Econometrics I

Ryan T. Godwin

Contents

Chapter 1

Introduction

1.1 This book

This book serves as the course notes for the introductory Econometrics graduate level course Econ 7010, at the University of Manitoba. These notes are heavily influenced by those of David Giles. Econometric Analysis by Greene, and Econometric Theory and Methods by Davidson and MacKinnon, are used as companions to these notes.

This book tries to cover the standard topics covered in introductory graduate level courses, at least in the U15 in Canada. The primary focus is on theory, with as much application as time allows. The algebraic and statistical properties of least squares are first established under standard assumptions. The

1.2 Objectives of this course

The primary objective is to cover the theory necessary to assess the merits of an econometric method given a model and data. We begin be assessing the basic and popular least-squares method. Then, by examining how the desirable properties of the least-squares method relies on various assumptions we are led to more complicated and refined models. This serves as an exercise in econometric modelling.

- Learn mathematical tools necessary to examine and discuss econometric methods.
- Study some classic topics in econometrics such as heteroskedasticity and instrumental variables.
- Learn how to estimate, interpret, and test.
- Have basic proficiency with R.

1.3 R Statistical Environment and R Studio

The theory and concepts presented in this course will be illustrated by analysing several data sets. Data analysis will be accomplished through the R Statistical Environment and RStudio. Both are free, and R is fast becoming the best and most widely used statistical software.

Chapter 2

The R Programming Language

2.1 What is R?

R is a programming language designed to analyse data. R is free and open-source, with many user contributed "add-on" packages that are readily downloadable by anyone. R is found in all areas of academia that encounter data, and in many private and public organizations.

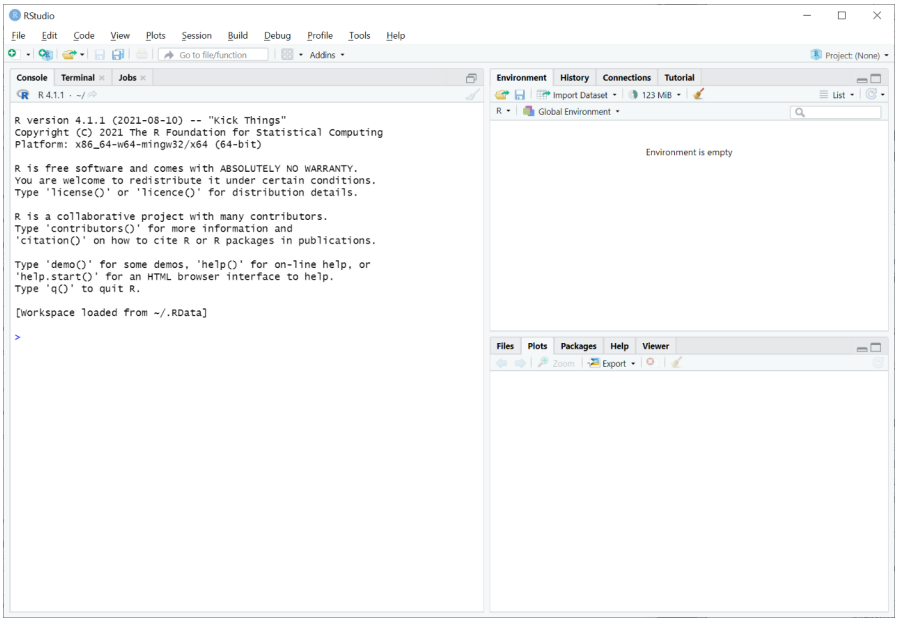
2.2 Where to get R

In this course we will use R and RStudio. Download and install R first: https://cran.r-project.org/bin/windows/base/ (for Windows) or https://cran.r-project.org/bin/macosx/ (for Mac). Then, download and install RStudio from https://www.rstudio.com/products/rstudio/download/.

2.3 Getting started with RStudio

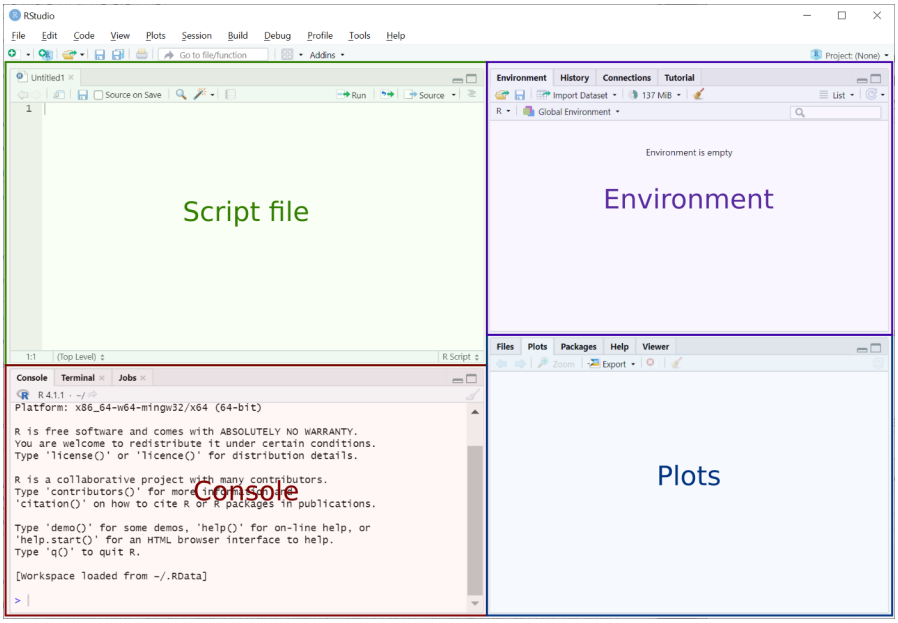
2.3.1 Open RStudio

After you open RStudio it should look something like this:



2.3.2 Create a "script" file

A script file is a file where you can type and save your R computer code. To open a script file, click on "File", "New File", "R Script".



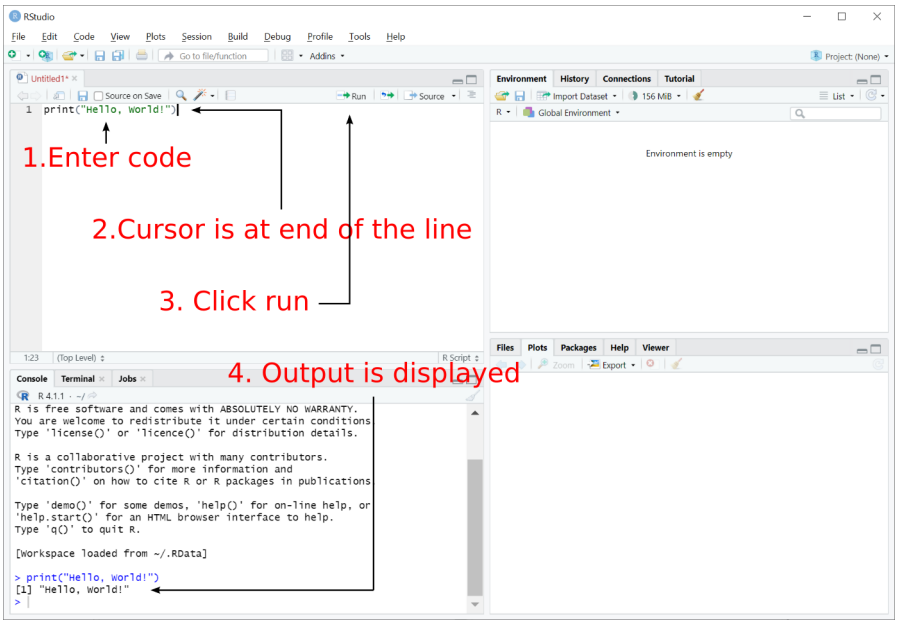
- In the top left is your Script file. R commands can be run from the R Script file, and saved at any time.
- In the bottom left is the Console window. Output is displayed here. R commands can be run from the Console, but not saved.
- In the top right is the Environment. Data and variables will be visible here.
- The bottom right will display graphics (e.g. histograms and scatterplots).

2.3.3 Running R code

Copy and paste the following R code into the script window:

```
print("Hello, World!")
```

Run the code by highlighting it, or making sure the cursor is active at the end of the line, and clicking "Run" (you can also press `Ctrl + Enter` on PC or `Cmd + Return` on Mac).



The output from the program is reproduced in the box below:

```
[1] "Hello, World!"
```

2.4 Arithmetic in R

R's arithmetic operators include:

Operator	Function
+	addition
-	subtraction
*	multiplication
/	division
^	exponentiation

Example 2.1 — Arithmetic in R. Use R to perform the following arithmetic operations:

1. 2×13

```
2 * 13
```

```
[1] 26
```

2. $16/4$

```
16 / 4
```

```
[1] 4
```

3. 2^8

```
2 ^ 8
```

```
[1] 256
```

4. $\frac{10+6}{2}$

```
(10 + 6) / 2
```

```
[1] 8
```

2.5 Create an object

You can create objects in R. Objects can be vectors, matrices, character strings, data frames, scalars etc. Create two different scalars. Give them any name you like, but object names cannot start with a number and cannot include certain characters like "`!`":

```
a <- 3
b <- 5
```

We have created two new objects called `a` and `b`, and have assigned them values using the assignment operator `<-` (the "less than" symbol followed by the "minus" symbol). Notice that `a` and `b` pop up in the top-right of your screen (the Environment window). We can now refer to these objects by name:

```
a * b
```

```
[1] 15
```

produces the output `15`. To create a vector in R we use the "combine" function, `c()`:

```
myvector <- c(1, 2, 4, 6, 7)
```

Notice that after creating it, the `myvector` object appears in the top-right Environment window. `myvector` is just a list of numbers:

$$\texttt{myvector} = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 6 \\ 7 \end{bmatrix}$$

2.6 Simple functions in R

Table 2.1: Simple R functions.

Function
`sum()`
`mean()`
`var()`
`lm()`
`summary()`

An R function takes an input, performs an operation, and then provides an output. Type the name of the function and then type the input inside of parentheses: `function.name(input)`. After we click the "Run" button, we get the output. There are thousands of functions in R, a few simple ones are in Table 2.1.

For example, to add up all of the numbers in `myvector` we would run:

```
sum(myvector)
```

```
[1] 20
```

which provides the output `20`. We have asked the computer to add up an object by calling the function `sum()`, and putting the name of the object `myvector` inside of the parentheses.

2.7 Logical operators

Logical operators are used to determine whether something is `TRUE` or `FALSE`. Some logical operators are:

Operator	Function
>	greater than
==	equal to
<	less than
>=	greater than or equal to
<=	less than or equal to
!=	not equal to

Logical operators are useful for creating "subsamples" or "subsets" from our data. Using logical operators, we can calculate statistics separately for ethnicities, treatment group vs. control group, developed vs. developing countries, etc. (we will see how to do this later). For now, let's try some simple logical operations. Try entering and running each of the following lines of code one by one:

```
8 > 4
```

```
[1] TRUE
```

```
b == 6
```

```
[1] FALSE
```

To check to see which elements in `myvector` are greater than 3 we use:

```
myvector > 3
```

```
[1] FALSE FALSE  TRUE  TRUE  TRUE
```

2.7.1 Multiple logical operators

Sometimes we would like to create subsets in our data based on multiple conditions or characteristics. For example, we might want to study a subset of our data consisting of only single or widowed women with 1 child or more. The "and" / "or" operators are useful in these situations:

Operator	Function
&	"and"
|	"or"

For example, the following line of code:

```
myvector > 3 & myvector < 7
```

```
[1] FALSE FALSE  TRUE  TRUE FALSE
```

checks to see whether each element in `myvector` is greater than 3 *and* less than 7.

2.8 Loading data into R

There are several ways to load data into R. In this course we work with *comma-separated values* file format (CSV format).

2.8.1 Directly from the internet

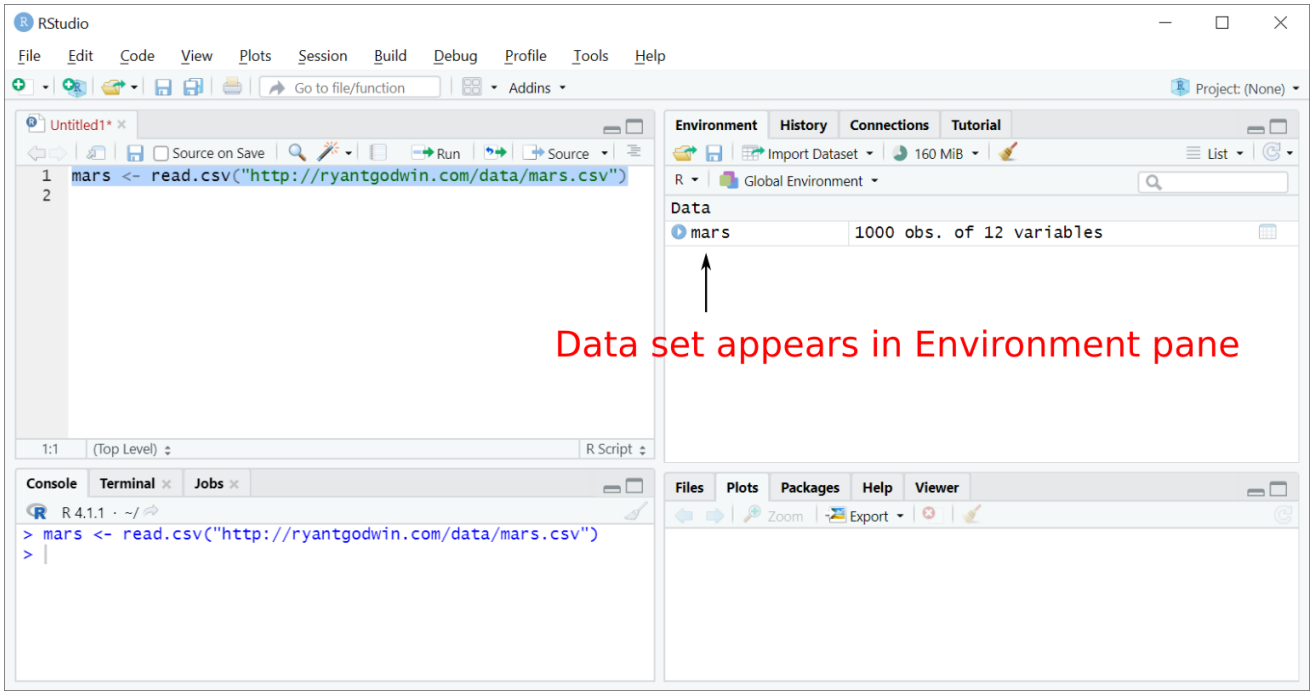
We can use the R code:

```
mydata <- read.csv("file location.csv")
```

We need to replace `file location` with the actual location of the file, either on the internet or on your computer. We can also replace the name of the data set `mydata` with any name we like. For example, to load data directly from the internet into R, try the following:

```
mars <- read.csv("http://ryantgodwin.com/data/mars.csv")
```

After running the above line of code, you should see the data set appear in the top-right of RStudio (the environment pane).



2.8.2 From a location on your computer

After saving a `.csv` file to your computer, you can use the `read.csv()` command to load the file from its location on your computer. For example:

```
mars <- read.csv("c:/data/mars.csv")
```

loads a file from the location `c:/data/`.

2.8.3 `file.choose()`

Using the `file.choose()` command will prompt you to select the file using file explorer:

```
mars <- read.csv(file.choose())
```

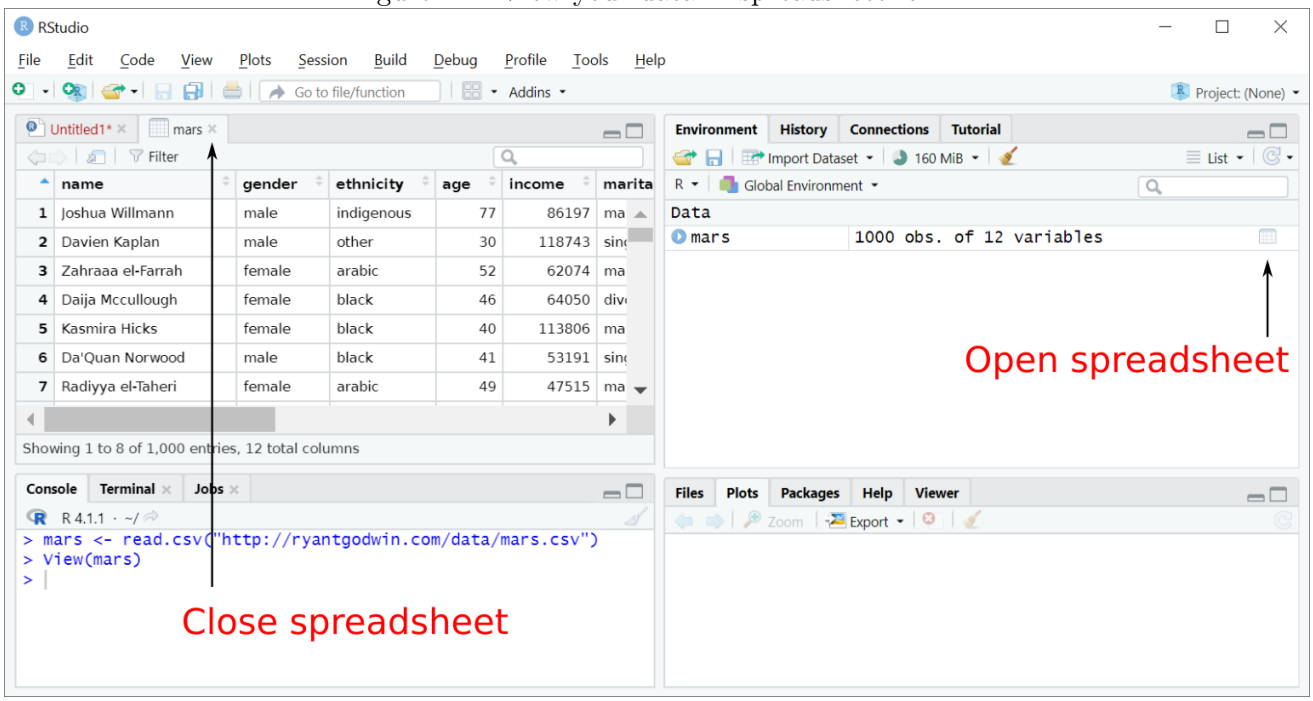
2.9 View your data in spreadsheet form

Click on the spreadsheet icon next to your `mars` data set, or run the following command:

```
View(mars)
```

Note the uppercase `V` (R is *case sensitive*). This command allows you to view your data in spreadsheet form. See Figure 2.1.

Figure 2.1: View your data in spreadsheet form.



Chapter 3

Basic Multiple Regression

A general population model is:

$$\boldsymbol{y} = f(x_1, x_2, \ldots, x_k; \boldsymbol{\theta}) + \boldsymbol{\epsilon} \tag{3.1}$$

- $\boldsymbol{y}$ is the dependent variable or "regressand".
- $x_1, x_2, \ldots, x_k$ are the explanatory variables or "regressors".
- $\boldsymbol{\theta}$ is a parameter vector.
- $\boldsymbol{\epsilon}$ is the disturbance term or the random "error".

Equations such as 3.1, although quite vague at this stage, are sometimes called the *data generating process*. There is some mathematical process behind the scenes that is governing or creating the data that we observe. A population model or data generating process[1] mathematically states how the variables are linked to one another in the physical world.

We'll focus on population models where f is parametric and (usually) linear in the parameters. The first estimation strategy that we'll consider, *Ordinary Least Squares*, requires that the model be linear in the parameters. In general, however, f may be:

- linear or non-linear in the variables
- linear or non-linear in the parameters
- parametric or non-parametric

Questions:

1. What is the role of the error term?
2. What is random, and what is deterministic?
3. What is observable, and what is unobservable?

3.1 Some classic population models in economics

These models are provided as examples of the general population model in equation 3.1. For each model, try to determine the components of the model, and whether or not it is linear/non-linear in the regressors/parameters. The background of these models are not of great importance to the course - they are merely examples.

Keynes' consumption function

$$C = \beta_1 + \beta_2 Y + \epsilon$$

[1] The terms "population model" and "data generating process" are mostly interchangeable.

Cobb-Douglas production function

$$Y = AK^{\beta_2}L^{\beta_3}e^{\epsilon}$$

By taking logs, the Cobb-Douglas production function can be rewritten as:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + \epsilon$$

where $\beta_1 = \log A$.

Gravity model of trade

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \epsilon_{ij}$$

Mincer earnings equation

$$\ln w = \ln w_0 + \rho s + \beta_1 x + \beta_2 x^2 + \epsilon$$

CES production function

$$Y = \varphi \left(aK^r + (1-a)L^r\right)^{1/r} e^{\epsilon}$$

Taking logs, the CES production function is written as:

$$\log Y = \log \varphi + \frac{1}{r} \log \left(aK^r + (1-a)L^r\right) + \epsilon$$

3.2 Sample information

Suppose that we have a *sample* of n observations:

$$\{y_i; x_{i1}, x_{i2}, \ldots, x_{ik}\}; \quad i = 1, 2, \ldots, n$$

i denotes an observation (individual, country, firm, etc.), and k is the number of *variables* in the model.

Assuming that the observed values are generated by the population model, and taking the case where the model is *linear in the parameters*, Equation 3.1 becomes:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \quad ; \quad i = 1, \ldots, n \tag{3.2}$$

Recall that the βs and ϵ are unobservable. So, y_i is generated by two components:

1. Deterministic component: $\Sigma_{j=1}^{k} \beta_j x_{ij}$
2. Stochastic component: ϵ_i

To begin with, we will make the unrealistic assumption that the x_{ij} are *non-random*. Regardless of this assumption, y_i is a linear function of a random variable ϵ and so is itself random. The y_i data that we observe are "realized values" of a random variable.

Some typical objectives are to:

(i) Estimate unknown parameters of Equation 3.2.
(ii) Test hypotheses about the parameters.
(iii) Predict values of y outside of the sample.

3.3 Interpreting the parameters in a model

Once we estimate $\boldsymbol{\theta}$ (e.g. all the βs), how do we interpret them? A major advantage of the linear model is the ease in which the parameters may be interpreted. That is, the βs in equation 3.2 have an important economics interpretation. For example:

$$\frac{\partial y_i}{\partial x_{i1}} = \beta_1$$

The parameters are the marginal effects of the x on y, with other factors held constant (ceteris paribus). For example, from Keynes' consumption function:

$$\frac{\partial C}{\partial Y} = \beta_2 = \text{Marginal Propensity to Consume}$$

We might wish to test the hypothesis that $\beta_2 = 0.9$, for example.

Depending on how the population model is specified, however, the β might not be interpreted as marginal effects. For example, after taking logs of the Cobb-Douglas production function in, we get the following population model:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + \epsilon,$$

and

$$\beta_2 = \frac{\partial \log Y}{\partial \log K} = \frac{\partial \log Y}{\partial Y} \times \frac{\partial Y}{\partial K} \times \frac{\partial K}{\partial \log K} = \frac{1}{Y} \times \frac{\partial Y}{\partial K} \times K = \frac{\partial Y/Y}{\partial K/K},$$

so that β_2 is the elasticity of output with respect to capital. The point is that we need to be careful about how the parameters of the model are interpreted in all but the most simple of cases.

Questions:

1. How are the parameters interpreted in a *log-linear* model? For example:

$$\log(y_i) = \beta_0 + \beta_1 x_{i1} + ... + \epsilon_i$$

2. How are the parameters interpreted in a *linear-log* model? For example:

$$y_i = \beta_0 + \beta_1 \log(x_{i1}) + ... + \epsilon_i$$

So, we have a stochastic model that might be useful as a starting point to represent economics relationships. We need to be especially careful about the way in which we specify both parts of the model (the deterministic and stochastic parts).

3.4 Assumptions of the Classical Linear Regression Model

In this section, we are going to state six "classical" assumptions, and refer back to them frequently throughout the course. These simplifying assumptions are a starting point, and are likely not satisfied in real data. One of the main objectives of this course is to re-consider these assumptions - are they realistic; can they be tested; what if they are wrong; can they be "relaxed"? When these assumptions are violated, and we consider how to fix the resulting consequence, we will be led to different *estimation strategies*, such as *instrumental variables* estimation or *generalized least squares*.

All "models" are simplifications of reality. Presumably we want our econometric model to be simple but "realistic" – at least in the sense that we can *identify* our objective.

Traditionally the objective of most econometric models was to describe an economic process. Assumptions, such as the ones to follow, were to ensure the "quality" of the estimated economic model (we

will soon measure quality in terms of *unbiasedness*, *efficiency*, and *consistency*). More recently, applied econometrics has been focused on obtaining *causal inference* from *observational data*. The emphasis is usually on estimating the marginal effect of just one of the x variables on y. From this perspective, the following assumptions can in part be viewed as necessary for estimating a causal relationship between x and y.

A.1: Linearity

The model is linear in the parameters:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i$$

Linearity in the parameters allows the model to be written in matrix notation. Let,

$$\underset{(n\times 1)}{\boldsymbol{y}} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}; \quad \underset{(k\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}; \quad \underset{(n\times k)}{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1k} \\ x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix};$$

$$\underset{(n\times 1)}{\boldsymbol{\epsilon}} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Then, we can write the model, for the full sample, as:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.3}$$

If we take the i^{th} row (observation) of this model we have:

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i \qquad \text{(scalar)}$$

Notational points

- Vectors are in bold.
- The dimensions of vectors/matrices are written (rows × columns).
- The first subscript denotes the row, the second subscript the column.
- Some texts (including Greene, 2011), use the convention that vectors are columns. Hence, when an observation (row) is extracted from the X matrix, it is transformed into a column. The above equation would then be expressed as $y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i$.

Question: Which of the equations in Section 3.1 are linear in the parameters? Which of these equations can be "linearized"?

A.2: Full Rank

We assume that there are no exact linear dependencies among the columns of X (if there were, then one or more regressor is redundant). Note that X is $(n \times k)$ and $\text{rank}(X) = k$. So we are also implicitly assuming that $n > k$, since $\text{rank}(A) \leq \min\{\#rows, \#cols\}$. What does this assumption really mean? Suppose we had:

$$y_i = \beta_1 x_{i1} + \beta_2 (2x_{i1}) + \epsilon_i$$

We can only identify, and estimate, the one function, $(\beta_1 + 2\beta_2)$. In this model, $\text{rank}(X) = k - 1 = 1$. An example which is commonly found in undergraduate textbooks, of where A.2 is violated, is the *dummy variable trap*.

A.3: Errors have a zero mean

Assume that, *in the population*, $\mathbb{E}(\epsilon_i) = 0 \quad ; \quad i = 1, 2, \ldots, n$. So,

$$\mathbb{E}(\boldsymbol{\epsilon}) = \mathbb{E}\begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix} = \mathbf{0}$$

This is arguably the least important assumption.

A.4: Spherical errors

Assume that, in the population, the disturbances are generated by a process whose variance is constant (σ^2), and that these disturbances are uncorrelated with each other:

$$\text{var}(\epsilon_i) = \sigma^2; i = 1, 2, \ldots, n \quad \text{(Homoskedasticity)}$$
$$\text{cov}(\epsilon_i, \epsilon_j) = 0; \forall i \neq j \quad \text{(no Autocorrelation)}$$

Putting these assumptions together we can determine the form of the "covariance matrix" for the random vector, $\boldsymbol{\epsilon}$.

$$\begin{aligned}\mathbb{V}(\epsilon) &= \mathbb{E}\left[(\boldsymbol{\epsilon} - \mathbb{E}(\boldsymbol{\epsilon}))(\boldsymbol{\epsilon} - \mathbb{E}(\boldsymbol{\epsilon}))'\right] \\ &= \mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] \\ &= \begin{bmatrix} \mathbb{E}(\epsilon_1\epsilon_1) & \ldots & \mathbb{E}(\epsilon_1\epsilon_n) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(\epsilon_n\epsilon_1) & \ldots & \mathbb{E}(\epsilon_n\epsilon_n) \end{bmatrix}\end{aligned}$$

but...

$$\mathbb{E}(\epsilon_i\epsilon_i) = \mathbb{E}\left(\epsilon_i^2\right) = \mathbb{E}\left[(\epsilon_i - 0)^2\right] = \text{var}(\epsilon_i) = \sigma^2$$

and

$$\mathbb{E}(\epsilon_i\epsilon_j) = \mathbb{E}\left[(\epsilon_i - 0)(\epsilon_j - 0)\right] = \text{cov}(\epsilon_i, \epsilon_j) = 0.$$

So:

$$\mathbb{V}(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

a scalar matrix.

A.5: Generating process for X

The classical regression model assumes that the regressors are "fixed in repeated samples" (laboratory situation). We can assume this, but it is a very strong and unrealistic assumption for most economics data.

Alternatively, allow X to be random, but restrict the form of their randomness: the process that generates X must be unrelated to the process that generates $\boldsymbol{\epsilon}$ in the population.

This is arguably the most important assumption. We will soon see that it is imperative that X and $\boldsymbol{\epsilon}$ are statistically independent. So, if X is random, we need to assume *strict exogeneity*:

$$\mathbb{E}(\boldsymbol{\epsilon}|X) = \mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0} \tag{3.4}$$

This says that $\boldsymbol{\epsilon}$ and X are *statistically independent*. Statistical independence implies that X and $\boldsymbol{\epsilon}$ are uncorrelated (have zero covariance):

$$\text{cov}(\boldsymbol{x_j}, \boldsymbol{\epsilon}) = \mathbf{0} \quad ; \quad \text{for } j = 1, \ldots, k$$

Note that statistical independence implies zero correlation, but not necessarily the other way around.

Prove that if X and $\boldsymbol{\epsilon}$ are uncorrelated, and if $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$, then:

$$\mathbb{E}(X'\boldsymbol{\epsilon}) = \mathbf{0} \tag{3.5}$$

A.6: Normality of errors

$$(\boldsymbol{\epsilon}|X) \sim N\left[\mathbf{0}, \sigma^2 I_n\right]$$

This assumption is not as strong as it seems:

- often reasonable due to the Central Limit Theorem (C.L.T.)
- often not needed
- when some distributional assumption is needed, often a more general one is ok

Summary

The classical linear regression model is:

- $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$
- $(\boldsymbol{\epsilon}|X) \sim N\left[\mathbf{0}, \sigma^2 I_n\right]$
- $\text{rank}(X) = k$
- The data generating process (DGP) of X and $\boldsymbol{\epsilon}$ are unrelated

Implications for y (if X is non-random; or conditional on X):

$$\mathbb{E}(\boldsymbol{y}) = X\boldsymbol{\beta} + \mathbb{E}(\boldsymbol{\epsilon}) = X\boldsymbol{\beta}$$
$$\mathbb{V}(\boldsymbol{y}) = \mathbb{V}(\boldsymbol{\epsilon}) = \sigma^2 I_n$$

Because linear transformations of a Normal random variable are themselves Normal, we also have:

$$\boldsymbol{y} \sim N\left[X\boldsymbol{\beta}, \sigma^2 I_n\right]$$

In subsequent chapters, we will return to these assumptions, and ask:

- How reasonable are the assumptions associated with the classical linear regression model?
- How do these assumptions affect the estimation of the model's parameters?
- How do these assumptions affect the way we test hypotheses about the model's parameters?
- Which of these assumptions are used to establish the various results we'll be concerned with?
- Which assumptions can be "relaxed" without affecting these results?

3.5 Least Squares Estimator

Our first task is to estimate the parameters of our population model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad ; \quad \boldsymbol{\epsilon} \sim N\left[\mathbf{0}, \sigma^2 I_n\right]$$

Note that there are $(k+1)$ parameters, including σ^2.

- There are many possible procedures for estimating parameters.
- Choice should be based mostly on the "sampling properties" of the resulting estimator (to be considered later). Other considerations include computational convenience, and the ease of interpretation of the estimated model.
- To begin with, we consider one possible estimation strategy – Least Squares.

For the i^{th} data-point, we have:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i.$$

Given the assumptions in Section 3.4 regarding the error term, the *expected* value of y_i conditional on $\mathbf{x}_i$:

$$\mathbb{E}\left(y_i|\boldsymbol{x}_i'\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

We'll estimate $\mathbb{E}\left(y_i|\boldsymbol{x}_i'\right)$ by

$$\hat{y}_i = \boldsymbol{x}_i'\boldsymbol{b},$$

where $\hat{y}$ is a least squares "predicted" or "fitted" value, and $\boldsymbol{b}$ is the vector of least squares estimates for $\boldsymbol{\beta}$. *In the population*, the true (unobserved) error disturbance is

$$\epsilon_i = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}.$$

When we use $\boldsymbol{b}$ to estimate $\boldsymbol{\beta}$, there will be some "estimation error", and the value,

$$e_i = y_i - \boldsymbol{x}_i'\boldsymbol{b}$$

will be called the i^{th} "residual". So,

$$y_i = \left(\boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i\right) = \left(\boldsymbol{x}_i'\boldsymbol{b} + e_i\right) = \left(\hat{y}_i + e_i\right)$$

Question: Which terms are unobserved (from the population) and which are observed (determined by the sample)?

After the $\boldsymbol{\beta}$ have been estimated, we can write the $\boldsymbol{y}$ values as a sum of two components (the above equation in matrix form):

$$\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$$

This says that the actual $\boldsymbol{y}$ values can be written as the sum of fitted (estimated or predicted) values $\widehat{\boldsymbol{y}}$, and residuals (prediction error) $\boldsymbol{e}$.

3.5.1 The Least Squares criterion

We will "choose $\boldsymbol{b}$ so as to minimize the sum of squared residuals". That is, $\boldsymbol{b}$ is defined as from the minimization problem:

$$\text{Min}_{(\boldsymbol{b})} \sum_{i=1}^{n} e_i^2$$

Before we proceed, let's consider some questions:

1. Why *squared* residuals? If we did not square the residuals, *positive* "distances" could cancel out *negative* ones.
2. Why not *absolute values* of residuals? This is valid, but leads to the *Least Absolute Deviations* (LAD) estimator.
3. Why not use a "minimum distance" criterion? Why vertical instead of horizontal? The choice of criterion seems quite arbitrary at this point, and is justified later in terms of the *statistical properties* of the resulting estimator.

3.5.2 Minimizing the sum of squared residuals: an optimization problem

We will solve a minimization problem using the least squares criterion in order to derive the "Least Squares" estimator. The problem that we are trying to solve can be stated as:

$$\begin{aligned}\text{Min}_{(\boldsymbol{b})} \sum_{i=1}^{n} e_i^2 &\Leftrightarrow \text{Min}_{(\boldsymbol{b})} \left(\boldsymbol{e}'\boldsymbol{e}\right) \\ &\Leftrightarrow \text{Min}_{(\boldsymbol{b})} \left[(\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b})\right]\end{aligned}$$

Now, let:

$$S = (\boldsymbol{y} - X\boldsymbol{b})'(\boldsymbol{y} - X\boldsymbol{b}) = \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{b}'X'\boldsymbol{y} - \boldsymbol{y}'X\boldsymbol{b} + \boldsymbol{b}'X'X\boldsymbol{b}$$

Note that:

$$\begin{aligned}\boldsymbol{b}'X'\boldsymbol{y} &= \boldsymbol{y}'X\boldsymbol{b} \\ (1 \times k)(k \times n)(n \times 1) &= (1 \times 1)\end{aligned}$$

So,

$$S = \boldsymbol{y}'\boldsymbol{y} - 2\left(\boldsymbol{y}'X\right)\boldsymbol{b} + \boldsymbol{b}'\left(X'X\right)\boldsymbol{b}$$

Two rules involving the differentiation of vectors and matrices that we need are:

(i) $\partial\left(\boldsymbol{a}'\boldsymbol{x}\right)/\partial\boldsymbol{x} = \boldsymbol{a}$
(ii) $\partial\left(\boldsymbol{x}'A\boldsymbol{x}\right)/\partial\boldsymbol{x} = 2A\boldsymbol{x}$; if A is symmetric

Applying these two results:

$$\partial S/\partial\boldsymbol{b} = \boldsymbol{0} - 2\left(\boldsymbol{y}'X\right)' + 2\left(X'X\right)\boldsymbol{b} = 2\left[X'X\boldsymbol{b} - X'\boldsymbol{y}\right]$$

Set this to zero (for a *turning point*):

$$\begin{aligned}X'X\boldsymbol{b} &= X'\boldsymbol{y} \\ (k \times n)(n \times k)(k \times 1) &= (k \times n)(n \times 1)\end{aligned}$$

This gives us k equations in k unknowns, sometimes called the "normal equations". Finally, provided that $(X'X)^{-1}$ exists:

$$\boldsymbol{b} = \left(X'X\right)^{-1}X'\boldsymbol{y} \tag{3.6}$$

Notice that $X'X$ is $(k \times k)$, and $\text{rank}(X'X) = \text{rank}(X) = k$ (by assumption). This implies that $(X'X)^{-1}$ exists. We need the "full rank" assumption for the Least Squares estimator, $\boldsymbol{b}$, to *exist*. None of our other assumptions have been used so far.

Check: have we *minimized* S?

$$\left(\frac{\partial^2 S}{\partial\boldsymbol{b}\partial\boldsymbol{b}'}\right) = \frac{\partial}{\partial\boldsymbol{b}'}\left[2X'X\boldsymbol{b} - 2X'\boldsymbol{y}\right] = 2\left(X'X\right) \quad ; \quad \text{a } (k \times k) \text{ matrix}$$

Note that $X'X$ is at least positive *semi-definite*:

$$\boldsymbol{\eta}'\left(X'X\right)\boldsymbol{\eta} = (X\boldsymbol{\eta})'(X\boldsymbol{\eta}) = (\boldsymbol{u}'\boldsymbol{u}) = \sum_{i=1}^{n} u_i^2 \geq 0$$

and so if $X'X$ has full rank, it will be *positive-definite*, not negative-definite.

So, our assumption that X has full rank has two implications:

(i) The Least Squares estimator, $\boldsymbol{b}$, *exists*.
(ii) Our optimization problem leads to the *minimization* of S, not its maximization!

3.5.3 Least Squares estimator in scalar form

For a population model with an intercept and a single regressor, you may have seen the following formulas used in undergraduate textbooks:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{s_{x,y}}{s_x^2}$$
$$b_0 = \overline{y} - b_1\overline{x}, \tag{3.7}$$

where $s_{x,y}$ is the sample covariance between x and y, and s_x^2 is the sample variance of x.

Questions: Why do population models typically include an intercept? How is the intercept included in the population model when in matrix form (3.3)?

3.6 Method of Moments

The least squares criterion may seem dubious and unmotivated. We have yet to see the benefits of using an estimator that minimizes the sum of squared residuals. Rather than starting from this seemingly arbitrary criterion, we can instead derive the least squared estimator using the *Method of Moments* (MM).

The Method of Moments relies on the principle that the *sample mean* is a good way of estimating a *population mean* (we will see this later in the *law of large numbers*). The MM is widely used in statistics, and many estimators in econometrics can be motivated using it or the closely related *Generalized Method of Moments*.

Take the simple population model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{3.8}$$

and take assumptions A.3 and A.5:

$$\mathbb{E}[\epsilon_i] = 0 \quad ; \quad \mathbb{E}[x_i\epsilon_i] = 0 \tag{3.9}$$

That is, assumptions A.3 and A.5 imply two *moment conditions*, expressed in equation 3.9. By replacing the expectations with sample averages, it can be seen that the MM estimator for the above population model 3.8 is identical to that in equation 3.7.

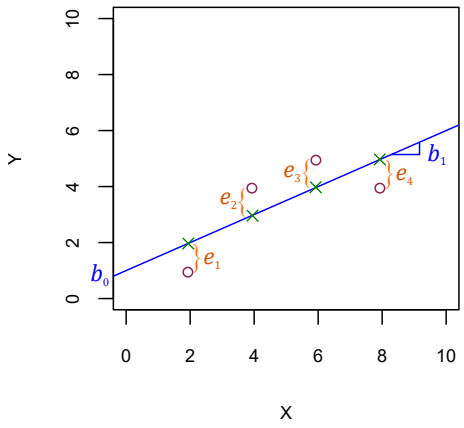
Note that the population model 3.8 above can be generalized to include k regressors; A.3 will provide one moment condition while A.5 will provide the remaining $(k-1)$ moment conditions necessary to solve for the unknown β. The MM estimator, in matrix form, will be identical to equation 3.6.

3.7 Exercises

1. Let the $\boldsymbol{y}$ and X data be:

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 4 \\ 5 \\ 4 \end{bmatrix}; \quad X = \begin{bmatrix} 1 & 2 \\ 1 & 4 \\ 1 & 6 \\ 1 & 8 \end{bmatrix}$$

Figure 3.1: A simple data set with the estimated OLS line in blue. b_0 is the OLS intercept, and b_1 is the OLS slope. The OLS residuals (e_i) are the vertical distances between the actual data points ($\circ$) and the OLS predicted values ($\times$).



(a) Calculate the Least Squares estimators for β_0 and β_1 for the population model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

(b) Calculate the predicted values, and residuals, for the above data and model.
(c) Verify that equation 3.6 and equation 3.7 are identical for the above situation.

The data points, LS estimates, predicted values, and residuals, are shown in Figure 3.1.

2. Suppose the population model is:

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

The $\boldsymbol{y}$ and $\boldsymbol{x}$ variables are:

$$\boldsymbol{y} = \begin{bmatrix} -1 \\ 2 \\ 5 \end{bmatrix}; \quad \boldsymbol{x} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Calculate the OLS estimators for β_1 and β_2.
Answer.
The full X matrix is:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

The $(X'X)^{-1}$ matrix is then:

$$(X'X)^{-1} = \left(\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \right)^{-1} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix}^{-1} = \frac{1}{6} \begin{bmatrix} 14 & -6 \\ -6 & 3 \end{bmatrix} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix}$$

The $(X'\boldsymbol{y})$ matrix is:

$$(X'\boldsymbol{y}) = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 6 \\ 18 \end{bmatrix}$$

Finally, the vector of OLS estimates is:

$$\boldsymbol{b} = \begin{bmatrix} \frac{7}{3} & -1 \\ -1 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 6 \\ 18 \end{bmatrix} = \begin{bmatrix} -4 \\ 3 \end{bmatrix}$$

That is, the estimated intercept is $b_1 = -4$ and the estimated slope is $b_1 = 3$.

3. Given the population model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

derive the OLS estimator for $\boldsymbol{\beta}$. Which assumptions do you need?
Answer.
The OLS estimator is defined as the vector of estimates $\boldsymbol{b}$ which minimizes the sum of squared residuals $\boldsymbol{e}'\boldsymbol{e}$, where $\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e}$.
The optimization problem can be stated as:

$$\min_{\boldsymbol{b}} \boldsymbol{e}'\boldsymbol{e}$$

Substituting $\boldsymbol{y} - X\boldsymbol{b}$ into $\boldsymbol{e}$, we get:

$$\boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - X\boldsymbol{b})' (\boldsymbol{y} - X\boldsymbol{b}) = \boldsymbol{y}'\boldsymbol{y} - 2 (\boldsymbol{y}'X) \boldsymbol{b} + \boldsymbol{b}' (X'X) \boldsymbol{b}$$

Taking the derivative of $\boldsymbol{e}'\boldsymbol{e}$ with respect to the vector $\boldsymbol{b}$, and setting it equal to zero, we get:

$$\frac{\partial \boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{b}} = \boldsymbol{0} - 2 (\boldsymbol{y}'X)' + 2 (X'X) \boldsymbol{b} = 0$$

or

$$X'X\boldsymbol{b} = X'\boldsymbol{y}.$$

Using assumption A.2 (full rank) so that $(X'X)^{-1}$ exists, we can solve for $\boldsymbol{b}$:

$$\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y}$$

The full rank assumption also ensures that we have minimized (not maximized) $\boldsymbol{e}'\boldsymbol{e}$.

4. Suppose that we have the population model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

but where the X matrix contains only a column of 1s. In this case, prove that $b = \bar{y}$.
Answer.
Again, the OLS estimator is:

$$\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y}$$

If the model contains only an intercept, then the X matrix is only a column of 1s. The OLS estimator then becomes:

$$\boldsymbol{b} = \left(\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \frac{1}{n} \sum_{i=1}^{n} y_i = \bar{y}$$

5. Prove that the Method of Moments estimator for the population model in equation 3.8 is identical to the least squared estimator in scalar form (equation 3.7). Hint: use the results that

$$\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})$$

and

$$\sum_{i=1}^{n}\left(x_i^2\right) - n\overline{x}^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

Answer. Using assumptions A.3 and A.5 give the two moment conditions $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[x_i\epsilon_i] = 0$ respectively. The sample counterparts to these two moment conditions are:

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \frac{1}{n}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0 \tag{3.10}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i e_i = \frac{1}{n}\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = 0 \tag{3.11}$$

Rearrange equation 3.10, and use the fact that $\sum_{i=1}^{n} x_i = n\bar{x}$:

$$\begin{aligned} nb_0 &= n\bar{y} + nb_1\bar{x} \\ b_0 &= \bar{y} - b_1\bar{x} \end{aligned} \tag{3.12}$$

Now substitute equation 3.12 into equation 3.11:

$$\sum_{i=1}^{n}\left(x_i y_i - (\bar{y} - b_1\bar{x})x_i - b_1 x_i^2\right) = 0$$

and distribute the sum to each term in the expression:

$$\sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i + b_1\bar{x}\sum_{i=1}^{N} x_i - b_1\sum_{i=1}^{n} x_i^2 = 0 \tag{3.13}$$

Again, use the fact that $\sum_{i=1}^{n} x_i = n\bar{x}$ and solve equation 3.13 for b_1:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i y_i) - n\bar{x}\bar{y}}{\sum_{i=1}^{n}(x_i^2) - n\bar{x}^2}$$

Finally, use the 2 hints in the question to write:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

6. Estimate a Cobb-Douglas production function by downloading data from Statistics Canada.
   - Go to the Statistics Canada web page
   - Click "Data"
   - Search "output capital labour"
   - Select "Multifactor productivity..."
   - Click "Add/Remove data"
   - Select "Multifactor productivity and related variables"

- Tick only:
  - "Labour input" → "Hours worked"
  - "Gross output"
  - "Capital cost", then click "Apply"
- Select "(NAICS)", and choose one industry (I chose "Oil and gas extraction"). Click "Apply"
- Select "Reference period", choose 1961 to 2019, click "Apply"
- Select "Customize layout", change "Display Reference period as" to "Row"
- Click "Download options", select top option
- Open the file in Excel. Delete all the junk at the top and bottom of the file. Label the columns: "year", "L", "Y" and "K" (or some such). Highlight all numerical values, right-click, format cells, select "Numeric". Save the file as a simple .csv file: "cobb.csv" (or some such).
- In RStudio, open the .csv file (see Section 2.8) using, for example:

```
mydata <- read.csv(file.choose())
```

- Finally, estimate the model via LS:

```
lm(log(Y) ~ log(L) + log(K), data = cobb_douglas)
```

```
Coefficients:
 (Intercept)   log(labour)  log(capital)
      0.1350         0.3477         0.8823
```

- Interpretation: if labour were to increase by 1%, output is predicted to increase by 0.35%, etc.

Chapter 4

Algebraic/geometric properties of least squares

4.1 Orthogonality property of residuals

First, note that the LS residuals are "orthogonal" to the regressors:

$$\underset{(k\times 1)}{X'X\boldsymbol{b} - X'\boldsymbol{y}} = \mathbf{0} \quad \text{"normal equations"}$$

So,

$$-X'(\boldsymbol{y} - X\boldsymbol{b}) = -X'\boldsymbol{e} = \mathbf{0}$$

or,

$$X'\boldsymbol{e} = \mathbf{0} \tag{4.1}$$

Note that, by definition, equation 4.1 must be true. Both the least-squares method of minimizing $\boldsymbol{e}'\boldsymbol{e}$ to derive $\boldsymbol{b}$, and the derivation through the Method of Moments condition that $\mathbb{E}[X'\boldsymbol{b}] = \mathbf{0}$, imply equation 4.1.

Question: What does equation 4.1 imply for our ability to test the important assumption A.5?

Orthogonality is an important concept in econometrics. If two vectors in 2-dimensional space are orthogonal, then they are at right-angles. The inner-product of orthogonal vectors ($\boldsymbol{x}$ and $\boldsymbol{y}$ for example) is the null vector: $\boldsymbol{x}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{x} = \mathbf{0}$. This is important because $\boldsymbol{x}'\boldsymbol{y} = \mathbf{0}$ implies that $\text{cov}(\boldsymbol{x}, \boldsymbol{y}) = \mathbb{E}\left[(\boldsymbol{x} - \mu_x)'(\boldsymbol{y} - \mu_y)\right] = \mathbf{0}$.

4.2 Implication of $X'e = 0$ and regression with a constant

If the model includes an intercept term, then one regressor (say, the first column of X) is a unit vector. In this case we get three further results.

4.2.1 The LS residuals sum to zero

$$X'\boldsymbol{e} = \begin{pmatrix} 1 & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_{nk} \end{pmatrix}' \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_i e_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

From the first element:

$$\sum_{i=1}^{n} e_i = 0 \tag{4.2}$$

Question: How does this property relate to A.3?

4.2.2 The fitted regression passes through the sample mean

$$X'\boldsymbol{y} = X'X\boldsymbol{b}$$

or,

$$\begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ x_{1k} & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & \dots & x_{1k} \\ \vdots & \ddots & \vdots \\ 1 & \dots & x_{nk} \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

So,

$$\begin{pmatrix} \sum_i y_i \\ ? \\ ? \end{pmatrix} = \begin{pmatrix} n & \sum_i x_{i2} & \dots \\ ? & \dots & ? \\ ? & \dots & ? \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix}.$$

From the first row of this vector equation:

$$\sum_i y_i = nb_1 + b_2 \sum_i x_{i2} + \cdots + b_k \sum_i x_{ik}$$

or

$$\bar{y} = b_1 + b_2\bar{x}_2 + \cdots + b_k\bar{x}_k$$

4.2.3 The sample mean of the fitted y-values equals the sample mean of actual y-values

Recall that the unobservable components of y_i can be replaced by estimates and residuals:

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i = \boldsymbol{x}_i'\boldsymbol{b} + e_i = \widehat{y}_i + e_i.$$

So,

$$\frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \widehat{y}_i + \frac{1}{n}\sum_{i=1}^{n} e_i,$$

or,

$$\bar{y} = \overline{\hat{y}} + 0 = \overline{\hat{y}}$$

These last three results use the fact that the model includes an intercept.

4.3 Frisch-Waugh-Lovell Theorem

We first establish some results and develop some tools, then show the Frisch-Waugh-Lovell (FWL) Theorem, and then briefly discuss some applications of the FWL theorem.

4.3.1 Partitioning

The usual population regression model:

$$\underset{(n\times 1)}{y} = \underset{(n\times k)\,(k\times 1)}{X\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}} \tag{4.3}$$

can be partitioned into "blocks". Assuming 2 blocks for simplicity:

$$\underset{(n\times 1)}{y} = \underset{(n\times k_1)\,(k_1\times 1)}{X_1\boldsymbol{\beta}_1} + \underset{(n\times k_2)\,(k_2\times 1)}{X_2\boldsymbol{\beta}_2} + \underset{(n\times 1)}{\boldsymbol{\epsilon}} \tag{4.4}$$

where $k_1 + k_2 = k$. Note that when multiplying or transposing matrices with a partition, the partitions behave as if they were elements in a matrix.

Equations 4.3 and 4.4 are equivalent. In this section, we will investigate how the formula for b_2 in the LS regression of models 4.3 and 4.4 depend on X_1, under what circumstances X_1 is not needed to determine b_2, and how X_1 may be used to transform X_2 in such a way that X_1 is not directly needed to obtain b_2 (a "two-stage" approach). Before we do this, we need some tools:

4.3.2 Two "orthogonal projection" matrices

Two orthogonal projection matrices, P_X and M_X, project any vector that they pre-multiply onto the subspace $\mathcal{S}(X)$ and $\mathcal{S}^{\perp}(X)$ respectively. Define these projection matrices to be:

$$P_X = X\left(X'X\right)^{-1}X'$$

and

$$M_X = \left(I - X\left(X'X\right)^{-1}X'\right)$$

The interpretations of these projection matrices are:

- When a vector ($\boldsymbol{y}$, for example) is pre-multiplied by P_X, the result is the predicted (fitted values $\hat{\boldsymbol{y}}$) of a LS regression of $\boldsymbol{y}$ on X.
- When a vector ($\boldsymbol{y}$, for example) is pre-multiplied by M_X, the result is the residuals ($\boldsymbol{e}$) from a LS regression of $\boldsymbol{y}$ on X.
- P_X is a "predicted-value-maker".
- M_X is a "residual-maker".

Prove that the above interpretations of P_X and M_X are true.

We can also project vectors onto subspaces using only subsets, or blocks of the X matrix. For example, $P_{X_1}\boldsymbol{y}$, where:

$$P_{X_1} = X_1\left(X_1'X_1\right)^{-1}X_1',$$

produces the fitted values from a regression of $\boldsymbol{y}$ on X_1 only. Similarly, $M_{X_1}\boldsymbol{y}$, where:

$$M_{X_1} = \left(I - X_1\left(X_1'X_1\right)^{-1}X_1'\right),$$

produces the residuals from a regression of the model $\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$. For some ease of notation, we will suppress the "X" in the subscript so that:

$$P_1 = P_{X_1}$$

and

$$M_1 = M_{X_1}$$

Note that P_X and M_X are *idempotent* matrices. For example, $M_X M_X = M_X M_X' = M_X = M_X' M_X$. This result is used frequently in algebra that follows.

4.3.3 Invariance of $\hat{y}$ and e to non-singular linear transformations

What would you expect to happen to the estimation results if you measured one of the $\boldsymbol{x}$ variables in different units? For example, 1000s of dollars instead of dollars? In a model of energy consumption, degrees Fahrenheit instead of degrees Celsius? The estimated values for the βs would change, but in a way that corresponds to the change of units.

Such a change in units, where an $\boldsymbol{x}$ variable is linearly transformed (just by multiplying it and/or adding to it), should not change the estimated model in a substantive way. It would be disturbing if the model's predictions $\hat{\boldsymbol{y}}$, and residuals, changed under such a transformation. Any such linear transformation can be represented in a $k \times k$ matrix (call it A) that post-multiplies X: XA. For example, if X contains $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, and we wanted $\boldsymbol{x}_2$ to be measured in 1000s of dollars instead of dollars, then:

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{1000} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The "invariance property" means that the models:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and

$$\boldsymbol{y} = XA\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

produce the same $\hat{\boldsymbol{y}}$ and $\boldsymbol{e}$ when estimated by LS.

Prove the above result.

4.3.4 Independence between LS estimators and orthogonal regressors

In general, in the model:

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{4.5}$$

$\boldsymbol{b}_1$ depends on the X_2 data. This is intuitive: often the very reason that we include many of the regressors in the model is to provide "controls". These are variables that are correlated with the $\boldsymbol{x}$ variables of interest, and that may also determine $\boldsymbol{y}$.

That is, if we "drop" X_1 from the model:

$$\boldsymbol{y} = X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}, \tag{4.6}$$

then, in general, the LS estimates for $\boldsymbol{\beta}_2$ are **not** equal between models 4.5 and 4.6.

However, an important situation is that, if X_1 and X_2 are orthogonal (uncorrelated), then the LS estimates for $\boldsymbol{\beta}_2$ from models 4.5 and 4.6 **are equivalent**. There is a messy, but useful, way to prove this.

Proof

Visualize the partitioned model:

$$\boldsymbol{y} = [X_1 : X_2] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \boldsymbol{\epsilon} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

and the LS estimator is still:

$$\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y}.$$

We can rewrite the LS estimator as:

$$\begin{aligned} \boldsymbol{b} &= \left\{ [X_1 : X_2]' [X_1 : X_2] \right\}^{-1} [X_1 : X_2]' \boldsymbol{y} \\ &= \left\{ \begin{bmatrix} X_1' \\ \dots \\ X_2' \end{bmatrix} \begin{bmatrix} X_1 & : & X_2 \end{bmatrix} \right\}^{-1} \begin{bmatrix} X_1' \\ \dots \\ X_2' \end{bmatrix} \boldsymbol{y} \end{aligned}$$

and

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1} \begin{pmatrix} X_1'\boldsymbol{y} \\ X_2'\boldsymbol{y} \end{pmatrix}.$$

The "normal equations" underlying this are

$$(X'X)\,\boldsymbol{b} = X'\boldsymbol{y}$$

or:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = \begin{pmatrix} X_1'\boldsymbol{y} \\ X_2'\boldsymbol{y} \end{pmatrix}$$

Now, we can solve the normal equations independently for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$! The main advantage is that some properties of $\boldsymbol{b}_1$ can be determined independently of $\boldsymbol{b}_2$. Solving the normal equations for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$:

$$X_1'X_1\boldsymbol{b}_1 + X_1'X_2\boldsymbol{b}_2 = X_1'\boldsymbol{y} \tag{4.7}$$

$$X_2'X_1\boldsymbol{b}_1 + X_2'X_2\boldsymbol{b}_2 = X_2'\boldsymbol{y} \tag{4.8}$$

From 4.7:

$$(X_1'X_1)\,\boldsymbol{b}_1 = X_1'\boldsymbol{y} - X_1'X_2\boldsymbol{b}_2,$$

or:

$$\begin{aligned} \boldsymbol{b_1} &= (X_1'X_1)^{-1} X_1'\boldsymbol{y} - (X_1'X_1)^{-1} X_1'X_2\boldsymbol{b}_2 \\ &= (X_1'X_1)^{-1} [X_1'\boldsymbol{y} - X_1'X_2\boldsymbol{b}_2] \end{aligned} \tag{4.9}$$

Now we see that only if $X_1'X_2 = 0$, then $\boldsymbol{b_1} = (X_1'X_1)^{-1} X_1'\boldsymbol{y}$. If a regressor is dropped from the model, the estimated values for $\boldsymbol{b}$ will change, unless the dropped variable is *orthogonal* to the other regressors.

4.3.5 FWL Theorem

We have proved two algebraic properties of LS estimation: (i) any linear transformation of X leaves $\bar{\boldsymbol{y}}$ and $\boldsymbol{e}$ unchanged; (ii) $\boldsymbol{b}_2$ remains unchanged when X_1 is dropped, only if $X_1'X_2 = 0$.

Now, in general $X_1'X_2 \neq 0$; the $\boldsymbol{x}$ variables in the model are typically correlated with each other. But consider that we could use X_1 in order to transform X_2 in a way such that the transformation is orthogonal to X_1. Then, X_1 is no longer needed to estimate $\boldsymbol{\beta}_2$. Furthermore, if we apply the same transformation to $\boldsymbol{y}$, then $\hat{\boldsymbol{y}}$ and $\boldsymbol{e}$ will also be identical. The FWL theorem says that such a transformation is M_1. That is, the two models:

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{4.10}$$

and

$$M_1\boldsymbol{y} = M_1X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{4.11}$$

yield identical results for $\hat{\boldsymbol{y}}$, $\boldsymbol{e}$, and $\boldsymbol{b}_2$ when estimated by LS.

Proof

Again, the LS estimator is:

$$\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} \tag{4.12}$$

Applying formula 4.12 to model 4.11, the LS estimator for $\boldsymbol{\beta}_2$ is:

$$\boldsymbol{b}_2^\star = (X_2'M_1X_2)^{-1}X_2'M_1\boldsymbol{y} \tag{4.13}$$

The trick is to show that $\boldsymbol{b}_2^\star$ is identical to $\boldsymbol{b}_2$ from model 4.10. The estimated model corresponding to model 4.10 is:

$$\boldsymbol{y} = X_1\boldsymbol{b}_1 + X_2\boldsymbol{b}_2 + \boldsymbol{e}$$

Pre-multiply both sides by $X_2'M_1$:

$$\begin{aligned} X_2'M_1\boldsymbol{y} &= X_2'M_1X_1\boldsymbol{b}_1 + X_2'M_1X_2\boldsymbol{b}_2 + X_2'M_1\boldsymbol{e} \\ &= \boldsymbol{0} + X_2'M_1X_2\boldsymbol{b}_2 + \boldsymbol{0} \end{aligned}$$

Finally, solve for $\boldsymbol{b}_2$:

$$\begin{aligned} X_2'M_1X_2\boldsymbol{b}_2 &= X_2'M_1\boldsymbol{y} \\ \boldsymbol{b}_2 &= (X_2'M_1X_2)^{-1}X_2'M_1\boldsymbol{y} \end{aligned}$$

Hence, $\boldsymbol{b}_2^\star = \boldsymbol{b}_2$.

The FWL theorem suggests that the LS estimates for $\boldsymbol{\beta}_2$ in a partitioned model such as 4.10 can be obtained through the following steps:

(i) Regress X_2 on X_1, get the residuals (M_1X_2).
(ii) Regress $\boldsymbol{y}$ on X_1, get the residuals $(M_1\boldsymbol{y})$.
(iii) Regress the residuals from (ii) onto (i).

4.4 Applications of FWL

Now that we know the implications of the FWL theorem, it is much easier to effect, and understand, common transformations of variables that are practised in econometrics:

- deviations from the mean (de-meaning or centring the variables)
- de-seasonalizing data
- de-trending data

All of the transformations above are non-singular linear transformations. From the invariance property in Section 4.3.3 we know that $\bar{\boldsymbol{y}}$ and $\boldsymbol{e}$ will be unchanged. If we can showed that the transformed data is orthogonal to any of the regressors, then the FWL theorem tells us those regressors may be dropped from the model.

In most cases of de-ZZZZing data, the transformations are not necessary and do not change anything substantive in the model. The transformations can, however, aid in the visualization of the data, interpretation of the estimated parameters, and in some cases ease the computational burden in calculating the estimates.

4.4.1 Centring data (deviations from the mean, or de-meaning)

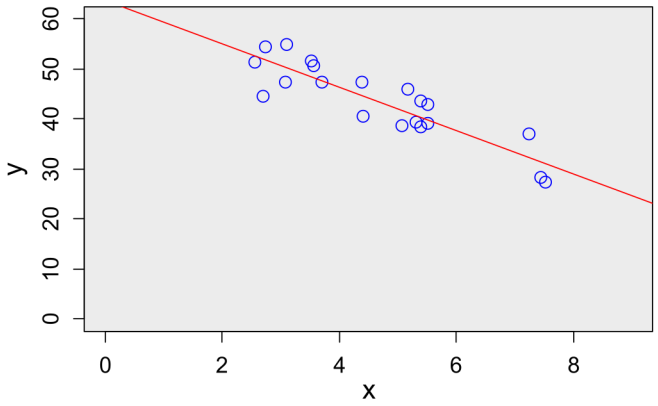
Take the simple model:

$$\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon} \tag{4.14}$$

Under what circumstances can we drop the constant β_1 from model 4.14? The LS estimators for β_2 would be different if we were to instead estimate the model:

$$\boldsymbol{y} = \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon} \tag{4.15}$$

Figure 4.1: A least-squares line fitted through some uncentred data. The estimated intercept of $b_1 = 63.7$ is outside the range of the data, and has little economic meaning in most models.



However, if we use "centred" data, we can drop the intercept. To illustrate this, download some *un*centred data:

```
un <- read.csv("http://rtgodwin.com/data/centrethis.csv")
```

Estimate the intercept and slope in model 4.14:

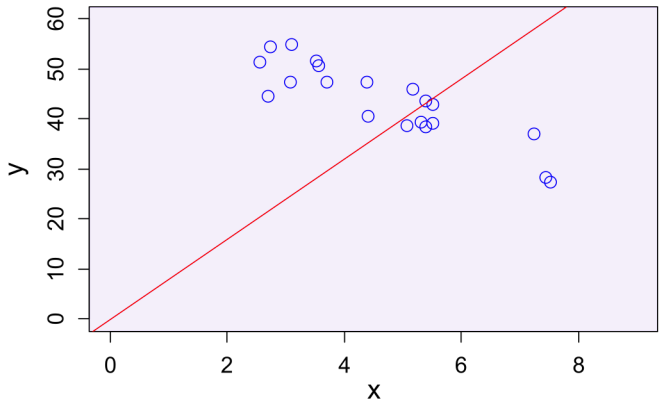
```
lm(y ~ x, data=un)
```

```
Coefficients:
(Intercept)            x
     63.744       -4.328
```

Plot the data and the estimated line:

```
plot(un$x, un$y, ylim=c(0,60), xlim=c(0,9))
abline(lm(y ~ x, data=un))
```

Figure 4.2: The least-squares line is forced through the origin if the model does not include an intercept.



The plot appears in Figure 4.1. If we instead estimate model 4.16 the estimated β_2 is of course different since the least-squares line must pass through the origin (see Figure 4.2):

```
abline(lm(y ~ x -1, data=un))
```

Recall that the regression line must pass through the sample means of the data (see Section 4.2.2). If the data all has mean zero, then the LS line must pass through the origin anyway, and dropping the intercept has no effect.

To centre data, we *transform* it by subtracting its sample mean:

$$\boldsymbol{y}^\star = \boldsymbol{y} - \boldsymbol{i}\bar{y} \quad ; \quad \boldsymbol{x}^\star = \boldsymbol{x} - \boldsymbol{i}\bar{x}$$

where $\boldsymbol{y}^\star$ and $\boldsymbol{x}^\star$ are the centred variables, and $\boldsymbol{i}$ is a column vector of 1s:

$$\underset{(n \times 1)}{\boldsymbol{i}} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

To centre the variables in R:

```
xstar <- un$x - mean(un$x)
ystar <- un$y - mean(un$y)
```

Now we can see that estimating the model:

$$\boldsymbol{y}^\star = \beta_2 \boldsymbol{x}^\star + \boldsymbol{\epsilon} \tag{4.16}$$

yields identical results to model 4.14, for the estimated β_2:

```
lm(ystar ~ xstar -1)
```

```
Coefficients:
 xstar
-4.328
```

The trick is that the variables $\boldsymbol{y}^\star$ and $\boldsymbol{x}^\star$ are orthogonal to the column vector $\boldsymbol{i}$. The FWL theorem says that exclusion of this regressor ($\boldsymbol{i}$) does not affect the LS estimates. To prove this, consider the residuals from regressions of $\boldsymbol{x}$ on a constant, and $\boldsymbol{y}$ on a constant. That is, consider the vectors $M_i\boldsymbol{y}$ and $M_i\boldsymbol{x}$. Let:

$$M_i = I - \boldsymbol{i}\left(\boldsymbol{i}'\boldsymbol{i}\right)^{-1}\boldsymbol{i}' = I - \frac{1}{n}\boldsymbol{i}\boldsymbol{i}' \tag{4.17}$$

Then,

$$M_i\boldsymbol{y} = \boldsymbol{y} - \boldsymbol{i}\bar{y} = \boldsymbol{y}^\star$$

The M_i matrix, when pre-multiplying a vector, creates the deviations-from-means. That is, it centres a variable. To see how this works, multiply out $M_i\boldsymbol{y}$:

$$\begin{aligned} M_i\boldsymbol{y} &= \left\{ \begin{bmatrix} 1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & 1 \end{bmatrix} - \begin{bmatrix} 1/n & \dots & 1/n \\ \vdots & \ddots & \vdots \\ 1/n & \dots & 1/n \end{bmatrix} \right\} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= \begin{bmatrix} y_1 - y_1/n - y_2/n - \dots - y_n/n \\ \vdots \\ y_n - y_1/n - y_2/n - \dots - y_n/n \end{bmatrix} = \begin{pmatrix} y_1 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} \end{aligned}$$

The transformed (centred) variables are now orthogonal to the regressor $\boldsymbol{i}$, that is $(M_i\boldsymbol{y})'\boldsymbol{i} = 0$ and $(M_i\boldsymbol{x})'\boldsymbol{i} = 0$, and so the intercept may be dropped from the model without any substantive effect.

4.4.2 De-seasonalizing

Time series data that is reported quarterly may have seasonal effects. Certain activities only take place in the summer, more presents are purchased in the 4th quarter, etc. It is often desirable to visualize, and work with, data that has been de-seasonalized. That is, we sometimes want to *purge* the seasonal patterns from data, and explain variation in the data that is independent from seasonal variation. For example, see Figure 4.3 for the quarterly residential demand for natural gas in Manitoba, from 1990 Q1 to the end of 2001 Q4[1]. There is a strong seasonal component.
To produce a graph similar to that in Figure 4.3, download the data in R:

```
gas <- read.csv("http://rtgodwin.com/data/MBgas.csv")
```

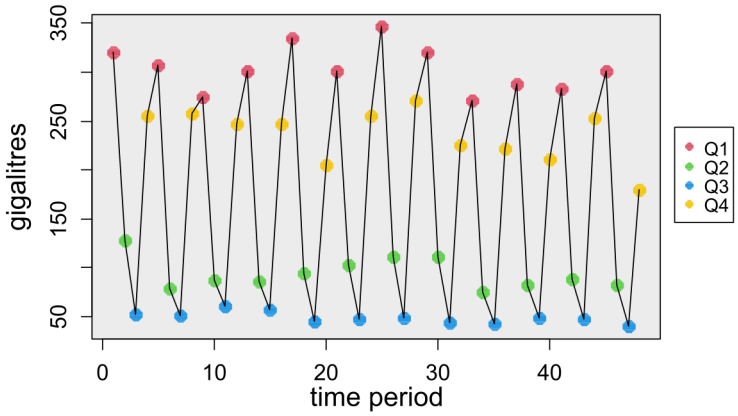
Create a time-trend variable, and plot the gas consumption over time:

```
gas$time <- 1:nrow(gas)
plot(gas$time, gas$gigalitres, type = "l",
        xlab = "time period", ylab = "gigalitres")
```

As mentioned, it may be desirable to *remove* the trend or *de-seasonalize* the data, in order that non-seasonal variation may be more easily visualized and examined. This can be accomplished through

[1]Data from: Statistics Canada. Table 25-10-0005-01 Supply and demand of primary and secondary energy in natural units, quarterly, with data for years 1990 - 2001

Figure 4.3: Quarterly seasonality in the residential demand for natural gas in Manitoba.



quarterly dummy variables. Let $\boldsymbol{q}_1$ be a dummy variable equal to 1 if the reference period is in the first quarter, and 0 otherwise. Similar definitions follow for the dummy variables $\boldsymbol{q}_2$, $\boldsymbol{q}_3$, and $\boldsymbol{q}_4$. That is, we need to create dummy variables that look like:

$$\boldsymbol{q}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \quad ; \quad \boldsymbol{q}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ \vdots \end{bmatrix} \quad ; \quad \boldsymbol{q}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ \vdots \end{bmatrix} \quad ; \quad \boldsymbol{q}_4 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \end{bmatrix}$$

Beware! We cannot use all four of these dummy variables in our regression model (otherwise we fall into the "dummy variable trap"). This is because there is an exact linear dependency between these four dummy variables, and the intercept. That is:

$$\boldsymbol{q}_1 + \boldsymbol{q}_2 + \boldsymbol{q}_3 + \boldsymbol{q}_4 = \boldsymbol{i} \tag{4.18}$$

Including all 4 dummies, and the intercept, would be a violation of A.2 (no perfect multicollinearity): the $(X'X)$ matrix is not invertible and the LS estimator is undefined. We must *exclude* one of the dummy variables, or exclude the intercept.

Question: Suppose we estimate a model where we exclude $\boldsymbol{q}_1$. What would change if we instead decided to exclude $\boldsymbol{q}_2$, or the intercept?

To de-seasonalize the data, we can regress the seasonal variable on the system of dummy variables, and extract the residuals. That is, the de-seasonalized variable is $M_Q\boldsymbol{y}$, where Q consists of the regressors representing the seasonal dummy variables. For example:

$$Q = \begin{bmatrix} \boldsymbol{i} & \boldsymbol{q}_2 & \boldsymbol{q}_3 & \boldsymbol{q}_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

$M_Q\boldsymbol{y}$ consists of variation that is independent of the seasonal component. The de-seasonalized variable is now *orthogonal* to the seasonal dummies. To accomplish this de-seasonalization in R, we first create the system of dummies:

```
gas$q4 <- gas$q3 <- gas$q2 <- gas$q1 <- 0
gas$q1[seq(1, n, 4)] <- 1
gas$q2[seq(2, n, 4)] <- 1
gas$q3[seq(3, n, 4)] <- 1
gas$q4[seq(4, n, 4)] <- 1
```

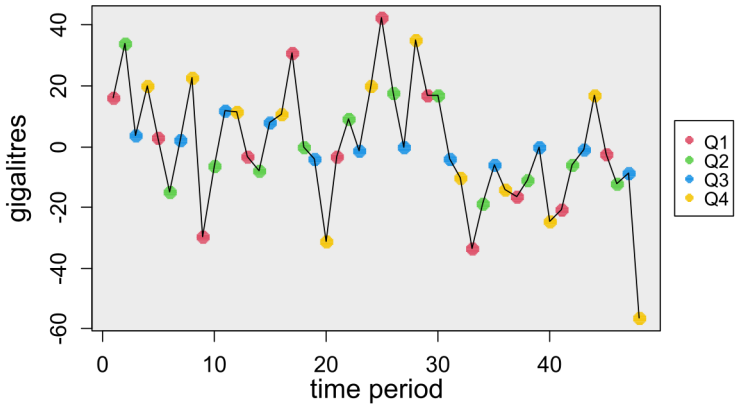
and then extract the residuals from the regression of the time series on the seasonal dummies:

```
Mgigalitres <- lm(gigalitres ~ q2 + q3 + q4, data = gas)$residuals
```

Now plot the de-seasonalized variable over time (see Figure 4.4):

```
plot(gas$time, Mgigalitres, type = "l")
```

Figure 4.4: De-seasonalized time series of the residential demand for gas in MB.



Questions:

1. What is the mean value of the de-seasonalized data?
2. What is the mean value of the de-seasonalized data, for the 1st quarter?
3. When is it acceptable to "drop" seasonal dummy variables from a model?

To answer question 3, consider what is *wrong* with estimating the model:

$$gas^{\star} = \beta_1 + \beta_2 temp + \epsilon \tag{4.19}$$

where $temp$ is the mean quarterly temperature at Richardson International airport obtained from Environment Canada, and $gas^{\star} = M_q gas$ is the de-seasonalized demand for MB gas. Estimate this model in R:

```
summary(lm(Mgigalitres ~ gas$temp))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.6650     2.8550   0.233    0.817
gas$temp      -0.2350     0.2397  -0.980    0.332

Residual standard error: 19.21 on 46 degrees of freedom
Multiple R-squared:  0.02046,    Adjusted R-squared:  -0.0008304
F-statistic: 0.961 on 1 and 46 DF,  p-value: 0.3321
```

Notice that the variable `gas$temp` is *insignificant*. Consider instead the regression model:

$$gas^\star = \beta_1 + \beta_2 temp^\star + \epsilon \tag{4.20}$$

where $temp^\star = M_Q temp$ and has likewise been de-seasonalized. Estimating this in R:

```
Mtemp <- lm(temp ~ q2 + q3 + q4, data = gas)$residuals
summary(lm(Mgigalitres ~ Mtemp - 1))
```

```
Coefficients:
      Estimate Std. Error t value Pr(>|t|)
Mtemp  -8.0933     0.7641  -10.59 4.83e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.44 on 47 degrees of freedom
Multiple R-squared:  0.7048,    Adjusted R-squared:  0.6985
F-statistic: 112.2 on 1 and 47 DF,  p-value: 4.835e-14
```

Notice that `Mtemp` is *significant*.

Questions:

1. Which model is the "correct" one to estimate, model 4.19 or model 4.20? Why?
2. Why has the intercept been dropped in model 4.20?
3. In general, what might be the problem with using de-seasonalized data?

4.5 Goodness of fit

One way of measuring the "quality" of a fitted regression model is by the extent to which the model "explains" the sample variation for $\boldsymbol{y}$. That is, our LS model seeks to explain the sample variance of $\boldsymbol{y}$, and our measure of fit tells how good a job the model does. Recall that the sample variance of $\boldsymbol{y}$ is:

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

Or, we could just use

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

to measure variability. One estimator for variance has an $(n-1)$ in the denominator while another just has n. Since we will be comparing the sample variance of two vectors with the same dimension $(n \times 1)$, the denominator will cancel out.

Two measures of fit that are often used in conjunction with LS are the R-squared R^2 and adjusted-R-squared $\bar{R}^2$. In this section, we present both these measures of fit, and argue why $\bar{R}^2$ is usually better.

4.5.1 Coefficient of Determination - R^2

R^2 is the ratio of variance in $\boldsymbol{y}$ that can be explained using the model (the X variables and LS estimates) over the total variance in $\boldsymbol{y}$. Start by writing our measure of sample variance in matrix form. Measures of variability use the squared *deviations-from-means*, so we can use the M_i matrix (see equation 4.17):

$$\boldsymbol{y}' M_i \boldsymbol{y} = \boldsymbol{y}' M_i' M_i \boldsymbol{y} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.21}$$

Recall that LS "decomposes" $\boldsymbol{y}$ into two components, fitted values and residuals:

$$\boldsymbol{y} = P_X\boldsymbol{y} + M_X\boldsymbol{y} = X\boldsymbol{b} + \boldsymbol{e} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$$

We can take the sample variance of *both* sides of $\boldsymbol{y} = \widehat{\boldsymbol{y}} + \boldsymbol{e}$. Start by taking the deviations in means of both sides:

$$M_i\boldsymbol{y} = M_i\widehat{\boldsymbol{y}} + M_i\boldsymbol{e} = M_i\widehat{\boldsymbol{y}} + \boldsymbol{e} \tag{4.22}$$

We have converted the components of the model into deviations from means.

Question: Why does $M_i e = e$?

Now, pre-multiply both sides of equation 4.22 by its own transpose:

$$\begin{aligned}\boldsymbol{y}'M_i\boldsymbol{y} = \boldsymbol{y}'M_i'M_i\boldsymbol{y} &= \left(M_i\widehat{\boldsymbol{y}} + \boldsymbol{e}\right)'\left(M_i\widehat{\boldsymbol{y}} + \boldsymbol{e}\right) \\ &= \widehat{\boldsymbol{y}}'M_i\widehat{\boldsymbol{y}} + \boldsymbol{e}'\boldsymbol{e} + 2e'M_i\widehat{y}\end{aligned}.$$

however,

$$e'M_i\widehat{y} = e'M_i'\widehat{y} = \left(M_i e\right)'\widehat{y} = e'\widehat{y} = e'X\left(X'X\right)^{-1}X'y = 0$$

We have "decomposed" the sample variance of $\boldsymbol{y}$ into two parts: that which is explained by the estimated model ($\widehat{\boldsymbol{y}}$), and that which is unexplained ($\boldsymbol{e}$).

Question: Why does $e'M_i\widehat{y} = 0$? Recall the rule for taking the variance of a sum of two random variables.

So, we have (recall that $\bar{\widehat{y}} = \bar{y}$):

$$\begin{array}{ccccc}\boldsymbol{y}'M_i\boldsymbol{y} & = & \widehat{\boldsymbol{y}}'M_i\widehat{\boldsymbol{y}} & + & \boldsymbol{e}'\boldsymbol{e} \\ \sum_{i=1}^n\left(y_i-\bar{y}\right)^2 & = & \sum_{i=1}^n\left(\widehat{y}_i-\bar{y}\right)^2 & + & \sum_{i=1}^n e_i^2 \\ \text{TSS} & = & \text{ESS} & + & \text{RSS}\end{array}$$

Where TSS = "total sum of squares", ESS = "explained sum of squares", and RSS = "residual sum of squares". This lets us define the R-squared:

$$R^2 = \left(\frac{ESS}{TSS}\right) = 1 - \left(\frac{RSS}{TSS}\right)$$

R^2 is the portion of variance in the dependent variable that can be explained by the estimated model.

- The second equality in the definition of R^2 holds only if model includes an intercept.
- $R^2 = \left(\frac{ESS}{TSS}\right) \geq 0$
- $R^2 = 1 - \left(\frac{RSS}{TSS}\right) \leq 1$
- So, $0 \leq R^2 \leq 1$
- R^2 is *unitless*.

Question: What is the interpretation of $R^2 = 0$ and $R^2 = 1$?

4.5.2 R^2 increases when a regressor is added to the model

What happens if we add any regressor(s) to the model? Consider the population model:

$$y = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon} \tag{4.23}$$

then consider adding regressors to it:

$$y = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{u} \tag{4.24}$$

Optimization problem A - apply LS to 4.24:

$$\min\left(\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}\right) \quad ; \quad \widehat{\boldsymbol{u}} = \boldsymbol{y} - X_1\boldsymbol{b}_1 - X_2\boldsymbol{b}_2$$

Optimization problem B - apply LS to 4.23:

$$\min\left(\boldsymbol{e}'\boldsymbol{e}\right) \quad ; \quad \boldsymbol{e} = \boldsymbol{y} - X_1\hat{\boldsymbol{\beta}}_1$$

Problem B is just Problem A, subject to the restriction: $\boldsymbol{\beta}_2 = \mathbf{0}$. Minimized value in A must be $\leq$ minimized value in B. So, $\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}} \leq \boldsymbol{e}'\boldsymbol{e}$.

- Adding any regressor(s) to the model cannot increase (and typically will decrease) the sum of squared residuals.
- So, adding any regressor(s) to the model cannot decrease (and typically will increase) the value of R^2.
- "Junk" variables could be added to the model to get the R^2 arbitrarily high.
- Means that R^2 is not really a very interesting measure of the "quality" of the regression model, in terms of explaining sample variability of the dependent variable.

For these reasons, we usually "adjust" the Coefficient of Determination.

4.5.3 Adjusted R-square: $\bar{R}^2$

Modify R^2:

$$R^2 = \left[1 - \frac{\boldsymbol{e}'\boldsymbol{e}}{\boldsymbol{y}'M_i\boldsymbol{y}}\right]$$

to become:

$$\bar{R}^2 = \left[1 - \frac{\boldsymbol{e}'\boldsymbol{e}/(n-k)}{\boldsymbol{y}'M_i\boldsymbol{y}/(n-1)}\right]$$

We're adjusting for "degrees of freedom" in the numerator and denominator. Now, when a regressor is added to the model, $\bar{R}^2$ increases due to the improved "fit" ($\boldsymbol{e}'\boldsymbol{e}$ decreases), and decreases due to the penalty imposed by k. $\bar{R}^2$ only increases if the improvement in model fit due to the additional regressor dominates the penalty for adding another regressor.

"Degrees of freedom" are the number of independent pieces of information.

When we calculate $\bar{y}$, we lose one degree of freedom. That is, the sample y, together with $\bar{y}$, only contains $(n-1)$ independent pieces of information. For example, if $\boldsymbol{y} = \{1, 3, z\}$, and $\bar{y} = 3$, then you know that $z = 5$. Similarly, in order to calculate $\boldsymbol{e}$, we first need to calculate the $(k \times 1)$ vector $\boldsymbol{b}$. Once we have $\boldsymbol{b}$, there are only $(n-k)$ pieces of independent information left in $\boldsymbol{e}$.

$\boldsymbol{e} = \boldsymbol{y} - X\boldsymbol{b}$. We estimate k parameters from the n data-points, before we can calculate $\boldsymbol{e}$. We have $(n-k)$ "degrees of freedom" associated with the fitted model.

Some final points about R^2 and $\bar{R}^2$:

- Possible for $\bar{R}^2 \leq 0$ (even with an intercept in the model).
- $\bar{R}^2$ can *increase* or *decrease* when we add regressors.
- In multiple regression, $\bar{R}^2$ will increase (decrease) if a variable is deleted, if and only if the associated t-statistic has *absolute value less than* (greater than) unity.
- If the model doesn't include an intercept, then TSS $\neq$ ESS + RSS, and in this case there is no longer any guarantee that $0 \leq R^2 \leq 1$.
- Must be careful comparing R^2 and $\bar{R}^2$ values across models. For example:

$$\begin{aligned} \widehat{C}_i &= 0.5 + 0.8Y_i \quad ; \quad R^2 = 0.90 \\ \log\left(\widehat{C}_i\right) &= 0.2 + 0.75Y_i \quad ; \quad R^2 = 0.80 \end{aligned}$$

The sample variation is in *different units*.

4.6 Exercises

1. Prove that the LS residuals sum to zero, that the regression line passes through the means of the data, and that the sample mean of the actual y values equals the sample mean of the fitted y values.
   Answer. See Section 4.2 for the proofs.
2. Derive the formula for $\boldsymbol{b}_2$ in the partitioned model. Show that it is equivalent to equation 4.13.
   Answer. The least-squares "normal equations" are:

$$(X'X)\,\boldsymbol{b} = X'\boldsymbol{y}$$

In the partitioned model, the normal equations can be partitioned as well:

$$\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} \begin{pmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{pmatrix} = \begin{pmatrix} X_1'\boldsymbol{y} \\ X_2'\boldsymbol{y} \end{pmatrix}$$

Solving the normal equations for $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$:

$$X_1'X_1\boldsymbol{b}_1 + X_1'X_2\boldsymbol{b}_2 = X_1'\boldsymbol{y} \tag{4.25}$$

$$X_2'X_1\boldsymbol{b}_1 + X_2'X_2\boldsymbol{b}_2 = X_2'\boldsymbol{y} \tag{4.26}$$

From 4.25:

$$(X_1'X_1)\,\boldsymbol{b}_1 = X_1'\boldsymbol{y} - X_1'X_2\boldsymbol{b}_2,$$

or:

$$\begin{aligned} \boldsymbol{b_1} &= (X_1'X_1)^{-1} X_1'\boldsymbol{y} - (X_1'X_1)^{-1} X_1'X_2\boldsymbol{b}_2 \\ &= (X_1'X_1)^{-1} [X_1'\boldsymbol{y} - X_1'X_2\boldsymbol{b}_2] \end{aligned} \tag{4.27}$$

Now, substitute 4.27 into 4.26:

$$(X_2'X_1)\left[(X_1'X_1)^{-1} X_1'\boldsymbol{y} - (X_1'X_1)^{-1} X_1'X_2\boldsymbol{b}_2\right] + (X_2'X_2)\,\boldsymbol{b}_2 = X_2'\boldsymbol{y},$$

or

$$\left[(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)\right]\boldsymbol{b}_2 = X_2'\boldsymbol{y} - (X_2'X_1)(X_1'X_1)^{-1}X_1'\boldsymbol{y}$$

and so:

$$\boldsymbol{b}_2 = \left[(X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2)\right]^{-1}\left[X_2'\left(I - X_1(X_1'X_1)^{-1}X_1'\right)\boldsymbol{y}\right]$$

Define:

$$M_1 = \left(I - X_1(X_1'X_1)^{-1}X_1'\right)$$

Then, we can write:

$$\boldsymbol{b}_2 = (X_2{}'M_1X_2)^{-1}X_2{}'M_1\boldsymbol{y}$$

Which is the same as the FWL estimator in equation 4.13.

3. The M matrix is sometimes referred to as a "residual-maker" matrix. That is, $M_Q\boldsymbol{q}$ produces the LS residuals from a regression of the vector $\boldsymbol{q}$ on the matrix Q. Let $M_1 = \left(I - X_1(X_1'X_1)^{-1}X_1'\right)$. Prove that $M_1\boldsymbol{y}$ is equal to the LS residuals of a regression of $\boldsymbol{y}$ on X_1.
Answer.

$$M_1\boldsymbol{y} = \left(I - X_1(X_1'X_1)^{-1}X_1'\right)\boldsymbol{y} = \boldsymbol{y} - X_1(X_1'X_1)^{-1}X_1'\boldsymbol{y}$$

But $(X_1'X_1)^{-1}X_1'\boldsymbol{y}$ is just the LS estimator from a regression of $\boldsymbol{y}$ on X_1, so:

$$M_1\boldsymbol{y} = \boldsymbol{y} - X_1\boldsymbol{b}_1 = \boldsymbol{e}_1$$

where $\boldsymbol{e}_1$ are the residuals from an LS regression of $\boldsymbol{y}$ on X_1.

4. Prove that if the data is transformed using M_1, then X_1 may be dropped from the regression model.
Answer. This is the same as the FWL proof in Section 4.3.5.

5. Download Canadian GDP data using:

```
cangdp <- read.csv("http://ryantgodwin.com/data/canada-gdp.csv")
```

De-trend the log of GDP. Compare the time plot of the log of GDP, with the time plot of the de-trended log of GDP. How do you interpret values of the de-trended time series? What might be a problem with using the de-trended variable in a regression model?
Answer. To de-trend the variable, extract the residuals from a regression of `log(GDP)` on the `Year` variable:

```
detrendedlogGDP <- lm(log(GDP) ~ Year, data = cangdp)$residuals
```

Then, compare the two time plots:

```
plot(cangdp$Year, log(cangdp$GDP), type="l")
plot(cangdp$Year, detrendedlogGDP, type="l")
```

Interpretation: since the `GDP` variable is in logs, and after the variable has been centred, each observation is now a *percentage deviation from the mean*. A potential problem: unless all variables in the model are de-trended in the same way, omitting the time trend and using de-trended GDP will not give the same estimates as including the time trend and using the trending GDP.

6. Take a simple population model: $\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon}$. Sketch some data, and the LS line, for which $R^2 = 0$ and $R^2 = 1$. What must be true of b_2 if $R^2 = 0$ and if $R^2 = 1$?
   Brief answer. When $b_2 = 0$, $R^2 = 0$. When the estimated LS "line" passes through each data point, $R^2 = 1$.
7. Estimate a LS model in R, and find the R^2 and $\bar{R}^2$. Randomly create a "junk" variable and add it to the model. What happens to the R^2 and $\bar{R}^2$?
   Brief answer. The R^2 will increase, but the $\bar{R}^2$ can either increase or decrease.

Chapter 5

Finite sample properties of the least squares estimator

In this chapter, we will derive some of the finite sample properties for the least squares estimator. By finite sample properties, we mean the statistical properties of $\boldsymbol{b}$ for some finite value of n. Later, we will consider the properties of $\boldsymbol{b}$ for when $n \to \infty$ (i.e. asymptotic properties). Recall that the population model is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and that by assumptions A.3, A.4 and A.6, the error term is:

$$\boldsymbol{\epsilon} \sim N\left[\mathbf{0}, \sigma^2 I_n\right]$$

and also recall that our LS estimator is:

$$\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y} = f(\boldsymbol{y})$$

That is, $\boldsymbol{b}$ is a function of the random sample data, so is itself random! This is a very important point.

$$\boldsymbol{\epsilon} \text{ is random } \longrightarrow \boldsymbol{y} \text{ is random } \longrightarrow \boldsymbol{b} \text{ is random}$$

- $\boldsymbol{b}$ is an estimator of $\boldsymbol{\beta}$. It is a function of the *random* sample data.
- $\boldsymbol{b}$ is a "statistic".
- $\boldsymbol{b}$ has a probability distribution – called its *sampling distribution*.

Interpretation of the sampling distribution:

- Repeatedly draw all possible samples of size n.
- Calculate values of $\boldsymbol{b}$ each time.
- Construct a relative frequency distribution for the $\boldsymbol{b}$ values and probability of occurrence.
- It is a hypothetical construct.

Question: Why is the sampling distribution a hypothetical construct?

The *sampling distribution* offers one basis for answering the question: "How good is $\boldsymbol{b}$ as an estimator of $\boldsymbol{\beta}$?"

We will be assessing the quality of the estimator in terms of its performance in *repeated samples*. The finite sample properties that we derive in this chapter tell us nothing about the quality of the estimator for *one particular sample*.

We will explore some of the properties of the LS estimator, $\boldsymbol{b}$, and build up its sampling distribution. We'll introduce some general results, and then apply them to our specific problem. Before we do so, let's look at a *simulation experiment* using R. For the experiment, I will estimate the simple model:

$$\boldsymbol{y} = \beta_1 + \beta_2 \boldsymbol{x} + \boldsymbol{\epsilon}$$

where I will maintain all of the usual assumptions (A.1-A.6) and choose:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \end{bmatrix} \quad ; \quad \sigma^2 = 1 \quad ; \quad n = 100$$

In addition I will use the unrealistic assumption that x is "fixed in repeated samples", and will generate the x variable from a $N(0, 1)$ distribution (I will use the same distribution for ϵ as well).

In R, run the following code to set the parameters of the experiment:

```
beta1 <- 2
beta2 <- -4
n <- 100
x <- rnorm(n)
```

and now run the following code several times:

```
epsilon <- rnorm(n)
y <- beta1 + beta2 * x + epsilon
lm(y ~ x)
```

```
Coefficients:
(Intercept)            x
      1.886       -4.060
```

Question: How can you use the above code to *simulate* the sampling distribution of b_2?

To simulate the sampling distribution, we need to run the above code many, many times, and collect the value of b_2 each time. Then, we can plot b_2 in a histogram, take the $\mathbb{E}[b_2]$, take the var$[b_2]$, etc. I set the *random seed* so that when you run the code on your computer, you get the same results as me. I use a `for` loop to repeat the experiment `nrep` $= 10000$ times:

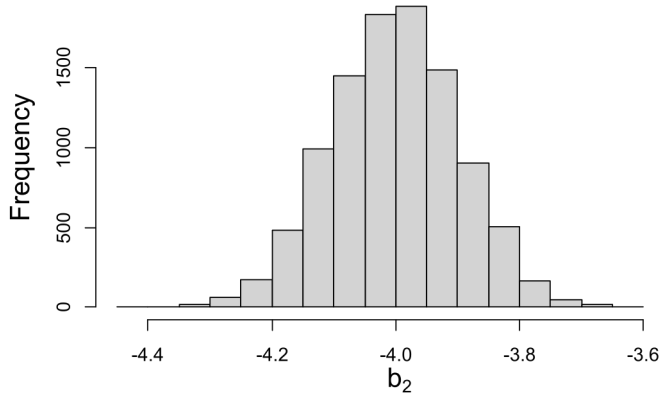
```
set.seed(7010)
nrep <- 10000
n <- 100
x <- rnorm(n)
beta1 <- 2
beta2 <- -4
b2 <- numeric(nrep)
for(i in 1:nrep){
  epsilon <- rnorm(n, mean=0, sd=1)
  y <- beta1 + beta2 * x + epsilon
  b2[i] <- lm(y ~ x)$coefficients[2]
}
mean(b2)
var(b2)
hist(b2)
```

```
> mean(b2)
[1] -4.000452
```

```
> var(b2)
[1] 0.01061148
```

The histogram of all 10,000 b_2 values is shown in Figure 5.1.

Figure 5.1: A simulated sampling distribution.



Questions:

1. What distribution appears to describe Figure 5.1?
2. Is the average value of the *estimator* close to the true *population* value?

5.1 Unbiased

Definition 5.1 — Unbiased estimator. An estimator, $\hat{\boldsymbol{\theta}}$, is an *unbiased* estimator of the parameter vector, $\boldsymbol{\theta}$, if:

$$\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$$

That is, if $\mathbb{E}[\hat{\boldsymbol{\theta}}(\boldsymbol{y})] = \boldsymbol{\theta}$, or $\int \hat{\boldsymbol{\theta}}(\boldsymbol{y}) p(\boldsymbol{y}|\boldsymbol{\theta}) d\boldsymbol{y} = \boldsymbol{\theta}$. The "Bias" of $\hat{\boldsymbol{\theta}}$ is the quantity:

$$\text{Bias}(\boldsymbol{\theta}, \boldsymbol{y}) = \mathbb{E}[\widehat{\boldsymbol{\theta}}(\boldsymbol{y}) - \boldsymbol{\theta}]$$

In words, an estimator is *unbiased* if it gives the right answer on average.

Example 5.1 — Unbiasedness of $\bar{y}$ Let $\{y_1, y_2, \ldots, y_n\}$ be a random sample from population with a finite mean, μ, and a finite variance, σ^2. Consider the *statistic*:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Then,

$$\mathbb{E}[\bar{y}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} y_i\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left(y_i\right) = \frac{1}{n} \sum_{i=1}^{n} \mu = \left(\frac{1}{n} n\mu\right) = \mu$$

So, $\bar{y}$ is an *unbiased estimator* of the parameter, μ. Here, there are lots of possible unbiased estimators

of μ. So, we will need to consider additional characteristics of estimators to help us choose from among them.

Return to our LS problem:

$$\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y}$$

Recall assumption A.5. We will use the strongest version of this assumption - that X is non-random. We could use the weaker versions of the assumption, and we'll get the same results, but the notation will be more cumbersome. Now, take the expected value of the random estimator $\boldsymbol{b}$:

$$\begin{aligned}\mathbb{E}(\boldsymbol{b}) &= \mathbb{E}\left[(X'X)^{-1} X'\boldsymbol{y}\right] = (X'X)^{-1} X'\mathbb{E}(\boldsymbol{y}) \\ &= (X'X)^{-1} X'\mathbb{E}[X\boldsymbol{\beta} + \boldsymbol{\epsilon}] = (X'X)^{-1} X'[X\boldsymbol{\beta} + \mathbb{E}(\boldsymbol{\epsilon})] \\ &= (X'X)^{-1} X'[X\boldsymbol{\beta} + \mathbf{0}] = (X'X)^{-1} X'X\boldsymbol{\beta} \\ &= \boldsymbol{\beta}\end{aligned} \tag{5.1}$$

This proves that:

The LS estimator of $\boldsymbol{\beta}$ is Unbiased.

5.2 Linear

Definition 5.2 — Linear estimator. Any estimator that is a *linear function* of the random sample data is called a *Linear Estimator*.

Example 5.2 — Sample average is linear. Let $\{y_1, y_2, \ldots, y_n\}$ be a random sample from population with a finite mean, μ, and a finite variance, σ^2. Consider the statistic:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\left[y_1 + y_2 + \cdots + y_n\right]$$

This statistic is a linear estimator of μ (note that the "weights" are non-random).

Return to our LS problem:

$$\underset{(k\times 1)}{\boldsymbol{b}} = (X'X)^{-1} X'\boldsymbol{y} = \underset{(k\times n)(n\times 1)}{A\boldsymbol{y}}$$

Note that, under our (strictest form of the) assumptions, A is a *non-random* matrix. So,

$$\begin{pmatrix} b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{bmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ldots & \vdots \\ a_{k1} & \ldots & a_{kn} \end{bmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

For example, $b_1 = [a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n]$, etc.

Thus:

The LS estimator, $\boldsymbol{b}$, is a linear (and unbiased) estimator of $\boldsymbol{\beta}$.

5.3 Efficient

Now let's consider the dispersion (variability) of $\boldsymbol{b}$, as an estimator of $\boldsymbol{\beta}$.

Definition 5.3 Suppose we have an $(n \times 1)$ random vector, $\boldsymbol{x}$. Then the *covariance matrix* of $\boldsymbol{x}$ is defined as the $(n \times n)$ matrix:

$$V(\boldsymbol{x}) = \mathbb{E}\left[(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))(\boldsymbol{x} - \mathbb{E}(\boldsymbol{x}))'\right]$$

- Diagonal elements of $V(\boldsymbol{x})$ are $\text{var}(x_1), \ldots, \text{var}(x_n)$.
- Off-diagonal elements are $\text{cov}(x_i, x_j)$; $i, j = 1, \ldots, n$; $i \neq j$.
- We have already made use of the *covariance matrix* when we made assumption A.4.

Return to our LS problem. We have a $(k \times 1)$ random vector, $\boldsymbol{b}$, and we know that $\mathbb{E}(\boldsymbol{b}) = \boldsymbol{\beta}$. The *covariance matrix* of $\boldsymbol{b}$, $V(\boldsymbol{b})$, is:

$$\begin{aligned} V(\boldsymbol{b}) &= \mathbb{E}\left[(\boldsymbol{b} - \mathbb{E}(\boldsymbol{b}))(\boldsymbol{b} - \mathbb{E}(\boldsymbol{b}))'\right] \\ &= \mathbb{E}\left[(\boldsymbol{b} - \boldsymbol{\beta})(\boldsymbol{b} - \boldsymbol{\beta})'\right] \end{aligned}$$

Now,

$$\begin{aligned} \boldsymbol{b} &= (X'X)^{-1} X'\boldsymbol{y} = (X'X)^{-1} X'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (X'X)^{-1} (X'X)\, \boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon} \\ &= I\boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon} \end{aligned}$$

So,

$$(\boldsymbol{b} - \boldsymbol{\beta}) = (X'X)^{-1} X'\boldsymbol{\epsilon} \tag{5.2}$$

Using the result from 5.2 in $V(\boldsymbol{b})$ we have:

$$\begin{aligned} V(\boldsymbol{b}) &= \mathbb{E}\left\{\left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right]\left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right]'\right\} \\ &= (X'X)^{-1} X'\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] X\, (X'X)^{-1} \end{aligned}$$

For assumption A.4, we showed earlier that because $\mathbb{E}(\boldsymbol{\epsilon}) = \mathbf{0}$:

$$V(\boldsymbol{\epsilon}) = \mathbb{E}\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right) = \sigma^2 I_n$$

So, we have:

$$\begin{aligned} V(\boldsymbol{b}) &= (X'X)^{-1} X'\mathbb{E}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right] X\, (X'X)^{-1} \\ &= (X'X)^{-1} X'\sigma^2 I X\, (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1} (X'X)\, (X'X)^{-1} \\ &= \sigma^2 (X'X)^{-1}. \end{aligned}$$

So, the covariance matrix of $\boldsymbol{b}$ is:

$$V(\boldsymbol{b}) = \sigma^2 (X'X)^{-1} \tag{5.3}$$

Question: What is the interpretation of the diagonal and off-diagonal elements of this matrix? What might the elements of this matrix be used for, in practice?

Finally, because the error term, $\boldsymbol{\epsilon}$ is assumed to be Normally distributed,

1. $y = X\boldsymbol{\beta} + \epsilon$: this implies that $\boldsymbol{y}$ is also Normally distributed.
2. $\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y} = A\boldsymbol{y}$: this implies that $\boldsymbol{b}$ is also Normally distributed.

Question: Why does the Normality of $\boldsymbol{\epsilon}$ transfer to $\boldsymbol{b}$?

We now have the full *sampling distribution* of the LS estimator, $\boldsymbol{b}$:

$$\boldsymbol{b} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1}\right]$$

Note:

- This result depends on our *rigid* assumptions about the various components of the regression model.
- The Normal distribution here is a "multivariate Normal" distribution.
- As with estimation of the population mean μ (see Example 5.1), there are lots of other unbiased estimators of $\boldsymbol{\beta}$.

Question: How might we choose between the many possible linear and unbiased estimators for $\boldsymbol{\beta}$? Why is linearity desirable?

We need to consider other desirable properties that these unbiased estimators may have. One option to help discern the "best" estimator for $\boldsymbol{\beta}$ is to take into account the estimators' *precisions*.

Definition 5.4 Suppose that we have two different *unbiased* estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, of the (scalar) parameter, θ. Then we say that $\widehat{\theta}_1$ is **at least as efficient** as $\widehat{\theta}_2$ if $\text{var}\left(\widehat{\theta}_1\right) \leq \text{var}\left(\widehat{\theta}_2\right)$

Note:

- The variance of an estimator is just the variance of its sampling distribution.
- "Efficiency" is a relative concept.

Question: What if there are 3 or more unbiased estimators being compared?

5.3.1 Mean Squared Error (MSE)

What if one or more of the estimators being compared is biased? In this case we can take account of both variance, and any bias, at the same time by using "mean squared error" (MSE) of the estimators.

Definition 5.5 Suppose that $\widehat{\theta}$ is an estimator of the (scalar) parameter, θ. Then the MSE of $\widehat{\theta}$ is defined as:

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

Note that:

$$\text{MSE}(\widehat{\theta}) = \text{var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

To prove this, write:

$$\text{MSE}(\widehat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}] + \mathbb{E}[\hat{\theta}] - \theta\right)^2\right],$$

expand out, and note that $\mathbb{E}[\mathbb{E}(\hat{\theta})] = \mathbb{E}(\hat{\theta})$ and $\mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})] = 0$. The above expression makes it clear that if the estimator is unbiased, the MSE equals the variance.

Definition 5.6 Suppose we have two (possibly) *biased* estimators, $\widehat{\theta}_1$ and $\widehat{\theta}_2$, of the (scalar) parameter, θ. Then we say that $\widehat{\theta}_1$ is **at least as efficient** as $\widehat{\theta}_2$ if $\text{MSE}\left(\widehat{\theta}_1\right) \leq \text{MSE}\left(\widehat{\theta}_2\right)$.

5.3.2 Efficiency and MSE for a vector of estimators

If we extend all of this to the case where we have a vector of parameters, $\boldsymbol{\theta}$, then we have the following definitions:

Definition 5.7 Suppose that we have two different *unbiased* estimators, $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\widehat{\boldsymbol{\theta}}_1$ is **at least as efficient** as $\widehat{\boldsymbol{\theta}}_2$ if $\Delta = V\left(\widehat{\boldsymbol{\theta}}_2\right) - V\left(\widehat{\boldsymbol{\theta}}_1\right)$ is *positive semi-definite*.

Definition 5.8 Suppose that we have two different (possibly) *biased* estimators, $\widehat{\boldsymbol{\theta}}_1$ and $\widehat{\boldsymbol{\theta}}_2$, of the parameter vector, $\boldsymbol{\theta}$. Then we say that $\widehat{\boldsymbol{\theta}}_1$ is **at least as efficient** as $\widehat{\boldsymbol{\theta}}_2$ if $\Delta = \text{MMSE}\left(\widehat{\boldsymbol{\theta}}_2\right) - \text{MMSE}\left(\widehat{\boldsymbol{\theta}}_1\right)$ is *positive semi-definite*.

Note: $\text{MMSE}(\widehat{\boldsymbol{\theta}}) = \mathbb{E}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})'\right] = V[\widehat{\boldsymbol{\theta}}] + \text{Bias}(\widehat{\boldsymbol{\theta}})\,\text{Bias}(\widehat{\boldsymbol{\theta}})'$.

5.3.3 Gauss-Markhov theorem

Taking account of its *linearity*, *unbiasedness*, and its *precision*, in what sense is the LS estimator, $\boldsymbol{b}$, of $\boldsymbol{\beta}$, optimal?

Gauss-Markhov Theorem: In the "standard" linear regression model, $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, the LS estimator, $\boldsymbol{b}$, of $\boldsymbol{\beta}$ is **Best Linear Unbiased** (BLU). That is, it is *efficient* in the class of all linear and unbiased estimators of $\boldsymbol{\beta}$.

Question: Why is this an interesting result?

Proof: Let $\boldsymbol{b}_0$ be any other *linear* estimator of $\boldsymbol{\beta}$:

$$\boldsymbol{b}_0 = C\boldsymbol{y} \quad ; \quad \text{for some non-random } C$$

The *covariance matrix* for $\boldsymbol{b}_0$ is:

$$V(\boldsymbol{b}_0) = CV(\boldsymbol{y})C' = C\left(\sigma^2 I_n\right)C' = \sigma^2 CC'$$

Take the difference between $\boldsymbol{b}_0$ and $\boldsymbol{b}$:

$$\boldsymbol{b}_0 - \boldsymbol{b} = C\boldsymbol{y} - (X'X)^{-1}X'\boldsymbol{y} = D\boldsymbol{y},$$

where

$$D = C - (X'X)^{-1}X'$$

is the difference between how the other estimator uses the X data to "weight" the y data (C), and how the LS estimator uses the X data ($(X'X)^{-1}X'$). Now restrict $\boldsymbol{b}_0$ to be unbiased, so that:

$$\mathbb{E}(\boldsymbol{b}_0) = \mathbb{E}(C\boldsymbol{y}) = CX\boldsymbol{\beta} = \boldsymbol{\beta}.$$

This requires that $CX = I$, which in turn implies that:

$$DX = \left[C - (X'X)^{-1}X'\right]X = CX - I = 0 \quad (\text{and } X'D' = 0)$$

Solve for C in terms of D:

$$C = D + (X'X)^{-1} X',$$

and return to the covariance matrix of $\boldsymbol{b}_0$:

$$\begin{aligned} V(\boldsymbol{b}_0) &= \sigma^2 CC' \\ &= \sigma^2 \left[D + (X'X)^{-1} X'\right] \left[D + (X'X)^{-1} X'\right]' \\ &= \sigma^2 \left[DD' + (X'X)^{-1} X'X (X'X)^{-1}\right] \quad ; \quad DX = X'D' = 0 \\ &= \sigma^2 DD' + \sigma^2 (X'X)^{-1} \\ &= \sigma^2 DD' + V(\boldsymbol{b}) \end{aligned}$$

or:

$$[V(\boldsymbol{b}_0) - V(\boldsymbol{b})] = \sigma^2 DD' \quad ; \quad \sigma^2 > 0 \tag{5.4}$$

Now we just have to "sign" this (matrix) difference:

$$\boldsymbol{\eta}' (DD') \boldsymbol{\eta} = (D'\boldsymbol{\eta})' (D'\boldsymbol{\eta}) = v'v = \sum_{i=1}^{n} v_i^2 \geq 0$$

So, $\Delta = [V(\boldsymbol{b}_0) - V(\boldsymbol{b})]$ is a p.s.d. matrix, implying that $\boldsymbol{b}_0$ is relatively less efficient than $\boldsymbol{b}$. Result:

The LS estimator is the Best Linear Unbiased estimator (BLUE) of $\boldsymbol{\beta}$.

Questions:

1. What assumptions did we use, and where?
2. Were there any standard assumptions that we *didn't* use?
3. What does this suggest?

5.4 Exercises

1. Explain why the LS estimator is a random variable.
   Answer. $\boldsymbol{b}$ is a linear function of $\boldsymbol{y}$, and $\boldsymbol{y}$ is a linear function of $\boldsymbol{\epsilon}$. Since $\boldsymbol{\epsilon}$ is random, so is $\boldsymbol{b}$.
2. Show that the LS estimator is linear.
   Answer. See Section 5.2.
3. Prove that the LS estimator is unbiased.
   Answer. See Section 5.1.
4. Derive the variance of the LS estimator, stating any assumptions that you use.
   Answer. See the derivation of equation 5.3 in Section 5.3.
5. Explain the result of the Gauss-Markov theorem, and why it is important.
   Answer. The Gauss-Markov theorem states that $\mathrm{V}(\hat{\boldsymbol{\beta}}) - \mathrm{V}(\boldsymbol{b})$ is a positive-semi-definite matrix, where $\hat{\boldsymbol{\beta}}$ is any linear and unbiased estimator for $\boldsymbol{\beta}$ and $\boldsymbol{b}$ is the LS estimator. That is, the LS estimator has the lowest variance out of the entire class of linear and unbiased estimators for $\boldsymbol{\beta}$. This is important because there are many such possible estimators. The G-M theorem explains a reason why LS might be the "best" choice among them.
6. Suppose that we have our usual linear model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

but that we partition the X matrix and $\boldsymbol{\beta}$ vector and write the model as:

$$\boldsymbol{y} = X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}.$$

All of the usual assumptions are satisfied, except that $\mathbb{E}[\boldsymbol{\epsilon}] = X_1\boldsymbol{\gamma}$. That is, the mean vector for the disturbances is a linear combination of a subset of the regressors (implying that $\boldsymbol{\epsilon}$ and X_1 are correlated). Let $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ be the OLS estimators for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Obtain the expressions for $\mathbb{E}[\boldsymbol{b}_1]$ and $\mathbb{E}[\boldsymbol{b}_2]$, and interpret your results.
Answer. The LS estimator for $\boldsymbol{\beta}_1$ is:

$$\boldsymbol{b}_1 = \left(X_1'M_2X\right)^{-1} X_1'M_2\boldsymbol{y},$$

where $M_2 = I - X_2\left(X_2'X_2\right)^{-1}X_2'$. Substituting in the population model for $\boldsymbol{y}$, and expanding, we have:

$$\begin{aligned}\boldsymbol{b}_1 &= \left(X_1'M_2X_1\right)^{-1} X_1'M_2\left(X_1\boldsymbol{\beta}_1 + X_2\boldsymbol{\beta}_2 + \epsilon\right)\\ &= \left(X_1'M_2X_1\right)^{-1} X_1'M_2X_1\boldsymbol{\beta}_1 + \left(X_1'M_2X_1\right)^{-1} X_1'M_2X_2\boldsymbol{\beta}_2\\ &\quad + \left(X_1'M_2X_1\right)^{-1} X_1'M_2\epsilon\end{aligned}$$

Using $M_2X_2 = 0$, and simplifying, we get:

$$\boldsymbol{b}_1 = \boldsymbol{\beta}_1 + 0 + \left(X_1'M_2X_1\right)^{-1} X_1'M_2\epsilon$$

Finally, taking the expectation of $\boldsymbol{b}_1$ and using $\mathbb{E}\left[\epsilon\right] = X_1\gamma$ we get:

$$\mathbb{E}\left[\boldsymbol{b}_1\right] = \boldsymbol{\beta}_1 + 0 + \gamma$$

We will get a similar result for $\mathbb{E}\left[\boldsymbol{b}_2\right]$, except that the third term will cancel since $M_1X_1 = 0$:

$$\mathbb{E}\left[\boldsymbol{b}_2\right] = 0 + \boldsymbol{\beta}_2 + \left(X_2M_1'X_2\right)^{-1} X_2'M_1X_1\gamma = \boldsymbol{\beta}_2$$

A critical assumption for the unbiasedness of the OLS estimator is that the regressors are unrelated to the error term (assumption A.5). The results above show that if some of the regressors (X_1 for example) are related to the error term, then only those corresponding LS estimators will be biased ($\boldsymbol{b}_1$). $\boldsymbol{b}_2$ remains unbiased as long as X_2 is unrelated to ϵ.

7. The least absolute deviations (LAD) estimator is an alternative to the LS estimator. LAD is less sensitive than LS to the presence of "outliers", since LAD minimizes the sum of the *absolute* residuals instead of the *squared* residuals. To perform a LAD regression in R, first install and load the `L1pack` package:

```
install.packages("L1pack")
library(L1pack)
```

Now, take the same R code used to generate Figure 5.1, but add the following line *before* the `for` loop:

```
b2.lad <- numeric(n)
```

and add the following line *inside* the `for` loop:

```
b2.lad[i] <- lad(y ~ x)$coefficient[2]
```

Run the simulation. Use the simulated distributions to answer the following questions: (i) Is the LAD estimator unbiased? (ii) How does the variance of the LAD estimator compare to that of the LS estimator?

Chapter 6

Simple Hypothesis Testing

An underlying principal of hypothesis testing is to compare an estimated value (such as b_j) to a hypothesized value (denoted $\beta_{j,0}$), and assess the plausibility of the hypothesis by taking the variance of the estimator into account.

In general, a hypothesis test begins with a null hypothesis, and an alternative hypothesis:

$$H_0 : \beta_j = \beta_{i,0}$$
$$H_A : \beta_j \neq \beta_{i,0}$$

H_0 is the null hypothesis. The hypothesized value of β_i is denoted $\beta_{i,0}$. The alternative hypothesis is denoted by H_A. One of the two situations must occur. This is called a "two-sided" hypothesis test: the null hypothesis is wrong if β_i is either "too small" or "too big" relative to the hypothesized value.

The hypothesis test concludes with either: (i) "reject" H_0 in favour of H_A, or (ii) "fail to reject" H_0. We should never say that we "accept" either of the hypotheses: we either have evidence to reject H_0, or we do not have enough evidence to reject H_0.

The decision to "reject" or "fail to reject" H_0 may begin by the researcher *subjectively* deciding on a *significance level* and then doing one or more of the following:

- Calculating a (p-value) and comparing it to the significance level.
- Seeing whether or not $\beta_{j,0}$ is contained in a confidence interval.
- Calculating a test statistic and seeing if it exceeds a critical value.

To proceed, we must first have an *estimator* for the *standard error of the estimator* being used to assess the hypothesis. For example, the t-test statistic for testing:

$$H_0 : \beta_i = 0$$
$$H_A : \beta_i \neq 0$$

is:

$$t = \frac{b_i}{\hat{se}.(b_i)}$$

We need this quantity $\hat{se}.(b_i)$, which is called the *estimated standard error*.

6.1 Estimating σ^2

We now know a lot about estimating $\boldsymbol{\beta}$. There's another parameter in the regression model, σ^2 – the variance of each ϵ_i. You are likely familiar with a t-test or an F-test. In order to perform a hypothesis test, using a t-test statistic for example, we need to estimate σ^2 so that we can get an estimate for the covariance matrix of the LS estimator: $V(\boldsymbol{b}) = \sigma^2 (X'X)^{-1}$.

Let's derive an estimator for σ^2. Begin by noting that

$$\sigma^2 = \text{var}(\epsilon_i) = \mathbb{E}\left[(\epsilon_i - \mathbb{E}(\epsilon_i))^2\right] = \mathbb{E}\left(\epsilon_i^2\right),$$

due to assumption A.3. The sample counterpart to this population parameter (σ^2) is the sample average of the "residuals":

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\boldsymbol{e}'\boldsymbol{e},$$

which is the method of moments, and the maximum likelihood estimator. However, there is a distortion in this estimator of σ^2. Although the mean of the e_i's is zero (if there is an intercept in the model), not all of e_i's are independent of each other: only $(n-k)$ of them are.

We should consider what properties $\hat{\sigma}^2$ has as an estimator of σ^2, before we use it. Is this a *good* estimator? What properties of the LS estimator did we evaluate? We will write $\boldsymbol{e}'\boldsymbol{e}$ in terms of only $\boldsymbol{\epsilon}$ (which we have made assumptions about), and then derive its expected value:

$$\boldsymbol{e} = (\boldsymbol{y} - \widehat{\boldsymbol{y}}) = (\boldsymbol{y} - X\boldsymbol{b}) = \boldsymbol{y} - X\left(X'X\right)^{-1}X'\boldsymbol{y} = M\boldsymbol{y}$$

where

$$M = I_n - X\left(X'X\right)^{-1}X' \quad ; \quad \text{idempotent, and } MX = \mathbf{0}$$

So,

$$\boldsymbol{e} = M\boldsymbol{y} = M(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = M\boldsymbol{\epsilon}$$

and

$$\boldsymbol{e}'\boldsymbol{e} = (M\boldsymbol{\epsilon})'(M\boldsymbol{\epsilon}) = \boldsymbol{\epsilon}'M\boldsymbol{\epsilon} \quad ; \quad \text{a scalar}$$

From this, it can be shown that:

$$\begin{aligned} E\left(\boldsymbol{e}'\boldsymbol{e}\right) = E\left[\boldsymbol{\epsilon}'M\boldsymbol{\epsilon}\right] &= E\left[\text{tr}\left(\boldsymbol{\epsilon}'M\boldsymbol{\epsilon}\right)\right] = E\left[\text{tr}\left(M\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\right] \\ &= \text{tr}\left[ME\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right)\right] = \text{tr}\left[M\sigma^2 I_n\right] = \sigma^2\,\text{tr}(M) \\ &= \sigma^2(n-k) \end{aligned}$$

We will not cover the trace operator (tr), we will not discuss why $\text{tr}(M) = \sigma^2(n-k)$, and you do not need to know how to obtain this expectation. However, you need to be aware that an important step in considering whether an estimator should be used is to examine its *bias*, and in the case of $\hat{\sigma}^2$:

$$E\left[\hat{\sigma}^2\right] = E\left[\frac{1}{n}\boldsymbol{e}'\boldsymbol{e}\right] = \frac{1}{n}(n-k)\sigma^2 < \sigma^2$$

The method of moments and maximum likelihood estimator, $\hat{\sigma}^2$, is *biased*.

It is easy to convert this biased estimator to an *unbiased* one:

$$s^2 = \frac{1}{(n-k)}e'e$$

Some notes:

- $(n-k)$ is the "degrees of freedom" – number of independent sources of information in the n residuals (the e_i's).
- We can use s as an estimator of σ, but it is a biased estimator. Even though it is biased, s is typically used in practice as the bias is small.
- s is called the "standard error of the regression", or the "standard error of estimate".
- s^2 is a statistic. It has its own sampling distribution, etc.

Let's see one immediate application of s^2 and s. Recall the sampling distribution for the LS estimator, $\boldsymbol{b}$:

$$\boldsymbol{b} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1}\right]$$

So, the variance of the i^{th} LS estimator is the i^{th} diagonal of the covariance matrix of $\boldsymbol{b}$: $\text{var}\left(b_i\right) = \sigma^2 \left[\left(X'X\right)^{-1}\right]_{ii}$, but σ^2 is *unobservable*. If we want to report the variability associated with b_i as an estimator of $\boldsymbol{\beta}_i$, we need to use an estimator of σ^2. The estimated variance of the i^{th} LS estimator is then:

$$\widehat{\text{var}\left(b_i\right)} = s^2 \left[\left(X'X\right)^{-1}\right]_{ii}$$

The square-root of the above is called the "standard error" of b_i. This quantity will be very important when it comes to constructing *interval estimates* of our regression coefficients, and when we construct *tests of hypotheses* about these coefficients.

Example 6.1 — Standard errors in R. Start by setting the random seed and sample size, and then generate some data:

```
set.seed(7010)
n <- 10
x <- rnorm(n)
y <- rnorm(n)
```

Estimate and summarize a model:

```
mod <- lm(y ~ x)
summary(mod)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.2472     0.3734   0.662    0.526
x            -0.2455     0.6348  -0.387    0.709

Residual standard error: 1.178 on 8 degrees of freedom
Multiple R-squared:  0.01835,   Adjusted R-squared:  -0.1044
F-statistic: 0.1495 on 1 and 8 DF,  p-value: 0.7091
```

R is reporting the standard errors of b_1 and b_2 as 0.3734 and 0.6348 respectively. We will see one way how statistical packages can calculate these numbers. Start by getting the estimate s^2:

```
s2 <- sum(mod$residuals ^ 2) / (n - 2)
```

Note that $k = 2$ above. If we take the square root, we get the "residual standard error" reported in the R output above:

```
sqrt(s2)
```

```
[1] 1.178469
```

Next we will calculate the $V(\boldsymbol{b})$ matrix. Start by arranging the x data into a matrix (and take a look at the X matrix):

```
X <- matrix(c(rep(1, n), x), n, 2)
X
```

```
      [,1]       [,2]
 [1,]    1 -0.2214732
 [2,]    1  0.6051370
 [3,]    1  0.7208573
 [4,]    1 -0.2230900
 [5,]    1 -0.2662395
 [6,]    1 -1.0890823
 [7,]    1 -0.5655553
 [8,]    1  0.5330395
 [9,]    1  0.6225352
[10,]    1 -0.4755066
```

The $s^2(X'X)^{-1}$ matrix is then:

```
s2 * solve(t(X) %*% X)
```

```
            [,1]       [,2]
[1,] 0.13939930 0.01448197
[2,] 0.01448197 0.40297318
```

I have used the `solve()` function to find the inverse. Taking the square root of any of the diagonal elements gives the standard error reported in the `summary()` output above:

```
sqrt(s2 * solve(t(X) %*% X))[1, 1]
```

```
[1] 0.3733622
```

6.2 Hypothesis testing and confidence intervals

Hypothesis testing, and confidence intervals, allow us to quantify the uncertainty around our "point" estimate. For example, a value of $b_2 = 2.1$ million may seem very far from a value of $\beta_2 = 0$; but this all depends on the *variability* of the estimator b_2. In order to assess the variability of the estimator in relation to a point estimate, we'll need the full sampling distributions of both $\boldsymbol{b}$ and s^2. Note that assumption A.6 will be particularly important in what follows. Recall that:

$$\boldsymbol{b} \sim N\left[\boldsymbol{\beta}, \sigma^2 \left(X'X\right)^{-1}\right]$$

and that because the marginal distribution from the multivariate-Normal distribution is still Normal:

$$b_i \sim N\left[\beta_i, \sigma^2 \left(\left(X'X\right)^{-1}\right)_{ii}\right]$$

Suppose that your null hypothesis is $H_0 : \beta_{2,0} = 0$[1] (you think that x_2 has *zero effect* on y), but that the value you estimate for β_2 is $b_2 = 2,100,000$. What can guide you in your decision to reject or fail-to-reject H_0? One possibility is to use a p-value.

Definition 6.1 — p-value. The probability of obtaining an estimate more adverse to the null hypothesis, compared to the estimate just obtained, provided the null hypothesis is true.

[1] The 0 in the subscript (after the comma) is to denote that this is the value for the parameter under the *null hypothesis*.

That is, the p-value provides the probability that "things could be worse", in terms of the currently observed discrepancy between the null and the estimated value. If this probability is subjectively large, then that would indicate that the we should *fail-to-reject* the null hypothesis (and *vice versa*).

What is the p-value for the above example? $Pr(b_2 > 2.1\text{mil})$ can be easily found using the Normal distribution, with a mean of $\beta_{2,0}$ and a variance of $\sigma^2 \left((X'X)^{-1}\right)_{ii}$. Then, if the *alternative hypothesis* is *two-sided*, $Pr(b_2 > 2.1\text{mil})$ is multiplied by two (most alternative hypotheses in econometrics are *two-sided*).

6.2.1 *z*-test statistic

Calculating a probability from the Normal distribution requires integrating under the Normal curve for some mean and variance, which presents a different integral for each hypothesis test, since the mean of the Normal distribution is determined by H_0 and the variance by the particular data set. Integrating under a Normal distribution is not an issue given modern computers, but it *was* an issue over a hundred years ago before the computer. In order to avoid having to integrate for each hypothesis test, instead we could: (i) *standardize* the estimator used in the null hypothesis so that it follows an $N \sim (0, 1)$ distribution; (ii) integrate many times under the standard Normal $N \sim (0, 1)$ distribution and tabulate some probabilities; (iii) obtain our p-value from this table. Note that this process is needlessly complicated given the modern computer, but is ingrained in statistics and still used.

Using properties of the mean and variance[2], we can *standardize* b_i:

$$z_i = \frac{(b_i - \beta_{i,0})}{\sqrt{\sigma^2 \left[(X'X)^{-1}\right]_{ii}}} \sim N(0, 1)$$

All we are accomplishing here is creating a *test statistic*; altering the distribution of b_i to one that is standard Normal, **provided the null hypothesis is true**. If H_0 is false, then the mean of z_i is not 0. Calculating a standardized test statistic is a very common way of obtaining a p-value, but it is not necessary. To see that standardization is not needed to obtain the p-value, note that $Pr(z_i > 0) = Pr(b_i > \beta_{i,0})$.

6.2.2 *t*-test statistic

The z-statistic typically can't be calculated in practice, since the value for σ^2 is usually unknown. We can't use z_i directly to draw inferences about b_i. Instead, we will have to replace the unknown σ^2 with an estimator: s^2 for example. The major issue is that when we introduce a random variable into the denominator of the z-statistic, we change the distribution to something that is *not* Normal. Given all of our assumptions A.1 to A.6, it turns out that the test statistics is transformed so that it follows the t-distribution. We will not fully prove this, but instead sketch out the proof.

- Definition: let $z_1, z_2, \ldots, z_m$ be independent $N(0, 1)$ random variables. Then the quantity $\sum_{i=1}^{m} (z_i^2)$ has a Chi-square distribution with m degrees of freedom, χ^2_m.
- Note: $\boldsymbol{e}'\boldsymbol{e} = \boldsymbol{\epsilon}' M \boldsymbol{\epsilon}$ is a sum of squared Normal variables.
- Note: not all of the $\boldsymbol{e}$ are independent, only $(n - k)$ of them. That is, the degrees of freedom in $\boldsymbol{e}$ is $(n - k)$.
- This leads to the distribution for s^2:

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$$

- Definition: let $z \sim N(0, 1)$, and let $x \sim \chi^2_v$, where z and x are independent. Then the statistic, $t = z/\sqrt{x/v}$ follows the Student's t-distribution, with "v" degrees of freedom.

[2] (i) $\mathbb{E}[c + Y] = c + \mathbb{E}[Y]$; (ii): $\text{var}[cY] = c^2 \text{var}[Y]$, where c is a constant and Y is a random variable

- Examine the formula for the t-statistic, and notice it has a Normal variable in the numerator (b_i) and the square root of a chi-square variable in the denominator (s^2). With some re-arranging, it can be shown that this statistic matches the definition for a t-distribution.

$$t_i = \frac{(b_i - \beta_{i,0})}{\sqrt{s^2 \left[(X'X)^{-1}\right]_{ii}}} = \frac{(b_i - \beta_{i,0})}{\widehat{s.e.(b_i)}} \sim t_{(n-k)}$$

- Finally, note that for large n, the t-distribution becomes the standard Normal distribution.

Example 6.2 — Simple hypothesis test. Suppose the estimated model is:

$$\begin{array}{rcccccc} \hat{y} & = & 1.4 & + & 0.25x_2 & + & 0.6x_3 \\ & & (0.7) & & (0.1) & & (1.4) \end{array}$$

with

$$H_0 : \beta_2 = 0 \quad \text{vs.} \quad H_A : \beta_2 \neq 0.$$

The t-statistic associated with this hypothesis test is:

$$t = \left[\frac{b_2 - \beta_2}{s.e.\,(b_2)}\right] = \left[\frac{0.25 - 0}{0.1}\right] = 2.5$$

Now we must determine the p-value, or compare this "2.5" to a *critical value*. Suppose that $n = 20$. Then, $t \sim t_{(17)}$. The p-value can be easily obtained from R using

```
2 * (1 - pt(2.5, 17))
```

```
[1] 0.02294781
```

What is the interpretation of 0.02295 here? If H_0 is true, there is only a 2.3% chance of obtaining a b_2 that is "further away" from 0 then what was just observed (a "distance" of 0.25). Things can't get much worse. Either: (i) we obtained a strange sample, or (ii) H_0 is false. With a *significance level* of $\alpha = 5\%$, we would reject H_0.

6.2.3 Critical values

We could also perform this hypothesis test using a *critical value*. Recall that for a significance level of 5%, the critical value from the Normal distribution is 1.96. This just means that if you obtain $|z_i| > 1.96$, you will obtain a p-value less than 5%. So, we *could* compare $t = 2.5$ to 1.96, except the sample size isn't large enough!

6.2.4 Confidence Intervals

We can also use our t-statistic to construct a confidence interval for β_i. Note that if H_0 is true, then the probability that the t-statistic lies within the α critical range is $(1 - \alpha)$:

$$\text{Pr.}\ [-t_c \leq t \leq t_c] = (1 - \alpha).$$

For example, if n is large and $\alpha = 0.05$, then there is a 0.95 probability that t lies within the values -1.96 and 1.96, provided H_0 is true. Now, substitute the formula for t into the above probability statement:

$$\text{Pr}.\left[-t_c \leq \left[\frac{b_i - \beta_i}{s.e.\,(b_i)}\right] \leq t_c\right] = (1 - \alpha),$$

and now solve the inequality so that β_i is in the centre:

$$\begin{aligned}
&\Pr.\left[-t_c \times s.e.\left(b_i\right) \leq \left(b_i - \beta_i\right) \leq t_c \times s.e.\left(b_i\right)\right] = (1-\alpha) \\
&\Pr.\left[-b_i - t_c \times s.e.\left(b_i\right) \leq \left(-\beta_i\right) \leq -b_i + t_c \times s.e.\left(b_i\right)\right] = (1-\alpha) \\
&\Pr.\left[b_i + t_c \times s.e.\left(b_i\right) \geq \beta_i \geq b_i - t_c \times s.e.\left(b_i\right)\right] = (1-\alpha) \\
&\Pr.\left[b_i - t_c \times s.e.\left(b_i\right) \leq \beta_i \leq b_i + t_c \times s.e.\left(b_i\right)\right] = (1-\alpha)
\end{aligned}$$

Interpretation of the confidence interval

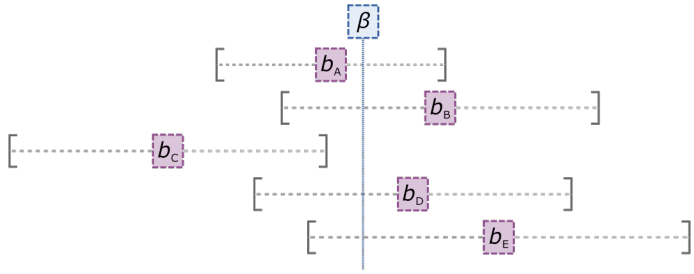
$$\left[b_i - t_c \times s.e.\left(b_i\right), \quad b_i + t_c \times s.e.\left(b_i\right)\right],$$

is *random*. The parameter, β_i is *fixed*, but unknown.

- If we were to take a sample of n observations, and construct such an interval, and then repeat this exercise many, many times, then $100(1-\alpha)\%$ of such intervals would cover the true value of β_i.
- Such an interval contains all the values for the parameter under the null hypothesis, that will not be rejected.

If we just construct an interval, for our given sample of data, we'll never know if this particular interval covers β_i, or not.

Figure 6.1: Each hypothetical sample of size n that we could draw (sample A, sample B, etc.) provides a 95% confidence interval that has a 95% probability of containing the true population β. In practice, we will only draw one sample from the population, and calculate one interval.



6.3 Some Properties of Tests

Classical hypothesis testing:

- Assume that H_0 is *true*
- Compute value of test statistic using random sample of data
- Determine distribution of the test statistic (when H_0 is true)
- Check if observed value of test statistic is likely to occur, if H_0 is true
- If this event is sufficiently unlikely, then reject H_0 (in favour of H_A)

Note:

1. Can never accept H_0. Why not?
2. What constitutes "unlikely" – subjective?
3. There are two types of errors we might incur with this process.

Definition 6.2 — Type I and II error.

- Type I Error: **Reject** H_0 when in fact it is **true**.

- Type II Error: **Do not reject** H_0 when in fact it is **false**.

The probability of a type I error is denoted by α, and is also the *significance* level of the test (sometimes also called the "size" of the test). The significance level of the test is a predetermined maximum p-value before which the null hypothesis is rejected. That is, if the p-value is less than α, H_0 is rejected. In deciding on this maximum acceptable p-value, we are also determining the type I error. Even when the null hypothesis is true, there is an $\alpha\%$ probability of drawing an "extreme" sample that will lead to a incorrect rejection of H_0.

The probability of a type II error is sometimes denoted by β, but this is less useful of a construct than α. This is because β typically will not be known. Why is this? β will depend on *how* H_0 is false. Usually, there are many ways. For example, H_0 could be false because the true parameter value is very far away from the value under the null; or it could be false because the truth is only a little bit different from the null. β will be smaller in the former situation relative to the latter situation.

Although the β of a test is typically unknown and not useful in practice, the concept is theoretically useful for designing or *choosing* a testing procedure. You are likely familiar with the t-test and F-test, but why do we use these tests? It is because, under certain assumptions, these tests have *desirable properties*. Similar to how we want to use an estimator that is unbiased and efficient, we want to use a test that is *powerful*, *unbiased*, and *consistent*. Note, however, that the properties *unbiased* and *consistent* have different meanings depending on whether we are talking about a *test* or an *estimator*.

Definition 6.3 — Power. The "power" of a test is $Pr.[\text{Reject } H_0 | H_0 \text{ is false}]$. So, Power $= 1 - Pr.[\text{Do not reject } H_0 | H_0 \text{ is false}] = 1 - \beta$.

Depending on the *way* that H_0 is false, we typically have a **Power Curve**. For a fixed value of α, this curve plots Power against parameter value(s). We want our tests to have *high power*, and we want the power of our tests to *increase* as H_0 becomes *increasingly false*. Now let's consider some desirable properties for a test.

Property 1 - UMP

Consider a fixed sample size, n, and a fixed significance level, α. Then, a test is "Uniformly Most Powerful" if its power exceeds (or is no less than) that of any other test, for all possible ways that H_0 could be false.

Property 2 - consistent (test)

Consider a fixed significance level, α. Then, a test is "consistent" if its Power $\to 1$, as $n \to \infty$, for all possible ways that H_0 is false.

Property 3 - unbiased (test)

Consider a fixed sample size, n, and a fixed significance level, α. Then, a test is said to be "unbiased" if its power never falls below the significance level.

Property 4 - LMP

Consider a fixed sample size, n, and a fixed significance level, α. Then, a test is said to be "Locally Most Powerful" if the slope of its Power curve is greater than the slope of the power curves of all other size α tests, in a neighbourhood of H_0.
Note:

- For many testing problems, no UMP test exists. This is why LMP tests are important.
- Why do we use our "t-test" in the regression model? Because it has properties 1 - 3 against one-sided alternative hypotheses, and has properties 2 - 4 against two-sided alternatives. Similar to how the LS estimator is "best" among other estimation alternatives (given certain assumptions), so is the t-test "best" among other potential testing strategies.

6.4 Exercises

1. Given that $E\left[\hat{\sigma}^2\right] = \frac{(n-k)\sigma^2}{n}$, prove that s^2 is unbiased.
2. Explain why the t-statistic follows a t-distribution and not a Normal distribution.
3. Using the R output from Example 6.1, calculate the 95% confidence interval around the estimated slope parameter, and test the hypothesis that this parameter is equal to zero.
4. Which of our assumptions are required in order for the t-test to be *appropriate*?
5. What are some desirable *statistical properties* for a test to have?

Chapter 7

Asymptotic Properties of Various Estimators

So far our results apply for any finite sample size n. In more general models we often can't obtain exact results for estimators' properties (for example, models that are estimated via *maximum likelihood*. In these cases, we might instead consider the estimator's properties as $n \to \infty$, as a way of "approximating" results. Asymptotic properties of estimators are also of interest in their own right: inferential procedures should "work well" when we have lots of data. We have already seen one example of an asymptotic property: hypothesis tests that are "consistent".

Definition 7.1 — Weak consistency. An estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$, is said to be (weakly) *consistent* if

$$\lim_{n\to\infty}\left\{\Pr.\left[\left|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right| < \boldsymbol{\epsilon}\right]\right\} = 1$$

for all $\boldsymbol{\epsilon} > \mathbf{0}$.

A sufficient condition for *weak* consistency to hold is the stronger mean-square consistency:

Definition 7.2 — Mean-square consistency. An estimator, $\hat{\boldsymbol{\theta}}$, for $\boldsymbol{\theta}$, is said to be mean-square *consistent* if its bias and variance go to zero as n goes to infinity:

$$\begin{aligned}&\text{Bias}\left(\widehat{\boldsymbol{\theta}}_n\right) \to \mathbf{0};\ \text{as } n \to \infty,\\ &V\left(\widehat{\boldsymbol{\theta}}_n\right) \to 0;\ \text{as } n \to \infty.\end{aligned}$$

Mean-square consistency is often useful for checking consistency, since it is easier to prove than weak consistency.

If $\hat{\boldsymbol{\theta}}$ is weakly consistent for $\boldsymbol{\theta}$, we say that the "probability limit" of $\hat{\boldsymbol{\theta}}$ equals $\boldsymbol{\theta}$. We denote this by using the "plim" operator, and we write:

$$\text{plim}\left(\widehat{\boldsymbol{\theta}}_n\right) = \boldsymbol{\theta} \quad \text{or,} \quad \widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}$$

Loosely speaking, consistency means that, given an infinitely large sample of data, the estimator provides the true parameter value *exactly* (there is zero bias and variance).

Example 7.1 — Consistency of the sample average. Let $x_i \sim [\mu, \sigma^2]$ be a random i.i.d. variable.

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
E[\bar{x}] &= \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}(n\mu) = \mu \quad \text{(unbiased, for all } n\text{)} \\
\text{var}[\bar{x}] &= \frac{1}{n^2}\,\text{var}\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(x_i) \\
&= \frac{1}{n^2}\left(n\sigma^2\right) = \sigma^2/n
\end{aligned}$$

So, $\bar{x}$ is an unbiased estimator of μ, and $\lim_{n\to\infty}\{\text{var}[\bar{x}]\} = 0$. This implies that $\bar{x}$ is both a mean-square consistent, and a weakly consistent estimator of μ.

Note:

- If an estimator is inconsistent, then it is a pretty useless estimator!
- There are many situations in which our LS estimator is inconsistent! For example, when:
  - there is an "omitted" variable that is relevant, and that is correlated to a variable included in the estimated model;
  - there is *simultaneous causality*;
  - a time series model includes lags of the dependent variable, and the errors are *autocorrelated*.

7.1 Slutsky's Theorem

Let $\text{plim}\left(\widehat{\boldsymbol{\theta}}_n\right) = \boldsymbol{c}$, and let $f(.)$ be any continuous function. Then, $\text{plim}\left[f\left(\widehat{\boldsymbol{\theta}}_n\right)\right] = f(\boldsymbol{c})$. For example:

$$\text{plim}\left(\frac{1}{\widehat{\theta}}\right) = \frac{1}{c}$$

where $\widehat{\theta}$ and c are scalars;

$$\text{plim}\left(e^{\hat{\boldsymbol{\theta}}}\right) = e^{\boldsymbol{c}}$$

where $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{c}$ are vectors;

$$\text{plim}\left(\hat{\Theta}^{-1}\right) = C^{-1}$$

where $\hat{\Theta}$ and C are matrices. Slutsky's Theorem is a very useful result: the "plim" operator can be used very flexibly.

7.2 Asymptotic Properties of LS Estimator

Consider LS estimator of β under our standard assumptions, in the "large n" asymptotic case.

- Can relax some assumptions:
  - i Don't need Normality assumption for the error term of our model.
  - ii Columns of X can be random, just assume that $\{\boldsymbol{x}_i', \epsilon_i\}$ is a random and *independent* sequence; $i = 1, 2, 3, \ldots$.
  - iii The above assumption implies that $\text{plim}\left[n^{-1}X'\epsilon\right] = \mathbf{0}$.

- Amend (extend) our assumption about X having full column rank. Assume instead that plim $\left[n^{-1}X'X\right] = Q$, where Q is a finite, positive-definite and symmetric $(k \times k)$ matrix that is *unobservable*.

Question: In words, what are we assuming about the elements of X, as n increases without limit?

Theorem: The LS estimator of β is weakly consistent.
Proof:

$$\begin{aligned}\boldsymbol{b} = (X'X)^{-1}X'\boldsymbol{y} &= (X'X)^{-1}X'(X\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (X'X)X'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + \left[\frac{1}{n}(X'X)\right]^{-1}\left[\frac{1}{n}X'\boldsymbol{\epsilon}\right].\end{aligned}$$

If we now apply Slutsky's Theorem repeatedly, we have:

$$\text{plim}(\boldsymbol{b}) = \boldsymbol{\beta} + Q^{-1}\mathbf{0} = \boldsymbol{\beta}$$

We can also show that s^2 is a consistent estimator for σ^2. There are at least two ways to do this (each uses different assumptions). First, assume the errors are Normally distributed, and get a strong result. We can also relax this assumption and get a weaker result.

Theorem: If the regression model errors are Normally distributed, then s^2 is a *mean-square* consistent estimator for σ^2.
Proof: If the errors are Normal, then we know that

$$\frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{(n-k)}$$

The mean and variance of a χ^2 distributed random variable are:

$$E\left[\chi^2_{(n-k)}\right] = (n-k)$$

$$\text{var}\left[\chi^2_{(n-k)}\right] = 2(n-k)$$

So,

$$E\left(s^2\right) = \frac{\sigma^2 E\left[\chi^2_{(n-k)}\right]}{n-k} = \sigma^2 \quad ; \quad \text{unbiased}$$

and

$$\text{var}\left[\frac{(n-k)s^2}{\sigma^2}\right] = 2(n-k)$$

$$\left[\frac{(n-k)^2}{\sigma^4}\right]\text{var}\left(s^2\right) = 2(n-k)$$

$$\text{var}\left(s^2\right) = 2\sigma^4/(n-k)$$

So, $\text{var}\left(s^2\right) \to 0$, as $n \to \infty$, and the estimator is unbiased. This implies that s^2 is a mean-square consistent estimator for σ^2. (This implies that it is also a weakly consistent estimator.)

- With the addition of the (relatively) strong assumption of Normally distributed errors, we get the (relatively) strong result.
- Note that $\hat{\sigma}^2 = (e'e)/n$ is also a consistent estimator, even though it is biased.

What can we say if we relax the assumption of Normality? We need a preliminary result to help us (Khintchine's theorem; or the Weak Law of Large Numbers).

7.2.1 Khinchin's Theorem; Weak Law of Large Numbers (WLLN)

Suppose that $\{x_i\}_{i=1}^n$ is a sequence of random variables that are uncorrelated, and all drawn from the same distribution with a finite mean, μ, and a finite variance, σ^2. Then, $\text{plim}(\bar{x}) = \mu$. Khinchin's theorem says that a sample average of i.i.d. variables is at least *weakly* consistent. We can use this result to establish the consistency of s^2.

Theorem 7.1 — Weak consistency of s^2. In our regression model, s^2 is a weakly consistent estimator for σ^2. (Notice that this also means that $\hat{\sigma}^2$ is a weakly consistent estimator, so start with the latter estimator.)

Proof:

$$\begin{aligned}
\hat{\sigma}^2 &= \left(\frac{e'e}{n}\right) = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \\
&= \frac{1}{n}(M\epsilon)'(M\epsilon) = \frac{1}{n}\epsilon' M \epsilon \\
&= \frac{1}{n}\left[\epsilon'\epsilon - \epsilon' X \left(X'X\right)^{-1} X'\epsilon\right] \\
&= \left[\left(\frac{1}{n}\epsilon'\epsilon\right) - \left(\frac{1}{n}\epsilon' X\right)\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\epsilon\right)\right]
\end{aligned}$$

So, $\text{plim}\left(\hat{\sigma}^2\right) = \text{plim}\left(\frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}\right) - \mathbf{0}'Q^{-1}\mathbf{0} = \text{plim}\left[\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\right]$ (if the errors are not autocorrelated, neither are the squared values). Also, $E\left[\epsilon_i^2\right] = \text{var.}(\epsilon_i) = \sigma^2$. By Khintchine's Theorem, we immediately have the result:

$$\text{plim}\left(\hat{\sigma}^2\right) = \sigma^2$$

and

$$\text{plim}\left(s^2\right) = \sigma^2$$

Relaxing the assumption of Normally distributed errors led to a weaker result for the consistent estimation of the error variance.

7.3 Asymptotic efficiency

Suppose we want to compare the (large n) asymptotic behaviour of our LS estimators with those of other potential estimators. These other estimators will presumably also be consistent. This means that

in each case the sampling distributions of the estimators collapse to a "spike", located exactly at the true parameter values. So, how can we compare such estimators when n is very large: aren't they indistinguishable? If the limiting density of any consistent estimator is a degenerate "spike", it will have zero variance, in the limit. Can we still compare large-sample variances of consistent estimators? In other words, is it meaningful to think about the concept of asymptotic efficiency?

The key to asymptotic efficiency is to "control" for the fact that the distribution of any consistent estimator is "collapsing", as $n \to \infty$.

The rate at which the distribution collapses is crucially important. This is probably best understood by considering an example.

Example 7.2 Let $\{x_i\}_{i=1}^n$ be a random sample from a population with mean and variance $[\mu, \sigma^2]$. We know from a previous example that:

$$E[\bar{x}] = \mu; \quad \text{var}[\bar{x}] = \sigma^2/n$$

Observe how $\text{var}[\bar{x}] \to 0$ as $n \to \infty$ (the sampling distribution collapses to a "spike" at the true parameter value). Now, construct: $y = \sqrt{n}(\bar{x} - \mu)$. Note that:

$$E(y) = \sqrt{n}(E(\bar{x}) - \mu) = 0$$

and

$$\text{var}[y] = (\sqrt{n})^2 \,\text{var}(\bar{x} - \mu) = n \,\text{var}(\bar{x}) = \sigma^2$$

The scaling we've used results in a finite, non-zero, variance. $E(y) = 0$, and $\text{var}[y] = \sigma^2$; *unchanged* as $n \to \infty$. So, $y = \sqrt{n}(\bar{x} - \mu)$ has a well-defined "limiting" (asymptotic) distribution. The asymptotic mean of y is zero, and the asymptotic variance of y is σ^2. Why did we scale by $\sqrt{n}$, and not (say), by n itself?

In fact, because we had independent x_i's (random sampling), we have the additional result that $y = \sqrt{n}(\bar{x} - \mu) \xrightarrow{d} N[0, \sigma^2]$, the Lindeberg-Lévy Central Limit Theorem.

Definition 7.3 Let $\hat{\theta}$ and $\tilde{\theta}$ be two consistent estimator of θ; and suppose that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \left[0, \sigma^2\right], \text{ and } \sqrt{n}(\tilde{\theta} - \theta) \xrightarrow{d} \left[0, \varphi^2\right].$$

Then $\hat{\theta}$ is "asymptotically efficient" relative to $\tilde{\theta}$ if $\sigma^2 < \varphi^2$. In the case where $\boldsymbol{\theta}$ is a vector, $\widehat{\boldsymbol{\theta}}$ is "asymptotically efficient" relative to $\widetilde{\boldsymbol{\theta}}$ if $\Delta = \text{asy}.V(\widetilde{\boldsymbol{\theta}}) - \text{asy}.V(\widehat{\boldsymbol{\theta}})$ is positive definite.

7.4 Asymptotic Distribution of the LS Estimator

Let's consider the full asymptotic distribution of the LS estimator, $\boldsymbol{b}$, for $\boldsymbol{\beta}$ in our linear regression model. We'll actually have to consider the behaviour of $\sqrt{n}(\boldsymbol{b} - \boldsymbol{\beta})$:

$$\begin{aligned}\sqrt{n}(\boldsymbol{b} - \boldsymbol{\beta}) &= \sqrt{n}\left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right] \\ &= \left[\frac{1}{n}(X'X)\right]^{-1} \left(\frac{1}{\sqrt{n}} X'\boldsymbol{\epsilon}\right)\end{aligned}$$

It can be shown, by the Lindeberg-Feller Central Limit Theorem, that

$$\left(\frac{1}{\sqrt{n}} X'\boldsymbol{\epsilon}\right) \xrightarrow{d} N\left[0, \sigma^2 Q\right]$$

where $Q = \text{plim}\left[\frac{1}{n}(X'X)\right]$. So, the asymptotic covariance matrix of $\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})$ is

$$\text{plim}\left[\frac{1}{n}(X'X)\right]^{-1}\left(\sigma^2 Q\right)\text{plim}\left[\frac{1}{n}(X'X)\right]^{-1} = \sigma^2 Q^{-1}.$$

In full, the asymptotic distribution of $\boldsymbol{b}$ is correctly stated by saying that:

$$\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta}) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 Q^{-1}\right]$$

The asymptotic covariance matrix is unobservable, for two reasons:

1. σ^2 is typically unknown.
2. Q is unobservable.

- We can estimate σ^2 consistently, using s^2.
- To estimate $\sigma^2 Q^{-1}$ consistently, we can use $ns^2(X'X)^{-1}$:

$$\text{plim}\left[ns^2(X'X)^{-1}\right] = \text{plim}\left(s^2\right)\text{plim}\left[\frac{1}{n}(X'X)\right]^{-1} = \sigma^2 Q^{-1}$$

The square roots of the diagonal elements of $ns^2(X'X)^{-1}$ are the asymptotic standard errors for the elements of $\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})$. Loosely speaking, the asymptotic covariance matrix for $\boldsymbol{b}$ itself is $s^2(X'X)^{-1}$; and the square roots of the diagonal elements of this matrix are the asymptotic standard errors for the b_i's themselves.

7.5 Exercises

1. Prove that the LS estimator is consistent, stating any assumptions that you use.
2. Provide an example of when the LS estimator is inconsistent.
3. $\tilde{\theta}$ and $\hat{\theta}$ are both unbiased estimators for θ. The variances of the two estimators are:

   $$\text{var}\left(\tilde{\theta}\right) = \frac{2\sigma^2}{n}$$

   and

   $$\text{var}\left(\hat{\theta}\right) = \frac{\sigma^2}{2n}$$

   (a) Are the estimators consistent?
   (b) What are the asymptotic distributions of the two estimators?
   (c) Which estimator is asymptotically more efficient?

4. Given that $\text{plim}(\hat{\sigma}^2) = \sigma^2$, prove that s^2 is also consistent.

Chapter 8

Instrumental Variables

We have been assuming either that the columns of X are non-random; or that the sequence $\{\boldsymbol{x}_i, \epsilon_i\}$ is independent. Often, neither of these assumptions are tenable. This implies that plim $\left(\frac{1}{n}X'\epsilon\right) \neq \mathbf{0}$, and then the LS estimator is inconsistent.

Prove that the LS estimator is inconsistent when plim $\left(\frac{1}{n}X'\epsilon\right) \neq \mathbf{0}$.

Inconsistency of the LS estimator is a serious issue. Inconsistency means that the estimation results can be wrong, and that no matter how large n is, the problem does not go away. If the LS estimator is inconsistent, it should not be used. In this chapter, we motivate the situation through a *missing variable* that is *correlated* with a variable(s) in the X matrix. Then, we discuss instrumental variable (IV) estimation as a potential solution to the problem.

8.1 Correlation between the error term and regressors

There are several ways in which to motivate the situation where X and $\boldsymbol{\epsilon}$ are correlated. Davidson and MacKinnon (2004) discuss "Errors in Variables" (pg. 312), and "Simultaneous Equations" (pg. 314). The very problem which motivated IV estimation is the simultaneity of price and quantity through demand and supply equations in a competitive market. For example, a linear model of demand and supply is:

$$q_i = \gamma_d p_i + \boldsymbol{x}_i^d \boldsymbol{\beta}_d + \epsilon_i^d \tag{8.1}$$

$$q_i = \gamma_s p_i + \boldsymbol{x}_i^s \boldsymbol{\beta}_s + \epsilon_i^s \tag{8.2}$$

where equation 8.1 is the demand function and equation 8.2 is the supply function, the x variables are predetermined or exogenous, and the γ are the slopes of the demand or supply curves. There are two equations in two unknowns (p_i and q_i), and it is easy to solve for the equilibrium values by, for example, solving for p_i in equation 8.2 and substituting into equation 8.1. In doing so, we see that the solutions for p_i and q_i depends not only on both x_i^d and x_i^s, but also on both ϵ_i^d and ϵ_i^s. In any linear simultaneous equations model, the endogenous variables are necessarily correlated with the error terms. This is a violation of A.5, leading to inconsistency of the LS estimator.

Another way to motivate a situation where X and $\boldsymbol{\epsilon}$ are dependent, is to consider omitted, unobservable, or *missing* variables, M. Consider also that these missing variables are correlated with the regressors X and the dependent variable $\boldsymbol{y}$. The true population model is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + M\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

Since M is unobservable, the observable model that we can estimate is:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{8.3}$$

Notice that in 8.3, $\boldsymbol{\epsilon}$ contains $M\boldsymbol{\gamma}$, so that X and $\boldsymbol{\epsilon}$ are not independent (X is endogenous), since X and M are correlated.

Prove that OLS is biased and inconsistent under this data generating process.

In such cases of endogeneity, we want a safe (consistent) way of estimating $\boldsymbol{\beta}$. One general family of such estimators is the family of Instrumental Variables (IV) estimators.

8.2 Instrumental variable

A variable(s), Z, qualifies as an instrument if it satisfies two conditions.

An instrumental variable, Z, must be:

1. Correlated with the endogenous variables X.
   - This is sometimes called the "relevance" of an IV.
   - This condition can be tested.
2. Uncorrelated with the error term, or equivalently, uncorrelated with the dependent variable other than through its correlation with X.
   - This is sometimes called the "exclusion" restriction.
   - This restriction cannot be easily tested.

The "exclusion" restriction implies k moment conditions, which allows us to derive the IV estimator:

$$E(Z'\boldsymbol{\epsilon}) = \mathbf{0}$$

or

$$E(Z'\boldsymbol{\epsilon}) = E(Z'\boldsymbol{y} - Z'X\boldsymbol{\beta}) = E(Z'\boldsymbol{y}) - E(Z'X)\boldsymbol{\beta} = 0$$

Solving the k moment conditions for $\boldsymbol{\beta}$, replacing the expected values with sample averages, yields the IV estimator:

$$\hat{\boldsymbol{b}}_{IV} = \left(n^{-1}\sum_{i=1}^{n}\mathbf{z}_i'\mathbf{x}_i\right)^{-1}\left(n^{-1}\sum_{i=1}^{n}\mathbf{z}_i'y_i\right)$$

or, in matrix notation:

$$b_{IV} = \left(Z'X\right)^{-1}Z'\boldsymbol{y}$$

In general, this estimator is biased. We can show it is consistent, however:

$$\begin{aligned}
&\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
&\operatorname{plim}\left(\tfrac{1}{n}X'X\right) = Q \quad ; \quad \text{p.d. and finite} \\
&\operatorname{plim}\left(\tfrac{1}{n}X'\boldsymbol{\epsilon}\right) = \gamma \neq 0
\end{aligned}$$

Full rank of the instrument matrix, the *relevancy* of the I.V., and the *exclusion restriction* imply respectively that :

$$\begin{aligned}
&\operatorname{plim}\left(\tfrac{1}{n}Z'Z\right) = Q_{ZZ} \quad ; \quad \text{p.d. and finite} \\
&\operatorname{plim}\left(\tfrac{1}{n}Z'X\right) = Q_{ZX} \quad ; \quad \text{p.d. and finite} \\
&\operatorname{plim}\left(\tfrac{1}{n}Z'\boldsymbol{\epsilon}\right) = \mathbf{0}
\end{aligned}$$

Then, the IV estimator is *consistent*:

$$\begin{aligned}\boldsymbol{b}_{IV} &= (Z'X)^{-1} Z'\boldsymbol{y} = (Z'X)^{-1} Z'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (Z'X)^{-1} Z'X\boldsymbol{\beta} + (Z'X)^{-1} Z'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + (Z'X)^{-1} Z'\boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + \left(\frac{1}{n}Z'X\right)^{-1}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right)\end{aligned}$$

and so:

$$\begin{aligned}\operatorname{plim}(\boldsymbol{b}_{IV}) &= \boldsymbol{\beta} + \left[\operatorname{plim}\left(\frac{1}{n}Z'X\right)\right]^{-1} \operatorname{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) \\ &= \boldsymbol{\beta} + Q_{ZX}^{-1}\mathbf{0} = \boldsymbol{\beta}\end{aligned}$$

Note that choosing different Z matrices generates different members of the IV family.

Although we won't derive the full asymptotic distribution of the IV estimator, note that it can be expressed as:

$$\sqrt{n}(\boldsymbol{b}_{IV} - \boldsymbol{\beta}) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right]$$

where $Q_{XZ} = Q'_{ZX}$.

Questions: How would you estimate this asymptotic covariance matrix? How would you estimate the covariance matrix for $\boldsymbol{b}_{IV}$?

8.3 Interpreting IV as two-stage least squares (2SLS)

IV estimation is also called two-stage least squares. Before modern computers, IV estimates were calculated by two (or more) least squares estimations. Two stage least squares offers an intuitive interpretation of the IV estimator. The idea behind IV estimation is that the instrument Z may be used to "extract" the "clean" variation from the endogenous variables X. When X is endogenous, a change in X is associated with a change in $\boldsymbol{\epsilon}$, so it is impossible to *identify* how much of the observed change in X led to the observed change in $\boldsymbol{y}$. However, if we could *extract* the variation in X that is uncorrelated with variation in $\boldsymbol{\epsilon}$, then we could use this *clean* variation to estimate $\boldsymbol{\beta}$ consistently. If we find an instrument Z that is correlated with X but uncorrelated with $\boldsymbol{\epsilon}$, then we can use changes in X *due to changes in* Z *only*, to extract this clean variation.

1st stage of 2SLS

In the first stage, we regress each column of X on Z using LS, and get $\hat{X}$. That is, we get $\hat{X} = P_Z X$.

- $\hat{X} = P_Z X$ contains the variation in X due to Z *only*.
- $P_Z X$ is not correlated with $\boldsymbol{\epsilon}$.
- Recall that $\hat{X} = P_Z X = Z(Z'Z)^{-1} Z'X$.

Question: Why is $P_Z X$ uncorrelated with $\boldsymbol{\epsilon}$?

2nd stage of 2SLS

In the second stage, we estimate the model: $y = \hat{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = P_Z X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, using LS. Applying the LS formula to this model, we get:

$$\boldsymbol{b}_{IV} = \left[X'Z(Z'Z)^{-1} Z'X\right]^{-1} X'Z(Z'Z)^{-1} Z'\boldsymbol{y}$$

or just:

$$\boldsymbol{b}_{IV} = \left[X' P_Z X\right]^{-1} X' P_Z \boldsymbol{y}$$

In fact, this is the *Generalized* I.V. estimator of $\boldsymbol{\beta}$. We can actually use more instruments than regressors (the "Over-Identified" case). Why might we want to do this? (Efficiency). Note that if X and Z have the same dimensions, then the generalized estimator collapses to the simple one.

Now, let's check the consistency of this Generalized IV estimator.

$$\begin{aligned}\boldsymbol{b_{IV}} &= \left[X' P_Z X\right]^{-1} X' P_Z \boldsymbol{y} = \left[X' P_Z X\right]^{-1} X' P_Z (X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \left[X' P_Z X\right]^{-1} X' P_Z X\boldsymbol{\beta} + \left[X' P_Z X\right]^{-1} X' P_Z \boldsymbol{\epsilon} \\ &= \boldsymbol{\beta} + \left[X' Z \left(Z'Z\right)^{-1} Z' X\right]^{-1} X' Z \left(Z'Z\right)^{-1} Z' \boldsymbol{\epsilon}\end{aligned}$$

So,

$$\boldsymbol{b_{IV}} = \boldsymbol{\beta} + \left[\left(\frac{1}{n} X'Z\right)\left(\frac{1}{n} Z'Z\right)^{-1}\left(\frac{1}{n} Z'X\right)\right]^{-1}\left(\frac{1}{n} X'Z\right)\left(\frac{1}{n} Z'Z\right)^{-1}\left(\frac{1}{n} Z'\boldsymbol{\epsilon}\right).$$

Modify our assumptions. We have a (random) $(n \times L)$ matrix, Z, such that:

1. $\text{plim}\left(\frac{1}{n} Z'Z\right) = Q_{ZZ} \quad ; \quad (L \times L)$, p.s.d. and finite
2. $\text{plim}\left(\frac{1}{n} Z'X\right) = Q_{ZX} \quad ; \quad (L \times k)$, rank $= k$, and finite
3. $\text{plim}\left(\frac{1}{n} Z'\boldsymbol{\epsilon}\right) = \mathbf{0} \quad ; \quad (L \times 1)$

So,

$$\text{plim}\left(\boldsymbol{b_{IV}}\right) = \boldsymbol{\beta} + \left[Q_{XZ} Q_{ZZ}^{-1} Q_{ZX}\right]^{-1} Q_{XZ} Q_{ZZ}^{-1} \mathbf{0} = \boldsymbol{\beta} \quad ; \quad \text{consistent}$$

Similar to before, a *consistent estimator* of σ^2 is

$$s_{IV}^2 = \left(\boldsymbol{y} - X\boldsymbol{b_{IV}}\right)' \left(\boldsymbol{y} - X\boldsymbol{b_{IV}}\right)/n$$

- Recall that each choice of Z leads to a *different* IV estimator.
- Z must be chosen in way that ensures consistency of the IV estimator.
- How might we choose a suitable set of instruments, in practice?
- If we have several "valid" sets of instruments, how might we choose between them?

For the "simple" IV regression model, recall that:

$$\sqrt{n}\left(\boldsymbol{b_{IV}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right]$$

so if $k = 1$,

$$Q_{ZZ} = \text{plim}\left(n^{-1} \sum_{i=1}^{n} z_i^2\right)$$

and

$$Q_{ZX} = \text{plim}\left(n^{-1} \sum_{i=1}^{n} z_i x_i\right) = Q_{XZ}$$

The asymptotic efficiency of $\boldsymbol{b}_{IV}$ will be higher, the more highly correlated are Z and X, asymptotically. We need to find instruments that are uncorrelated with the errors, but highly correlated with the regressors (asymptotically). This is not easy to do! A good instrument comes from an intimate understanding of the economics driving the regressor of interest. A good survey of some classic IV papers may be found here[1].

[1] Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic perspectives*, 15(4), 69-85.

8.4 IV tests

8.4.1 Testing if IV estimation is needed

This is a good situation to be in: it means you have found a potentially valid instrument! Now, should you use it? Recall that LS is BLUE, so we should use it where possible. Consider the following.

- Why does LS fail, and when do we need IV?
- If plim $\left(\frac{1}{n}X'\boldsymbol{\epsilon}\right) \neq \mathbf{0}$.
- We can *test* to see if this is a problem, and decide if we should use LS or IV.

The Hausman test

We want to test:

$$H_0 : \text{plim}\left(\frac{1}{n}X'\epsilon\right) = \mathbf{0} \quad \text{vs.} \quad H_A : \text{plim}\left(\frac{1}{n}X'\epsilon\right) \neq \mathbf{0}$$

- If we reject H_0, we will use IV estimation.
- If we cannot reject H_0, we'll use LS estimation.
- The Hausman test is a general "testing strategy" that can be applied in many situations, not just for this particular situation!
- Basic idea: construct 2 estimators of $\boldsymbol{\beta}$:
  1. $\boldsymbol{b}$ (LS estimator): the estimator is both *consistent* and *asymptotically efficient* if H_0 is true.
  2. $\boldsymbol{b}_{IV}$: the estimator is at least *consistent*, even if H_0 is false.
- If H_0 is true, we'd expect $(\boldsymbol{b}_{IV} - \boldsymbol{b})$ to be "small", at least for large n, as both estimators are consistent in that case.
- Hausman shows that $\hat{V}(\boldsymbol{b}_{IV} - \boldsymbol{b}) = \hat{V}(\boldsymbol{b}_{IV}) - \hat{V}(\boldsymbol{b})$, if H_0 is true.
- So, the test statistic is, $H = (\boldsymbol{b}_{IV} - \boldsymbol{b})' \left[\hat{V}(\boldsymbol{b}_{IV}) - \hat{V}(\boldsymbol{b})\right]^{-1} (\boldsymbol{b}_{IV} - \boldsymbol{b})$.
- $H \xrightarrow{d} \chi^2_J$, if H_0 is true.
- Here, J is the number of columns in X which may be correlated with the errors, and for which we need instruments.

Note that there are other asymptotically equivalent tests for the same null and alternative hypothesis; for example, the Durbin-Wu test.

8.4.2 Testing the exogeneity of instruments

The key assumption that ensures the consistency of IV estimators is that

$$\text{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) = \mathbf{0}.$$

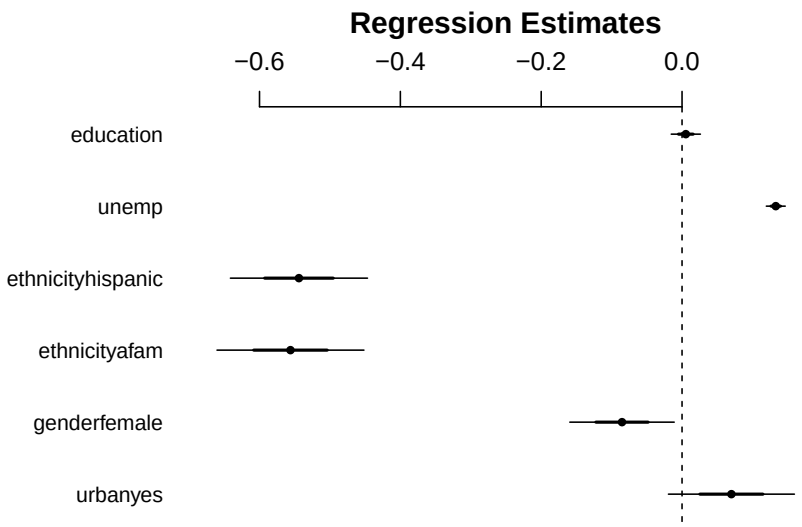
This condition involves the *unobservable* $\boldsymbol{\epsilon}$. It is difficult to test. If there are more instruments than regressors (the over-identified case) than the Sargan-Hansen or J test may be used. In an applied economics paper, establishing the exogeneity of the instruments is more likely a matter of arguing for this key assumption through an understanding of the variables, rather than relying on a statistical test.

8.4.3 Weak instruments

Problems arise if the instruments are *not* well correlated with the regressors (not relevant).

- These problems go beyond loss of asymptotic efficiency.
- Small-sample bias of IV estimator can be greater than that of LS!
- Sampling distribution of IV estimator can be bi-modal!

Figure 8.1: Results of LS regression using Card (1993) data. Dependent variable is *wage*. Notice that *education* is statistically insignificant.



- Fortunately, we can again test to see if we have these problems.

Tests aimed at detecting weak instruments revolve around detecting correlation between the instruments and endogenous regressors. Although it doesn't quite work, we could draw an analogy to the R^2, or the significance, of the 1^{st} stage in 2SLS.

8.5 Empirical example

Let's look at data from Card (1993).[2]

- Data contains *wage*, *years of education*, and demographic variables.
- Goal: estimate the returns to education in terms of *wage*.
- Problem: ability (intelligence) may be correlated with (cause) both wage and education.
- Since ability is unobservable, it is contained in the error term.
- The education variable is then correlated with the error term (endogenous).
- LS estimation of the returns to education may be inconsistent.

First, let's try LS (see the estimation results visualized in figure 8.1).

```
library(AER)
library(arm)
data("CollegeDistance")
ls <- lm(wage ~ urban + gender + ethnicity + unemp + education,
         data = CollegeDistance)
coefplot(ls)
```

Now let's try using *distance from college* (while attending high school) as an instrument for education. The argument for the validity of this instrument is that *distance from college* is correlated with *education*, since the closer a student is, the cheaper it is to get an education. For the instrument to be valid, we require that *distance* and *education* be correlated:

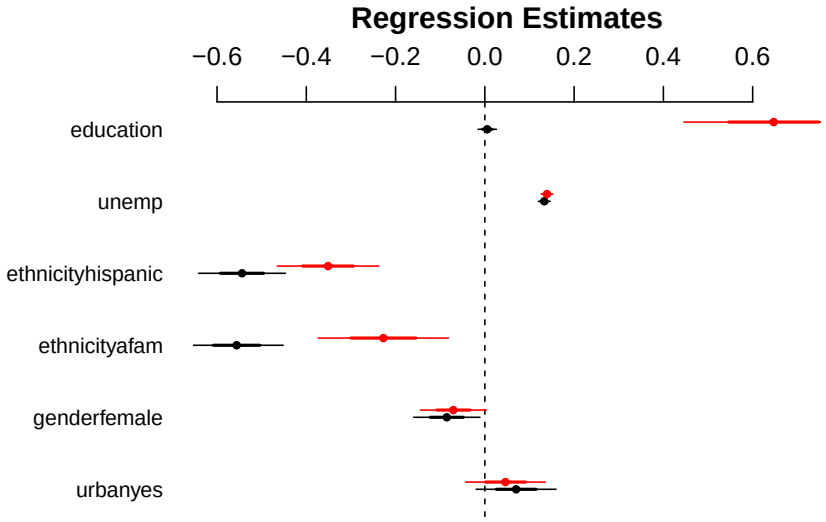
[2]Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (No. w4483). National Bureau of Economic Research.

```
summary(lm(education ~ distance, data = CollegeDistance))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.93861      0.03290 423.683    < 2e-16 ***
distance     -0.07258     0.01127   -6.441   1.3e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

While distance appears to be statistically significant, this isn't quite enough to test for validity (a testing problem we won't address here). Now, let's estimate the model using IV.

Figure 8.2: Results of LS and IV (in red) regression using Card (1993) data. Dependent variable is *wage*; *distance from college* is an instrument for *education*. Notice that the returns to education are now significant!




```
library(ivreg)
iv <- ivreg(wage ~ urban + gender + ethnicity + unemp + education |
             urban + gender + ethnicity + unemp + distance,
             data = CollegeDistance)
summary(iv, diagnostics = TRUE)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.35903    1.90830  -0.188   0.8508
urbanyes           0.04614    0.06039   0.764   0.4449
genderfemale      -0.07075    0.04997  -1.416   0.1569
ethnicityafam     -0.22724    0.09863  -2.304   0.0213 *
ethnicityhispanic -0.35129    0.07706  -4.559 5.28e-06 ***
unemp              0.13916    0.00912  15.259  < 2e-16 ***
education          0.64710    0.13594   4.760 1.99e-06 ***

Diagnostic tests:
                 df1  df2 statistic  p-value
Weak instruments   1 4732     50.31 1.51e-12 ***
Wu-Hausman         1 4731     41.12 1.57e-10 ***
Sargan             0   NA        NA       NA
```

In the `ivreg` function, the population model precedes the vertical line |, and the instruments follow the | (each exogenous regressor is an instrument for itself). Notice the "Weak instruments" and "Wu-Hausman" tests. What are these p-values telling you? The "Sargan" test is only applicable when we have an "over-identified" IV estimator. See figure 8.2 for a visual comparison with LS.

Chapter 9

Multiple Hypothesis Testing

So far, we have seen the z and t test, and a few tests to do with IV. In this chapter, we will take a general approach to testing a set of restrictions, and consider the consequences of imposing such restrictions on an estimated model. Some examples of **multiple restrictions** that we might want to test:

1.

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$$
$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{vs.} \quad H_A : \beta_1 \neq 0, \beta_2 \neq 0, \ldots, \beta_k \neq 0$$

2.

$$\log(Y) = \beta_1 + \beta_2 \log(K) + \beta_3 \log(L) + \epsilon$$
$$H_0 : \beta_2 + \beta_3 = 1 \quad \text{vs.} \quad H_A : \beta_2 + \beta_3 \neq 1$$

3.

$$\log(q) = \beta_1 + \beta_2 \log(p) + \beta_3 \log(y) + \epsilon$$
$$H_0 : \beta_2 + \beta_3 = 0 \quad \text{vs.} \quad H_A : \beta_2 + \beta_3 \neq 0$$

If the null hypothesis is true, then this implies a *restricted model*. If we can obtain one model from another by imposing restrictions on the parameters of the first model, we say that the two models are "nested".

We'll be concerned with (several) possible restrictions on β, in the usual model:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad ; \quad \boldsymbol{\epsilon} \sim N\left[0, \sigma^2 I_n\right]$$

and let's return to our simplifying assumption that X is non-random. Let's focus on J *linear restrictions*:

$$\begin{aligned} r_{11}\beta_1 + r_{12}\beta_2 + \cdots + r_{1k}\beta_k &= q_1 \\ r_{21}\beta_1 + r_{22}\beta_2 + \cdots + r_{2k}\beta_k &= q_2 \\ &\vdots \\ r_{J1}\beta_1 + r_{J2}\beta_2 + \cdots + r_{Jk}\beta_k &= q_J \end{aligned}$$

where many of the r_{jk}'s will likely be zero. We can combine these J restrictions:

$$R\boldsymbol{\beta} = \boldsymbol{q}$$

where R and $\boldsymbol{q}$ are known and non-random. We'll assume that $\text{rank}(R) = J < k$, so that there are no conflicting or redundant restrictions.

Question: What if $J = k$?

Example 9.1 — Take a set of linear restrictions and write them in terms of R and q.

1. $\beta_2 = \beta_3 = \cdots = \beta_k = 0$.

$$R = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad ; \quad \boldsymbol{q} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

2. $\beta_2 + \beta_3 = 1$.

$$R = \begin{bmatrix} 0 & 1 & 1 & 0 & \dots & 0 \end{bmatrix} \quad ; \quad q = 1$$

3. $\beta_3 = \beta_4$ and $\beta_1 = 2\beta_2$.

$$R = \begin{bmatrix} 0 & 0 & 1 & -1 & 0 & \dots & 0 \\ 1 & -2 & 0 & 0 & 0 & \dots & 0 \end{bmatrix} \quad ; \quad \boldsymbol{q} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Suppose that we estimate the model by LS, and get $\boldsymbol{b} = (X'X)^{-1} X'\boldsymbol{y}$. It is very unlikely that $R\boldsymbol{b} = \boldsymbol{q}$! Denote the *difference* between what is estimated, and what is hypothesized as:

$$\boldsymbol{m} = R\boldsymbol{b} - \boldsymbol{q}$$

$\boldsymbol{m}$ is a $(J \times 1)$ *random* vector. Let's consider the sampling distribution of $\boldsymbol{m}$. It is a linear function of $\boldsymbol{b}$. If the errors in the model are Normal, then $\boldsymbol{b}$ is Normally distributed, and hence $\boldsymbol{m}$ is Normally distributed as well. Now, to the expected value:

$$E[\boldsymbol{m}] = RE[\boldsymbol{b}] - \boldsymbol{q} = R\boldsymbol{\beta} - \boldsymbol{q}$$

So,

$$E[\boldsymbol{m}] = \boldsymbol{0}; \quad \text{iff} \quad R\boldsymbol{\beta} = \boldsymbol{q}$$

In addition, the covariance matrix of $\boldsymbol{m}$ is:

$$\begin{aligned} V[\boldsymbol{m}] &= V[R\boldsymbol{b} - \boldsymbol{q}] = V[R\boldsymbol{b}] = RV[\boldsymbol{b}]R' \\ &= R\sigma^2 (X'X)^{-1} R' = \sigma^2 R (X'X)^{-1} R' \end{aligned}$$

Question: What assumptions were used to derive the expected value and variance of $\boldsymbol{m}$?

So, the full sampling distribution of $\boldsymbol{m}$ is:

$$m \sim N\left[0, \sigma^2 R (X'X)^{-1} R'\right]$$

Let's see how we can use this sampling distribution to test if $R\boldsymbol{\beta} = \boldsymbol{q}$.

9.1 Wald test

Definition 9.1 — Wald Test Statistic. The Wald Test Statistic for testing $H_0 : R\boldsymbol{\beta} = \boldsymbol{q}$ vs. $H_A : R\boldsymbol{\beta} \neq \boldsymbol{q}$ is $W = \boldsymbol{m}'[V(\boldsymbol{m})]^{-1}\boldsymbol{m}$.

So:

$$\begin{aligned} W &= (R\boldsymbol{b} - \boldsymbol{q})' \left[\sigma^2 R \left(X'X\right)^{-1} R'\right]^{-1} (R\boldsymbol{b} - \boldsymbol{q}) \\ &= (R\boldsymbol{b} - \boldsymbol{q})' \left[R \left(X'X\right)^{-1} R'\right]^{-1} (R\boldsymbol{b} - \boldsymbol{q})/\sigma^2 \end{aligned}$$

and if H_0 is true then $m \sim N\left[\mathbf{0}, \sigma^2 R \left(X'X\right)^{-1} R'\right]$ and:

$$W \sim \chi^2_{(J)}$$

provided that σ^2 is *known*. Note that:

- This result is valid only asymptotically if σ^2 is unobservable, and we replace it with any consistent estimator.
- We would reject H_0 if $W >$ critical value (i.e., when $\boldsymbol{m} = R\boldsymbol{b} - \boldsymbol{q}$ is sufficiently "large").
- The Wald test is a very general testing procedure and is used in other testing problems.
- The Wald test statistic is always constructed using an estimator that *ignores* the restrictions being tested.

As we'll see in the next section, for this particular testing problem, we can modify the Wald test slightly and obtain a test that is exact in finite samples, and has excellent power properties.

9.2 F-test statistic and its distribution

Definition 9.2 — F-distribution Let $x_1 \sim \chi^2_{(v_1)}$, $x_2 \sim \chi^2_{(v_2)}$, and x_1 and x_2 be independent. χ^2 denotes the chi-square distribution, and v_1 and v_2 denotes degrees of freedom. Then:

$$F = \frac{\left[x_1/v_1\right]}{\left[x_2/v_2\right]} \sim F_{(v_1,v_2)}$$

$F_{(v_1,v_2)}$ denotes Snedecor's F-distribution, with degrees of freedom v_1 and v_2.

The issue with the Wald test, is that when we replace σ^2 with s^2 (For example), the Chi-square distribution only *approximately* describes the test statistic. The F-distribution tells us the exact distribution of the test statistic, which depends on the sample size n. As n grows to infinity however, the Wald test and F-test become identical.

The F-statistic can be derived from the Wald statistic by replacing σ^2 with s^2, and dividing by the number of restrictions J in the null hypothesis:

$$F = \left(\frac{W}{J}\right)\left(\frac{\sigma^2}{s^2}\right)$$

Theorem 9.1 — Distribution of the F-test statistic. $F = \left(\frac{W}{J}\right)\left(\frac{\sigma^2}{s^2}\right) \sim F_{(J,(n-k))}$, if the null hypothesis $H_0 : R\boldsymbol{\beta} = \boldsymbol{q}$ is true.

Proof.

$$F = \frac{(R\boldsymbol{b}-\boldsymbol{q})' \left[R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b}-\boldsymbol{q})}{\sigma^2}\left(\frac{1}{J}\right)\left(\frac{\sigma^2}{s^2}\right)$$
$$= \frac{(R\boldsymbol{b}-\boldsymbol{q})' \left[\sigma^2 R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b}-\boldsymbol{q})/J}{\left[\frac{(n-k)s^2}{\sigma^2}\right]/(n-k)} = \left(\frac{N}{D}\right)$$

where $D = \left[\frac{(n-k)s^2}{\sigma^2}\right]/(n-k) = \chi^2_{(n-k)}/(n-k)$. Consider the numerator:

$$N = (R\boldsymbol{b}-\boldsymbol{q})' \left[\sigma^2 R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b}-\boldsymbol{q})/J$$

Suppose that H_0 is true, so that $R\boldsymbol{\beta} = \boldsymbol{q}$, and then:

$$(R\boldsymbol{b}-q) = (R\boldsymbol{b}-R\boldsymbol{\beta}) = R(\boldsymbol{b}-\boldsymbol{\beta})$$

Recalling that $\boldsymbol{b} = \boldsymbol{\beta} + (X'X)^{-1}X'\boldsymbol{\epsilon}$:

$$R(\boldsymbol{b}-\boldsymbol{\beta}) = R\left(X'X\right)^{-1}X'\boldsymbol{\epsilon}$$

and

$$N = \left[R\left(X'X\right)^{-1}X'\boldsymbol{\epsilon}\right]' \left[\sigma^2 R\left(X'X\right)^{-1}R'\right]^{-1}\left[R\left(X'X\right)^{-1}X'\boldsymbol{\epsilon}\right]/J$$
$$= (1/J)(\boldsymbol{\epsilon}/\sigma)'[Q](\boldsymbol{\epsilon}/\sigma)$$

where $Q = X\left(X'X\right)^{-1}R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}X'$, and $(\boldsymbol{\epsilon}/\sigma) \sim N\left[\boldsymbol{0}, I_n\right]$. Now, $(\boldsymbol{\epsilon}/\sigma)'[Q](\boldsymbol{\epsilon}/\sigma) \sim \chi^2_{(r)}$ since Q is idempotent, where $r = \text{rank}(Q)$. So,

$$\begin{aligned}
\text{rank}(Q) &= \text{tr}.(Q) \\
&= \text{tr}.\left\{X\left(X'X\right)^{-1}R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}X'\right\} \\
&= \text{tr}.\left\{\left(X'X\right)^{-1}R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}X'X\right\} \\
&= \text{tr}.\left\{R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}\right\} \\
&= \left\{\left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}R'\right\} \\
&= \text{tr}.\left(I_J\right) = J.
\end{aligned}$$

So, $N = (1/J)(\boldsymbol{\epsilon}/\sigma)'[Q](\boldsymbol{\epsilon}/\sigma) = \chi^2_{(J)}/J$. In the construction of F we have a ratio of two Chi-Square statistics, each divided by their degrees of freedom. We need to establish that N and D are independent. The Chi-Square statistic in N is: $(\boldsymbol{\epsilon}/\sigma)'[Q](\boldsymbol{\epsilon}/\sigma)$ and the Chi-Square statistic in D is: $\frac{(n-k)s^2}{\sigma^2}$. Re-write this:

$$\frac{(n-k)s^2}{\sigma^2} = \frac{(n-k)}{\sigma^2}\left(\boldsymbol{e}'\boldsymbol{e}/(n-k)\right) = \left(\boldsymbol{e}'\boldsymbol{e}/\sigma^2\right)$$
$$= (M\boldsymbol{\epsilon}/\sigma)'(M\boldsymbol{\epsilon}/\sigma) = (\boldsymbol{\epsilon}/\sigma)'M(\boldsymbol{\epsilon}/\sigma)$$

So, we have $N = (\boldsymbol{\epsilon}/\sigma)' [Q] (\boldsymbol{\epsilon}/\sigma)$ and $D = (\boldsymbol{\epsilon}/\sigma)' [M] (\boldsymbol{\epsilon}/\sigma)$. These two statistics are independent if and only if $MQ = 0$.

$$\begin{aligned} MQ &= \left[I - X (X'X)^{-1} X'\right] X (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} X' \\ &= Q - X (X'X)^{-1} X'X (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} X' \\ &= Q - X (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} X' \\ &= Q - Q = 0 \end{aligned}$$

So, if H_0 is true, the statistic F is the ratio of two independent Chi-square variables, each divided by their degrees of freedom. This implies that, if H_0 is true:

$$F = \frac{(R\boldsymbol{b} - \boldsymbol{q})' \left[R (X'X)^{-1} R'\right]^{-1} (R\boldsymbol{b} - \boldsymbol{q})/J}{s^2} \sim F_{(J,(n-k))}$$

In the proof of the distribution of the F-test statistic, it is important to note that this result relies on A.6, the Normality of the error term. If the errors are not Normal, the F-test statistic does not follow the F distribution.

The F-test is used for testing a set of linear restrictions because it is *uniformly most powerful*.

Finally, note that:

$$\left(t_{(v)}\right)^2 = F_{(1,v)}$$

9.3 Implementing the F-test

One way of implementing the F-test is to estimate the population model by LS and then calculate the F statistic according to the formula:

$$F = (R\boldsymbol{b} - \boldsymbol{q})' \left[s^2 R (X'X)^{-1} R'\right]^{-1} (R\boldsymbol{b} - \boldsymbol{q})/J \tag{9.1}$$

and then reject the null hypothesis if the p-value from $F_{J,(n-k)}$ is less than the significance level. We will see a more convenient and intuitive way to perform an F-test, but first, let us consider a simple example.

9.3.1 Simple F-test in a Cobb-Douglas model

Suppose that we want to estimate the simple Cobb-Douglas model:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + \epsilon$$

Download some data (the data is from table F7.2, Greene, 2012), estimate a model by LS, and view the results:

```
cobbdata <- read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/cobb.csv")
mod1 <- lm(log(y) ~ log(k) + log(l), data = cobbdata)
summary(mod1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8444     0.2336   7.896 7.33e-08 ***
log(k)        0.2454     0.1069   2.297   0.0315 *
log(l)        0.8052     0.1263   6.373 2.06e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2357 on 22 degrees of freedom
Multiple R-squared:  0.9731,    Adjusted R-squared:  0.9706
F-statistic: 397.5 on 2 and 22 DF,  p-value: < 2.2e-16
```

What is the F-statistic of 397.5 for?
Let's get the RSS (residual sum of squares, $\boldsymbol{e}'\boldsymbol{e}$ for later use:

```
sum(mod1$residuals ^ 2)
```

```
[1] 1.22226
```

Now, test the hypothesis of *constant returns to scale*:

$$H_0 : \beta_2 + \beta_3 = 1 \quad \text{vs.} \quad H_A : \beta_2 + \beta_3 \neq 1$$

The R and $\boldsymbol{q}$ associated with this null hypothesis are:

$$R = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix} \quad ; \quad q = 1$$

We can calculate the F-test statistic from equation 9.1 in R using (this is *not* the way you will do it in practice):

```
R <- matrix(c(0, 1, 1), 1, 3)
b <- matrix(mod1$coef, 3, 1)
q <- 1
m <- R %*% b - q
Fstat <- t(m) %*% solve(R %*% vcov(mod1) %*% t(R)) %*% m
Fstat
```

```
         [,1]
[1,] 1.540692
```

and then calculate the p-value using:

```
1 - pf(Fstat, 1, 22)
```

```
           [,1]
[1,] 0.2275873
```

With a p-value of 0.23 we fail to reject the null hypothesis. However, the F-test statistic only follows the F-distribution if $\boldsymbol{\epsilon}$ is Normally distributed. Test the Normality assumption using the Jarque-Bera test:

```
install.packages("tseries")
library(tseries)
jarque.bera.test(mod1$residuals)
```

```
        Jarque Bera Test

data:  mod1$residuals
X-squared = 5.5339, df = 2, p-value = 0.06285
```

With a p-value of 0.063, the Normality assumption is dicey, and we may want to use the *Wald* test instead. The Wald test statistic can be calculated from the F-stat by: $W = J \times F$. In this example, there is only one restriction ($J = 1$) so that the Wald and F-statistic coincide. The Wald statistic is asymtptoically (in other words, approximately) Chi-square distributed even if A.6 is violated (as long the LS estimators are asymptotically Normal). To get the p-value from the Chi-square distribution we can use:

```
1 - pchisq(Fstat, 1)
```

```
           [,1]
[1,] 0.2145148
```

In this instance, the Wald and F-statistics are similar and our decision to "fail to reject" does not change. The Wald and F-tests have supported the validity of the *restrictions* specified in the null hypothesis. We could *impose* the restriction of constant returns to scale, by estimating the *restricted* model:

$$\log(Y/L) = \beta_1 + \beta_2 \log(K/L) + \epsilon$$

In the next section, we consider (i) the benefit of estimating a "restricted" model, and (ii) the cost of using a restricted model, when the restrictions (null hypothesis) are false. To do this, we develop the idea of the *restricted least squares* (RLS) estimator.

9.4 Restricted Least Squares

The RLS estimator is developed here, in order to:

- Examine the benefits (in terms of the properties of the estimator) of imposing restrictions on a model
- Examine the costs of imposing false restrictions (again in terms of properties like unbiasedness, efficiency, and consistency)

The basic idea is that the null hypothesis implies *restrictions* on the initial model under the *alternative hypothesis*. If we fail to reject H_0, then the restrictions could be imposed on the initial model simply by substituting the values for (say) β from the null. A model obtained by imposing parameter restrictions on a larger un-restricted model is often said to be "nested".

Model B is "nested" in model A if B can be obtained by imposing parameter restrictions on A.

Many tests in econometrics can be interpreted as comparing the "fit" or "performance" of the restricted (nested model, under H_0) to that of the unrestricted model (under H_A). The F-test is no exception. We will see that we can calculate the F-statistic by estimating two models and comparing their RSS or R^2. Before we do so, we develop the RLS estimator.

Definition 9.3 — Restricted least squares estimator (RLS) The RLS estimator of $\boldsymbol{\beta}$, in the model $\boldsymbol{y} = X\boldsymbol{\beta}+\boldsymbol{\epsilon}$, is the vector, $\boldsymbol{b}_*$, which minimizes the sum of the squared residuals, subject to the constraint(s) $R\boldsymbol{b}_* = \boldsymbol{q}$.

We will obtain the expression for this new estimator, and derive its sampling distribution. Set up the Lagrangian:

$$\mathcal{L} = (\boldsymbol{y} - X\boldsymbol{b}_*)'(\boldsymbol{y} - X\boldsymbol{b}_*) + 2\boldsymbol{\lambda}'(R\boldsymbol{b}_* - \boldsymbol{q})$$

For the first order conditions, set $(\partial\mathcal{L}/\partial\boldsymbol{b}_*) = \boldsymbol{0}$; $(\partial\mathcal{L}/\partial\lambda) = \boldsymbol{0}$, and solve:

$$\mathcal{L} = \boldsymbol{y}'\boldsymbol{y} + \boldsymbol{b}_*'X'X\boldsymbol{b}_* - 2\boldsymbol{y}'X\boldsymbol{b}_* + 2\boldsymbol{\lambda}'(R\boldsymbol{b}_* - \boldsymbol{q})$$

$$(\partial\mathcal{L}/\partial\boldsymbol{b}_*) = 2X'X\boldsymbol{b}_* - 2X'\boldsymbol{y} + 2R'\boldsymbol{\lambda} = \boldsymbol{0} \tag{9.2}$$

$$(\partial\mathcal{L}/\partial\boldsymbol{\lambda}) = 2(R\boldsymbol{b}_* - \boldsymbol{q}) = \boldsymbol{0} \tag{9.3}$$

From equation 9.2

$$\begin{aligned} R'\boldsymbol{\lambda} &= X'(\boldsymbol{y} - X\boldsymbol{b}_*) \\ R(X'X)^{-1}R'\boldsymbol{\lambda} &= R(X'X)^{-1}X'(\boldsymbol{y} - X\boldsymbol{b}_*) \\ \boldsymbol{\lambda} &= \left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'(\boldsymbol{y} - X\boldsymbol{b}_*) \end{aligned} \tag{9.4}$$

Inserting equation 9.4 into equation 9.2, and dividing by 2:

$$\begin{aligned} (X'X)\boldsymbol{b}_* &= X'\boldsymbol{y} - R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'(\boldsymbol{y} - X\boldsymbol{b}_*) \\ (X'X)\boldsymbol{b}_* &= X'\boldsymbol{y} - R'\left[R(X'X)^{-1}R'\right]^{-1}R(\boldsymbol{b} - \boldsymbol{b}_*) \\ \boldsymbol{b}_* &= (X'X)^{-1}X'\boldsymbol{y} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\boldsymbol{b} - R\boldsymbol{b}_*) \end{aligned} \tag{9.5}$$

and finally substituting equation 9.3 into equation 9.5, the RLS estimator is:

$$\boldsymbol{b}_* = \boldsymbol{b} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\boldsymbol{b} - \boldsymbol{q}) \tag{9.6}$$

From the above formula, we can see that RLS = LS + "adjustment factor".

Questions: What if R$\boldsymbol{b} = \boldsymbol{q}$? What is the interpretation of this?

We will use the RLS formula in order to derive the *statistical properties* of RLS estimation, thus informing us the benefits of imposing true restrictions, and costs of imposing false ones.

Theorem 9.2 — RLS estimator is unbiased iff H_0 is true. The RLS estimator of $\boldsymbol{\beta}$ is unbiased iff $R\boldsymbol{\beta} = \boldsymbol{q}$ is true. Otherwise, the RLS estimator is *biased*.

Proof.

$$\begin{aligned} E(\boldsymbol{b}_*) &= E(\boldsymbol{b}) - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(\boldsymbol{R}E(\boldsymbol{b}) - \boldsymbol{q}) \\ &= \boldsymbol{\beta} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\boldsymbol{\beta} - \boldsymbol{q}) \end{aligned}$$

So, if $R\boldsymbol{\beta} = \boldsymbol{q}$, then $E(\boldsymbol{b}_*) = \boldsymbol{\beta}$. Similarly, the LS estimator is only *consistent* if the restrictions (null hypothesis) are true.

The cost of imposing false restrictions (for example "dropping" a variable that is actually significant) is incurring bias and inconsistency in the LS estimator.

The potential cost of estimating a restrictive model is very high. The potential benefit, however, is an improvement in *efficiency*. That is, the RLS estimator has smaller variance than the LS estimator (regardless of the veracity of the restrictions). Intuitively, if restrictions are imposed on a model, than there are fewer parameters to estimate. The same amount of information (the sample size) can focus on estimating fewer parameters; this translates to smaller standard errors, narrower confidence intervals, etc. Another intuitive way to think of it is as follows. The null hypothesis contains information about the parameters. Using this information while estimating the model improves efficiency.

Below, we prove that the RLS estimators has lower variance than the LS estimators (this does not contradict the GM theorem since the RLS is estimating a *different* model from the LS estimator). The proof does not require that the *restrictions* are actually *true*.

Theorem 9.3 — The RLS estimator has smaller variance than the LS estimator. Proof:
First, we derive the covariance matrix of the RLS estimator of $\boldsymbol{\beta}$.

$$\begin{aligned}\boldsymbol{b}_* &= \boldsymbol{b} - (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} (R\boldsymbol{b} - \boldsymbol{q}) \\ &= \left\{I - (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R\right\} \boldsymbol{b} + \boldsymbol{\alpha}\end{aligned}$$

where $\boldsymbol{\alpha}$ is non-random and $\boldsymbol{\alpha} = (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} \boldsymbol{q}$. So, $V(\boldsymbol{b}_*) = AV(\boldsymbol{b})A'$, where $A = \left\{I - (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R\right\}$. That is, $V(\boldsymbol{b}_*) = AV(\boldsymbol{b})A' = \sigma^2 A (X'X)^{-1} A'$. Now, examine $A (X'X)^{-1} A'$:

$$\begin{aligned}A (X'X) A' = &\left\{I - (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R\right\} (X'X)^{-1} \\ &\times \left\{I - R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1}\right\} \\ = &(X'X)^{-1} + (X'X)^{-1} R' \\ &\times \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} \\ &- 2 (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} \\ = &(X'X)^{-1} \left\{I - R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1}\right\}\end{aligned}$$

So,

$$\begin{aligned}V(\boldsymbol{b}_*) &= \sigma^2 (X'X)^{-1} \left\{I - R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1}\right\} \\ &= \sigma^2 (X'X)^{-1} - \sigma^2 (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1} \\ &= V(\boldsymbol{b}) - \sigma^2 (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1}\end{aligned}$$

So, $V(\boldsymbol{b}) - V(\boldsymbol{b}_*) = \sigma^2 \Delta$, where $\Delta = (X'X)^{-1} R' \left[R (X'X)^{-1} R'\right]^{-1} R (X'X)^{-1}$. Δ is square, symmetric, and of full rank. So, Δ is at least positive semi-definite.

The above proof tells us that the variability of the RLS estimator is no more than that of the LS estimator, whether or not the restrictions are true. Generally, the RLS estimator will be "more precise" than the LS estimator.

In addition, we know that the RLS estimator is unbiased if the restrictions are true. So, if the restrictions are true, the RLS estimator, $\boldsymbol{b}_*$, is more efficient than the LS estimator, $\boldsymbol{b}$, of the coefficient vector, $\boldsymbol{\beta}$. If the restrictions are false, and we consider the MSE, then the relative efficiency can go either way.

The RLS estimator that we have discussed in this section (equation 9.6) is used to determine the consequences of imposing restrictions on a model, whether the restrictions are true or false. In practice, equation 9.6 is not used to effect RLS estimation. Rather, the restrictions are simply imposed on the model (such as dropping a regressor), and the model is re-estimated using LS. In practice, we:

- Estimate the unrestricted model, using LS.
- Test $H_0 : R\boldsymbol{\beta} = \boldsymbol{q}$ vs. $H_A : R\boldsymbol{\beta} \neq \boldsymbol{q}$.

- If the null hypothesis can't be rejected, substitute the restrictions into the model and re-estimate using LS.
- Otherwise, retain the initial estimates under the unrestricted model.

9.5 Testing by comparing unrestricted and restricted models

We can rewrite the F-test statistic formula in equation 9.1 in terms of the residuals or R^2 from the restricted and unrestricted model. This may be a more intuitive way of looking at the F-test.

Let $\boldsymbol{e}_*$ be the residuals from RLS estimation. By substituting in equation 9.6 for $\boldsymbol{b}_*$, we can write $\boldsymbol{e}_*$ as:

$$\begin{aligned}\boldsymbol{e}_* &= (\boldsymbol{y} - X\boldsymbol{b}_*) = \boldsymbol{y} - X\boldsymbol{b} + X\left(X'X\right)^{-1}R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b} - \boldsymbol{q}) \\ &= \boldsymbol{e} + X\left(X'X\right)R'\left[R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b} - \boldsymbol{q})\end{aligned}$$

Recalling that $X'\boldsymbol{e} = \boldsymbol{0}$:

$$\boldsymbol{e}_*'\boldsymbol{e}_* = \boldsymbol{e}'\boldsymbol{e} + (R\boldsymbol{b} - \boldsymbol{q})'A(R\boldsymbol{b} - \boldsymbol{q})$$

where A is full rank and is positive semi-definite, and is:

$$\begin{aligned}A &= \left[R\left(X'X\right)^{-1}R'\right]^{-1}R\left(X'X\right)^{-1}\left(X'X\right)\left(X'X\right)^{-1}R'\left[R\left(X'X\right)^{-1}R'\right]^{-1} \\ &= \left[R\left(X'X\right)^{-1}R'\right]^{-1}\end{aligned}$$

Question: $\boldsymbol{e}_*'\boldsymbol{e}_* > \boldsymbol{e}'\boldsymbol{e}$, because $(R\boldsymbol{b} - \boldsymbol{q})'A(R\boldsymbol{b} - \boldsymbol{q}) > 0$. This inequality will always hold. What's the intuition behind this?

Now, take the difference:

$$\begin{aligned}(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}) &= (R\boldsymbol{b} - \boldsymbol{q})'A(R\boldsymbol{b} - \boldsymbol{q}) \\ &= (R\boldsymbol{b} - \boldsymbol{q})'\left[R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b} - \boldsymbol{q})\end{aligned} \tag{9.7}$$

Recalling that:

$$F = \frac{(R\boldsymbol{b} - \boldsymbol{q})'\left[R\left(X'X\right)^{-1}R'\right]^{-1}(R\boldsymbol{b} - \boldsymbol{q})/J}{s^2}$$

we can rewrite the F-statistic formula in an alternative and more convenient form:

$$F = \frac{\left(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}\right)/J}{s^2} = \frac{\left(\boldsymbol{e}_*'\boldsymbol{e}_* - \boldsymbol{e}'\boldsymbol{e}\right)/J}{\boldsymbol{e}'\boldsymbol{e}/(n-k)} \tag{9.8}$$

In retrospect, we can see further why R^2 increases when we add any regressor to the model: deleting a regressor is equivalent to imposing a zero restriction on one of the coefficients. The RSS increases and so R^2 decreases. In fact, using the definition for R^2, we can also rewrite the F-test statistic as:

$$F = \frac{\left(R^2 - R_*^2\right)/J}{\left(1 - R^2\right)/(n-k)} \tag{9.9}$$

where R_*^2 is from the restricted model.

Equations 9.8 and 9.9 provide an intuitive interpretation of the F-test (which apply to many other tests used in econometrics): if the restricted model (under the null hypothesis) "fits" the data much more poorly than the unrestricted model (under the alternative hypothesis), then the F-statistic will be large and the restrictions (null) will be rejected.

9.5.1 F-test in Cobb-Douglas again

We revisit the example from section 9.3.1 applying the version of the F-test from equation 9.8. To accomplish this, estimate two models: one under the null hypothesis (restricted model), and one under the alternative hypothesis (unrestricted model):

```
cobbdata <- read.csv("http://home.cc.umanitoba.ca/~godwinrt/7010/cobb.csv")
unrestricted <- lm(log(y) ~ log(k) + log(l), data = cobbdata)
restricted <- lm(log(y/l) ~ log(k/l), data = cobbdata)
```

Notice that, to implement the restriction of CRTS, the left-hand-side variable has changed. This means that equation 9.9, and some "canned" commands in Stata and R (for example `anova()`) for performing the F-test, will not work! We can use the residuals from the two models, however:

```
RSS <- sum(unrestricted$residuals ^ 2)
RSSstar <- sum(restricted$residuals ^ 2)
```

and calculate the F-stat according to equation 9.8:

```
Fstat2 <- ((RSSstar - RSS) / 1) / (RSS / 22)
Fstat2
```

```
> Fstat2
[1] 1.540692
```

where we notice that the F-statistic is the same as in section 9.3.1.

9.6 Testing for differences

Dummy variables are important and are used very commonly in economics research. Although a dummy variable can take many values, in economics it is almost always refers to a binary indicator variable. Typically the values that the dummy variable can take are 0 or 1, where each value corresponds to a certain quality, or to membership in a group. A common convention is to name the dummy variable so that a value of 1 indicates "yes", and a value of 0 indicates "no". For example, the variable `foreign` might equal 1 when a firm is foreign-owned, and 0 when domestically owned.

Dummy variables provide a convenient way to test for *differences* between groups, regions, countries, types of firms, moments in time, *treatment* status, etc.

Let D denote our dummy variable. Then, consider a model of the form:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k D_i + \epsilon_i \tag{9.10}$$

we could then think of testing:

$$H_0 : \beta_k = 0 \quad \text{vs.} \quad H_A : \beta_k \neq 0$$

using a t-test or z-test. Rejection of H_0 implies there is a particular type of *structural change* in the model. In particular, the dummy variable in equation 9.10 allows the *mean* of y_i to differ depending on the value of D_i, that is:

$$\mathbb{E}\left(y_i | D_i = 1\right) - \mathbb{E}\left(y_i | D_i = 0\right) = \beta_k$$

So, rejection of H_0 implies that D_i has a significant effect on the mean of y_i. More generally, consider a model of the form:

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_{k-1}\left(D_i \times x_{i2}\right) + \beta_k D_t + \epsilon_i \tag{9.11}$$

where we could use an F or Wald test for:

$$H_0 : \beta_{k-1} = \beta_k = 0 \quad \text{vs.} \quad H_A : \text{Not } H_0$$

Rejection of H_0 implies a different type of structural change in the model: a shift in the mean and one of the marginal effects. Allowing the dummy variable to interact with the "x" variables, such as in model 9.11, allows for variables to have different marginal effects depending on regions, times, or treatment status. For example, in model 9.11, the change in y associated with a change in x_2 is:

$$\frac{\partial y_i}{\partial x_{i2}} = \beta_2 + D_i \beta_{k-1}$$

At the extreme, we could allow the dummy variable to interact fully with every variable in the model:

$$\boldsymbol{y} = X\boldsymbol{\beta}_1 + DX\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{9.12}$$

Testing the joint significance of every term that includes D_i ($H_0 : \boldsymbol{\beta_2} = 0$) is equivalent to the Chow test (Gregory Chow, 1960). Note the parameter estimates from model 9.12 can also be obtained by fitting two separate models for the two separate subsamples (as defined by the dummy variable).

Example 9.2 — Life Expectancy - F-test. This example follows Greene (2012), example 6.10 (pg. 173). The WHO released a report in 2000 that gained much attention, and was controversial. It suggested that health care expenditure by OECD countries had more of an impact on life expectancy than spending by non-OECD countries.

The dependent variable is `DALE` - disability-adjusted life expectancy. Consider a simple model:

$$DALE = \beta_1 + \beta_2 HEXP + \beta_3 HC3 + \beta_4 HC3^2 + \epsilon \tag{9.13}$$

where `HEXP` is health expenditure, and `HC3` is a measure of educational attainment. Download data from Greene, choose to use only the year 1997, and delete missing values (to follow Greene):

```
health <- read.csv("http://www.stern.nyu.edu/~wgreene/Text/Edition7/TableF6-3.csv")
health <- subset(health, health$YEAR == 1997)
health <- na.omit(health)
```

Next create the squared term (we can't square inside `lm()`):

```
health$HC3sq <- health$HC3^2
```

Estimate model 9.13, and view the results:

```
mod1 <- lm(DALE ~ HEXP + HC3 + HC3sq, data = health)
summary(mod1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.237039   2.439645  10.345  < 2e-16 ***
HEXP         0.006291   0.001056   5.960 1.23e-08 ***
HC3          7.930951   0.902432   8.788 9.73e-16 ***
HC3sq       -0.438988   0.077157  -5.690 4.85e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.984 on 187 degrees of freedom
Multiple R-squared:  0.6824,    Adjusted R-squared:  0.6773
F-statistic: 133.9 on 3 and 187 DF,  p-value: < 2.2e-16
```

Now, consider a model that has a dummy variable `OECD` that *interacts* with all of the regressors:

$$\begin{aligned} DALE =& \beta_1 + \beta_2 HEXP + \beta_3 HC3 + \beta_4 HC3^2 + \beta_5 OECD + \beta_6 (OECD \times HEXP) \\ &+ \beta_7 (OECD \times HC3) + \beta_8 (OECD \times HC3^2) + \epsilon \end{aligned} \tag{9.14}$$

Model 9.14 allows for the effects of both health expenditure and education to differ, depending on whether the country is OECD or not. Estimate this model in R:

```
mod2 <- lm(DALE ~ HEXP + HC3 + HC3sq + OECD + OECD*HEXP + OECD*HC3 + OECD*HC3sq,
        data = health)
summary(mod2)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 26.812047   2.651604  10.112  < 2e-16 ***
HEXP         0.009551   0.001934   4.938 1.77e-06 ***
HC3          7.043319   1.074319   6.556 5.46e-10 ***
HC3sq       -0.373804   0.096020  -3.893 0.000139 ***
OECD        15.916263  25.678970   0.620 0.536149
HEXP:OECD   -0.006872   0.002619  -2.624 0.009435 **
HC3:OECD    -0.866530   6.261351  -0.138 0.890082
HC3sq:OECD  -0.011024   0.378032  -0.029 0.976768
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.86 on 183 degrees of freedom
Multiple R-squared:  0.7002,    Adjusted R-squared:  0.6887
F-statistic: 61.06 on 7 and 183 DF,  p-value: < 2.2e-16
```

Now, test the hypothesis that there is *no difference* between OECD and non-OECD countries (we are testing the joint significance of all interaction terms):

$$H_0 : \beta_5 = 0, \beta_6 = 0, \beta_7 = 0, \beta_8 = 0 \quad \text{vs.} \quad H_A : \text{ not } H_0$$

To calculate the F-test, we can use the R-squared from the unrestricted and restricted models:

$$F = \frac{\left(R^2 - R_*^2\right)/J}{\left(1 - R^2\right)/\left(n - k\right)} = \frac{(0.7002 - 0.6824)/4}{(1 - 0.7002)/183} = 2.716$$

While this formula for the F-test statistic highlights its dependence on the "fit" of the models under the null and alternative hypotheses, we can easily perform this test in R using the `anova()` function:

```
anova(mod1, mod2)
```

```
Analysis of Variance Table

Model 1: DALE ~ HEXP + HC3 + HC3sq
Model 2: DALE ~ HEXP + HC3 + HC3sq + OECD + OECD * HEXP + OECD * HC3 +
    OECD * HC3sq
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1    187 9121.8
2    183 8611.0  4    510.83 2.714 0.0314 *
```

We get a similar F-stat of 2.714, with associated p-value of 0.0314. We reject the null hypothesis at 5% significance. It would appear that OECD countries might differ from non-OECD. However, this difference is being driven by the variable `HEXP`. We would fail to reject the null that the effect of education on life expectancy is the same for both OECD and non-OECD (see exercise 2), and could perhaps drop the interaction terms $OECD \times HC3$ and $OECD \times HC3^2$ from the model.

9.7 Exercises

1. Using R, verify that estimating a model where a dummy variable fully interacts with every variable, provides identical parameter estimates to dividing the data into two subsamples (where the subsamples are defined by the dummy variable).
2. Using the data from example 9.2, and using model 9.14 as the unrestricted model, test the hypothesis that the effect of education on life expectancy is the same for OECD and non-OECD countries.

Chapter 10

Non-Linear Relationships and Non-Linear Least Squares

The models we've worked with so far have been linear in the parameters, and have been of the form:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \epsilon$$

In this chapter we'll be interested in a more general model:

$$\boldsymbol{y} = f(\boldsymbol{\theta}; X) + \boldsymbol{\epsilon} \tag{10.1}$$

where $f()$ can be non-linear (the linear model is just a special case).

Many relationships between variables are non-linear. Simple examples are diminishing marginal utility, or increasing returns to scale. If the data generating process specifies a non-linear relationship between the dependent and explanatory variables, LS may be biased and inconsistent.

10.1 Transforming a non-linear population model

In some situations where f in model 10.1 is non-linear, we may be able to transform the *variables* in order to linearize the relationship between y and X (for example taking logs of a Cobb-Douglas production function). Cobb-Douglas production function:

$$Y = AK^{\beta_2}L^{\beta_3}\varepsilon$$

By taking logs, the Cobb-Douglas production function can be rewritten as:

$$\log Y = \beta_1 + \beta_2 \log K + \beta_3 \log L + \log(\varepsilon)$$

This model now satisfies A.1 (linear in the parameters), however, it is not always advisable to estimate multiplicative models by LS! Silva and Tenreyro (2006)[1] if $\log(\varepsilon)$ is heteroskedastic[2] (it likely is), then X and $\log(\varepsilon)$ are *not* independent (violation of A.5).

10.2 Polynomial regression model

In other situations, where the model is non-linear in the parameters (model 10.1), we can approximate the non-linear relationship. A common practice is to add additional variables to X, that are just non-linear transformations of the original X matrix. For example, including in the regression model higher order polynomials of the x variables (i.e. squared and cubed terms), makes f in model 10.1 approximately linear. The validity of such an approach is based on the Taylor series approximation.

[1] Silva and Tenreyro (2006). The Log of Gravity. *The Review of Economics and Statistics*.
[2] We cover heteroskedasticity in the next chapter.

One way to characterize the non-linear relationship between y and x is to say that the marginal effect of x on y depends on the value of x itself. By just including powers of the regressors on the right-hand-side of the regression model (this is not a violation of A.2), we can allow the marginal effect of x on y to vary in a way that depends on x. For example:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \epsilon \tag{10.2}$$

Taking the derivative of y with respect to x gives:

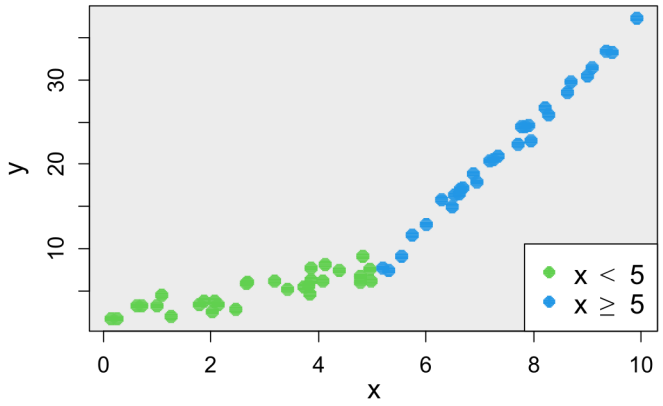
$$\frac{\partial y}{\partial x} = \beta_1 + 2x\beta_2 + 3x^2\beta_3 + \cdots$$

We can choose $\boldsymbol{\beta}$ to give an arbitrarily good approximation of any non-linear relationship $y = f(x)$. The appropriate order of the polynomial may be determined through a series of t-tests, where if the highest order term is found to be insignificant, it is dropped and the model re-estimated.

The drawback of polynomial regression models are that the parameters become difficult to interpret, and the model may not generalize well outside of the data. Often we are interested in a causal effect that is represented by parameters in a model, and so we might want to estimate the non-linear model directly.

10.3 Splines

Figure 10.1: What model would you specify for this data?



There may be a "break" in the model so that it is "piecewise" linear (see Figure 10.1 where there is a break at $x = 5$). From section 9.6 we know that we can use a dummy variable to fully interact with all of the variables in the model, allowing the intercept and slope coefficients to differ by the value of the dummy. That is, we could define a dummy variable:

$$D = 0 \quad \text{if } x < 5;$$
$$D = 1 \quad \text{if } x \geq 5$$

and then estimate a model for the data in Figure 10.1 as:

$$y = \beta_1 + \beta_2 x + \beta_3 D + \beta_4 D x + \epsilon \tag{10.3}$$

Model 10.3 is estimated, and drawn in figure 10.2. Notice that the two separate regression lines do not connect at the break point.

In section 9.4, we discussed how we can impose *restrictions* on the parameters in the model. If we want to estimate a *piecewise* linear regression function, then we can *force* the two lines in Figure 10.2 to join

Figure 10.2: A dummy variable allows for two separate regression lines to be estimated, on either side of the "break" point. Notice that the two lines do not connect.

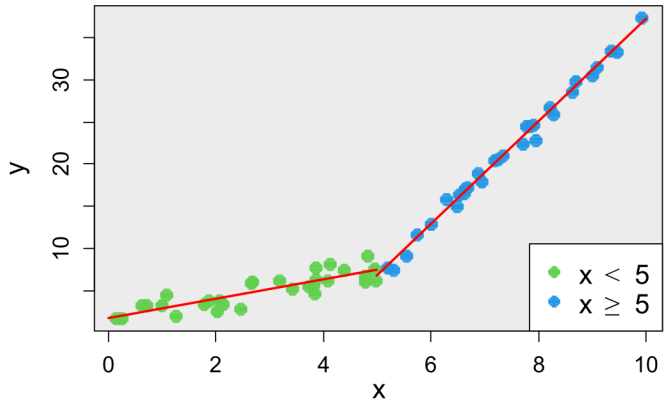
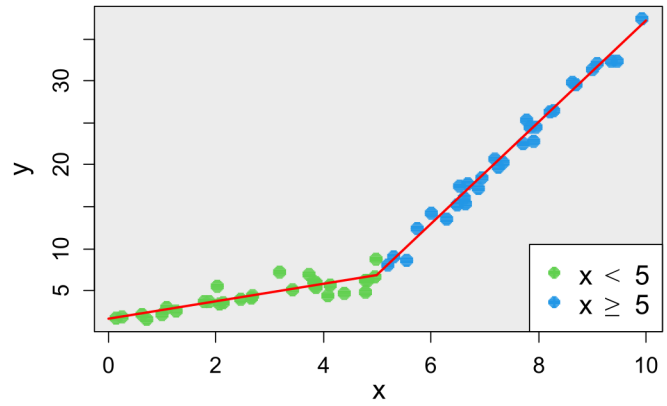


Figure 10.3: Restricting the parameter values in a model where a dummy variable fully interacts with every variable can force the piecewise linear function to join at knots. This is known as spline regression.



at the break-point (called a "knot"). We can do so by imposing a restriction on the model: the y value at the end of the first line should be equal to the y value at the beginning of the second line. That is, the equations for the two separate lines should equal when $x = 5$. Imposing such a restriction gives us the model:

$$y = \beta_1 + \beta_2 x + \beta_4 D(x - 5) + \epsilon \tag{10.4}$$

Estimating the restricted model 10.4 (see Figure 10.3) is called *spline* regression. We can have multiple "knots" (break points) in the spline regression, where the piecewise linear functions are joined at the knots through a set of restrictions on the parameter values. This model can be extended to allow an arbitrary number of knots, the locations of which can be determined by the data. This extension is known as non-parametric kernel regression (not covered in these notes).

10.4 Non-linear least squares

In yet other cases, when the model is inherently non-linear in the parameters, an approximation may be inadequate. In addition, we may desire to estimate the economic model in its natural form, to have a more direct interpretation of the parameter estimates. Then, we must use a *different* estimation methodology, such as non-linear least squares (NLS). An alternative option could be *maximum likelihood*.

Take for example a CES production function:

$$Y_i = \gamma \left[\delta K_i^{-\rho} + (1-\delta) L_i^{-\rho}\right]^{-v/\rho} \exp(\epsilon_i)$$

This is a function where there is no known transformation to make it linear in the parameters. In general, suppose we have a single non-linear equation:

$$\begin{aligned} y_i &= f(x_{i1}, x_{i2}, \ldots, x_{ik}; \theta_1, \theta_2, \ldots, \theta_p) + \epsilon_i \\ \boldsymbol{y} &= f(X; \boldsymbol{\theta}) + \boldsymbol{\epsilon} \end{aligned} \tag{10.5}$$

We can still consider a "least squares" approach to estimate model 10.5 (which can be motivated by the method of moments). The Non-Linear Least Squares estimator is the vector, $\hat{\boldsymbol{\theta}}$, that minimizes the quantity: $S(X, \boldsymbol{\theta}) = \sum_i \left[y_i - f_i(X, \widehat{\boldsymbol{\theta}})\right]^2$. (Note that the usual LS estimator is a special case of this). To obtain the estimator, we differentiate S with respect to each element of $\hat{\boldsymbol{\theta}}$; set up the "$p$" first-order conditions, and solve.

We won't derive the results here, but the properties of the NLS estimator are as follows:

- consistent
- asymptotically efficient
- asymptotically Normal

The NLS estimator has good asymptotic properties, however, there is a serious computational difficulty. Usually there is no *exact* solution for $\hat{\boldsymbol{\theta}}$! The first-order conditions are themselves non-linear in the unknown parameters. This is true for the vast majority of estimators in non-linear models, including maximum likelihood. There is (generally) no exact, closed-form solution. That is, we can't write down an explicit formula for the estimators of the parameters.

Example 10.1 — No closed form solution. Take the model:

$$y_i = \theta_1 + \theta_2 x_{i2} + \theta_3 x_{i3} + (\theta_2 \theta_3) x_{i4} + \epsilon_i$$

The sum of squared errors are:

$$S = \sum_i \left[y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - (\theta_2 \theta_3) x_{i4}\right]^2$$

and the first-order conditions are:

$$\begin{aligned} \frac{\partial S}{\partial \theta_1} &= -2 \sum_i \left[y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - (\theta_2 \theta_3) x_{i4}\right] \\ \frac{\partial S}{\partial \theta_2} &= -2 \sum_i \left[(\theta_3 x_{i4} + x_{i2}) (y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - \theta_2 \theta_3 x_{i4})\right] \\ \frac{\partial S}{\partial \theta_3} &= -2 \sum_i \left[(\theta_2 x_{i4} + x_{i3}) (y_i - \theta_1 - \theta_2 x_{i2} - \theta_3 x_{i3} - \theta_2 \theta_3 x_{i4})\right] \end{aligned}$$

Setting these 3 equations to zero, we can't solve analytically for the estimators of the three parameters.

In situations such as example 10.1, and the majority of non-linear models, we need to use a *numerical algorithm* to obtain a solution to the first-order conditions. There are lots of methods for doing this, most of which are based on *Newton*'s algorithm.

10.4.1 Taylor series approximation

A Taylor series is an expansion of a function $f(x)$ about a point $x = a$, and is given by:

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

where $f'(a)$ and $f''(a)$ are the first and second derivatives of $f(x)$ evaluated at the point a, for example.

Taylor's theorem says that certain functions can be expressed as a Taylor series. The right-hand-side of the Taylor series becomes an *approximation* when we truncate the infinite sum, for example:

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2$$

10.4.2 Newton-Raphson algorithm

Suppose we want to minimize some function, $f(\boldsymbol{\theta})$. We can approximate the function using a Taylor's series expansion about $\widetilde{\boldsymbol{\theta}}$, the vector value that minimizes $f(\boldsymbol{\theta})$:

$$f(\boldsymbol{\theta}) \approx f(\widetilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})' \left(\frac{\partial f}{\partial \boldsymbol{\theta}}\right)_{\widetilde{\boldsymbol{\theta}}} + \frac{1}{2!}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})' \left[\frac{\partial^2 f}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right]_{\widetilde{\boldsymbol{\theta}}} (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$$

or:

$$f(\boldsymbol{\theta}) \approx f(\widetilde{\boldsymbol{\theta}}) + (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})' g(\widetilde{\boldsymbol{\theta}}) + \frac{1}{2!}(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})' H(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$$

Now, the derivative of the function $f(\boldsymbol{\theta})$ that we want to minimize is:

$$\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \approx 0 + (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})' g(\widetilde{\boldsymbol{\theta}}) + \frac{1}{2!} 2 H(\widetilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}})$$

However, the gradient at the minimum located by $\widetilde{\boldsymbol{\theta}}$ is zero ($g(\widetilde{\boldsymbol{\theta}}) = 0$), so:

$$(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \approx H^{-1}(\widetilde{\boldsymbol{\theta}}) \left(\frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)$$

or

$$\widetilde{\boldsymbol{\theta}} \approx \boldsymbol{\theta} - H^{-1}(\widetilde{\boldsymbol{\theta}}) g(\boldsymbol{\theta}) \tag{10.6}$$

Equation 10.6 suggests a numerical algorithm: set $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ to begin, and then iterate:

$$\begin{aligned}
\boldsymbol{\theta}_1 &= \boldsymbol{\theta}_0 - H^{-1}(\boldsymbol{\theta}_1)\, g(\boldsymbol{\theta}_0) \\
\boldsymbol{\theta}_2 &= \boldsymbol{\theta}_1 - H^{-1}(\boldsymbol{\theta}_2)\, g(\boldsymbol{\theta}_1) \\
&\vdots \\
\boldsymbol{\theta}_{n+1} &= \boldsymbol{\theta}_n - H^{-1}(\boldsymbol{\theta}_{n+1})\, g(\boldsymbol{\theta}_n)
\end{aligned}$$

or, *approximately*:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - H^{-1}(\boldsymbol{\theta}_n)\, g(\boldsymbol{\theta}_n) \tag{10.7}$$

The algorithm stops when the difference between iterations is within a certain *tolerance*, that is, if:

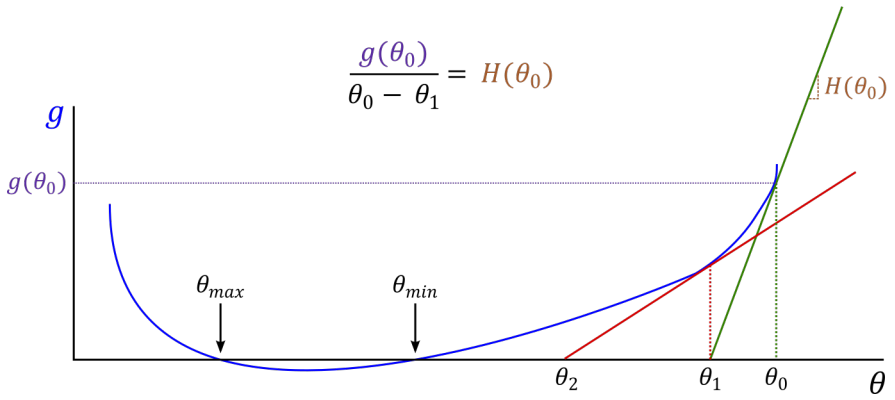
$$\left| \frac{\left(\theta_{n+1}^{(i)} - \theta_n^{(i)}\right)}{\theta_n^{(i)}} \right| < \mathcal{E}^{(i)}; \quad i = 1, 2, \ldots, p$$

When the algorithm stops because the changes between successive iterations are small, the algorithm is said to have *converged*. Note:

- The algorithm fails if H ever becomes *singular* at any iteration.
- The algorithm will achieve a minimum if H is positive definite.
- The algorithm may locate only a local minimum.
- The algorithm may oscillate.

The algorithm can be given a nice geometric interpretation in the case of a scalar θ.

Figure 10.4: Illustration of Newton-Raphson algorithm for scalar θ.



In general, different choices of θ_0 may lead to different solutions, or no solution at all.

Example 10.2 — Newton's method. Consider locating the minimum of the following function:

$$f(\theta) = 3\theta^4 - 4\theta^3 + 1$$

We can actually solve for the minimum analytically (the solution is 1):

$$g(\theta) = 12\theta^3 - 12\theta^2 = 12\theta^2(\theta - 1)$$
$$H(\theta) = 36\theta^2 - 24\theta = 12\theta(3\theta - 2)$$

Let's instead use the algorithm. Choose an initial value $\theta_0 = 2$ for example. Then:

$$\theta_1 = 2 - \left(\frac{48}{96}\right) = 1.5$$
$$\theta_2 = 1.5 - \left(\frac{13.5}{45}\right) = 1.2$$
$$\theta_3 = 1.2 - \left(\frac{3.456}{23.040}\right) = 1.05$$

We can see the solution converging to 1. What if we try $\theta_0 = -2$?

Example 10.3 — NLS in R. In this example, we generate data according to the following non-linear model:

$$y = \beta_1 x^{\beta_2} + \epsilon$$

Set some parameter values and generate some data:

```
set.seed(1)
n <- 50
x <- rnorm(n, 5, 2)
beta1 <- 3
beta2 <- 2.5
y <- beta1 * x ^ beta2 + rnorm(n, 0, 40)
```

Now, estimate the model using NLS:

```
mod <- nls(y ~ b1 * x ^ b2, start = list(b1 = 0, b2 = 0))
summary(mod)
```

```
Formula: y ~ b1 * x^b2

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b1   2.6193     0.6124   4.277 8.97e-05 ***
b2   2.5618     0.1230  20.824  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.46 on 48 degrees of freedom

Number of iterations to convergence: 9
Achieved convergence tolerance: 1.443e-07
```

Notice that we need to choose starting values for the parameters, and that the estimation output refers to iterations, convergence, and tolerance.

10.5 The Log of Gravity

Beginning in 1962, many theories of trade, even when based on different foundations, all predicted a gravity relationship for trade flows (analogous to Newton's law of gravitation):

$$T_{ij} = \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij} \tag{10.8}$$

The gravity model in equation 10.8 suggests that trade flow from i to j is directly proportional to the GDP of i and j (Y_i and Y_j), and inversely proportional to their (broadly-defined) distance (D_{ij}). $\alpha_0, \alpha_1, \alpha_2, \alpha_3$ are the parameters to be estimated. η_{ij} is an error term with $E(\eta_{ij} \mid Y_i, Y_j, D_{ij}) = 1$, which is independent from the regressors.

For decades, equation 10.8 was log-linearized (as we have done with the Cobb-Douglas model) so that LS can be used to estimate the unknown parameters:

$$\begin{aligned} \ln T_{ij} &= \ln \alpha_0 + \alpha_1 \ln Y_i + \alpha_2 \ln Y_j + \alpha_3 \ln D_{ij} + \ln \eta_{ij} \\ &= \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_j + \beta_3 \ln D_{ij} + \epsilon_{ij} \end{aligned} \tag{10.9}$$

An immediate issue with log-linearizing the gravity equation is that trade flows of zero ($T_{ij} = 0$) are not allowed, since $\ln 0$ is undefined. The practice was to drop observations with 0 trade flows from the data set, even though in many cases there were lots of 0s in the data.

Further to the problem of 0s in the data, in a famous paper called "The log of gravity"[3], Silva and Tenreyo (2006) showed that the LS estimator of equation 10.9 is biased and inconsistent in the fairly common situation where the error term η_{ij} is heteroskedastic (we cover heteroskedasticity in the next chapter). This was surprising; although η_{ij} may be independent from the regressors, $\ln \eta_{ij}$ is generally not independent. This is a violation of A.5, suggesting that LS should not be used, and that the gravity equation in 10.8 should be estimated in it's non-linear form (NLS can be used, but Silva and Tenreyo actually propose a more efficient estimator in their paper).

[3]Silva, J. S., & Tenreyro, S. (2006). The log of gravity. The Review of Economics and Statistics, 88(4), 641-658.

Figure 10.5: A table of estimates from Silva and Tenreyo (2006). I have edited the table, removing three columns.

TABLE 3.—THE TRADITIONAL GRAVITY EQUATION

Estimator: Dependent Variable:	OLS $\ln(T_{ij})$	NLS T_{ij}	PPML T_{ij}
Log exporter's GDP	0.938**	0.738**	0.733**
	(0.012)	(0.038)	(0.027)
Log importer's GDP	0.798**	0.862**	0.741**
	(0.012)	(0.041)	(0.027)
Log exporter's GDP per capita	0.207**	0.396**	0.157**
	(0.017)	(0.116)	(0.053)
Log importer's GDP per capita	0.106**	−0.033	0.135**
	(0.018)	(0.062)	(0.045)
Log distance	−1.166**	−0.924**	−0.784**
	(0.034)	(0.072)	(0.055)
Contiguity dummy	0.314*	−0.081	0.193
	(0.127)	(0.100)	(0.104)
Common-language dummy	0.678**	0.689**	0.746**
	(0.067)	(0.085)	(0.135)
Colonial-tie dummy	0.397**	0.036	0.024
	(0.070)	(0.125)	(0.150)
Landlocked-exporter dummy	−0.062	−1.367**	−0.864**
	(0.062)	(0.202)	(0.157)
Landlocked-importer dummy	−0.665**	−0.471**	−0.697**
	(0.060)	(0.184)	(0.141)
Exporter's remoteness	0.467**	1.188**	0.660**
	(0.079)	(0.182)	(0.134)
Importer's remoteness	−0.205*	1.010**	0.561**
	(0.085)	(0.154)	(0.118)
Free-trade agreement dummy	0.491**	0.443**	0.181*
	(0.097)	(0.109)	(0.088)
Openness	−0.170**	0.928**	−0.107
	(0.053)	(0.191)	(0.131)
Observations	9613	18360	18360
RESET test p-values	0.000	0.000	0.331

10.5.1 Estimate gravity by LS

Let's reproduce the first column in Table 10.5. Load the gravity data using:

```
grav <- read.csv("https://rtgodwin.com/data/gravity.csv")
```

Drop all observations where trade is zero (that's almost half the sample!) and estimate the model by log-linearizing and using LS:

```
grav.no.zeros <- subset(grav, grav$trade > 0)
m1 <- lm(log(trade) ~ lypex + lypim + lyex + lyim + ldist + border + comlang
         + colony + landl_ex + landl_im + lremot_ex + lremot_im + comfrt_wto
         + open_wto,
         data = grav.no.zeros)
summary(m1)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.49202    1.08804 -26.187  < 2e-16 ***
lypex         0.93782    0.01163  80.624  < 2e-16 ***
lypim         0.79779    0.01110  71.881  < 2e-16 ***
lyex          0.20731    0.01664  12.462  < 2e-16 ***
lyim          0.10613    0.01670   6.355 2.18e-10 ***
```

```
ldist        -1.16601    0.03391 -34.390  < 2e-16 ***
border        0.31400    0.14252   2.203 0.027603 *
comlang       0.67804    0.06398  10.597  < 2e-16 ***
colony        0.39680    0.06810   5.827 5.84e-09 ***
landl_ex     -0.06197    0.06459  -0.959 0.337363
landl_im     -0.66452    0.06313 -10.526  < 2e-16 ***
lremot_ex     0.46707    0.07776   6.007 1.96e-09 ***
lremot_im    -0.20496    0.08077  -2.538 0.011177 *
comfrt_wto    0.49082    0.10533   4.660 3.20e-06 ***
open_wto     -0.16964    0.04902  -3.460 0.000542 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.9 on 9598 degrees of freedom
Multiple R-squared:  0.6623,    Adjusted R-squared:  0.6618
F-statistic:  1345 on 14 and 9598 DF,  p-value: < 2.2e-16
```

Notice that these results are very similar to those reported in Figure 10.5.

10.5.2 Estimate gravity by NLS

Taking logs and using LS is a bad idea (bias and inconsistency), so we should *undo* the logs:

$$\begin{aligned} T_{ij} &= \alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3} \eta_{ij} \\ \ln T_{ij} &= \beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_j + \beta_3 \ln D_{ij} + \ln \eta_{ij} \\ \exp(\ln T_{ij}) &= \exp(\beta_0 + \beta_1 \ln Y_i + \beta_2 \ln Y_j + \beta_3 \ln D_{ij} + \ln \eta_{ij}) \\ \boldsymbol{T} &= \exp(X\boldsymbol{\beta}) + \boldsymbol{\epsilon} \end{aligned} \tag{10.10}$$

We write the gravity model in this way[4] so that we can see that NLS is applicable (it matches $\boldsymbol{y} = f(X;\boldsymbol{\beta}) + \boldsymbol{\epsilon}$ in equation 10.5), and so that we see that we can ask the computer to find $\boldsymbol{\beta}$ in $\exp(X\boldsymbol{\beta})$ instead of α in $\alpha_0 Y_i^{\alpha_1} Y_j^{\alpha_2} D_{ij}^{\alpha_3}$. Asking for $\boldsymbol{\beta}$ is much easier than the α (it is very common to make transformations in non-linear estimation).

To estimate the model by NLS, on the RHS we put the linear equation inside of the exponent: $\exp(X\boldsymbol{\beta})$.[5] In the R code below, notice that we need to choose starting values for each of the parameters:

```
m2 <- nls(trade ~ exp(b0 + b1 * lypex + b2 * lypim + b3 * lyex + b4 * lyim
                      + b5 * ldist + b6 * border + b7 * comlang + b8 * colony
                      + b9 * landl_ex + b10 * landl_im + b11 * lremot_ex
                      + b12 * lremot_im + b13 * comfrt_wto + b14 * open_wto),
          data = dat,
          start = list(b0 = 0, b1 = 0.5, b2 = 0.5, b3 = 0.5, b4 = 0.5,
                       b5 = 0.5, b6 = 0.5, b7 = 0.5, b8 = 0.5, b9 = 0.5,
                       b10 = 0.5, b11 = 0.5, b12 = 0.5, b13 = 0.5, b14 = 0.5))
summary(m2)
```

```
Formula: trade ~ exp(b0 + b1 * lypex + b2 * lypim + b3 * lyex + b4 * lyim +
    b5 * ldist + b6 * border + b7 * comlang + b8 * colony + b9 *
    landl_ex + b10 * landl_im + b11 * lremot_ex + b12 * lremot_im +
    b13 * comfrt_wto + b14 * open_wto)

Parameters:
      Estimate Std. Error  t value Pr(>|t|)
b0  -45.098657   0.239150 -188.579  < 2e-16 ***
b1    0.737755   0.004358  169.282  < 2e-16 ***
b2    0.861871   0.004517  190.811  < 2e-16 ***
```

[4] We have used $\eta_{ij} = 1 + \epsilon_{ij}/\exp(x_{ij}\beta)$ (see Silva and Tenreyro (2006)) to get to the last line.
[5] This is a common "link" function to use when the LHS variable needs to be non-negative.

```
b3     0.395645    0.009664    40.940  < 2e-16 ***
b4    -0.032511    0.006660    -4.881 1.06e-06 ***
b5    -0.923709    0.008487  -108.844  < 2e-16 ***
b6    -0.081309    0.009858    -8.248  < 2e-16 ***
b7     0.689402    0.015900    43.358  < 2e-16 ***
b8     0.035797    0.017787     2.013   0.0442 *
b9    -1.367068    0.030514   -44.802  < 2e-16 ***
b10   -0.471462    0.022312   -21.130  < 2e-16 ***
b11    1.187801    0.018258    65.056  < 2e-16 ***
b12    1.009682    0.017882    56.462  < 2e-16 ***
b13    0.442547    0.013734    32.224  < 2e-16 ***
b14    0.928022    0.023823    38.955  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 567000 on 18345 degrees of freedom

Number of iterations to convergence: 47
Achieved convergence tolerance: 6.762e-06
```

Again note that these results are very similar as those reported in Figure 10.5.

10.6 Exercises

1. Show that a polynomial regression model allows for regressors to have non-linear effects on the outcome ($\boldsymbol{y}$) variable.
2. Show how to get model 10.4 from model 10.3.
3. Explain how you could decide whether to use model 10.4 or model 10.3.
4. Describe, but do not derive, the properties of the NLS estimator.
5. Derive the Newton-Raphson algorithm graphically, for the scalar case.
6. Explain the terms: iterations, convergence, and tolerance, in reference to the Newton-Raphson algorithm.
7. Estimate the gravity model yourself, using LS and NLS.

Chapter 11

Heteroskedasticity

In this chapter we revisit assumption A.4, which says:

$$\mathrm{V}\left[\boldsymbol{\epsilon}\right] = \sigma^2 I_n$$

The term "non-spherical disturbances" refers to the situation where $\mathrm{V}\left[\boldsymbol{\epsilon}\right] \neq \sigma^2 I_n$. In this chapter, we instead generalize the specification of the error term in the population model:

$$E[\boldsymbol{\epsilon}] = \mathbf{0} \quad ; \quad V[\boldsymbol{\epsilon}] = \sigma^2 \Omega = \Sigma \tag{11.1}$$

Equation 11.1 allows for the possibility of one or both of *heteroskedasticity* and *autocorrelation*. In this chapter we examine the situation of heteroskedasticity, and how this more general situation for the covariance matrix of the error term affects our LS estimator, and hypothesis testing.

Definition 11.1 — Heteroskedasticity. The error term is said to be heteroskedastic when $\mathrm{var}[\epsilon_i] = \sigma_i^2$, and there are some $\sigma_i^2 \neq \sigma_j^2$. That is, each observation can have a different variance, and the term "heteroskedasticity" means "differing dispersion." The alternative to heteroskedasticy is *homoskedasticity* (which we have been assuming via A.4), where $\mathrm{var}[\epsilon_i] = \sigma^2$.

In the case of heteroskedasticity, the covariance matrix for the error term takes the form:

$$\sigma^2\Omega = \sigma^2 \begin{bmatrix} \omega_{11} & 0 & \cdots & 0 \\ 0 & \omega_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_{nn} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} = \mathrm{diag}\left(\sigma_i^2\right)$$

When the error term $\boldsymbol{\epsilon}$ exhibits heteroskedasticity, we will find that:

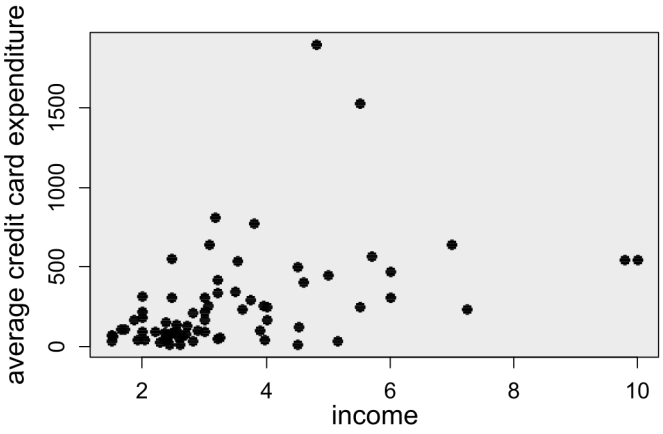
- The LS estimator is still unbiased and consistent.
- The LS estimator is now *inefficient*.
- The usual estimator for $\mathrm{V}[\boldsymbol{b}]$ (which has been $s^2(X'X)^{-1}$ in previous chapters) is now *inconsistent*, which invalidates hypothesis testing.

A solution to the inefficiency of LS is to use the generalized least squares (GLS) or feasible (FGLS) estimator, which also takes care of the inconsistency of the standard errors of $\boldsymbol{b}$. A common practice, however, is to ignore the inefficiency of LS and use a *robust* estimator for $\mathrm{V}[\boldsymbol{b}]$ (such as White's heteroskedastic robust covariance estimator).

Load and plot a dataset that potentially has heteroskedasticity (see Figure 11.1):

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
plot(ccard$income, ccard$avgexp)
```

Figure 11.1: Credit card expenditure data possibly exhibits heteroskedasticity.



11.1 Statistical properties of LS estimation in the presence of heteroskedasticity

Recall the proof that the LS estimator is unbiased:

$$\begin{aligned} \boldsymbol{b} &= (X'X)^{-1} X'\boldsymbol{y} = (X'X)^{-1} X'(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= \boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon} \\ E(\boldsymbol{b}) &= \boldsymbol{\beta} + (X'X)^{-1} X'E(\boldsymbol{\epsilon}) = \boldsymbol{\beta} \end{aligned}$$

We need to use assumption A.3 and A.5 to establish this result, but we do not need A.4. Hence, heteroskedasticity does not affect the unbiasedness property of LS. Similarly, $\text{plim}[\boldsymbol{b}] = \boldsymbol{\beta}$ whether the error term is heteroskedastic or not. (The IV and NLS estimators will also be consistent in the presence of heteroskedasticity).

Now, let's consider the covariance matrix of our LS estimator under heteroskedasticity:

$$\begin{aligned} V(\boldsymbol{b}) &= V\left[\boldsymbol{\beta} + (X'X)^{-1} X'\boldsymbol{\epsilon}\right] = V\left[(X'X)^{-1} X'\boldsymbol{\epsilon}\right] \\ &= \left[(X'X)^{-1} X'V(\boldsymbol{\epsilon})X\,(X'X)^{-1}\right] \\ &= \left[(X'X)^{-1} X'\sigma^2\Omega X\,(X'X)^{-1}\right] \\ &\neq \left[\sigma^2\,(X'X)^{-1}\right] \end{aligned}$$

where we have used assumption 11.1 instead of A.4. We can see that if $\Omega = I_n$ then we get the usual expression for $V(\boldsymbol{b})$.

The usual computer output (for example from `summary()`), will be using $s^2\,(X'X)^{-1}$, which is the *wrong* formula! The standard errors, t-statistics, confidence intervals, will all be incorrect. The usual estimator for the covariance matrix of $\boldsymbol{b}$, namely $s^2\,(X'X)^{-1}$, will be an *inconsistent* estimator of the true covariance matrix of $\boldsymbol{b}$.

The LS estimator will turn out to be *inefficient* under heteroskedasticity, but it is easiest to show this after we develop the generalized least squares (GLS) estimator, and so we postpone this discussion for later. For now, we turn to the most pressing issue - the inconsistency of the estimator for the covariance matrix of $\boldsymbol{b}$.

11.2 White's heteroskedastic consistent covariance matrix

If we knew Σ, then the "estimator" of the covariance matrix for $\boldsymbol{b}$ would just be:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = \left[(X'X)^{-1} X'\Sigma X (X'X)^{-1}\right] \tag{11.2}$$

The covariance matrix in equation 11.2 is known as a *sandwich covariance matrix*. In practice, $\mathrm{V}[\boldsymbol{\epsilon}] = \Sigma$ will usually be unknown and need to be estimated. But since Σ is $n \times n$ and explodes as $n \to \infty$, it seems hopeless to try to get a consistent estimator for Σ. However, we can find a consistent estimator when we consider the entire *middle* of the sandwich.

For asymptotic theory, what we actually need is an estimator for the covariance matrix of $\sqrt{n}(\boldsymbol{b}-\boldsymbol{\beta})$, not $\boldsymbol{\beta}$. By distributing the factor n we can rewrite equation 11.2 as:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = \frac{1}{n}\left[\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\Sigma X\right)\left(\frac{1}{n}X'X\right)^{-1}\right]$$

where we see that we need to find a consistent estimator of $\frac{1}{n}X'\Sigma X$. While Σ is $n \times n$ and explodes as $n \to \infty$, the matrix $\frac{1}{n}X'\Sigma X$ is a $k \times k$ symmetric matrix, and has k distinct elements in the diagonal (with autocorrelation there would be $\frac{1}{2}(k^2+k)$ distinct elements).

Let $Q^* = \left(\frac{1}{n}X'\Sigma X\right)$. In the case of just heteroskedasticity (for autocorrelation we would have $\boldsymbol{x}_i\boldsymbol{x}_j'$ terms), Q^* becomes:

$$Q^* = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\boldsymbol{x}_i\boldsymbol{x}_i'$$

White (1980) showed that if we define

$$S_0 = \frac{1}{n}\sum_{i=1}^{n}e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'$$

then

$$\operatorname{plim}(S_0) = Q^*$$

Since LS is still consistent under heteroskedasticity, the residuals $\boldsymbol{e}$ are still consistent estimators for $\boldsymbol{\epsilon}$. This means that we can estimate the model by LS, get the residuals $\boldsymbol{e}$, and then a consistent estimator of $\mathrm{V}[\boldsymbol{b}]$ will be:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = \frac{1}{n}\left[\left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)\left(\frac{1}{n}X'X\right)^{-1}\right]$$

In practice we ignore the n^{-1} and use:

$$\hat{\mathrm{V}}[\boldsymbol{b}] = (X'X)^{-1}\left(\sum_{i=1}^{n}e_i^2\boldsymbol{x}_i\boldsymbol{x}_i'\right)(X'X)^{-1} \tag{11.3}$$

which amounts to replacing every diagonal element of Σ with a squared residual. The sandwich estimator in 11.3 is called a *heteroskedasticity-consistent covariance matrix estimator*, and is valid regardless of the unknown form of the heteroskedasticity. Taking the square roots of the diagonal elements of 11.3 gives us the het-consistent, or "robust" standard errors.

There are alternatives to the sandwich estimator in 11.3. Alternate versions include multiplying the entire matrix by $n/(n-k)$ as a degrees of freedom correction, or using $e_i^2/(1-h_i)$ instead of just e_i^2,

where h_i is the i^{th} diagonal element of the P_X matrix. All of the alternatives are consistent estimators, and differ in their *finite* sample properties, which vary depending on the data.

As a result of using a sandwich estimator such as in 11.3, the t-statistics, F-statistic, standard errors, confidence intervals, etc. will be modified, but only in a manner that is appropriate asymptotically. This means that the usual test statistics will be unreliable in finite samples, and instead of the t-distribution and F-distribution we should use their asymptotic approximations: the standard Normal and Chi-square distributions.

Example 11.1 — Robust standard errors. Use the credit card expenditure data to estimate the model:

$$\text{avgexp} = \beta_1 + \beta_2 \text{age} + \beta_3 \text{ownrent} + \beta_4 \text{income} + \beta_5 \text{income}^2 + \epsilon$$

Download the data:

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
```

Estimate the model assuming *homoskedasticity*:

```
ccard.mod <- lm(avgexp ~ age + ownrent + income + I(income^2), data = ccard)
summary(ccard.mod)
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -237.147    199.352  -1.190  0.23841
age           -3.082      5.515  -0.559  0.57814
ownrent       27.941     82.922   0.337  0.73721
income       234.347     80.366   2.916  0.00482 **
I(income^2)  -14.997      7.469  -2.008  0.04870 *
```

If we have heteroskedasticity, then the standard errors, t-statistics, and associated p-values, are all wrong! Install and load a package capable of "sandwich" covariance matrix estimation:

```
install.packages("sandwich")
install.packages("lmtest")
library(sandwich)
library(lmtest)
```

and get White's heteroskedastic consistent covariance matrix estimator from equation 11.3 (we can change the `type` to use alternate estimators):

```
coeftest(ccard.mod, vcov = vcovHC(ccard.mod, "HC1"))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -237.1465   220.7950 -1.0741  0.28665
age           -3.0818     3.4226 -0.9004  0.37112
ownrent       27.9409    95.5657  0.2924  0.77090
income       234.3470    92.1226  2.5439  0.01328 *
I(income^2)  -14.9968     7.1990 -2.0832  0.04105 *
```

The standard errors have either decreased or increased, and some are quite different! The significance of one of the regressors has changed, for example. Ignoring the possibility of heteroskedasticity, and thus using the wrong standard errors, can invalidate hypothesis testing.

11.3 Testing for homoskedasticity

Heteroskedasticity reduces the efficiency of the LS estimator of $\boldsymbol{\beta}$ (we still haven't showed this) and has serious implications for the properties of the associated standard errors, confidence intervals, and tests. It would be very useful to have a test of the hypothesis that the errors in our regression model

are homoskedastic, against the alternative that they exhibit some sort of heteroskedasticity. Because LS is still a consistent estimator of $\boldsymbol{\beta}$ even if the errors are heteroskedastic, we can use the LS residuals to construct tests that will still be (at least) asymptotically valid.

11.3.1 White's test

Consider the following null and alternative hypotheses under the standard population model:

$$H_0 : \sigma_i^2 = \sigma^2 \quad ; \quad i = 1, 2, \ldots, n \quad \text{vs.} \quad H_A : \text{ Not } H_0$$

The alternative hypothesis is very general, and no specific form of heteroskedasticity has been declared. To implement the test:

1. Estimate the model by LS, and get the residuals, $e_i \quad ; \quad i = 1, 2, \ldots, n$.
2. Using LS again, regress the e_i^2 values on each of the x's in the original model; their squared values; all of the cross-products of the regressors; and an intercept. We are using the information in X to approximate any possible unknown form of heteroskedasticity.
3. The nR^2 from the regression in Step 2 is asymptotically $\chi^2_{(p)}$ (Chi-square distributed) if H_0 is true; where p is the number of parameters that are estimated at Step 2.
4. Reject H_0 in favour of H_A if the p-value for the nR^2 statistic from the chi-square distribution is small.

Note the limitation of this test:

- It is valid only asymptotically.
- The test is "non-constructive", in the sense that if we reject H_0, we don't know what form of heteroskedasticity we may have.
- This means that it won't be clear what form the GLS estimator (in the next section) should take.

Even though White's test is non-constructive, it can provide enough information to alert us to use White's heteroskedasticity-consistent estimator of $\mathrm{V}(\boldsymbol{b})$. In fact, there is little, if anything, to be lost in using this covariance matrix estimator, as long as the sample is large. This is because homoskedasticity is just a *special case* of heteroskedasticity. That is, the heteroskedastic consistent covariance matrix estimators do not rule out the possibility of homoskedasticity.

Example 11.2 — White's test in R Use the data and model from Example 11.1 to test for the presence of heteroskedasticity:

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
ccard.mod <- lm(avgexp ~ age + ownrent + income + I(income^2), data = ccard)
```

Install and load a package:

```
install.packages("skedastic")
library(skedastic)
```

and calculate White's test using the `white()` function:

```
white(ccard.mod, interactions = TRUE)
```

```
  statistic p.value parameter method       alternative
      <dbl>   <dbl>     <dbl> <chr>        <chr>
1      14.3   0.426        14 White's Test greater
```

The White test statistic is 14.3, with a Chi-square p-value of 0.426. We fail to reject the null of

homoskedasticity after all! Are the degrees of freedom for the Chi-square distribution right?

Example 11.3 — White's test by hand Use the data from Example 11.1 and 11.2 to test for the presence of heteroskedasticity "by hand". Get the squared residuals from the estimated model:

```
ccard <- read.csv("https://rtgodwin.com/data/creditcard.csv")
ccard.mod <- lm(avgexp ~ age + ownrent + income + I(income^2), data = ccard)
ccard.res.sq <- ccard.mod$residuals ^ 2
```

and regress the squared residuals on all regressors, squared regressors, and cross-products:

```
summary(lm(ccard.res.sq ~ age + ownrent + income + I(income^2)
           + I(age^2) + age*ownrent + age*income + age*I(income^2)
           + ownrent^2 + ownrent*income + ownrent*I(income^2)
           + I(income^2) + I(income^3)
           + I(income^4), data=ccard))
```

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          1637390.4  1290979.7   1.268   0.2097
age                     5366.2    48893.8   0.110   0.9130
ownrent                812036.8   991630.2   0.819   0.4161
income               -2021697.6  1053559.1  -1.919   0.0598 .
I(income^2)            669055.3   365666.7   1.830   0.0724 .
I(age^2)                 -424.1      627.5  -0.676   0.5018
I(income^3)            -86805.3    51162.6  -1.697   0.0950 .
I(income^4)              3762.7     2277.4   1.652   0.1038
age:ownrent              4661.7    14424.6   0.323   0.7477
age:income              11499.9    15614.3   0.736   0.4643
age:I(income^2)         -1093.3     1568.1  -0.697   0.4884
ownrent:income        -510192.3   469792.6  -1.086   0.2819
ownrent:I(income^2)     51835.1    61799.8   0.839   0.4050
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 274600 on 59 degrees of freedom
Multiple R-squared:  0.199,     Adjusted R-squared:  0.0361
F-statistic: 1.222 on 12 and 59 DF,  p-value: 0.2905
```

We are essentially looking at the "overall fit" of this auxiliary regression, or the "joint significance" of all of the explanatory variables. Typically we would look at the p-value of 0.2905 for the joint significance. But this is an F-test, and we are in an asymptotic setting. So, instead of the F-test and F-distribution we use the Wald test and the Chi-square distribution. The Wald test statistic is $nR^2 = 72 \times 0.199 = 14.3$ (same as from example 11.2) and the associated p-value is:

```
1 - pchisq(72 * 0.199, 14)
```

```
[1] 0.4255717
```

But the degrees of freedom of 14 is wrong! Two of the cross-products are redundant and have been dropped from the auxiliary regression, leaving us with $p = 12$ and the proper p-value is:

```
1 - pchisq(72 * 0.199, 12)
```

```
[1] 0.280255
```

The `white()` provides the wrong p-value. In any case, we cannot reject the null of homoskedasticity using White's test, even though heteroskedasticity seems apparent from Figure 11.1. What would be

the safe thing to do in this case?

11.4 Generalized least squares

We now turn to the estimation of $\boldsymbol{\beta}$, taking into account when the error term is heteroskedastic. Using this information should enable us to improve the efficiency of the LS estimator.

In the present context, (Ordinary) LS ignores some important information, and we'd anticipate that this will result in a loss of efficiency when estimating $\boldsymbol{\beta}$. Intuitively, observations with less variance should be given more *weight* than observations with high variance. The observations with smaller variance are more "valuable".

Let's see how to obtain an efficient (also linear and unbiased) estimator. Recall that we are allowing a general form for the covariance matrix of the error term: $\mathrm{V}(\boldsymbol{\epsilon}) = \Sigma$. Generally Σ will be unknown. However, to begin with, let's consider the case where it is actually known.

Clearly, Σ must be symmetric, as it is a covariance matrix. Assume that Σ is also positive-definite. Then, Σ^{-1} is also positive-definite, and so there exists a non-singular matrix, P, such that $\Sigma^{-1} = P'P$. Now, if:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

then

$$P = \begin{bmatrix} \frac{1}{\sigma_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{\sigma_n} \end{bmatrix}$$

Consider the population model where all usual assumptions hold *except* A.4:

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad ; \quad \boldsymbol{\epsilon} \sim [0, \Sigma]$$

Recall the Gauss-Markhov theorem in Section 5.3.3. A critical assumption in the Gauss-Markhov theorem is A.4. So, the LS estimator, under heteroskedasticity, is inefficient. However, consider pre-multiplying the population model by P:

$$P\boldsymbol{y} = PX\boldsymbol{\beta} + P\boldsymbol{\epsilon} \tag{11.4}$$

or write:

$$y^* = X^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$$

where * denotes that the variable has been transformed using P. Note that, because P is a diagonal matrix, we are simply scaling the i^{th} observation of all variables by the reciprocal of the square root of each error term's variance:

$$\frac{1}{\sigma_i} y_i = \beta_1 \frac{1}{\sigma_i} + \beta_2 \frac{1}{\sigma_i} x_{i2} + \cdots + \beta_k \frac{1}{\sigma_i} x_{ik} + \frac{1}{\sigma_i} x_{i2} \epsilon_i \tag{11.5}$$

Applying LS to the transformed model in 11.4 (or equivalently model 11.5), yields the generalized least squares (GLS) estimator. This is also known as "weighted least squares", since we are *weighting* each observation using the inverse of it's standard deviation. Note that observations with high variance receive less weight.

Applying LS to the transformed model 11.4 gives the formula for the GLS estimator:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_{\text{GLS}} &= \left[X^{*\prime}X^{*}\right]^{-1}X^{*\prime}\boldsymbol{y}^{*}\\ &= \left[(PX)'(PX)\right]^{-1}(PX)'(P\boldsymbol{y})\\ &= \left[X'P'PX\right]^{-1}X'P'P\boldsymbol{y}\\ &= \left[X'\Sigma^{-1}X\right]^{-1}X'\Sigma^{-1}\boldsymbol{y}\end{aligned} \tag{11.6}$$

11.4.1 Properties of the GLS estimator

Since Σ is a non-random matrix, so is P. This means that if we have assumptions A.3 and A.5 to begin with, transforming the data using P will have no effect on these assumptions. For example:

$$E\left[\boldsymbol{\epsilon}^{*}\right] = E[P\boldsymbol{\epsilon}] = PE[\boldsymbol{\epsilon}] = \mathbf{0}$$

Very importantly, the transformed model attains assumption A.4 (this was the whole point of transforming the model in the first place; to recover A.4):

$$\begin{aligned}V\left[\boldsymbol{\epsilon}^{*}\right] &= V[P\boldsymbol{\epsilon}]\\ &= PV(\boldsymbol{\epsilon})P'\\ &= P\left(\Sigma\right)P' = P\Sigma P'\end{aligned}$$

Because P is both square and non-singular, note that:

$$\begin{aligned}P\Sigma P' &= P\left(\Sigma^{-1}\right)^{-1}P'\\ &= P\left(P'P\right)^{-1}P'\\ &= PP^{-1}\left(P'\right)^{-1}P' = I\end{aligned}$$

and so:

$$V\left[\epsilon^{*}\right] = I \tag{11.7}$$

The transformed model, $y^{*} = X^{*}\boldsymbol{\beta} + \boldsymbol{\epsilon}^{*}$, has an error-term that satisfies the usual assumptions. In particular, the transformed model is homoskedastic. So, if we apply (ordinary) least squares to the model, $y^{*} = X^{*}\boldsymbol{\beta} + \boldsymbol{\epsilon}^{*}$, we'll get the BLU estimator of $\boldsymbol{\beta}$, by the Gauss-Markhov Theorem. This means that ordinary LS, under heteroskedasticity, is inefficient.

Moreover, all of the results that we established with regard to testing for linear restrictions and incorporating them into our estimation, also apply to GLS if we make some obvious modifications. For example, we would estimate estimate $sigma^2$ using $\hat{\sigma}^2 = \boldsymbol{e}'_{GLS}\boldsymbol{e}_{GLS}/(n-k)$, and to test $H_0 : R\boldsymbol{\beta} = \boldsymbol{q} \quad vs. \quad H_A : R\boldsymbol{\beta} \neq \boldsymbol{q}$ we could use the F-statistic $F = (R\widehat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{q})'\left[R\left(X^{*\prime}X^{*}\right)^{-1}R'\right]^{-1}(R\widehat{\boldsymbol{\beta}}_{GLS} - \boldsymbol{q})/J\hat{\sigma}^2$, where X^{*} is the transformed data.

11.4.2 Unknown σ^2

In this section we highlight that, in order to perform GLS, all we really need to know is the *proportionality* of the variances between observations, not the actual variance.

An important difference between the GLS estimator in equation 11.6, and the ordinary LS estimator, is that it appears to require that σ^2 is known to be 1 (in equation 11.7, $\sigma^2 = 1$). But, if we write:

$$V[\boldsymbol{\epsilon}] = \Sigma = \sigma^2\Omega$$

then we will see that as long as Ω is known, we can obtain the GLS estimates. That is, all we need to know is the *proportionality* of the difference in variance between observations. For example, if we knew that some observations had twice as much variance as others, we could perform GLS without knowing the exact magnitude of the variances. If σ^2 is unknown but Ω is known to be:

$$\Omega = \begin{bmatrix} \omega_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \omega_{nn} \end{bmatrix}$$

then the P matrix used to transform the data can instead be written as:

$$P = \begin{bmatrix} {\omega_{11}}^{-1/2} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & {\omega_{nn}}^{-1/2} \end{bmatrix}$$

The GLS estimator becomes:

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_{\text{GLS}} &= \left[X'\Sigma^{-1}X\right]^{-1} X'\Sigma^{-1}\boldsymbol{y} \\ &= \left[X'\left(\sigma^2\Omega\right)^{-1}X\right]^{-1} X'\left(\sigma^2\Omega\right)^{-1}\boldsymbol{y} \\ &= \left[X'\Omega^{-1}X\right]^{-1} X'\Omega^{-1}\boldsymbol{y} \end{aligned} \tag{11.8}$$

which amounts to weighting the data by the inverse of the square root of the proportionality constants ω_{ii}:

$$\omega_{ii}^{-\frac{1}{2}} y_i = \beta_1 \omega_{ii}^{-\frac{1}{2}} + \beta_2 \left(\omega_{ii}^{-\frac{1}{2}} x_{i2}\right) + \cdots + \left(\omega_{ii}^{-\frac{1}{2}} \epsilon_i\right)$$

Now we can see that the error term in the transformed model has variance $\text{V}[\boldsymbol{\epsilon}^*] = \sigma^2 I_n$, when Σ is unknown, but Ω is known.

11.4.3 Clustering

In some cases we will actually know the form of the heteroskedasticity, so we can apply GLS directly. Suppose that we have the usual population model:

$$\begin{aligned} &y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &E\left[\epsilon_i\right] = 0 \quad ; \quad \text{var}.\left[\epsilon_i\right] = \sigma^2 \quad ; \quad \text{i.i.d} \end{aligned}$$

However, suppose that we only observe "grouped" data, rather than the observations on the individual agents. This happens frequently in practice, when data are released in this way to preserve confidentiality.

Suppose there are m groups (e.g., income groups), with n_j observations in the j^{th} group; $j = 1, 2, \dots, m$. The model that we can *actually* estimate is of the form:

$$\bar{y}_j = \beta_1 + \beta_2 \bar{x}_{j2} + \cdots + \beta_k \bar{x}_{jk} + \bar{\epsilon}_j \quad ; \quad j = 1, 2, \dots, m$$

That is, rather than seeing the data at an individual level, we are seeing the data averaged over each group. Clearly, averaging the data over the groups does not change the fact that the error term is still mean zero:

$$E\left[\bar{\epsilon}_j\right] = E\left[\frac{1}{n_j}\sum_{i=1}^{n_j}\epsilon_i\right] = \left[\frac{1}{n_j}\sum_{i=1}^{n_j} E\left(\epsilon_i\right)\right] = 0$$

but it does change the *variance* of the error term across groups:

$$\begin{aligned}\text{var}\cdot[\bar{\epsilon}_i] &= \text{var}\cdot\left[\frac{1}{n_j}\sum_{i=1}^{n_j}\epsilon_i\right] = \left[\frac{1}{n_j^2}\sum_{i=1}^{n_j}\text{var}\,.\,(\epsilon_i)\right] \\ &= \left(n_j\sigma^2/n_j^2\right) \\ &= \left(\sigma^2/n_j\right).\end{aligned}$$

The n_j values are generally reported, so we know the *proportionality* of the error covariance matrix:

$$\sigma^2\Omega = \sigma^2\begin{bmatrix} 1/n_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/n_m \end{bmatrix}$$

Because Ω is known, we can compute the GLS estimator for $\boldsymbol{\beta}$ directly using $\widehat{\boldsymbol{\beta}}_{GLS} = \left[X'\Omega^{-1}X\right]^{-1}X'\Omega^{-1}\boldsymbol{y}$.

11.5 Feasible generalized least squares (FGLS)

In many cases the Ω matrix will be unknown. In order to be able to implement the GLS estimator, in practice, we're usually going to have to provide a suitable estimator of Ω. We'll want to obtain an estimator that is at least consistent, and place this into the formula for the GLS estimator, yielding:

$$\widehat{\boldsymbol{\beta}}_{FGLS} = \left[X'\widehat{\Omega}^{-1}X\right]^{-1}X'\widehat{\Omega}^{-1}\boldsymbol{y}$$

A problem is that the Ω matrix is $(n \times n)$, and it has n distinct elements. However, we have only n observations on the data. This precludes obtaining a consistent estimator. We need to constrain the elements of Ω in some way. In practice, this won't be a big problem, because usually there will be lots of "structure" on the form of Ω. Typically, we'll have $\Omega = \Omega(\boldsymbol{\theta})$, where the vector, $\boldsymbol{\theta}$ has low dimension.

For example, we can specify the *skedastic* function (a function that determines a variable's conditional variance):

$$\text{var}(\epsilon_i) = \exp(\boldsymbol{z}_i\boldsymbol{\theta}) \tag{11.9}$$

where $\boldsymbol{z}_i$ are regressors that may contain some or all of $\boldsymbol{x}_i$, $\boldsymbol{\theta}$ is a parameter vector to be estimated, and the exponent keeps the values of the function positive for any $\boldsymbol{\theta}$ (variances must be positive). To obtain a consistent estimator of $\boldsymbol{\theta}$, we can run the auxiliary regression using LS:

$$\log e_i^2 = \boldsymbol{z}_i\boldsymbol{\theta} + \varepsilon_i$$

to find the estimates $\widehat{\boldsymbol{\theta}}$. The elements of the Ω matrix can then be estimated using the predicted values from the above regression:

$$\widehat{\omega}_{ii} = \left(\exp\left(\boldsymbol{z}_i\widehat{\boldsymbol{\theta}}\right)\right)^{1/2} \tag{11.10}$$

finally, the FGLS estimates are obtained by applying LS to data that has been weighted using the fitted values $\widehat{\omega}_{ii}$ from equation 11.10.

Approaches to obtain $\widehat{\Omega}$ using the LS residuals, such as above, are valid because LS is consistent even in the presence of heteroskedasticity. This extends to the residuals: they are consistent estimators for the unknown error term. The residuals can be used to test for the presence of heteroskedasticity (as in White's test), construct heteroskedastic-robust covariance estimators, and as we have now just seen, estimate the unknown form of heteroskedasticity.

Obtaining a *consistent* estimator for Ω in turn ensures that the FGLS estimator is also consistent (the proof is difficult and not shown here). The FGLS estimator will also be asymptotically efficient. Little can be said about its finite sample properties however, and it will usually be biased, with the bias depending on the form of Ω and our choice of $\widehat{\Omega}$.

11.6 Exercises

1. Explain the difference between homoskedasticity and heteroskedasticity.
2. What are the consequences of heteroskedasticity?
3. Briefly explain some of the ways that we can deal with heteroskedasticity.
4. Derive the covariance matrix for $\boldsymbol{b}$ under the assumption of heteroskedasticity.
5. What is White's heteroskedastic consistent covariance matrix estimator?
6. Should the t-distribution / F-distribution be used in hypothesis testing, after using White's estimator for standard errors?
7. Explain White's test for heteroskedasticity. What is the null and alternative? How can you implement the test? What is the intuition behind the test?
8. Derive the GLS estimator, by transforming the error term so that it is homoskedastic.
9. Show that LS is inefficient in the presence of heteroskedasticity, by arguing that GLS is efficient.
10. Explain how we only need to know the *relative* differences in variance between observations (the proportionality of the variances), in order to perform GLS.
11. Show how to implement GLS when the data is "clustered".
12. Explain the difference between GLS and FGLS.
13. Using the data from Example 11.1, compare the p-value for the significance of the *age* variable using the regular covariance matrix estimator (under homoskedasticity), to the p-value using White's robust standard errors.
14. Using the following code to estimate a wage model:

```
cps <- read.csv("http://rtgodwin.com/data/cps1985.csv")
cps.mod <- lm(log(wage) ~ education + gender + age + experience
                        + gender * education, data = cps)
summary(cps.mod)
```

    test for heteroskedasticity. Use White's heteroskedastic robust standard errors, regardless of the results of the test. Which variables are significant under homoskedasticity vs. heteroskedasticity?

Chapter 12

Introduction to Time Series

This is not a comprehensive introduction. This is a short collection of topics that I find important and interesting, namely (i) the inconsistency of LS under lagged dependent variables and autocorrelated errors, and (ii) spurious regressions due to random walks.

12.1 What is a time series

A time series is a single occurrence of a random event. The sequence of observations, $\{y_t\}_{t=-\infty}^{t=+\infty}$, is a time series process. There is no counterpart to repeated sampling for a time series. We observe realizations of this process in a time window, $t = 1, \ldots, T$. The frequency of observations are important, but the length of the window is arguably more important. Asymptotics involves considering an increasingly longer window.

Quarterly time series data on Canadian GDP, CPI, unemployment rate, and average target interest rates from 1993Q1 to 2022Q2 (see Figure 12.1):

```
can <- read.csv("https://rtgodwin.com/data/canseries.csv")
plot(can$time, can$CPI, type="l")
plot(can$time, can$GDP, type="l")
plot(can$time, can$unemployment, type="l")
plot(can$time, can$interest, type="l")
```

Time series models typically seek to *forecast* (predict) future values of the series, and sometimes look to estimate causal effects of policies or interventions (like the effect of interest rates on inflation). In order to model a time series, y_t, it is typical to include:

- contemporaneous factors, x_t
- lagged factors, $x_{t-1}, x_{t-2}, \ldots$
- its own past, $y_{t-1}, y_{t-2}, \ldots$
- a time trend, t
- seasonal dummies, d
- disturbances (innovations), ϵ_t

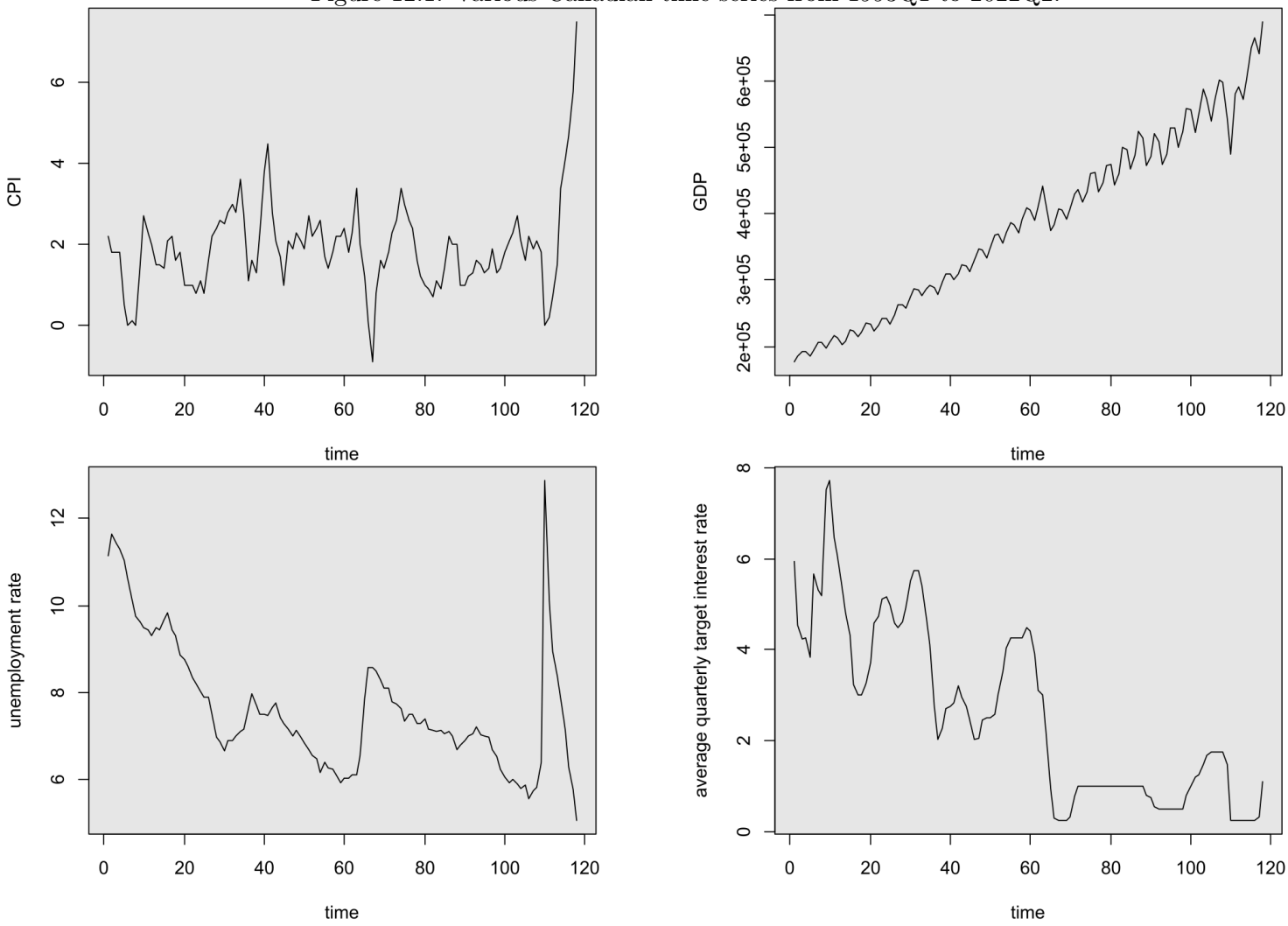
For example:

$$y_t = \beta_1 + \beta_2 t + \beta_3 d_t + \beta_4 x_t + \beta_5 y_{t-1} + \epsilon_t$$

12.2 Autocorrelation

In time series models, the past is very important! Often a very good predictor of y_t are the $y_{t-1}, y_{t-2}, \ldots$. This suggests that the error term, ϵ_t, also depends on its own past values. That is, even after including

Figure 12.1: Various Canadian time series from 1993Q1 to 2022Q2.



other time series (x_t) in the regression, there are likely still missing time series in ϵ_t that depend on their own past. This means that the error term will very likely violate A.4. The error term is correlated to its own past values. This is called serial correlation, or autocorrelation.

In the case of autocorrelation, the off-diagonal elements of $V(\boldsymbol{\epsilon})$ will be non-zero. The particular values they take will depend on the form of autocorrelation. That is, they will depend on the pattern of the correlations between the elements of the error vector. For example, the covariance matrix of the error term will look like:

$$V(\boldsymbol{\epsilon}) = \begin{bmatrix} \sigma^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma^2 & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma^2 \end{bmatrix}$$

If the errors themselves are autocorrelated, often this will be reflected in the regression residuals also being autocorrelated. That is, the residuals will follow some sort of pattern, rather than just being random. Let's take a look at the quarterly Canadian GDP series. We know from Section 4.4.2 how to de-seasonalize (and de-trend) a time series:

```
# Create quarterly dummies
n <- 118
```

```
can$q4 <- can$q3 <- can$q2 <- can$q1 <- 0
can$q1[seq(1, n, 4)] <- 1
can$q2[seq(2, n, 4)] <- 1
can$q3[seq(3, n, 4)] <- 1
can$q4[seq(4, n, 4)] <- 1

# Regress GDP on a time trend and quarterly dummies
gdp.mod <- lm(can$GDP ~ can$time + can$q2 + can$q3 + can$q4)

# Collect the residuals, which are de-trended, de-seasonalized GDP
gdp.resid <- gdp.mod$residuals
```

Now that we have de-trended and de-seasonalized GDP, let's try to explain GDP in terms of its past values. We'll estimate the equation (where $GDP^\star$ is the de-trended, de-seasonlaized series):

$$GDP_t^\star = \beta_0 + \beta_1 GDP_{t-1}^\star + \beta_2 GDP_{t-2}^\star + \beta_3 GDP_{t-3}^\star + \beta_4 GDP_{t-4}^\star + \beta_5 GDP_{t-5}^\star + \epsilon_t$$

We need to regress GDP on it's lagged values (notice that we lose 5 observations):

```
gdp.mod.lag <- lm(gdp.resid[6:118] ~ gdp.resid[5:117] + gdp.resid[4:116]
                + gdp.resid[3:115] + gdp.resid[2:114] + gdp.resid[1:113])
summary(gdp.mod.lag)
```

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.7951  1181.7694  -0.005  0.99610
gdp.resid[5:117]    0.9013     0.1003   8.985    1e-14 ***
gdp.resid[4:116]   -0.3254     0.1336  -2.435  0.01655 *
gdp.resid[3:115]    0.2336     0.1389   1.682  0.09551 .
gdp.resid[2:114]    0.1455     0.1343   1.083  0.28111
gdp.resid[1:113]   -0.3140     0.1050  -2.989  0.00347 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12330 on 107 degrees of freedom
Multiple R-squared:  0.5273,    Adjusted R-squared:  0.5052
F-statistic: 23.87 on 5 and 107 DF,  p-value: 4.666e-16
```

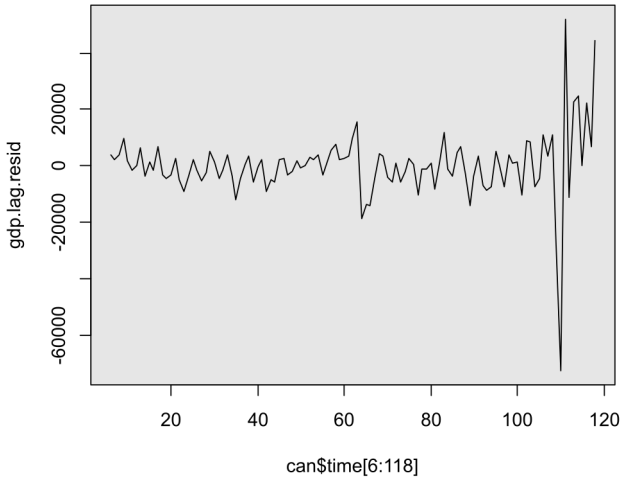
Now, we'll once again get the residuals from this model, and plot them over time (see Figure 12.2):

```
gdp.lag.resid <- gdp.mod.lag$residuals
plot(can$time[6:118], gdp.lag.resid, type="l")
```

Figure 12.2 shows some sort of patern in the residuals, indicating that the error term is autocorrelated. The time series model that we estimated is simple, and we could include explanatory $x_t, x_{t-1}, \ldots$ variables. Even when we do so, however, we are unlikely to be able to completely account for the autocorrelation apparent in the residuals.

If the errors of our model are autocorrelated, then the OLS estimator of $\boldsymbol{\beta}$ usually will be unbiased and consistent, but it will be inefficient. In addition $V(\boldsymbol{\beta})$ will be computed incorrectly, and the standard errors, etc., will be inconsistent. (Same situation as with heteroskedasticity). In general time series models are concerned with testing for the presence/absence of autocorrelation, estimating the form of autocorrelation (the $V(\boldsymbol{\beta})$ matrix), and then estimating models where the errors are autocorrelated. In this introduction, we will not look at these methods. We will consider two ways of modelling autocorrelation: an AR process and an MA process. We will also consider a limiting form of an AR process.

Figure 12.2: The residuals from our time series model are autocorrelated.



12.2.1 Autoregressive process

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \quad ; \quad u_t \sim \text{ i.i.d. } N\left[0, \sigma_u^2\right] \quad ; \quad |\rho| < 1$$

This is an AR(1) model for the error process. More generally:

$$\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + \cdots + \rho_p\epsilon_{t-p} + u_t \quad ; \quad u_t \sim \text{ i.i.d. } N\left[0, \sigma_u^2\right]$$

This is an AR(p) model for the error process. [e.g., p = 4 with quarterly data.] An AR process can be used to model any time series, not just the error term. Notice that we used an AR(5) process in our GDP example above.

12.2.2 Moving average process

$$\epsilon_t = u_t + \phi u_{t-1} \quad ; \quad u_t \sim \text{ i.i.d.N } \left[0, \sigma_u^2\right]$$

This is an MA(1) model for the error process. More generally:

$$\epsilon_t = u_t + \phi_1\epsilon_{t-1} + \cdots + \phi_q u_{t-q} \quad ; \quad u_t \sim \text{ i.i.d. } N\left[0, \sigma_u^2\right]$$

This is an MA(q) model for the error process. We can combine both types of process into an ARMA(p, q) model:

$$\epsilon_t = \rho_1\epsilon_{t-1} + \rho_2\epsilon_{t-2} + \cdots + \rho_p\epsilon_{t-p} + u_t + \phi_1 u_{t-1} + \cdots + \phi_q u_{t-q} \quad ; \quad u_t \sim \text{ i.i.d. } N\left[0, \sigma_u^2\right]$$

12.2.3 Stationarity

Note that in the AR(1) process, we said that $|\rho| < 1$. This condition is needed to ensure that the process is "stationary." Suppose that the following three conditions are satisfied:

1. $E\left[\epsilon_t\right] = 0 \quad ; \quad \text{for all } t$
2. $\text{var.}\left[\epsilon_t\right] = \sigma^2 \quad ; \quad \text{for all } t$
3. $\text{cov}\cdot\left[\epsilon_t, \epsilon_s\right] = \gamma_{|t-s|} \quad ; \quad \text{for all } t, s; t \neq s$

Then we say that the time-series sequence, ϵ_t, is "Covariance Stationary"; or "Weakly Stationary".

Unless a time-series is stationary, we can't identify and estimate the parameters of the process that is generating its values.

Let's see how this notion relates to the AR(1) model, introduced above. We have:

$$\begin{aligned}
\epsilon_t &= \rho\epsilon_{t-1} + u_t \\
\epsilon_t &= \rho\left[\rho\epsilon_{t-2} + u_{t-1}\right] + u_t \\
&= \rho^2\epsilon_{t-2} + \rho u_{t-1} + u_t \\
&= \rho^2\left[\rho\epsilon_{t-3} + u_{t-2}\right] + \rho u_{t-1} + u_t \\
&= \rho^3\epsilon_{t-3} + \rho^2 u_{t-2} + \rho u_{t-1} + u_t
\end{aligned}$$

Continuing in this way, eventually, we get:

$$\epsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \cdots \tag{12.1}$$

This is an infinite order MA process. The value of ϵ_t embodies the entire past history of the u_t values.

From equation 12.1, $\mathbb{E}(\epsilon_t) = 0$, and:

$$\begin{aligned}
\text{var}\,(\epsilon_t) &= \text{var}\,.\,(u_t) + \text{var}\,.\,(\rho u_{t-1}) + \text{var}\,.\left(\rho^2\epsilon_{t-2}\right) + \cdots \\
&= \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \cdots
\end{aligned}$$

Question: Under what conditions will this series *converge*?

The series will converge to $\sigma_u^2\left(1-\rho^2\right)^{-1}$, as long as $|\rho^2| < 1$, and this in turn requires that $|\rho| < 1$.

This is a necessary condition needed to ensure that the process ϵ_t is stationary, because if this condition isn't satisfied, then $\text{var}(\epsilon_t)$ is infinite.

If we have a (stationary) AR(1) process, for example, then it can be shown that the covariance matrix for $\boldsymbol{\epsilon}$ is:

$$V(\boldsymbol{\epsilon}) = \sigma_u^2\Omega = \frac{\sigma_u^2}{(1-\rho^2)}\begin{bmatrix} 1 & \rho & \cdots & \rho^{n-1} \\ \rho & 1 & \ddots & \rho^{n-2} \\ \vdots & \ddots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \ldots & 1 \end{bmatrix} \tag{12.2}$$

So, if we can estimate ρ (for example by looking at the residuals), we can estimate the entire covariance matrix and use FGLS to obtain an estimator that is efficient and has consistent standard errors.

12.3 Inconsistency of LS with AR errors and lagged dependent variables

In the presence of autocorrelation, in general $\boldsymbol{b}$ will still be a consistent estimator. However, there is one important situation where it will be inconsistent. This will be the case if the errors are autocorrelated, and one or more lagged values of the dependent variable enter the model as regressors. A quick way to observe that inconsistent estimation will result in this case is as follows. Suppose that:

$$\begin{aligned}
y_t &= \beta y_{t-1} + \epsilon_t \quad ; \quad |\beta| < 1 \\
\epsilon_t &= \rho\epsilon_{t-1} + u_t \quad ; \quad u_t \sim \text{ i.i.d. } \left[0, \sigma_u^2\right] \quad ; \quad |\rho| < 1
\end{aligned} \tag{12.3}$$

Now subtract ρy_{t-1} from the expression for y_t in equation 12.3:

$$(y_t - \rho y_{t-1}) = (\beta y_{t-1} + \epsilon_t) - \rho (\beta y_{t-2} + \epsilon_{t-1}) \tag{12.4}$$

or:

$$\begin{aligned} y_t &= (\beta + \rho) y_{t-1} - \beta\rho y_{t-2} + (\epsilon_t - \rho\epsilon_{t-1}) \\ &= (\beta + \rho) y_{t-1} - \beta\rho y_{t-2} + u_t \end{aligned} \tag{12.5}$$

So, if we estimate the model with just y_{t-1} as the only regressor, then we are effectively omitting a relevant regressor, y_{t-2}, from the model. This amounts to imposing a false (zero) restriction on the coefficient vector, and we know that this causes OLS to be not only biased, but also inconsistent.

12.4 Random walk

Working with non-stationary data is dangerous. A non-stationary time series where:

$$y_t = y_{t-1} + \epsilon_t$$

is said to be integrated of order one I(1), or is said to follow a *random walk*. The danger arises as we can easily find what is called a "spurious relationship" between two unrelated random walks. For example, suppose that x_t also follows a random walk:

$$x_t = x_{t-1} + \varepsilon_t$$

Remember that both of these series have infinite variance! Now, what happens if we regress one random walk on another unrelated random walk? We shouldn't find any relationship, because the two series are unrelated, right? Right?? Granger and Newbold[1] performed a simple simulation showing that such a regression yields spurious results.

Try running the following code in R, with increasing sample sizes:

```
n <- 100
y <- x <- 0
for(i in 2:n){
 y[i] <- y[i - 1] + rnorm(1)
 x[i] <- x[i - 1] + rnorm(1)
}
plot(y, type = "l", col = "red", ylim = c(min(x,y),max(x,y)))
points(x, type = "l", col = "blue")
summary(lm(y ~ x))
```

As the sample size increases, the t-statistic on β_1 goes to infinity, and the R^2 goes to 1, even though there is no relationship! This is because both series have infinite variance. For example, let's try (see Figure 12.3 for the two generated random walks):

```
set.seed(7010)
n <- 5000
y <- x <- 0
for(i in 2:n){
  y[i] <- y[i - 1] + rnorm(1)
  x[i] <- x[i - 1] + rnorm(1)
}
plot(y, type = "l", col = "red", ylim = c(min(x,y),max(x,y)))
points(x, type = "l", col = "blue")
```

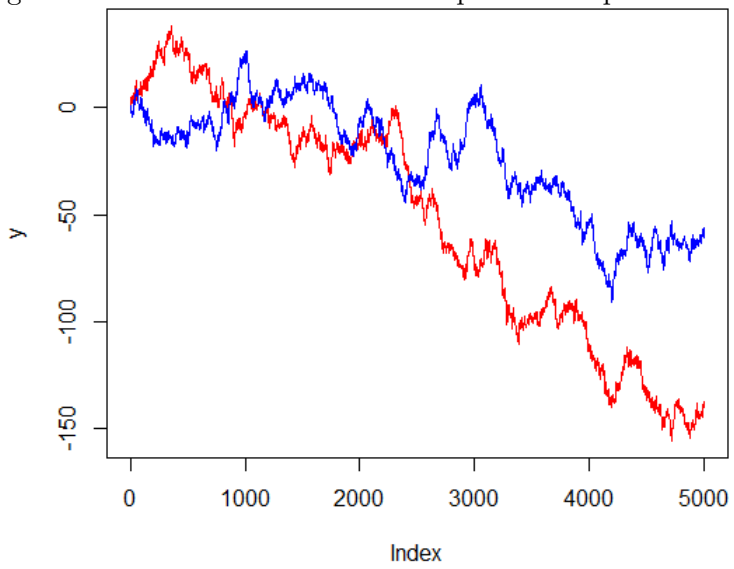
[1] Granger, C. W., & Newbold, P. (1974). Spurious regressions in econometrics. Journal of econometrics, 2(2), 111-120.

```
summary(lm(y ~ x))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.94998    0.58101  -18.85   <2e-16 ***
x             1.72176    0.01613  106.74   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 30.54 on 4998 degrees of freedom
Multiple R-squared:  0.6951,    Adjusted R-squared:  0.695
F-statistic: 1.139e+04 on 1 and 4998 DF,  p-value: < 2.2e-16
```

Figure 12.3: Two random walks that produce a spurious relationship.



A common way to deal with a random walk is to *first difference* the data. This will ensure that we don't find a spurious relationship (as long as we have the order of integration correct):

```
yd <- y[2:5000] - y[1:4999]
xd <- x[2:5000] - x[1:4999]
summary(lm(yd ~ xd))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.02783    0.01414  -1.968   0.0491 *
xd           0.01928    0.01418   1.359   0.1741
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9998 on 4997 degrees of freedom
Multiple R-squared:  0.0003697, Adjusted R-squared:  0.0001696
F-statistic: 1.848 on 1 and 4997 DF,  p-value: 0.1741
```

12.5 Exercises

1. How can you de-seasonalize and de-trend a time series? Why would you want to do so?
2. How can we tell if the error term is autocorrelated?
3. What are some of the consequences of autocorrelation?

4. What is the difference between an AR(1) and MA(1) process?
5. Show that an AR process has "infinite memory"? In words, what does this mean?
6. Derive the variance of an AR(1) error term, and note the condition on ρ so that the series converges.
7. Show that LS is inconsistent when the model includes a lagged dependent variable, and the error term follows an AR(1) process.
8. Show that the variance of a random walk is infinite.
9. Explain a "spurious regression" in the context of two random walks.

Chapter 13

Maximum likelihood estimation

Least squares does not work when the dependent variable ($\boldsymbol{y}$) is:

1. The length of time it takes for something to happen (a strike, an insurance claim, an unemployment spell)
2. The number of things that happen (doctor visits, number of customers, bank failures, number of patents)
3. A yes/no dummy variable (whether person is in the labour force, whether a customer makes a purchase)

In these cases, the dependent variable is *limited* in some way. In (1) the time must be positive, $y_i \geq 0$. In (2) the counts are not continuous and non-negative, $y_i = 0, 1, 2, \ldots$. In (3) the values take on only 0 or 1. The linear model, and least-squares, has no way of recognizing or accounting for the limited nature of the dependent variable. A model estimated by LS will provide predicted values that are not allowed for y, and in most cases is misspecified so that the LS estimator is inconsistent.

In cases such as above, if we are willing to specify a *probability* distribution for y, then we can use maximum likelihood estimation (MLE). (Anytime we are willing to choose a distribution for y, we can use MLE). MLE is not the only option available: GMM, Bayesian, non-parametric, and others; but MLE has excellent properties and is a popular estimation strategy.

- MLE proposed by R. A. Fisher, 1921-1925.
- MLE is a parametric method.
- That is, we assume each sample data is generated from a known probability distribution function (pdf), $p(y_i \mid \boldsymbol{\theta})$. i.e. y_i comes from a "family".

Consider that we have random data $\boldsymbol{y} = \{y_1, \ldots, y_n\}$, and a parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$. Our objective is to estimate $\boldsymbol{\theta}$. The probability of jointly observing the data is:

$$p(y_1, \ldots, y_n \mid \boldsymbol{\theta}) \qquad \text{"joint pdf"}$$

We can view $p(y_1, \ldots, y_n \mid \boldsymbol{\theta})$ in two different ways:

1. As a function of $\{y_1, \ldots, y_n\}$, given $\boldsymbol{\theta}$.
2. As a function of $(\theta_1, \ldots, \theta_k)$, given $\boldsymbol{y}$. i.e., the data are given, the parameters vary.

The latter is called the **likelihood function**. Note:

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta} \mid y_1, \ldots, y_n) = p(y_1, \ldots, y_n \mid \boldsymbol{\theta})$$

Definition 13.1 — Maximum likelihood estimator (MLE) The MLE of $\boldsymbol{\theta}$ (call it $\tilde{\boldsymbol{\theta}}$) is that value of $\boldsymbol{\theta}$ such that $L(\tilde{\boldsymbol{\theta}}) > L(\hat{\boldsymbol{\theta}})$, for all other $\hat{\boldsymbol{\theta}}$.

The idea behind MLE: "given the y_i's, what is the most likely $\boldsymbol{\theta}$ to have generated such a sample?"

Note:

- $\tilde{\boldsymbol{\theta}}$ need not be unique.
- $\tilde{\boldsymbol{\theta}}$ should locate the global max. of $L(\boldsymbol{\theta})$.
- If the sample data are independent then $L(\boldsymbol{\theta} \mid \boldsymbol{y}) = p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta})$.
- Any monotonic transformation of $L(\boldsymbol{\theta})$ leaves the location of the extremum unchanged, e.g. $\log L(\boldsymbol{\theta})$

13.1 Some basic concepts and notation

1. Gradient/score vector: $\left[\frac{\partial \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] \quad (k \times 1)$
2. Hessian matrix: $\left[\frac{\partial^2 \log L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}\right] \quad (k \times k)$
3. Likelihood equations: $\frac{\partial \log L(\theta)}{\partial \theta} = 0 \quad (k \times 1)$

The optimization problem is:

$$\max_{\boldsymbol{\theta}} \prod_{i=1}^{n} L(\boldsymbol{\theta} \mid y_i)$$

To obtain the MLE, $\widetilde{\boldsymbol{\theta}}$, we solve the likelihood equation(s) and then check the second-order condition(s) to make sure we have maximized (not minimized) $L(\boldsymbol{\theta})$. If the Hessian matrix is at least n.s.d., then $L(\boldsymbol{\theta})$ is concave, and this is sufficient for a maximum. So, MLE is accomplished by:

1. Specifying the likelihood function.
   - This involves writing down an equation which states the joint likelihood (or joint probability) of observing the sample data, conditional on the unknown parameter values of the probability distribution function.
   - Independence of the y data is usually assumed (and will be for the purposes of this course).
   - Given independence, the likelihood function is obtained by multiplying together the probability of each y_i occurring.
2. Taking the natural log of the likelihood function. This usually simplifies the next step. The location of the maximum will not change.
3. Taking the first derivative of the log-likelihood function with respect to all parameters, setting each derivative equal to zero, and solving for the parameter values. The solution of the FOCs provides the formulas for the MLEs.
4. Checking to make sure the estimator in (3) attains a maximum (not a minimum). This involves taking the second derivatives of the log-likelihood function with respect to all parameters, so as to construct the Hessian matrix. If the Hessian is n.s.d., then the MLE achieves a global max.
5. Obtaining the variance of the MLEs for use in hypothesis testing. A variance-covariance matrix can be found by inverting the negative of the expected Hessian.

13.2 Properties of MLE

- MLE has very desirable asymptotic properties.

- Namely, MLE is Best Asymptotically Normal.
- That is, under mild assumptions, ML estimators are consistent, asymptotically efficient, and asymptotically Normally distributed.
- These properties are obtained by examining the asymptotic distribution of the MLE (which we will not derive in class):

$$\sqrt{n}\left(\tilde{\theta}-\theta_0\right) \xrightarrow{d} N\left[0, IA^{-1}(\theta)\right]$$

where

$$IA^{-1}(\theta) = \lim_{n\to\infty}\left(\frac{1}{n}[-E[H(\theta)]]^{-1}\right)$$

- $IA^{-1}(\theta)$ is the asymptotic information matrix, and $H(\theta)$ is the Hessian.
- The statement of the asymptotic distribution shows that the MLEs are consistent, asymptotically normal, and asymptotically efficient.
- The efficiency result relies on the Cramer-Rao lower bound. The Cramer-Rao lower bound is a theoretical minimum variance that any estimator can obtain. The MLE attains this minimum, that is, $IA^{-1}(\theta)$ is equal to the asymptotic Cramer-Rao lower bound.

The asymptotic distribution also allows us to see the variance of the MLEs in finite samples. The variance-covariance of $\widetilde{\boldsymbol{\theta}}$ for finite samples can be solved from the asymptotic variance:

$$\text{var}[\sqrt{n}(\tilde{\theta})] = n \times \text{var}(\tilde{\theta}) = \frac{1}{n}[-E[H(\theta)]]^{-1}$$

so,

$$\text{var}(\tilde{\theta}) = [-E[H(\theta)]]^{-1}$$

The matrix $-E[H]$ is termed the "Information Matrix" and is denoted by $I(\theta)$.

A very useful property of MLEs is their "invariance." That is, the estimator for $g(\theta)$ is $g(\tilde{\theta})$. Hence, an estimator for the variance-covariance of $\tilde{\theta}$ is:

$$\widetilde{\text{var}(\tilde{\theta})} = [-E[H(\tilde{\theta})]]^{-1}$$

Note that if misspecification occurs (if we have selected the wrong probability density function to begin with), we are not assured of any of the asymptotic properties.

13.2.1 Finite sample properties of MLEs

MLEs can be biased in finite samples (and typically are). We can evaluate bias much like we have done in previous parts of the course; by taking $E(\tilde{\theta})$. This knowledge can be used to correct for any bias (as in the case of $\tilde{\sigma}^2$). However, in most cases, there is no closed-form solution for the MLE itself, and numerical methods must be used to solve for the estimate. When the estimator does not have a closed form solution, we cannot take $E(\tilde{\theta})$, and we will not be able to "see" whether or not the estimator is biased. In this case, approximations or Monte Carlo experiments may be used to evaluate bias.

13.3 Application of MLE: count data

There are many instances in econometrics where the variable that we want to explain is a count variable, i.e. $y = 0, 1, 2, \ldots$. Examples of some cases are:

- calls at a call-centre
- number of customers

- doctor visits
- bank failures
- insurance claims
- patents

LS can be inconsistent when the the dependent variable is a count, and does not provide a "fitted" model that is useful. Instead of LS, we could use a "count data" model. The most basic of count data models is the Poisson regression model.

13.3.1 Poisson distribution

When estimating a model by maximum likelihood, we first need to pick a suitable distribution for describing the y variable. We start with the Poisson distribution, which describes the number of events that will happen over some fixed interval (usually an interval of time). If the events are assumed to be independent from one another, and the times between events are exponentially distributed, then we get a Poisson distribution:

$$f(y_i \mid \lambda) = \frac{\lambda^{y_i}}{e^{\lambda} y_i!} \quad ; \quad y_i = 0, 1, 2, \ldots \quad ; \quad \lambda > 0 \tag{13.1}$$

The mean and variance of this distribution is λ.[1] If we had an estimate for the mean λ then we would know everything about the counting process; e.g. the probability of more than 4 customers showing up in a day.

13.3.2 Maximum likelihood estimation of the Poisson distribution

Suppose we have a sample of data, $\boldsymbol{y}$. The MLE will tell us the value of λ that is most likely to have "generated" the sample that we have observed.
The first step is to determine the *joint log likelihood*. Assuming independence of the y_is, the joint log-likelihood is:

$$l(\lambda \mid y_i) = \sum_{i=1}^{n} (y_i \log \lambda - \lambda - \log y_i!)$$

or:

$$l(\lambda \mid y_i) = \sum_{i=1}^{n} y_i \log \lambda - n\lambda - \sum_{i=1}^{n} \log y_i! \tag{13.2}$$

Now, taking the derivative of 13.2 with respect to λ we get:

$$\frac{\partial l}{\partial \lambda} = \frac{\sum_{i=1}^{n} y_i}{\lambda} - n \tag{13.3}$$

Setting 13.3 equal to zero for the FOC, and solving for λ, yields:

$$\tilde{\lambda} = \bar{y} \tag{13.4}$$

In order to verify that 13.4 is the MLE for λ we take the second derivative of 13.3:

$$\frac{\partial^2 l}{\partial \lambda^2} = -\frac{\sum_{i=1}^{n} y_i}{\lambda^2} \tag{13.5}$$

Since 13.5 is negative, the log-likelihood is concave and 13.4 solves for the global maximum. Note that 13.5 is the (scalar) Hessian matrix, H.

[1] This equi-dispersion property proves too restrictive for most applications.

13.3.3 The variance of $\tilde{\lambda}$

The variance of an MLE may be found by taking the inverse of the negative of the expected Hessian matrix (the matrix of second order derivatives and cross derivatives of the log-likelihood). In the present context:

$$\text{var}(\tilde{\lambda}) = [-E(H)]^{-1} = \frac{\lambda^2}{\sum E(y_i)} = \frac{\lambda^2}{n\lambda} = \frac{\lambda}{n} \tag{13.6}$$

Using the invariance property of MLEs, an MLE for the variance of $\tilde{\lambda}$ is found by substituting $\tilde{\lambda}$ into 13.6:

$$\widetilde{\text{var}(\tilde{\lambda})} = \frac{\tilde{\lambda}}{n}$$

Example 13.1 — Flying-bomb hits on London during WWII The following data is on number of bomb hits in south London during WWII (Feller, 1957). The city was divided into 576 areas, and the number of areas hit exactly y times was counted. What does the assumption of independence of the data imply here?

Hits	0	1	2	3	4	5+
Observed	229	211	93	35	7	1
Expected	228	211	98	30	7	1

What is $\tilde{\lambda}$? What is $\widetilde{\text{var}(\tilde{\lambda})}$? How are the "Expected" values in the table calculated? How would you test the hypothesis that the expected number of bomb hits for an area is less than 1?

13.3.4 Specification testing for the Poisson distribution

Originally, we had to make the assumption that the data were Poisson distributed. The distributional assumption is an important first step, and one that can be tested. Similar to R^2 in the linear regression model, we can examine how well the model estimated by MLE "fits" the data. If the fit is good, the distributional assumption is generally considered to be good.

Goodness-of-fit tests for the Poisson distribution can be achieved by comparing the observed and expected counts. For example, consider the following χ^2 statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i are the observed frequencies and E_i are the expected frequencies (obtained by plugging the MLE into the probability function). If the Poisson model is correctly specified, then the expected value of the above statistic is 0. If the above chi-square statistic becomes too large, we may reject the null that Poisson is the correct distribution, however, rejection does not indicate the appropriate distribution, only that the Poisson model is misspecified (and we lose of some or all of MLs asymptotic properties).

The main limitation of the Poisson distribution in applications is it's property of equidispersion. Most count data are overdispersed, i.e. the variance exceeds the mean. Hence, there are several tests based on this restriction. In many cases, there are other candidate distributions that the data may follow (e.g. negative binomial or zero-inflated Poisson), that nest the Poisson distribution. Wald, likelihood ratio, and score testing procedures may be used.

13.4 The Poisson regression model

Once we include explanatory variables, or "regressors" into the model, it becomes a "regression model". How do we accomplish this? A common modelling strategy is to "link" the mean of the distribution

to regressors. We simply write an equation where the parameters of the distribution are determined by explanatory variables. Usually, one of the parameters of the distribution is the *mean*, and this is where we form the link. This allows, for example, different people with different characteristics to have different means. For Poisson, the link is usually:

$$E\left[y_i \mid X_i\right] = \lambda_i = \exp\left(X_i'\boldsymbol{\beta}\right) \tag{13.7}$$

That is, the mean of $\boldsymbol{y}$ is conditional on X and can vary by individual or observation. The specific form of the link function is somewhat arbitrary, but ensures that $\lambda_i > 0$. For example, consider the number of doctor visits. An individual's doctor visits may depend on age, underlying health conditions, genetics, and insurance status. The economist may be interested in moral hazard or adverse selection.

By substituting 13.7 into 13.1, multiplying across all observations (by independence of the data), and taking logs, we have the following joint log-likelihood function:

$$l\left(\boldsymbol{\beta} \mid y_i, X_i\right) = \sum_{i=1}^{n} y_i X_i'\boldsymbol{\beta} - \exp X_i'\boldsymbol{\beta} - \log y_i! \tag{13.8}$$

The derivative of 13.8 with respect to the vector $\boldsymbol{\beta}$ is:

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{n} \left(y_i - \exp X_i'\beta\right) X_i \tag{13.9}$$

Setting 13.9 equal to zero does not admit a closed form solution for $\boldsymbol{\beta}$! Hence, numerical methods, such as Newton-Raphson, must be used to obtaining the ML estimate. Note that asymptotic standard errors for the $\boldsymbol{\beta}$ can again be estimated by inverting the expected Hessian matrix.

13.4.1 Interpreting the β

Due to the exponent in the link function, the β do not have as simple of an interpretation as they do in LS (this is one of the reason people hesitate to depart from LS). For example, a one unit change in the jth regressor leads to a *proportionate* change in $\mathbb{E}[y_i \mid X_i]$. Note that while standard errors for the β can be estimated by inverting the Hessian, estimating standard errors of the semi-elasticities would require the *delta method*.

13.5 Application: badhealth

Load some data:

```
install.packages("COUNT")
library(COUNT)
data(badhealth)
```

The variables in the data set are `numvisits` - the number of visits to the doctor, `badh` - a dummy equal to 1 if the individual self-reports that they are in "bad health", and `age` - the age of the individual. Plot the data:

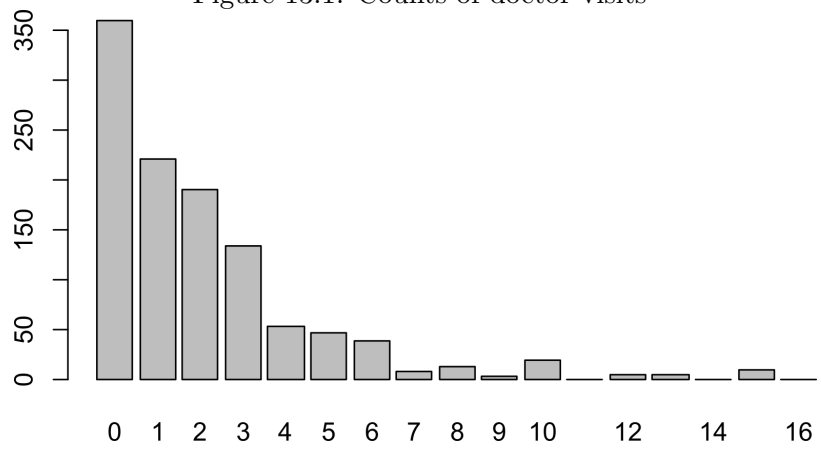
```
barplot(table(badhealth$numvisit[badhealth$numvisit < 17]))
```

The goal is to figure out how `badh` effects the number of doctor visits. Try LS:

```
summary(lm(numvisit ~ badh + age, data=badhealth))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.436780   0.345927   4.153 3.52e-05 ***
badh        4.081386   0.327616  12.458  < 2e-16 ***
```

Figure 13.1: Counts of doctor visits




```
age          0.013720   0.009055   1.515     0.13
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.227 on 1124 degrees of freedom
Multiple R-squared:  0.1323,    Adjusted R-squared:  0.1307
F-statistic: 85.68 on 2 and 1124 DF,  p-value: < 2.2e-16
```

The LS model implies that doctor visits increase by 4 when a person is in “badhealth”. But the model that we really want to estimate is:

$$\mathbb{E}[y_i] = \exp\left(\beta_0 + \beta_1 badh + \beta_2 age\right)$$

where y_i is recognized to be a count variable and follows the Poisson distribution (or some other counting distribution). To estimate this model in R:

```
pois.mod <- glm(numvisit ~ badh + age, family=poisson, data=badhealth)
summary(pois.mod)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.447022   0.071428   6.258 3.89e-10 ***
badh        1.108331   0.046169  24.006  < 2e-16 ***
age         0.005822   0.001822   3.195   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 4020.3  on 1126  degrees of freedom
Residual deviance: 3465.3  on 1124  degrees of freedom
AIC: 5638.6
```

To interpret the estimated coefficients, we can take exponents. For example:

```
exp(0.447022 + 1.108331*1 + 0.005822*30) / exp(0.447022 + 1.108331*0 + 0.005822*30)
```

```
[1] 3.029298
```

tells us that people with “badhealth” are 3.03 *times* more likely to visit the doctor. (Note that the above exponent is the same as `exp(1.108331)`).

13.6 MLE with a Normal distribution

Consider a random sample of n observations on a variable x, where $x_i \sim N(\mu, \sigma^2)$ for all i. This implies that the density function for a single x_i value is the familiar "bell-shaped" curve, the formula for which is:

$$f\left(x_i \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}\left(x_i - \mu\right)^2\right\}$$

Now consider the joint density function of the n sample values, given the parameters. Using the independence of these values:

$$\begin{aligned} f\left(x_1, x_2, \ldots, x_n \mid \mu, \sigma^2\right) &= f\left(x_1 \mid \mu, \sigma^2\right) \times f\left(x_2 \mid \mu, \sigma^2\right) \times \ldots \times f\left(x_n \mid \mu, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}\sum_i \left(x_i - \mu\right)^2\right). \end{aligned}$$

Here, the joint density function is viewed as a function of the x data, given the values of the two parameters, μ and σ^2. Without changing the form of this expression, we could also view it from a different perspective: as a function of the parameters, given the values of the data. When we view it in this alternative way we give the function a different name, even though the algebraic expression is the same – we call it the "Likelihood Function". That is:

$$L\left(\mu, \sigma^2 \mid x_1, x_2, \ldots x_n\right) = f\left(x_1, x_2, \ldots\ldots, x_n \mid \mu, \sigma^2\right)$$

One general estimation principle that we could adopt is to try and choose the values of the parameters that are "most likely" to have generated the observed sample of data. Note that all of the (data-related) information that we have about the parameters is summarized in the joint density function for the sample observations (i.e., the likelihood function). In other words, we could choose a formula for the estimator which maximizes the value of the likelihood function (or the value of the logarithm of the likelihood function, as this will yield the same result, because the logarithmic transformation is strictly monotonic increasing). The estimator that we obtain by following this logic is called the Maximum Likelihood Estimator (MLE), and is perhaps the most widely used estimation principle in statistics (and econometrics).

To get the MLE's of μ and σ^2 in our example, we will have to take the partial derivatives of the likelihood function, or its logarithm (denoted $\log L(\mu, \sigma^2)$) for convenience, with respect to μ and σ^2, and set these derivatives equal to zero to obtain the first-order conditions. We then solve this pair of simultaneous equations for μ and σ^2:

$$\log L\left(\mu, \sigma^2\right) = -(n/2)\log\left(2\pi\sigma^2\right) - \left(1/2\sigma^2\right)\sum_{i=1}^{n}\left(x_i - \mu\right)^2$$

so,

$$\frac{\partial \log L}{\partial \mu} = 0 - \frac{1}{2\sigma^2} \times 2(-1)\sum_{i=1}^{n}\left(x_i - \mu\right) \tag{13.10}$$

and

$$\frac{\partial \log L}{\partial \sigma^2} = -\left(n/2\sigma^2\right) + \frac{1}{2\sigma^4}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \tag{13.11}$$

(Note: to get 13.11 we differentiated with respect to σ^2, not with respect to σ.) Setting 13.10 equal to zero:

$$\sum_{i=1}^{n}\left(x_1 - \widetilde{\mu}\right) = 0$$

or

$$n\tilde{\mu} = \sum_{i=1}^{n} x_i$$

we get:

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

Similarly, setting 13.11 equal to zero:

$$\sum_{i=1}^{n} (x_i - \widetilde{\mu})^2 / \widetilde{\sigma}^2 = n$$

$$\tilde{\sigma}^2 = (1/n) \sum_{i=1}^{n} (x_i - \widetilde{\mu})^2$$

In this case, the MLE for the population mean is just the sample mean, and the MLE for the population variance is biased (it has n in the denominator instead of $n - 1$).

13.7 Exercises

1. Explain what a limited dependent variable is, and list some examples.
2. What is the likelihood function?
3. Define the maximum likelihood estimator.
4. What are the properties of the maximum likelihood estimator?
5. What is the crucial assumption backing the "good" properties of the maximum likelihood estimator?
6. Why do we typically work with the *log*-likelihood, instead of just the likelihood?
7. In the context of the Poisson distribution, explain how to create a regression model.
8. The geometric distribution may be used to describe the probability of a number of failures occurring before the first success. The probability function for a random variable, y_i, which follows a geometric distribution, is given by:

   $$\Pr(Y_i = y_i) = (1 - p)^{y_i} p$$

   where p is the probability of success. The mean of the distribution (the expected value of y_i) is given by:

   $$E(y_i) = \frac{1 - p}{p}$$

   (a) Specify the joint likelihood function.
   Answer.
   The likelihood function is:

   $$L(p \mid y_i, \ldots, y_n) = \prod_{i=1}^{n} (1 - p)^{y_i} p = (1 - p)^{y_i} p^n$$

   (b) Take the log of the likelihood function.
   Answer.

   $$\log L = \sum y_i \log(1 - p) + n \log p$$

(c) Find the MLE for p.
Answer.

$$\frac{\partial \log L}{\partial p} = -\frac{\sum y_i}{1-p} + \frac{n}{p} = 0$$
$$\tilde{p} = \frac{1}{1+\bar{y}}$$

(d) Find the Hessian, and make sure it is n.s.d.
Answer.
In this case, there is only one parameter, so the Hessian is scalar:

$$\frac{\partial^2 \log L}{\partial p^2} = \frac{-\sum y_i}{(1-p)^2} - \frac{n}{p^2}$$

Since the Hessian is ≤ 0 for all $p \geq 0$, the MLE $\tilde{p}$ solves for the global max.

9. The exponential distribution describes the time between events in a Poisson process. The probability function for the exponential distribution is:

$$p\left(y_i \mid \lambda\right) = \lambda e^{-y_i \lambda}$$

Find the maximum likelihood estimator for λ.
Answer.
The joint likelihood function is:

$$L\left(\lambda \mid y_1, \ldots, y_n\right) = \prod_{i=1}^{n} p(y_i \mid \lambda) = \lambda^n e^{-\lambda \sum y_i}$$

and the log-likelihood is:

$$\log L = n \log \lambda - \lambda \Sigma y_i$$

Taking the first derivative of the log-likelihood and setting it equal to zero for the FOC, and solving, gives:

$$\frac{\partial \log L}{\partial \lambda} = \frac{n}{\lambda} - \sum y_i = 0$$
$$\tilde{\lambda} = n / \sum y_i = 1/\bar{y}$$

Taking the second derivative:

$$\frac{\partial^2 \log L}{\partial \lambda^2} = \frac{-n}{\lambda^2}$$

We see that it is negative for $\lambda > 0$, so the MLE finds the global maximum.

10. Refer to section 13.6. Consider our usual linear regression model:

$$y = X\boldsymbol{\beta} + \epsilon \quad ; \quad \epsilon \sim \mathrm{N}\left(0, \sigma^2 \mathrm{I}\right)$$

(a) Write down the likelihood function, $L\left(\boldsymbol{\beta}, \sigma^2 \mid y_1, y_2, \ldots, y_n\right)$. Note that if the ϵ_i are uncorrelated and normally distributed, then they are independent.

Answer.

$$\begin{aligned}
L &= p\left(y_1, \ldots, y_n \mid \beta, \sigma\right) \\
&= \prod_{i=1}^{n} p\left(y_i \mid \beta, \sigma\right); \text{ given independence} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2\sigma^2}\left(y_i - x_i'\beta\right)^2}\right] \\
&= \frac{1}{\sigma^n(\sqrt{2\pi})^n} \cdot e^{-\frac{1}{2\sigma^\sigma}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2} \\
&= \sigma^{-n}(2\pi)^{-\frac{n}{2}} e^{\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[y_i - x_i'\beta\right]^2\right)}
\end{aligned}$$

(b) Derive the maximum likelihood estimator for $\boldsymbol{\beta}$ in this model. Is this estimator biased or unbiased?

Answer.

$$\begin{aligned}
\log L &= -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - x_i'\beta\right)^2 \\
&= -n\log\sigma - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log L}{\partial \beta} &= -\frac{1}{2\sigma^2} \cdot \frac{\partial}{\partial \beta}\left[y'y + \beta'X'X\beta - 2y'x\beta\right] \\
&= -\frac{1}{2\sigma^2}\left(2X'x\beta - 2X'y\right) = 0
\end{aligned}$$

Hence,

$$\left(X'X\beta - x'y\right) = 0 \quad \Rightarrow \quad \hat{\beta} = \left(X'X\right)^{-1}X'y$$

and this is just the LS estimator, which is unbiased under the usual set of assumptions.

(c) Compare the assumptions that we used in class to derive the LS estimator of $\boldsymbol{\beta}$ with the assumptions used here to obtain the MLE.

Answer.

With LS, the derivation of $\boldsymbol{b}$ requires: (i) a linear model; (ii) that X has full rank (the properties of $\boldsymbol{b}$ then depend on additional assumptions). With the MLE, the derivation of the estimator requires (i) and (ii) above, and an appropriate assumption about the distribution of the errors - here, Normality was used. The properties of the MLE are the same as those of $\boldsymbol{b}$, if the assumptions hold.

(d) Derive the MLE for σ^2 in this model. Is this estimator biased or unbiased?

Answer.

$$\begin{aligned}
\frac{\partial \log L}{\partial \sigma^2} &= \frac{\partial}{\partial \sigma^2}\left[-\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log 2\pi - \frac{1}{2\sigma^2}\left(y - X\beta\right)\left(y - X\beta\right)\right] \\
&= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'\left(y - X\beta\right) = 0
\end{aligned}$$

so

$$\tilde{\sigma}^2 = \frac{1}{n}(y - X\tilde{\beta})'(y - X\tilde{\beta}) = \frac{1}{n}e'e$$

where e is the MLE *and* LS residual vector. Compare this estimator with $s^2 = e'e/(n-k)$, our usual estimator. We know that s^2 is unbiased, so clearly $\tilde{\sigma}^2$ is biased.

(e) One of the two estimators you have obtained is biased. Show that this bias vanishes if the sample size is sufficiently large.
Answer.

$$E\left(\hat{\sigma}^2\right) = \frac{1}{n}E\left(e'e\right) = \frac{1}{n}(n-k)\sigma^2 = (1-\frac{k}{n})\sigma^2 \neq \sigma^2$$

and

$$\text{Bias}\left(\tilde{\sigma}^2\right) = E\left(\tilde{\sigma}^2\right) - \sigma^2 = (1-\frac{k}{n})\sigma^2 - \sigma^2 = (-\frac{k}{n})\sigma^2$$

So, the bias is negative. For fixed k, if $n \to \infty$, then this bias vanishes.