t),\ \sigma^2 = Var(Y_t).$$

Stationary time series

A time series process is stationary if its stochastic properties and its temporal dependence stucture do not change over time (see definition above). For this course it will suffice to be familiar with a weaker form of stationarity, called **covariance stationarity**. Covariance stationarity requires that the first two moments of the stochastic process – that is, the mean and variance – do not change (are identical) over time. Moreover, it requires that the covariance between x_t and x_{t+h} depends only on the distance in time between the two terms, h, and not on the location of the initial time period, t. This, of course, implies that the correlation between x_t and x_{t+h} also depends only on h:

Covariance stationarity: A stochastic process $\{x_t : t = 1, 2, \ldots\}$ is covariance stationary if:

(i) $E(x_t)$ is constant and finite

(ii) $Var(x_t)$ is constant and finite

(iii) $Cov(x_t, x_{t+h})$ depends only on the distance in time h, not location in time t.

Weakly dependent time series

Stationarity describes the stability of the joint distribution of a process as it moves through time. A very different concept is that of **weak dependence**, which places restrictions on how strongly related the random variables x_t and x_{t+h} can be as the time distance between them, h, increases.

To describe the dependence over time, we can use either the covariance, $Cov(x_t, x_{t+h})$, or correlation $Corr(x_t, x_{t+h})$ as a function of h. Recall:

$$Corr(x_t, x_{t+h}) = \frac{Cov(x_t, x_{t+h})}{\sqrt{Var(x_t)}\sqrt{Var(x_{t+h})}}.$$

The assumption of independence in the process implies $Corr(x_t, x_s) = 0$ for any $t \neq s$, which is a condition which is too restrictive for time series. Loosely speaking (the formal definition is technical), **weak dependence** means:

$$Corr(x_t, x_{t+h}) \to 0 \quad \text{as } h \to \infty$$

that is, $Corr(x_t, x_{t+h})$ should vanish eventually or asymptotically. Weak dependence is also referred to as **asymptotic independence**.

Two leading examples discussed in Section 11.1b that describe, quite distinct, dependence structures in time series processes are the moving error process of order 1, MA(1) and the autoregressive process of order 1, AR(1). You should be able to define these processes and be able to analyse covariance stationarity and weak dependence properties for them. This discussion will be a bit mathematical but it is an extremely useful tool not only for econometrics but also for any science talking about dynamics (macroeconomics, finance, biology, physics etc.). Before the journey, let's recall:

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)\,E(Y) \\ Var(X) &= E(X^2) - E(X)^2. \end{aligned}$$

If X and Y are independent, then:

$$Cov(X,Y) = 0.$$

MA(1) process

The process $\{y_t\}$ is a moving average process of order 1 when:

$$y_t = e_t + \theta e_{t-1}, \quad \text{for } t = 1, 2, \ldots$$

where $\{e_t : t = 0, 1, \ldots\}$ is an i.i.d. sequence with zero mean and variance σ^2.

Note that this model only has error terms on the right-hand side of the equation. Intuitively, our outcome is formed of only a random shock in the current and previous period.

To show that $\{y_t\}$ is *covariance stationary* we need to check the three conditions defining covariance stationarity.

(i) Mean:

$$E(y_t) = E(e_t) + \theta E(e_{t-1}) = 0 \quad \text{for all } t.$$

(ii) Variance:

$$\begin{aligned} Var(y_t) = Var(e_t + \theta e_{t-1}) &\overset{indep}{=} Var(e_t) + \theta^2\,Var(e_{t-1}) \\ &= (1+\theta^2)\sigma^2 \quad \text{for all } t \end{aligned}$$

as $\{e_t\}$ is an i.i.d. sequence with zero mean and variance σ^2.

(iii) Finally, we can show:

$$Cov(y_t, y_{t+h}) = \begin{cases} \theta\sigma^2 & \text{if } h = 1 \\ 0 & \text{if } h \geq 2 \end{cases}$$

hence the MA(1) process is **covariance stationary**. Covariance only changes as a function of distance, h, not location in time, t.

To calculate the covariance $Cov(y_t, y_{t+h})$, the key is to note that:

$$Cov(e_t, e_s) = E(e_t e_s) = \begin{cases} 0 & \text{if } t \neq s \\ \sigma^2 & \text{if } t = s \end{cases}$$

because $\{e_t\}$ is i.i.d.

To compute the covariance between y_t and y_{t+1}, we note:

$$\begin{aligned} Cov(y_t, y_{t+1}) = E(y_t y_{t+1}) &= E[(e_t + \theta e_{t-1})(e_{t+1} + \theta e_t)] \\ &= E(e_t e_{t+1}) + \theta E(e_t^2) + \theta E(e_{t-1} e_{t+1}) + \theta^2 E(e_{t-1} e_t) \\ &= \theta \sigma^2. \end{aligned}$$

Similarly, for $Cov(y_t, y_{t+2})$, we obtain:

$$\begin{aligned} Cov(y_t, y_{t+2}) = E(y_t y_{t+2}) &= E[(e_t + \theta e_{t-1})(e_{t+2} + \theta e_{t+1})] \\ &= E(e_t e_{t+2}) + \theta E(e_t e_{t+1}) + \theta E(e_{t-1} e_{t+2}) + \theta^2 E(e_{t-1} e_{t+1}) \\ &= 0. \end{aligned}$$

Analogously, we can show that $Cov(y_t, y_{t+h}) = 0$ for any $h \geq 2$.

We can conclude from these derivations that $\{y_t\}$ is *weakly dependent*. In particular:

$$Corr(y_t, y_{t+1}) = \frac{Cov(y_t, y_{t+1})}{\sqrt{Var(y_t)}\sqrt{Var(y_{t+1})}} = \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{1+\theta^2}$$

and:

$$Corr(y_t, y_{t+h}) = \frac{Cov(y_t, y_{t+h})}{\sqrt{Var(y_t)}\sqrt{Var(y_{t+h})}} = 0, \quad \text{for } h \geq 2.$$

As variables in the process that are more than one period apart are uncorrelated, we have established weak dependence.

All MA(1) processes (for any value of θ) are stationary and weakly dependent. In fact, this result generalised to all finite order MA(q) processes.

AR(1) process

The process $\{y_t\}$ is an autoregressive process of order 1 when:

$$y_t = \rho y_{t-1} + e_t \quad \text{for } t = 1, 2, \ldots$$

where $\{e_t : t = 1, \ldots\}$ is an i.i.d. sequence with zero mean and variance σ^2.

If we assume that the starting point in the sequence is y_0 (at $t = 0$) and we use so-called repeated substitution we can write:

$$\begin{aligned} y_t &= \rho y_{t-1} + e_t = \rho(\rho y_{t-2} + e_{t-1}) + e_t \\ &= \rho^2 y_{t-2} + \rho e_{t-1} + e_t = \rho^2(\rho y_{t-3} + e_{t-2}) + \rho e_{t-1} + e_t \\ &= \rho^3 y_{t-3} + \rho^2 e_{t-2} + \rho e_{t-1} + e_t \\ &\vdots \\ y_t &= \rho^t y_0 + \rho^{t-1} e_1 + \cdots + \rho e_{t-1} + e_t. \end{aligned} \tag{*}$$

We assume that the e_ts are independent of y_0 and $E(y_0) = 0$.

By taking expectation of the expression in (*), we can then show $E(y_t) = 0$.

By taking the variance of the expression in (*), we get:

$$\begin{aligned} Var(y_t) &= \rho^{2t} Var(y_0) + \rho^{2(t-1)} Var(e_1) + \cdots + \rho^2 Var(e_{t-1}) + Var(e_t) \\ &= \rho^{2t} Var(y_0) + (1 + \rho^2 + \cdots + \rho^{2(t-1)})\sigma^2. \end{aligned}$$

To ensure the variance of y_t exists, we will need to assume that $|\rho| < 1$. With $|\rho| < 1$, ρ^s gets closer to zero when s increases.

In order for the AR(1) process $\{y_t\}$ to be *covariance stationary* we need to assume $|\rho| < 1$. Required to ensure the variance exists as discussed above.

The easiest way to derive the mean, variance and covariances of an AR(1) process *makes use of the properties of a covariance stationary process.*

- **Mean:** By the definition of y_t: $E(y_t) = E(\rho y_{t-1} + e_t)$. As $E(e_t) = 0$, this gives:

  $$E(y_t) = \rho E(y_{t-1}) \quad \text{or} \quad E(y_t) - \rho E(y_{t-1}) = 0.$$

  As $E(y_t) = E(y_{t-1})$ by covariance stationarity, we can write this as $(1 - \rho)E(y_t) = 0$. Since $\rho \neq 1$, this reveals $E(y_t) = 0$.

- **Variance:** Using the independence between e_t and y_{t-1}:

  $$Var(y_t) = Var(\rho y_{t-1} + e_t) = \rho^2 Var(y_{t-1}) + Var(e_t).$$

  As $Var(y_t) = Var(y_{t-1})$ by covariance stationarity and $Var(e_t) = \sigma^2$, this yields:

  $$Var(y_t) = \rho^2 Var(y_t) + \sigma^2 \quad \text{or} \quad (1 - \rho^2) Var(y_t) = \sigma^2.$$

  Since $\rho \neq 1$, this reveals $Var(y_t) = \sigma^2/(1 - \rho^2)$.

- **Covariance:** Let us look at $Cov(y_t, y_{t+1})$. Using the definition of y_{t+1}:

  $$Cov(y_t, y_{t+1}) = Cov(y_t, \rho y_t + e_{t+1}) = \rho Var(y_t) + Cov(y_t, e_{t+1}).$$

  Using the independence of e_{t+1} from y_t, we obtain:

  $$Cov(y_t, y_{t+1}) = \rho Var(y_t)$$

  which gives us:

  $$Corr(y_t, y_{t+1}) = \rho.$$

Covariance stationary AR(1) processes are weakly dependent. With:

$$Corr(y_t, y_{t+h}) = \rho^h \quad \text{where } |\rho| < 1$$

it is clear that as $h \to \infty$, $Corr(y_t, y_{t+h}) \to 0$ as required (quite fast).

Violations of stationarity and weak dependence will be discussed in the last chapter. For now, we will assume that our time series processes are stationary and weakly dependent to permit large-sample analysis and inference.

> **Activity 13.7** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 13.8** Let $\{e_t : t = -1, 0, 1, \ldots\}$ be a sequence of independent and identically distributed random variables with mean zero and variance σ^2.
>
> Consider an MA(2) process:
>
> $$y_t = e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}.$$
>
> (i) Find $E(y_t)$ and $Var(y_t)$. Do either of these depend on t?
>
> (ii) Find $Cov(y_t, y_{t+h})$ for any $h \geq 1$.
>
> (iii) Find $Corr(y_t, y_{t+h})$ for any $h \geq 1$. Is this process weakly dependent?

13.2.5 Asymptotic properties of OLS

Read: Wooldridge, Section 11.2.

As discussed, in order to analyse the large sample properties of the OLS estimator with time series data we need to introduce the assumption that our stochastic processes $\{(y_t, x_t) : t = 1, 2, \ldots\}$ are stationary and weakly dependent (TS.1$'$) which permits us to rely on the usual LLN and CLT. The latter is important as we also no longer will rely on a normality assumption of the errors.

In Section 11.2 in Wooldridge it is established that the consistency of our OLS estimator in time series models does *not* require the strong assumption of strict exogeneity (TS.3), but instead permits us to rely on the more reasonable assumption of contemporaneous exogeneity (TS.3$'$). You should be able to show this result in a simple linear regression setting (using the plim operator and standard LLN arguments).

Let us briefly look at Assumption TS.3$'$ here.

This assumption says that all regressors are contemporaneous exogenous, i.e. we have:

$$E(u_t \,|\, \mathbf{x}_t) = 0$$

where $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$.

Assumption TS.3$'$ implies uncorrelatedness between the errors (u_t) and the contemporaneous regressors $(\mathbf{x}_t)$.

- If $\mathbf{x}_t = (z_t, z_{t-1}, z_{t-2})$, u_t needs to be uncorrelated with z_t, z_{t-1} and z_{t-2}.
- If $\mathbf{x}_t = (y_{t-1}, z_t, z_{t-1}, z_{t-2})$, u_t needs to be uncorrelated with y_{t-1}, z_t, z_{t-1} and z_{t-2}.

Contemporaneous exogeneity **does not restrict** correlations between the error and explanatory variables across other time periods. In the same examples, there may be correlation between u_t and z_{t+1}.

The advantage of TS.3′ is that it allows for lagged dependent variables and explanatory variables that react to *past* changes in y.

Consistent but biased under TS.1′ to TS.3′

To give a specific example. Consider the following ADL(1, 1) model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 z_{t-1} + u_t$$

where all variables are weakly dependent (which requires $|\beta_1| < 1$), and:

$$E(u_t \,|\, y_{t-1}, z_{t-1}) = 0.$$

Assumption TS.3′ is satisfied, hence the OLS estimator is consistent. As the model includes the lagged dependent variable y_{t-1} as a regressor, the OLS estimator will be biased as the assumption of strict exogeneity, TS.3, cannot possibly hold.

Comment: The consistency result provides justification for estimating models where regressors are **not strictly exogenous** (models with lagged dependent variables and other non-strictly exogenous variables). While we cannot have unbiased estimators, they are consistent under fairly weak assumptions.

For statistical inference we again rely on some form of homoskedasticity and no serial correlation. The exact form of these conditions, presented in TS.4′ and TS.5′, is not that important. The main result is given in Theorem 11.2 (Wooldridge). The upshot of this theorem is that **we can use approximate inference for regression with time series like we do with cross-sections.** Empirical examples are provided in Section 11.2.

It is important to recognise that the standard errors (and hence inference) are invalid in the presence of serial correlation (or heteroskedasticity) in the errors. We look at this in more detail in the next chapter.

> **Activity 13.9** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 13.10** Consider an AR(1) model:
>
> $$y_t = \rho y_{t-1} + e_t$$
>
> where $\{e_t : t = 1, 2, \ldots\}$ is an i.i.d. sequence with zero mean and variance σ^2 and $|\rho| < 1$.
>
> The starting point in the sequence is y_0 (at $t = 0$). We also assume that y_0 is independent of $\{e_t\}$ and $E(y_0) = 0$.
>
> Let $\widehat{\rho}$ denote the OLS estimator of ρ in this model.
>
> (i) Is estimator $\widehat{\rho}$ consistent? Explain in detail.
>
> (ii) Is estimator $\widehat{\rho}$ unbiased? Explain in detail.

Activity 13.11 **Strict and contemporaneous exogeneity**

Think about a regression of sales of ice cream on advertising, where the error comes from the effect of weather (weather is unobserved by the researcher, but not to the ice cream seller or their customers). In other words, we want to study the effect of advertising on sales:

$$\log(sales_t) = \beta_0 + \beta_1 advert_t + u_t$$

where u_t are weather conditions which are unobserved by the researcher. Assume for the sake of simplicity that weather has no persistence across days.

(i) Suppose the ice cream seller uses advertising to smooth the effect of weather, so days with high advertising are also cooler days and low advertising days are hotter. Discuss whether assumptions TS.3 and TS.3′ hold. Are the OLS estimators of β_0 and β_1 unbiased? Are these OLS estimators consistent?

(ii) Now suppose that advertising in period t cannot respond to weather in the same period. In other words, knowing the advertising level in period t tells us nothing about the weather in that period, on average. This seems reasonable as advertising takes time to roll out. Flyers need to be designed and printed and someone must be hired to hand them out. But advertising in period $t+1$ could be correlated with weather in t. It is also possible that advertising in period t is correlated with weather (or rather beliefs about it) in period $t+1$. Discuss whether assumptions TS.3 and TS.3′ hold. Are the OLS estimators of β_0 and β_1 unbiased? Are these OLS estimators consistent?

(iii) Now suppose that in addition to the current period t, advertising in future and past periods s does not tell us anything about weather in period t, so the ice cream seller cannot, for example, select advertising in period $t+1$ to compensate for the effect of cool weather in period t. Discuss whether assumptions TS.3 and TS.3′ hold. Are the OLS estimators of β_0 and β_1 unbiased? Are these OLS estimators consistent?

Activity 13.12 Exercise 11.6 part (i) and (ii) from Wooldridge.

13.3 Answers to activities

Solution to Activity 13.2

(i)
- The value of $unem_t$ for $t = 1992$ is 6.9.
- The value of $unem_t$ for $t = 1993$ is 7.2.
- The value of $unem_t$ for $t = 1999$ is 5.9.

(ii)
- The value of $unem_{t-1}$ for $t = 1992$ is not available (there is no data in the table for 1991).
- The value of $unem_{t-1}$ for $t = 1993$ is 6.9.
- The value of $unem_{t-1}$ for $t = 1999$ is 6.0.

(iii)
- The value of $unem_{t-2}$ for $t = 1992$ is not available (there is no data in the table for 1990).
- The value of $unem_{t-2}$ for $t = 1993$ is not available (there is no data in the table for 1991).
- The value of $unem_{t-2}$ for $t = 1999$ is 6.2.

Solution to Activity 13.4

We want to analyse whether for any $t \geq 1$, we have:

$$E(u_t \mid \mathit{inf}_0, \ldots, \mathit{inf}_t, \ldots, \mathit{inf}_n, unem_0, \ldots, unem_t, \ldots, unem_n) = 0. \qquad (*)$$

In this condition the mean of error u_t is conditioned on the regressors across all the time periods. One implication of this condition is that unobservable u_t in period t is uncorrelated with all current, past and future values of the regressors.

Note that inf_t is among $\mathit{inf}_0, \ldots, \mathit{inf}_t, \ldots, \mathit{inf}_n$. Since in the equation for period t:

$$\mathit{inf}_t = \beta_0 + \beta_1 \mathit{inf}_{t-1} + \beta_2 unem_{t-1} + u_t$$

variable inf_t is the dependent variable and is directly affected by u_t, then u_t cannot possibly be mean independent of inf_t.

Moreover, due to the presence of the *inf* lag on the right-hand side, error u_t, through its effect on inf_t, will also affect all future values inf_{t+s} (to see this, write the regression equation for periods $t + 1$, then $t + 2$, etc.) and, thus, cannot possibly be mean independent of inf_{t+1}, inf_{t+2}, ... either.

Another way to see that (*) cannot possibly hold is to note that (*) implies that u_t and inf_t are uncorrelated. However, using our model we can see that:

$$\begin{aligned} Cov(\mathit{inf}_t, u_t) &= Cov(\beta_0 + \beta_1 \mathit{inf}_{t-1} + \beta_2 unem_{t-1} + u_t, u_t) \\ &= Cov(\beta_0, u_t) + \beta_1 Cov(\mathit{inf}_{t-1}, u_t) + \beta_2 Cov(unem_{t-1}, u_t) + Cov(u_t, u_t) \\ &= 0 + 0 + 0 + Var(u_t) \\ &\neq 0. \end{aligned}$$

Therefore, it is not possible for (*) to be true.

Similar conclusions can be reached for AR(p) models and any ADL(p, q) model with $p \geq 1$.

Solution to Activity 13.5

(Exercise 10.6 from Wooldridge.)

(i) Given $\delta_j = \gamma_0 + \gamma_1 j + \gamma_2 j^2$, for $j = 0, 1, \ldots, 4$, we can write:

$$\begin{aligned} y_t &= \alpha_0 + \gamma_0 z_t + (\gamma_0 + \gamma_1 + \gamma_2) z_{t-1} + (\gamma_0 + 2\gamma_1 + 4\gamma_2) z_{t-2} + (\gamma_0 + 3\gamma_1 + 9\gamma_2) z_{t-3} \\ &\quad + (\gamma_0 + 4\gamma_1 + 16\gamma_2) z_{t-4} + u \\ &= \alpha_0 + \gamma_0(z_t + z_{t-1} + z_{t-2} + z_{t-3} + z_{t-4}) + \gamma_1(z_{t-1} + 2z_{t-2} + 3z_{t-3} + 4z_{t-4}) \\ &\quad + \gamma_2(z_{t-1} + 4z_{t-2} + 9z_{t-3} + 16z_{t-4}) + u_t. \end{aligned}$$

(ii) This is suggested in part (i). For clarity, define three new variables:

$$\begin{aligned} z_{t0} &= z_t + z_{t-1} + z_{t-2} + z_{t-3} + z_{t-4} \\ z_{t1} &= z_{t-1} + 2z_{t-2} + 3z_{t-3} + 4z_{t-4} \\ z_{t2} &= z_{t-1} + 4z_{t-2} + 9z_{t-3} + 16z_{t-4}. \end{aligned}$$

Then, $\alpha_0, \gamma_0, \gamma_1, \gamma_2$ are obtained from the OLS regression of y_t on z_{t0}, z_{t1}, and z_{t2}, where $t = 1, 2, \ldots, T$. (We can agree to let $t = 1$ denote the first time period where we have a full set of regressors.)

(iii) The unrestricted model is the original equation, which has six parameters (α_0 and the five δ_js). The polynomial distributed lag (PDL) model has four parameters. Therefore, there are two restrictions on moving from the general model to the PDL model. (Note how we do not have to actually write out what the restrictions are.)

The *df* in the unrestricted model is $T - 6$. Therefore, we would obtain the unrestricted R-squared, R^2_{ur}, from the regression of y_t on $z_t, z_{t-1}, \ldots, z_{t-4}$ and the restricted R-squared from the regression in part (ii), R^2_r.

The F statistic is:

$$F = \frac{R^2_{ur} - R^2_r}{1 - R^2_{ur}} \frac{T-6}{2}.$$

Under H_0 and the CLM assumptions, $F \sim F_{2,T-6}$.

Solution to Activity 13.8

(i) Let us first look at the expected value of y_t:

$$\begin{aligned} E(y_t) &= E(e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}) \\ &= E(e_t) + \theta_1 E(e_{t-1}) + \theta_2 E(e_{t-2}) \\ &= 0 + \theta_1 0 + \theta_2 0 \\ &= 0. \end{aligned}$$

We conclude from here that the expected value of y_t does not depend on t.

Next, we look at the variance of y_t:

$$Var(y_t) = Var(e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}).$$

As we know that $\{e_t\}$ is a sequence of random variables, we can write this (using the properties of the variance) as:

$$\begin{aligned}
Var(y_t) &= Var(e_t) + Var(\theta_1 e_{t-1}) + Var(\theta_2 e_{t-2}) \\
&= \underbrace{Var(e_t)}_{\sigma^2} + \theta_1^2 \underbrace{Var(e_{t-1})}_{\sigma^2} + \theta_2^2 \underbrace{Var(e_{t-2})}_{\sigma^2} \\
&= (1 + \theta_1^2 + \theta_2^2)\sigma^2.
\end{aligned}$$

The variance of y_t therefore also does not depend on t.

(ii) So let's think about this for a second before we just jump into computations. So y_t is defined as the linear combination of three innovations – so the one from the same time period and then from two previous periods. So, therefore, we can expect that the effect of any shock will last for two periods and then will completely disappear. That is we will expect $Cov(y_t, y_{t+h}) = 0$ for $h \geq 3$, whereas for h equal to one or two it will not be zero.

Let us show this formally, we start with $h = 1$. To find the covariance between y_t and y_{t+1} we just plug in the expressions for y_t and y_{t+1} from the definition of an MA(2) process.

$$Cov(y_t, y_{t+1}) = Cov(e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, e_{t+1} + \theta_1 e_t + \theta_2 e_{t-1}).$$

Because our innovation sequence $\{e_t\}$ is a sequence of independent random variables, we know that any two innovations in two different time periods are independent and therefore uncorrelated. So, for instance, you see that e_t in the first expression will be uncorrelated with e_{t+1} and will be uncorrelated with e_{t-1} in the second expression. We only have to pay attention to the variables in the first expression and the second expression that have the same indices:

$$\begin{aligned}
Cov(y_t, y_{t+1}) &= \theta_1 \underbrace{Cov(e_t, e_t)}_{\sigma^2} + \theta_1\theta_2 \underbrace{Cov(e_{t-1}, e_{t-1})}_{\sigma^2} \\
&= \theta_1(1 + \theta_2)\sigma^2.
\end{aligned}$$

Similarly, for $h = 2$:

$$\begin{aligned}
Cov(y_t, y_{t+2}) &= Cov(e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, e_{t+2} + \theta_1 e_{t+1} + \theta_2 e_t) \\
&= \theta_2 Cov(e_t, e_t) \\
&= \theta_2 Var(e_t) \\
&= \theta_2 \sigma^2.
\end{aligned}$$

While, for $h = 3$:

$$Cov(y_t, y_{t+3}) = Cov(e_t + \theta_1 e_{t-1} + \theta_2 e_{t-2}, e_{t+3} + \theta_1 e_{t+2} + \theta_2 e_{t+1}) = 0$$

as there are no common time indices in the two expressions.

Analogously, $Cov(y_t, y_{t+h}) = 0$ for $h \geq 3$.

(iii) We recall the definition of correlation:

$$Corr(y_t, y_{t+h}) = \frac{Cov(y_t, y_{t+h})}{\sqrt{Var(y_t)\, Var(y_{t+h})}}.$$

As the variance of y_t does not depend on t, this simplifies to:

$$Corr(y_t, y_{t+h}) = \frac{Cov(y_t, y_{t+h})}{Var(y_t)}.$$

Using the results from (i) and (ii) this yields:

$$\begin{aligned} Corr(y_t, y_{t+1}) &= \frac{\theta_1(1+\theta_2)\sigma^2}{(1+\theta_1^2+\theta_2^2)\sigma^2} = \frac{\theta_1(1+\theta_2)}{1+\theta_1^2+\theta_2^2} \\ Corr(y_t, y_{t+2}) &= \frac{\theta_2\sigma^2}{(1+\theta_1^2+\theta_2^2)\sigma^2} = \frac{\theta_2}{1+\theta_1^2+\theta_2^2} \\ Corr(y_t, y_{t+h}) &= 0 \quad \text{for } h \geq 3. \end{aligned}$$

So what we can conclude is that the dependence in the process only lasts for two periods. Therefore, we can conclude that the process is indeed weakly dependent.

Solution to Activity 13.10

(i) Let us verify TS.1′–TS.3′. If they hold, then $\widehat{\rho}$ is consistent.

TS.1′ (Linearity and Weak Dependence):

- AR(1) is a linear model. In period t, the dependent variable is y_t and the independent variable is y_{t-1}.
- As we discussed earlier, with $|\rho| < 1$ the AR(1) process $\{y_t\}$ is stationary and weakly dependent.
- Thus TS.1′ holds.

TS.2′ (No perfect collinearity):

- In the AR(1) model it means that y_t does not take the same value for each t.
- This is extremely likely in practice.
- In fact, it is almost impossible to imagine a scenario when this would not hold in AR(1).
- Thus TS.2′ holds.

TS.3′ (Contemporaneous exogeneity): $E(e_t \,|\, y_{t-1}) = 0$.

- In our earlier discussion of AR(1) we established that AR(1) implied that e_t is independent of $y_{t-1}, y_{t-2}, y_{t-3}$ (if you don't recall it, please go back and review that discussion).
- This implies that $E(e_t \,|\, y_{t-1})$ does not depend on y_{t-1}.

- Thus, $E(e_t \,|\, y_{t-1}) = E(e_t) = 0$ and, thus, we established TS.3$'$.

Since TS.1$'$–TS.3$'$ hold, the OLS estimator $\widehat{\rho}$ is consistent.

(ii) We need to verify assumptions TS.1–TS.3. If they all hold, then $\widehat{\rho}$ is unbiased.

TS.1 (Linearity):

- AR(1) is a linear model. In period t the dependent variable is y_t and the regressor is y_{t-1}.
- Note that TS.1 does not require checking stationarity and weak dependence.
- Thus TS.1 holds.

TS.2 (No Perfect Collinearity): same as TS.2$'$.

- We already checked it. Thus TS.2 holds.

TS.3 (Strict exogeneity): $E(e_t \,|\, y_0, y_1, \ldots, y_t, \ldots, y_n) = 0$.

- We already discussed several times that whenever we have lagged dependent variables as regressors, TS.3 cannot hold. It is the case here once again.
- Indeed, if strict exogeneity were true it would imply that e_t is uncorrelated with y_t. But this is certainly false because:

  $$Cov(y_t, e_t) = Cov(\rho y_{t-1} + e_t, e_t) = \rho Cov(y_{t-1}, e_t) + Cov(e_t, e_t) = 0 + Var(e_t) > 0.$$

  Thus TS.3 does not hold.

The OLS estimator $\widehat{\rho}$ is biased.

Solution to Activity 13.11

(i) In this situation the advertising $advert_t$ in period t is correlated with the error u_t (which contains weather):

$$Cov(advert_t, u_t) \neq 0.$$

This implies that neither TS.3 nor TS.3$'$ hold. Indeed, each of these assumptions would imply uncorrelatedness between advertising today and the weather today but this is clearly false in the described setting. Since TS.3 is key for the unbiasedness of OLS estimators and TS.3$'$ is key for the consistency of OLS estimators, we conclude that the OLS estimators are biased and inconsistent.

(ii) In this setting, for each period t:

$$E(u_t \,|\, advert_t) = 0$$

but $advert_{t+1}$ can be correlated with past u_t and also can be correlated with future u_{t+2}.

This means that contemporaneous exogeneity TS.3$'$ holds whereas strict exogeneity TS.3 does not (indeed, TS.3 does not allow u_t to be correlated with future or past *advert*). The violation of TS.3 immediately implies that the OLS estimators are biased. However, the OLS estimators are consistent if TS.1′ and TS.2′ hold in addition to TS.3′ – that is, if sequences of $\log(sales_t)$ and $advert_t$ are stationary and weakly dependent (we already have linearity in the model) and $advert_t$ is not the same value across time periods in the sample.

(iii) In this setting, for each period t:

$$E(u_t \mid advert_1, \ldots, advert_t, \ldots, advert_n) = 0.$$

This means that strict exogeneity TS.3 holds. Contemporaneous exogeneity TS.3′ holds too as it is implied by TS.3.

We can conclude that the OLS estimators are unbiased if TS.1 and TS.2 hold in addition to TS.3 – we already have linearity in the model and, thus, would only require that $advert_t$ is not the same value across time periods in the sample.

Also, the OLS estimators are consistent if TS.1′ and TS.2′ hold in addition to TS.3′ – that is, if sequences of $\log(sales_t)$ and $advert_t$ are stationary and weakly dependent (we already have linearity in the model) and $advert_t$ is not the same value across time periods in the sample.

Solution to Activity 13.12

(i) The t statistic for $H_0 : \beta_1 = 1$ is $t = (1.104 - 1)/.039 \approx 2.67$. We must rely on asymptotic results.

- Under H_0 and TS.1′–TS.5′, t has an approximate t_{121} distribution.
- The 1% critical value against a two-sided alternative is $t_{121,\,0.005} \approx 2.58$ and since $|2.67| > t_{121,\,0.005}$, we reject $H_0 : \beta_1 = 1$ against $H_1 : \beta_1 \neq 1$ at the 1% level.
- It is hard to know whether the estimate is practically different from one without comparing investment strategies based on the theory ($\beta_1 = 1$) and the estimate ($\widehat{\beta}_1 = 1.104$). But the estimate is 10% higher than the theoretical value.

(ii)
- The t statistic for testing $H_0 : \beta_1 = 1$ is now $(1.053 - 1)/.039 \approx 1.36$, so under TS.1′–TS.5′, $H_0 : \beta_1 = 1$ is no longer rejected against a two-sided alternative unless we are using quite a high significance level (which is not conventional).
- But the lagged spread is very significant under TS.1′–TS.5′ (contrary to what the expectations hypothesis predicts): $t = .480/.109 \approx 4.40$ for the null $H_0 : \beta_{r6_{t-1} - r3_{t-1}} = 0$.
- Based on the estimated equation, when the lagged spread is positive, the predicted holding yield on six-month T-bills is above the yield on three-month T-bills (even if we impose $\beta_1 = 1$), and so we should invest in six-month T-bills.

13.4 Overview of chapter

By now you have a firm grasp of multiple regression mechanics and inference. Therefore, in this segment we mostly focused on those features that make time series applications different from cross-sectional ones.

We first introduced and discussed *static and finite distributed lag* (FDL, or just DL) models. Many interesting examples have distributed lag dynamics. We then expanded our knowledge of time series models by introducing models that allow for direct dynamic effects by using lagged dependent variables as regressors. Examples of such models are *autoregressive models of order p* (AR(p)) and autoregressive distributed lag (ADL) models.

We covered basic regression analysis with time series data. Under assumptions that parallel those for cross-sectional analysis, OLS is unbiased (under TS.1 through TS.3), OLS is BLUE (under TS.1 through TS.5), and the usual OLS standard errors, t statistics, and F statistics can be used for statistical inference (under TS.1 through TS.6).

We paid particular attention to the strict exogeneity assumption TS.3 as it is key for unbiasedness of the OLS estimators. Because of the temporal correlation in most time series data, this assumption is explicit about how the errors are related to the explanatory variables in all time periods. We discussed that this assumption is very strong as, for example, it does not allow for a feedback effect – that is, it does not allow policy variables in future time periods to respond to shocks in previous time periods. It is also pretty apparent that, in many applications, there are likely to be some explanatory variables that are not strictly exogenous. Thus, TS.3 is pretty restrictive for time series applications and cannot hold in models that have lagged dependent variables as regressors.

Overall, the classical linear model assumptions TS.1–TS.6 in time series models are quite restrictive but they are a natural starting point. Static and finite distributed lag models have a shot at satisfying these assumptions.

We also considered the large sample properties of OLS estimators with time series data. We discussed that OLS can be justified using asymptotic analysis, provided certain conditions are met. When the data are *weakly dependent* and *stationary* (with weak dependence being particularly important) and the explanatory variables are *contemporaneously exogenous*, OLS is consistent. Contemporaneous exogeneity (TS.3$'$) required for consistency is a much weaker and a more realistic restriction than the strict exogeneity condition (TS.3), which was required for the unbiasedness of the OLS.

The consistency result discussed in this topic has many applications, including the stationary AR(1) regression model. When we add the appropriate homoskedasticity and no serial correlation assumptions, the usual test statistics are asymptotically valid.

13.4.1 Key terms and concepts

- Stochastic process
- Autocorrelation/serial correlation

- Static model
- Finite Distributed Lag (FDL) model
- Impact propensity
- Long Run Propensity (LRP)
- Lag distribution
- Autoregressive process/autoregressive model
- Autoregressive Distributed Lag (ADL) model
- Strict exogeneity
- Stationary/strictly stationary stochastic process
- Covariance stationary stochastic process
- Weakly dependent stochastic process
- Contemporaneous exogeneity
- Autoregressive process of order one [AR(1)]
- Moving average process of order one [MA(1)]
- Moving average process of order two [MA(2)]

13.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand essential differences between time series data and cross-sectional data
- discuss limitations of static models in a time series setting
- describe a general form of finite distributed lag (FDL or just DL) models, interpret their coefficients and explain how to express immediate and cumulative effects on the response variable
- discuss the role of multicollinearity in FDL models and its consequences for the precision of estimates
- formulate and interpret models with lagged dependent variables as regressors
- explain which conditions guarantee the unbiasedness of OLS in regressions with time series data, explain the strict exogeneity assumption, and describe what economic situations it eliminates
- explain which conditions guarantee that the usual OLS variance formulas are valid in models with regressors correlated across time

- describe conditions that guarantee that the OLS estimator is BLUE
- estimate and interpret static and FDL in practice
- understand notions of stationarity, covariance stationarity and weak dependence
- analyse covariance stationarity and weak dependence properties of an AR(1) process
- analyse covariance stationarity and weak dependence properties of an MA(1) process
- explain which conditions guarantee the consistency of OLS in regressions with time series data, formulate the contemporaneous exogeneity assumption, and describe how it is different from the strict exogeneity assumption
- explain which conditions guarantee that the OLS estimator is approximately normally distributed and the usual t tests, F tests and confidence intervals are all asymptotically valid.

13.5 Test your knowledge and understanding

13.5.1 Additional examination based exercises

13.1E Ever since the oil price increases of the 1970s, macroeconomists have been interested in estimating the effect of an increase in the international price of crude oil on the U.S. rate of inflation. To be more specific, consider the model:

$$\mathit{Infl}_t = \alpha_0 + \delta_0 X_t + \delta_1 X_{t-1} + u_t$$

where X_t denotes the change in the crude oil price in period t. Imagine you attend a conference where a presenter talks about this model and says the following: ‘Because oil prices are set in world markets in large part by foreign oil-producing countries, I think that strict exogeneity (and, hence, contemporaneous exogeneity) in this model is satisfied.’ Would you agree with the presenter’s statement? Discuss briefly.

13.2E Increases in oil prices have been blamed for several recessions in developed countries. Let GDP_t denote the value of quarterly gross domestic product in the US and let $Y_t = 100\log(GDP_t/GDP_{t-1})$ be the quarterly percentage change in GDP.

Arguably, oil prices adversely affect the economy only when they jump above their values in the recent past. Let O_t equal the maximum of the ‘percentage point difference between oil prices at date t and their maximum value during the past year’ and ‘zero’. A distributed lag regression relating Y_t and O_t, estimated over the period 1955:Q1–2000:Q4 is:

$$\begin{aligned}\widehat{Y}_t = \underset{(0.1)}{1.0} \underset{(0.054)}{-0.055}\, O_t \underset{(0.057)}{-0.026}\ O_{t-1} \underset{(0.048)}{-0.031}\ O_{t-2} \underset{(0.042)}{-0.109}\ O_{t-3} \underset{(0.053)}{-0.128}\ O_{t-4}\\ \underset{(0.025)}{+0.038}\ O_{t-5} \underset{(0.048)}{+0.025}\ O_{t-6} \underset{(0.039)}{-0.019}\ O_{t-7} \underset{(0.042)}{+0.067}\, O_{t-8}.\end{aligned}$$

The standard errors are reported in parentheses.

(a) What is the effective sample size used in the estimation of the above model?

(b) Suppose that oil prices jump 25% above their previous peak value and stay at this new higher level (so that $O_t = 25$ and $O_{t+1} = O_{t+2} = \cdots = O_{t+8} = 0$). What is the predicted effect on output growth for each quarter over the next 2 years?

(c) Construct 95% confidence intervals for your answers to (b).

(d) What is the predicted cumulative change in GDP growth over eight quarters?

(e) The HAC F statistic testing whether the coefficients on O_t and its lags are zero is 3.49. Are these coefficients significantly different from zero? Explain your answer. Briefly indicate why a robust F statistic was used instead of the usual F statistic.

(f) Are any of the coefficients on O_{t-j}, for $j = 0, \ldots, 8$, individually statistically significant? What do you think drives the imprecision of the OLS estimators in this model?

13.3E Consider the stationary AR(1) model:

$$y_t = \beta_1 y_{t-1} + \varepsilon_t \quad \text{with} \quad |\beta_1| < 1, \quad t = 1, \ldots, T$$

where ε_t is i.i.d. $(0, \sigma^2)$, and ε_t is independent of $y_{t-1}, y_{t-2}, \ldots$. Let $y_0 = 0$.

(a) Derive the OLS estimator of β_1. Denote it as $\widehat{\beta}_1$.

(b) Is $\widehat{\beta}_1$ an unbiased estimator of β_1? Is it BLUE?

(c) Is $\widehat{\beta}_1$ a consistent estimator of β_1? Clearly explain your answer.

13.5.2 Solutions to additional examination based exercises

13.1E Because oil prices are set in world markets in large part by foreign oil-producing countries, initially one might indeed think that oil prices are exogenous. But oil prices are not like the weather: Members of OPEC set oil production levels strategically, taking many factors, including the state of the world economy, into account. To the extent that oil prices (or quantities) are set based on an assessment of current and future world economic conditions, including inflation in the United States, oil prices are not strictly exogenous or even contemporaneously exogenous.

13.2E (a) Overall we have 184 quarterly observations. But because we have eight lags of variable O_t on the right-hand side, in the regression equation we use 176 observations only (starting with period 9 or, equivalently, Q1 of 1957).

(b) The predicted effect on output growth is $\widehat{\delta}_i \times 25$.

Period ahead: 0 $\widehat{\delta}_1 \times 25 = -0.055 \times 25 = -1.375$.

Period ahead: 1 $\widehat{\delta}_2 \times 25 = -0.026 \times 25 = -0.65$.

Period ahead: 2 $\widehat{\delta}_3 \times 25 = -0.031 \times 25 = -0.775$.

Period ahead: 3 $\widehat{\delta}_4 \times 25 = -0.109 \times 25 = -2.725$.

Period ahead: 4 $\widehat{\delta}_5 \times 25 = -0.128 \times 25 = -3.2$.

Period ahead: 5 $\widehat{\delta}_6 \times 25 = 0.038 \times 25 = 0.95$.

Period ahead: 6 $\widehat{\delta}_7 \times 25 = 0.025 \times 25 = 0.625$.

Period ahead: 7 $\widehat{\delta}_8 \times 25 = -0.019 \times 25 = -0.475$.

Period ahead: 8 $\widehat{\delta}_9 \times 25 = 0.067 \times 25 = 1.675$.

(c) Observe that $se(\widehat{\delta}_j \times 25) = 25 \times se(\widehat{\delta}_j)$. The 95% confidence interval is obtained as:

$$\left[\widehat{\delta}_j \times 25 - 1.96 \times se(\widehat{\delta}_j \times 25), \widehat{\delta}_j \times 25 + 1.96 \times se(\widehat{\delta}_j \times 25)\right]$$

$$= 25 \times \left[\widehat{\delta}_j - 1.96 \times se(\widehat{\delta}_j), \widehat{\delta}_j + 1.96 \cdot \text{se}(\widehat{\delta}_j)\right].$$

So:

Period ahead: 0
$[-1.375 - 1.96 \times 25 \times .054, -1.375 + 1.96 \times 25 \times .054] = [-4.021, 1, 271]$.

Period ahead: 1
$-0.65 - 1.96 \times 25 \times .057, -0.65 + 1.96 \times 25 \times .057] = [-3.443, 2.143]$.

Period ahead: 2
$[-0.775 - 1.96 \times 25 \times .048, -0.775 + 1.96 \times 25 \times .048] = [-3.127, 1.577]$.

Period ahead: 3
$[-2.725 - 1.96 \times 25 \times .042, -2.725 + 1.96 \times 25 \times .042] = [-4.783, -0.667]$.

Period ahead: 4
$[-3.2 - 1.96 \times 25 \times .053, -3.2 + 1.96 \times 25 \times .053] = [-5.797, -0.603]$.

Period ahead: 5
$[0.95 - 1.96 \times 25 \times .025, 0.95 + 1.96 \times 25 \times .025] = [-0.275, 2.175]$.

Period ahead: 6
$[0.625 - 1.96 \times 25 \times .048, 0.625 + 1.96 \times 25 \times .048 = [-1.727, 2, 977]$.

Period ahead: 7
$[-0.475 - 1.96 \times 25 \times .039, -0.475 + 1.96 \times 25 \times .039] = [-2.386, 1, 436]$.

Period ahead: 8
$[1.675 - 1.96 \times 25 \times .042, 1.675 + 1.96 \times 25 \times .042] = [-0.015, 0.149]$.

Observation: Most of these intervals contain the value zero, a clear sign of the imprecision with which these effects are obtained due to the presence of near multicollinearity.

(d) The predicted cumulative change is given by:

$$\begin{aligned}(\widehat{\delta}_1 + \cdots + \widehat{\delta}_9) \times 25 &= (-0.055 - 0.026 - 0.031 - 0.109 - 0.128 + 0.038 \\ &\quad + 0.025 - 0.019 + 0.067) \times 25 \\ &= -5.95\%.\end{aligned}$$

That is, the predicted cumulative change in GDP growth is a decline of nearly 6%.

Observation: Even though the predicted effect on output growth for each quarter is estimated very imprecisely, we do expect that we can estimate this cumulative change precisely.

(e) The HAC F statistic tests the hypothesis:

$$H_0 : \delta_1 = \cdots = \delta_9 = 0$$

against the alternative:

$$H_1 : \text{ at least one of } \delta_j, \text{ for } j = 1, \ldots, 9, \text{ is not zero.}$$

The fact that we use the HAC F test indicates that we use a robust estimate of the variances and covariances needed for its computation (robust to the presence of heteroskedasticity and serial correlation). We use the F distribution with $(9, T - 10)$ degrees of freedom (9 restrictions and $T - 10$ degrees of freedom of the unrestricted model). Here $T = 46 \times 4 - 8 = 176$. At the 5% level of significance our critical value is approximately equal to 1.88. Since the statistic exceeds this critical value, we reject the null, indicating that the coefficients are jointly significantly different from zero, that is, we expect increases in oil prices to influence the output growth for at least 2 years.

(f) As is clear from confidence intervals constructed in (c), none of the coefficients on O_{t-j}, for $j = 0, \ldots, 8$, is individually statistically significant at the 5% significance level. Such imprecision of individual estimators is caused by the presence of near multicollinearity.

13.3E (a) To get the OLS estimator of β_1, we minimise:

$$\min_{b_1} \sum_{t=1}^{n} (y_t - b_1 y_{t-1})^2.$$

The FOC is given by:

$$-2 \sum_{t=1}^{n} (y_t - \widehat{\beta}_1 y_{t-1}) y_{t-1} = 0.$$

Solving for $\widehat{\beta}_1$ yields:

$$\widehat{\beta}_1 = \frac{\sum_{t=1}^{n} y_{t-1} y_t}{\sum_{t=1}^{n} y_{t-1}^2}.$$

We need to assume that $\sum_t y_{t-1}^2 > 0$ (so that we have sample variation in y_t).

(b) $\widehat{\beta}_1$ cannot be BLUE, because it is not unbiased.

After plugging in the model, we can rewrite:

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_t y_{t-1} \varepsilon_t}{\sum_t y_{t-1}^2}.$$

Taking its expectation gives:

$$E(\widehat{\beta}_1) = \beta_1 + E\left(\frac{\sum_t y_{t-1} \varepsilon_t}{\sum_t y_{t-1}^2} \right).$$

$\widehat{\beta}_1$ is not likely to be unbiased, i.e. $E(\widehat{\beta}_1) \neq \beta_1$ as $E\left(\frac{\sum_t y_{t-1}\varepsilon_t}{\sum_t y_{t-1}^2}\right) \neq 0$. Why is that? If one wanted to try to consider conditioning arguments, that will not help because, even though ε_t is uncorrelated with the past values of y_t, the current and future values of y_t will be related to ε_t. This is a setting where it is clearly unreasonable to assume strict exogeneity.

Observe: If you want to evaluate $E\left(\frac{y_{t-1}\varepsilon_t}{\sum_t y_{t-1}^2}\right)$ it would be convenient to be able to take the ys out of the expectation operator. To do this we need to condition on $y_1, \ldots, y_{n-1}$. Then:

$$E\left(\frac{y_{t-1}\varepsilon_t}{\sum_t y_{t-1}^2}\,\middle|\, y_1, \ldots, y_{n-1}\right) = \frac{y_{t-1}E(\varepsilon_t \,|\, y_1, \ldots, y_{n-1})}{\sum_t y_{t-1}^2}.$$

The fact that ε_t will be correlated with $y_t, y_{t+1}, \ldots$ means that $E(\varepsilon_t \,|\, y_1, \ldots, y_{n-1}) \neq 0$. Then we would also expect $E\left(\frac{y_{t-1}\varepsilon_t}{\sum_t y_{t-1}^2}\right) \neq 0$.

Very important: Note that:

$$E\left(\frac{\sum_t y_{t-1}\varepsilon_t}{\sum_t y_{t-1}^2}\right) \neq \frac{E(\sum_t y_{t-1}\varepsilon_t)}{E\left(\sum_t y_{t-1}^2\right)}.$$

Had that been the case one might have continued to argue that:

$$E\left(\sum_t y_{t-1}\varepsilon_t\right) = \sum_t E(y_{t-1}\varepsilon_t) = 0$$

as the independence of ε_t and y_{t-1} gives:

$$E(y_{t-1}\varepsilon_t) = E(y_{t-1})\,E(\varepsilon_t) = 0.$$

But it is not this simple.

(c) $\widehat{\beta}_1$ is consistent.

After plugging in the model, we can rewrite:

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_t y_{t-1}\varepsilon_t}{\sum_t y_{t-1}^2}$$

and further rewrite this in terms of averages to yield:

$$\widehat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum_t y_{t-1}\varepsilon_t}{\frac{1}{n}\sum_t y_{t-1}^2}.$$

By properties of the plim operator:

$$\text{plim } \widehat{\beta}_1 = \beta_1 + \frac{\text{plim } \frac{1}{n}\Sigma_t y_{t-1}\varepsilon_t}{\text{plim } \frac{1}{n}\Sigma_t y_{t-1}^2}.$$

Given we have weak dependence, $|\beta_1| < 1$, we can then use the LLN to get:

$$\text{plim } \widehat{\beta}_1 = \beta_1 + \frac{E(y_{t-1}\varepsilon_t)}{E(y_{t-1}^2)}.$$

As $E(y_{t-1}\varepsilon_t) = 0$ we have established that plim $\widehat{\beta}_1 = \beta_1$, provided $E(y_{t-1}^2) \neq 0$ that is we have proven consistency.

Chapter 14
Autocorrelation in time series regression

14.1 Introduction

14.1.1 Aims of the chapter

In this chapter we discuss the consequences of using the OLS estimator when the regression error exhibits serial correlation (that is, where the errors show dependence over time). As long as the regressors are exogenous (strict or contemporaneous) the OLS estimator remains consistent. For valid statistical inference, we will need to use robust standard errors and we will discuss the use of HAC standard errors to account for serial correlation (and heteroskedasticity) in the errors. When our model includes lagged dependent variables as regressors, the OLS estimator will typically no longer be consistent when the regression error exhibits serial correlation. Hence, we should not use OLS in this case!

As there are clear implications to the presence of autocorrelation in errors in both settings, you will also learn how to test for the presence of autocorrelation.

14.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- describe the consequences of serial correlation in the regression error for the OLS estimation in models with strictly exogenous or contemporaneously exogenous regressors
- give the intuition for heteroskedasticity and serial correlation adjustment of standard errors
- describe the consequences of serial correlation in the regression error for the OLS estimation in models with lags of the dependent variable as regressors
- outline an approach to testing for serial correlation in regression errors in models under strict exogeneity
- outline an approach to testing for serial correlation in regression errors in models under contemporaneous exogeneity

- use statistical software to estimate time series regressions and conduct a test for serial correlation in regression errors.

.

14.1.3 Essential reading

Readings: Wooldridge, Sections 12.1–12.3 (you may skip Section 12.3b).

14.1.4 Synopsis of this chapter

In this chapter we consider the setting where the regression error exhibits serial correlation. The implications of this for the properties of the OLS estimator are analysed.

In Sections 12.1a–12.1b in Wooldridge the focus is on situations where the OLS estimator remains consistent. As autocorrelation in the error poses a violation of one of the Gauss–Markov assumptions, the estimator is no longer efficient (BLUE). Moreover, the standard errors need to be adjusted for heteroskedasticity and/or serial correlation (so-called heteroskedasticity and autocorrelation-consistent (HAC) standard errors). Section 12.2 in Wooldridge discusses serial correlation-robust inference after OLS in these settings. The use of the goodness of fit measure in these settings is briefly discussed in Section 12.1c.

In Section 12.1d in Wooldridge, we learn that the consequences for the OLS estimator of serial correlation in the regression error of models that have lagged dependent variables as regressors is more severe. Here the OLS estimator is inconsistent (should not use OLS!). Various approaches to estimating such models are discussed.

Finally, Section 12.3 in Wooldridge discusses how to test for the presence of serial correlation. The discussion centres around a test for the null hypothesis of the absence of serial correlation against the alternative hypothesis which specifies that the serial correlation is given by an autoregressive process. We will focus in this course on the use of the (asymptotic) t or F test, where the F test is required when higher-order autoregressive processes are considered.

14.2 Content of chapter

14.2.1 Properties of OLS with serially correlated errors; Serial correlation-robust inference after OLS

Read: Wooldridge, Sections 12.1–12.2.

When discussing the properties of the OLS estimator in the presence of serially correlated errors it is important to establish whether this serial correlation in the errors gives rise to a correlation between the error and regressor (that is causes a violation of TS.3 or TS.3$'$).

- **Case 1**: If serial correlation does *not* cause a correlation between the error and regressors, we can still use OLS but for inference we will need to correct the standard errors. As serial correlation is a violation of the Gauss–Markov assumptions (TS.5) the OLS estimator will not be efficient (BLUE).
- **Case 2**: If serial correlation does cause a correlation between the error and regressor(s), we should no longer apply OLS. The OLS estimator will not be unbiased nor consistent even when stationarity and weak dependence are assumed (TS.1′). A classic example of this is when we have lagged dependent variables as regressors in our model and the errors are serially correlated.

Further discussion Case 1

For concreteness, consider the regression model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

where u_t exhibits serial correlation.

Serial correlation means that the errors, $\{u_t : t = 1, 2, \ldots\}$ are correlated. This, of course, immediately implies that Assumption TS.5 ($E(u_t u_s \,|\, \mathbf{x}_1, \ldots, \mathbf{x}_n) = 0$, for $t \neq s$) is violated and Assumption TS.5′ ($E(u_t u_s \,|\, \mathbf{x}_t, \mathbf{x}_s) = 0$, for $t \neq s$) is violated as well.

- For this model, as long as TS.3 is satisfied, the OLS estimator will be unbiased. If TS.3 is violated, but TS.3′ is satisfied the OLS estimator will be consistent (assuming stationarity and weak dependence are satisfied).
- As serial correlation is a violation of the Gauss–Markov assumptions, the estimator will not be BLUE (efficient).
- The presence of serial correlation will invalidate the usual OLS statistical inference (t and F tests, confidence intervals) even in large samples. Intuitively, the issue of serial correlation is similar to heteroskedasticity in the case of cross-sections and so our approach will be to use the OLS estimator but to modify the formula for the OLS variance and use that modified formula for the computation of standard errors, confidence intervals, t and F test statistics.

The technical details provided in Section 12.2 are not examinable. Let us provide here the main ideas and intuition around serial correlation-robust inference. Statistical packages will easily allow us to obtain the **heteroskedasticity and autocorrelation robust (HAC) standard errors**. For simplicity, consider the bivariate regression:

$$y_t = \beta_0 + \beta_1 x_t + u_t, \quad t = 1, \ldots, T$$

and suppose Assumptions TS.1–TS.3 hold. (OLS unbiased and consistent!)

- To get $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ with $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, you want to recall (see the topic on bivariate regressions) that the OLS estimator $\widehat{\beta}_1$ can be decomposed as:

  $$\widehat{\beta}_1 = \beta_1 + \sum_{t=1}^{n} w_t u_t \quad \text{with} \quad w_t = \frac{x_t - \bar{x}}{SST_x}$$

  i.e. truth + noise decomposition.

- Because β_1 is constant:

$$Var(\widehat{\beta}_1 \,|\, \mathbf{X}) = Var\left(\sum_{t=1}^{n} w_t u_t \,|\, \mathbf{X}\right).$$

- Let us recall the following key fact (property of the variance):

$$Var(A+B) = Var(A) + Var(B) + Cov(A,B) + Cov(B,A).$$

  If A and B are independent, then $Cov(A,B) = Cov(B,A)$ is zero.

  Applying this to $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ under serial correlation of $\{u_t\}$, we obtain:

$$\begin{aligned} Var\left(\sum_{t=1}^{n} w_t u_t \,|\, \mathbf{X}\right) &= \sum_{t=1}^{n} Var(w_t u_t \,|\, \mathbf{X}) + \sum_{t=1}^{n}\sum_{s\neq t}^{n} Cov(w_t u_t, w_s u_s \,|\, \mathbf{X}) \\ &= \sum_{t=1}^{n} w_t^2\, Var(u_t \,|\, \mathbf{X}) + \underbrace{\sum_{t=1}^{n}\sum_{s\neq t}^{n} w_t w_s Cov(u_t, u_s \,|\, \mathbf{X})}_{not\, zero} \end{aligned}$$

  (note, once you condition on $\mathbf{X}$, w_t are fixed constants). Due to the autocorrelation we can no longer argue that the second term vanishes (equals zero) as we did in the cross-sectional setting.

- Because we assume weak dependence, we will be able to argue that as $s-t$ increases, the covariances will become zero. Let us assume below, that we can ignore $Cov(u_t, u_s), s \neq t$ when $|t-s| \geq q$ (truncation lag).

- Intuitively, a good estimator of $Var(\widehat{\beta}_1 \,|\, \mathbf{X})$ then is given by:

$$\widehat{Var}\left(\sum_{t=1}^{n} w_t u_t \,|\, \mathbf{X}\right) = \underbrace{\sum_{t=1}^{n} w_t^2 \widehat{u}_t^2}_{\text{Heterosk Robust}} + \underbrace{\sum_{t=1}^{n}\sum_{\substack{t\neq s \\ |t-s|<q}}^{n} w_t w_s \widehat{u}_t \widehat{u}_s}_{\text{Serial Correl Robust}}.$$

The standard errors computed based on this estimator (or another good estimator of the above variance) are sometimes called **HAC (heteroskedasticity and autocorrelation consistent) standard errors**. This allows us to compute valid confidence intervals and test statistics that are robust to general forms of serial correlation.

It is important to realise that we are still estimating the parameters by OLS. We are only changing how we estimate their precision and perform inference. Just as with the heteroskedasticity-robust inference, we can apply the HAC inference whether or not we have evidence of serial correlation. Large differences in the HAC standard errors and the usual standard errors suggest the presence of serial correlation (autocorrelation) and/or heteroskedasticity.

Further discussion Case 2

Let us consider a simple linear regression model:

$$\begin{aligned} &y_t = \beta_0 + \beta_1 y_{t-1} + u_t, \quad \text{with } |\beta_1| < 1 \\ &u_t = \rho u_{t-1} + e_t, \quad e_t \text{ is i.i.d. } (0, \sigma_e^2) \\ &|\rho| < 1 \text{ (stationary AR(1))} \end{aligned}$$

where e_t is independent of the past and, therefore, independent of y_{t-1}, y_{t-2}, ... (This independence implies that $E(e_t \,|\, y_{t-1}, y_{t-2}, \ldots) = 0$.)

In this model the OLS estimators of β_0 and β_1 are not only inefficient and biased, but **inconsistent** as well. The main reason for this inconsistency is that in the equation in period t regressor y_{t-1} is correlated with the regression error u_t:

$$Cov(y_{t-1}, u_t) \neq 0$$

so the assumption TS.3$'$ cannot be satisfied.

The reasons why $Cov(y_{t-1}, u_t) \neq 0$ are the following: (i) y_{t-1} depends on u_{t-1} (this can be directly seen from the regression model in period $t-1$), and (ii) u_t depends on u_{t-1} if $\rho \neq 0$ (this can be directly seen from the AR(1) definition of u_t).

The dependence of *both* y_{t-1} and u_t on u_{t-1} results in y_{t-1} and u_t being correlated.

In the above model there is actually a simple way to remove the problem altogether by adding a further lag, y_{t-2}, to the model. To show this, it is useful to recognise that e_t is a 'nice' error which does not exhibit serial correlation, and by rewriting our AR(1) definition of u_t:

$$e_t = u_t - \rho u_{t-1}.$$

The fact that when we subtract ρu_{t-1} from u_t we obtain e_t, suggests that we should subtract ρy_{t-1} from y_t as well:

$$\begin{array}{rcl} y_t = & \beta_0 + \beta_1 y_{t-1} + u_t & \\ \rho y_{t-1} = & \rho\beta_0 + \rho\beta_1 y_{t-2} + \rho u_{t-1} & - \\ \hline y_t - \rho y_{t-1} = & (1-\rho)\beta_0 + \beta_1 y_{t-1} - \rho\beta_1 y_{t-2} + \underbrace{u_t - \rho u_{t-1}}_{e_t} & \end{array}$$

By moving ρy_{t-1} to the right-hand side we obtain:

$$\begin{aligned} y_t &= (1-\rho)\beta_0 + (\beta_1 + \rho) y_{t-1} - \rho\beta_1 y_{t-2} + e_t \\ &= \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + e_t \end{aligned}$$

where $\alpha_0 = (1-\rho)\beta_0$, $\alpha_1 = \beta_1 + \rho$, $\alpha_2 = -\rho\beta_1$.

This new equation has no serial correlation in errors. Since $E(e_t \,|\, y_{t-1}, y_{t-2}) = 0$ (explained earlier), we can estimate the parameters using OLS. They will be consistent even though they are still biased (recall that TS.3 is violated in models with lagged dependent variables as regressors). It should be noted that in this equation we, of course, are not estimating the original parameters $\beta_0, \beta_1, \beta_2$ but rather parameters α_0, α_1, α_2. However, for many practical purposes, including forecasting, it is this new equation that we would be most interested in. Alternatively, if one is interested in obtaining consistent estimators of $\beta_0, \beta_1, \beta_2$, we can consider the Instrumental Variable method to be applied in the original model.

Activity 14.1 Take the MCQs related to this section on the VLE to test your understanding.

Activity 14.2 An interesting economic model that leads to an econometric model with a lagged dependent variable relates y_t to the expected value of x_t, say, x_t^*, where the expectation is based on all observed information at time $t-1$:

$$y_t = \alpha_0 + \alpha_1 x_t^* + u_t. \tag{1}$$

A natural assumption on $\{u_t\}$ is that $E(u_t \,|\, I_{t-1}) = 0$, where I_{t-1} denotes all information on y and x observed at time $t-1$; this means that:

$$E(y_t \,|, I_{t-1}) = \alpha_0 + \alpha_1 x_t^*.$$

To complete this model, we need an assumption about how the expectation x_t^* is formed. We consider the following adaptive expectations scheme:

$$x_t^* - x_{t-1}^* = \lambda(x_{t-1} - x_{t-1}^*) \tag{2}$$

where $0 < \lambda < 1$. This equation implies that the change in expectations reacts to whether last period's realised value was above or below its expectation. The assumption $0 < \lambda < 1$ implies that the change in expectations is a fraction of last period's error.

(i) Show that the two equations imply that:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_{t-1} + v_t$$

where $\beta_0 = \lambda\alpha_0$, $\beta_1 = 1-\lambda$, $\beta_2 = \lambda\alpha_1$, and $v_t = u_t - (1-\lambda)u_{t-1}$. [*Hint*: Lag equation (1) one period, multiply it by $1-\lambda$, and subtract this from (1). Then, use (2).]

(ii) Under $E(u_t \,|\, I_{t-1}) = 0$, $\{u_t\}$ is serially uncorrelated. What does this imply about the new errors, $v_t = u_t - (1-\lambda)u_{t-1}$? What name do we give such a process?

(iii) Comment on the following statement: 'In order to deal with the autocorrelation in v_t, we need to use HAC standard errors when estimating the β parameters by OLS.'

(iv) Discuss how you can consistently estimate the βs.

(v) Given consistent estimators of the βs, how would you consistently estimate λ and α_1?

14.2.2 Testing for autocorrelation

Read: Wooldridge, Section 12.3 (you may skip 12.3b, Durbin–Watson test).

As we have discussed in the last section, the presence of autocorrelation in the errors has important consequences for estimating the multiple regression model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t$$

and conducting statistical inference. It is therefore important that we learn how to formally test for the presence of autocorrelation to see whether or not we should deal with these consequences or not. All tests for autocorrelation considered make use of the OLS residuals, $\{\widehat{u}_t\}$. Intuitively, if the true errors exhibit autocorrelation (show dependence over time) then our OLS residuals will display dependence over time as well. (Similar to the case of testing for heteroskedasticity, where we used $\widehat{u}_t^2$ in place of u_t^2.)

When specifying our test for autocorrelation, we need to clearly distinguish between the setting where we have strict exogeneity (TS.3 satisfied) and the setting where we have regressors that are not strict exogenous (TS.3$'$ satisfied). We will assume stationarity and weak dependence, which will permit us to rely on large sample inference.

When testing for autocorrelation we need to start by considering a particular autocorrelation model, say, the errors exhibit autoregressive dependence of order 1, AR(1):

$$u_t = \rho u_{t-1} + e_t \tag{*}$$

where $\{e_t : t = 1, 2, \ldots\}$ is an i.i.d. sequence with zero mean and variance σ^2. e_t is independent of $u_{t-1}, u_{t-2}, \ldots$. Here we should test:

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho \neq 0 \text{ (or } \rho > 0).$$

Under the null, therefore, there is no autocorrelation, whereas rejection of the null will give evidence for the presence of AR(1) autocorrelation in the errors.

As discussed in Wooldridge, if we could observe $\{u_t\}$, we would just estimate (*) by OLS and use a t test for $\rho = 0$. Instead, we base a test on the OLS residuals, $\widehat{u}_t$ and rely on an asymptotic t test. Specifically:

- when the regressors are strict exogeneous, to test whether $\rho = 0$ we should regress:

$$\widehat{u}_t = \alpha_0 + \rho \widehat{u}_{t-1} + \varepsilon_t$$

- when the regressors are not all strict exogenous, we should use:

$$\widehat{u}_t = \alpha_0 + \rho \widehat{u}_{t-1} + \alpha_1 x_{t1} + \cdots + \alpha_k x_{tk} + \varepsilon_t$$

- it is not necessary to estimate an intercept, but it is harmless to do so.

In Sections 12.3a (strict exogeneity) and 12.3c (no strict exogeneity) in Wooldridge step-by-step instructions are provided for these procedures. The main difference lies in the inclusion of all regressors $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})$ in our test equation (given above) when we do not have strict exogeneity. We use the asymptotic t test given by $\widehat{\rho}/se(\widehat{\rho})$. Under the null the test is asymptotic $Normal(0, 1)$. In Section 12.3d, the tests are extended to a higher order of serial correlation. Here we need to perform an asymptotic F test, which is needed to test the joint hypothesis:

$$H_0 : \rho_1 = 0, \rho_2 = 0, \ldots, \rho_q = 0$$

against the alternative that at least one ρ_j, for $j = 1, \ldots, q$ is non-zero.

Comment: You are only responsible for the tests of autocorrelation as discussed above. The Durbin–Watson test (discussed in Section 12.3) is a finite sample test that permits testing for no autocorrelation against AR(1) dependence. It requires strict exogeneity and the full set of CLM assumptions (including normality). As these assumptions are deemed to be quite restrictive in time series models, we will not discuss this test here. You may also ignore the discussion of the LM form of the test for autocorrelation (called the Breusch–Godfrey test).

Activity 14.3 Let us consider the following multiple linear regression model:

$$y_t = \beta_0 + \beta_1 z_t + \beta_2 z_{t-1} + u_t, \quad t = 2, 3, \ldots, T$$

where $E(u_t \,|\, z_1, z_2, \ldots, z_T) = 0$ and $Var(u_t \,|\, z_1, z_2, \ldots, z_T) = \sigma^2$.

(i) Discuss how you can test for autocorrelation in the error against the alternative that the error exhibits an AR(2) process (assumed to be stationary and weakly dependent).

(ii) Discuss how your result in (i) needs to be modified if all we can say about the errors is that they satisfy the conditions: $E(u_t \,|\, z_t, z_{t-1}) = 0$ and $Var(u_t \,|\, z_t, z_{t-1}) = \sigma^2$.

14.2.3 R application for stationary and weakly dependent time series models

Activity 14.4 Complete 'R Task – Activity 7' on VLE.

Here we show how to use time series data in R. It starts by explaining how to tell R that we are working with time series data and how to plot time series processes. We discuss how to run OLS using the command **dynlm**, how to test for autocorrelation, and how to obtain heteroskedasticity and autocorrelation robust (HAC) standard errors.

14.3 Answers to activities

Solution to Activity 14.2

(i) Lag equation (1) once:

$$y_{t-1} = \alpha_0 + \alpha_1 x_{t-1}^* + u_{t-1}$$

then multiply it by $1 - \lambda$:

$$(1-\lambda)y_{t-1} = (1-\lambda)\alpha_0 + (1-\lambda)\alpha_1 x_{t-1}^* + (1-\lambda)u_{t-1}$$

and subtract this from (1) to obtain:

$$y_t - (1-\lambda)y_{t-1} = \lambda\alpha_0 + \alpha_1(x_t^* - (1-\lambda)x_{t-1}^*) + u_t - (1-\lambda)u_{t-1}.$$

But we can rewrite (2) as:

$$x_t^* - (1-\lambda)x_{t-1}^* = \lambda x_{t-1}.$$

When we plug this into the first equation we obtain the desired result.

(ii) If $\{u_t\}$ is serially uncorrelated, then $\{v_t = u_t – (1–\lambda)u_{t-1}\}$ must be serially correlated. In fact, $\{v_t\}$ is an MA(1) process (recall the definition of an MA(1) process). We can find that:

$$Cov(v_t, v_{t-1}) = -(1-\lambda)\sigma_u^2$$

and the correlation between v_t and v_{t-h} is zero for $h > 1$.

(iii) The serial correlation in v_t will give rise to correlation between v_t and the lagged dependent variable y_{t-1} (indeed, v_t is correlated with v_{t-1}, and v_{t-1} affects y_{t-1} directly, which can be seen from the equation for period $t-1$). Therefore, the OLS estimators of the β_js will be inconsistent (and biased, of course) and we should not use OLS to estimate the βs. The use of HAC robust standard errors does not resolve the problem of the inconsistency of OLS.

(iv) To estimate the βs consistently, we can use an Instrumental Variables approach here. Indeed, we can use x_{t-2} as an IV for y_{t-1} because x_{t-2} is uncorrelated with v_t (because u_t and u_{t-1}, which compose v_t, are both uncorrelated with x_{t-2}) and x_{t-2} is partially correlated with y_{t-1}.

(v) Because $\beta_1 = 1-\lambda$, then λ is naturally estimated as $\widehat{\lambda} = 1 - \widehat{\beta}_1$. It is consistent using the properties of the plim operator. In particular:

$$\text{plim}(\widehat{\lambda}) = \text{plim}(1-\widehat{\beta}_1) = 1 - \text{plim}(\widehat{\beta}_1) = 1 - \beta_1 = \lambda.$$

Because $\beta_2 = \lambda\alpha_1$, we estimate $\widehat{\alpha}_1 = \frac{\widehat{\beta}_2}{\widehat{\lambda}}$. Consistency of $\widehat{\alpha}_1$ follows as:

$$\text{plim}(\widehat{\alpha}_1) = \text{plim}\left(\frac{\widehat{\beta}_2}{\widehat{\lambda}}\right) = \frac{\text{plim}(\widehat{\beta}_2)}{\text{plim}(\widehat{\lambda})} = \frac{\beta_2}{\lambda} = \alpha_1.$$

None of the estimators is unbiased. For one, the OLS estimators of the β_js are not unbiased given that there is a lagged dependent variable. Of course, $\widehat{\alpha}_1$ would not be unbiased even if $\widehat{\beta}_1$ and $\widehat{\beta}_2$ were as $E(\widehat{\alpha}_1) \neq \frac{E(\widehat{\beta}_2)}{1-E(\widehat{\beta}_1)}$.

Solution to Activity 12.3

(i) First we need to define the AR(2) process for u_t. It is given by:

$$u_t = \rho_1 u_{t-1} + \rho_2 u_{t-2} + e_t$$

where $\{e_t : t = 1, 2, \ldots\}$ is an i.i.d. sequence with zero mean and variance σ^2. e_t is independent of $u_{t-1}, u_{t-2}, \ldots$.

Then we need to clearly specify the null and the alternative hypotheses:

$$H_0 : \rho_1 = 0 \text{ and } \rho_2 = 0 \quad \text{against} \quad H_1 : \rho_1 \neq 0 \text{ and/or } \rho_2 \neq 0.$$

As our regressors are strictly exogenous (satisfy TS.3), we proceed as follows.

- Compute the OLS residuals, $\widehat{u}_t$ for all $t = 1, 2, \ldots, T$:

$$\widehat{u}_t = y_t - \widehat{\beta}_0 - \widehat{\beta}_1 z_t - \widehat{\beta}_2 z_{t-1}.$$

- Run the test regression:

$$\widehat{u}_t = \rho_0 + \rho_1 \widehat{u}_{t-1} + \rho_2 \widehat{u}_{t-2} + error_t \quad \text{for } t = 3, 4, \ldots, T.$$

- Use the (asymptotic) F test to test the joint significance of ρ_1 and ρ_2. We reject when the realisation of our test statistic exceeds the critical value for a given level of significance. A rejection of the null suggests evidence of autocorrelation. You should clearly indicate the fact that the test uses the strict exogeneity assumption.

(ii) As we now only assume contemporaneous exogeneity, we need to add z_t and z_{t-1} as additional regressors when estimating the test regression, that is, using the OLS residuals we need to estimate:

$$\widehat{u}_t = \rho_0 + \rho_1 \widehat{u}_{t-1} + \rho_2 \widehat{u}_{t-2} + \gamma_1 z_t + \gamma_2 z_{t-2} + v_t \quad \text{for } t = 3, 4, \ldots, T.$$

We reject when the p-value of the (asymptotic) F test for joint significance of ρ_1 and ρ_2 is lower than the significance level.

14.4 Overview of chapter

In this chapter we covered the important problem of serial correlation in the errors of multiple regression models. Positive correlation between adjacent errors is common, especially in static and finite distributed lag models. Serial correlation in the errors generally causes the usual OLS standard errors and statistics to be misleading (although the OLS estimators will still be consistent if TS.1′–TS.3′ hold and will be unbiased if TS.1–TS.3 hold). Typically, the OLS standard errors underestimate the true uncertainty in the parameter estimates. We discussed an idea behind computing HAC (heteroskedasticity and autocorrelation consistent) standard errors which are robust to general serial correlation and also robust to heteroskedasticity of unknown form. Using these robust standard errors, one could compute test statistics (t and F) that are also robust in this sense.

In addition, we considered the consequences of serial correlation in the regression error in models that include lagged dependent variables as regressors. In such models OLS estimators are often inconsistent and we discuss possible approaches to estimating such models in a reliable way.

Finally, we described an approach to testing for serial correlation. Our method tests the null hypothesis of the absence of serial correlation when the alternative hypothesis specifies that the serial correlation is given by an autoregressive process. We can conduct such testing either under strict exogeneity or contemporaneous exogeneity of explanatory variables.

14.4.1 Key terms and concepts

- Heteroskedasticity and Autocorrelation Consistent (HAC) standard errors
- Serial correlation-robust inference

14.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- describe the consequences of serial correlation in the regression error for the OLS estimation in models with strictly exogenous or contemporaneously exogenous regressors
- give the intuition for heteroskedasticity and serial correlation adjustment of standard errors
- describe the consequences of serial correlation in the regression error for the OLS estimation in models with lags of the dependent variable as regressors
- outline an approach to testing for serial correlation in regression errors in models under strict exogeneity
- outline an approach to testing for serial correlation in regression errors in models under contemporaneous exogeneity
- use statistical software to estimate time series regressions and conduct a test for serial correlation in regression errors.

14.5 Test your knowledge and understanding

14.5.1 Additional examination based exercises

14.1E For the US economy, let *gprice* denote the monthly growth in the overall price level and let *gwage* be the monthly growth in hourly wages. (These are both obtained by taking the difference of the logarithms: $gprice = \Delta \log(price)$). We have estimated the following distributed lag model (ADL$(0,12)$):

$$\begin{aligned}\widehat{gprice}_t = & \underset{(.00057)}{-.00093} + \underset{(.052)}{.119}\, gwage_t + \underset{(.039)}{.097}\, gwage_{t-1} + \underset{(.039)}{.040}\, gwage_{t-2} \\ & + \underset{(.039)}{.038}\, gwage_{t-3} + \underset{(.039)}{.081}\, gwage_{t-4} + \underset{(.039)}{.107}\, gwage_{t-5} + \underset{(.039)}{.095}\, gwage_{t-6} \\ & + \underset{(.039)}{.104}\, gwage_{t-7} + \underset{(.039)}{.103}\, gwage_{t-8} + \underset{(.039)}{.159}\, gwage_{t-9} + \underset{(.039)}{.110}\, gwage_{t-10} \\ & + \underset{(.039)}{.103}\, gwage_{t-11} + \underset{(.052)}{.016}\, gwage_{t-12}\end{aligned}$$

$$T = 273,\ R^2 = .317,\ \bar{R}^2 = .283.$$

The usual standard errors are in parentheses.

(a) We want to test whether the long-run propensity is significantly different from one. Clearly indicate the null and the alternative hypotheses, give the formula for the test statistic and the rejection rule. What is the unknown quantity in your test statistic? What regression would you run to obtain the standard error of the the long-run propensity directly?

(b) Discuss how you would conduct a formal test for serial correlation.

(c) Suppose a formal test in (b) suggests the presence of serial correlation in the regression errors. Discuss the consequence of this for the test you conducted in (a) and suggest two ways to deal with this problem.

14.2E In this part we consider the expectations augmented Phillips curve (see also Mankiw, 1994):

$$\mathit{infl}_t - \mathit{infl}_t^e = \beta_1(\mathit{unem}_t - \mu_0) + e_t$$

where μ_0 is the natural rate of unemployment (assumed to be constant over time) and infl_t^e is the expected rate of inflation formed in $t-1$. This model suggests that there is a trade-off between unanticipated inflation ($\mathit{infl}_t - \mathit{infl}_t^e$) and cyclical unemployment (the difference between actual unemployment and the natural rate of unemployment).

We assume that e_t (also called a supply shock) is an i.i.d. random variable with zero mean. You are told that expectations are formed using the following adaptive expectations model:

$$\mathit{infl}_t^e - \mathit{infl}_{t-1}^e = \lambda(\mathit{infl}_{t-1} - \mathit{infl}_{t-1}^e).$$

Given this information, it can be shown (you are not asked to do this) that the model can be written as:

$$\Delta \mathit{infl}_t = \gamma_0 + \gamma_1 \mathit{unem}_t + \gamma_2 \mathit{unemt}_{t-1} + v_t \qquad (*)$$

where $\Delta \mathit{infl}_t = \mathit{infl}_t - \mathit{infl}_{t-1}$, $\gamma_0 = -\beta_1\lambda\mu_0$, $\gamma_1 = \beta_1$, $\gamma_2 = -(1-\lambda)\beta_1$, and:

$$v_t = e_t - (1-\lambda)e_{t-1}. \qquad (**)$$

(a) What name do we give the process v_t in (**)? Is this process covariance stationary and/or weakly dependent? Discuss.

(b) Discuss what (minimal) assumptions you need to make about e_t (the supply shock) to guarantee the consistency of the OLS estimator of $(\gamma_0, \gamma_1, \gamma_2)$ in (*). (You are not expected to prove its consistency.) Indicate how you can use the consistency of the OLS estimator of $(\gamma_0, \gamma_1, \gamma_2)$ to obtain a consistent estimator of the long-run effect of unemployment on the change in inflation. Prove your claim.

(c) Let the assumptions for consistency of the OLS estimator of $(\gamma_0, \gamma_1, \gamma_2)$ in (*) be satisfied. How can you obtain valid standard errors for the long-run effect of unemployment on the change in inflation?

14.5.2 Solutions to additional examination based exercises

14.1E (a) The estimated long-run propensity is given by:

$$\widehat{\theta} = \widehat{\beta}_1 + \cdots + \widehat{\beta}_{13} = 1.172$$

where $\widehat{\beta}_i$ is the estimated coefficient on $gwage_{t-i+12}$.

We want to test $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$. The test statistic we should use is:

$$t = \frac{\widehat{\theta} - 1}{se(\widehat{\theta})} \sim t_{T-14} \quad \text{under } H_0 \text{ (GM + normality).}$$

Reject if $|t| > 1.96$ a 5% level of significance. (Important: $se(\widehat{\theta}) \neq se(\widehat{\beta}_1) + \cdots + se(\widehat{\beta}_{13})$). If we reject we find evidence that LRP is significantly smaller than one.

But $se(\widehat{\theta})$ is not available from the above output. Note that we can rewrite our model to allow us to obtain $se(\widehat{\theta})$ automatically from regression output. Specifically:

$$gprice_t = \beta_0 + \theta gwage_t + \beta_2(gwage_{t-1} - gwage_t) + \cdots + \beta_{13}(gwage_{t-12} - gwage_t) + u_t.$$

The model is exactly the same, observe $\theta - \beta_2 - \cdots - \beta_{13} = \beta_1$. Because both models are the same their R^2s are the same, it is only the interpretations of the parameters (here it is the parameter of $gwage_t$) that are different.

(b) We will describe how to conduct the test in the context of the AR(1) model:

$$u_t = \rho u_{t-1} + e_t$$

where $\{e_t\}$ is serially uncorrelated, has a zero mean, and a constant variance. Then we test:

$$H_0 : \rho = 0 \quad \text{against} \quad H_1 : \rho \neq 0 \text{ (or } \rho > 0).$$

We will describe how to test for serial correlation in errors under contemporaneous exogeneity.

- After estimating the model given in the question by OLS we save the residuals $\{\widehat{u}_t : t = 1, \ldots, T\}$.
- We then run the regression:

$$\widehat{u}_t \text{ on } \widehat{u}_{t-1},\ gwage_t, \ldots, gwage_{t-12}$$

  and compute the usual or heteroskedasticity-robust t statistic for $\widehat{\rho}$. We then carry out the test $H_0 : \rho = 0$ in the usual way (if the t statistic in the absolute value exceeds the critical value for a chosen significance level, we reject the null of no serial correlation).

(c) If there is autocorrelation we cannot trust the test performed in (a). We can use the robust (HAC) standard errors and recompute the test statistic (so we stick with the original model and OLS estimation). Another approach would be to try to remove the serial correlation by introducing more dynamics in the model, for example, by introducing lagged dependent variables.

14.2E (a) This is an MA(1) process. Textbook discussion of the process being covariance stationary and weakly dependent. There is no dependence when observations are more than 1 period apart, which is clearly evidence of weak dependence.

(b) Let x_t denote $unem_t$. Consistency requires the error term in the new model v_t and both regressors, x_t and x_{t-1} to be uncorrelated. Since e_t has a zero mean this means:

$$E[x_t(e_t - (1-\lambda)e_{t-1})] = 0$$
$$E[x_{t-1}(e_t - (1-\lambda)e_{t-1})] = 0.$$

A sufficient condition for this would be that $E[x_t e_s] = 0$ or uncorrelatedness between x_t and e_s $\forall\, s, t$. Obviously a more reasonable (minimal) assumption would only require $E[e_t x_{t+1}] = 0$ and $E[e_t x_t] = 0$.

The LR effect is $\gamma_1 + \gamma_2$. A consistent estimator of it is given by $\widehat{\gamma_{1,OLS}} + \widehat{\gamma_{2,OLS}}$ under these assumptions. We prove this by applying the plim rules.

(c) To get its standard error directly we could regress:

$$\Delta \mathit{infl}_t = \delta_0 + \delta_1 unem_t + \delta_2(unem_{t-1} - unem_t) + v_t$$

as δ_1 gives the long-run effect. Because of the serial correlation in the error, robust (HAC) standard errors should be used.

Chapter 15

Trends, seasonality and highly persistent time series in regression analysis

15.1 Introduction

15.1.1 Aims of the chapter

In this chapter we learn how to deal with the fact that many economic processes are nonstationary and may exhibit strong dependence (also called persistence). A key issue we need to avoid in these settings is the so-called spurious regression problem.

Violations of stationarity are commonplace as many time series processes exhibit seasonality and display trends. While the inclusion of seasonal dummies and deterministic trends may enable us to use standard statistical inference in these cases, the presence of strong dependence (persistence) poses a more serious issue for statistical analysis. We will discuss how to test for strong dependence (presence of unit roots) using non-standard (augmented) Dickey–Fuller tests. Two examples of highly persistent processes are a random walk and a random walk with drift (common in financial data), and we will show that by differencing such processes (instead of detrending and deseasonalising) we can remove the problem of nonstationarity inherent in these processes.

To avoid the spurious regression problem when running a regression involving highly persistent processes, we discuss a test of cointegration. If we find evidence of cointegration (that is reject the null of a spurious regression) we expect there to be a meaningful, long-run relationship between the variables whereby the variables co-move over time. We learn how a cointegrating relationship permits us to obtain an error correction model which describes the short-run dynamics in their relationship.

15.1.2 Learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the concept of trending time series and seasonality
- distinguish linear versus exponential trends
- explain the spurious regression problem

- understand the importance of including a time trend in regressions with deterministically trending variables
- understand what we mean by a detrended time series and a seasonally adjusted time series
- understand the concept of a highly persistent (strongly dependent) time series process
- explain the importance of the correlogram in identifying whether a process is highly persistent
- formulate a random walk and a random walk with drift
- explain the terminology of a unit root process, an integrated process and the order of integration
- discuss the Dickey–Fuller (DF) and augmented Dickey–Fuller (ADF) unit root tests and explain their diference
- discuss what we mean by a 'trend stationary' and a 'difference stationary' process, and discuss how we can use the unit root test to distinguish between them
- understand the importance of testing for cointegration when using highly persistent variables in regression
- explain the concept of cointegration in regression models
- conduct a unit root test on the residuals to test for cointegration (Engle–Granger test)
- construct an error correction model (ECM) and describe the advantage over differencing
- explain the Engle–Granger two-step procedure for estimating the ECM
- use statistical software to estimate time series regressions in the presence of nonstationarity and implement unit root test on real data.

15.1.3 Essential reading

Readings: Wooldridge, Sections 10.5 (skip 10.5d), 11.3, and 18.2–18.4.

15.1.4 Synopsis of this chapter

Many economic time series exhibit characteristics that require special attention when applying regression analysis. While assumptions of covariance stationarity and weak dependence are convenient assumptions we may want to impose in order to guarantee consistency of OLS and permit us to apply the usual inference procedures this may not always be satisfied by economic time series processes.

In Section 10.5 in Wooldridge, we learn that violations of stationarity are commonplace. Indeed many time series processes exhibit seasonality and show clear upward (or downward) trends. By including seasonal dummies and deterministic trends (whichever ones are appropriate), usual statistical inference will hold. The inclusion of deterministic trends (seasonal dummies) in regression models is equivalent to using detrended (seasonally adjusted) stochastic processes. These transformations (detrending/seasonal adjustments) effectively allow us to remove the nonstationarity in these processes. We introduce the concept of trend-stationary processes here.

In Section 11.3 in Wooldridge, we learn that the presence of strong dependence (persistence) poses a more serious issue for statistical analysis. Two examples of highly persistent processes are a random walk and a random walk with drift (common in finance). We will discuss these processes and show that by differencing such processes (instead of detrending) we can remove the problem of nonstationarity inherent in these processes. When we can make a process stationary by differencing a process once, we say that we have a process that is integrated of order 1, denoted $I(1)$. We introduce the concept of difference-stationary here.

Section 18.2 in Wooldridge discusses how to test for persistence (presence of unit roots) with a non-standard (augmented) Dickey–Fuller test. When we run regressions involving highly persistent processes, we need to be careful to avoid spurious (meaningless) regressions, whereby the only reason we appear to have a relationship is due to their persistence. This problem is discussed in Section 18.3. Regressions involving highly persistent time series that describe a meaningful (as opposed to spurious) relationship describe the long-run, cointegrating relationship between the variables as discussed in Section 18.4 in Wooldridge. Over time these variables co-move in a way that their difference has a zero mean and their differences return to zero with some regularity. In the presence of a long-run relationship between economic variables, we can describe an error correction model that enables us to study the short-run dynamics in their relationship.

15.2 Content of chapter

We have argued that the classical linear regression model assumptions for time series analysis, TS.1 to TS.6, are rather restrictive. To establish the consistency of OLS and justify the use of standard (robust) inference procedures under weaker assumptions (TS.1$'$ to TS.5$'$), we introduced two convenient assumptions: (covariance) stationarity and weak dependence. **In this chapter we will discuss violations from (covariance) stationary and weak dependence that are common in economic time series.**

Nonstationary but weakly dependent

We start with the simplest setting, where the only problem associated with the requirements of stationarity and weak dependence of time series is the **presence of deterministic trends and seasonality**. As the mean of such weakly dependent processes will be changing over time, they will be nonstationary. Nevertheless, such nonstationary processes can easily be made stationary by detrending (deseasonalising).

We introduce the concept **trend-stationarity** to represent this setting (which encompasses weak dependence). By including an appropriate trend and seasonal dummies in our regression analysis we can use standard (robust) inference procedures as trend stationary processes also satisfy TS.1′.

Nonstationary and persistent (strong dependent)

Thereafter, we turn to the more complex problem of **persistence**. Highly persistent time series are processes that display strong dependence, a common problem with many economic time series. Statistical analysis in the presence of strong dependence unfortunately becomes non-standard. We will show how we can test for persistence by considering **'unit root' processes**. Instead of detrending our process, we can remove the problem of nonstationarity and persistence inherent in these processes by differencing. We introduce the concept of **difference-stationarity** to represent this setting.

15.2.1 Trends and seasonality

Read: Wooldridge, Section 10.5.

Trending variables

Many economic time series have the tendency to grow over time as discussed in Section 10.5a, and a simple way to model such trending behaviour is to assume either:

- **linear time trend**:

  $$y_t = \alpha_0 + \alpha_1 t + e_t, \quad E(e_t) = 0.$$

  (Abstracting from random deviations, the time series increases linearly with time.)

- **exponential time trend**:

  $$\log(y_t) = \beta_0 + \beta_1 t + e_t, \quad E(e_t) = 0.$$

  (Abstracting from random deviations, the time series has a constant growth rate.)

  Comment: In the simplest definition of these models, we may assume that the error term is an i.i.d. sequence with $E(e_t) = 0$ and $Var(e_t) = \sigma^2$, but without loss of generality we may allow $\{e_t\}$ to be stationary and weakly dependent.

You need to be able to interpret α_1 and β_1 clearly, and explain why these processes violate the covariance stationarity assumption. It should be clear that the mean of a process $\{y_t\}$ that exhibits a linear or exponential time trend is changing over time (using a graphical discussion is fine).

We need to be careful when using trending variables in regression analysis to avoid the so-called **spurious regression problem** as discussed in Section 10.5b in Wooldridge.

This is the phenomenon of finding a relationship between two (or more) independent variables simply because of the fact that they are trending.

A classic example of the spurious regression problem appeared in Yule (1926). 'Why do we sometimes get nonsense-correlations between time-series?' *Journal of the Royal Statistical Society*, 89, 1–63.

Yule observed a strong correlation between the proportion of UK marriages in church (*ChurchMarr*) and the UK mortality rate (*Mortality*) using data for the years 1866 to 1911. When running the following regression:

$$Mortality_t = \alpha_0 + \alpha_1 ChurchMarr_t + u_t$$

the estimate of α_1 was significantly different from zero. Obviously, as he argued, it is very hard to explain how the marriages in church can possibly affect the mortality ('nonsense correlations in time-series'). As their graph reveals, both variables show a clear downward trend in this period. The high correlation, and significance of α_1, is purely a result of the trending nature in both variables. In other words, it is a spurious regression.

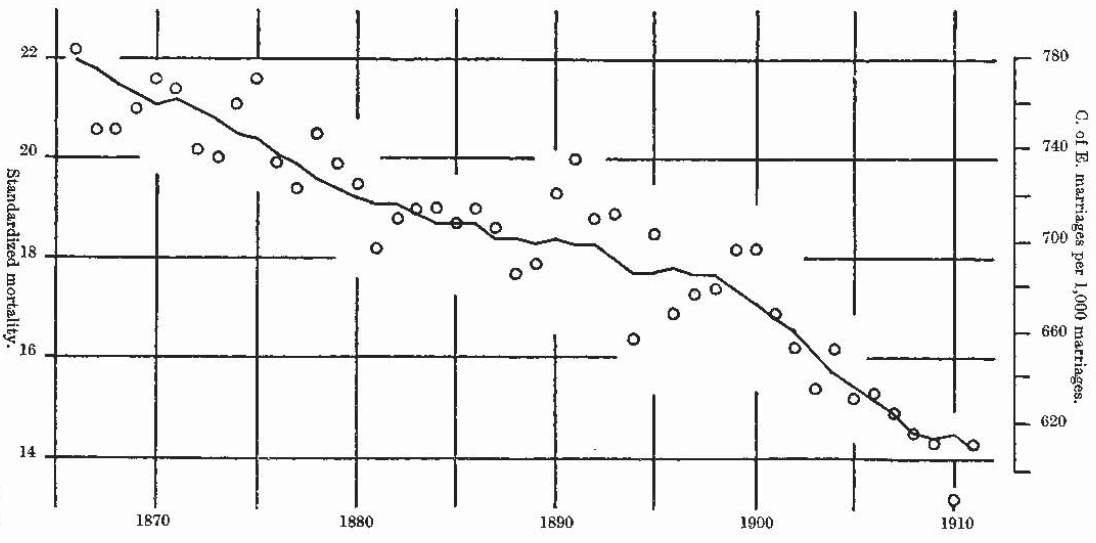


Fig. 1.—Correlation between standardized mortality per 1,000 persons in England and Wales (circles), and the proportion of Church of England marriages per 1,000 of all marriages (line), 1866–1911. $r = +0{\cdot}9512$.

When working with deterministically trending variables, we **avoid the spurious regression problem by including a time trend in our regression**.

In Section 10.5b, we learn that running a regression that includes a time trend yields parameter estimates that are identical to running a regression that uses so-called detrended variables. This important result relies on the *partialling-out interpretation of OLS*, which recognises that we can first control explicitly for the trending behaviour of our variables (a process called detrending) before estimating the parameters of interest (a result of the Frisch–Waugh–Lovell Theorem). In Wooldridge this procedure is described clearly when applied to the following model which includes a linear trend, t:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 t + u_t.$$

The detrending variables, such as $\ddot{x}_{t1}$, are obtained by running a regression of the variable on an intercept and the time trend and computing its residuals.

When conducting statistical inference using time series, we have advocated the use of assumptions TS.1′ to TS.5′ (asymptotic inference). Whereas deterministically trending variables do not satisfy the requirement of stationarity imposed by TS.1′, statistical inference will continue to hold **as long as we include the time trend in our regression**.

This result is due to the fact that the inclusion of the time trend ensures that we are effectively using ***detrended variables*** which **are stationary and weakly dependent (as required in TS.1′)**.

If we can model the trending behaviour using a deterministic trend, such as:

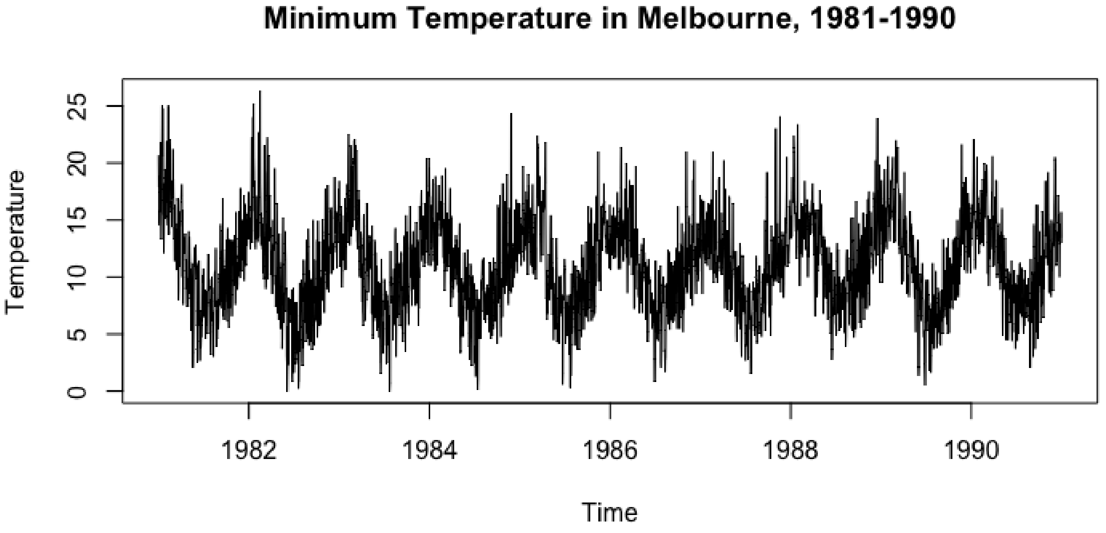
$$z_t = \alpha_0 + \alpha_1 t + e_t, \quad E(e_t) = 0$$

then the detrended variable (or its estimate) $z_t - \alpha_0 - \alpha_1 t$ will be stationary and weakly dependent as long as e_t is stationary and weakly dependent.

We say that the **deterministically trending variable z_t is trend-stationary**. It is a nonstationary variable that can be rendered stationary and weakly dependent by detrending.

Seasonality

If a time series is observed at monthly or quarterly intervals (or even weekly or daily), it may exhibit seasonality. An example of this is the 'Minimum Temperature in Melbourne, 1981 to 1990'. The units are in degrees Celsius and there are 3,650 observations (the source is the Australian Bureau of Meteorology). A seasonality component is quite obvious in these data.



A simple approach to deal with seasonality in regression is to include a set of **seasonal dummy variables** to account for seasonality in the dependent variable, the independent variables, or both.

If we have *quarterly data*, and we think the seasonal patterns within a year are roughly constant over time, we may consider the following regression model:

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 s_{t1} + \beta_4 s_{t2} + \beta_5 s_{t3} + u_t$$

where s_{1t}, s_{2t}, s_{3t} are seasonal dummies, such that:

$$s_{t1} = \begin{cases} 1 & \text{if } t \text{ falls in 1st quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$s_{t2} = \begin{cases} 1 & \text{if } t \text{ falls in 2nd quarter} \\ 0 & \text{otherwise} \end{cases}$$

$$s_{t3} = \begin{cases} 1 & \text{if } t \text{ falls in 3rd quarter} \\ 0 & \text{otherwise} \end{cases}$$

and x_{t1} and x_{t2} are explanatory variables of y_t. The reason for leaving out one of the seasonal dummies (base category) is to **avoid perfect collinearity**.

If we have *monthly data*, then we would include dummy variables for eleven of the twelve months, with the omitted category being the base category.

As discussed in Section 10.5e in Wooldridge, running a regression that includes seasonal dummies is identical to running a regression that uses deseasonalised variables. This important result uses the *partialling out interpretation of OLS* which recognises that we can first control explicitly for seasonality (a process called deseasonalisation) before estimating the parameters of interest.

> **Activity 15.1** Take the MCQs related to this section on the VLE to test your understanding.

> **Activity 15.2** (Based on Exercise 10.C6 from Wooldridge.)
>
> Using the data FERTIL3 (annual data, 1913 to 1984), we can obtain the following regression result to evaluate the effect of personal exemptions on the general fertility rate:
>
> $$\widehat{gfr} = \underset{(7.44)}{124.09} + \underset{(.078)}{.348}\, pe - \underset{(8.60)}{35.88}\, ww2 - \underset{(6.96)}{10.12}\, pill - \underset{(.74)}{2.53}t + \underset{(.0089)}{.0196}t^2$$
>
> $$T = 72,\ R^2 = .727.$$
>
> Robust (HAC) standard errors in parentheses.
>
> This regression allows for a *more general trending behaviour of our variables* (includes both a linear and quadratic time trend). Since a plot of *gfr* over time shows that the general fertility rate displayed both upward and downward trends during the period this is more appropriate than simply assuming a linear trend.
>
> Discuss how in this setting you can obtain the detrended variables ($\ddot{gfr}_t$, $\ddot{pe}_t$, $\ddot{ww2}_t$, $\ddot{pill}_t$) that would enable us to obtain the above result using detrended variables only:
>
> $$\ddot{gfr}_t = \beta_1 \ddot{pe}_t + \beta_2 \ddot{ww2}_t + \beta_3 \ddot{pill}_t + v_t.$$

Activity 15.3 (Based on Exercise 10.C6 from Wooldridge.)

Consider the regression result reported in Activity 15.2 for this question.

(i) Discuss the reason why robust (HAC) standard errors are reported.

(ii) You are interested in testing the hypothesis $H_0 : \beta_{pe} = 0$ against $H_1 : \beta_{pe} > 0$. Implement the test. Briefly discuss the assumptions that underlie your test.

Activity 15.4 (Based on Exercise 10.C11 from Wooldridge.)

Let us use the data TRAFFIC2, which contains 108 monthly observations on automobile accidents, traffic laws, and other variables from California (January 1981 through December 1989). Figures reported in parentheses are robust (HAC) standard errors.

(i) A regression of the variable log(*totacc*) on a time trend and 11 monthly dummy variables yields the following results:

$$\widehat{\log(totacc)} = \underset{(.035)}{10.47} + \underset{(.0004)}{.0027}t - \underset{(.015)}{.042}\,feb + \underset{(.015)}{.080}\,mar + \underset{(.016)}{.018}\,apr + \underset{(.017)}{.032}\,may + \underset{(.017)}{.020}\,jun$$

$$+ \underset{(.020)}{.038}\,jul + \underset{(.021)}{.054}\,aug + \underset{(.023)}{.042}\,sep + \underset{(.023)}{.082}\,oct + \underset{(.025)}{.071}\,nov + \underset{(.031)}{.096}\,dec$$

$$T = 108,\ R^2 = .797.$$

Interpret the coefficient estimate on the time trend and discuss its significance. Would you say there is seasonality in total accidents? A robust F test on the seasonal dummies has a p-value $< .001$.

(ii) What would happen to the coefficient on the time trend if we run a regression of log(*totacc*) on a time trend and all monthly dummy variables leaving out the intercept?

(iii) Consider the following regression aimed at evaluating the effect on accidents of the introduction of the seatbelt law (*beltlaw* = 1 from January 1986 onwards, 0 otherwise) and the highway speed law which permitted an increase from 55 to 65 miles per hour (*spdlaw* = 1 from May 1987 onwards, 0 otherwise), the number of weekends per month *wkends*, and the unemployment in the state *unem*.

$$\widehat{\log(totacc)} = \underset{(.062)}{10.64} + \underset{(.0026)}{.003}\ wkends - \underset{(.005)}{.021}\,unem - \underset{(.018)}{.054}\,spdlaw + \underset{(.020)}{.095}\,beltlaw$$

$$+ \underset{(.0004)}{.0011}t - \underset{(.016)}{.034}feb + \underset{(.017)}{.077}\,mar + \cdots + \underset{(.024)}{.099}\,dec$$

$$n = 108,\ R^2 = .910.$$

Discuss the importance of including the seasonal dummies and the time trend in this regression.

(iv) Discuss the coefficient (and its significance) on *unem*. Does its sign and magnitude make sense? Explain.

(v) Discuss the coefficients (and their significance) on *spdlaw* and *beltlaw*. Are the estimated effects as expected? Explain.

15.2.2 Using highly persistent time series in regression analysis

Read: Wooldridge, Section 11.3.

Unfortunately, *many economic time series exhibit strong dependence (persistence)* and we will need to learn how to deal with such data. Examples of models that describe this behaviour (persistence) are given by the **random walk** and **random walk with drift**. They are (i) not (covariance) stationary, and (ii) exhibit strong dependence.

Example of a highly persistent process: random walk

A random walk process is defined as:

$$y_t = y_{t-1} + e_t, \quad t = 1, 2, \ldots$$

where $\{e_t\}_{t=1}^T$ is i.i.d. with zero mean and variance σ^2. This is a simple AR(1) process $y_t = \rho y_{t-1} + e_t$, where $\rho = 1$.

This process is nonstationary, as the requirement for the AR(1) process to be covariance stationary is given by $|\rho| < 1$, which is clearly not satisfied here.

To show that this process $\{y_t\}_{t=1}^T$ exhibits strong dependence (persistence), we want to show that $Corr(y_t, y_{t+h})$ does not go to zero when h increases. Let $y_0 = 0$. Using repeated substitution we can then write our random walk as a sum of i.i.d. errors:

$$y_t = e_1 + e_2 + \cdots + e_{t-1} + e_t, \quad t = 1, 2, \ldots.$$

Using this result, you should be able to establish the following.

- $E(y_t) = 0$ and $Var(y_t) = t\sigma^2$ (sum of i.i.d. random variables). The variance increases with time (not constant), hence y_t is nonstationary.
- $Cov(y_t, y_{t+h}) = t\sigma^2$. The covariance (and therefore also the correlation) does not die out as $h \to \infty$ which is required for weak dependence. We have:

$$Cov(y_t, y_{t+h}) \equiv E(y_t y_{t+h}) \quad \text{as} \quad E(y_t) = 0$$

and:

$$\begin{aligned} E(y_t y_{t+h}) &= E((e_1 + \cdots + e_t)(e_1 + \cdots + e_t + \cdots + e_{t+h})) \\ &= E(e_1^2) + \cdots + E(e_t^2) \quad \text{by independence } E(e_t e_s) = 0, \ t \neq s \\ &= t\sigma^2. \end{aligned}$$

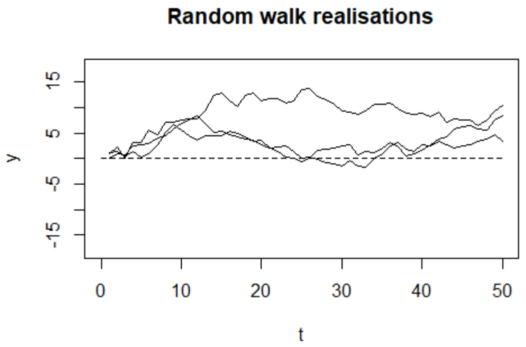
Example of a highly persistent and trending process: random walk with drift

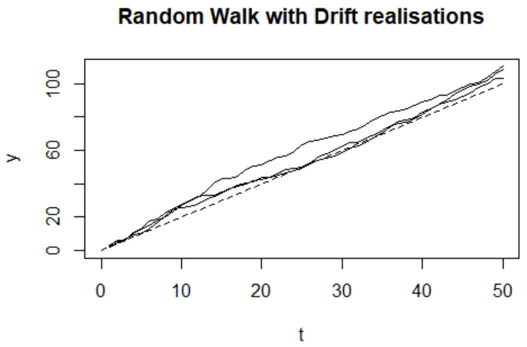
It is often the case that highly persistent series also exhibit a clear trend. A highly persistent process that displays an obvious upward or downward trend is the random walk with drift:

$$y_t = \alpha_0 + y_{t-1} + e_t, \quad t = 1, 2, \ldots$$

where $\{e_t\}_{t=1}^{T}$ is i.i.d. with zero mean and variance σ^2; α_0 is called the **drift term**. When $\alpha_0 = 0$ this is the **random walk**, which does not display a clear trend.

Below we show three realisations of these highly persistent processes: the random walk (on the left) and the random walk with drift (on the right, with $\alpha_0 = 2$) with $y_0 = 0$ and $e_t \sim Normal(0, 1)$:





The dashed lines indicate $E(y_t)$. Unlike the *random walk*, the *random walk with drift* is clearly trending. (Unlike a trend stationary process, the series does not regularly return to the trend line.)

To show the *random walk with drift is trending*, we use repeated substitutions:

$$\begin{aligned} y_t &= \alpha_0 + \underbrace{(\alpha_0 + y_{t-2} + e_{t-1})}_{y_{t-1}} + e_t \\ &= 2\alpha_0 + \underbrace{(\alpha_0 + y_{t-3} + e_{t-2})}_{y_{t-2}} + e_{t-1} + e_t \\ &= \cdots \end{aligned}$$

until we obtain $y_t = \alpha_0 t + y_0 + e_1 + e_2 + \cdots + e_t$. For simplicity, let $y_0 = 0$, so:

$$y_t = \underbrace{\alpha_0 t}_{\text{trend}} + \underbrace{e_t + e_{t-1} + \cdots + e_1}_{\text{random walk behaviour}}.$$

Importance of determining that our stochastic process is highly persistent

From an **economic point of view** determining whether a process exhibits persistence is important as it will indicate whether *shocks or policy changes will have a lasting/permanent effect* (strong dependence) or will only be transitory (weak dependence).

From a **statistical point of view** it is important as it will *render statistical inference (large sample) non-standard.* We will need to be careful to rule out that we have obtained a spurious (meaningless) regression, when running regressions involving highly persistent processes.

To clarify this spurious regression problem, whereby the only reason we have the appearance of a relationship is due to their highly persistent behaviour, let us consider two **independent** random walks (both persistent):

$$y_t = y_{t-1} + e_t \quad \text{with } e_t \sim Normal(0,1)$$
$$x_t = x_{t-1} + u_t \quad \text{with } u_t \sim Normal(0,1),\ e_t, u_t \text{ independent.}$$

We consider the outcome of running the following simple regression:

$$y_t = \alpha_0 + \alpha_1 x_t + v_t.$$

- Typically, this regression will result in an apparent significant t statistic for $\widehat{\alpha}_1$ despite the fact that y_t and x_t are totally unrelated.
- As there is no relationship, $\alpha_1 = 0$, the error term $v_t = y_t - \alpha_0$ will be highly persistent as well which renders standard inference invalid. The distribution of the t statistic for the hypothesis $H_0 : \alpha_1 = 0$ does *not* have anything close to a t distribution. It rejects H_0 much too often, and this problem only gets worse if the sample size increases.

Important: The spurious regression problem cannot be resolved here by including a time trend. When we regress $y_t = \alpha_0 + \alpha_1 x_t + \alpha_2 t + \varepsilon_t$, $\widehat{\alpha}_1$ still often appears significant. The error remains highly persistent and the t *statistic is invalid.*

We will be able to avoid the spurious regression problem, as discussed in Section 11.3b in Wooldridge by working with differenced processes instead. Let $\Delta y_t = y_t - y_{t-1}$ and $\Delta x_t = x_t - x_{t-1}$. When we regress $\Delta y_t = \alpha_1 \Delta x_t + \nu_t$, $\widehat{\alpha}_1$ no longer often appears significant. The error term no longer exhibits persistence and the t *statistic is valid.*

Terminology

A **unit root process** is defined as:

$$y_t = \alpha_0 + y_{t-1} + e_t, \quad t = 1, 2, \ldots$$

where $\{e_t\}_{t=1}^{T}$ is a *weakly dependent process.* The random walk and random walk with drift are special cases of a unit root process where the error term e_t is assumed to be i.i.d. (which is clearly weakly dependent!).

While unit root processes are strongly dependent (persistent), by taking first differences $y_t - y_{t-1}$, denoted Δy_t, we can obtain a weakly dependent process. A process $\{y_t\}$ that becomes weakly dependent after differencing $\{\Delta y_t\}$ is called **integrated of order one** (denoted $I(1)$).

To see this subtract y_{t-1} from both sides to reveal:

$$\Delta y_t = y_t - y_{t-1} = \alpha_0 + e_t, \text{ which is weakly dependent because } e_t \text{ is.}$$

We also call such processes **difference-stationary**.

The terminology ‘integrated of order 0’, or $I(0)$, is used to describe a weakly dependent process. Observe, when $y_t = I(1)$ then $\Delta y_t = I(0)$.

Comment: Running regressions using ‘differenced $I(1)$ variables’ permits us to use standard statistical inference (asymptotic) as they satisfy the required weak dependence.

An informal method that can be used to help decide whether a process is $I(1)$, discussed in Section 11.3c in Wooldridge, uses the **first-order autocorrelation**. In the next section, we consider a formal test to detect a unit root.

Activity 15.5 Take the MCQs related to this section on the VLE to test your understanding.

Activity 15.6 (Based on Exercise 10.C14 from Wooldridge.)

Let us consider the following model:

$$\begin{aligned} approve_t &= \beta_0 + \beta_1 \log(cpifood)_t + \beta_2 \log(rgasprice)_t + \beta_3 unemploy_t \\ &\quad + \beta_4 sep11_t + \beta_5 iraqinvade_t + u_t. \end{aligned}$$

(i) The first-order correlation for the variables *approve* and log(*rgasprice*) equal .931 and .939, respectively. What does this information tell us regarding the dependence inherent in these two time series processes?

(ii) Why might your result in (i) make you hesitant to estimate this model by OLS?

(iii) Estimating the model in first differences results in the following regression line:

$$\begin{aligned} \Delta\widehat{approve}_t &= \underset{(.674)}{.019} - \underset{(219.91)}{125.39}\Delta\log(cpifood_t) - \underset{(7.21)}{13.96}\Delta\log(rgasprice_t) \\ &\quad - \underset{(1.50)}{2.32}\Delta unemploy_t + \underset{(2.98)}{15.57}\Delta sep11_t + \underset{(2.99)}{1.35}\Delta iraqinvade_t \\ &\quad T = 77,\ R^2 = .31. \end{aligned}$$

The numbers in parentheses are standard errors. How do you interpret the estimate of β_2? Is it statistically significant? The p-value is 0.057.

(iv) Interpret the estimate of β_4 and discuss its statistical significance.

(v) Upon adding the variable log(*sp500*) to our model and estimating the equation by first differencing beforehand, we obtain:

$$\begin{aligned} \Delta\widehat{approve}_t &= \underset{(.679)}{.014} - \underset{(221.41)}{127.59}\Delta\log(cpifood_t) - \underset{(7.27)}{13.81}\Delta\log(rgasprice_t) \\ &\quad - \underset{(1.51)}{2.29}\Delta unemploy_t + \underset{(3.05)}{15.76}\Delta sep11_t + \underset{(3.01)}{1.36}\Delta iraqinvade_t \\ &\quad + \underset{(12.78)}{4.18}\,\Delta\log(sp500_t) \\ &\quad T = 77,\ R^2 = .31. \end{aligned}$$

Discuss what this says about the effect of the stock market on approval ratings (*sp500* is a well-known stock market index, the S&P 500).

15.2.3 Testing for unit roots

Read: Wooldridge, Section 18.2.

Section 18.2 in Wooldridge discusses how to test for the presence of a unit root (persistence). The test is non-standard and is called the **Dickey–Fuller (DF) test**.

The discussion of the Dickey–Fuller test begins with the following AR(1) model:

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

where $\{e_t : t = 1, 2, \ldots\}$ is an i.i.d. sequence with zero mean and variance σ^2, and e_t is independent of $y_{t-1}, y_{t-2}, \ldots$. Under the null, $\{y_t\}$ has a unit root and under the alternative we have a stationary, weakly dependent AR(1) process. That is:

$$\begin{aligned} H_0 &: \rho = 1 \quad \text{(non-stationary, highly persistent)} \\ H_1 &: \rho < 1 \quad \text{(stationary, weakly dependent).} \end{aligned}$$

When $\rho > 1$, y_t is non-stationary. When testing for a unit root, it is common to transform the model first by subtracting y_{t-1} from both sides. That is, to test for a unit root we will use the test equation:

$$\Delta y_t = \theta y_{t-1} + e_t, \quad \text{where } \theta = \rho - 1.$$

In this transformed model, testing for a unit root becomes:

$$\begin{aligned} H_0 &: \theta = 0 \quad \text{(non-stationary, highly persistent)} \\ H_1 &: \theta < 0 \quad \text{(stationary, weakly dependent).} \end{aligned}$$

The Dickey–Fuller test for a unit root is given by the familiar t statistic:

$$\text{DF test} = \frac{\widehat{\theta}}{se(\widehat{\theta})}.$$

Due to the one-sided nature of our test ($H_1 : \theta < 0$), our decision rule becomes:

$$\begin{aligned} \text{Not Reject } H_0 &: \text{DF test} \geq c_\alpha \\ \text{Reject } H_0 &: \text{DF test} < c_\alpha \end{aligned}$$

where c_α denotes the critical value given significance level α. It is important to note that the critical value is different from the conventional t test.

To find the critical value, we need the (asymptotic) distribution of our test under H_0. As we have high persistence under the null, $I(1)$, we cannot rely on standard CLT. The asymptotic distribution of our test statistic is derived by Dickey and Fuller (1979). It has come to be known as the **Dickey–Fuller distribution**. A table with critical values is provided below (and will be provided when required in the examination).

Depending on the process of interest, we may want to use an *alternative test equation* to conduct the Dickey–Fuller test. Let us distinguish here the following three possible test equations:

$$\begin{aligned} \Delta y_t &= \theta y_{t-1} + e_t && \blacktriangleleft \text{ No constant} \\ \Delta y_t &= \alpha + \theta y_{t-1} + e_t && \blacktriangleleft \text{ Constant} \\ \Delta y_t &= \alpha + \delta t + \theta y_{t-1} + e_t && \blacktriangleleft \text{ Constant \& Trend} \end{aligned}$$

The selection of the required test equation is fairly intuitive. *We should include an intercept* if our process $\{y_t\}$ has a **non-zero mean**. *We should also include the time trend* if our process $\{y_t\}$ is clearly **trending**.

Regardless of the specific test equation used, we test for the presence of a unit root by testing:

$$H_0 : \theta = 0 \quad \text{against} \quad H_1 : \theta < 0$$

using the familiar test statistic $\widehat{\theta}/se(\widehat{\theta})$. Because of the one-sided nature of our test, we will reject H_0 if the realisation of this DF test statistic is smaller than the relevant critical value.

The critical values are non-standard and are somewhat different depending on which test equation was used to detect unit roots. A table with critical values for different levels of significance is given below:

	Significane level			
	1%	2.5%	5%	10%
Critical value (no constant)	−2.58	−2.23	−1.95	−1.62
Critical value (constant)	−3.43	−3.12	−2.86	−2.57
Critical value (constant & trend)	−3.96	−3.66	−3.41	−3.12

Examples implementing the Dickey–Fuller test

Example 18.2 in Wooldridge, discusses a test for a unit root in the 3-month T-bill rate, $r3_t$. As the process is clearly not trending, the following test equation was used:

$$\Delta r3_t = \alpha + \theta\; r3_{t-1} + e_t.$$

The test failed to reject $H_0 : \theta = 0$, recall $\theta = 1 - \rho$, providing some support that the 3-month T-bill rate may be persistent (at least persistence was not rejected!).

The realisation of the test statistic $\widehat{\theta}/se(\widehat{\theta})$ equals -2.46 which is not smaller than the critical value at the 10% level of significance given by -2.57. We conclude that care should be taken when using this variable in regression analysis.

When testing for a unit root in log(GDP) (see also Example 18.4 in Wooldridge), we may consider the test equation:

$$\Delta \log(GDP)_t = \alpha + \theta\; \log(GDP)_{t-1} + \delta t + e_t.$$

It is useful to recall that $\log(GDP)_t - \log(GDP)_{t-1}$ denotes the growth rate in GDP, which is denoted $gGDP$ in Wooldridge.

Using yearly data in INVEN, the following result is obtained:

$$\Delta \widehat{\log(GDP)}_t = \underset{(.650)}{1.121} - \underset{(.084)}{.140} \log(GDP)_{t-1} + \underset{(.003)}{.0037t}$$

$$T = 36.$$

The DF test is the t statistic on the coefficient on $\log(GDP)_{t-1}$, which equals $-.140/.084 \approx -1.657$. As the DF test statistic is not smaller than the critical value at the 5% level (-3.41), we fail to reject the null, which is suggestive that $\log(GDP)$ has a unit root, i.e. $\log(GDP) = I(1)$. It should be noted that the sample is quite small.

Augmented Dickey–Fuller test

A key assumption underlying the **Dickey–Fuller test** is that **there is no autocorrelation in the errors**.

The Dickey–Fuller test assumes that the process of interest could be represented by an AR(1), where the errors do not have any autocorrelation (recall the definition of an AR(1) process).

If the true process is given by a higher order AR process (more complicated dynamics), the error in the DF test equations will exhibit autocorrelation (invalidates the standard errors). To resolve this problem, the Augmented Dickey–Fuller test (ADF) includes lagged differences $\Delta y_{t-1}, \Delta y_{t-2}, \ldots$ to our test equation, where the analyst needs to select the number of lagged differences to include.

When the true process can be described by an AR(2) process, we simply add Δy_{t-1} and consider the following three possible test equations:

$$\Delta y_t = \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t \qquad \blacktriangleleft \text{ No constant}$$

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t \qquad \blacktriangleleft \text{ Constant}$$

$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + e_t \qquad \blacktriangleleft \text{ Constant \& Trend}$$

everything else remains the same.

Examples implementing the Augmented Dickey–Fuller test

Example 18.3 in Wooldridge provides a test for a unit root in US inflation, inf_t. As the process is clearly not trending, the following test equation was used:

$$\Delta inf_t = \alpha + \theta inf_{t-1} + \gamma_1 \Delta inf_{t-1} + e_t.$$

Example 18.4 in Wooldridge considers the ADF when testing for a unit root in $\log(GDP)$. The testing equation is given by:

$$\Delta \log(GDP)_t = \alpha + \theta \log(GDP)_{t-1} + \gamma_1 \Delta \log(GDP)_{t-1} + \delta t + e_t.$$

You should study these and other examples presented in Wooldridge.

Activity 15.7 Take the MCQs related to this section on the VLE to test your understanding.

Activity 15.8 Let us consider the following model:

$$y_t = \alpha + u_t \quad \text{where} \quad u_t = \rho u_{t-1} + e_t$$

and e_t is i.i.d. with zero mean and variance σ^2.

(i) What condition will ensure that $\{u_t\}_{t=1}^T$ is stationary and weakly dependent? Explain why under these conditions $\{y_t\}_{t=1}^T$ is also stationary and weakly dependent.

(ii) Show that you can rewrite the equation in the form:

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + e_t$$

where $\beta_0 = \alpha(1 - \rho)$ and $\beta_1 = \rho - 1$.

Hint: Subtract from the original model ρ times the model one period lagged.

(iii) Show that under the null of a unit root, y_t is a random walk (not a random walk with drift).

(iv) Discuss the Dickey–Fuller test in detail. (Provide the null and alternative hypotheses, test statistic, and rejection rule.)

Activity 15.9 Let us consider the following model:

$$y_t = \alpha + \delta t + u_t \quad \text{where} \quad u_t = \rho u_{t-1} + e_t$$

and e_t is i.i.d. with zero mean and variance σ^2.

(i) Discuss the following statement 'If $\{u_t\}_{t=1}^T$ is stationary and weakly dependent then $\{y_t\}_{t=1}^T$ is trend-stationary (and weakly dependent).'

(ii) Show that you can write the equation in the following form:

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 t + e_t$$

where $\beta_0 = \alpha(1 - \rho) + \rho\delta$, $\beta_1 = \rho - 1$, and $\beta_2 = \delta(1 - \rho)$.

Hint: Subtract from the original model ρ times the model one period lagged.

(iii) Show that under the null of a unit root, y_t is a random walk with drift.

(iv) Discuss the Dickey–Fuller test in detail. (Provide the null and alternative hypotheses, test statistic, and rejection rule.)

15.2.4 Spurious regression, cointegration and error correction models

Read: Wooldridge, Sections 18.3 and 18.4.

Let us recall when running a regression using two (or more) highly persistent processes we will need to rule out that we have obtained a spurious relationship, where the only reason we have the appearance of a relationship is due to their persistence.

If y_t and x_t are both $I(1)$, we have a spurious regression problem if the error ε_t is $I(1)$ as well, that is:

$$y_t = \alpha_0 + \alpha_1 x_t + \varepsilon_t \quad \text{with} \quad \varepsilon_t \sim I(1)$$

results in large R^2 and a seemingly significant α_1 even though y_t and x_t are unrelated. Inference is not standard as usual CLT does not apply.

By differencing these processes we can avoid this problem. However, differencing $I(1)$ variables beforehand is not always needed and, indeed, may limit the scope of questions that we can answer.

Engle and Granger (1987) introduced the notion of **cointegration** to describe meaningful regressions involving $I(1)$ regressors.

If y_t and x_t are both $I(1)$, we have a cointegrating relationship if the error ε_t is $I(0)$, that is:

$$y_t = \alpha_0 + \alpha_1 x_t + \varepsilon_t \quad \text{with} \quad \varepsilon_t \sim I(0).$$

It recognises the fact that there may be a **long-run relationship** between the highly persistent processes. In this case y and x move together.

Statistically, in this case there exists a linear combination of the $I(1)$ processes y_t and x_t that is $I(0)$, here $y_t - \alpha_1 x_t$. We use the term cointegration to describe the setting where the order of integration of a linear combination of processes is lower than the order of integration of the original processes.

α_1 is called the **cointegration parameter**.

Testing for cointegration

Say we have determined that y_t and x_t are $I(1)$. How do we test whether our regression:

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

is a spurious (meaningless) regression or a cointegrating regression, reflective of a long-run relationship? In light of the discussion above, *our test will be based on testing whether the residuals from the above regression*:

$$\widehat{u}_t = y_t - \widehat{\beta}_0 - \widehat{\beta}_1 x_t$$

have a unit root or not. Formally, we want to test:

$$H_0 : \{u_t\} \text{ is } I(1) \text{ (spurious regression)}$$

$$H_1 : \{u_t\} \text{ is } I(0) \text{ (cointegrating regression).}$$

Under H_0, our regression is spurious and the residuals will display the presence of a unit root. Under H_1, our regression is cointegrating and the residuals will not display the presence of a unit root. We can perform this test by applying a DF (or ADF) test to the residuals. The test is called the **Engle–Granger test**. The test proceeds as follows.

- Using the OLS residuals, $\widehat{u}_t$, obtain the t statistic for θ in the model:

$$\Delta\widehat{u}_t = \mu + \theta\widehat{u}_{t-1} + e_t.$$

- With α denoting the significance level:

$$\text{Not Reject } H_0 : \frac{\widehat{\theta}}{se(\widehat{\theta})} \geq c_\alpha \text{ spurious regression}$$

$$\text{Reject } H_0 : \frac{\widehat{\theta}}{se(\widehat{\theta})} < c_\alpha \text{ cointegrating regression.}$$

- Due to the estimation error in $\widehat{u}_t$, the critical values are somewhat different from those used in the DF test.
- In principle we can add lags of $\Delta\widehat{u}_t$ (ADF) and include a trend here too.

Error correction model (ECM)

If y_t and x_t are $I(1)$ and they are *not cointegrated*, then we can only specify models based on first differences Δy_t and Δx_t as they are $I(0)$. For example:

$$\Delta y_t = \delta_0 + \delta_1\Delta x_t + \varepsilon_t.$$

(You could also add lags like Δx_{t-j}, Δy_{t-j}, so on.) On the other hand, if y_t and x_t are $I(1)$ and they *are cointegrated*, then we can enrich our model by specifying an error correction model (ECM).

- **Idea**: When there is a long-run relationship, there has to be some mechanism that will bring about equilibrium after a shock. By estimating the ECM, we can quantify this process.

Specifically, if y_t and x_t are $I(1)$ and they are cointegrated, then we can include an additional $I(0)$ variable:

$$s_t = y_t - \alpha - \beta x_t$$

which represents the deviation from the long-run equilibrium. This suggests the **error correction model (ECM)**:

$$\Delta y_t = \delta_0 + \delta_1\Delta x_t + \gamma s_{t-1} + \varepsilon_t$$

whereby changes in y at time t, Δy_t, respond to changes in x at time t, Δx_t and a disequilibrium at time $t-1$, s_{t-1}. (You could try adding lags like Δx_{t-j}, Δy_{t-j}, and so on.)

- The term *γs_{t-1} is called the error correction term*, where s_{t-1} represents a deviation from the long-run equilibrium (error) at time $t-1$:
$$s_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}.$$
- As long as $\gamma < 0$, we get the following error correction mechanism:
  - If $s_{t-1} > 0$ (y overshoots equilibrium level at $t-1$), then γs_{t-1} will push y back toward equilibrium level at t. (Same comment applies to $s_{t-1} < 0$.)
  - We should expect $-1 \leq \gamma < 0$. With a smaller γ, the adjustment pushing us back towards equilibrium is stronger. When $\gamma > 0$, we will see a movement further away from the equilibrium level which is nonsensical, and when $\gamma = 0$ there is no ECM (which implies that there is no cointegration).
- The ECM, therefore, *allows us to study the short-run dynamics in the relationship between y and x* required to re-establish equilibrium.

Estimating the error correction model

Let us discuss how we can estimate the error correction model:

$$\Delta y_t = \delta_0 + \delta_1 \Delta x_t + \gamma s_{t-1} + \varepsilon_t$$

where:

$$s_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}.$$

Recall x_t and y_t are $I(1)$ and β is the cointegrating parameter, representing the long-run relationship between x_t and y_t. s_{t-1} represents the disequilibrium at time $t-1$. If the cointegating relationship is known, we can compute $s_{t-1} = y_{t-1} - \alpha - \beta x_{t-1}$ and can apply OLS using the differenced processes Δy_t, Δx_t and s_{t-1} to obtain the ECM. *If the cointegating relationship is unknown*, we will need to estimate it first. This procedure is called the **Engle–Granger two-step procedure.**

- **Step 1**: Estimate the cointegrating relationship by OLS, and obtain the residuals:
$$\widehat{s}_t = y_t - \widehat{\alpha} - \widehat{\beta} x_t.$$
- **Step 2**: Estimate the ECM by OLS using $\widehat{s}_{t-1}$ in place of s_{t-1}:
$$\Delta y_t = \delta_0 + \delta_1 \Delta x_t + \gamma \widehat{s}_{t-1} + e_t.$$

  Conveniently, for hypothesis testing we can ignore the fact that we use an estimate of s_{t-1} in place of s_{t-1} for inference purposes (related to the non-examinable concept of super-consistency).

The discussion in Wooldridge (seventh edition), deviates slightly from our discussion here. In Wooldridge, the assumption is made that $s_t = y_t - \beta x_t$ has zero mean. By describing the cointegrating relationship as $y_t = \alpha + \beta x_t$ (adding an intercept) we avoid this.

Example: estimating the error correction mechanism

From Wooldridge: Let $hy6_t$ be the three-month holding yield (in percentage) from buying a six-month T-bill and selling it three months later as a three-month T-bill, and let $hy3_t$ be the three-month holding yield (in percentage) from buying a three-month T-bill. **We assume that both processes are $\boldsymbol{I(1)}$**.

Using INTQRT, the Dickey Fuller test on $hy3_t$ finds support for the process being $I(1)$; no support was found for $hy6_t$ being $I(1)$.

We expect there is a cointegrating relationship between $hy6_t$ and $hy3_{t-1}$. In fact, the *expectations hypothesis* suggests that these two different three-month investments should be the same on average. Using INTQRT we obtain the following cointegrating regression:

$$\widehat{hy6_t} = -\underset{(.070)}{.058} + \underset{(.039)}{1.104}hy3_{t-1}, \quad T = 123.$$

The **Engle–Granger test** confirms that this is a cointegrating relationship.

Recall, this test performs the Dickey–Fuller test on the residuals from this regression. The t statistic is -10.86, which strongly suggests cointegration:

$$\widehat{\Delta resid_t} = -\underset{(.030)}{.0005} - \underset{(.091)}{0.992}\, resid_{t-1}.$$

To estimate the ECM for holding yields we use the **Engle–Granger two-step procedure**:

- **Step 1**: Using the cointegrating relationship (estimated by OLS):

  $$\widehat{hy6_t} = -\underset{(.070)}{.058} + \underset{(.039)}{1.104}hy3_{t-1}, \quad T = 123$$

  we obtain the residuals from this regression $\widehat{u}_t$.

- **Step 2**: We estimate the ECM by OLS, which yields:

  $$\widehat{\Delta hy6_t} = -\underset{(.030)}{.001} + \underset{(.222)}{1.193}\Delta hy3_{t-1} - \underset{(.211)}{.916}\widehat{u}_{t-1}.$$

  The error correction parameter is statistically significant (can use the asymptotic t test $\widehat{\gamma}/se(\widehat{\gamma})$, revealing the existence of an error correction mechanism ($\gamma \neq 0$)). In fact, the adjustment to a disequilibrium is not significantly different from -1, which would permit the adjustment to be immediate.

Activity 15.10 Take the MCQs related to this section on the VLE to test your understanding.

Activity 15.11 Suppose the relationship between two $I(1)$ variables y_t and x_t can be characterised by the ADL$(1,1)$ model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + e_t$$

where $|\beta_1| < 1$, e_t is i.i.d. and $E(e_t \,|\, x_{t-1}, x_{t-2}, \ldots, y_{t-1}, y_{t-2}, \ldots) = 0$.

(i) Show that in equilibrium:

$$y = \alpha_0 + \alpha_1 x$$

where $\alpha_0 = \frac{\beta_0}{1-\beta_1}$ and $\alpha_1 = \frac{\beta_2+\beta_3}{1-\beta_1}$.

(ii) Hence, discuss that the cointegrating relationship between $\{y_t\}$ and $\{x_t\}$ is given by:

$$y_t = \alpha_0 + \alpha_1 x_t.$$

(iii) Show that you can rewrite the ADL$(1,1)$ model to obtain the following error correction model (ECM):

$$\Delta y_t = \beta_2 \Delta x_t + (\beta_1 - 1)\, diseq_{t-1} + e_t$$

where $diseq_{t-1} = y_{t-1} - \alpha_0 - \alpha_1 x_{t-1}$. *Hint:* First subtract y_{t-1} from both sides of the first equation. Then, add and subtract $\beta_2 x_{t-1}$ from the right-hand side and rearrange.

(iv) Discuss how you can estimate the ECM given in (iii).

15.2.5 R application for nonstationary and persistent time series models

Activity 15.12 Complete 'R Task – Activity 8' on VLE.

Here we show how to account for nonstationarity when using time series data in R. We discuss how to account for trends and seasonality. We also show how to simulate a random walk and test for the presence of a unit root using the ADF test.

15.3 Answers to activities

Solution to Activity 15.2

(Based on Exercise 10.C6 from Wooldridge.)

- To obtain the detrended *gfr* variable we need to run the following regression:

$$gfr_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + v.$$

- The detrended variable $\ddot{gfr}_t$ is given by the residuals from this regression:

$$\ddot{gfr}_t = gfr_t - \widehat{\alpha}_0 - \widehat{\alpha}_1 t - \widehat{\alpha}_2 t^2.$$

- The procedure to obtain the detrended variables $\ddot{pe}_t$, $\ddot{ww}2_t$, and $\ddot{pill}_t$ is identical.
- When we run a regression using the detrended variables (no intercept), we obtain:

$$\widehat{\ddot{gfr}} = .348\ddot{pe} - 35.88\ddot{ww}2 - 10.12\ddot{pill}.$$

(You can convince yourself by running the accompanying R script.)

- **Conclusion**: We can obtain the effects of interest by first partialling out the trend. How we partial out the trend depends on our assumption regarding the deterministic trend (linear or quadratic).

Solution to Activity 15.3

(Based on Exercise 10.C6 from Wooldridge.)

(i) The use of robust (HAC) standard errors accounts for the possible presence of weak dependence and heteroskedaticity in the errors.

The usual standard errors are only valid under the assumption of homoskedasticity and absence of serial correlation where the latter in particular is likely to be violated in a time series setting.

The fact that the robust standard errors are quite different from the regular standard errors is indicative of the presence of heteroskedasticity and/or autocorrelation.

(ii) Using the robust standard errors we can simply apply the **asymptotic t test**:

$$\frac{\widehat{\beta}_{pe}}{se(\widehat{\beta}_{pe})} = \frac{.347}{.078} = 4.42.$$

Under H_0 its distribution is approximately $Normal(0, 1)$.

We find strong evidence against H_0, as our one-sided critical value at the 1% level of significance equals $z_{.01} = 2.326$.

We assume that our detrended variables satisfy assumption TS.1′ (stationarity and weak dependence), we do not have perfect multicollinearity, and our regressors are contemporaneously exogeneous TS.3′ (weaker than strict exogeneity TS.3).

As we are using robust standard errors we do not assume homoskedasticity or absence of serial correlation.

Solution to Activity 15.4

(Based on Exercise 10.C11 from Wooldridge.)

(i) When multiplied by 100, the coefficient on t gives roughly the average monthly percentage growth in *totacc*, *ceteris paribus*.

Once seasonality was eliminated, *totacc* grew by about .275% per month over this period, or 12(.275) = 3.3% at an annual rate.

Only February has a lower number of total accidents than the base month, January. The peak is in December: roughly, there are 9.6% more accidents in December than January in the average year.

The fact that the robust F statistic on the seasonal dummies has a p-value $< .001$ ensures that the dummies are jointly highly significant (under $H_0 : F \overset{a}{\sim} F_{11,\,108-13}$).

(ii) The estimate on the time trend, and its interpretation, would remain unchanged if instead we had run a regression that included all 12 seasonal dummies, the time trend but excluded the intercept.

To avoid perfect multicollinearity we either have to leave out one seasonal dummy, or the intercept!

The parameters on the dummy variables would be affected; their interpretation changes. (*jan* will have the estimate 10.47, the estimate for *feb* is $10.47 - .042$, for *mar* $10.47 + .080$ etc.)

(iii) The fact that the number of accidents exhibit trends and seasonality requires us to include these variables so as not to give rise to bias arising from confounders and permit us to assume TS.1′.

Their exclusion is likely to lead to biased parameter estimates as we expect positive correlation between the traffic laws and time (both laws implemented later in the sample) and seasonality in unemployment rates.

(iv) The negative coefficient on *unem* makes sense if we view *unem* as a measure of economic activity. As economic activity increases – *unem* decreases – we expect more driving and therefore more accidents.

The estimate indicates that a one-unit increase in the unemployment rate (a 1% point increase) reduces the total accidents by about 2.1% (semi-elasticity), *ceteris paribus*.

The asymptotic t test $\widehat{\beta}_{unem}/se(\widehat{\beta}_{unem})$ shows it is statistically significant.

(v) Using asymptotic t tests both can be shown to be statistically significant.

$\widehat{\beta}_{spdlaw} = -.054$: Accidents dropped by about 5.4% after the highway speed limit was increased from 55 to 65 miles per hour, *ceteris paribus*. Perhaps, people became safer drivers after the speed limit was increased.

$\widehat{\beta}_{beltlaw} = .095$: Accidents increased by about 9.5% after the introduction of the seatbelt law, *ceteris paribus*. Perhaps, people became less cautious once they were forced to wear seatbelts. It could also be that some other change caused changes in the total number of accidents that we have not properly accounted for (OVB).

Solution to Activity 15.6

(Based on Exercise 10.C14 from Wooldridge.)

(i) The first-order autocorrelation values for *approve* and *lrgasprice* are about .931 and 0.939, respectively, which are fairly close to unity. This may be indicative of a unit root process, which is nonstationary and highly persistent (strong dependence).

(ii) The fact that it is likely we are dealing with highly persistent (and nonstationary) processes makes us hesitant to estimate the model by OLS because it may give rise to a spurious regression.

(iii) Notice that the slopes in this regression that uses all variables in first differences provides the estimates of the slopes in our original model. Recall:

$$y_t = \alpha_0 + \alpha_1 x_t + e_t, \quad \text{can be rewritten as} \quad \Delta y_t = \alpha_1 \Delta x_t + v_t.$$

The interpretation of $\widehat{\beta}_2$ is as follows: An approximate 1% growth in *rgasprice* ($\Delta \log(\mathit{rgasprice}) = .01$) reduces the approval rate with 1.4 units (1.4 percentage points), *ceteris paribus.*

The t statistic $\widehat{\beta}_2 / se(\widehat{\beta}_2) \approx -1.94$. With a p-value $= .057$, the estimate is not statistically significant at the 5% level, but it is significant at the 6% level.

(iv) The estimate of β_4 is given by the coefficient on $\Delta sep11_t$. The parameter is statistically significant with a t statistic of 5.2 and a p-value $< .001$.

To interpret, it is easier to recall the equation in levels:

$$\begin{aligned} \mathit{approve}_t = \beta_0 &+ \beta_1 \log(\mathit{cpifood})_t + \beta_2 \log(\mathit{rgasprice})_t + \beta_3 \mathit{unemploy}_t \\ &+ \beta_4 sep11_t + \beta_5 \mathit{iraqinvade}_t + u_t. \end{aligned}$$

Hence, we observe that the approval rate was 15.6 percentage points higher in September 2001 and the following two months, *ceteris paribus.*

(v) The estimate of the effect on $\log(sp500)$ equals 4.18. The effect is statistically insignificant with a t statistic of .33 and a p-value of .745. After controlling for the other variables, there is no effect of the stock market valuation on approval ratings.

An approximate 1% growth in the value of the S&P 500 ($\Delta \log(sp500) = .01$) increases the approval rate by only .04 percentage points (economically insignificant too).

Solution to Activity 15.8

(i) This describes the process $\{y_t\}$ as the sum of the constant α and an AR(1) process given by $\{u_t\}$.

- $\{u_t\}$ is an AR(1) process that is stationary and weakly dependent when $|\rho| < 1$.
- As $\{y_t\}$ simply shifts over $\{u_t\}$ with a constant α, it is also a stationary and weakly dependent process when $|\rho| < 1$.

(ii) Let us follow the hint.

- Subtract from the original model ρ times the model one period lagged

$$\begin{array}{ll} y_t & = \alpha + u_t \\ \rho y_{t-1} & = \rho\alpha + \rho u_{t-1} \quad - \\ \hline y_t - \rho y_{t-1} & = \alpha(1-\rho) + \underbrace{u_t - \rho u_{t-1}}_{e_t}. \end{array}$$

- We can write this as:

$$y_t = \alpha(1-\rho) + \rho y_{t-1} + e_t \quad \text{where} \quad e_t \text{ i.i.d.}(0, \sigma^2).$$

- Subtracting y_{t-1} from both sides yields the answer:

$$\Delta y_t = \underbrace{\alpha(1-\rho)}_{\beta_0} + \underbrace{(\rho-1)}_{\beta_1} y_{t-1} + e_t.$$

(iii) When $\rho = 1$, we will have $\beta_0 = 0$ and $\beta_1 = 0$, so that we obtain:

$$\Delta y_t = e_t \quad \text{or} \quad y_t = y_{t-1} + e_t, \quad e_t \text{ i.i.d.}(0, \sigma^2).$$

This shows that y_t is a random walk.

(iv) The Dickey–Fuller test requires us to estimate:

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + e_t.$$

We will test:

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 < 0$$

(even though $\beta_0 = 0$ at the same time) using the t statistic:

$$\text{DF test} = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)}.$$

At the 5% level of significance, our decision rule is given by:

$$\begin{aligned} \text{Not Reject } H_0 &: \text{DF test} \geq -2.86 \\ \text{Reject } H_0 &: \text{DF test} < -2.86. \end{aligned}$$

Solution to Activity 15.9

(i) This describes the process $\{y_t\}$ as the sum of α, a deterministic trend δt, and an AR(1) process given by $\{u_t\}$.

- If $\{u_t\}$ is stationary and weakly dependent, $|\rho| < 1$ then the only 'problem' with the $\{y_t\}$ process is the presence of the deterministic trend ($\delta \neq 0$).
- By detrending $\{y_t\}$ can be made stationary and weakly dependent:

$$y_t - \delta t = \alpha + u_t.$$

We have called such a process **trend-stationary**.

(ii) Let us follow the hint:

- Subtract from the original model ρ times the model one period lagged:

$$\begin{array}{lcccccccl} y_t & = & \alpha & + & \delta t & + & u_t & & \\ \rho y_{t-1} & = & \rho\alpha & + & \rho\delta(t-1) & + & \rho u_{t-1} & & - \\ \hline y_t - \rho y_{t-1} & \multicolumn{8}{l}{= \alpha(1-\rho) + \rho\delta + \delta(1-\rho)t + \underbrace{u_t - \rho u_{t-1}}_{e_t}.} \end{array}$$

- We can write this as:

$$y_t = \alpha(1-\rho) + \rho\delta + \rho y_{t-1} + \delta(1-\rho)t + e_t.$$

- Subtracting y_{t-1} from both sides yields the answer:

$$\Delta y_t = \underbrace{\alpha(1-\rho) + \rho\delta}_{\beta_0} + \underbrace{(\rho-1)}_{\beta_1} y_{t-1} + \underbrace{\delta(1-\rho)}_{\beta_2} t + e_t.$$

(iii) When $\rho = 1$, we will have $\beta_0 = \delta$, $\beta_1 = 0$, and $\beta_2 = 0$, so that we obtain:

$$\Delta y_t = \delta + e_t \quad \text{or} \quad y_t = \delta + y_{t-1} + e_t, \quad e_t \text{ i.i.d.}(0, \sigma^2)$$

This defines a random walk with drift.

In this case $\{y_t\}_{t=1}^T$ is $I(1)$. We have called such a process **difference-stationary**.

(iv) The Dickey–Fuller test requires us to estimate:

$$\Delta y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 t + e_t.$$

We will test:

$$H_0 : \beta_1 = 0 \quad \text{against} \quad H_1 : \beta_1 < 0$$

(even though $\beta_2 = 0$ at the same time) using the t statistic:

$$\text{DF test} = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)}.$$

At the 5% level of significance level, our decision rule is given by:

$$\text{Not Reject } H_0 : \text{DF test} \geq -3.41$$

$$\text{Reject } H_0 : \text{DF test} < -3.41.$$

■ Recall, we started with the following process:

$$y_t = \alpha + \delta t + u_t, \quad \text{where } u_t = \rho u_{t-1} + e_t$$

which we were asked to rewrite as follows:

$$\Delta y_t = \underbrace{\alpha(1-\rho) + \rho\delta}_{\beta_0} + \underbrace{(\rho-1)}_{\beta_1} y_{t-1} + \underbrace{\delta(1-\rho)}_{\beta_2} t + e_t.$$

Solution to Activity 15.11

(i) In equilibrium (y^e, x^e), using $E(e) = 0$, we have:

$$y^e = \beta_0 + \beta_1 y^e + \beta_2 x^e + \beta_3 x^e.$$

We can rewrite this relationship as:

$$y^e = \underbrace{\frac{\beta_0}{1-\beta_1}}_{\alpha_0} + \underbrace{\frac{\beta_2+\beta_3}{1-\beta_1}}_{\alpha_1} x^e.$$

(ii) The cointegrating relationship reveals the long-run relationship between the $I(1)$ processes, hence:

$$y_t = \alpha_0 + \alpha_1 x_t.$$

$\{y_t : \ t = 1, 2, \ldots, \}$ and $\{x_t : \ t = 1, 2, \ldots, \}$ are $I(1)$, but there is an α_1 such that $y_t - \alpha_0 - \alpha_1 x_t$ is $I(0)$. This α_1 is the cointegrating parameter.

(iii) Rewriting the above equation in ECM; we follow the hint.

- Subtract y_{t-1} from both sides:

$$\Delta y_t = \beta_0 + (\beta_1 - 1) y_{t-1} + \beta_2 x_t + \beta_3 x_{t-1} + e_t.$$

- Next, add and subtract $\beta_2 x_{t-1}$ and rearrange:

$$\begin{aligned}
\Delta y_t &= \beta_0 + (\beta_1 - 1) y_{t-1} + \underbrace{\beta_2 x_t \ [-\beta_2 x_{t-1}}_{} + \underbrace{\beta_2 x_{t-1}] \ + \beta_3 x_{t-1}}_{} + e_t \\
&= \beta_0 + (\beta_1 - 1) y_{t-1} \quad + \quad \beta_2 \Delta x_t + \quad (\beta_2 + \beta_3) x_{t-1} + e_t \\
&= \beta_2 \Delta x_t + \underbrace{(\beta_1 - 1)\left(y_{t-1} - \frac{\beta_0}{1 - \beta_1} - \frac{\beta_2 + \beta_3}{1 - \beta_1} x_{t-1} \right)}_{diseq_{t-1}} + e_t.
\end{aligned}$$

$\beta_1 - 1$ indicates the speed of adjustment.

β_2 indicates the short-run effect of a change in x on y.

(iv) Let us write the ECM as:

$$\Delta y_t = \beta_2 \Delta x_t + \gamma \left(y_{t-1} - \alpha_0 - \alpha_1 x_{t-1} \right) + e_t$$

where $\gamma = \beta_1 - 1 < 0$.

The Engle–Granger two-step procedure is given by the following.

- **Step 1**: We estimate the cointegrating relationship by OLS:

$$y_t = \alpha_0 + \alpha_1 x_t + v_t$$

from which we obtain:

$$\widehat{v}_t = y_t - \widehat{\alpha}_0 - \widehat{\alpha}_1 x_t.$$

- **Step 2**: We estimate the ECM by running the following OLS regression:

$$\Delta y_t = \beta_2 \Delta x_t + \gamma \widehat{v}_{t-1} + \varepsilon_t, \quad \text{where } \gamma = \beta_1 - 1.$$

15.4 Overview of chapter

Many economic time series exhibit characteristics that require special attention when applying regression analysis. While assumptions of covariance stationarity and weak dependence are convenient assumptions we may want to impose in order to guarantee consistency of OLS and allow us to apply the usual inference procedures (as discussed before) this may not always be satisfied by economic time series.

Violations of stationarity are commonplace as many time series processes exhibit seasonality and display trends and we discussed in what settings the introduction of seasonal dummies and the inclusion of a time trend will permit the usual statistical inference to hold.

The presence of strong dependence (persistence) poses a more serious issue for statistical analysis. We discussed how to test for persistence (presence of unit roots) with a non-standard (Augmented) Dickey–Fuller test. Two examples of highly persistent processes are the random walk and random walk with drift, and we showed that by differencing such processes (instead of detrending) we can remove the problem of nonstationarity inherent in these processes. These processes are also called difference-stationary, or integrated of order 1, denoted $I(1)$ (by differencing them once, they can be made stationary and, more importantly, weakly dependent).

When dealing with highly persistent processes, we can avoid the spurious regression problem by differencing our processes beforehand. Simply differencing $I(1)$ processes before running regression analysis does limit the scope of questions we can answer, as there may be meaningful (as opposed to spurious) relationships that describe the long-run (cointegrating) relationship between highly persistent processes. Over time these variables co-move in a way that their difference has a zero mean and their differences return to zero with some regularity. The presence of a long-run relationship between economic variables, permits us to describe an error correction model (ECM), which enables us to study the short-run dynamics in their relationship.

15.4.1 Key terms and concepts

- Nonstationary process
- Spurious regression problem
- Time trend (linear and exponential)
- Detrending
- Seasonality
- Seasonally adjusted processes
- Trend-stationary process
- Autoregressive process of order 1, AR(1)
- Strong dependence (persistence)
- Unit roots
- Random walk
- Random walk with drift
- Dickey–Fuller test and Augmented Dickey–Fuller test
- First difference

- Integrated of order d, $I(d)$
- Difference-stationary process
- Cointegration and spurious regression
- Engle–Granger test
- Error correction model
- Engle–Granger two-step procedure

15.4.2 A reminder of your learning outcomes

After completing this chapter, and having completed the accompanying reading and activities, you should be able to:

- understand the concept of trending time series and seasonality
- distinguish linear versus exponential trends
- explain the spurious regression problem
- understand the importance of including a time trend in regressions with deterministically trending variables
- understand what we mean by a detrended time series and a seasonally adjusted time series
- understand the concept of a highly persistent (strongly dependent) time series process
- explain the importance of the correlogram in identifying whether a process is highly persistent
- formulate a random walk and a random walk with drift
- explain the terminology of a unit root process, an integrated process and the order of integration
- discuss the Dickey–Fuller (DF) and augmented Dickey–Fuller (ADF) unit root tests and explain their diference
- discuss what we mean by a 'trend stationary' and a 'difference stationary' process, and discuss how we can use the unit root test to distinguish between them
- understand the importance of testing for cointegration when using highly persistent variables in regression
- explain the concept of cointegration in regression models
- conduct a unit root test on the residuals to test for cointegration (Engle–Granger test)

- construct an error correction model (ECM) and describe the advantage over differencing
- explain the Engle–Granger two-step procedure for estimating the ECM
- use statistical software to estimate time series regressions in the presence of nonstationarity and implement unit root test on real data.

15.5 Test your knowledge and understanding

15.5.1 Additional examination based exercises

15.1E This question is based on 'An Empirical Comparison of Alternative Models of the Short-Term Interest Rate', by Chan et al. (*The Journal of Finance*, 1992). In this article, the authors are interested in a regression model for the short-term interest rate r_t given by:

$$r_t - r_{t-1} = \beta_0 + \beta_1 r_{t-1} + \varepsilon_t.$$

Assume that the errors ε_t are i.i.d.

(a) Provide sufficient conditions (on β_1) for r_t to be stationary and weakly dependent. Will the OLS estimator for a linear regression of r_t on r_{t-1} in this case be unbiased and consistent? Explain your answer.

(b) A few well-known models in finance hypothesise that $\beta_1 = 0$.

i. With $r_0 = 0$, show that if $\beta_1 = 0$ we can write:

$$r_t = t\beta_0 + \varepsilon_t + \varepsilon_{t-1} + \cdots + \varepsilon_1.$$

What does this equation reveal regarding the properties of the $\{r_t\}$ process, and what name do we give the process r_t?

ii. Explain how you would test this hypothesis.

15.2E Consider the following time series model for $\{y_t\}_{t=1}^T$:

$$y_t = \alpha + \beta t + u_t, \quad t = 1, \ldots, T \quad \text{with } u_t = \rho u_{t-1} + \varepsilon_t \text{ and } 0 \leq \rho \leq 1$$

where ε_t is an i.i.d. $(0, \sigma_\varepsilon^2)$ error that is uncorrelated with anything in the past (white noise). In this model $\{y_t\}_{t=1}^T$ will be nonstationary if $\beta \neq 0$ or $\rho = 1$.

(a) Discuss the concept 'trend stationarity' and contrast it to the concept 'difference stationarity'.

(b) Show that y_t is trend stationary when $|\rho| < 1$.

(c) Show that y_t is difference stationary when $\rho = 1$.

(d) Discuss the importance of distinguishing between trend stationary and difference stationary processes.

(e) What assumptions do you make if you want to argue that we can use OLS to get a consistent estimator of α and β? Will the OLS estimator be efficient in this case, and can you use the usual standard errors for inference?

15.3E (a) Consider the following regression model:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \quad t = 1, \ldots, T.$$

Both $\{Y_t\}_{t=1}^{T}$ and $\{X_t\}_{t=1}^{T}$ are $I(1)$ processes.

How can you test whether the above relation is spurious or cointegrating? In your answer you are expected to explain the difference between a spurious and cointegrating relationship. What implication does your answer have when you are interested in testing the hypothesis $\beta_1 = 0$?

(b) Let us consider the following OLS regression output regarding the above model:

$$Y_t = \underset{(0.1305)}{3.0920} + \underset{(0.0103)}{0.6959} X_t + \widehat{\varepsilon}_t$$

$$R^2 = .99,\ F = 4523.5,\ s = .0236,\ DW = .557,\ T = 740.$$

$$\Delta\widehat{\varepsilon}_t = -\underset{(0.0845)}{.216}\,\widehat{\varepsilon}_{t-1} + \underset{(0.1592)}{.2349}\,\Delta\widehat{\varepsilon}_{t-1} + \underset{(0.1631)}{.2029}\,\Delta\widehat{\varepsilon}_{t-2} + \widehat{u}_t$$

$$R^2 = .1799,\ s = .0115,\ T = 737$$

where s is the standard error of the residuals, T is the number of observations, $\widehat{\varepsilon}_t$ and $\widehat{u}_t$ are OLS residuals, and standard errors are in parentheses.

Do the results above indicate that Y_t and X_t are cointegrated? Specify clearly all the assumptions you have made.

[Note: Critical value at 5% significance level from MacKinnon table is -3.34.]

15.4E Let us consider the relationship between the natural logarithm of GDP, GDP_t, and the lagged long-term interest rates $rate_t$ and $rate_{t-1}$.

Assume that GDP_t and $rate_t$ are $I(1)$.

Consider the following regression model, where ε_t has mean zero and is uncorrelated with GDP_{t-1}, GDP_{t-2}, $rate_t$, and $rate_{t-1}$:

$$GDP_t = \delta + \theta_1 GDP_{t-1} + \theta_2 GDP_{t-2} + \phi_1 rate_t + \phi_2 rate_{t-1} + \varepsilon_t.$$

(a) Provide the long-run relationship between GDP_t and $rate_t$.

(b) The model can be written in the following error correction form:

$$\begin{aligned}\Delta GDP_t = {} & \phi_1 \Delta rate_t - \theta_2 \Delta GDP_{t-1} \\ & - (1-\theta_1-\theta_2)\left(GDP_{t-1} - \frac{\delta}{1-\theta_1-\theta_2} - \frac{\phi_1+\phi_2}{1-\theta_1-\theta_2} rate_{t-1}\right) + \varepsilon_t.\end{aligned}$$

Interpret the above equation and discuss how you can estimate the ECM.

15.5.2 Solutions to additional examination based exercises

15.1E (a) We can rewrite this process in the usual AR(1) form:

$$r_t = \beta_0 + (1 + \beta_1)r_{t-1} + \varepsilon_i.$$

The condition for this AR(1) process to be stationary and weakly dependent is $\beta_1 < 0$ (and $\beta_1 > -2$, but this may be ignored) so that the coefficient is smaller than 1.

We can obtain consistent estimates using OLS as ε_t is uncorrelated with r_{t-1}, but the OLS estimator will be biased as we use a lagged dependent variable which means that strict exogeneity (required for unbiasedness) cannot be satisfied.

(b) i. We need to show this result using recursive substitution.

Using the linearity of the expectation operator and using $E(\varepsilon_t) = 0$, we obtain $E(r_t) = t\beta_0$, the mean is trending. Using the i.i.d. assumption on ε_t, we can show the variance is changing over time too: $Var(r_t) = t\sigma^2$. Using independence we can also show $Cov(r_t, r_{t+h}) = h\sigma^2$, and:

$$Corr(r_t, r_{t+h}) \equiv \frac{Cov(r_t, r_{t+h})}{\sqrt{Var(r_t)\ Var(r_{t+h})}} = \frac{h\sigma^2}{\sqrt{(t\sigma^2)(t+h)\sigma^2}} = \sqrt{\frac{h}{t+h}}$$

where the latter reveals the strong dependence (as $Corr(r_t, r_{t+h})$ stays close to 1 as h increases for large t).

ii. We need to apply the DF test. We should test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 < 0$, for which we use the usual test statistic:

$$\text{DF test} = \frac{\widehat{\beta}_1}{se(\widehat{\beta}_1)}.$$

Under the null (nonstationarity) the DF statistic has a non-standard distribution, also called the Dickey–Fuller distribution. For a given level of significance we should reject the null if the realisation of our DF test statistic is lower than the critical value (setting where we include an intercept but not a time trend).

15.2E (a) Processes that can be made stationary and weakly dependent by detrending are called trend stationary processes, they exhibit **deterministic trends**.

Processes that can be made stationary and weakly dependent by differencing are called difference stationary processes. The random walk and random walk with drift are examples. As we can write these processes as sums of i.i.d. shocks, we also say that these processes have **stochastic trends**.

We call a process covariance stationary if the mean and variance of the process do not change over time, and the covariance is only a function of distance in time not location in time.

(b) We can write the detrended process as:

$$y_t - \beta t = \alpha + u_t$$

where u_t is a stationary AR(1) process if $\rho < 1$; $\alpha + u_t$ is then also a stationary AR(1) process (adding a constant does not affect the mean, variance or covariance of a process). As stationary AR(1) processes are weakly dependent this proves that the detrended process is stationary and weakly dependent as required for trend stationarity. (It is fine to call $y_t - \alpha - \beta t$ a detrended process.)

(c) To obtain the differenced process, we need to subtract y_{t-1} from y_t, yielding:

$$\Delta y_t = (\alpha + \beta t + u_t) - (\alpha + \beta(t-1) + u_{t-1}) = \beta + \Delta u_t.$$

If $\rho = 1$, $\Delta u_t = \varepsilon_t$ which is an i.i.d. process. As i.i.d. processes are stationary and weakly dependent this shows that u_t is difference stationary (u_t itself is nonstationary as it is an AR(1) process with a unit root).

Adding β to an i.i.d. process, i.e. $\beta + \Delta u_t$, is still i.i.d. As this means that Δy_t is stationary and weakly dependent we have proven that y_t is difference stationary.

(d) Difference stationary processes are highly persistent (strong dependence), whereas trend stationary processes are weakly dependent.

In highly persistent processes, shocks or policy changes will have a lasting/permanent effect (strong dependence), under weak dependence their impact is only transitory.

When running a regression using two (or more) highly persistent processes we will need to rule out that we have obtained a meaningless, spurious relationship, where the only reason we have the appearance of a relationship is due to their highly persistent behaviour. Inference (large sample) non-standard.

(e) To ensure we get a consistent estimator of α and β we will need to rule out $\rho = 1$; the laws of large numbers used to establish large sample properties such as consistency require that the error is weakly dependent.

The OLS estimator will not be efficient, unless $\rho = 0$. With $0 < \rho < 1$, there will be autocorrelation and the usual standard errors are not valid (assume homoskedasticity and no autocorrelation). We need to use robust (HAC) standard errors for inference.

15.3E (a) To test whether the above relation is spurious or cointegrating, we need to test whether the errors are $I(1)$. We can do that by conducting a DF test on the OLS residuals (should provide details).

A cointegrating relationship indicates that there is a long-run relation between the two processes. If the only reason that there is an apparent relationship is due to the persistence of the two processes we have a spurious, meaningless relationship.

We can only test the hypothesis in settings where we can reject the null hypothesis that the above relation is spurious.

(b) Let us denote the parameter on $\widehat{\varepsilon}_{t-1}$ as δ. We want to test here the hypothesis $H_0 : \delta = 0$ against $H_1 : \delta < 0$. Our test is the Augmented Dickey–Fuller test (as we have the included lagged difference $\Delta\widehat{\varepsilon}_{t-1}$). The ADF test statistic is:

$$\text{ADF-test} = \frac{\widehat{\delta}}{se(\widehat{\delta})}.$$

The realisation of this test statistic $-.216/.0845 = -2.56$. As this realisation is not smaller than the 5% critical value, we cannot reject the null of the relationship being spurious. We do not find evidence for the relationship being cointegrating at the 5% level.

15.4E (a) In equilibrium (GDP^*, $rate^*$), using $E(\varepsilon_t) = 0$, we have:

$$GDP^* = \delta + \theta_1 GDP^* + \theta_2 GDP^* + \phi_1 rate^* + \phi_2 rate^*$$

$$(1 - \theta_1 - \theta_2) GDP^* = \delta + (\phi_1 + \phi_2) rate^*.$$

Rewriting yields, the long-run relationship:

$$GDP^* = \frac{\delta}{1 - \theta_1 - \theta_2} + \frac{\phi_1 + \phi_2}{1 - \theta_1 - \theta_2}\, rate^*.$$

(b) This ECM describes how GDP_t changes as a result of changes in $rate_t$, GDP_{t-1}, and the existence of a disequilibrium at time $t - 1$ (when $\Delta GDP_t = \Delta rate_t = 0$):

$$GDP_{t-1} - \frac{\delta}{1 - \theta_1 - \theta_2} - \frac{(\phi_2 + \phi_1)}{1 - \theta_1 - \theta_2}\, rate_{t-1}.$$

The parameter $(\theta_1 + \theta_2 - 1)$ indicates how quickly the variable GDP_t will respond to the long-run disequilibrium: if GDP_{t-1} is larger than the equilibrium value, then $\Delta GDP_t < 0$ to bring GDP_t back to equilibrium.

The parameters $-\theta_2$ and ϕ_1 are the direct effects of a change in the past GDP level and the current interest rate has on the the current GDP level.

We can estimate the ECM using the Engle–Granger two-step procedure.

- **Step 1**: Estimate the cointegrating relationship by OLS:

$$GDP_t = \beta_0 - \beta_1 rate_t + e_t$$

  and compute $diseq_{t-1} = GDP_{t-1} - \widehat{\beta}_0 - \widehat{\beta}_1 rate_{t-1}$.
- **Step 2**: Estimate by OLS:

$$\Delta GDP_t = \phi_1 \Delta rate_t - \theta_2 \Delta GDP_{t-1} - \theta_3 diseq_{t-1} + v_t.$$

We can derive the ECM by subtracting lagged terms:

$$GDP_t = \delta + \theta_1 GDP_{t-1} + \theta_2 GDP_{t-2} + \phi_1 rate_t + \phi_2 rate_{t-1} + \varepsilon_t$$

$$GDP_t - GDP_{t-1} = \delta + (\theta_1 - 1) GDP_{t-1} + \theta_2 GDP_{t-2} + \phi_1 rate_t + \phi_2 rate_{t-1} + \varepsilon_t$$

$$GDP_t - GDP_{t-1} = \delta + (\theta_1 + \theta_2 - 1) GDP_{t-1} + \theta_2 (GDP_{t-2} - GDP_{t-1}) +$$
$$+ \phi_1 (rate_t - rate_{t-1}) + (\phi_2 + \phi_1) rate_{t-1} + \varepsilon_t.$$

Simplifying the above expression, we have the following ECM model:

$$\begin{aligned}\Delta GDP_t &= -\theta_2 \Delta GDP_{t-1} + \phi_1 \Delta rate_t \\ &\quad + (\theta_1 + \theta_2 - 1)\left(GDP_{t-1} - \frac{\delta}{1-\theta_1-\theta_2} - \frac{(\phi_2+\phi_1)}{1-\theta_1-\theta_2} rate_{t-1} \right) + \varepsilon_t.\end{aligned}$$