Introduction to Econometrics

Ryan T. Godwin

Copyright © 2024 by Ryan T. Godwin
Winnipeg, Manitoba, Canada
ISBN 978-1-77284-004-9
This work, as a whole, is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (the "License"). You may not use this file except in compliance with the License. You may obtain a copy of the License at https://creativecommons.org/licenses/by-nc-sa/4.0/.

Contents

Chapter 1

Introduction

1.1 This book

This book serves as the course notes for the introductory econometrics course Econ 3040, at the University of Manitoba. This book covers the standard topics in an introductory econometrics course, and tries to balance theory and application.

1.2 What is Econometrics?

Econometrics is the study of statistical methods applied to economics data. It is a subset of statistics. Similarly, biology has "biometrics", psychology has "psychometrics", etc. Econometrics uses those methods most suited to economics data.

Econometrics can be used to test economics theories. Economics is a social *science*, and economics benefits from the scientific method. Theories are formed and tested using observations from the real world. The testing part mostly relies on econometrics.

Econometrics can be used to estimate *causal effects*, though it should not be used to find them. That is, the theoretical model (e.g. from Micro or Macro) should specify which variable causes which. It is then up to the econometrician to estimate *how much* of an effect one variable has on another. Econometrics may also be used to forecast or predict economic variables, although forecasting is not covered in this course.

Econometrics specializes in dealing with *observational data*. Observational data is in contrast to *experimental data*. In an experiment, there is some element of *control* - a variable can be changed by the researcher, and the effect of the change on another variable can be more easily measured. In observational data the causal variable is changing on its own, and this can be very problematic. Typically there are important *omitted variables* in observational data. An experiment provides a better way to estimate a causal effect, since the missing variables are not a problem in a well constructed experiment.

Economic models often suggest that one variable causes another. This often has *policy implications*. The economic models, however, do not provide quantitative magnitudes of the causal effects. For example:

- How would a change in the *price* of alcohol or cigarettes effect the *quantity* consumed?
- If *income* increases, how much of the increase will be *consumed*?
- If an additional fireplace is added to a house, how much will the price of the house increase?
- How does another year of *education* change *earnings*?

How would you use an experiment to determine the above four causal effects? You will likely conclude that using an experiment would be too costly and/or unethical. Hence, we must rely on observational data, and try to sort out the associated problems.

It is important to be aware of the limitations of statistics. It can never be used to determine *causation*. Causation must be theorized. If two variables are correlated, statistics alone cannot tell

which variable causes which, or if there is any causation at all. That is, *correlation does not imply causation*. If, however, we find that two variables are statistically independent from each other, one variable can not cause the other.

Objectives

Some objectives of this text are the following:

- Learn a method for estimating causal effects (LS)
- Understand some theoretical properties of LS
- Learn about hypothesis testing
- Learn to read regression analyses, so as to understand empirical economics papers in other courses
- Practice OLS using data sets

1.3 R Statistical Environment and RStudio

1.3.1 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.

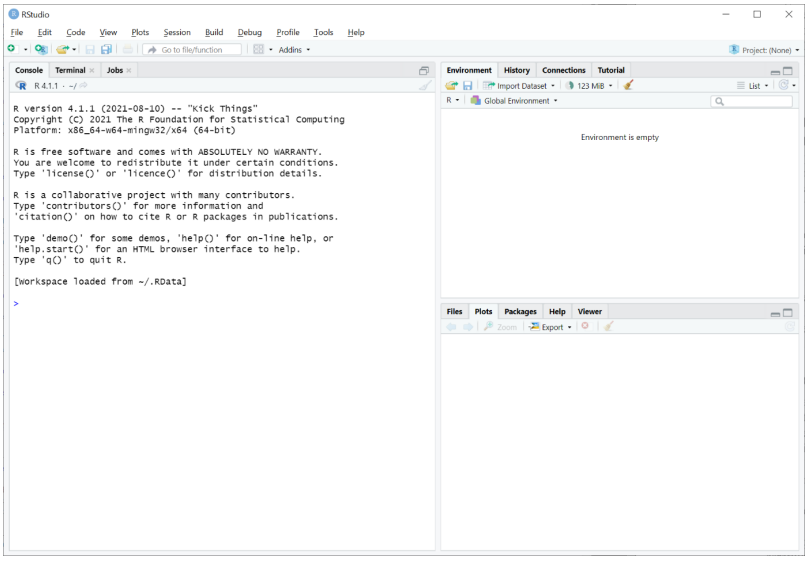
1.3.2 Where to get R

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/.

1.4 Getting started with RStudio

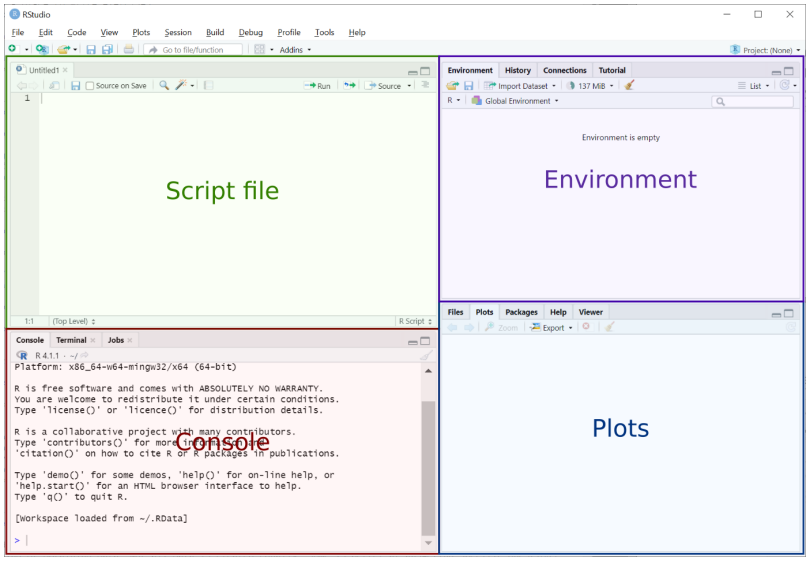
1.4.1 Open RStudio

After you open RStudio it should look something like this:



1.4.2 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".



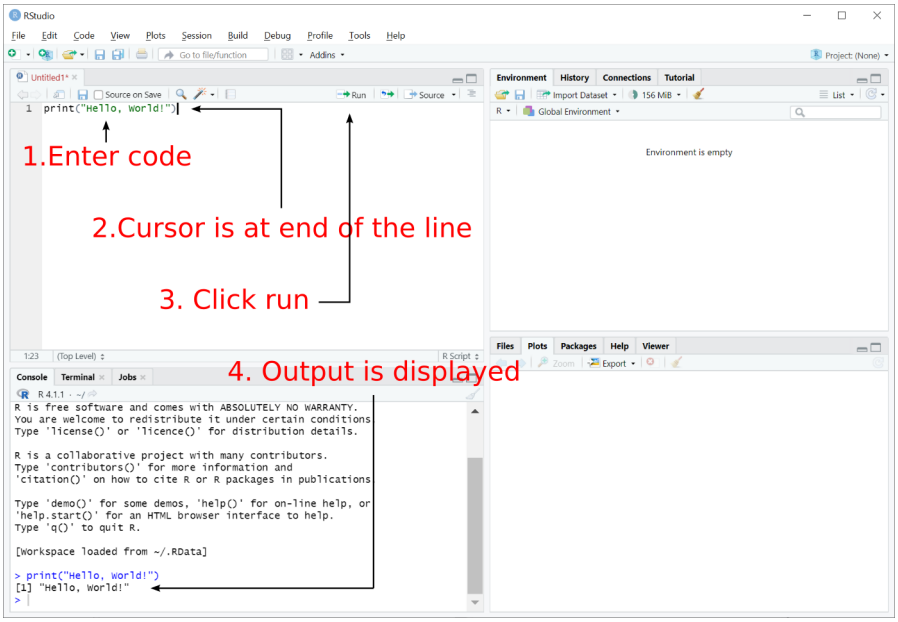
- In the top left is your Script file. R commands can be run from the R Script file, and saved at any time.
- In the bottom left is the Console window. Output is displayed here. R commands can be run from the Console, but not saved.
- In the top right is the Environment. Data and variables will be visible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

1.4.3 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!"
```

1.5 Arithmetic in R

R's arithmetic operators include:

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

Example 1.1 — **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
```

1.6 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:

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

1.7 Simple functions in R

Table 1.1: 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 1.1. 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.

1.8 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
```

1.8.1 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.

1.9 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).

1.9.1 Directly from the internet

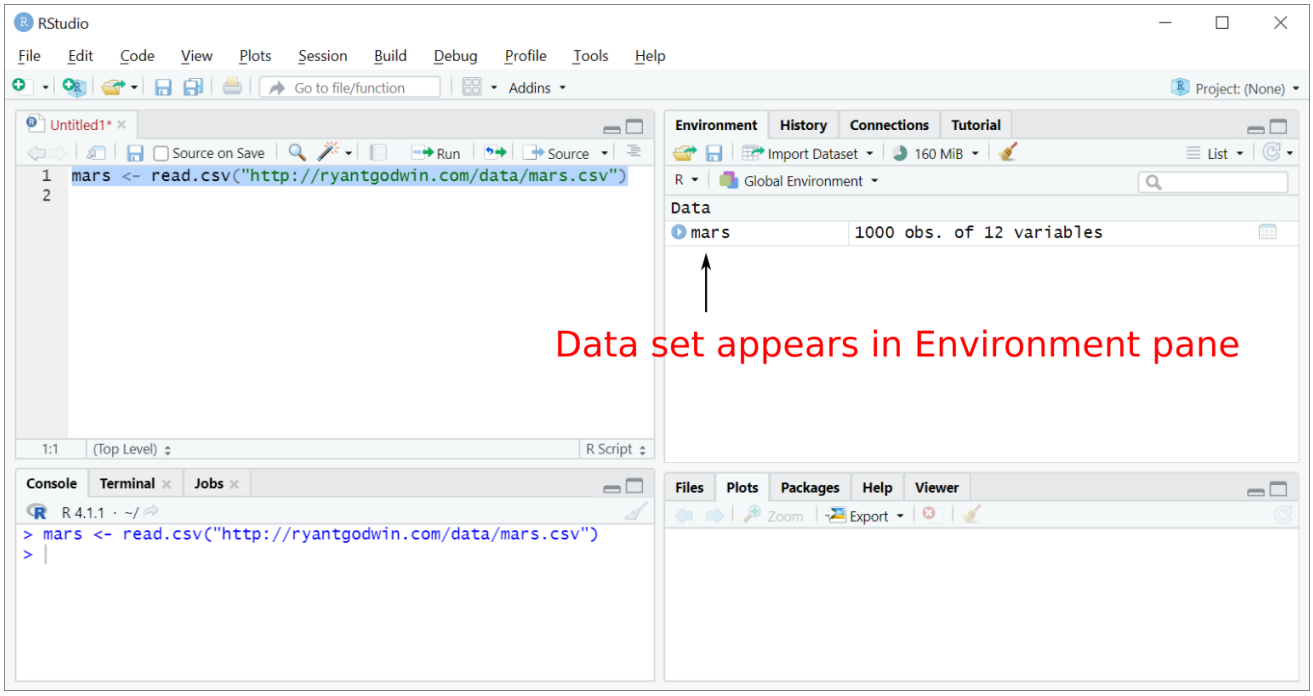
We can use the R code:

```
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).



1.9.2 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/`.

1.9.3 `file.choose()`

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

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

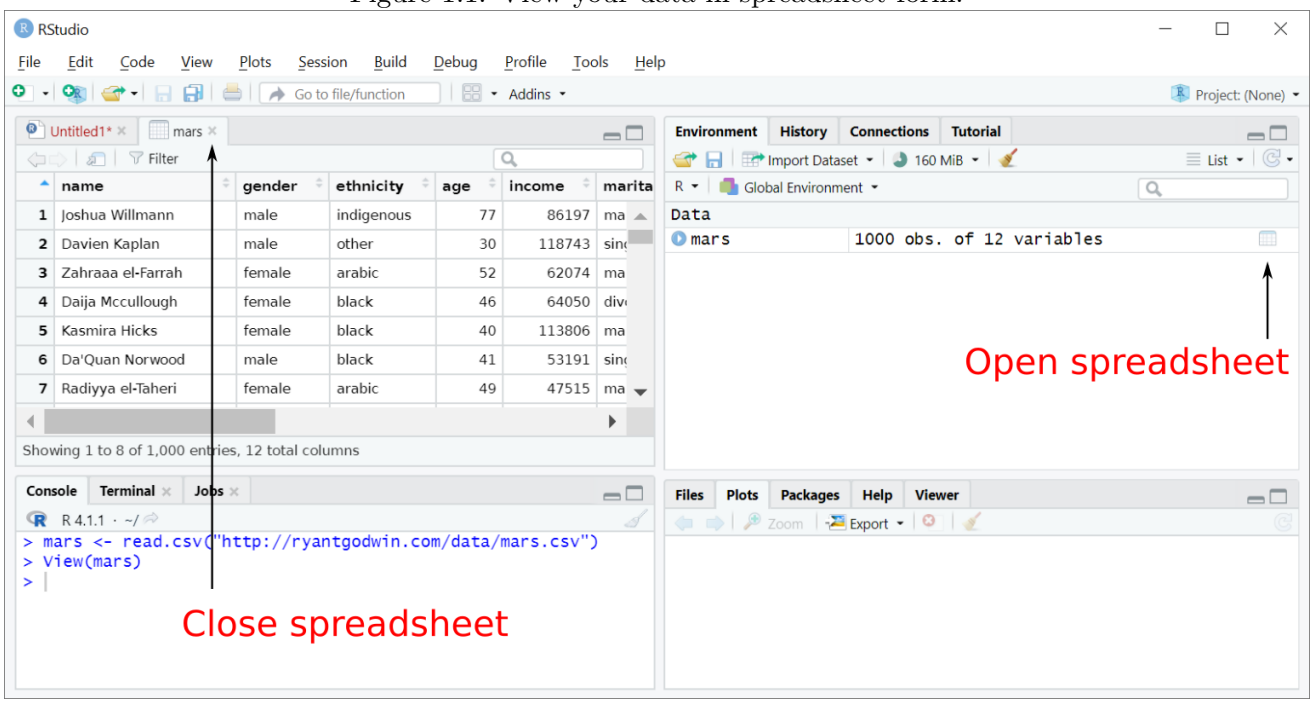
1.10 View your data in spreadsheet form

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 1.1.

Figure 1.1: View your data in spreadsheet form.



1.11 Scientific notation in R output

R's default is to report numbers with many digits in *scientific notation*. For example, the number 1 million (1000000) is written in scientific notation as 1×10^6. We can see this notation in R using:

```
my.number <- 1000000
my.number
```

```
[1] 1e+06
```

The `e` in the output signifies an exponent to base 10. Similarly, the number 0.0000001 would be output as `1e-06` (note the negative sign on the exponent).

The scientific notation can be difficult to read at times, and you can suppress this notation using `options(scipen=999)`. Try this option, and print out `my.number` again:

```
options(scipen=999)
my.number
```

```
[1] 1000000
```

Chapter 2

Probability Review

It is important for us review some basic concepts in probability. Ultimately, we will be calculating "statistics" using datasets: averages, correlations, t-statistics, slope estimators, for example. All of these "statistics" are random variables! Understanding some concepts in probability can help guide us as we choose what to do with a dataset.

These are concepts that you should know from your previous statistics courses, and this chapter is meant as only a review of some important concepts.

2.1 Fundamental Concepts

2.1.1 Randomness

Randomness is unpredictability. Outcomes that we cannot predict are random. Randomness represents our inability as humans to accurately predict things. For example, if I roll two dice, the outcome is random because I am not smart enough or skilled enough to predict what the roll will be. Things that I cannot predict, or are to difficult to predict, are random. We cannot know everything. However, we can attempt to model the randomness mathematically.

Randomness: the inability to predict an outcome.

This definition of randomness does not oppose a deterministic world view (fate). While many things in our lives *appear* to be random, I still think that at some fundamental level the world is deterministic, and that all events are potentially predictable. In the dice example, it is not far-fetched to believe that a computer could analyze my hand movements and perfectly predict the outcome of the roll.

The sample space is the set of all possibilities (all outcomes) that can occur as a result of the random process.

It is sometimes useful to construct a set, or *sample space* of the possible *outcomes* of interest. In the dice example, the sample space is { ⚀⚀, ⚀⚁, ⚀⚂, ..., ⚅⚅ }. An *event* is a subset of the sample space, and consists of one or more of the possible outcomes. For example, rolling higher than ten is an event consisting of three outcomes { ⚄⚅, ⚅⚄, ⚅⚅ }.

An outcome is a single point, or possibility, in the sample space.

An event is a collection of outcomes. An event is a subset of the sample space.

2.1.2 Probability

A probability is a number between 0 and 1 that is assigned to an event (sometimes expressed as a percentage). A standard definition is: the probability of an event is the proportion of times it occurs in

the long run. This is fine for the dice example, and you may be aware that the probability of rolling a seven is 1/6 or of rolling higher than ten is 1/12. This definition works for this example because we can imagine rolling the dice repeatedly under similar conditions and observing that a seven occurs one-sixth of the time.

What about events that occur seldomly or only once? What is the probability that you will obtain an A+ in this course? What is the probability that Donald Trump will be imprisoned? For these events, the former definition of probability is less satisfactory. A more general definition is: probability is a mathematical way of quantifying uncertainty. For the Trump example, the probability of imprisonment is *subjective*. I may think the probability is 0.1, but someone else may assign a probability of 0.9. Which is right? These problems are better suited to a *Bayesian* framework, which is not discussed in this book. The first definition of probability will suffice for the topics covered in this book.

Probability: a number between 0 and 1 representing the portion of times an event will occur, if the event could occur repeatedly.

2.2 Random variables

A *random variable* translates outcomes into numerical values. For example, a die roll only has numerical meaning because someone has etched numbers onto the sides of a cube. A random variable is a human-made construct, and the choice of numerical values can be arbitrary. Different choices can lead to different properties of the random variable. For example, I could measure temperature in Celsius, Fahrenheit, Kelvin or something new (degrees Ryans). The probability that it will be above 20° tomorrow depends critically on how I have constructed the random variable.

2.2.1 Discrete and continuous random variables

Random variables can be separated into two categories, *discrete* and *continuous*. A discrete random variable takes on a countable number of values, e.g. {0, 1, 2, ...}. The result of the dice roll is a discrete random variable. A continuous random variable takes on a continuum of possible values (an infinite number of possibilities).

Even when the random variable has lower and upper bounds, there are still infinite possibilities. The temperature tomorrow is a continuous random variable. It may be bound between -50°C and 50°C, but there are still infinite possibilities. What is the probability that it is 20°C? What about 20.1°C? What about 20.0001°C? We could keep adding 0s after the decimal. In fact, the probability of the temperature taking on any one value approaches 0. Instead, we must talk about the probability of a *range* of numbers. For example, the probability that the temperature is between 19°C and 21°C.

The continuum of possibilities makes it more difficult to discuss continuous random variables than it does discrete random variables. We will use discrete random variables for examples and try to extend the logic to continuous random variables.

2.2.2 Realization of a random variable

Finally, note the difference between a *random variable* and the *realization of a random variable*. Before I roll the die, the outcome is random. After I roll the die and get a ⚃ (for example), the 4 is just a number - a *realization* of a random variable.

2.2.3 Key points

- A random variable can take on different values (or ranges of values), with different probabilities
- There are discrete and continuous random variables
- Continuous random variables can take on an infinite number of possible values, so we can only assign probabilities to *ranges* of values

- We can assign probabilities to all possible values for a discrete random variable
- The *realization* of a random variable is just a number, it used to be random, but now we've seen the outcome

2.3 Probability function

A *probability function* is also called a *probability distribution*, or a *probability distribution function* (PDF). Sometimes a distinction is made: *probability mass function* (PMF) for discrete variables instead of PDF for continuous variables. I will use *probability function* for both.

A probability function is an equation (it can also be a graph or table), which contains information about a random variable. The nature and properties of the randomness determines what type of equation is appropriate. A different equation would be used for a dice roll than would be used for the wage of a worker. The probability function is very important. The probability function accomplishes two things: (i) it lists all possible numerical values that the random variable can take, and (ii) assigns probability to values. Note that the probabilities of all outcomes must sum to 1 (something must happen). The probability function contains all possible knowledge that we can have about the random variable (before we observe its realization).

Probability function. The probability function accomplishes two things: (i) it lists all possible numerical values that the random variable can take, and (ii) assigns probabilities to ranges of values.

Example 2.1 — Probability function for a die roll. Let $Y =$ the result of a die roll. The probability function for Y is:

$$Pr(Y=1)=\frac{1}{6}, Pr(Y=2)=\frac{1}{6}, \ldots, Pr(Y=6)=\frac{1}{6} \tag{2.1}$$

Note how the function lists all possible numerical outcomes and assigns a probability to each. A more compact way of expressing (2.1) is:

$$Pr(Y=y)=\frac{1}{6}\ ;\ y=1,\ldots,6 \tag{2.2}$$

The probability function in (2.2) may also be expressed in a graph (see Figure 2.1).

Example 2.2 — Probability function for a normally distributed random variable. The normal distribution is an important probability distribution. Later, we will discuss why it is so important and prevalent. For now, I will present the probability function for a random variable (you do not need to memorize this).

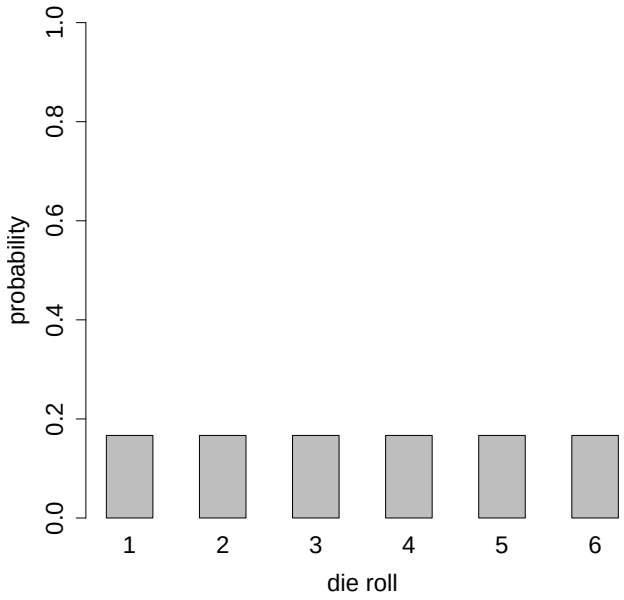
$$f(y|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp-\frac{(y-\mu)^2}{2\sigma^2}\ ;\ -\infty<y<\infty \tag{2.3}$$

y is the random variable, μ and σ^2 are the *parameters* that govern the probability of y. μ turns out to be the *mean* or *expected value* of y, and σ^2 turns out to be the *variance* of y. If μ and σ^2 are known (usually they aren't), then you can determine the probability that y takes on any range of values. However, this requires integration (you won't have to integrate in this course).

2.3.1 Probabilities of events

Recall that the probability function contains all possible information about the random variable (all the outcomes, and a probability assigned to each outcome), and that an event is a collection of outcomes. The probability function can be used to calculate the probability of different events occurring.

Figure 2.1: Probability function for the result of a die roll.



Example 2.3 — **Probability of an event** Let Y be the result of a die roll. What is the probability of rolling higher than 3?

$$Pr(Y > 3) = Pr(Y = 4) + Pr(Y = 5) + Pr(Y = 6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

2.3.2 Cumulative distribution function

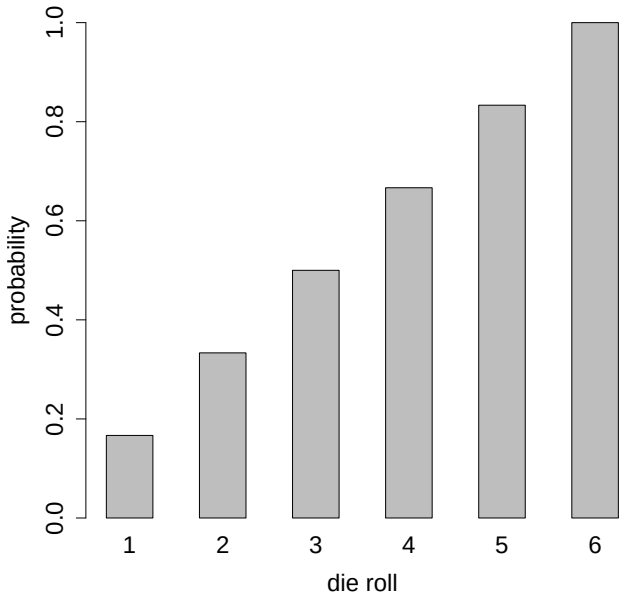
The *cumulative distribution function* (CDF) is related to the probability function. It is the probability that the random variable is *less than or equal to* a particular value. While every random variable has a probability function, it does not always have a CDF (but usually does). Again, let Y be the result of a die roll, then the CDF for Y is expressed as equation 2.4 or as figure 2.2.

$$\begin{aligned} Pr(Y \leq 1) &= 1/6 \\ Pr(Y \leq 2) &= 2/6 \\ Pr(Y \leq 3) &= 3/6 \\ Pr(Y \leq 4) &= 4/6 \\ Pr(Y \leq 5) &= 5/6 \\ Pr(Y \leq 6) &= 1 \end{aligned} \tag{2.4}$$

2.4 Moments of a random variable

The term "moment" is related to a concept in physics. The first moment of a random variable is the mean, the second (central) moment is the variance, the third the skewness, and the fourth the kurtosis. In this book, we will make extensive use of mean and variance, as well as the "mixed" moment: covariance (and it's close friend, correlation).

Figure 2.2: Cumulative density function for the result of a die roll



2.4.1 Mean / expected value

The *mean*, also called the *expected value*, of a random variable is the value that is expected, or the value that occurs on average through many realizations of the random variable. An equation for the mean (of a discrete random variable Y) is:

$$\mathrm{E}[Y] = \sum_{i=1}^{K} p_i Y_i \tag{2.5}$$

There are K possible events that can occur. Each event is labelled by i, and i goes from 1 to K. The probability of each event is p_i. Y_i is the numerical value of event i. For such discrete random variables as Y, the mean is determined by taking a weighted average of all possible outcomes, where the weights are the probabilities.

The mean of a random variable can be determined from its probability function. Recall that the probability function contains all possible information we could hope to have about the random variable. So, it should be no surprise that if we want to determine the mean we have to do some math using the probability function. The mean (and variance, etc.) is just information contained in the probability function.

Notation for the mean of Y or expected value of Y is μ_Y and $\mathrm{E}[Y]$.

Equation 2.5 is valid for any discrete random variable Y. For the die roll example, using the probability function, we have that $K = 6$ and each $p_i = 1/6$, so the mean of Y is:

$$\mathrm{E}(Y) = \frac{1}{6} \times (1) + \frac{1}{6} \times (2) + ... + \frac{1}{6} \times (6) = 3.5$$

The mean / expected value:

- Might not even be a possible value for Y_i, as in the die roll example above. (3.5 can't be rolled).

- The mean is not necessarily the value that "usually" occurs (that's called the *mode*).
- The mean is not necessarily the value that is "most likely" to occur (that's called the *median*).
- The mean / expected value is not the same thing as the sample average!

Calculating the mean of a continuous random variable is analogous, but more difficult. Again, the mean is determined from the probability function, but instead of *summing* across all possible outcomes we have to *integrate* (since the random variable can take on a continuum of possibilities). Let y be a continuous random variable. The mean of y is

$$\mathrm{E}[y] = \int y f(y)\,\mathrm{d}y$$

If y is normally distributed, then $f(y)$ is equation (2.3), and the mean of y turns out to by μ. You do not need to integrate for this course, but you should have some idea about how the mean of a continuous random variable is determined from its probability function.

Some properties of the mean are:

- $\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]$
- $\mathrm{E}[cY] = c\mathrm{E}[Y]$, where c is a constant
- $\mathrm{E}[c + Y] = c + \mathrm{E}[Y]$
- $\mathrm{E}[c] = c$

2.4.2 Median and Mode

The *mean* of a random variable is not to be confused with the *median* or *mode* of a random variable, although all three are measures of "central tendency". The *median* is the "middle" value, where 50% of values will be above and below. The *mode* is the value which occurs the most.

For variables that are normally distributed, the *mean*, *median* and *mode* are all the same, but this is not always true. For a die roll, the mean and median are 3.5, but there either is no mode or all of the values are the mode (depending on the particular definition of mode).

2.4.3 Variance

The variance of a random variable is a measure of its *spread* or *dispersion*. Variance is often denoted by σ^2. In words, variance is the expected squared difference of the random variable from its mean.

An equation for the variance of a random variable Y is:

$$\mathrm{Var}(Y) = \mathrm{E}[(Y - \mathrm{E}[Y])^2] \tag{2.6}$$

When Y is a discrete random variable, then equation (2.6) becomes:

$$\mathrm{Var}(Y) = \sum_{i=1}^{K} p_i \times (Y_i - \mathrm{E}[Y_i])^2 \tag{2.7}$$

where p_i, Y_i, and K are defined as before. Note that equation 2.7 is a weighted averaged of squared distances. The variance is measuring how far, on average, the variable is from its mean. The higher the variance, the higher the probability that the random variable will be far away from its expected value.

When the random variable is continuous, equation (2.6) becomes:

$$\mathrm{Var}(y) = \int (y - \mathrm{E}[y])^2 f(y)\,\mathrm{d}y$$

but you don't need to know this for the course.

Some properties of the variance are:

- $\text{Var}[X+Y] = \text{Var}[X] + \text{Var}[Y] + 2 \times \text{Cov}[X,Y]$
- $\text{Var}[cY] = c^2\text{Var}[Y]$, where c is a constant
- $\text{Var}[c+Y] = \text{Var}[Y]$
- $\text{Var}[c] = 0$

2.4.4 Skewness and Kutosis

Notice in the variance formula (2.6), that there is an expectation of a squared term ($\text{E}[]^2$). This partly explains why the variance is called the *second* (central) moment. Similarly, we could take the expectation of the Y to the third power, or fourth power, etc. Doing so would (almost) give us the third and fourth moments.

The third (central) moment is called *skewness* and the fourth is called *kurtosis*. Much less attention is paid to these moments than is to the mean and the variance. However, it is worth noting that if a random variable is normally distributed, it has a skewness of 0 and a kurtosis of 3.

2.4.5 Covariance

Covariance is a measure of the relationship between two random variables. Random variables Y and X are said to have a *joint* probability distribution. The joint probability distribution is like the probability functions we have seen before (equations 2.1 and 2.3), except that it involves two random variables. The joint probability function for Y and X would (i) list all possible combinations that Y and X could take, and (ii) assign a probability to each combination. A useful summary of the information contained in the joint probability function, is the *covariance*.

The covariance between Y and X is the expected difference of Y from its mean, multiplied by the expected value of X from its mean. Covariance tells us something about how two variables *move* together. That is, if the covariance is positive, then when one variable is larger (or smaller) than its mean, the other variable tends to be larger (or smaller) as well. The larger the magnitude of covariance, the more often this statement tends to be true. Covariance tells us about the direction and strength of the relationship between two variables.

The formula for the covariance between Y and X is

$$\text{Cov}(Y,X) = \text{E}[(Y-\mu_Y)(X-\mu_X)] \tag{2.8}$$

The covariance between Y and X is often denoted as σ_{YX}. Note the following properties of σ_{YX}:

- σ_{YX} is a measure of the *linear* relationship between Y and X. Non-linear relationships will be discussed later.
- $\sigma_{YX} = 0$ means that Y and X are linearly independent.
- If Y and X are independent (neither variable causes the other), then $\sigma_{YX} = 0$. The converse is not necessarily true (because of non-linear relationships).
- The $\text{Cov}(Y,Y)$ is the $\text{Var}(Y)$.
- A positive covariance means that the two variables tend to differ from their mean in the *same* direction.
- A negative covariance means that the two variables tend to differ from their mean in the *opposite* direction.

2.4.6 Correlation

Correlation is similar to covariance. It is usually denoted with the Greek letter ρ. Correlation conveys all the same information that covariance does, but is easier to interpret, and is frequently used instead of covariance when summarizing the linear relationship between two random variables. The formula for

correlation is

$$\rho_{YX} = \frac{\text{Cov}(Y,X)}{\sqrt{\text{Var}(Y)\text{Var}(X)}} = \frac{\sigma_{YX}}{\sigma_Y \sigma_X} \tag{2.9}$$

The difficulty in interpreting the value of covariance is because $-\infty < \sigma_{YX} < \infty$. Correlation transforms covariance so that it is bound between -1 and 1. That is, $-1 \leq \rho_{YX} \leq 1$.

- $\rho_{YX} = 1$ means perfect positive linear association between Y and X.
- $\rho_{YX} = -1$ means perfect negative linear association between Y and X.
- $\rho_{YX} = 0$ means no linear association between Y and X (linear independence).

2.4.7 Conditional distribution and conditional moments

When we introduced covariance, and began to talk about the relationship between two random variable, we introduced the concept of the joint probability distribution function. Recall that the joint probability function lists all combinations of the random variables, assigning a probability to each combination.

Sometimes, however, it is useful to obtain a *conditional* distribution from the joint distribution. The conditional distribution just fixes the value of one of the variables, while providing a probability function for the other. This probability function may change depending on the fixed value.

We need this concept for the *conditional expectation*, which will be important later when we discuss dummy variables. The *conditional expectation* is just the expected or mean value of one variable, conditional on some value for the other variable.

Let Y be a discrete random variable. Then, the conditional mean of Y given some value for X is

$$\text{E}(Y|X=x) = \sum_{i=1}^{K} (p_i|X=x) Y_i \tag{2.10}$$

Example 2.4 — Joint distribution Suppose that you have a midterm tomorrow, but that there is a possibility of a blizzard. You are wondering if the midterm might be cancelled. If there is a blizzard, there is a strong chance of cancellation. If there is no blizzard, then you can only hope that the professor gets severely ill, but that still only gives a small chance of cancellation. The joint probability distribution for the two random events (occurrence of the blizzard, and occurrence of the midterm) is given in table (2.1). Note how all combinations of events have been described, and a probability assigned to each combination, and that all probabilities in the table sum to 1.

What is $\text{E}[Y]$? It is 0.77. This means there is a 77% chance you will have a midterm. $\text{E}[Y]$ is an *unconditional* expectation; it is the mean of Y before you look out the window in the morning and see if there is a blizzard. The conditional expectations, however, are $\text{E}[Y|X=1] = 0.20$ and $\text{E}[Y|X=0] = 0.96$. This means there is only a 20% chance of a midterm if you see a blizzard in the morning, but a 96% chance with no blizzard. Some other review questions using table (2.1) are at the end of this chapter.

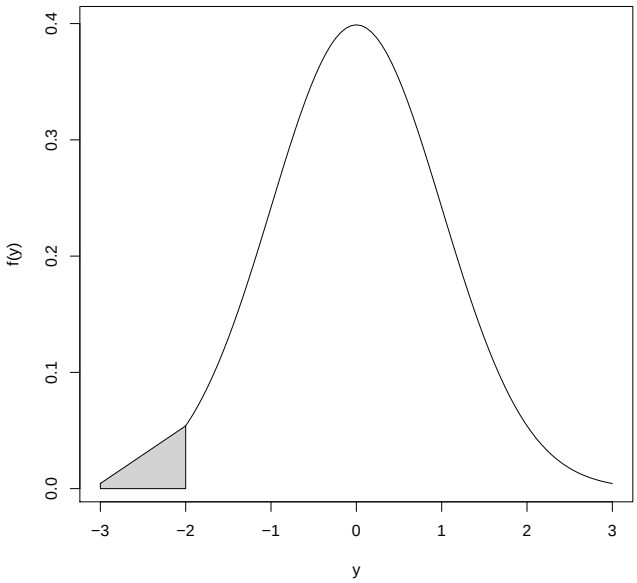
Table 2.1: Joint distribution for snow and a canceled midterm

	Midterm ($Y=1$)	No Midterm ($Y=0$)
Blizzard ($X=1$)	0.05	0.20
No Blizzard ($X=0$)	0.72	0.03

2.5 Some Special Probability Functions

In this section, we present some common probability functions that we will reference in this course. We start with the normal distribution, and a discussion of the *central limit theorem*.

Figure 2.3: Probability function for a standard normal variable, $p_{y<-2}$ in gray



2.5.1 The Normal distribution

The probability function for a normally distributed random variable, y, has already been given in equation (2.3). What is the use of knowing this? If we know that y is normal, and if we knew the parameters μ and σ^2 (we will likely have to estimate them) then we know all we can possibly hope to about y. That is, we can use equation (2.3) to determine the mean and variance of y. We can draw out equation (2.3), and calculate areas under the curve. These areas would tell us about the probability of events occurring.

Suppose that we knew y had mean 0 and variance 1. What is the probability that $y < -2$? Using equation (2.3), we could draw out the probability function, and calculate the area under the curve, to the left of -2. See figure (2.3). This area, and probability, is 0.023.

2.5.2 The standard Normal distribution

The probability function drawn out in figure (2.3) is actually the probability function for a standard normal variable. A variable is standard normal when its mean is 0 and variance is 1. When $\mu = 0$ and $\sigma^2 = 1$, the probability function for a normal variable (equation 2.3) becomes:

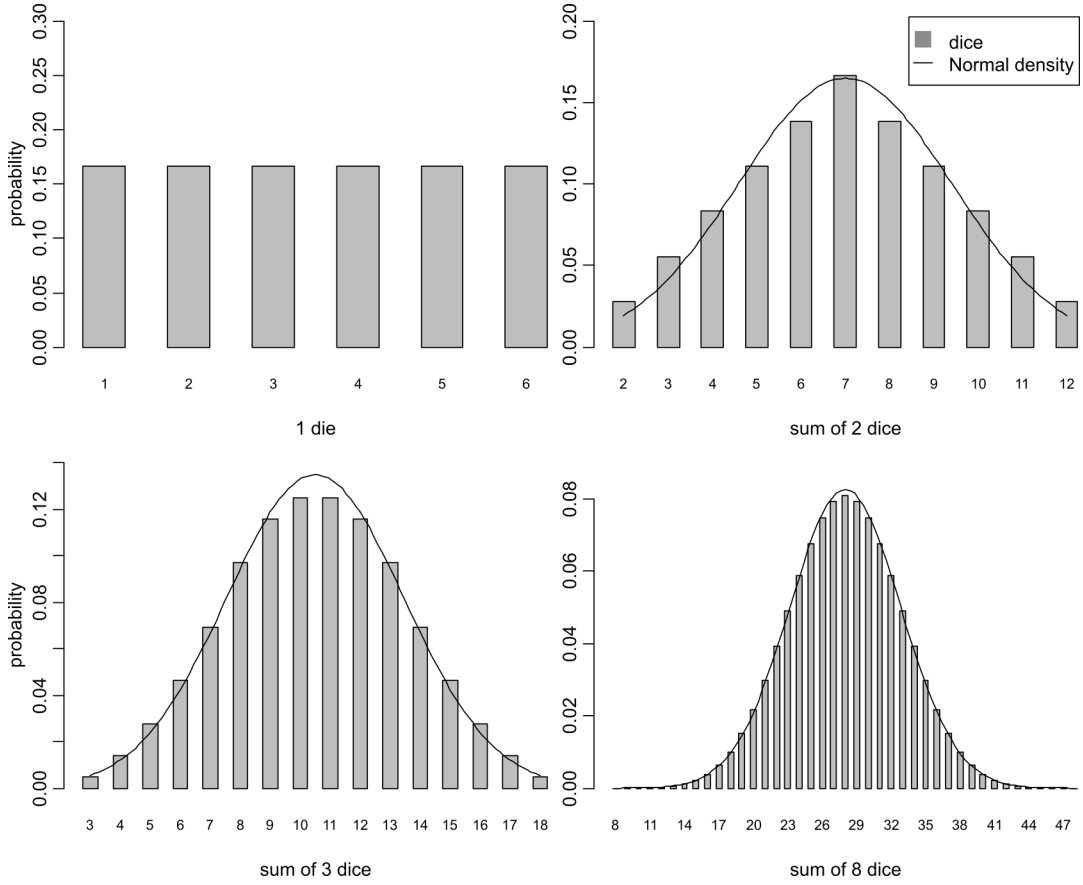
$$f(y) = \frac{1}{\sqrt{2\pi}} \exp \frac{-y^2}{2} \tag{2.11}$$

Note that any random normal variable can be "standardized". That is, if we subtract the variable's mean, and divide by it's standard deviation, then we change the mean to 0, and variance to 1. It becomes "standard normal". This practice is useful in hypothesis testing, as we shall see.

2.5.3 The central limit theorem

So why do we care so much about the normal distribution? There are hundreds of probability functions, that are appropriate in various situations. The heights of waves might be described by the Nakagami distribution. The probability of successfully drawing a certain number of red balls out of a hat of red and blue balls is described by the binomial distribution. The number of customers that visit a store in

Figure 2.4: Probability function for the sums of dice, with Normal density functions superimposed. As the number of random variables that we sum increases, the distribution of the sum becomes Normal. This is due to the central limit theorem (CLT).



an hour might be described by the Poisson distribution. The result of a die roll is uniformly distributed. So why should we pay so much attention to the normal distribution?

The answer is the *central limit theorem* (CLT). Loosely speaking, the CLT says that if we add up enough random variables, the resulting sum tends to be normal. It doesn't matter if some are Poisson and some are uniform. It only matters that we add up enough. If the random outcomes that we seek to model using probability theory are the results of many random factors all added together, then the central limit theorem applies. This turns out to be plausible for the types of economic models we are going to consider. This has been a very casual explanation of the CLT; you should be aware that there are several conditions required for it to hold, and several versions.

Example 2.5 — **Sum of two uniforms.** Let Y be the result of summing two die rolls. What is the

probability function for Y?

$$\begin{aligned} Pr(Y=2) &= 1/36 \\ Pr(Y=3) &= 2/36 \\ Pr(Y=4) &= 3/36 \\ Pr(Y=5) &= 4/36 \\ Pr(Y=6) &= 5/36 \\ Pr(Y=7) &= 6/36 \\ Pr(Y=8) &= 5/36 \\ &\vdots \\ Pr(Y=12) &= 1/36 \end{aligned}$$

The above equation lists all the possibilities, and assigns a probability to each. The probability function is also represented in Figure (2.4). Notice that while each individual die has a uniform (flat) distribution, summed together it begins to get a "curve".

Now, let's add a third die, and see if the probability function looks more normal. Let $Y =$ the sum of *three* dice. It turns out the mean of Y is 10.5 and the variance is 8.75. The probability function for Y is shown in figure (2.4). Also in figure (2.4), the probability function for a normal distribution with $\mu = 10.5$ and $\sigma^2 = 8.75$. Notice the similarity between the two probability functions.

The CLT says that if we add up the result of *enough* dice, the resulting probability function should become normal. Finally, we add up *eight* dice, and show the probability function for both the dice and the normal distribution in figure(2.4), where the mean and variance of the normal probability function has been set equal to that of the sum of the dice.

2.5.4 The Chi-square (χ^2) distribution

Suppose that y is normally distributed. If we add or subtract from y we change the mean of y, but it still will follow a normal distribution. If we multiply or divide y by a number, we change its variance, but y will still be normal. In fact, this is how we standardize a normal variable (we subtract its mean, and divide by its standard deviation).

While a linear transformation (addition, multiplication, etc.) of a normal variable leaves the variable normally distributed, normal variables are not invariant to *non-linear* transformations. If we square a standard normal variable (e.g. y^2), it becomes a χ^2 distributed variable. We will use this distribution for the F-test in a later chapter.

2.6 Review Questions

1. Define the following terms:

outcome	event	random variable
discrete variable	continuous variable	parameter
CLT	mean	variance
probability function	covariance	correlation

2. The joint probability function for X and Y is:

	$Y=-1$	$Y=0$	$Y=1$
$X=1$	0.25	0	0.25
$X=-1$	0	0.5	0

   (a) What is the $\text{Cov}(X,Y)$?
   (b) Are X and Y independent?

3. Use table (2.1).
   (a) What are the probability functions for Y and X (independent from each other)?
   (b) What are the mean and variance of X?
   (c) What is $\text{Cov}(X, Y)$?
   (d) What is ρ_{XY}?
4. Can the mean of a random variable be a value that is *not* in the sample space?
5. Let Y be the number of "heads" that occur from flipping two coins.
   (a) What is the probability distribution for Y?
   (b) Derive the mean and variance of Y.
6. The random variable Y has a mean of 2 and a variance of 4. What is the mean and variance of Z, where $Z = \frac{1}{2}(Y - 2)$?
7. Let Y be the sum of 3 dice.
   (a) What is the mean and variance of Y?
   (b) Suppose that on each of the 3 dice, the numbers on each side are *tripled*. What is the mean and variance of the sum of these 3 dice (where all number have been tripled)?
8. Suppose that Z is a standard Normal random variable.
   (a) What is the distribution for Y, where $Z = 2 \times Y + 1$?
   (b) What is the distribution for Z^2?

2.7 Answers

2. a) The formula for the covariance of X and Y is:

$$\text{Cov}(X, Y) = \text{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$$

The mean of X and Y are:

$$\mu_X = 0.5(1) + 0.5(-1) = 0$$

$$\mu_Y = 0.25(-1) + 0.5(0) + 0.25(1) = 0$$

Finally, the covariance is:

$$\text{Cov}(X, Y) = \text{E}[XY] = 0.25(1)(-1) + 0.5(-1)(0) + 0.25(1)(1) = 0$$

b) Even though the covariance is 0, X and Y are not independent! We can see this by looking at the joint probability function. If we observe the value of Y, then we know, with certainty, the value of X. That is, if we observe $Y = -1$ or $Y = 1$, then we know that $X = 1$. If we observe $Y = 0$, then we know that $X = -1$. Y can predict the value of X, so X and Y are not independent. The point is that covariance measures *linear* association between two variables. In this example, the relationship between X and Y is *non-linear*. If we were to graph the relationship between the two variables, we would see a "U" shape.

3. a) To get the *uncoditional* probabilities for Y we can sum the columns, and for the probabilities of X we can sum the rows, of table (2.1). The probability function for Y is:

$$\Pr(Y = 1) = 0.77 \quad ; \quad \Pr(Y = 0) = 0.23$$

and for X is:

$$\Pr(X = 1) = 0.25 \quad ; \quad \Pr(Y = 0) = 0.75$$

b)

$$\text{E}[X] = 0.25(1) + 0.75(0) = 0.25$$

$$\text{Var}\left[X\right] = 0.25(1-0.25)^2 + 0.75(0-0.25)^2 = 0.1875$$

c) To get the covariance, we will need the mean of Y:

$$\text{E}\left[Y\right] = 0.77(1) + 0.23(0) = 0.77$$

Now, the covariance is:

$$\begin{aligned}\text{Cov}\left(X,Y\right) &= 0.05(1-0.25)(1-0.77)\\ &\quad + 0.20(1-0.25)(0-0.77)\\ &\quad + 0.72(0-0.25)(1-0.77)\\ &\quad + 0.03(0-0.25)(0-0.77)\\ &= -0.1425\end{aligned}$$

d) The formula for correlation is given in equation (2.9). We have already calculated $\text{Cov}\left(X,Y\right)$ and $\text{Var}\left[X\right]$, but we need $\text{Var}\left[Y\right]$:

$$\text{Var}\left[Y\right] = 0.77(1-0.77)^2 + 0.23(0-0.77)^2 = 0.1771$$

Now, the correlation is:

$$\rho_{YX} = \frac{\text{Cov}(Y,X)}{\sqrt{\text{Var}(Y)\text{Var}(X)}} = \frac{-0.1425}{\sqrt{0.1875 \times 0.1771}} = -0.7820$$

Chapter 3

Statistics Review

A statistic is any mathematical function using a *sample* of data. It is just an equation applied to the data. When a statistic is used to estimate a *population* parameter, it is called an *estimator*. One of the main goals of this course is to become familiar with a particular estimator - the *ordinary least squares* estimator, but for this chapter we will review some simpler estimators.

We will discuss the population, and why the sample y should be considered random. Then, we will discuss some estimators. A very important point is that, because y is random, functions of y are also random. Since an estimator is just an equation applied to y, the estimator itself is also random. As we know from the previous chapter, random variables have probability functions.

The probability function for an estimator is given a special name - the *sampling distribution*. Obtaining some properties of the estimator from its sampling distribution, such as mean and variance, will tell us whether or not the estimator is "good", and will guide our choice of which estimator to use.

3.1 Random Sampling from the Population

A sample of data is a collection of variables. In econometrics, most of these variables are *realizations* of a random process. The numbers that make up (at least some of) the sample values came from a random process. The sample typically appears to us on our computer screen as a "spreadsheet" where each column is a different variable and each row is a different *sample unit*. The sampling "units" could be people, countries, firms, etc.

There are at least two ways to think about where a random sample, y, comes from. Both ways make use of the idea of a *population*. The population holds all of the information, the truth. If we knew the entire population, our jobs as statisticians or econometricians would be much easier. Instead we will obtain only a piece of the puzzle, a *sample* of data from the population.

The first way to think about the population, is that it is a data generating process (dgp). It is a random process that generates the y variables that we observe. It is as if a die is being rolled, generating the numbers in the sample, but we can't quite see what the die looks like. Alternatively, if y is normally distributed, then values in y are generated from equation (2.3), but where μ and σ^2 are unknown. This might be a difficult way to think about things.

A second, possibly easier way to think about the population, is to imagine it consisting of all of the data possible. When we obtain economic data, we typically do not observe everyone or everything in the *population* of interest. Instead we observe a *sample* of the population. Hopefully, members of the population will be selected randomly into the sample (otherwise we will have problems).

Suppose we want to know the mean height of a male U of M student. We can not afford to measure the height of every student, so we collect a sample, and hope that it represents the population. Suppose we stand in the University Centre for an hour and measure heights of students. The sample that we will collect is random - we don't know what the heights will be yet. On a different day, at a different time, or in a parallel universe, we will randomly select different students, get different heights, and a

different sample.

We will want this sample to be independently and identically distributed (iid). *Indpendent* - none of the random variables in the sample have any connection. Independence would be violated if a basketball team walked through the University Centre and I sampled all of their heights. *Identical* - all of the random variables in the sample come from the same population (or probability function). The *identical* assumption would be violated if I accidentally sampled some Mini U students (grade school students touring campus).

Table 3.1: Entire population of heights (in cm). The true (unobservable) population mean and variance are $\mu_y = 176.8$ and $\sigma_y^2 = 39.7$.

177.3	170.2	187.2	178.3	170.3	179.4	181.2	180.0	**173.9**
178.7	**171.7**	160.5	183.9	175.7	175.9	**182.6**	181.7	180.2
181.5	176.5	**162.1**	180.3	175.6	**174.9**	**165.7**	172.7	178.9
175.3	178.7	175.6	166.4	173.1	173.2	175.6	183.7	181.3
174.2	180.9	179.9	171.2	171.0	178.6	181.4	175.2	**182.2**
171.7	178.4	**168.1**	186.0	**189.9**	173.4	168.7	180.0	175.1
175.7	180.8	176.2	170.8	177.3	**163.4**	**186.3**	177.1	191.2
171.0	180.3	**169.5**	167.2	178.0	172.9	176.0	176.5	**171.9**
175.1	184.2	165.3	180.2	178.3	183.4	**173.9**	178.6	177.9
184.5	184.1	180.9	187.1	179.9	167.1	**172.0**	167.4	**172.7**
171.6	186.6	182.4	185.5	174.8	178.8	192.8	179.3	**172.0**

As an example, let's pretend that the entire population of heights is in table (3.1). This is a simplified example of a population - the table should be much larger - usually we assume the population is near-infinite. Let's collect a random sample from this population, say 20 observations (the bold numbers in the table). Our sample is then denoted $y = \{173.9, 171.7, 182.6, 181.5, 162.1, 174.9, 165.7, 182.2, 171.7, 168.1, 189.9, 175.7, 163.4, 186.3, 169.5, 171.9, 173.9, 172.0, 172.7, 172.0\}$. y is random because we *could* have selected different heights from the table.

3.2 Estimators and Sampling Distributions

An *estimator* is a way of using the sample data y in order to "guess" something about the population that y comes from. In the example of the heights of male U of M students, we might be interested in knowing the mean height. The mean height would provide the best prediction for the height of the next random student that walks through the door. So, we collect our sample, $y = \{173.9, 171.7, 182.6, 181.5, 162.1, 174.9, 165.7, 182.2, 171.7, 168.1, 189.9, 175.7, 163.4, 186.3, 169.5, 171.9, 173.9, 172.0, 172.7, 172.0\}$. How should we use this sample to *estimate* the mean height?

The difference between a population value (such as the population mean or variance), and an estimator (such as the sample mean or variance), is very important. The population mean is the unobservable truth, and is a constant (non-random). The sample mean is an estimator for the population mean, and as we shall see, is a random variable. In this section we want to build up the idea of the *sampling distribution* of an estimator, in order to determine its properties. This will help us to determine if the estimator is "good".

3.2.1 Sample mean

A popular choice for estimating the population mean ($\mathrm{E}[y]$ or μ_y) is the *sample mean* (or *sample average*, or just *average*). The sample mean of y is usually denoted by $\bar{y}$. You have seen the equation for the sample mean before:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3.1}$$

where y_i denotes the i^{th} observation, and where n denotes the sample size. If we plug in our sample of heights into equation (3.1) we get $\bar{y} = 174.1$.

An important question is: how good is the estimator? That is, how good of a job is the estimator doing at "guessing" the true unobservable thing in the population? In our specific example: how good is the sample mean at estimating the true population mean of heights? This is an importannt question, because there are many ways that we could use the information in y to try to estimate the mean height. Why is equation (3.1) so popular?

To answer these questions, we need to enter a hypothetical situation, which will likely not be the case in the real world. Let's pretend we can "see" the entire population of heights (all of Table 3.1). If we can see all of Table (3.1), and not just the sample y, then we know the true mean height. We just take the average of the entire population, and get 176.8. So, $\bar{y} = 174.1$ is wrong!

Recall that the sample, y, is random. Each element of y was selected randomly from the population. We could have selected a different sample of size $n = 20$. For example, in a parallel universe, we could have gotten $y^* = \{175.9, 175.3, 182.2, 178.6, 175.2, 180.3, 178.3, 183.7, 176.0, 167.4, 178.7, 178.7, 186.0, 175.6, 180.0, 168.7, 178.6, 173.1, 173.2, 187.1\}$, where the * in y^* denotes that we are in the parallel universe. In this parallel universe, we got $\bar{y}^* = 177.6$. But in every universe, the population (table 3.1), is the same.

So, $\bar{y}$ is a random variable. $\bar{y}$ is random because y is random. We *could* have drawn a different random sample, in which case we would have gotten a different $\bar{y}$. In our example, there are a near infinite number (about 4×10^{20}) of different samples of size $n = 20$, and $\bar{y}$s, that we could get from the same population. Some of the $\bar{y}$s will be close to the true population mean height of 176.8, others far away. Whether or not $\bar{y}$ is a good idea for estimating the population mean $\mathrm{E}(y)$ can be determined by analyzing all the possible values that $\bar{y}$ can take.

3.2.2 Sampling distribution of the sample mean

Recall the discussion on *probability functions* in Chapter 2. A random variable (usually) has a probability function. This probability function describes all the possible values that the random variable can take, assigning a probability to each possibility. The form of the probability function depends on the nature of the random variable.

When the random variable is an *estimator*, then the probability function gets a special name - the *sampling distribution*. That is, a *sampling distribution* is just a fancy name for the probability function of an estimator. The sampling distribution is a hypothetical construct. It describes the probability of outcomes of $\bar{y}$, but in the real world we only get one sample y and one estimate $\bar{y}$.

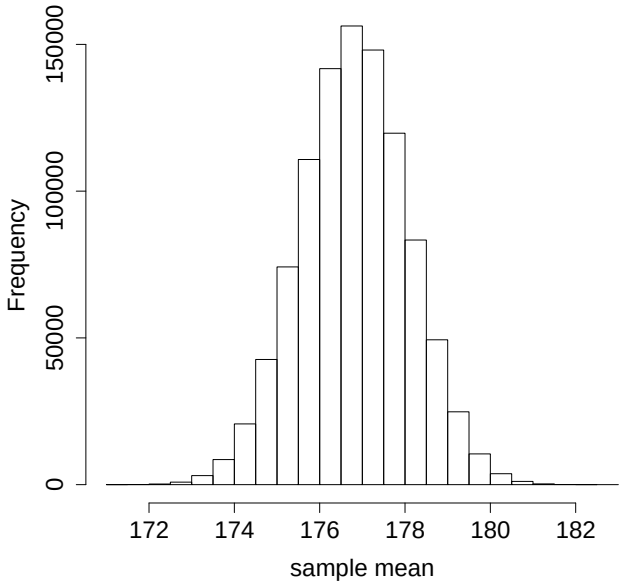
An alternative way of defining the sampling distribution follows. Imagine that you could draw all possible random samples of size $n = 20$ from the population, calculate $\bar{y}$ each time, and construct a relative frequency diagram (a histogram) for all of the $\bar{y}$s. This relative frequency diagram would be the sampling distribution of the estimator $\bar{y}$ for $n = 20$.

This alternative definition of the sampling distribution can be approximated using a computer. Using a computer, I have drawn 1 million different random samples of size $n = 20$ from table (3.1), and have calculated $\bar{y}$ each time. (This takes about 10 seconds on a fast computer). I have drawn a histogram using all of the $\bar{y}$s (figure 3.1). Figure (3.1) is a simulated sampling distribution.

Which probability function describes $\bar{y}$? Look again at equation (3.1). Notice the summation operator. $\bar{y}$ involves taking the sum of random variables (the y_is). It turns out that if the sample size is large enough (our $n = 20$ might be a bit too small) then the central limit theorem applies, and $\bar{y}$ is normally distributed (recall the summation of dice). Notice also that figure (3.1) resembles a normal distribution.

We will derive some features of an estimator from its sampling distribution. These features will tell us whether the estimator is "good" or "bad". Some important properties of the estimator are its mean (expected value) and its variance. This may be a strange idea at first. For example, we will take the expected value of the sample mean (which is an estimator for an expected value). That is, we will take

Figure 3.1: Histogram for 1 million $\bar{y}$s



the mean of the sample mean (meta!).

Three important properties of an estimator, that will largely guide whether the estimator is "good" or not, are *bias*, *efficiency*, and *consistency*. These properties are partly determined from the sampling distribution of the estimator, and we will now discuss each property in turn.

3.2.3 Bias

What happens if we consider the expected value, or the mean, of an estimator? An estimator is random, so it should have a mean. What would we want the expected value of the estimator to be? The thing we are trying to estimate, of course. So, if we are estimating the population mean using the sample mean (equation 3.1), then we want to get the "right" answer on average. That is, we want $\mathrm{E}[\bar{y}] = \mathrm{E}[y]$. If this is true, then I can "expect" to get the right answer when using $\bar{y}$ in many situations.

If $\mathrm{E}[\bar{y}] = \mathrm{E}[y]$, then $\bar{y}$ is said to be unbiased. If $\mathrm{E}[\bar{y}] \neq \mathrm{E}[y]$, then $\bar{y}$ would be a biased estimator; it would not give us the "right" answer on average. Given the popularity of $\bar{y}$ as an estimator for the population mean, you might anticipate that it is an unbiased estimator. The following is a short proof of the unbiasedness of the sample average.

Assume that $y_i \sim (\mu_y, \sigma_y^2)$, and that the y_is are iid. This just says that each random variable, y_i, in the sample, has the same population mean (μ_y) and population variance (σ_y^2). Now, take the expected

value of the estimator:

$$\begin{aligned}
\mathrm{E}\left[\bar{y}\right] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n}\mathrm{E}\left[y_1 + y_2 + \cdots + y_n\right] \\
&= \frac{1}{n}\left(\mathrm{E}\left[y_1\right] + \mathrm{E}\left[y_2\right] + \cdots + \mathrm{E}\left[y_n\right]\right) \\
&= \frac{1}{n}\left(\mu_y + \mu_y + \cdots + \mu_y\right) \\
&= \frac{n\mu_y}{n} = \mu_y
\end{aligned} \tag{3.2}$$

We find that the expected value of $\bar{y}$ is equal to the true unobservable population mean, and so $\bar{y}$ is an unbiased estimator.

3.2.4 Efficiency

Suppose that the estimator is unbiased. What happens now if we consider the variance of an estimator? What do want this variance to be? We would want it to be as small as possible. That is, we would want the estimator to have a high probability of being close to the thing we are trying to estimate. In the case of $\bar{y}$, we should hope that the $\mathrm{Var}[\bar{y}]$ is small so that on average, $\bar{y}$ is close to μ_y.

Efficiency is when an estimator has the smallest variance, compared to all other potential estimators. We will restrict our attention to other estimators that are also linear and unbiased. So, $\bar{y}$ is efficient if $\mathrm{Var}[\bar{y}] \leq \mathrm{Var}[\hat{\mu}_y]$, where $\hat{\mu}_y$ is any other linear unbiased estimator for the population mean of y. It turns out that there are many linear and unbiased estimators for the population mean, but that the sample mean has the smallest variance. So, we say that $\bar{y}$ is efficient.

The proof of the efficiency of $\bar{y}$ is omitted, however, an important part of the proof is included. In order to compare the variance of $\bar{y}$ to other potential estimators, we first have to be able to derive it:

$$\begin{aligned}
\mathrm{Var}\left[\bar{y}\right] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} y_i\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[y_1 + y_2 + \cdots + y_n\right] \\
&= \frac{1}{n^2}\left(\mathrm{Var}\left[y_1\right] + \mathrm{Var}\left[y_2\right] + \cdots + \mathrm{Var}\left[y_n\right]\right) \\
&= \frac{1}{n^2}\left(\sigma_y^2 + \sigma_y^2 + \cdots + \sigma_y^2\right) \\
&= \frac{n\sigma_y^2}{n^2} = \frac{\sigma_y^2}{n}
\end{aligned} \tag{3.3}$$

Note that the n in the denominator means the variance gets smaller as the sample size grows. That is, a larger sample provides an estimate that is on average closer to the true population mean. This is one reason why larger samples are better than smaller ones.

Now that we have derived the mean and variance of $\bar{y}$, and have used the central limit theorem to say that $\bar{y}$ is normally distributed, we can write the full sampling distribution: $\bar{y} \sim N(\mu_y, \sigma_y^2/n)$. Recall that this sampling distribution contains all the knowledge that we can have about the random variable $\bar{y}$. This sampling distribution is not only useful to determine the properties of unbiasedness, efficiency, and consistency, but will also be useful for hypothesis testing.

3.2.5 Consistency

Consistency is the last statistical property of an estimator that we will consider. An estimator is consistent if, having all possible information in the population, it provides the "right answer" every time. That is, as the sample size grows to infinity, the estimator provides the thing it's trying to estimate with probability 1. Two conditions are required for $\bar{y}$ to be (strongly) consistent: $\lim_{n\to\infty} \mathrm{E}[\bar{y}] = \mu_y$ and $\lim_{n\to\infty} \mathrm{Var}[\bar{y}] = 0$. The first condition says that the bias should disappear as the sample size grows. Since $\bar{y}$ is unbiased this condition is easily met. The second condition says that the variance of the estimator should go to 0 as the sample size grows; this is easily verified by noting the n in the denominator of $\mathrm{Var}[y]$.

Consistency is the most important property for an estimator to have. Without consistency, the estimator is useless. In all, we have shown that $\bar{y}$ is unbiased, efficient, and consistent. Sometimes the acronym BLUE (best linear unbiased estimator) is used to describe such an estimator. That $\bar{y}$ is BLUE is a very good reason to use it as an estimator for μ_y, among the many possibilities.

3.3 Hypothesis Tests (known σ_y^2)

The types of hypotheses we are talking about concern statements about the unobservable population. For example, we might hypothesize that the true population mean height of U of M students is 173 cm. A hypothesis test uses the information in the sample to assess the plausibility of the hypothesis. In general, a hypothesis test begins with a null hypothesis, and an alternative hypothesis. For example:

$$\begin{aligned} H_0 &: \mu_y = \mu_{y,0} \\ H_A &: \mu_y \neq \mu_{y,0} \end{aligned} \tag{3.4}$$

H_0 is the null hypothesis. The null hypothesis is "choosing" a value for the population mean, μ_y. The hypothesized value of the population mean is denoted $\mu_{y,0}$. The alternative hypothesis (H_A) is two-sided; the null hypothesis is wrong if the population mean (μ_y) is either "too small" or "too big" relative to the hypothesized value. Since most tests in econometrics are two-sided, we will not consider one-sided tests here, although they are very similar.

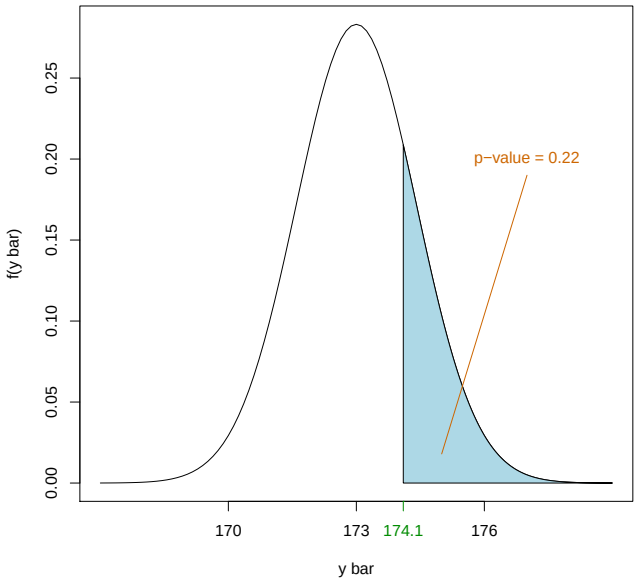
The hypothesis test concludes with either: (i) "reject" H_0 in favour of H_A, or (ii) "fail to reject" H_0. Which decision is reached ultimately depends on a probability (p-value), and on the researcher (you) deciding subjectively whether this probability is small or large. The sample data, and our knowledge of the sampling distribution of the estimator, will determine this probability.

Let's go back to the heights example. From our sample of $n = 20$ we estimated the population mean to be $\bar{y} = 174.1$. Suppose that the null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \mu_y = 173 \\ H_A &: \mu_y \neq 173 \end{aligned} \tag{3.5}$$

Our estimate of 174.1 is clearly different from our hypothesis that the true population mean height is 173. Notice that the difference between what we actually estimated from the sample, and our null hypothesis, is $174.1 - 173 = 1.1$. This difference of 1.1 does not necessarily imply we should reject the

Figure 3.2: Normal distribution with $\mu = 173$ and $\sigma^2 = {}^{39.7}/_{20}$. Shaded area is the probability that the normal variable is greater than 174.1.



null hypothesis. Rather, is this difference big enough to warrant rejection of H_0? More accurately, we should only reject H_0 if the probability of getting a $\bar{y}$ further away than 1.1 from H_0, is small. This probability is called a p-value.

Recall once again that $\bar{y}$ is a random variable. Its value depends on the random sample that we draw from the population. A different sample might give us $\bar{y} = 190$. This would be "worse" for the null hypothesis of 173, than getting the value $\bar{y} = 174.1$. Out of all the samples that we could draw, out of all the parallel universes, what proportion of them would provide a $\bar{y}$ that is further than 1.1 from H_0? Imagine that only 4.3% of possible samples from the population were further than 1.1 from H_0. We have to decide one of two things. Either we have witnessed a rare event (are living in a strange universe) and the null is true, or the null is false. The actual p-value for this example is not 4.3%. We will now discuss how to determine the actual p-value for this problem, and for other problems in general.

As we have repeatedly stated, $\bar{y}$ is a random variable. It has a probability function, which we call a sampling distribution (because it's an estimator). We have derived the sampling distribution: $\bar{y} \sim N(\mu_y, \sigma_y^2/n)$. The sampling distribution can be used to calculate various events involving $\bar{y}$. For example, if we want to know the probability that $\bar{y} > 18$, we can draw out the normal curve (provided that we know μ_y and σ_y^2/n) and calculate the area under the curve, to the right of 18.

Classical hypothesis testing proceeds by assuming that H_0 is true. If H_0 is true, then the sampling distribution of $\bar{y}$ is $N(\mu_{y,0}, \sigma_y^2/n)$. That is, if the null hypothesis is correct, the true mean of $\bar{y}$ is $\mu_{y,0}$. To calculate the p-value, we still need to know σ_y^2. For now, we will assume that it is know, but this is an unrealistic assumption. In the real world, we will have to estimate σ_y^2.

Assuming that we know that $\sigma_y^2 = 39.7$ (again, this is very unrealistic) then we have the variance of the sample average ($\sigma_y^2/n = {}^{39.7}/_{20} = 2.0$), and so the full sampling distribution of the sample mean under the null hypothesis is: $\bar{y} \sim N(173, 2)$. This probability function is drawn in figure (3.2). All that remains is to calculate the probability of obtaining a $\bar{y}$ that is *more adverse* to the null hypothesis than the one we just calculated. Half of this probability is represented by the shaded region in figure (3.2). This is a two sided test, so it doesn't matter if $\bar{y}$ is too large or too small: we need to multiply the

one-sided p-value by 2. So, the p-value for our two-sided test is $0.22 \times 2 = 0.44$.

The interpretation of the p-value of 0.44 is as follows. If the null hypothesis of $H_0 = 173$ is true, then there is a 44% chance of observing a $\bar{y}$ that is further away from 173 than the difference of $174.1 - 173 = 1.1$ that we just observed. Would you "reject" or "fail to reject" based on this? Most researchers would fail to reject. There is a high probability of getting a $\bar{y}$ much more adverse to the null, so the null seems plausible.

3.3.1 Significance of a test

At what point should we decide that the p-value is too small, and reject the null hypothesis? The choice is somewhat arbitrary, and is up to the researcher (you). Standard choices have been 10%, 5%, and 1%. A pre-decided maximum p-value under which H_0 will be rejected is called the *significance level* of the test. It is sometimes denoted by α. In the previous example, we fail to reject the null at the 10% significance level. Note that failing to reject at the 10% level implies that we also fail to reject H_0 at the 5% and 1% significance levels.

3.3.2 Type I error

Take another look at figure (3.2). Even when the null hypothesis is true and figure (3.2) is the correct sampling distribution for $\bar{y}$, we will sometimes randomly draw a weird sample that makes H_0 appear to be "wrong". That is, even when the null is true, in some of the parallel universes we will draw a sample that gives a $\bar{y}$ that is very far from the truth. In these cases, we will erroneously reject the null. If the null hypothesis is falsely rejected, it is called a *type I error*. Type I error is the probability that H_0 is rejected when the null is true:

$$\Pr(\text{type I error}) = \Pr(\text{reject } H_0 \mid H_0 \text{ is true}) \tag{3.6}$$

How do we determine what this type I error will be? As soon as we pick the significance of the test, it has been determined. That is, type I error $= \alpha$. When we decide that 5% of $\bar{y}$s that are furthest from H_0 are just too rare, we are deciding that we will make a type I error in 5% of the parallel universes (or in 5% of other similar situations). That is, if we conduct thousands of scientific studies where we always use $\alpha = 5\%$, in 5% of those studies where we reject the null, we will be doing so falsely.

In reality, we do not know the population values, so we will never know if we have made a type I error or not. That is, the idea of type I error tells us nothing about the particular sample that we are working with. It only tells us something about what happens through repeated applications of our tested procedure.

3.3.3 Type II error

There is another type of error we can make. There are two possibilities for H_0: either it is true or false. In type I error, we considered that H_0 is actually true. If we consider that H_0 is actually false, then we make a *type II error* if we *fail to reject*. The probability of a type II error is:

$$\Pr(\text{type II error}) = \Pr(\text{fail to reject } H_0 \mid H_0 \text{ is false}) \tag{3.7}$$

If H_0 is actually false, one of two things can happen: we "reject" or we "fail to reject". The probabilities of both of these events must sum to 1 (something must happen). So:

$$\Pr(1 - \text{type II error}) = \Pr(\text{reject } H_0 \mid H_0 \text{ is false}) \tag{3.8}$$

Equation (3.8) is called the *power* of the test. We want the power to be as high as possible. That is, we do not want to make a type II error, and we want the probability of rejection to be as high as possible when H_0 is actually false.

Determining the type II error (and power) of a test is difficult or impossible. This is because power depends on knowing the unobservable population. The concept is useful, however, when we are trying to find the "best" test available. In may be possible to determine that some ways of testing are more powerful than others, even though we may not know what the actual numbers are.

3.3.4 Test statistics

A *test statistic* is a convenient way of assessing the null hypothesis, and provides an easier way to obtain a p-value. If we wanted to use the above testing procedure for different problems, we would have to "graph" a different normal curve (similar to the one in figure 3.2), and calculate a different area under the curve, for each testing problem. Decades ago, calculating an area under the normal curve was difficult (now it is easily done by computers). Consequently, a method was devised so that every such testing problem would use the *standard normal curve*. That way, different areas under the curve could be tabulated for various values on the x-axis.

To *standardize* a variable, we subtract its mean and divide by its standard deviation. This creates a new normal random variable from the old one, called a "standard normal" variable. For example, let $y \sim N(\mu_y, \sigma_y^2)$. Create a new variable z where:

$$z = \frac{y - \mu_y}{\sigma_y} \tag{3.9}$$

Now, z is still normally distributed, but has mean 0 and variance 1 since

$$\mathrm{E}[z] = \mathrm{E}[y - \mu_y] = \mathrm{E}[y] - \mu_y = \mu_y - \mu_y = 0$$

and

$$\mathrm{Var}[z] = \mathrm{Var}\left[\frac{y}{\sigma_y}\right] = \frac{\mathrm{Var}[y]}{\sigma_y^2} = \frac{\sigma_y^2}{\sigma_y^2} = 1$$

(refer to the rules of mean and variance).

How is this helpful? Recall the sampling distribution of $\bar{y}$ under the null hypothesis: $\bar{y} \sim N(\mu_{y,0}, \sigma_y^2/n)$. Create a new variable z. Subtract $\mu_{y,0}$ (the mean of $\bar{y}$ if the null is true) from $\bar{y}$. Now z has mean 0 (if the null is actually true). Divide by the standard error (standard error = the standard deviation of an estimator) of $\bar{y}$, and z has variance of 1. That is:

$$z = \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2/n}} \sim N(0,1) \tag{3.10}$$

This is the "z test statistic" for the null hypothesis that $\mu_y = \mu_{y,0}$. If the null is true, then $\bar{y}$ should be close to $\mu_{y,0}$, implying that z should be close to 0. The probability of observing a $\bar{y}$ further away from H_0 than what we just observed from the sample is obtained by plugging $\bar{y}$ and $\mu_{y,0}$ into the z statistic formula, and calculating a probability using the standard normal distribution. From our heights example, the z statistic is:

$$z = \frac{174.1 - 173}{\sqrt{\frac{39.7}{20}}} = 0.78$$

Now, the question: "what is the probability of getting further away than 174.1 from the null hypothesis of 173?" has just been translated to: "What is the probability of a $N(0,1)$ variable being greater than 0.78 (or less than -0.78)?" So, as you may have guessed:

$$\Pr(z > 0.78) = 0.22 \tag{3.11}$$

Since all such testing problems can be standardized, we only need to calculate the area under the curve for several possible z values. These were tabulated long ago, and are reproduced in Table (3.2).

3.3.5 Critical values

Critical values are the most extreme values allowable for the test statistic, before the null hypothesis is rejected. Suppose that we choose a 5% significance level for our test. This means that if we receive a p-value that is less than 0.0250 in Table 3.2, we should reject the null hypothesis (since $2.5\% \times 2 = 5\%$). If we use Table 3.2 to find the z statistic that corresponds to a significance level, we are finding the critical value for the test. According to Table 3.2, we see that a p-value of 0.0250 corresponds to a z statistic of 1.96. This is the 5% critical value. We know that if the z statistic that we calculate for our test end up being greater than 1.96 or less than -1.96, we will get a p-value that is less than 0.05, and we will reject the test.

3.3.6 Confidence intervals

A confidence interval corresponds to a significance level. Suppose that the significance level is 5%. Then, the 95% confidence interval contains all of the values for $\mu_{y,0}$ (all values for null hypotheses) that will not be rejected at 5% significance.

What is the probability that our z statistic will be within a certain interval, if the null hypothesis is true? For example, what is the following probability?

$$\Pr\left(-1.96 \leq z \leq 1.96\right)? \tag{3.12}$$

Using Table 3.2, we can figure out that this probability is 0.95. Note that -1.96 and 1.96 are the left and right critical values, respectively, for a test with 5% significance. Now, to solve for the confidence interval around $\bar{y}$, we will first substitute the formula for the z statistic into equation 3.12:

$$\Pr\left(-1.96 \leq \frac{\bar{y} - \mu_{y,0}}{\sqrt{\sigma_y^2/n}} \leq 1.96\right) = 0.95 \tag{3.13}$$

Finally, we solve equation 3.13 so that the null hypothesis $\mu_{y,0}$ is in the middle of the probability statement:

$$\Pr\left(\bar{y} - 1.96 \times \sqrt{\frac{\sigma_y^2}{n}} \leq \ \mu_{y,0} \ \leq \bar{y} + 1.96 \times \sqrt{\frac{\sigma_y^2}{n}}\right) = 0.95 \tag{3.14}$$

This just says that $1.96 \times \sigma_y^2/n$ is the maximum distance that the null hypothesis can be from the sample average that we calculate, before we would get a p-value less than 0.05, and reject the test at the 5% significance level.

An alternative interpretation of the confidence interval (other than containing the set of values for the null that won't be rejected), is the following. Out of many such 95% confidence intervals that we construct in many hypothesis tests, 95% of such intervals will include the true population mean, μ_y. Two common *misinterpretations* of a confidence interval are: (i) there's a 95% probability that μ_y lies within the interval; and (ii) the confidence interval includes μ_y 95% of the time. The reason these last two interpretation are wrong has to do with the fact that the confidence interval is *random* and μ_y is fixed.

3.4 Hypothesis Tests (unknown σ_y^2)

So far we have assumed that σ_y^2 is known. We needed this σ_y^2 in order to calculate the variance of $\bar{y}$ (which is σ_y^2/n), and calculate our p-value.

But, if we have to estimate μ_y, it is unlikely that we would know σ_y^2. That is, if the population mean is unknown, it is likely that the population variance would be unknown as well. Hence, we now need to figure out how to estimate σ_y^2 from our sample of data, y.

3.4.1 Estimating σ_y^2

Recall that the variance for a discrete random variable is defined as:

$$\text{Var}(Y) = \sum_{i=1}^{K} p_i \times (Y_i - \text{E}[Y_i])^2$$

where Y_i are the different values that the random variable can take, and p_i are the probabilities of those values occurring. A sensible way of estimating σ_y^2 may be to take the *sample average* of the squared distances, but replacing $\text{E}[Y_i]$ with $\bar{y}$. That is, a natural estimator for σ_y^2 might be:

$$\hat{\sigma}_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.15}$$

When we considered whether or not $\bar{y}$ was a good estimator for μ_y, we first took the expected value of $\bar{y}$, and determined that it was *unbiased*. That is, it turned out that $\text{E}[\bar{y}] = \mu_y$. Well, it turns out that $\hat{\sigma}_y^2$ is a *biased* estimator! We won't derive the expected value here, we will only state it:

$$\text{E}\left[\hat{\sigma}_y^2\right] = \frac{n-1}{n}\sigma_y^2 \tag{3.16}$$

Equation 3.16 says that if we were to use equation 3.15 to estimate the variance of y, on average our estimate would be a little bit too small compared to the truth (by a factor of $(n-1)/n$). However, armed with this knowledge, we can construct what is called a *bias corrected* estimator. If we just multiply the right-hand-side of 3.16 by $n/(n-1)$, the bias disappears! That is, if we multiply the estimator $\hat{\sigma}_y^2$ by $n/(n-1)$, the resulting estimator is unbiased. This bias corrected estimator is usually denoted s_y^2, where:

$$s_y^2 = \frac{n}{n-1} \times \hat{\sigma}_y^2 = \frac{n}{n-1} \times \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.17}$$

3.4.2 The *t*-test

Now that we know how to estimate σ_y^2, we can estimate the variance of the sample average using:

$$\text{Estimated variance of } \bar{y} = \frac{s_y^2}{n}$$

We can implement hypothesis testing by replacing the unknown σ_y^2 with its estimator s_y^2. The z test statistic now becomes:

$$\frac{\bar{y} - \mu_{y,0}}{\sqrt{s_y^2/n}} = t$$

This is the t statistic. Because we have replaced σ_y^2 with s_y^2 (a random estimator) in the z statistic formula, the form of the randomness of z has changed. The t statistic is no longer a standard normal variable. It follows its own probability distribution, called the t distribution. When performing a t test, the p-values are different than in Table 3.2. However, as the sample size grows, the t distribution becomes the standard normal distribution. This means that, for sample sizes of approximately $n > 100$, using the standard normal distribution (Table 3.2) instead of the t distribution, makes very little difference. For the purposes of this course, we will assume that the sample size is large enough that the t statistic follows a standard normal distribution.

Finally, note that confidence intervals can be constructed, in practice, by replacing the unknown σ_y^2 in equation 3.14 with the estimator s_y^2. As long as the sample size is reasonably large, we do not have to worry about replacing the critical values in the confidence interval formula (for example, 1.96) with critical values from the t distribution. An example of performing a t test and constructing a confidence interval, is left for the Review Questions.

3.5 Review Questions

1. Prove that $\bar{y}$ is a random variable. Why might $\bar{y}$ follow a Normal distribution? What is the sampling distribution for $\bar{y}$?
2. Derive the mean and variance of $\bar{y}$. How does this help us determine if $\bar{y}$ is: (i) unbiased; (ii) efficient; and (iii) consistent?
3. Assume that $y_i \sim (\mu_y, \sigma_y^2)$, and that y_i is i.i.d. Let $\tilde{\mu}_y = \frac{y_1+y_n}{2}$. Is $\tilde{\mu}_y$ an unbiased estimator for μ_y? Compare the variance of $\tilde{\mu}_y$ to the variance of $\bar{y}$.
4. Assume that $y_i \sim (\mu_y, \sigma_y^2)$, that y_i is i.i.d., and that the sample size, n, is even. Let

$$\hat{\mu}_y = \frac{1}{2n}y_1 + \frac{3}{2n}y_2 + \frac{1}{2n}y_3 + \frac{3}{2n}y_4 + \cdots + \frac{1}{2n}y_{n-1} + \frac{3}{2n}y_n$$

   Is $\hat{\mu}_y$ an unbiased estimator for μ_y? Compare the variance of $\hat{\mu}_y$ to the variance of $\bar{y}$.
5. Refer to the above two questions. Are $\tilde{\mu}_y$ and $\hat{\mu}_y$ consistent estimators for μ_y?
6. Perform a t test of the null hypothesis in equation (3.5), using the heights data from table 3.1. Also, construct 95% and 90% confidence intervals around $\bar{y}$.

3.6 Answers

1. The formula for $\bar{y}$ is $1/n \sum_{i=1}^{n} y_i$. It is a linear function of the random y_i values, so it is a random variable itself. $\bar{y}$ might follow a Normal distribution due to the central limit theorem, which (loosely speaking) says that if we add up random variables the resulting sum tends to be Normally distributed. Note the summation operator in the formula for $\bar{y}$. Finally, the full sampling distribution can be written as: $\bar{y} \sim N(\mu_y, \sigma_y^2/n)$.
2. The mean of $\bar{y}$ is derived in equation (3.2) and the variance in equation (3.3). (i) The mean of $\bar{y}$ tells us that the estimator is unbiased. (ii) The variance of $\bar{y}$ allows us to compare to the variance of all other possible linear and unbiased estimators of μ_y, and determine that σ_y^2/n is smallest, and thus $\bar{y}$ is efficient. (iii) The n in the denominator of σ_y^2/n shows us that $\bar{y}$ is consistent. We know that the estimator is unbiased, and as the sample size grows, the variance of $\bar{y}$ goes to zero. This means that with a infinitely large sample size, our estimator would give the value μ_y with probability 1.
3. To derive the bias of the estimator $\tilde{\mu}_y$, we compare its expected value to μ_y:

$$\mathrm{E}\left[\tilde{\mu}_y\right] = \mathrm{E}\left[\frac{y_1 + y_n}{2}\right] = \frac{1}{2}\mathrm{E}\left[y_1 + y_n\right] = \frac{2\mu_y}{2} = \mu_y$$

   Since the expected value of the estimator is equal to μ_y, the estimator is unbiased.
   The variance of $\tilde{\mu}_y$ is:

$$\mathrm{Var}\left[\tilde{\mu}_y\right] = \mathrm{Var}\left[\frac{y_1 + y_n}{2}\right] = \frac{1}{4}\mathrm{Var}\left[y_1 + y_n\right]$$

   The i.i.d. assumption gives us the independence of the y_i values, allowing us to expand within the variance operator:

$$\frac{1}{4}\mathrm{Var}\left[y_1 + y_n\right] = \frac{1}{4}\left(\mathrm{Var}\left[y_1\right] + \mathrm{Var}\left[y_n\right]\right) = \frac{2\sigma_y^2}{4} = \frac{\sigma_y^2}{2}$$

   Comparing this variance to the variance of the sample average, we find:

$$\frac{\sigma_y^2}{2} > \frac{\sigma_y^2}{n} \quad ; n > 2$$

   which is not surprising result, since we know that $\bar{y}$ is an efficient estimator.

4. Again, we start by taking the expected value of the estimator:

$$\begin{aligned}\mathrm{E}\left[\hat{\mu}_y\right] &= \mathrm{E}\left[\tfrac{1}{2n}y_1 + \tfrac{3}{2n}y_2 + \tfrac{1}{2n}y_3 + \cdots + \tfrac{3}{2n}y_n\right] \\ &= \tfrac{1}{2n}\mu_y + \tfrac{3}{2n}\mu_y + \tfrac{1}{2n}\mu_y + \cdots + \tfrac{3}{2n}\mu_y \\ &= \mu_y\end{aligned}$$

So, $\hat{\mu}_y$ is an unbiased estimator.
Next, we find the variance of $\hat{\mu}_y$, again making use of the independence assumption:

$$\begin{aligned}\mathrm{Var}\left[\hat{\mu}_y\right] &= \mathrm{Var}\left[\tfrac{1}{2n}y_1 + \tfrac{3}{2n}y_2 + \tfrac{1}{2n}y_3 + \cdots + \tfrac{3}{2n}y_n\right] \\ &= \frac{1}{4n^2}\mathrm{Var}\left[y_1\right] + \frac{9}{4n^2}\mathrm{Var}\left[y_2\right] + \ldots \\ &= \frac{1}{4n^2}\,\sigma_y^2 + \frac{9}{4n^2}\,\sigma_y^2 + \ldots \\ &= \frac{5}{4n}\,\sigma_y^2\end{aligned}$$

We can see that this variance is larger than the variance of $\bar{y}$, which is another illustration of the efficiency property of $\bar{y}$.

5. $\tilde{\mu}_y$ (for example) is a consistent estimator if $\lim_{n\to\infty}\mathrm{E}[\tilde{\mu}_y] = \mu_y$ and $\lim_{n\to\infty}\mathrm{Var}[\tilde{\mu}_y] = 0$. We have already shown that the estimator is unbiased, so the first condition is satisfied. However, the variance of this estimator does not go to 0 as the sample size increases, so this estimator is not consistent! That is:

$$\lim_{n\to\infty}\frac{\sigma_y^2}{2} = \frac{\sigma_y^2}{2}$$

On the other hand, the estimator $\hat{\mu}_y$ is consistent, since there is an n in the denominator of $\frac{5}{4n}\sigma_y^2$.

6. The null and alternative hypotheses are:

$$\begin{aligned}H_0 &: \mu_y = 173 \\ H_A &: \mu_y \neq 173\end{aligned}$$

The sample mean and the sample variance are $\bar{y} = 174.1$ and $s_y^2 = 53.0$. The sample size is $n = 20$. The t statistic is:

$$t = \frac{174.1 - 173}{\sqrt{53.0/20}} = 0.68$$

Assuming that the sample size is large enough (even though $n = 20$ is too small), we can use the standard Normal distribution, and table 3.2 to find that the $p-$value $= 0.2483 \times 2 = 0.5$. We fail to reject the null hypothesis.
The 95% confidence interval is:

$$\bar{y} \pm 1.96 \times \sqrt{s_y^2/n} = 174.1 \pm 1.96 \times 1.63 = [170.9, 177.3]$$

For the 90% confidence interval, we need to change the critical value of 1.96. Using table 3.2, we find the z value which has 5% area under the curve (5% $\times$ 2 = 10% significance, 100% - 10% = 90% confidence). The 10% critical value is 1.64, so the 90% confidence interval is:

$$\bar{y} \pm 1.64 \times \sqrt{s_y^2/n} = 174.1 \pm 1.64 \times 1.63 = [171.4, 176.8]$$

Table 3.2: Area under the standard normal curve, to the right of z.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002

Chapter 4

Ordinary Least Squares (OLS)

In this chapter, we discuss a method to estimate the marginal effect of one variable on another. Economic models typically imply that one variable *causes* or *determines* another variable. Seldom (or never) does the economic model *quantify* the marginal effect, however. We need data and econometrics in order to estimate a *number* for the marginal effect.

We begin the chapter with two motivating examples. They are meant to show that many simple economic models can be represented through the equation for a line. We then proceed to estimate this line uses data. The method that we use to fit a straight line through data points is *ordinary least squares* (OLS) or just *least squares*. We will make some simplifying assumptions, and discuss the properties of the OLS estimator.

4.1 Motivating Example 1: Demand for Liquor

How much less alcohol will people consume if we raise the price? In first-year microeconomics you learned about the law of demand. The quantity demanded of a product should depend on its price (and other things):

$$Q_d = a + bP \tag{4.1}$$

where a is the intercept of the demand "curve", and b is the slope. See figure 4.1. You learned that the slope of the demand curve, b, depends on the type of good. For example, necessities such as medicine should have relatively flatter demand curves than luxuries such as a diamonds.

Estimating the slope of the demand curve is important for policy makers who might want to affect the quantity demanded of a good. For example, we might want to reduce consumption of alcohol or cigarettes by increasing price (taxing them). But before we fiddle with the price of these products, we should estimate how much quantity demanded will change given a change in price (if it changes at all).

Figure 4.1: A typical demand "curve". Note this is an "inverse" demand curve (quantity demanded is on the vertical axis, and price on the horizontal axis).

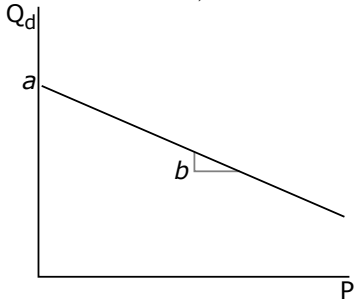
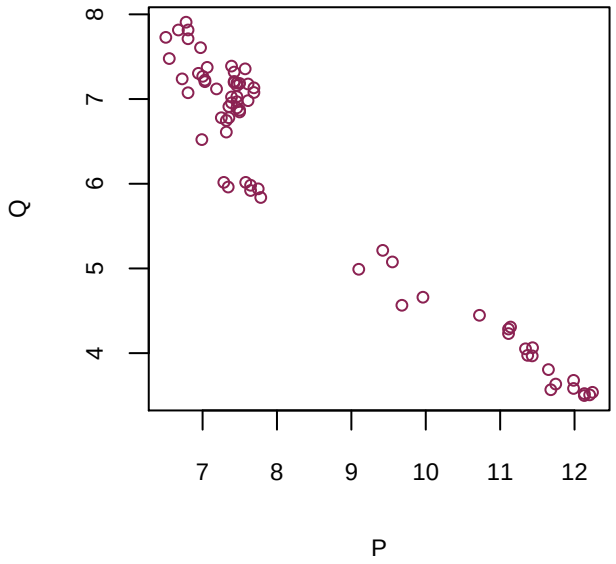


Figure 4.2: Per capita consumption, and price, of spirits. Choosing a line through the data necessarily chooses the slope of the line, b, which determines how much Q_d decreases for an increase in P.



Using data from Prest (1949), we plot the yearly (from 1870 to 1938) per-capita consumption of spirits (in proof gallons), and the relative price of spirits (deflated by a cost-of-living index). See figure 4.2. How should we fit a line through the data in figure 4.2? If we can pick a "good" line, then we will have a good estimate for the slope, b. This estimated b could then be used to determine how much alcohol consumption will decrease if we increase the tax on alcohol by \$1, for example. Note that b is the marginal effect of a change in price of spirits, on the quantity demanded of spirits, holding all else constant.

4.2 Motivating Example 2: Marginal Propensity to Consume

This example uses data on total disposable income and consumption (in millions of Pounds) from 1971-1985 (quarterly) in the U.K. (Verbeek and Marno, 2008). The data is shown in figure 4.3.

An increase in consumption is induced by an increase in income, but not all of the increase in income is consumed. Marginal propensity to consume is the proportion of an increase in disposable income that individuals spend on consumption:

$$MPC = \frac{\Delta C}{\Delta Y} \tag{4.2}$$

where ΔC is the change in consumption "caused" by the change in income, ΔY. John Maynard Keynes supposed that the MPC should be less than one, but without data and econometrics there is no way to put an actual number to the MPC.

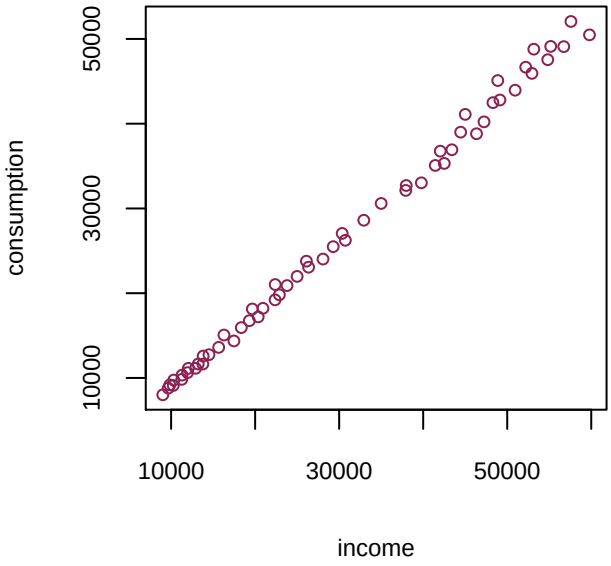
We can also write the relationship between consumption and disposable income through the equation of a line:

$$C = a + MPC \times Y \tag{4.3}$$

where a is again the intercept of the line (representing the amount of consumption with disposable income of zero), and where this time MPC is the *slope* of the line. Remember that MPC is the thing we are trying to estimate.

One of the points we are trying to make here is that many economics models can be represented by the equation of a straight line. If we can figure out how to estimate the line, then we have an estimate for the slope (the marginal effect), which is of great practical usefulness.

Figure 4.3: Income and consumption in the U.K. (Verbeek and Marno, 2008).



The next question is: how should we fit a line through data points (like the ones in figures 4.2 and 4.3)? Before we determine how to pick the line, however, we need to introduce some definitions and general notation.

4.3 The Linear Population Regression Model

The general regression model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{4.4}$$

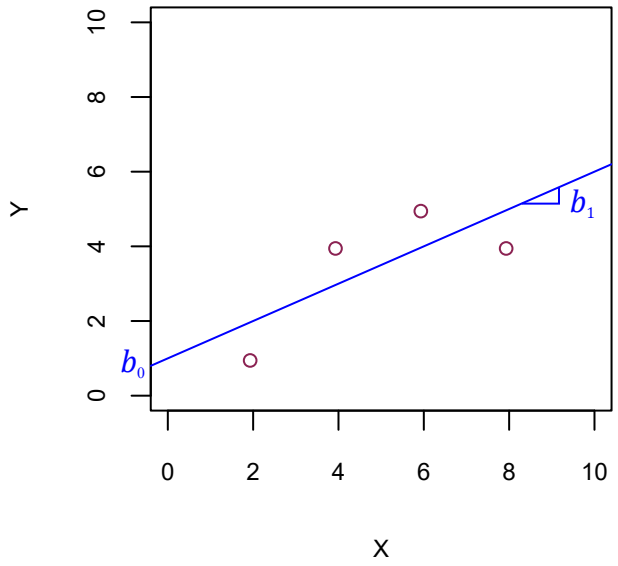
- X is called the *independent* variable or *regressor*. It is the variable that is assumed to *cause* the Y variable. In the "Demand for Liquor" example, this variable was *price* (P). See equation 4.1. In the MPC example the regressor was *income*. See equation 4.3.
- Y is the *dependent* variable. This variable is assumed to be caused by X (it *depends* on X). In the demand example the dependent variable was *quantity demanded* (Q_d) and in the MPC example it was *consumption* (C).
- β_0 is the population intercept. It was labelled a in both examples. It is unobservable, but we can try to estimate it.
- β_1 is the population slope. When X increases by 1, Y increases by β_1. This is the primary object of interest, and is unobservable. We want to estimate β_1. β_1 is interpreted as the marginal effect in many economics models.
- ϵ is the regression *error* term. It consists of all the other factors or variables that determine Y, other than the X variable. All of these other variables causing Y are combined into ϵ. ϵ is considered to be a random variable since we can not observe it.
- $i = 1, \ldots, n$. The subscript i denotes the observation. n is the sample size. For example, Y_4 refers to the fourth Y observation in the data set.

4.3.1 The importance of β_1

Note that in equation 4.4, the object of interest is β_1. It is the thing we are trying to estimate. It is the causal, or marginal effect, of X on Y. That is, a change in X of ΔX causes a β_1 change in Y:

$$\frac{\Delta Y}{\Delta X} = \beta_1$$

Figure 4.4: A simple data set with the estimated OLS line in blue. b_0 is the OLS intercept, and b_1 is the OLS slope.



4.3.2 The importance of ϵ

ϵ (epsilon) is the random component of the model. Without ϵ, statistics/econometrics is not required. ϵ represents all of the other things that determine Y, other than X. They are all added up and lumped into this one random variable. Because we can not observe all of these other factors, we consider them to be random. The fact that ϵ is random makes Y random as well.

Later, we will make some assumptions about the randomness of ϵ, that will ultimately determine the properties of the way that we choose to estimate β_1.

4.3.3 Why it's called a population model

Equation 4.4 is called a "population" model because it represents the true, but unknown way in which the Y variable is "created" or "determined". β_0 and β_1 are unknown (and so is ϵ). We will observe a sample of Y and X, and use the sample to try to figure out the βs.

4.4 The estimated model

Our primary goal is to estimate β_1 (the marginal effect of X on Y), but to do so we'll also have to estimate β_0. This estimated intercept and slope will define a straight line. These estimates will be denoted b_0 and b_1, the OLS intercept and slope.

Let's start with a very simple example using data that I made up: $Y = \{1, 4, 5, 4\}$, $X = \{2, 4, 6, 8\}$. The data, and estimated OLS line, are shown in figure 4.4. The OLS estimated intercept is $b_0 = 1$, and the estimated slope is $b_1 = 0.5$.

We still don't know how to get b_0 and b_1! Before we decide how to fit a straight line through some data points, we need to define two terms first.

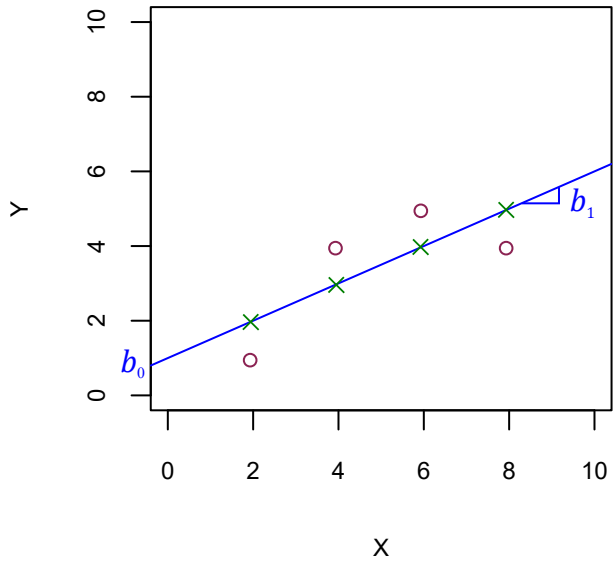
4.4.1 OLS predicted values ($\hat{Y}_i$)

The OLS predicted (or fitted) values, are the values for Y that we get when we "plug" the X values back into the estimated OLS line. These predicted Y values are denoted by $\hat{Y}$. We can find each predicted value, $\hat{Y}_i$, by plugging each X_i into the estimated equation.

In general, the estimated equation (or line) is written as:

$$\hat{Y}_i = b_0 + b_1 X_i. \tag{4.5}$$

Figure 4.5: The OLS predicted values shown by ×.



For our simple example, equation 4.5 becomes $\hat{Y}_i = 1 + 0.5X_i$, and each OLS predicted values is:

$$\hat{Y}_1 = 1 + 0.5(2) = 2$$
$$\hat{Y}_2 = 1 + 0.5(4) = 3$$
$$\hat{Y}_3 = 1 + 0.5(6) = 4$$
$$\hat{Y}_4 = 1 + 0.5(8) = 5$$

These OLS predicted values are added to the plot in figure 4.5. Notice how each predicted value lies on the blue line, directly above or below the data point.

4.4.2 OLS residuals (e_i)

An OLS predicted value tells us what the estimated model predicts for Y when given a particular value of X. When we plug in the sample values for X (as we did in the previous section), we see that the predicted values ($\hat{Y}_i$) don't quite line up with the actual Y_i values. The differences between the two are the OLS *residuals*. The OLS residuals are like prediction errors, and are determined by:

$$e_i = Y_i - \hat{Y}_i \tag{4.6}$$

Using equation 4.6 for our simple example, each OLS residual is:

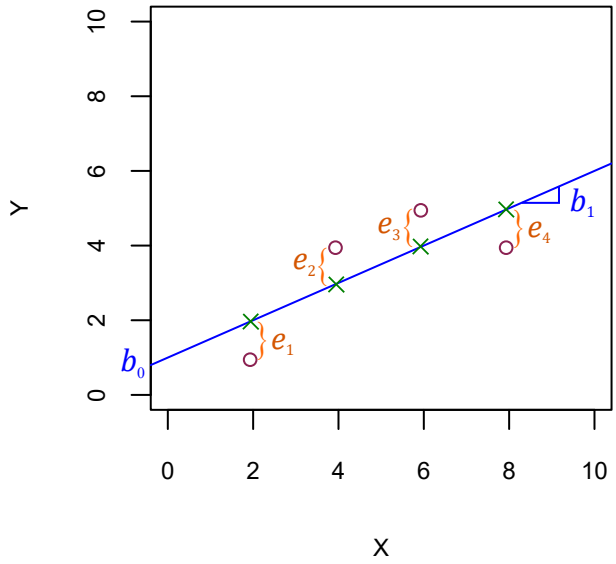
$$e_1 = 1 - 2 = -1$$
$$e_2 = 4 - 3 = 1$$
$$e_3 = 5 - 4 = 1$$
$$e_4 = 4 - 5 = -1$$

These OLS residuals are indicated in figure 4.6. They are the vertical distances between the actual data points (the circles) and the OLS predicted values (the ×).

Each data point (Y_i) is equal to its predicted value, plus its residual. That is, we can rearrange equation 4.6 and write:

$$Y_i = \hat{Y}_i + e_i$$

Figure 4.6: The OLS residuals (e_i) are the vertical distances between the actual data points (circles) and the OLS predicted values ($\times$).



or, using equation 4.5 for the definition of $\hat{Y}_i$:

$$Y_i = b_0 + b_1 X_i + e_i, \tag{4.7}$$

which will be useful in the next chapter. Note that equation 4.7 is the observable counterpart to the unobservable population model in equation 4.4.

4.5 How to choose b_0 and b_1, the OLS estimators

Now that we have defined the OLS residuals (e_i), we can define the OLS estimators b_0 and b_1 by coming up with an equation that will tell us how to use the X and Y data.

The OLS estimators are defined in the following way. They are the values for b_0 and b_1 that minimize the sum of squared vertical distances between the OLS line and the actual data points (Y_i). These vertical distances have already been defined as the OLS residuals (e_i). So the "objective" is to choose b_0 and b_1 so that $\sum_{i=1}^{n} e_i^2$ is minimized. This is an optimization problem from calculus. Formally stated, the OLS estimator is the solution to the minimization problem:

$$\min_{b_0, b_1} \sum_{i=1}^{n} e_i^2 \tag{4.8}$$

Substituting the value for e_i (equation 4.6) into equation 4.8:

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

and substituting in the value for $\hat{Y}_i$ (from equation 4.5) we get:

$$\min_{b_0, b_1} \sum_{i=1}^{n} \left(Y_i - b_0 - b_1 X_i\right)^2 \tag{4.9}$$

To solve this minimization problem, we take the partial derivatives of $\sum_{i=1}^{n} e_i^2$ with respect to b_0 and b_1, set those derivatives equal to zero, and solve for b_0 and b_1. That is, we need to solve the two equations:

$$\frac{\partial \left(\sum_{i=1}^{n} e_i^2\right)}{\partial b_0} = 0$$

$$\frac{\partial \left(\sum_{i=1}^{n} e_i^2\right)}{\partial b_1} = 0$$

We leave the derivation for an exercise, and only write the solution here:

$$b_1 = \frac{\sum_{i=1}^{n} \left[\left(Y_i - \bar{Y}\right)\left(X_i - \bar{X}\right)\right]}{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2} \tag{4.10}$$

$$b_0 = \bar{Y} - b_1 \bar{X}$$

These equations tell us how to pick a line (by picking an intercept and slope) in order to minimize the sum of squared vertical distances between the chosen line and each data point. The next question is, why should we choose a line in such a way?

4.6 The Assumptions and Properties of OLS

So, what's so great about OLS? There are many other ways that we could fit a line through some data points:

- instead of *vertical* distances, we could minimize the sum of *horizontal* or *orthogonal* distances
- instead of taking the sum of *squared* distances, we could take the sum of *absolute* distances
- we could divide the sample into two parts, get the average Y and X coordinates, and connect the dots
- we could pick (randomly or not) any two different data points and connect them

The main point here is that there are many ways that we could fit a line, so we should wonder why OLS is so special. Some of these alternatives above are obviously silly, but some lead to alternative estimators that have merit in various situations.

Recall that estimators are random variables (see Chapter 3). The OLS slope and intercept estimators have sampling distributions, with a mean and a variance. The reason why we use OLS is because these random estimators have good statistical properties (under certain assumptions). Here, we list the assumptions, and return to them at various stages throughout the book.

4.6.1 The OLS assumptions

A1 The population model is linear in the βs.
A2 There is no perfect multicollinearity between the X variables.
A3 The random error term, ϵ, has mean zero.
A4 ϵ is identically and independently distributed.
A5 ϵ and X are independent.
A6 ϵ is Normally distributed.

4.6.2 The properties of OLS

Provided that the above six assumptions hold:

- The OLS estimator is unbiased.

- The OLS estimator is efficient.
- The OLS estimator is consistent.
- The OLS estimator is Normally distributed.

Note that not all assumptions are needed for each of the above four properties. Additionally, some of the assumptions A1 - A6 are often unrealistic. Testing for the validity of these assumptions, re-evaluating the properties of the OLS estimator in the absence of each assumption, and figuring out how to recover unbiasedness, efficiencyand consistency, would lead to some different estimators, and would form the basis for future econometrics courses.

4.7 Review Questions

1. Let the sample data be $Y = \{5, 2, 2, 3\}$ and $X = \{5, 3, 5, 3\}$.
   a) Write down the population model.
   b) Calculate the OLS estimated slope and intercept, using equation 4.10.
   c) Interpret these estimates.
   d) Calculate the OLS predicted values and residuals.
   e) Using R, verify your answer in part (b).
2. How are the formulas for b_1 and b_0 derived?
3. Explain why, even if assumption A.6 does not hold, the OLS estimator may still be normally distributed.
4. Why is the ϵ term needed in equation 4.4?
5. Download the MPC data using:

   ```
   mpcdata <- read.csv("https://rtgodwin.com/data/mpc.csv")
   ```

   Use R to aid in the following exercises.
   a) Write down the population model you are trying to estimate. Describe the components of this model.
   b) Plot the data.
   c) Calculate the OLS estimated slope and intercept.
   d) Interpret these estimates.
   e) Add the estimated regression line to the plot of the data.

4.8 Answers

1. a) The assumed population model is $Y_i = \beta_0 + \beta_1 + \epsilon$. It is assumed that the X variable "causes" the Y variable. The Y and X data has been given to us. β_0 and β_1 are unknown parameters to be estimated. ϵ represents all the other factors (or variables) that cause Y but that are unobserved.

   b)

$$\begin{aligned}
&\bar{Y} = 3, \quad \bar{X} = 4 \\
&b_1 = \frac{(5-3)(5-4) + (2-3)(3-4) + (2-3)(5-4) + (3-3)(3-4)}{(5-4)^2 + (3-4)^2 + (5-4)^2 + (3-4)^2} \\
&\quad = 0.5 \\
&b_0 = 3 - 0.5 \times 4 = 1
\end{aligned}$$

   c) b_1 is the estimated slope, or marginal effect. Numerically, the values $b_1 = 0.5$ means that it is estimated that when X increases by 1, Y will increase by 0.5. b_0 is the estimated intercept. Numerically, when X is 0, it is estimated that Y is 1.

d)

$$\bar{Y}_1 = 1 + 0.5(5) = 3.5$$
$$\bar{Y}_1 = 1 + 0.5(3) = 2.5$$
$$\bar{Y}_1 = 1 + 0.5(5) = 3.5$$
$$\bar{Y}_1 = 1 + 0.5(3) = 2.5$$
$$e_1 = 5 - 3.5 = 1.5$$
$$e_2 = 2 - 2.5 = -0.5$$
$$e_3 = 2 - 3.5 = -1.5$$
$$e_4 = 3 - 2.5 = -0.5$$

e) In R, enter the following three commands:

```
y <- c(5,2,2,3)
x <- c(5,3,5,3)
lm(y ~ x)
```

and you should see the following output:

```
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
        1.0          0.5
```

2. The formulas for the OLS estimator are derived by minimizing the sum of squared OLS residuals. This involves solving an optimization problem in calculus. The derivatives of the sum of squared residuals, with respect to b_0 and b_1, are set equal to 0 and solved, providing the formulas in equation 4.10.
3. If assumption A.6 holds, then the OLS estimators will be Normally distributed. This is because, by the population model (equation 4.4), Y is a linear function of ϵ, hence Y is also Normally distributed. Furthermore, because b_1 and b_0 are linear functions of Y, they are also Normally distributed.
   However, even without A.6, the OLS estimator may still be Normally distributed. This is again due to the central limit theorem. Look again at the formula for the OLS estimator (equation 4.10) and note the summation sign. Since the OLS estimator involves summing the random variable Y, as long as the sample size is large enough, the resulting sum should be Normally distributed.
4. The error term is needed in order to represent all of the other factors that influence Y, besides the X variable. Since these other factors (or variables) are unobserved, we consider them to be random, and add them all up into one term. ϵ represents the randomness in the population model, without which there would be no need for statistics or econometrics.
5. a) The population model that we are trying to estimate is the consumption model from equation 4.3: $C = \beta_0 + \beta_1 \times Y + \epsilon$, where C is the independent variable (the "Y" variable), β_1 is the MPC, Y is the independent variable (the "X" variable), ϵ represents all the other variables that determine C, and where β_0 doesn't have much economic interest.
   b) First, you must load the data into R using the following command:

```
mpcdata <- read.csv("https://rtgodwin.com/data/mpc.csv")
```

Once the data has been loaded, enter the following command in order to plot the data:

```
plot(mpcdata$income, mpcdata$consumption, main="Consumption and
     Income in the U.K.")
```

c) In order to calculate the OLS estimates for the intercept and slope, run the following command in R:

```
lm(consumption ~ income, data=mpcdata)
```

d) The estimated slope on income is the estimated marginal propensity to consume. That is, when *income* increases by 1, it is estimated that *consumption* will increase by 0.869. The estimated intercept of 176.848 is the amount of consumption when income (or GDP) is zero, and since GDP is never zero, the intercept doesn't hold much economic interest.

e) In order to add the estimated regression line to your plot of data, use the following command (choose your own colour!):

```
abline(lm(consumption ~ income, data=mpcdata), col = "red")
```

Chapter 5

OLS Continued

In this chapter, we discuss three extensions of OLS. First, we introduce the regression R-square, which is a way to evaluate how well the estimated OLS regression line fits the data. Second, we discuss how to test a null hypothesis involving the βs (usually β_1). Third, we discuss the use of dummy variables in econometric models.

5.1 R-squared

R-squared is a "measure of fit" of the regression line. It is a number between 0 and 1 (as long as the model contains an intercept) that indicates how close the data points are to the estimated line. More accurately, the regression R-squared (R^2) is the portion of variance in the Y variable that can be explained by variation in the X variable.

Look again at the assumed population model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The assumption is that changes in X lead to changes in Y. We are using the observed changes in both variables to choose the regression line (via OLS). But, changes in X aren't the only reason that Y changes. There are unobservable variables in the error term (ϵ) that lead to changes in Y. How much of the changes in Y are coming from X (not ϵ)? R^2 helps answers this question.

The R^2 can also be thought of as an overall measure of how well the model explains the Y variable. That is, we are using information in X to explain or *predict* Y by estimating a model. How well does the estimated regression line "fit" the data? How well does the model explain the Y variable? R^2 provides a measure to address these questions. Let's reiterate the interpretations of R^2 before we derive it. R^2 measures:

- how well the estimated model explains the Y variable.
- how well changes in X explain changes in Y.
- how well the estimated regression line "fits" the data.
- the portion of the variance in Y that can be explained using the estimated model.

Figure 5.1 shows the estimated OLS regression line fitted to both the demand for spirits and demand for cigarettes data. The estimated regression line seems to fit the data better, or explain more of the variation in Q, for spirits rather than for cigarettes. We will find that the R^2 is indeed higher for the spirits data. In some sense, the R^2 can be used to compare OLS regressions.

Figure 5.2 shows a hypothetical situation where, if all data moves vertically further away from the estimated regression line, the regression line stays the same, but the R^2 decreases. That is, both the red (triangles) and blue (circles) provide the same estimated b_1, but the line fits the red data better. Changes in X account for more of the changes in Y for the red data. For the blue data, the *unobserved factors* (in ϵ) are accounting for more of the changes (or variation) in Y.

Figure 5.1: Which estimated regression line fits better? Demand for spirits (left) and demand for cigarettes (right). We might expect the regression on the left to have a higher R^2.

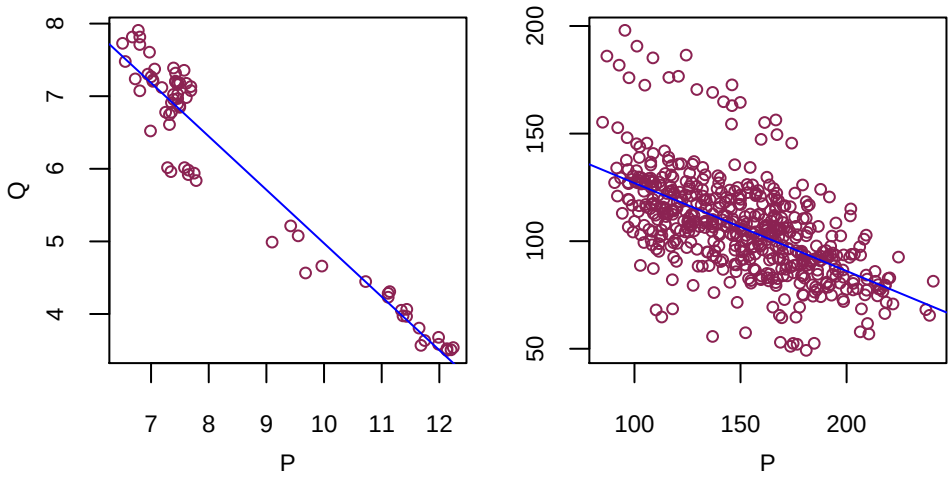
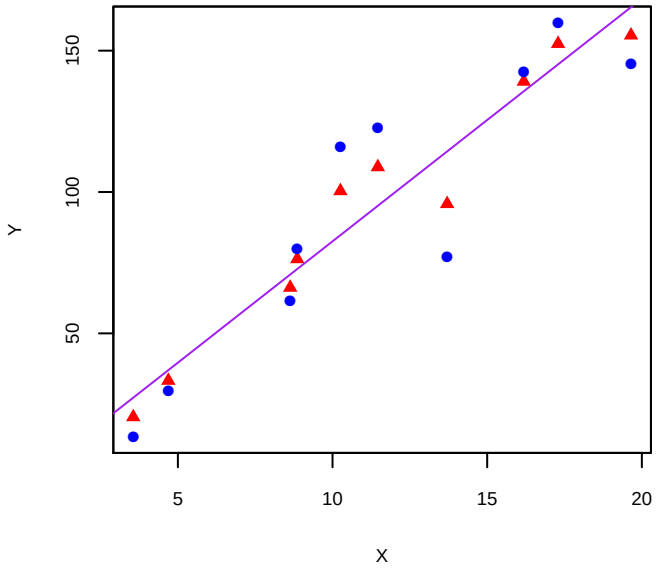


Figure 5.2: Two different data sets. The estimated regression line for both data sets is the same. The blue data points (circles) are twice as far (vertically) from the regression line as are the red data points (triangles). For red data, $R^2 = 0.95$. For blue data, $R^2 = 0.82$.



5.1.1 The R^2 formula

Now, we will derive the R^2 statistic, beginning with the definition: "R-squared is the portion of variance in Y that can be explained using the estimated model." The population model is (equation 4.4):

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

The estimated model is (equation 4.7):

$$Y_i = b_0 + b_1 X_i + e_i$$

Recall that the OLS predicted value is (equation 4.5):

$$\hat{Y}_i = b_0 + b_1 X_i$$

So:

$$Y_i = \hat{Y}_i + e_i \tag{5.1}$$

Equation 5.1 shows that each Y_i value has two parts: a part that can be explained by OLS ($\hat{Y}_i$), and a part that cannot (e_i). To get R^2, we'll start by taking the sample variance of both sides of equation 5.1. This will break the variance in Y up into two parts: variance the we *can* explain (variance in $\hat{Y}_i$), and variance that we *can't* explain (variance in e_i).

Recall that in Chapter 3, when we wanted to estimate the variance of y, we used equation 3.17, which is the sample variance:

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

Taking the sample variance of both sides of equation 5.1 we get (there is no sample covariance because $\hat{Y}_i$ and e_i are independent):

$$s_Y^2 = s_{\hat{Y}}^2 + s_e^2$$

Or:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{\hat{Y}}\right)^2 + \frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2 \tag{5.2}$$

To simplify equation 5.2, we'll make use of three algebraic properties:

- the $(n-1)$ cancel out
- $\bar{\hat{Y}} = \bar{Y}$
- $\bar{e} = 0$

Using these three properties, equation 5.2 becomes:

$$\sum\left(Y_i - \bar{Y}\right)^2 = \sum\left(\hat{Y}_i - \bar{Y}\right)^2 + \sum(e_i)^2 \tag{5.3}$$

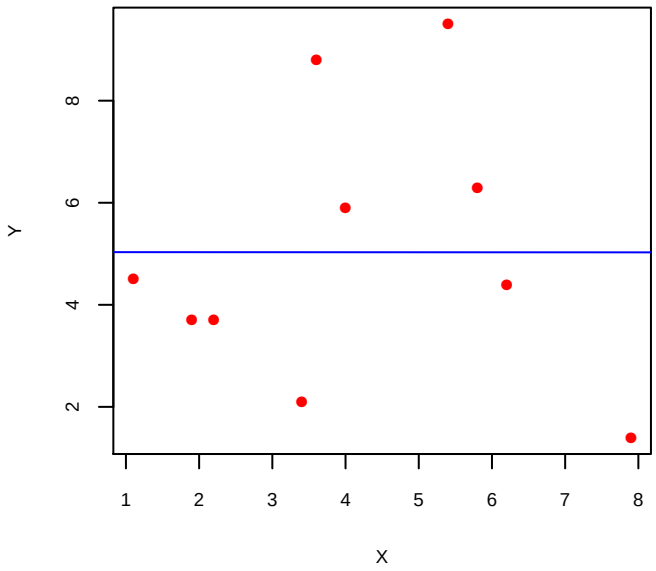
Notice that the terms in equation 5.3 are "sums of squares", and equation 5.3 is often written as:

$$TSS = ESS + RSS \tag{5.4}$$

where:

- TSS - total sum of squares

Figure 5.3: The estimated regression line is essentially flat: $b_1 = 0$. Observed changes in X are not at all helpful in predicting changes in Y. There is "no fit", and $R^2 = 0.00$.



- ESS - explained sum of squares
- RSS - residual sum of squares

Now, we return to our definition of R^2: "the portion of variance in Y that can be explained using the estimated model." This portion is written as:

$$R^2 = \frac{ESS}{TSS} \tag{5.5}$$

We can also re-write the formula for R^2 using equation 5.4:

$$R^2 = 1 - \frac{RSS}{TSS} \tag{5.6}$$

5.1.2 "No fit" and "perfect fit"

What is the worst possible situation, in terms of the "fit" of the estimated regression line? If the X variable cannot explain any of the changes/variation in the Y variable, then the estimated model (the estimated regression line) will be useless.

If the X observations are not useful in explaining changes in the Y observations (that is, if the sample X and Y data are *independent*), then $b_1 = 0$. In this case, we have a situation of "no fit", where $R^2 = 0$. See figure 5.3.

To see algebraically why $R^2 = 0$ when $b_1 = 0$, we start by looking at equation 4.5 again:

$$\hat{Y}_i = b_0 + b_1 X_i$$

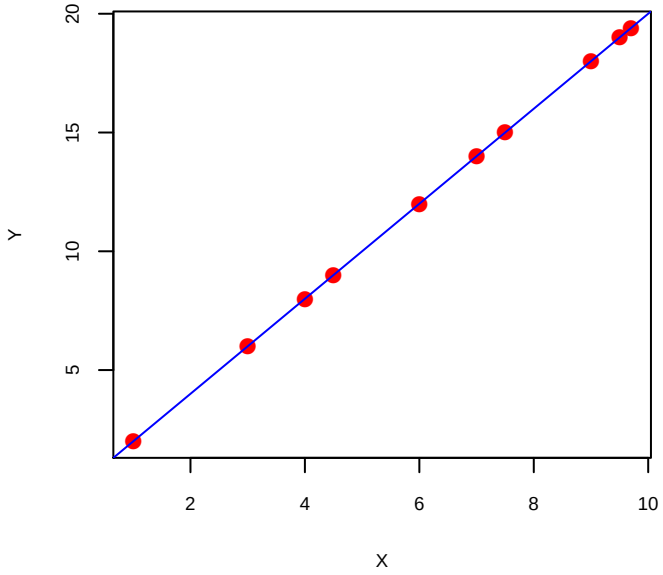
So, if $b_1 = 0$ then each predicted $\hat{Y}_i$ value is equal to just b_0 (all the predicted values are the same). Additionally, when $b_1 = 0$, by looking at the equation for the OLS intercept estimator, we see that:

$$b_0 = \bar{Y} - b_1 X_i = \bar{Y}$$

This mean that, if $b_1 = 0$, each predicted value is equal to the sample average of Y: $\hat{Y}_i = \bar{Y}$. Hence, $ESS = 0$:

$$ESS = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 = \sum \left(\bar{Y} - \bar{Y}\right)^2 = 0,$$

Figure 5.4: The estimated regression line exactly passes through each data point. Observed changes in X perfectly predict changes in Y. There is "perfect fit", and $R^2 = 1$.



and $R^2 = 0$.

Now, let's consider the opposite extreme: a situation where we have a "perfect fit". Imagine that observed changes in X could perfectly predict a change in Y. That is, if we knew the value of X, we would exactly know the value of Y with certainty. What would our sample of data have to look like in order for this to be the case? See figure 5.4.

In order for the estimated regression line to fit the data perfectly, all of the observed data points must line up in a straight line. If this were so, the estimated line would pass through each data point, the OLS predicted values ($\hat{Y}_i$) would be exactly equal to the actual values (Y_i), and there would be no prediction error ($e_i = 0 \; \forall \; i$). Algebraically, $\hat{Y}_i = Y_i$, so that $ESS = TSS$, and $R^2 = 1$.

The two cases that we have just considered, "no fit" and "perfect fit", are extremes. They should not actually occur in practice. In reality, the fit of the line will be somewhere between these two extremes. If the worst that can happen is "no fit" and the best is " perfect fit", then $0 \leq R^2 \leq 1$.

5.2 Hypothesis testing

We'll begin this section by looking at the variance of the OLS slope estimator ($\text{Var}[b_1]$). There are three reasons to get this formula:

1. Looking at it will provide insight into what determines the accuracy (a smaller variance) of the estimator.
2. It is required to prove that OLS is an efficient estimator, and therefore is BLUE.
3. It is needed for hypothesis testing.

5.2.1 The variance of b_1

In chapter 3, we derived the variance of the estimator, $\bar{y}$. Similarly, b_1 is a random variable, since it is obtained from a formula involving the random sample $\{Y_i, X_i\}$, and it is common to consider the variance of a random variable. However, deriving the variance of the OLS estimator is too difficult for this course, and we simply write the result:

$$\text{Var}[b_1] = \frac{\sigma_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}, \tag{5.7}$$

where σ_ϵ^2 is the variance of the error term ϵ, n is the sample size, and in the denominator we see something that looks like the sample variance of X_i. From equation 5.7, it can be seen that:

- $\text{Var}[b_1]$ decreases as n increases.
- $\text{Var}[b_1]$ decreases as the sample variation in X increases.
- $\text{Var}[b_1]$ decreases as variation in ϵ decreases.

We want our estimator to have as low a variance as possible! A lower variance means that, on average, we have a higher probability of being close to the "rights answer" (provided the estimator is unbiased). These factors that lead to a lower $\text{Var}[b_1]$ make sense:

- If we have more information (larger n), it should be "easier" to pick the right regression line.
- Since we are using changes in X to try to explain changes in Y, the bigger changes in X that we observe, the easier it is to pick the regression line.
- The less unobservable changes there are (in ϵ that are causing changes in Y, the easier it is to pick the regression line.

We could discuss a similar formula for $\text{Var}[b_0]$ as well, however, there is rarely any economic interest in the model's intercept that we omit the discussion.

A final note. $\text{Var}[b_1]$ is required in order to prove that OLS is *efficient* (the Gauss-Markov theorem). Proving that an estimator is efficient requires that its variance is shown to be the smallest among all other possible candidate estimators (in the Gauss-Markov theorem other candidate estimators are linear and unbiased ones). The Gauss-Markov theorem is very important because it provides the reason for why OLS should be used: provided (some of) assumptions A1-A6 hold, OLS is the best linear unbiased estimator (BLUE) possible for estimating β_1.

5.2.2 Test statistics and confidence intervals

Hypothesis testing in the context of OLS usually involves β_1. That is, usually we want to test if a marginal effect is equal to some value. For example, do similarly qualified women earn less than men? Are the returns to education the same for men and women? If we raise the taxes on cigarettes, will consumption decrease? These are all questions that can be answered by forming a null and alternative hypothesis, collecting data, estimating, and rejecting or failing to reject the null. In the context of OLS, a two-sided null and alternative hypothesis looks like:

$$H_0 : \beta_1 = \beta_{1,0}$$
$$H_A : \beta_1 \neq \beta_{1,0}$$

A common hypothesis in economics is where the marginal effect is zero (X does not cause Y), so that the above null and alternative become:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

As in chapter 3, we will begin with the z-test. In general, the z-statistic is determined by:

$$z\text{−statistic} = \frac{\text{estimate} - \text{value of } H_0}{\sqrt{\text{Var}[\text{estimator}]}} \tag{5.8}$$

This z-statistic is Normally distributed with mean 0 and variance 1 ($z \sim N(0,1)$), if H_0 is true and $\bar{Y}$ is Normal. In chapter 3, when our test involved the population mean, equation 5.8 became:

$$z = \frac{\bar{y} - \mu_{Y,0}}{\sqrt{\sigma_Y^2/n}}$$

In OLS, when we are testing the slope (marginal effect) of the model, equation 5.8 becomes:

$$z = \frac{b_1 - \beta_{1,0}}{\sqrt{\text{Var}\,[b_1]}},$$

where b_1 is the estimate that we actually get from the sample, $\beta_{1,0}$ is the hypothesized value of the slope, and $\text{Var}\,[b_1]$ is given by equation 5.7.

As was the case in chapter 3, however, it is not realistic that we would know the variance of b_1. By looking again at equation 5.7, we see that the unknown part is the variance of the error term, σ_ϵ^2. If we could estimate σ_ϵ^2, we would have an estimate for the variance of b_1, and we could use a t-test instead of a z-test.

Recall that the population model is:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

and that the estimated model is:

$$Y_i = b_0 + b_1 X_i + e_i$$

Each unobservable part in the population model (β_0, β_1, ϵ_i) has an observable counter-part in the estimated model. So, if we want to know something about ϵ we can use e. In fact, an estimator for the variance of ϵ is the *sample variance* of the OLS residuals:

$$s_\epsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 \tag{5.9}$$

Why is the -2 in the denominator of equation 5.9? Recall that, in chapter 3, when we wanted to estimate σ_y^2 we used the sample variance of y:

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

and that the -1 in the denominator was a degrees-of-freedom correction, so that the estimator is unbiased. We only had $(n-1)$ pieces of information available to estimate σ_y^2, after we had used up a piece of information to get $\bar{y}$. The story is similar in equation 5.9. In order to get the OLS residuals, we first have to estimate *two* things (b_0 and b_1):

$$e_i = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$$

This uses up two pieces of information, leaving $(n-2)$ remaining when we are using the e_i. Now that we have an estimator for σ_ϵ^2, we have an estimator for $\text{Var}\,[b_1]$ (we just replace the unknown σ_ϵ^2 with s_ϵ^2):

$$\hat{\text{Var}}\,[b_1] = \frac{s_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

And now, the t-statistic for testing β_1 is obtained by substituting $\hat{\text{Var}}\,[b_1]$ for $\text{Var}\,[b_1]$ in the z-statistic formula:

$$t = \frac{b_1 - \beta_{1,0}}{\sqrt{\hat{\text{Var}}\,[b_1]}} \tag{5.10}$$

The denominator of 5.10 is often called the *standard error* of b_1 (like a standard deviation), and equation 5.10 is often written instead as:

$$t = \frac{b_1 - \beta_{1,0}}{\text{s.e.}[b_1]} \tag{5.11}$$

where s.e. $[b_1]$ stands for the estimated standard error of b_1.

If the null hypothesis is true, the t-statistic in equation 5.11 follows a t-distribution with degrees of freedom $(n-k)$, where k is the number of βs we have estimated (two). To obtain a p-value we should use the t-distribution, however, if n is large, then the t-statistic follows the standard Normal distribution. For the purposes of this course, we shall always assume that n is large enough such that $t \sim N(0,1)$. To obtain a p-value, we can use the same table that we used at the end of chapter 3 (see Table 3.2).

5.2.3 Confidence intervals

Confidence intervals are obtained very similarly to how they were in chapter 3. The 95% confidence interval for b_1 is:

$$b_1 \pm 1.96 \times \text{s.e.}[b_1] \tag{5.12}$$

The 95% confidence interval can be interpreted as follows: (i) if we were to construct many such intervals (hypothetically), 95% of them would contain the true value of β_1; (ii) all of the values that we could choose for $\beta_{1,0}$ that we would fail to reject at the 5% significance level.

We can get the 90% confidence interval by changing the 1.96 in equation 5.12 to 1.65, and the 99% C.I. by changing it to 2.58, for example.

5.3 Dummy Variables

A *dummy variable* is a variable that takes on one of two values (usually 0 or 1). A dummy variable is also sometimes called a *binary variable* or a *dichotomous variable*. We will consider that the independent variable (the regressor or "X" variable) in our population model (equation 4.4) is a dummy variable, where:

$$D_i = \begin{cases} 0, & \text{if individual } i \text{ belongs to group } A \\ 1, & \text{if individual } i \text{ belongs to group } B \end{cases}$$

Dummy variables are useful for estimating differences between groups, where groups "A" and "B" can take on many definitions. For example, in labour economics and many other areas of economics, it is common to use a dummy variable to identify the *gender* of the individual.

5.3.1 A population model with a dummy variable

Now, let's consider a population model with a dummy:

$$Y_i = \beta_0 + \beta_1 D_i + \epsilon_i, \tag{5.13}$$

where $D_i = 0$ if the individual is female, $D_i = 1$ if the individual is male, and Y_i is the wage of the individual. How do we interpret β_1 from equation 5.13? Since D_i is not a continuous variable, β_1 is not a marginal effect, and we cannot take the derivative of Y with respect to D when D is non-continuous. Instead, let's use *conditional expectations* to find the interpretation of β_1.

Let's consider the expected wage of a male worker:

$$\text{E}[Y_i | D_i = 1] = \beta_0 + \beta_1(1) + \text{E}[\epsilon_i] = \beta_0 + \beta_1 \tag{5.14}$$

We have simply substituted in the population model (equation 5.13) for Y_i, substituted in $D_i = 1$, and made use of assumption A.3 ($\mathrm{E}[\epsilon_i] = 0$). Now, let's consider the expected wage of a female worker:

$$\mathrm{E}[Y_i|D_i = 0] = \beta_0 + \beta_1(0) + \mathrm{E}[\epsilon_i] = \beta_0 \tag{5.15}$$

What is the difference between these two conditional expectations (equations 5.14 and 5.15)? β_1! That is:

$$\mathrm{E}[Y_i|D_i = 1] - \mathrm{E}[Y_i|D_i = 0] = \beta_1 \tag{5.16}$$

So, when the "X" variable is a dummy variable, the attached β is interpreted as the difference in population means between the two groups.

5.3.2 An estimated model with a dummy variable

OLS works just fine when the right-hand-side variable is a dummy variable. The estimated model will be the same as it was before:

$$Y_i = b_0 + b_1 D_i + e_i, \tag{5.17}$$

where everything has the same interpretation as before, except that b_1 is the *estimated* difference in population mean of Y between the two groups as defined by the dummy variable. In fact, it turns out that:

- b_0 is the *sample* mean ($\bar{Y}$) for $D_i = 0$
- $b_0 + b_1$ is the *sample* mean for $D_i = 1$
- b_1 is the difference in sample means (be careful of the sign)

This means that, instead of using OLS, we could just divide the sample into two parts (using D_i), and calculate two sample averages! So why should we use OLS? At this stage, it looks like we are making things more complicated than they need to be. However, in the next chapter, we will add more X variables, so that we will not be able to get the same results by dividing the sample into two.

5.3.3 Example: Gender and wages using the CPS

The current population survey (CPS) is a monthly detailed survey conducted in the United States. It contains information on many labour market and demographic characteristics. In this section, we will use a subset of data from the 1985 CPS, to estimate the differences in wages between men and women. To load the data, use:

```
cps <- read.csv("https://rtgodwin.com/data/cps1985.csv")
```

You will see many variables in the dataset. For now, we look at only a few:

- wage - hourly wage
- education - number of years of education
- gender - dummy variable for gender

To run an OLS regression of `wage` on `gender`, use the following command:

```
summary(lm(wage ~ gender, data = cps))
```

You should see the following output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   7.8789     0.3216   24.50  < 2e-16 ***
gendermale    2.1161     0.4372    4.84  1.7e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.034 on 532 degrees of freedom
Multiple R-squared:  0.04218,   Adjusted R-squared:  0.04038
F-statistic: 23.43 on 1 and 532 DF,  p-value: 1.703e-06
```

From this output, you should be able to answer the following questions:

- What is the sample mean wage for men and for women?
- What is the interpretation of b_1?

We stated earlier that the results we obtain from regressing on a dummy variable are equivalent to what we would obtain by dividing the sample into two parts (by gender). Let's verify this using the CPS data. In R, create subsets for men and women:

```
cps.m <- subset(cps, gender == "male")
cps.f <- subset(cps, gender == "female")
```

then take the difference in the sample mean wage between men and women:

```
mean(cps.m$wage) - mean(cps.f$wage)
```

```
[1] 2.116056
```

The difference is equal to b_1, which is 2.1161! Also, note that the sample mean wage for women is b_0:

```
mean(cps.f$wage)
```

```
[1] 7.878857
```

and the sample mean wage for men is $b_0 + b_1$:

```
mean(cps.m$wage)
```

```
[1] 9.994913
```

So, OLS is not needed in this case! All the estimates can be obtained by dividing the sample in two, and taking sample averages. However, as soon as we have more than one "X" variable in the model, we can no longer obtain the OLS estimates by simply dividing the sample into two.

5.4 Reporting regression results

We end this chapter with a concise and conventional way of reporting regression results. If you were to see the results of an OLS regression in an economics paper or report, you would not see the ugly R output above. If there are many variables in the regression (see the next chapter), the results may be displayed in a table. However, if there are only a few variables in the regression, it is convenient to report results in an equation with two lines.

For example, when we regress `wage` on `gender`:

```
summary(lm(wage ~ gender), data = cps)
```

we could report the regression results as follows:

$$\begin{aligned} \hat{wage} &= 7.88 \ + \ 2.12 \times gendermale, \ R^2 = 0.042 \\ &\quad (0.32) \quad (0.44) \end{aligned} \tag{5.18}$$

Equation 5.18 conveys the estimated βs, as well as the estimated standard errors, and the R^2. Verify that you know where all of these numbers are coming from in the R output.

5.5 Review Questions

1. Derive the following expression for R^2:

   $$R^2 = \frac{ESS}{TSS},$$

   and show that R^2 can be rewritten as:

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

2. Using diagrams, explain why $0 \leq R^2 \leq 1$.
3. Using equation 5.7, explain why having a larger sample is better.
4. Explain what s.e.$[b_1]$ is.
5. Using equation 5.13, explain how to interpret β_0 and β_1.
6. The following question refers to the regression of `wage` on `gender` using the CPS data. The estimated results, equation 5.18, are repeated here:

   $$\begin{aligned} \hat{wage} &= 10.00 - 2.12 \times gender, \ R^2 = 0.042 \\ &\quad (0.30) \quad (0.44) \end{aligned}$$

   a) What is the estimated wage-gender gap?
   b) What is the sample mean wage for males and for females?
   c) Test the hypothesis that there is no wage-gender gap.
   d) Construct a 90% confidence interval for the wage-gender gap.
   e) Interpret the value for R^2.
   f) Another researcher uses the same data, but defines the dummy variable in the *opposite* way. What will be the estimated values for b_0 and b_1?

7. This question uses the CPS data set, which can be loaded into R using the following commands:

   ```
   install.packages("AER")
   library(AER)
   data("CPS1985")
   attach(CPS1985)
   ```

   a) Estimate the returns (in hourly wages) of an additional year of education. Summarize your results concisely in an equation.
   b) Test the hypothesis that the returns to education are zero.
   c) Construct a 95% confidence interval for the returns to education.
   d) Interpret the value of R^2.
   e) What does the estimated model predict the hourly wages will be for high school graduates and for university graduates?
   f) What is the estimated value, in terms of hourly wage, of obtaining an undergraduate degree?

5.6 Answers

1. A definition for R^2, in words, is: the portion of variance in Y that can be explained by the estimated model. Each Y observation can be written as a sum of two parts (a part that can be explained using the X variable, and the left over unexplainable part):

$$Y_i = \hat{Y}_i + e_i$$

Taking the sample variance of both sides we get:

$$\hat{\text{var}}[Y_i] = \hat{\text{var}}[\hat{Y}_i] + \hat{\text{var}}[e_i]$$

Note that there is no sample covariance between $\hat{Y}$ and e because they are *independent*. Using the formula for sample variance (from chapter 3, equation 3.17) into the above equation, we get:

$$\frac{\sum(Y_i - \bar{Y})^2}{n-1} = \frac{\sum(\hat{Y}_i - \bar{\hat{Y}})^2}{n-1} + \frac{\sum(e_i - \bar{e})^2}{n-1} \tag{5.19}$$

Now, we make three simplifications to the above:

- the $(n-1)$ cancel
- $\bar{\hat{Y}} = \bar{Y}$ (the sample mean of the OLS predicted values equals the sample mean of the actual values)
- $\bar{e} = 0$ (the OLS residuals sum to 0)

Equation 5.19 becomes:

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y}_i)^2 + \sum e_i^2$$

The terms in the above equation are "sums-of-squares", so that:

$$TSS = ESS + RSS \tag{5.20}$$

Where TSS is the total sum-of-squares (from the total sample variance of Y), ESS is the explained sum-of-squares (from the sample variance of the OLS predicted values), and RSS is the residual sum-of-squares (from the sample variance of the OLS residuals).
Returning to our original definition of R^2: "the portion of variance in Y that can be explained by the estimated model", we get:

$$R^2 = \frac{ESS}{TSS}. \tag{5.21}$$

To get an alternate equation, we solve 5.20 for ESS:

$$ESS = TSS - RSS$$

and substitute into R^2:

$$R^2 = \frac{ESS}{TSS} = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{5.22}$$

2. This question is answered by considering two extreme cases: (i) the X variable has no explanatory power, and (ii) the X variable can perfectly explain Y. (i) is a situation of "no fit", drawn in figure 5.3, and would occur if $b_1 = 0$. In this situation, each OLS predicted value will be equal to $\bar{Y}$, so ESS will equal 0, and so R^2 will also equal 0. (ii) is a situation of "perfect fit", drawn in figure 5.4. All data points are on the estimated regression line. $ESS = TSS$, $RSS = 0$, and so $R^2 = 1$.

3. Using equation 5.7, we just need to see that as n increases, the variance of the OLS estimator decreases.
4. In order to perform hypothesis testing, an estimate for the variance of the OLS estimator is required. If equation 5.7 is to be used in practice, we must replace the unknown σ_ϵ^2 with the estimator $s_e^2 psilon = \sum e_i^2/n-2$. When we take the square-root of this quantity, it is called the *standard error* of b_1 (or $s.e.[b_1]$ for short). That is,

$$s.e.[b_1] = \sqrt{\frac{s_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}}$$

5. The interpretation of β_1, when the independent variable is a dummy variable, is obtained by taking the conditional expectation of Y for each of the two possible values that the dummy variable can take. We repeat equation 5.16:

$$\mathrm{E}\left[Y_i|D_i = 1\right] - \mathrm{E}\left[Y_i|D_i = 0\right] = \beta_1$$

6. a) The estimated wage-gender gap is the coefficient in front of the `gender` dummy variable (where it is understood that `gender` = 1 if the worker is female). So, the estimated wage-gender gap is -2.12, meaning that on average, women earn \$2.12 less than men, according to this sample data.
   b) The sample mean wage for mean is $b_0 = 10.00$, and for women is $b_0 + b_1 = 10.00 - 2.12 = 7.78$.
   c) The null hypothesis is that the differences in wages between men and women is zero. In terms of the population model, this would mean that $\beta_1 = 0$.

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

   The t-test statistic for this null hypothesis is:

$$t = \frac{b_1 - \beta_{1,0}}{\text{s.e.}\left[b_1\right]} = \frac{-2.12 - 0}{0.44} = -4.82$$

   The associated p-value is 0.00. We reject the null hypothesis. The estimated wage-gender gap is statistically significant.
   d) The 90% confidence interval for the wage-gender gap is:

$$-2.12 \ \pm \ 1.65 \times 0.44 = (-2.85, -1.39)$$

   e) Gender explains 4.2% of the variation in wages.
   f) $b_0 = 7.78$ and $b_1 = 2.12$.
7. a) Use the following command:

```
summary(lm(wage ~ education))
```

   and you should see the following output:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.74598    1.04545  -0.714    0.476
education    0.75046    0.07873   9.532   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.754 on 532 degrees of freedom
Multiple R-squared:  0.1459,    Adjusted R-squared:  0.1443
```

```
F-statistic: 90.85 on 1 and 532 DF,  p-value: < 2.2e-16
```

Some of this information is summarized as follows:

$$\hat{wage} = -0.75 + 0.75 \times education, \; R^2 = 0.146$$
$$(1.05) \quad (0.08)$$

The estimated returns to education are \$0.75 in hourly wages per year of education.

b) From the R output we can see that the `education` variable is highly statistically significant. The p-value for the test is 0 (to sixteen decimal places).

c) The 95% confidence interval is:

$$0.75 \; \pm \; 1.96 \times 0.079 = (0.60, 0.91)$$

d) Years of education can explain 14.6% of the differences in wages.

e) Assuming that a high school graduate has 12 years of education, the predicted wage is:

$$\hat{wage} = -0.75 + 0.75(12) = 8.25$$

and assuming that university graduates have 16 years of education the predicted wage is:

$$\hat{wage} = -0.75 + 0.75(16) = 11.25$$

f) The predicted difference in wages between university and high school graduates is \$11.25 - \$8.25 = \$3.

Chapter 6

Multiple Regression

Multiple regression refers to having more than one "X" variable (more than one regressor). From now on, we will typically be dealing with population models of the form:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i \tag{6.1}$$

where k is the number of regressors in the model, and the total number of βs to be estimated is $(k+1)$. This new model allows for Y to be explained used *multiple* variables. That is, there can now be many Xs that are causal determinants of Y.

6.1 House prices

Should I build a fireplace in my home before I sell it? To motivate the need for a multiple regression model, we begin with an example. Let's try to determine the value of a fireplace using data on house prices. The data are from the New York area, 2002-2003, and are from Richard De Veaux of Williams College.

To load the data into R, use the following two commands:

```
house <- read.csv("https://rtgodwin.com/data/houseprice.csv")
```

The variables in the dataset are shown in Table 6.1.

Table 6.1: Description of the variables in the house price data set.

Price	the price of the house in dollars
Lot.Size	the size of the property in acres
Waterfront	dummy variable equal to 1 if house is on the water
Age	number of years since the house was built
Central.Air	dummy variable equal to 1 if house has air conditioning
Living.Area	the size of the house in square feet
Bedrooms	number of bedrooms
Fireplaces	number of fireplaces
Bathrooms	number of bathrooms (half-bathrooms are 0.5)
Rooms	total number of rooms in the house

We are interested in the effect of the variable `Fireplaces` on `Price`. Let's get some summary statistics for `Fireplaces`:

```
summary(house$Fireplaces)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.6019  1.0000  4.0000
```

The houses in the sample have anywhere from 0 to 4 fireplaces, with the average being 0.6.
Let's now turn our attention to the `Price` variable. For easier interpretation, we'll changethe units of `Price` form dollars to thousands of dollars:

```
house$Price <- house$Price / 1000
```

Next, let's see the sample mean price, conditional on the number of fireplaces:

```
mean(house$Price[house$Fireplaces == 0])
```

```
[1] 174.6533
```

```
mean(house$Price[house$Fireplaces == 1])
```

```
[1] 235.1629
```

```
mean(house$Price[house$Fireplaces == 2])
```

```
[1] 318.8214
```

```
mean(house$Price[house$Fireplaces == 3])
```

```
[1] 360.5
```

```
mean(house$Price[house$Fireplaces == 4])
```

```
[1] 700
```

We see that the average house price increases quite dramatically as the number of fireplaces increase. It's looking like I should build that fireplace! It should be no surprise that the two variables are correlated:

```
cor(house$Price, house$Fireplaces)
```

```
[1] 0.3767862
```

Now, let's estimate the population model:

$$Price = \beta_0 + \beta_1 Fireplaces + \epsilon$$

where β_0 would be the price of a house with 0 fireplaces, and β_1 is the increase in house price for an additional fireplace. The R code to estimate this model via OLS in R, and the resulting output, are as follows:

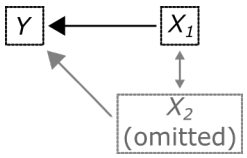
```
summary(lm(Price ~ Fireplaces, data=house))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  171.824      3.234   53.13   <2e-16 ***
Fireplaces    66.699      3.947   16.90   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 91.21 on 1726 degrees of freedom
Multiple R-squared:  0.142,     Adjusted R-squared:  0.1415
F-statistic: 285.6 on 1 and 1726 DF,  p-value: < 2.2e-16
```

What is the estimated marginal effect of `Fireplaces` on `Price`? Take a minute to google the cost of fireplace installation. As an economist, this should trouble you deeply. If the estimated value of an additional fireplace is $66,700, and if it only costs $10,000 to install a fireplace, we should see lots of houses with many fireplaces. Something is wrong here. To conclude this section, think about what the main determinant of house price should be.

6.2 Omitted variable bias

Figure 6.1: An *omitted* X_2 variable that is correlated with X_1, and that also determines Y, will make estimation of the true effect of X_1 on Y impossible.



The above OLS estimator (b_1 in the house prices example) is suffering from omitted variable bias. Omitted variable bias (OVB) occurs when one or more of the variables in the random error term (ϵ) are related to one or more of the X variables. Recall that ϵ contains all of the variables that determine Y, but that are unobserved (or omitted). Also, recall that one of the assumptions required for OLS to be a "good" estimator is A.5: ϵ and X are independent. If A.5 is not true, the OLS estimator can be biased (giving the wrong answer on average).

Suppose that there are two variables that determine Y: X and Z. Also suppose that X and Z are correlated (not independent). When X changes, Y changes. But when X changes, Z changes too (because Z and X are related), and this change in Z also causes a change in Y. If Z is omitted so that we only observe X and Y, then we cannot attribute changes in X directly to changes in Y. The changes in Z will "channel" through X. The OLS estimator for the effect of X on Y will be biased, unless the Z variable is included.

6.2.1 House prices revisited

What is the important omitted variable from the above house prices example? It seems like the estimated effect of `Fireplaces` on `Price` is too large. In fact, it may be that the number of fireplaces is just indicating the *size* of the house, which is really important for price!

Let's add the `Living.Area` variable to our population model:

$$Price = \beta_0 + \beta_1 Fireplaces + \beta_2 Living.Area + \epsilon$$

The R command and associated output is:

```
summary(lm(Price ~ Fireplaces + Living.Area, data = house))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.730146   5.007563   2.942  0.00331 **
Fireplaces   8.962440   3.389656   2.644  0.00827 **
Living.Area  0.109313   0.003041  35.951  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.98 on 1725 degrees of freedom
Multiple R-squared:  0.5095,    Adjusted R-squared:  0.5089
F-statistic: 895.9 on 2 and 1725 DF,  p-value: < 2.2e-16
```

Several results have changed with the addition of the `Living.Area` variable:

- The estimated value of an additional fireplace has dropped from $66,699 to $8,962.
- The R^2 has increased from 0.142 to 0.5095.
- The estimated intercept has changed by a lot (but this is unimportant).
- There is a new estimated β: $b_2 = 0.11$. This means that, it is estimated that an additional square-foot of house size increases price by $110.

So, what is going on here? From the first regression, the results are:

$$\begin{aligned}\hat{Price} =& 171.82 + 66.70 \times Fireplaces,\ R^2 = 0.142\\ & (3.23)\quad (3.95)\end{aligned}$$

and from the second regression:

$$\begin{aligned}\hat{Price} =& 14.73 + 8.96 \times Fireplaces + 0.11 \times Living.Area,\ R^2 = 0.511\\ & (5.01)\ \ (3.39)\qquad\qquad\quad (0.003)\end{aligned}$$

Why has the estimated effect of `Fireplace` on `Price` changed so much? `Living.Area` is an important variable. Arguably, the most important factor in determining house price is the size of the house. Houses that have more fireplaces tend to be larger. (There usually aren't two fireplaces in one room, for example). So, `Fireplaces` and `Living.Area` are correlated:

```
cor(house$Fireplaces, house$Living.Area)
```

```
[1] 0.4737878
```

When `Living.Area` is *omitted* from the regression, its effect on `Price` becomes mixed up in the effect of `Fireplaces` on `Price`. That is, when the house has more fireplaces, that means it's a larger house, so there are two reasons for a higher price. Lots of fireplaces is just indicating the house is large!

This is an example of omitted variable bias (OVB). When `Living.Area` is omitted, the OLS estimator is biased (in this case the effect of more fireplaces on house price is estimated to be way too large). OVB provides an important motivation for the multiple regression model: even though we may only be interested in estimating one marginal effect, we still should include other variables that are correlated to X, otherwise our estimator is biased. OVB is solved by adding the extra variables to the equation, thus *controlling* for their effect.

6.3 OLS in multiple regression

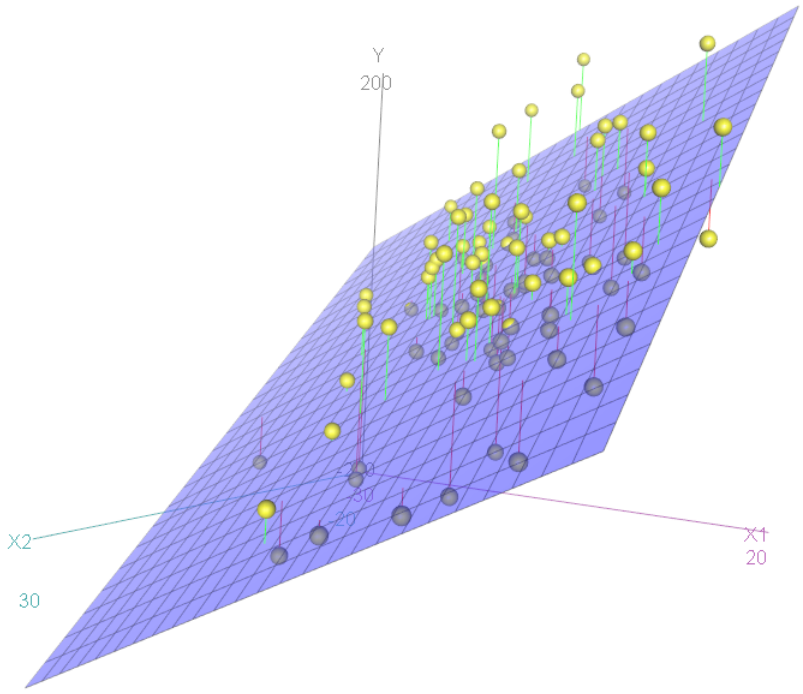
6.3.1 Derivation

The OLS estimators, $b_0, b_1, \ldots, b_k$, are derived similarly to how they were in chapter 4 (when we only had one X variable). The formulas are obtained by choosing $b_0, b_1, \ldots, b_k$ so that the sum of squared residuals is minimized:

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} e_i^2$$

This involves taking $(k+1)$ derivatives, setting them all equal to zero, and solving the system of equations. The formulas become too complicated to write, unless we use matrices (which we won't do here).

Figure 6.2: An OLS estimated regression plane (two X variables). The plane is chosen so as to minimize the sum of squared vertical distances indicated in the figure. The figure was drawn using the `scatter3d` function from the `rgl` package.



Now that we have multiple X variables, many concepts that we have already discussed become much more difficult to *visualize*. For example, the estimated model:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \cdots + b_k X_{ki} \tag{6.2}$$

can not be interpreted as a line! A line (with an intercept and slope) can be drawn in two dimensional space. The estimated model in equation 6.2 has k dimensions (and is a k-dimensional hyperplane). However, if we have only two X variables:

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i}$$

then we can still represent the estimated model in 3-dimensional space (see figure 6.2).

6.3.2 Interpretation

Let's look at a population model with two X variables:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{6.3}$$

- Y is still the dependent variable
- X_1 and X_2 are the independent variables (the regressors)
- i still denotes an observation number
- β_0 is the population intercept
- β_1 is the effect of X_1 on Y, holding all else constant (X_2)
- β_2 is the effect of X_2 on Y, holding all else constant (X_1)
- ϵ is the regression error term (containing all the omitted factors that effect Y)

Nothing substantial has changed. β_1, for example, is the marginal effect of X_1 on Y, while holding X_2 constant. In the fireplaces example, by inlcuding `Living.Area` in the regression we are able to find the marginal effect of fireplaces while holding house size constant. When we add more variables to the model, the interpretation of the βs remains the same.

6.4 OLS assumption A2: no perfect multicollinearity

In this section, we pay special attention to assumption A2, which has only now become relevant in the context of the multiple regression model.

A2 There is no perfect multicollinearity between the X variables.
This assumptions means that no two X variables (or combinations of the variables) can have an exact linear relationship. For example, exact linear relationship between Xs are:

- $X_1 = X_2$
- $X_1 = 100X_2$
- $X_1 = 1 + X_2 - 3X_3$

In these examples, you can figure out what one of the Xs will be, if you know the other Xs. This situation is usually called perfect multicollinearity. The data contains redundant information. This shouldn't be much of a problem, except that the OLS formula doesn't allow all of the estimators to be calculated (the problem is similar to trying to divide by zero).

Using R, let's see what happens when we try to include an X variable that is a perfectly linear relationship with another X variable. We'll use the house price data again. The `Living.Area` variable measures the size of the house in square feet. Suppose that there was another variable in the data set that measured house size in square metres (1 square foot = 0.0929 square metre). We can create this variable in R using:

```
house$House.Size <- house$0.0929 * Living.Area
```

and now let's include it in our OLS estimation:

```
summary(lm(Price ~ Fireplaces + Living.Area + House.Size, data = house))
```

```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.730146   5.007563   2.942  0.00331 **
Fireplaces   8.962440   3.389656   2.644  0.00827 **
Living.Area  0.109313   0.003041  35.951  < 2e-16 ***
House.Size         NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68980 on 1725 degrees of freedom
Multiple R-squared:  0.5095,    Adjusted R-squared:  0.5089
F-statistic: 895.9 on 2 and 1725 DF,  p-value: < 2.2e-16
```

Notice the error message "1 not defined because of singularities", and the row of "NA"s (not available). So, R recognized that there was a problem, and dropped the redundant variable, but not all econometric software has been this clever.

Some common examples of where the assumption of "no perfect multicollinearity" is violated in practice are when the same variable is measure in different units (such as square feet and square metres, or dollars and cents), and in the *dummy variable trap*.

6.4.1 The dummy variable trap

The dummy variable trap occurs when one too many dummy variables are included in the equation. For example, suppose that we have a dummy variable `female` that equals 1 if the worker is female. Suppose that we also have a variable `male` that equals 1 if the worker is male. There is an exact linear combination between the two variables:

$$female = 1 - male$$

If you know the value for the variable `male`, then you automatically know the value for `female`. Including both `male` and `female` in the equation would be a violation of assumption A2, and would be referred to as the dummy variable trap for this example. That is, OLS would not be able to estimate all of the βs in the equation:

$$wage = \beta_0 + \beta_1 \times male + \beta_2 \times female + \epsilon$$

The `male` and `female` dummy variables is a simple example, in other situations it is much easier to fall into the "trap". For example, suppose that you are provided data on a worker's location by province or territory. That is, each worker has a `Location` variable that takes on one of the values: $\{AB, BC, MB, NB, NL, NS, NT, NU, ON, PE, QC, SK, YT\}$. How should this variable be used? Typically, a series of dummy variables would be created from the `Location` variable:

$$\begin{aligned} Alberta &= 1 \text{ if } Location = AB; 0 \text{ otherwise} \\ British.Columbia &= 1 \text{ if } Location = BC; 0 \text{ otherwise} \\ Manitoba &= 1 \text{ if } Location = MB; 0 \text{ otherwise} \\ &\vdots \\ Yukon &= 1 \text{ if } Location = YT; 0 \text{ otherwise} \end{aligned}$$

So, we could create 13 dummy variables from the `Location` variable, but if we included all of them in the regression, we would fall into the dummy variable trap! Instead, one of the provinces/territories must be left out of the equation. Whichever group is left out, it becomes the *base group*, to which comparisons are made.

The solution to perfect multicollinearity, then, is to identify the redundant variable(s), and simply drop it from the equation.

As a final note, it is *not* a violation of "no perfect multicollinearity" if we take a non-linear transformation of a variable in the data set. For example, if we create a new variable X_2 where $X_2 = X_1^2$, this is ok! In fact, we will make use of non-linear transformations in chapter 8.

6.4.2 Imperfect multicollinearity

Imperfect multicollinearity is when two (or more) variables are *almost* perfectly related. That is, they are very highly correlated. Suppose that the true population model is (remember, we don't actually know this in practice):

$$Y = 2X_1 + 2X_2 + \epsilon$$

and that the correlation between X_1 and X_2 is 0.99. Regress Y on X_1:

```
summary(lm(Y ~ X1))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.4165     3.8954  -1.134    0.263
X1            4.0762     0.4698   8.676 2.13e-11 ***
```

The estimated standard error is small, so that the t-statistic is large, and we are sure that X_1 is statistically significant. However, the estimated β_1 is twice as big as it should be. This is because of omitted variable bias. So, we add X_2 to the equation:

```
summary(lm(Y ~ X1 + X2))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4.676      3.956  -1.182    0.243
X1              1.958      4.075   0.481    0.633
X2              2.128      4.066   0.523    0.603
```

Now, the estimated βs are closer to their true value of 2, but both appear to be statistically insignificant! (Note the large standard errors and small t-statistics.)

The problem here is that, because X_1 and X_2 are highly correlated, it is difficult to attribute changes in one of the X variables to changes in Y, because both X_1 and X_2 are almost always changing together in a similar fashion. That is, the *ceteris paribus* assumption (all else equal), is not feasible when the variables are highly correlated. β_1 is the effect of X_1 on Y, *holding* X_2 *constant*. But, because of the correlation, the data can not provide us such a *ceteris paribus* environment.

The problem of imperfect multicollinearity shows up in the large standard errors for the estimated βs of the affected variables. Adding and dropping the affected variables may result in large swings in the estimated coefficients. Imperfect multicollinearity makes us unsure of our estimated results. The problem is difficult to address. We cannot drop one of the correlated variables, due to the problem of omitted variable bias. In fact, there is very little to be done here. We need more *information*, but presumably the sample size n cannot be increased. As long as the variables we are interested in studying are not part of the multicollinearity problem (and the ones that are part of the problem are there to avoid OVB), then multicollinearity is not an issue.

6.5 Adjusted R-squared

We should no longer use R^2 in the multiple regression model. This is because when we add a new variable to the model, R^2 must always increase (or at best stay the same). This means that we could keep adding "junk" variables to the model to arbitrarily inflate the R^2. This is not a good property for a "measure of fit" to have. Instead, we will use "adjusted R-squared", denoted by $\bar{R}^2$.

6.5.1 Why R^2 must increase when a variable is added

To see why R^2 must always increase when a variable is added, we begin by looking again at the formula:

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

and again at the minimization problem that defines the OLS estimators:

$$\min_{b_0, b_1, \ldots, b_k} \sum_{i=1}^{n} e_i^2$$

When we add another X variable, the minimized value of $\sum_{i=1}^{n} e_i^2$ must get smaller! OLS picks the values for the bs so that the sum of squared vertical distances are minimized. If we give OLS another option for minimizing those distances, the distances have to get smaller (or at the worst stay the same). So, adding a variable means RSS decreases, so R^2 increases. The only way that R^2 stays the same is if OLS chooses a value of 0 for the associated slope coefficient, which never happens in practice.

As an example, let's try adding a nonsense variable to the house price model: random dice rolls. Using R, 1728 die rolls are simulated (to match the house price sample size of $n = 1728$), are recorded as a variable `Dice`, and added to the regression. Notice the difference in "Multiple R-squared" (R^2) and "Adjusted R-squared" ($\bar{R}^2$) between the two regressions:

```
summary(lm(Price ~ Fireplaces + Living.Area, data = house))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.730146   5.007563   2.942  0.00331 **
Fireplaces   8.962440   3.389656   2.644  0.00827 **
Living.Area  0.109313   0.003041  35.951  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.98 on 1725 degrees of freedom
Multiple R-squared:  0.5095,    Adjusted R-squared:  0.5089
F-statistic: 895.9 on 2 and 1725 DF,  p-value: < 2.2e-16
```

```
summary(lm(Price ~ Fireplaces + Living.Area + Dice, data = house))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.105383   6.072084   1.994  0.04635 *
Fireplaces   8.829436   3.394526   2.601  0.00937 **
Living.Area  0.109378   0.003042  35.954  < 2e-16 ***
Dice         0.743506   0.972575   0.764  0.44469
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 68.99 on 1724 degrees of freedom
Multiple R-squared:  0.5097,    Adjusted R-squared:  0.5088
F-statistic: 597.3 on 3 and 1724 DF,  p-value: < 2.2e-16
```

The variable `Dice` has no business being in the regression of house prices, and we fail to reject the null hypothesis that its effect is zero, yet the R^2 increases. The adjusted R-squared ($\bar{R}^2$) decreases, however.

6.5.2 The $\bar{R}^2$ formula

Adjusted R-squared ($\bar{R}^2$) is a measure-of-fit that can either increase or decrease when a new variable is added. $\bar{R}^2$ is a slight alteration of the R^2 formula. It introduces a penalty into R^2 that depends on the number of X variables in the model. (Remember that the number of Xs in the model is denoted by k.)

$$\bar{R}^2 = 1 - \frac{RSS \ / \ (n-k-1)}{TSS \ / \ (n-1)} \tag{6.4}$$

The $\bar{R}^2$ formula is such that when a variable is added to the model, k goes up, which tends to make $\bar{R}^2$ smaller. We know from the previous discussion, however, that whenever a variable is added, RSS must decrease. So, whether or not $\bar{R}^2$ increases or decreases depends on whether the new variable improves the fit of the model enough to beat the penalty incurred by k.

The justification for the $(n-k-1)$ and $(n-1)$ terms is from a degrees-of-freedom correction. How many things do we have to estimate before we can calculate RSS? $k+1$ βs must first be estimated before we can get the OLS residuals, and RSS. If you want to use RSS for something else (such as a measure of fit), we recognize that we don't have n pieces of information left in the sample, we have $(n-k-1)$. A similar argument can be made for the $(n-1)$ term in equation 6.4.

6.6 Review Questions

1. Explain why the estimated value for β_1 changes so much between the equations:

   $$Price = \beta_0 + \beta_1 Fireplaces + \epsilon$$

   and

   $$Price = \beta_0 + \beta_1 Fireplaces + \beta_2 Living.Area + \epsilon$$

2. What are the two conditions that will make an omitted variable cause OLS to be biased?
3. Explain how the OLS estimators, $b_0, b_1, \ldots, b_k$, are derived in the multiple regression model. (Explain how the equations for $b_0, b_1, \ldots, b_k$ are obtained.)
4. For the model:

   $$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon,$$

   explain the interpretation of β_1 and β_2.
5. Why is perfect multicollinearity a problem for OLS estimation?
6. Explain how the "dummy variable trap" is a situation of perfect multicollinearity.
7. Explain what imperfect multicollinearity is, and how it poses a problem for OLS estimation.
8. Why does R^2 always increase when a variable is added to the model?
9. Explain where the $(n-k-1)$ and $(n-1)$ terms in $\bar{R}^2$ come from.
10. An estimated model with two X variables, and from a sample size of $n = 27$, yields $R^2 = 0.5882$. Calculate $\bar{R}^2$.
11. This question again uses the CPS data set, which can be loaded into R using the following commands:

```
cps <- read.csv("htpps://rtgodwin.com/data/cps1985.csv")
```

    a) Regress `wage` on `education`, `age`, and `gender`, and report your results.
    b) Why has the estimated returns to education changed from the exercise in chapter 5?
    c) Are the variables statistically significant?
    d) Test the hypothesis that there is no wage-gender gap.
    e) What is the predicted wage for a 40 year-old female worker with 12 years of education?
    f) What is the predicted wage for a 40 year-old male worker with 12 years of education? What is the difference from the previous question?
    g) Why are the R^2 and $\bar{R}^2$ so similar for this regression?
    h) Interpret the value of $\bar{R}^2$.
    i) Try adding the variable `experience` to the regression. Are all the variables still statistically significant? What is going on here?

6.7 Answers

1. The estimated value changes so much due to *omitted variable bias*. `Living.Area` is an important determination of house price, and is correlated with `Fireplaces` (larger houses have more fireplaces). The effect of house size is "channeling" through the number of fireplaces. The omission of `Living.Area` is causing the OLS estimator in the first equation to be biased (and inconsistent).
2. If the omitted variable is (i) a determinant of the dependent (Y) variable; and (ii) is correlated with one or more of the included (X) variables.
3. The OLS estimators in the multiple regression model are derived similarly to how they were in chapter 4. $b_0, b_1, \ldots, b_k$ are chosen so as to minimize the sum of squared residuals. Solving for $b_0, b_1, \ldots, b_k$ involves solving a calculus minimization problem.

4. β_1 is the marginal effect of X_1 on Y, holding X_2 constant. Similar for β_2. To prove this, we can take the partial derivative of Y with respect to (say) X_1:

$$\frac{\partial Y}{\partial X_1} = 0 + \beta_1 + 0 + 0 = \beta_1$$

This tells us that the change in Y resulting from a change in X_1, is β_1, and that these changes are independent from changes in X_2.

5. Perfect multicollinearity is a problem because the OLS estimator is not *defined*. That is, our computer software will be unable to calculate all of the OLS estimators.
6. The "dummy variable trap" is when a redundant dummy variable is included in the regression. This is a case of perfect multicollinearity: there is an exact linear relationship between the dummy variables. For example, suppose that I had a two dummy variables:

$$attended = \begin{cases} 1, & \text{if the student attended class} \\ 0, & \text{if the student did not attend class} \end{cases}$$

and

$$skipped = \begin{cases} 1, & \text{if the student skipped class} \\ 0, & \text{if the student did not skip class} \end{cases}$$

Including both of these variables in the equation would result in perfect multicollinearity because there is an exact linear relationship between the two variables:

$$attended = 1 - skipped$$

7. Imperfect multicollinearity is when two (or more) variables are highly correlated. In this situation, OLS can be imprecise (have high variance) because it is difficult to tell which of the two correlated variables is causing the change in Y. The problem of imperfect multicollinearity shows up in large standard errors and confidence intervals, and large swings in the estimated βs as the affected variables are added to and dropped from the model.
8. The bs in OLS are chosen so as to minimize the sum of squared residuals. When a variable is added to the model, a b is added to the minimization problem, giving one more way to minimize RSS. So, RSS must decrease (or possibly stay the same) when another b is added. By the formula for R^2, it can easily be seen that R^2 must increase.
9. The justification for the $(n-k-1)$ and $(n-1)$ terms are due to degrees-of-freedom. The amount of information in the RSS statistic is $(n-k-1)$ since $k+1$ βs must first be estimated by OLS. In the TSS statistic, one thing must be estimated first ($\bar{Y}$), so the amount of information left over is $(n-1)$.
10.

$$\begin{aligned} R^2 &= 1 - \frac{RSS}{TSS} = 0.5882 \\ \frac{RSS}{TSS} &= 1 - R^2 = 1 - 0.5882 = 0.4118 \\ \bar{R}^2 &= 1 - \frac{RSS \ / \ (n-k-1)}{TSS \ / \ (n-1)} \\ &= 1 - 0.4118\frac{(n-1)}{(n-k-1)} \\ &= 1 - 0.4118\left(\frac{26}{24}\right) = 0.5539 \end{aligned}$$

11. a)
```
summary(lm(wage ~ education + age + gender, data = cps1985))
```

Table 6.2: Regression results using the CPS data.

Dependent variable: wage	
Regressor	Estimate (standard error)
education	0.827***
	(0.075)
age	0.113***
	(0.017)
female	-2.335***
	(0.388)
intercept	-4.843***
	(1.244)
$n = 534$	
$\bar{R}^2 = 0.249$	
*** denotes significance at the 0.1% level	

b) The estimated returns to education have changed from 0.751 to 0.827. The formula for each OLS estimator (b) depends on all of the variables in the regression. So, when the X variables change the estimated results will change (unless the sample correlation between the variables is exactly 0, which is never the case in practice). The fact that the results change may indicate that the regression from chapter 5 was suffering from omitted variable bias.

c) Yes (see the p-values in R).

d) This hypothesis has already been tested for us. We reject at the 0.1% significance level.

e)

$$\hat{wage} = -4.843 + 0.827(12) + 0.113(40) - 2.335(1) = 7.266$$

f)

$$\hat{wage} = -4.843 + 0.827(12) + 0.113(40) - 2.335(0) = 9.601$$

The difference between the two predicted values ($9.601 - 7.266 = 2.335$) is equal to the estimated gender-wage gap.

g) R^2 and $\bar{R}^2$ differ by the term:

$$\frac{(n-1)}{(n-k-1}$$

As n grows, the difference between R^2 and $\bar{R}^2$ disappears. In the CPS data, the sample size is reasonably large at $n = 534$, and k is only equal to 3, making the two measures-of-fit quite similar.

h) 24.9% of the variation in wages can be explained using the three variables in the model.

i) When we add `experience` to the model:

```
summary(lm(wage ~ education + age + gender
                          + experience))
```

all variables except the female dummy variable lose statistical significance. This is due to imperfect multicollinearity. Age, education, and experience, are all closely related.

Chapter 7

Joint Hypothesis Tests

Now that we have multiple X variables and βs in our population model, we might want to test hypotheses that involves two or more of the βs at once. In these cases, we (typically) do not use t-tests. Instead, we will use the F-test.

7.1 Joint hypotheses

The types of hypotheses we are now considering involve multiple coefficients (βs). For example:

$$\begin{aligned} H_0 &: \beta_1 = 0,\ \beta_2 = 0 \\ H_A &: \beta_1 \neq 0 \text{ and/or } \beta_2 \neq 0 \end{aligned} \tag{7.1}$$

and

$$\begin{aligned} H_0 &: \beta_1 = 1, \beta_2 = 2, \beta_4 = 5 \\ H_A &: \beta_1 \neq 1 \text{ and/or } \beta_2 \neq 2 \text{ and/or } \beta_4 \neq 5 \end{aligned} \tag{7.2}$$

Note that the null hypothesis is wrong if *any* of the individual hypotheses about the βs are wrong. In the latter example, if $\beta_2 \neq 2$, then the whole thing is wrong. Hence the use of the "and/or" operator in H_A. It is common to omit all the "and/or" and simply write "not H_0" for the alternative hypothesis.

A joint hypothesis specifies a value (imposes a restriction) for two or more coefficients. Use q to denote the number of restrictions ($q = 2$ for hypothesis 7.1, and $q = 3$ for hypothesis 7.2).

7.1.1 Model selection

If we fail to reject hypothesis 7.1, this implies that we should drop X_1 and X_2 from the model. That is, if variables are insignificant, we might want to exclude them from the model. If we wish to drop multiple variables from the model at once, that means we are hypothesizing that all of the associated βs are jointly equal to zero.

Why would we want to drop (or omit) variables from the model? There are two main reasons:

- A simpler model is always better. The same reasons that we wish to have simple models in economics also apply to econometrics. Simple models are easier to understand, easier to work with. They focus on the things we are trying to explain.
- The fewer βs that we try to estimate, the more information is available for each. That is, the variance of the remaining OLS estimators will be smaller after we drop X variables.

We have to be careful when we drop variables, however! The cost of wrongly dropping a variable is high. We can end up with omitted variable bias. So, we should be careful and err on the side of caution, since it is generally held that the cost of *wrongly omitting a variable* (omitted variable bias) is higher than the cost of *wrongly including a variable* (a loss of efficiency).

7.2 Example: CPS data

Load the CPS data:

```
install.packages("AER")
library(AER)
data("CPS1985")
attach(CPS1985)
```

Regress `wage` on `education`, `gender`, `age`, and `experience`:

```
summary(lm(wage ~ education + gender + age + experience))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.9574     6.8350  -0.286    0.775
education      1.3073     1.1201   1.167    0.244
genderfemale  -2.3442     0.3889  -6.028 3.12e-09 ***
age           -0.3675     1.1195  -0.328    0.743
experience     0.4811     1.1205   0.429    0.668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.458 on 529 degrees of freedom
Multiple R-squared:  0.2533,    Adjusted R-squared:  0.2477
F-statistic: 44.86 on 4 and 529 DF,  p-value: < 2.2e-16
```

In the above regression, both `age` and `experience` appear to be statistically *insignificant* (the p-values in the table are 0.743 and 0.668, respectively). That is, the null hypothesis $H_0 : \beta_3 = 0$ cannot be rejected, and neither can the null hypothesis $H_0 : \beta_4 = 0$. This suggests that `age` and `experience` could be dropped from the model. However, to drop both of these variables we actually need to test the joint hypothesis:

$$\begin{aligned} H_0 &: \beta_3 = 0,\ \beta_4 = 0 \\ H_A &: \beta_3 \neq 0 \text{ and/or } \beta_4 \neq 0 \end{aligned}$$

t-tests won't work for this hypothesis. Instead we will use the F-test.

7.3 The failure of the t-test in joint hypotheses

A natural idea for testing $H_0 : \beta_3 = 0,\ \beta_4 = 0$ (for example), is to reject H_0 if either $|t_3| > 1.96$ and/or $|t_4| > 1.96$. There are two problems with this. First, the type I error will not be 5%, unless we increase the critical value (showing this is left as an exercise). A much bigger problem is that t_3 and t_4 are likely *not* independent (they are correlated).

For example, in the population model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \tag{7.3}$$

if X_3 and X_4 are correlated, then the OLS estimators b_3 and b_4 will also be correlated with each other (recall OVB and how adding a variable to the model changes all the estimates - the formula for each b depends on *all* the X variables). If b_3 and b_4 are correlated then t_3 and t_4 are correlated!

In population model 7.3, suppose that X_3 and X_4 are positively correlated. Consider the null $H_0 : \beta_3 = 0,\ \beta_4 = 0$. Given the sign of the correlation between X_3 and X_4 (positive), it is more likely that b_3 and b_4 have the same sign (both positive or both negative). It is less likely that one of the coefficients would be estimated to be negative, and the other positive. Seeing opposite signs in the

estimated coefficients would be additional evidence against the null hypothesis that is not taken into account by looking at the individual t-statistics.

We need a test that will take into account the correlations between all the variables that are involved in the test. Such a test is the F-test.

7.4 The *F*-test

The F-test takes into account the correlations between the OLS estimators. Suppose the null hypothesis is still $H_0 : \beta_3 = 0,\ \beta_4 = 0$. Since we are testing exactly two βs, the F-statistic formula can be written as:

$$F = \frac{1}{2}\frac{t_3^2 + t_4^2 - 2r_{t_3,t_4}t_3t_4}{1 - r_{t_3,t_4}^2}$$

where r_{t_3,t_4} is the estimated correlation between t_3 and t_4. The larger the F-statistic, the more likely we are to reject the null. The purpose of showing this formula here is to highlight that the F-test takes into account the correlation between t_3 and t_4. The formula becomes much too complicated when we are testing more than two βs.

To obtain a more convenient formula for the F-test statistic, we need the idea of a *restricted* and *unrestricted* model. The *restricted* model is obtained by incorporating the values chosen for the βs in the null hypothesis into the population model. That is, the null hypothesis chooses certain values for some of the βs, and when those values are substituted into the full population model, we get a restricted model. In the alternative hypothesis, the population model is fully unrestricted. That is, none of the βs are chosen beforehand, and all values can be chosen by OLS. To summarize:

- restricted model - the model under the null hypothesis. Some βs are chosen in the null, and substituted into the population model.
- unrestricted model - the model under the alternative hypothesis. All βs are free to be chosen by the estimation procedure (OLS).

The F-test can be implemented by estimating these two models, and using some summary statistics from the regression. The intuition is that, if the restrictions are true (if H_0 is true), then the "fit" of the two models should be similar. Alternatively, if the restrictions are false (the null is false), then the unrestricted model should "fit" much better than the restricted model. We can measure the fit of the two models using the residual sum-of-squares, or the R^2.

One version of the F-statistic formula is:

$$F = \frac{(RSS_r - RSS_u)/q}{RSS_u/(n - k_u - 1)} \tag{7.4}$$

where:

- RSS_r is the residual sum-of-squares from the restricted model
- RSS_u is the residual sum-of-squares from the unrestricted model
- q is the number of restrictions being tested
- k_u is the number of X variables in the unrestricted model, or the number of βs (not counting the intercept)

Recall that the unrestricted model *must* fit better than the restricted model (OLS has more options for minimizing RSS). Also, note that the F-statistic must be a positive number, since RSS is a sum-of-squares.

If the restrictions are true, then OLS should (approximately) choose values for the βs that are already in the null hypothesis. The restricted and unrestricted models will be similar, $(RSS_r - RSS_u)$

will be small (close to zero), the F-statistic will be close to zero, and we will tend to fail to reject the null. Alternatively, if the null is false, $(RSS_r - RSS_u)$ will be large, leading to a large F-statistic, and a tendency to reject.

Another (possibly more convenient and intuitive) formulation of the F-statistic involves the R^2 (*not* the adjusted R^2). We can solve R^2 for RSS using the formula:

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

and re-write the F-statistics formula in equation 7.4 as:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k_u - 1)} \tag{7.5}$$

where:

- R_r^2 is the (unadjusted) R^2 from the restricted model
- R_u^2 is the (unadjusted) R^2 from the unrestricted model
- q and k_u are as before

Table 7.1: χ^2 critical values for the F-test statistic.

q	5% critical value
1	3.84
2	5.99
3	7.82
4	9.49
5	11.07

Remember that whenever we add a β to the model that R^2 has to increase. This was the whole reason that we needed to use adjusted R-square ($\bar{R}^2$) instead. However, if the fit of the model doesn't change much when the restrictions are imposed, the R^2 will be similar between the two models, leading to a small F-statistic, and a tendency to fail to reject H_0. Alternatively, if imposing the restrictions makes a big difference in terms of the fit of the model, the F-statistic will be large and we will tend to reject H_0.

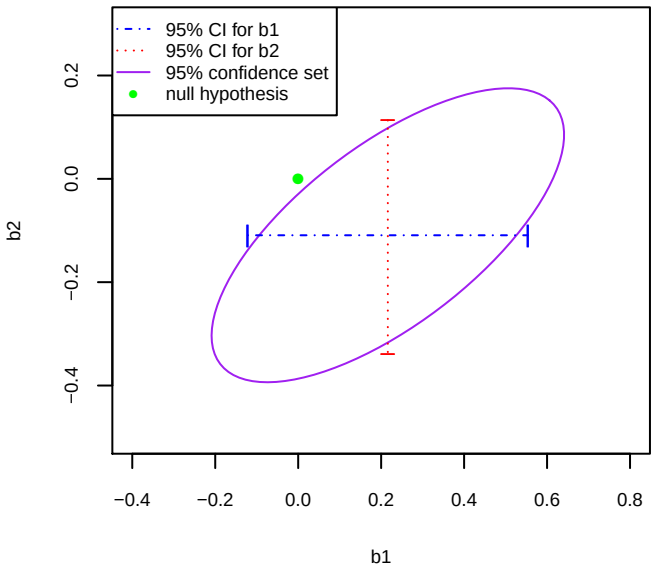
The F-test statistic that we have been discussing follows an F distribution with q and $(n - k_u - 1)$ degrees of freedom. If the sample size n is large, however, the F-statistic follows a χ^2 (chi-square) distribution with q degrees of freedom (similar to how the t-statistic follows a Normal distribution for large n). In this book we assume that n is large enough for this to be true. The F-statistic critical values for 5% significance, and for large n, are given in table 7.1. If the F-statistic exceeds the 5% critical value, the null hypothesis should be rejected at 5% significance.

7.5 Confidence sets

Confidence intervals may be used to test hypotheses that involve only one β. If the value chosen for β by the null hypothesis is within the confidence interval, we will fail to reject. In fact, one of the definitions for a confidence interval is that it is the interval that contains all values that can be chosen for a null hypothesis, that won't be rejected.

If our null hypothesis involves two βs, as in $H_0 : \beta_1 = 0,\ \beta_2 = 0$ for example, then the idea of a confidence interval would be extended to a *confidence set*. The confidence set would contain all the pairs of values for β_1 and β_2 that could be jointly chosen under the null hypothesis, where the null hypothesis would not be rejected.

Figure 7.1: Individual confidence intervals, and the confidence set.



7.5.1 Example: confidence intervals and a confidence set

Consider the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

which has been estimated by OLS:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6246     0.4660  -1.340    0.182
X1            0.2161     0.1723   1.255    0.211
X2           -0.1092     0.1153  -0.946    0.345
X3            2.9384     0.1092  26.914   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The 95% confidence interval around b_1 is $0.2161 \pm 1.96 \times 0.1723 = [-0.12,\ 0.55]$. The null hypothesis of $H_0 : \beta_1 = 0$ cannot be rejected at the 5% significance level since the value 0 is contained in the interval. By looking at the R output, we can tell that the 95% confidence interval contains 0 given that the p-value of 0.211 is greater than 0.05. Similarly, the confidence interval around b_2 is $-0.1092 \pm 1.96 \times 0.1153 = [-0.34,\ 0.12]$, and contains 0. Both X_1 and X_2 appear to be statistically insignificant, according to their individual confidence intervals.

Similar to why individual t-tests should not be used to test a joint hypothesis, neither should individual confidence intervals be used. In order to test the hypothesis:

$$H_0 : \beta_1 = 0,\ \beta_2 = 0$$
$$H_A : \text{not } H_0$$

using a predetermined set of values, we should use a confidence set containing all the *pairs* of β_1 and β_2 that won't be rejected. For this example, it turns out that the null hypothesis is not within the 95% confidence set, so that we reject the null hypothesis that both variables are statistically insignificant. We should not drop them from the model. This is a bit surprising considering the individual confidence intervals. The individual confidence intervals, and the confidence set for b_1 and b_2, are shown in figure 7.1.

The confidence set in figure 7.1 is a rotated ellipse. The angle of rotation is determined by the correlation between X_1 and X_2. Calculating the confidence intervals is easy, calculating the confidence set is not. Confidence sets are not typically used in practice in econometrics. The purpose of discussing them in this section was to reinforce the idea that the correlation between the variables must be considered when performing a joint hypothesis test.

7.6 Calculating the *F*-test statistic

To implement an F-test, we can estimate the *restricted* and *unrestricted* model, and compare the two. Using the previous data, we will test the hypothesis:

$$H_0 : \beta_1 = 0,\ \beta_2 = 0$$
$$H_A : \text{not } H_0$$

The full unrestricted model (under the alternative hypothesis) is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

The restricted model (under the null hypothesis) is:

$$Y = \beta_0 + \beta_3 X_3 + \epsilon$$

In R, we start by estimating these two models, and saving them:

```
unrestricted <- lm(Y ~ X1 + X2 + X3)
restricted <- lm(Y ~ X3)
```

Then, we can use the `anova` command to perform the F-test directly:

```
anova(restricted, unrestricted)
```

```
Analysis of Variance Table

Model 1: Y ~ X3
Model 2: Y ~ X1 + X2 + X3
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    198 8805.1
2    196 8472.7  2    332.37 3.8444 0.02303 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-statistic is 3.84, which is larger than the 5% critical value of 3.00 (see table 7.1). The p-value is 0.02303. We reject the null hypothesis at the 5% significance level.

To calculate the F-statistic using equation 7.5:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k_u - 1)}$$

we need the R^2 from the two models. From the unrestricted model, the R^2 is 0.7921:

```
summary(unrestricted)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6246     0.4660  -1.340    0.182
X1            0.2161     0.1723   1.255    0.211
X2           -0.1092     0.1153  -0.946    0.345
X3            2.9384     0.1092  26.914   <2e-16 ***
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.575 on 196 degrees of freedom
Multiple R-squared:  0.7921,    Adjusted R-squared:  0.7889
F-statistic: 248.9 on 3 and 196 DF,  p-value: < 2.2e-16
```

and from the restricted model the R^2 is 0.784:

```
summary(restricted)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5924     0.4719  -1.255    0.211
X3            2.9604     0.1104  26.804   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.669 on 198 degrees of freedom
Multiple R-squared:  0.784,     Adjusted R-squared:  0.7829
F-statistic: 718.5 on 1 and 198 DF,  p-value: < 2.2e-16
```

We are testing two restrictions ($q = 2$), and $n = 200$, so that the F-statistic is:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k_u - 1)} = \frac{(0.7921 - 0.784)/2}{(1 - 0.7921)/(200 - 3 - 1)} = 3.82$$

The number that we get by calculating the F-statistic using R^2 is a little different than from the `anova` command due to rounding.

7.7 The overall *F*-test

Regression software almost always reports the results of an "overall" F-test, whenever a model is estimated. The null and alternative hypotheses for this overall F-test is:

$$\begin{aligned} H_0 &: \ \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_A &: \text{at least one } \beta \neq 0 \end{aligned} \tag{7.6}$$

Again, k denotes the number of X variables in the model. This null hypothesis says that none of the X variable can explain the Y variable. It is a test to see if the estimated model is garbage. The intercept (β_0) is not included in the null hypothesis, otherwise there would be nothing to estimate, and if $\beta_0 = 0$ then the mean of Y is also zero (a somewhat silly hypothesis in most cases). The overall F-test statistic is reported in the bottom line of R ouptut. In the previous two examples the overall F-test statistic is 248.9 and 718.5, with associated p-values of 0 (to 16 decimal places). There is evidence that at least one X variable explains Y.

We also take this opportunity to point out that, when $q = 1$, the t-test and F-test provide identical results. In fact, when $q = 1$, $F = t^2$. This can be verified from the previous R output. The t-statistic on `X3` is 26.804, and $26.804^2 = 718.5$ (the overall F-statistic).

7.8 R output for OLS regression

We can now understand all of the R output from OLS estimation, except for "residual standard error". This is just the sample standard deviation of the OLS residuals. It is also used as a measure of fit, and is also sometimes called the root mean-squared-error. The residual standard error is:

$$\sqrt{\frac{\sum e_i^2}{n - k - 1}}$$

We have not discussed this elsewhere in the book, but mention it here as a matter of finality. We now know what everything is in the standard R output for OLS estimation.

7.9 Review Questions

1. Explain what is meant by a joint hypothesis, and provide an example.
2. Explain what the restricted and unrestricted models are in a joint hypothesis test.
3. Explain why t-tests can't be used to test a joint hypothesis.
4. Calculate the type I error (which is also the significance) when testing:

$$H_0 : \beta_3 = 0,\ \beta_4 = 0$$
$$H_A : \text{not } H_0$$

   using two individual t-tests with critical value 1.96, and assuming that the t-statistics are independent.
5. Use the CPS data. Let the full unrestricted population model be:

$$wage = \beta_0 + \beta_1 education + \beta_2 gender + \beta_3 age + \beta_4 experience + \epsilon$$

   a) Use t-tests to test the null hypothesis: $H_0 : \beta_3 = 0,\ \beta_4 = 0$.
   b) Use the `anova` command to test the null hypothesis from part (a).
   c) Use the R^2 from the unrestricted and restricted models to calculate the F-statistic for the null hypothesis in part (a). Use table 7.1 to decide whether to reject or fail to reject.
   d) Roughly sketch the confidence set for b_3 and b_4.
   e) Test the null hypothesis: $H_0 : \beta_1 = 0,\ \beta_2 = 0,\ \beta_3 = 0,\ \beta_4 = 0$.
   f) Using this data, and a null hypothesis of your choosing, verify that $t^2 = F$.

7.10 Answers

1. A joint hypothesis is a null hypothesis that involves two or more parameters (βs). That is, the null hypothesis is *jointly* specifying the values of two or more βs. See equations 7.1 and 7.2 for examples.
2. One way of conducting a joint hypothesis test is to estimate two separate models. The population model can be considered as the *unrestricted* model under the alternative hypothesis. It is unrestricted since none of the values are chosen (by H_0), and all βs are free to be estimated. The null hypothesis, H_0, however, is choosing (restricting) some of the values of the βs. When the restrictions under H_0 are incorporated into the population model, we get a *restricted* model.
3. t-tests are typically not used to test joint hypotheses for two reasons. (i) The usual critical values (such as 1.96 for 5% significance) would have to be adjusted. (ii) The estimators that are used in the hypothesis test (the OLS estimators b) are likely not-independent (e.g. correlated). This means that the individual t-statistics are also likely to be correlated. Unless this correlation is taken into account,
4. We will calculate the type I error assuming that the t-statistics are independent. Using two individual t-tests, the null hypothesis would be rejected if either, or both, of the t-statistics exceed 1.96 in absolute value. There are four possible outcomes: (i) both t-statistics are less than 1.96 (in absolute value), (ii) both are greater than 1.96, (iii) $|t_3| > 1.96$ and $|t_4| \leq 1.96$, (iv) $|t_3| \leq 1.96$ and $|t_4| > 1.96$. Only in (i) do we fail to reject the null. The probability of (i) occurring is $0.95 \times 0.95 = 0.9025$. So the probability of rejecting H_0 when it is true (the type I error) is the probability of (ii), (iii) and (iv), which is 1 minus the probability of (i), or 0.0975 (not 0.05). We could get the "right" type I error by increasing the critical value from 1.96. This, however, does not solve the larger problem of the dependence between the t-statistics.
5. Load the CPS data:

```
cps <- read.csv("http://rtgodwin.com/data/cps1985.csv")
```

a) First we need to estimate the model. Regress `wage` on `education`, `gender`, `age`, and `experience` (put the R code all on one line):

```
summary(lm(wage ~ education + gender + age + experience, data=cps))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.9574     6.8350  -0.286    0.775
education      1.3073     1.1201   1.167    0.244
genderfemale  -2.3442     0.3889  -6.028 3.12e-09 ***
age           -0.3675     1.1195  -0.328    0.743
experience     0.4811     1.1205   0.429    0.668
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.458 on 529 degrees of freedom
Multiple R-squared:  0.2533,    Adjusted R-squared:  0.2477
F-statistic: 44.86 on 4 and 529 DF,  p-value: < 2.2e-16
```

From the R output, we see that the individual t-statistics on `age` and `experience` are small (-0.328 and 0.429, with p-values 0.743 and 0.668). This indicates that we should fail to reject the null hypothesis.

b) We need to estimate a restricted model (under the null hypothesis):

```
restricted <- lm(wage ~ education + gender, data=cps)
```

and an unrestricted model (under the alternative hypothesis):

```
unrestricted <- lm(wage ~ education + gender
                                        + age + experience)
```

and use the `anova` command to get the relevant F-statistic:

```
anova(restricted, unrestricted)
```

```
Analysis of Variance Table

Model 1: wage ~ education + gender
Model 2: wage ~ education + gender + age + experience
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    531 11425
2    529 10511  2    914.27 23.007 2.625e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The F-statistic is 23.007 with a p-value of 0.000. We reject the null hypothesis. This is the opposite result of what the t-statistics would indicate.

c) We can find the R^2 from the restricted model using the command:

```
summary(restricted)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)   0.21783    1.03632   0.210    0.834
education     0.75128    0.07682   9.779  < 2e-16 ***
genderfemale -2.12406    0.40283  -5.273 1.96e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.639 on 531 degrees of freedom
Multiple R-squared:  0.1884,    Adjusted R-squared:  0.1853
F-statistic: 61.62 on 2 and 531 DF,  p-value: < 2.2e-16
```

So, $R_r^2 = 0.1884$. The R^2 from the unrestricted model is $R_u^2 = 0.2533$ (see the R output in part (a)). We are testing two restrictions, so that $q = 2$. The sample size is $n = 534$. The number of X variables in the unrestricted model is 4, so that $k_u = 4$. We can now calculate the F-statistic using equation 7.5:

$$F = \frac{(R_u^2 - R_r^2)/q}{(1 - R_u^2)/(n - k_u - 1)} = \frac{(0.2533 - 0.1884)/q}{(1 - 0.2533)/(534 - 4 - 1)} = 22.989$$

This is very close to the F-statistic that was obtained using the `anova` command in part (b). Using table 7.1, we see that the relevant 5% critical value is 3.00. Since $22.989 > 3.00$, we reject the null hypothesis at the 5% significance level.

d) The main feature of the confidence ellipse is that it should be rotated about the origin. See figure 7.1 for an example. The rotation of the ellipse reflects the non-independence of the estimators, b_3 and b_4.

e) The null hypothesis in this question is referring to the "overall F-test". This F-test statistic is calculated for us when we use the `summary` command. From the output in part (a), this F-statistic is 44.86 with p-value 0.000. We reject the null hypothesis.

f) The F-test and t-test are equivalent when $q = 1$. Specifically, $t^2 = F$. Note that the 5% critical value for $q = 1$ in the F-test (3.84) is the square of the 5% critical value in the t-test (1.96).
To verify the equivalence of the F-test and t-test, we'll calculate the F-statistic for a null hypothesis where $q = 1$, and make sure that it is the square of the corresponding t-statistic. Note that, in the R output in part (a), the t-statistic on `education` is 1.167. So, for the test:

$$H_0 : \beta_1 = 0$$
$$H_A : \beta_1 \neq 0$$

The F-statistic should be $F = 1.167^2 = 1.362$. Estimate the restricted model under this null hypothesis, and use the `anova` command:

```
restricted2 <- lm(wage ~ gender + age + experience)
anova(restricted2, unrestricted)
```

```
Analysis of Variance Table

Model 1: wage ~ gender + age + experience
Model 2: wage ~ education + gender + age + experience
  Res.Df   RSS Df Sum of Sq     F Pr(>F)
1    530 10538
2    529 10511  1    27.063 1.362 0.2437
```

Chapter 8

Non-Linear Effects

Many models in economics involve *non-linear effects*. A non-linear effect just means that the effect of one variable on another is *not constant*. For example, diminishing marginal utility says that as more is consumed, eventually there is less of an increase to utility than previous. The effect of quantity consumed on utility is *not constant* (there is a non-linear relationship between quantity and utility). Increasing and decreasing returns to scale is another example of a non-linear effect that you may have encountered in your first-year economics courses. Increasing returns to scale implies that when the inputs of production are doubled, output would more than double. The prevalence of the terms "marginal" and "increasing" or "decreasing" in many of our economic models would suggest a need to handle non-linearity.

8.1 The linear model

The models we have seen so far have been linear. In the population model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k + \epsilon$$

the change in Y due to a change in X_1 (for example) is: $\Delta Y / \Delta X_1 = \beta_1$. This effect of X_1 on Y is *constant*. For many relationships between variables, this is unreasonable.

As an example of how the linear model does *not* work, we use the `Diamond` data from the `Ecdat` R package (data originally from Chu, 2001). A plot of the *price* and *carats* of diamonds are shown in figure 8.1, with the OLS estimated line included in the plot. The relationship between *price* and *carats* appears to be non-linear. The effect of *carat* on *price* appears to be small when then diamond is small, and gets large as the size of the diamond grows. The reason for this might be that large diamonds are more *rare*. A larger diamond can always be cut into smaller diamonds, but two diamonds cannot be combined to make a larger one. The linear model says that the effect of *carat* on *price* is constant, no matter how large or small the diamond is to begin with.

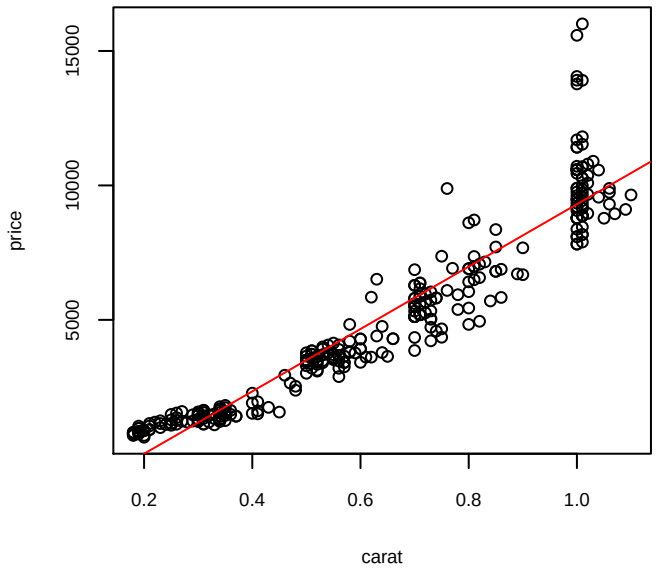
Ideally, we would like an estimated model that is capable of capturing the half "U" shape that we see in the diamonds plot, and other such non-linear shapes. If the true relationship between the two variables is non-linear, then the linear model is *misspecified*. OLS is biased and inconsistent. For situations like this, we need to specify a population model that allows for the marginal effect of X on Y to change depending on the value of X.

8.2 Polynomial regression model

A non-linear relationship between two variables can be approximated using a *polynomial* function. The validity of the approximation is based on a Taylor series expansion. A population model with a polynomial is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \cdots + \beta_r X_1^r + \epsilon \tag{8.1}$$

Figure 8.1: Price of diamonds, and carats, with OLS estimated line.



Equation 8.1 has a polynomial of degree r in X_1. If $r = 2$ we get a quadratic equation, and if $r = 3$ we get a cubic equation. Note that this is just the linear model that we have been using all along, but some of the regressors are powers of X_1. Other variables (X_2, X_3, etc.) can be added as usual. With the polynomial, estimation by OLS, and hypothesis testing, is the same as usual. Including powers of X_1 in the model as additional regressors is *not* a violation of no perfect multicollinearity (assumption A.2), because the relationship between the regressors is not linear.

8.2.1 Interpreting the βs in a polynomial model

The βs in the polynomial model become much more difficult to interpret. This is the point in including them. We are trying to model a (more complicated) non-linear relationship. Let's take a population model with a quadratic term (usually squaring is sufficient to model the non-linear effect):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_2^2 + \epsilon \tag{8.2}$$

In equation 8.2, β_1 is the marginal effect of X_1 on Y, but the marginal effect of X_2 on Y depends on both β_2 *and* β_3. That is, β_2 and β_3 don't make much sense by themselves. If we take the partial derivative of Y with respect to X_2, we get:

$$\frac{\partial Y}{\partial X_2} = \beta_2 + 2\beta_3 X_2$$

This derivative tells us that the squared term (X_2^2) allows the effect of X_2 on Y to *depend on the value of* X_2. A change in Y due to a change in X_2 is not constant, but depends on the value of X_2.

Including the squared term is just a mathematical "trick" for approximating the non-linear relationship. For example, if β_2 is positive, then a negative β_3 means there is a diminishing effect, and a positive β_3 means there is an increasing effect. OLS is free to choose values for β_2 and β_3 to best capture any non-linear relationship.

In order to obtain an interpretation for our estimated polynomial model, we can consider specific OLS predicted values. If we consider a lot of predicted values, we can plot them out in the data and see our estimated equation. If we calculate at least two pairs of predicted values, and take the differences between them, we can get an idea about how the estimated effect depends on the value of the X variable. This is illustrated in a following example.

8.2.2 Determining r

To determine the degree (r) of the polynomial, we can use a series of t-tests. We can start with a polynomial of degree r, and test the null hypothesis $H_0 : \beta_r = 0$. If we fail to reject (implying that X^r is not needed) then we re-estimate the model with a polynomial of degree $r - 1$. The process repeats until the null hypothesis is rejected. However, in most econometrics models only squared terms are used if needed; very rarely are there cubed (or higher) terms. Testing for the degree of r is illustrated in the following example.

8.2.3 Modelling the non-linear relationship in the Diamond data

We start by loading the Diamond data:

```
diamond <- read.csv("https://rtgodwin.com/data/diamond.csv")
```

and estimating the *linear* model, $price = \beta_0 + \beta_1 carat + \epsilon$:

```
summary(lm(price ~ carat, data=diamond))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2298.4       158.5  -14.50   <2e-16 ***
carat        11598.9       230.1   50.41   <2e-16 ***
```

It is estimated that an increase in *carat* of 1 is associated with an increase in the *price* of a diamond by \$11598.9. It might be more sensible to consider the smaller increase of 0.1 carats: an increase of 0.1 carats is associated with an increase in price of \$1160. This effect is the same whether the diamond is small or large to begin with.

In order to allow for the effect of *carat* on *price* to depend on the size of the diamond, we can include a quadratic term, and estimate the population model $price = \beta_0 + \beta_1 carat + \beta_2 carat^2 + \epsilon$. We can include the new variable $carat^2$ in the model using the `I()` function (where `^` is the power operator; shift-6 on most keyboards):

```
summary(lm(price ~ carat + I(carat^2), data=diamond))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -42.51     316.37  -0.134   0.8932
carat        2786.10    1119.61   2.488   0.0134 *
I(carat^2)   6961.71     868.83   8.013  2.4e-14 ***
```

Notice that `carat2` is highly statistically significant. There is evidence that the effect is non-linear.

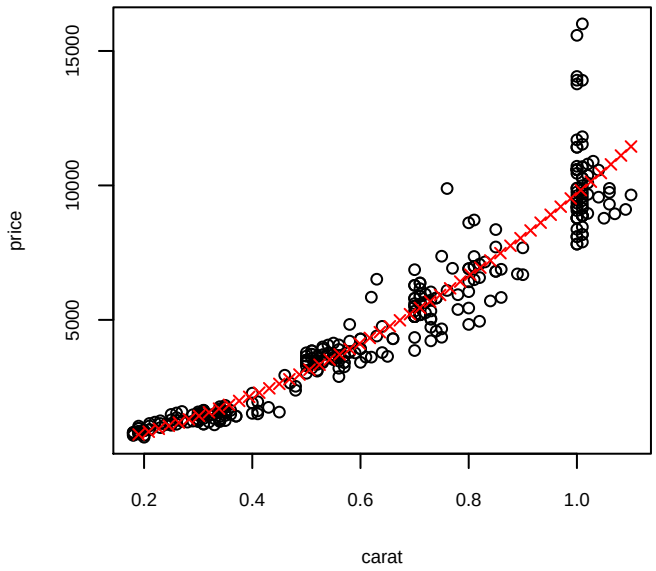
The positive sign on `carat2` means that we have estimated an *increasing* marginal effect. How do we interpret our estimated βs further? That is, what is the estimated effect of *carats* on *price*? The key is to calculate some OLS *predicted values*, to consider some specific scenarios. In figure 8.2, I calculate 50 OLS predicted values by choosing values for *carat* at regular intervals, and plot them over the Diamond data. Notice that our estimated equation captures the half "U" shape, and seems to fit the data well.

The predicted values used in figure 8.2 were obtained by substituting different values for *carat* into the estimated equation:

$$\hat{price} = -42.51 + 2786.10 carat + 6967.71 carat^2 \tag{8.3}$$

Now, let's focus on two specific scenarios: the effect of an increase in *carats* when (i) the diamond is small, and (ii) the diamond is large. Let's consider an increase of 0.1 in *carats* when the diamond is (i) 0.2 *carats* in size, and (ii) 1 *carat* in size. We need two predicted values for each scenario. For (i),

Figure 8.2: Diamond data, with estimated quadratic model.



we get the predicted values for $carat = 0.2$ and for $carat = 0.3$:

$$\hat{price}|_{carat=0.2} = -42.51 + 2786.10(0.2) + 6967.71(0.2)^2 = 793$$
$$\hat{price}|_{carat=0.3} = -42.51 + 2786.10(0.3) + 6967.71(0.3)^2 = 1420$$

and take the difference between these two predicted values:

$$\hat{price}|_{carat=0.3} - \hat{price}|_{carat=0.2} = 1419.88 - 793.18 = 627$$

So, the predicted effect of an increase in *carats* of 0.1, when the diamond is 0.2 *carats*, is \$627. We can also get this value by using the R code:

```
predict(quadmod, data.frame(carat = 0.3)) - predict(quadmod, data.frame(carat = 0.2))
```

```
> 626.6952
```

Now we consider the effect of a 0.1 increase in *carats* for (ii) a large diamond:

$$\hat{price}|_{carat=1} = -42.51 + 2786.10(1) + 6967.71(1)^2 = 9705$$
$$\hat{price}|_{carat=1.1} = -42.51 + 2786.10(1.1) + 6967.71(1.1)^2 = 11446$$

and again take the difference between the two predicted values:

$$\hat{price}|_{carat=1.1} - \hat{price}|_{carat=1} = 11446 - 9705 = 1741$$

The predicted effect of an increase in *carats* is larger, when the diamond is larger. That is, the estimated effect of a 0.1 increase in *carats* is \$1741.

The important point of this exercise is the following. The estimated effect of *carats* on *price* is much different depending on whether the diamond is large or small (\$627 when $carats = 0.2$ vs. \$1741 when $carats = 1$. The linear model estimates a constant effect of \$1160, which misses out on important non-linearities.

Finally, we determine the appropriate degree of the polynomial in `carat` (in practice, we should begin with this step). Let's estimate a cubic model: $price = \beta_0 + \beta_1 carat + \beta_2 carat^2 + \beta_3 carat^3 + \epsilon$. To estimate the model, use:

```
summary(lm(price ~ carat + I(carat^2) + I(carat^3), data=diamond))
```

```
(Intercept)    786.3       765.4    1.027    0.3051
carat        -2564.2      4636.9   -0.553    0.5807
I(carat^2)   16638.9      8185.3    2.033    0.0429 *
I(carat^3)   -5162.5      4341.9   -1.189    0.2354
```

The cubed variable, `carat3`, is insignificant (with p-value 0.2354). The quadratic specification is sufficient for capturing the non-linear relationship between *carat* and *price*. It is often the case that a quadratic specification is good enough.

8.3 Logarithms

Another way to approximate the non-linear relationship between Y and X is by using logarithms. Logarithms can be used to approximate a percentage change. If the relationship between two variables can be expressed in terms of proportional or percentage changes, then it is a type of non-linear effect. To see this, consider a 1% increase in 100 (which is 1), and a 1% increase in 200 (which is 2). The same 1% increase can be generated by different changes in the variable (e.g. a change of 1 or of 2).

For example, consider an increase in hourly wage of \$1. That is not a big increase for someone making \$50 per hour (an increase of only 2%). This change in wage is unlikely to have much effect on the behaviour of the individual. However, imagine an individual whose hourly wage is only \$1 per hour. An increase of \$1 doubles the wage (100% increase)! This is likely to have a big impact on behaviour.

It is desirable to measure thinks like wage in terms of proportional or percentage changes (regardless of whether it is included in a model as the dependent variable or as a regressor). This can be accomplished by using the log of the variable in the regression model, instead of the variable itself.

8.3.1 Percentage change

Let's be explicit about what is meant by a percentage change. A percentage change in X is:

$$\frac{\Delta X}{X} \times 100 = \frac{X_2 - X_1}{X_1} \times 100$$

where X_1 is the starting value of X, and X_2 is the final value.

8.3.2 Logarithm approximation to percentage change

The approximation to percentage changes using logarithms is:

$$\log(X + \Delta X) - \log(X) \times 100 \approx \frac{\Delta X}{X} \times 100$$

or

$$\log(X_2 - X_1) \times 100 \approx \frac{X_2 - X_1}{X_1} \times 100$$

So, when X changes, the change in $\log(X)$ is approximately equal to a percentage change in X. The approximation is more accurate the smaller the change in X. Table 8.1 shows variation percentage changes in X, and the approximate change using the log function. The approximation does not work well for changes above 10%.

8.3.3 Logs in the population model

The log function can be used in our population model so that the βs have various *percentage changes* interpretations. There are three ways we can introduce the log function into our models. The three

Table 8.1: Percentage change, and approximate percentage change using the log function.

Change in X	Percentage change: $\frac{X_2-X_1}{X_1} \times 100$	Approximated percentage change: $(\log X_2 - \log X_1) \times 100$
$X_1 = 1, X_2 = 2$	100%	69.32%
$X_1 = 1, X_2 = 1.1$	10%	9.53%
$X_1 = 1, X_2 = 1.01$	1%	0.995%
$X_1 = 5, X_2 = 6$	20%	18.23%
$X_1 = 11, X_2 = 12$	9.09%	8.70%
$X_1 = 11, X_2 = 11.1$	0.91%	0.91%

Table 8.2: Three population models using the log function.

Population model	Population regression function
I. linear-log	$Y = \beta_0 + \beta_1 \log X + \epsilon$
II. log-linear	$\log Y = \beta_0 + \beta_1 X + \epsilon$
III. log-log	$\log Y = \beta_0 + \beta_1 \log X + \epsilon$

different possibilities arise from taking logs of the left-hand-side variable, one or more of the right-hand-side variables, or both. Table 8.2 shows these three cases.
For each of the three different population models in table 8.2, β_1 has a different percentage change interpretation. We don't derive the interpretations of β_1, but instead list them for the three different cases in table 8.2:

- linear-log: a 1% change in X is associated with a $0.01\beta_1$ change in Y.
- log-linear: a change in X of 1 is associated with a $100 \times \beta_1\%$ change in Y.
- log-log: a 1% change in X is associated with a $\beta_1\%$ change in Y. β_1 can be interpreted as an *elasticity*.

8.3.4 A note on R^2

R^2 and $\bar{R}^2$ measure the proportion of variation in the dependent variable (Y) that can be explained using the X variables. When we take the log of Y in the log-linear or log-log model, the variance of Y changes. That is, $\text{Var}[\log Y] \neq \text{Var}[Y]$. We cannot use R^2 or $\bar{R}^2$ to compare models with different dependent variables. That is, we should not use R^2 to decide between two models, where the dependent variable is Y in one, and $\log Y$ in the other.

8.3.5 Log-linear model for the CPS data

It is common to use the log of *wage* as the dependent variable, instead of just *wage*. This allows for the factors that determine differences in wages to be associated with approximate percentage changes in *wage*. In the following, we'll see an example of a log-linear model estimated using the CPS data. Start by loading the data:

```
cps <- read.csv("https://rtgodwin.com/data/cps1985.csv")
```

and estimate a log-linear model:

$$\log(wage) = \beta_0 + \beta_1 education + \beta_2 gender + \beta_3 age + \beta_4 experience + \epsilon$$

```
summary(lm(log(wage) ~ education + gender + age + experience, data = cps))
```

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   1.15357     0.69387    1.663     0.097 .
education     0.17746     0.11371    1.561     0.119
genderfemale -0.25736     0.03948   -6.519  1.66e-10 ***
age          -0.07961     0.11365   -0.700     0.484
experience    0.09234     0.11375    0.812     0.417
```

The interpretation of the estimated coefficient on `education`, for example, is that a 1 year increase in *education* is associated with a 17.8% increase in *wage*. The interpretation of the coefficient on the dummy variable `genderfemale` is a bit more tricky. It is estimated that women make $(100 \times (\exp(-0.257) - 1) = -22.7\%)$ 22.7% less than men. For simplicity, however, we can say that women make approximately 25.7% less than men, but you should know that this interpretation is actually wrong.

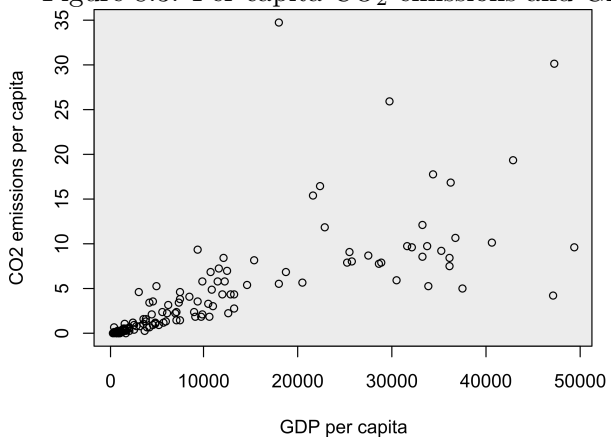
The advantage of using log *wage* as the dependent variable is that it allows the estimated model to capture non-linear effects. The 25.7% decrease in wages for women means that the dollar difference in wages between women and men in high-paying jobs (such as medicine) is larger than the dollar difference in wages between women and men in lower-paying jobs.

8.3.6 Log-log model for CO_2 emissions

In this section, we use data on per capita CO_2 emissions, and GDP per capita (data is from 2007). We will suppose that CO_2 emissions is the *dependent* variable. Load the data, and create the plot:

```
co2 <- read.csv("http://rtgodwin.com/data/co2.csv")
plot(co2$gdp.per.cap, co2$co2,
     ylab = "CO2 emissions per capita", xlab = "GDP per capita")
```

Figure 8.3: Per capita CO_2 emissions and GDP.



Consider this (possibly wrong) population model:

$$CO_2 = \beta_0 + \beta_1 GDP + \epsilon \tag{8.4}$$

As GDP gets larger, CO_2 emissions are all over the place. The problem with model 8.4 is that GDP has the same effect on CO_2 everywhere (for all levels of GDP). Since energy consumption (which produces CO_2 emissions) is a relatively inelastic good, it may be reasonable to think that an increase in GDP per capita of say $1000 has a much bigger impact on CO_2 emissions when GDP per capita is low. That is, their may be a non-linear relationship. If we take the *logs* of CO_2 and GDP per capita, then we are saying that percentage changes in per-capita GDP lead to percentage changes in CO_2:

$$\log(CO_2) = \beta_0 + \beta_1 \log(GDP) + \epsilon \tag{8.5}$$

Plot the data (see Figure 8.4):

```
plot(log(co2$gdp.per.cap), log(co2$co2),
     ylab = "log CO2 emissions per capita", xlab = "log GDP per capita")
```

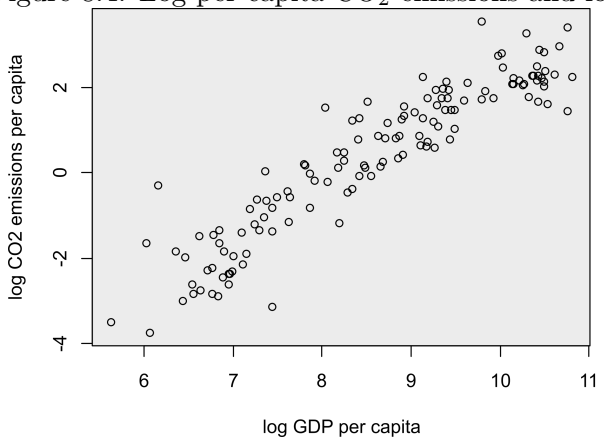
In Figure 8.4, it is much easier to see that there is a *strong* and *positive* relationship between per capita CO_2 emissions and per capita GDP. Now, let's estimate model 8.5:

```
co2mod <- lm(log(co2) ~ log(gdp.per.cap), data = co2)
summary(co2mod)
```

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -9.94045    0.36806  -27.01   <2e-16 ***
log(gdp.per.cap)  1.20212    0.04234   28.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6642 on 132 degrees of freedom
Multiple R-squared:  0.8593,    Adjusted R-squared:  0.8582
F-statistic: 806.1 on 1 and 132 DF,  p-value: < 2.2e-16
```

The interpretation of the results is that for every 1% increase in GDP per capita, it is estimated that CO_2 emissions increase by 1.2%.

Figure 8.4: Log per capita CO_2 emissions and log GDP.



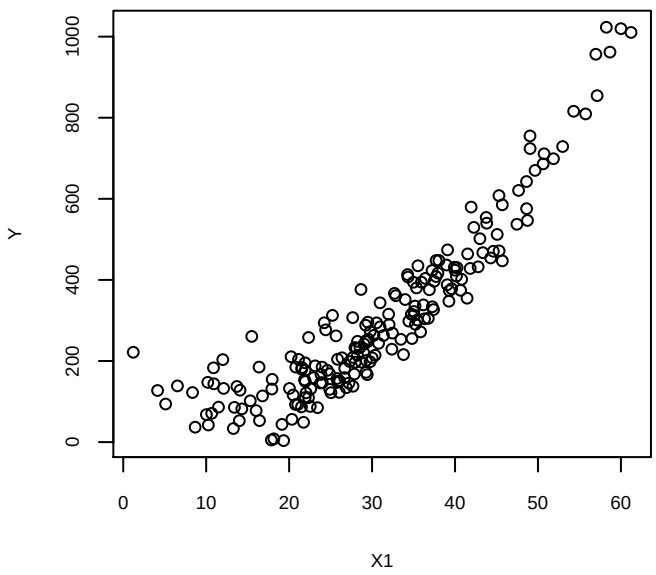
8.4 Review Questions

1. What is a polynomial regression model?
2. Why is it important to have a model that can capture non-linear effects?
3. Use the following in R to load the data necessary for this question:

```
dat <- read.csv("https://rtgodwin.com/data/chap8poly.csv")
```

   a) Plot the data. Which variable might have a non-linear relationship with Y?
   b) Estimate the population model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \beta_4 X_1^4 + \beta_5 X_2 + \epsilon$.
   c) Determine the appropriate degree of the polynomial in X_1 (determine the right r).
   d) What is the estimated effect of X_1 on Y?

Figure 8.5: Question 3, part (a).



4. Other than polynomials, what is another way to capture a non-linear effect in an OLS regression model?
5. What are the interpretations of the βs in population models that use logarithms?
6. Using the diamond data, estimate a linear-log, log-linear, and log-log model. Interpret your results in each case.

8.5 Answers

1. A polynomial regression model is one that includes powers of one or more of the X variables as additional regressors (e.g. X_3^2, X_3^3). This is done in order to approximate a non-linear relationship between the X and Y variables.
2. Many models in economics specify non-linear relationships between the variables. We want our econometric models to represent the features of the economic model. If non-linear relationships are ignored, the OLS estimator may be biased.
3. a) A plot of the data reveals that there is a possible non-linear relationship between X_1 and Y:

```
plot(X1, Y)
```

See figure 8.5. When using `plot(dat$X2, dat$Y)`, the relationship between X_2 and Y looks linear.

b) Estimate the model:

```
summary(lm(Y ~ X1 + I(X1^2) + I(X1^3) + I(X1^4) + X2, data = dat))
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.901e+02  1.809e+01  10.509  < 2e-16 ***
X1          -1.059e+01  3.135e+00  -3.380 0.000878 ***
I(X1^2)      5.076e-01  1.807e-01   2.810 0.005468 **
I(X1^3)     -3.431e-03  4.132e-03  -0.831 0.407262
I(X1^4)      3.141e-05  3.229e-05   0.973 0.331872
X2          -2.015e+00  6.118e-02 -32.944  < 2e-16 ***
```

c) In part (b), X_1^2, X_1^3, and X_1^4 were included in the regression, so that $r = 4$. We may not need to go as high as X_1^4 in order to adequately model the non-linear relationship between X_1

and Y. To determine the appropriate r, we can see if the highest power of X_1 is statistically significant. If not, we drop it from the model, and try again, stopping when the highest power *is* significant.
From the R output in part (b), we see that X_1^4 is "insignificant" (we fail to reject the null hypothesis that $\beta_4 = 0$). This indicates that X_1^4 is not needed in the polynomial, so we drop it from the model:

```
summary(lm(Y ~ X1 + I(X1^2) + I(X1^3) + X2, data = dat))
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.775e+02  1.260e+01  14.081  < 2e-16 ***
X1          -7.870e+00  1.409e+00  -5.586 7.71e-08 ***
I(X1^2)      3.382e-01  4.818e-02   7.020 3.60e-11 ***
I(X1^3)      5.584e-04  4.985e-04   1.120    0.264
X2          -2.023e+00  6.070e-02 -33.326  < 2e-16 ***
```

Now, we test to see if X_1^3 is insignificant (from the output above, it is). Dropping it from the model we get:

```
summary(lm(Y ~ X1 + I(X1^2) + X2, data = dat))
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 188.355857   8.024835   23.47   <2e-16 ***
X1           -9.337920   0.517857  -18.03   <2e-16 ***
I(X1^2)       0.391436   0.007933   49.34   <2e-16 ***
X2           -2.015532   0.060387  -33.38   <2e-16 ***
```

Finally, we see that the highest power of X_1 (now X_1^2) is statistically significant. We cannot drop it from the model. The appropriate degree of the polynomial in X_1 is $r = 2$.

d) In the estimated model

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_1^2 + b_3 X_2$$

one way to interpret the estimated effect of X_1 on Y is to consider specific OLS predicted values. The difficulty in interpretation arises because the effect of X_1 on Y now also depends on X_1^2, so that both b_1 and b_2 must be considered together.
The whole point of using the squared term (X_1^2) is to allow the change in Y due to a change in X_1 to depend on the value of X_1 itself. So, let's consider a change in X_1 of 1 unit, for two different starting values of X_1: 20 and 40.

$$\begin{aligned}\hat{Y}|_{X_1=21} - \hat{Y}|_{X_1=20} &= (-9.338 \times 21 + 0.391 \times 21^2) \\ &\quad - (-9.338 \times 20 + 0.391 \times 20^2) = 6.693\end{aligned}$$

When $X_1 = 20$, the effect of a 1 unit increase in X_1 is to increase Y by 6.693. Let's try for a larger value of X_1:

$$\begin{aligned}\hat{Y}|_{X_1=41} - \hat{Y}|_{X_1=40} &= (-9.338 \times 41 + 0.391 \times 41^2) \\ &\quad - (-9.338 \times 40 + 0.391 \times 40^2) = 22.333\end{aligned}$$

The estimated effect of X_1 on Y is much larger, for larger values of X_1.

4. Besides polynomials, we can also take the logarithms of the X and/or Y variables. Exploiting a property of logarithms that small changes in $\log X$ (or $\log Y$) are approximately equal to percentage changes in X (or Y. This leads the βs in the population regression model to have approximate percentage change interpretations of one variable on another. A percentage change is a non-linear change, since the actual amount of the change depends on the starting value.

5. See table 8.2 for the different population models using logs, and see the following discussion for the interpretations of the βs in the different models.
6. Load the diamond data:

```
diam <- read.csv("https://rtgodwin.com/data/diamond.csv")
```

The linear-log model:

```
summary(lm(price ~ log(carat)))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   8397.4      133.7   62.78   <2e-16 ***
log(carat)    5833.8      172.2   33.87   <2e-16 ***
```

The interpretation is that a 1% increase in *carats* is associated with an increase in *price* of \$58.34.
The log-linear model:

```
summary(lm(log(price) ~ carat, data = diam))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.44488    0.02938  219.40   <2e-16 ***
carat        2.84155    0.04264   66.64   <2e-16 ***
```

The interpretation is that an increase in *carats* of 1 is associated with an increase in price of 284% (it may be more sensible to instead say that a 0.1 increase in *carats* is associated with a 28.4% increase in *price*).
Finally, the log-log model:

```
summary(lm(log(price) ~ log(carat), data = diam))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.12775    0.01440  633.99   <2e-16 ***
log(carat)   1.53726    0.01854   82.92   <2e-16 ***
```

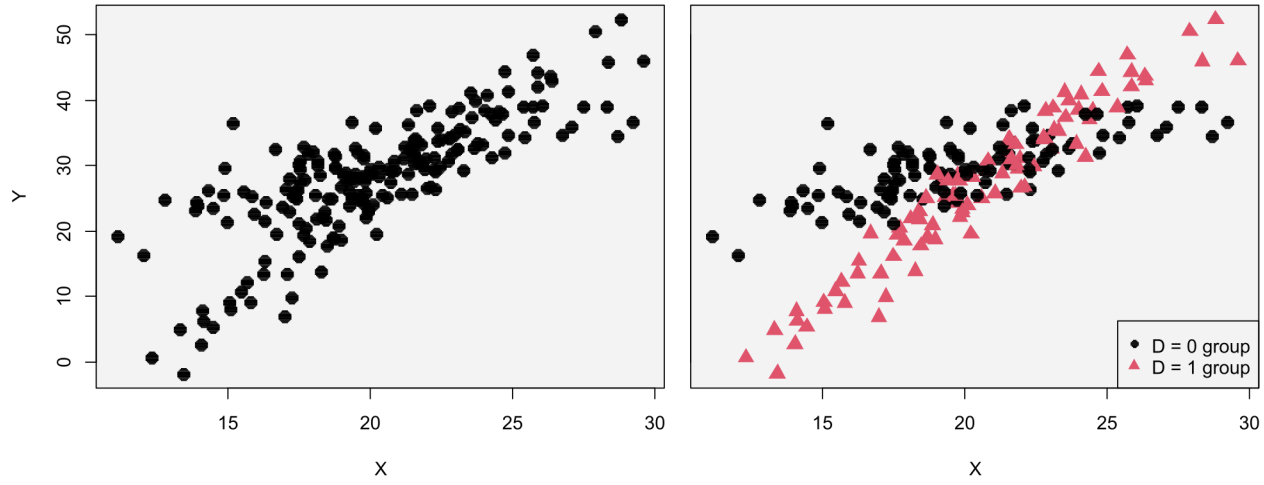
The interpretation is that a 1% increase in *carats* is associated with a 1.53% increase in *price*.

Chapter 9

Interaction terms

Interaction terms can model a type of non-linear effect between variables. They are useful when the effect of X on Y may depend on a different X or D variable. The interaction term $(D \times X)$ allows for a different linear effect between the two groups (the groups defined by D). Both of the variables in the interaction term can be dummy variables $(D_1 \times D_2)$, or both of the variables in the interaction can be continuous $(X_1 \times X_2)$, but the latter situation is somewhat rare and we do not discuss it here.

Figure 9.1: Same data is plotted in both panels. In the right panel, we use a dummy variable D to colour code the data points, revealing that there are separate regression lines for each group.



9.1 Simple example

To illustrate the usefulness of interaction terms, we use a *fake* data set. The variables are:

- Y - the dependent variable
- X - an explanatory variable
- D - a dummy variable

The data is plotted in Figure 9.1. When we use the dummy variable D to colour code the data points, we see that there are two different regression lines. Let's begin by estimating a simple model:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \epsilon \tag{9.1}$$

In R we can use:

```
summary(lm(Y ~ X + D), data=mydata)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.67535    2.00733  -4.820 2.86e-06 ***
X            1.99131    0.09807  20.304  < 2e-16 ***
D           -4.59618    0.72893  -6.305 1.85e-09 ***
```

Results:

- b_0 = -9.68. This is the intercept for the $D = 0$ group.
- b_1 = 1.99. An increase in X of 1 leads to an average increase in Y of 1.99. This is the marginal effect of X on Y.
- b_2 = -4.60. The $D = 1$ group Y values are 4.60 less than the $D = 0$ group, on average. The intercept shifts down by this amount for the $D = 1$ group, so that their intercept is $b_0 + b_2$ = -9.68 - 4.60 = -14.28.

The estimated model is shown in Figure 9.2 (left panel). The $D = 1$ group's regression line is 4.60 lower. We have two different regression lines for the two different groups, but they both have the same slope. We want them to have different slopes!

9.2 Dummy-continuous interaction

Ideally, we would like a separate regression line for the two groups, since the effect of X on Y may differ for the two. We need something new: an *interaction term*. This will allow for two separate marginal effects (slopes) for the two groups.

Dummy-continuous interaction term: When X is a continuous variable and D is a dummy variable, $D \times X$ is a new variable called an *interaction term*. It allows for the effect of X on Y to differ between the two groups defined by the dummy.

Putting the *interaction term* into the model gives us:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) + \epsilon \tag{9.2}$$

where $D \times X$ is the interaction term, and is a new variable that is created by multiplying the other two variables together. To see how model 9.2 allows for two separate lines, consider what the population model is for $D = 0$, and separately for $D = 1$.

Population model for $D = 0$
Let's substitute in the value $D = 0$ into equation 9.2 and get the population model for the first group:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 (0) + \beta_3 (0 \times X) + \epsilon \\ &= \beta_0 + \beta_1 X + \epsilon \end{aligned} \tag{9.3}$$

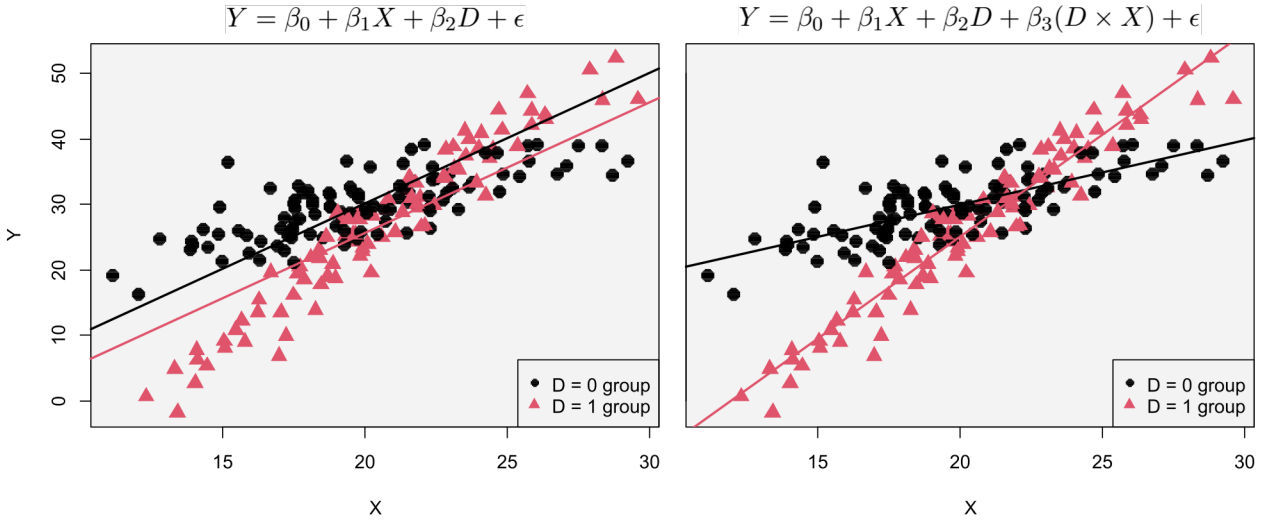
From equation 9.3, we can see that the intercept is β_0 and the slope is β_1.

Population model for $D = 1$
Substituting in the value $D = 1$ into equation 9.2, we get the population model for the other group:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 (1) + \beta_3 (1 \times X) + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X + \epsilon \end{aligned} \tag{9.4}$$

For the $D = 1$ group, the intercept is $\beta_0 + \beta_2$ and the slope is $\beta_1 + \beta_3$. The marginal effect of X on Y differs by β_3 between the two groups.

Figure 9.2: Left panel model (equation 9.1) uses a dummy variable, which allows for a different intercept for the two groups. Right panel model (equation 9.2) uses a dummy variable and an *interaction term*, which allows for a different intercept and *different slope*.



9.2.1 R code for an interaction term

We can include the interaction term by adding the term `I(D * X)` to the `lm()` function:

```
summary(lm(Y ~ X + D + I(D*X)), data=mydata)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.25251    1.73101   5.923  1.4e-08 ***
X             0.98663    0.08581  11.497  < 2e-16 ***
D           -47.61500    2.56503 -18.563  < 2e-16 ***
I(D * X)      2.13132    0.12499  17.052  < 2e-16 ***
```

The estimated value of $b_3 = 2.13$ means that the effect of X on Y (the slope) is 2.13 higher for the $D = 1$ group. That is, the effect of X on Y is 0.99 for $D = 0$, and $(0.99 + 2.13 = 3.12)$ for $D = 1$. The two different regression lines, with the two different slopes, are shown in the right panel of Figure 9.2.

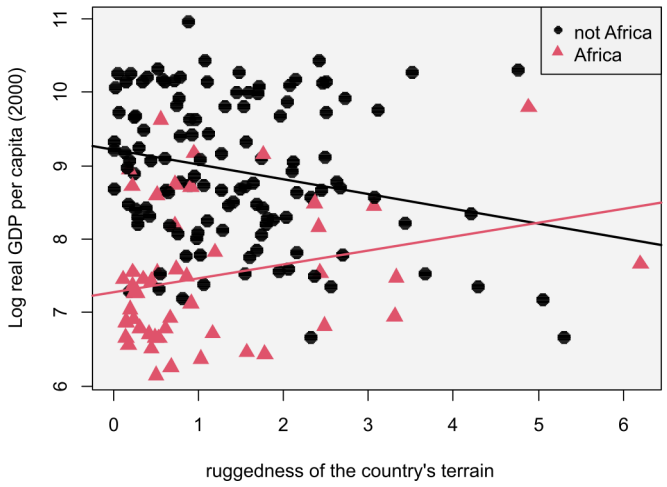
9.2.2 Example: land ruggedness and GDP

This example comes from "Ruggedness: The Blessing of Bad Geography in Africa", by Nunn and Puga (2012). The data is available from the authors here. The main variables in the study, for each of 170 countries, are:

Variable	Description
$\log(GDPpercap)$	Log real GDP per capita from 2000. This is the dependent variable, or y variable.
rugged	A Terrain Ruggedness Index that measures the amount of variation in the elevation of a country. It is a continuous variable. The higher the ruggedness, the more difficult the terrain is to traverse. This is the x variable.
Africa	A dummy variable equal to 1 if the country is in Africa. This is the D variable.

Rugged terrain hinders trade and productive activities, so the higher the ruggedness of a country, the lower the GDP (a negative relationship between x and y). However, the authors argue that the

Figure 9.3: Data is from Nunn and Puga (2012). Log real GDP per capita (from 2000) for 170 countries, and a measure of the ruggedness of the terrain in each country. A model with a dummy variable for African countries, and an interaction term with the dummy and ruggedness, is estimated. The interaction term allows for a different effect of difficult terrain on GDP, depending on whether the country is African or not.



relationship is opposite (positive) for African countries. The rationale is that rugged terrain offered protection from the slave trades. The slave trades hindered future economic development. For African countries, the higher the ruggedness, the higher the GDP.
The population model is:

$$\log(GDPpercap) = \beta_0 + \beta_1 rugged + \beta_2 Africa + \beta_3 (Africa \times rugged) + \epsilon$$

Download the data[1] and use `lm()` with an interaction term `I(cont_africa * rugged)`:

```
rug <- read.csv("https://rtgodwin.com/data/rugged.csv")
mod <- lm(log(rgdppc_2000) ~ rugged + cont_africa + I(cont_africa * rugged), data=rug)
summary(mod)
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.22323    0.13965  66.044  < 2e-16 ***
rugged                -0.20286    0.07739  -2.621  0.00958 **
cont_africa           -1.94805    0.22726  -8.572 6.79e-15 ***
I(cont_africa * rugged)  0.39339    0.13163   2.989  0.00323 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9438 on 166 degrees of freedom
Multiple R-squared:  0.3569,    Adjusted R-squared:  0.3453
F-statistic: 30.71 on 3 and 166 DF,  p-value: 7.595e-16
```

All variables are significant. The estimate -0.20286 means that for every increase in a country's ruggedness of 1, GDP is 20.286% lower on average. But, African countries are *significantly* different. The variable `cont_africa * rugged` allows for the effect of ruggedness to be different between the two groups, and it is significant with a p-value of 0.00323. For African countries, an increase of ruggedness of 1 leads to an *increase* in GDP of $-20.286\% + 39.339\% = 19.053\%$.

[1] As per Nunn and Puga (2012), the missing values for GDP were removed.

9.3 Dummy-dummy interactions

A dummy-dummy interaction is when two different dummy variables are multiplied together, creating a new variable. This new variable allows for an overlap or combination of the two categories. The two dummy variables by themselves allow for the groups to have different means, and the interaction term allows for a combination of the.

Two dummy variables without an interaction

As an example, we will use a version of the CPS data:

```
dat <- read.csv("https://rtgodwin.com/data/twodummies.csv")
```

In this example, the `university` variable is a **dummy** variable which equals to 1 if the individual has a university (BA) degree, and 0 otherwise. The other dummy variable in the data is `female`.

Variable	Description
`wage`	hourly wage of the worker
`female`	= 1 if the individual is female
	= 0 if male
`university`	= 1 if the individual has a university degree
	= 0 if no university degree
`age`	the age of the worker in years

First, we estimate the standard model without the interaction term, with $\log(wage)$ as the dependent variable:

$$\log(wage) = \beta_0 + \beta_1 female + \beta_2 university + \beta_3 age + \epsilon$$

Estimate this in R:

```
summary(lm(log(wage) ~ female + university + age, data = dat))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.016700   0.043388   46.48   <2e-16 ***
female      -0.152178   0.008503  -17.90   <2e-16 ***
university   0.337940   0.008409   40.19   <2e-16 ***
age          0.026435   0.001439   18.37   <2e-16 ***
```

The interpretation of the results is that women make 15% less than men, and that a university degree increases wage by 34%. However, this model does not allow for the possibility that education has a different effect for women than it does for men. There is a difference between men and women, and there is a difference for a university degree, but there is no difference in the effect of university for men vs. women. See Figure 9.4.

Two dummy variables and an interaction term

To allow for education to have a different effect for men than for women, we estimate the model:

$$\log(wage) = \beta_0 + \beta_1 female + \beta_2 university + \beta_3 (female \times university) + \beta_4 age + \epsilon$$

where β_3 is the additional percentage increase in wages for women with an education, versus men with an education. In R, we can do this by:

Figure 9.4: University makes a difference, and gender makes a difference, but there is not a separate difference for university educated women.

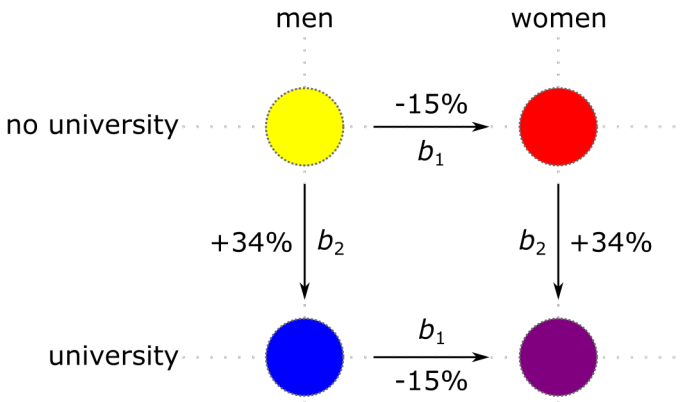
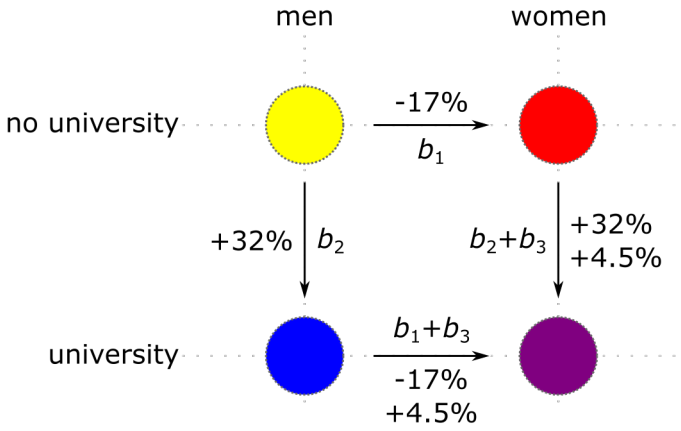


Figure 9.5: With a dummy-dummy interaction term, there is now a difference between all four groups.




```
summary(lm(log(ahe) ~ female + bachelor + I(female * bachelor) + age,
         data = cps))
```

```
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              2.01896    0.04338  46.541  < 2e-16 ***
female                  -0.17347    0.01173 -14.791  < 2e-16 ***
university               0.31895    0.01107  28.809  < 2e-16 ***
I(female * university)   0.04489    0.01704   2.635  0.00842 **
age                      0.02662    0.00144  18.479  < 2e-16 ***
```

It is estimated that women make 17% less than men, that men with a degree make 32% more than men without a degree, and that women with a degree make ($32\% + 4.5\% \approx 36\%$) more than women without a degree. There is a difference for men, a difference for women, and the difference between these two differences is β_3 (4.5%). See Figure 9.5.

9.3.1 Hypothesis tests involving dummy interactions

An important use of dummy interaction terms is to test whether there is a different effect between two groups. In the simple example, the interaction term measures the difference in the slope between the two groups. To test the hypothesis that the slope is the same for both groups, we could test the hypothesis:

$$H_0 : \beta_3 = 0$$
$$H_A : \beta_3 \neq 0$$

in the model:

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) + \epsilon$$

From the regression output from before, we see that the interaction term is highly significant, and we reject the null hypothesis. There is evidence that there is a different marginal effect for the two groups.

Similarly, testing $\beta_3 = 0$ in the model:

$$\log(wage) = \beta_0 + \beta_1 female + \beta_2 university + \beta_3 (female \times university) + \beta_4 age + \epsilon$$

is a test of whether there is a different effect of education for women than for men. From the regression output in the previous section, we see that the p-value for the estimated coefficient on `fem_bach` is 0.00842. We reject the null that there is no difference in the effect of education between men and women at the 1% significance level.

9.3.2 Some additional points

The third possibility, a continuous-continuous interaction term, was left out of the discussion. For example, the returns to education (measured in years as a continuous variable) may diminish as the worker ages (also a continuous variable). To capture this idea, we could multiply these two continuous variables together, and include the product in our regression.

Some of the models presented in this section had dummy variable interaction terms that resulted in completely separate regression functions for the different groups. This complete separation was due to the simplicity of the models. That is, no other variables were included. We can include other variables in the regression as usual. The interaction terms then have the interpretation of a difference between groups, *while controlling for other factors* (*ceteris paribus*).

Finally, the dummy interaction may involve *multiple variables*. This is particularly important when the polynomial regression model is used to capture a non-linear effect. For example, we might have used `education`^{2} as a variable to capture a non-linear effect. Using a dummy interaction with education should then involve both of the variables (`education` and `education`^{2}). A test for no differences between groups would then require the F-test.

9.4 Differences-in-Differences (DiD)

Differences-in-Differences (DiD) is a very popular framework for estimating a causal effect, that relies on the use of a dummy-dummy interaction term. In this section, we will look at a classic example involving minimum wage and employment.

9.4.1 Motivating example: minimum wage increase in New Jersey

Economists David Card and Alan Krueger sought to examine the effect of an impending minimum wage increase by measuring the employment in fast food restaurants before and after the wage increase. Contrary to conventional wisdom, they found an *increase* in employment, using DiD[2].

[2]Card, D., & Krueger, A. B. (1993). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania

Figure 9.6: In the early 90s, conventional economics wisdom held that an increase in the minimum wage would cause employment to decrease (image created by Stable Diffusion).

In 1992, New Jersey's minimum wage was set to increase from \$4.25 to \$5.05 per hour. Card and Krueger surveyed 410 fast-food restaurants before and after the minimum wage increase, recording each restaurant's number of employees before and after the wage increase. Download the data:

```
did <- read.csv("https://rtgodwin.com/data/card.csv")
```

The variables we'll look at in this example are:

Variable	Description
EMP	the number of full-time employees
TIME	= 1 if employment was measured after the minimum wage increase
	= 0 if employment was measured before the minimum wage increase
STATE	= 1 if employment was measured in New Jersey
	= 0 if employment was measured in Pennsylvania

To take a naive approach to estimate the impact of the minimum wage increase on employment, we could take the difference in average employment before and after the wage increase:

$$\bar{EMP}|_{STATE=1,TIME=1} - \bar{EMP}|_{STATE=1,TIME=0} = 0.47$$

which in R can be obtained using:

```
mean(did$EMP[did$STATE == 1 & did$TIME == 1])
       - mean(did$EMP[did$STATE == 1 & did$TIME == 0])
```

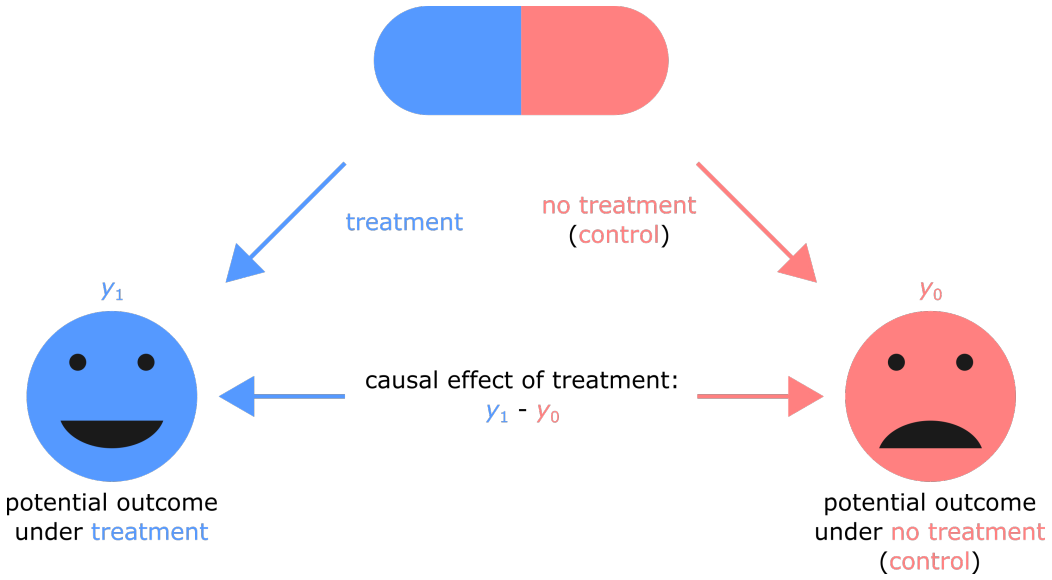
```
[1] 0.4666667
```

From our earlier discussion on dummy variables in Section 5.3, we know that we can obtain the same result using least-squares and a dummy variable, by estimating the model:

$$EMP = \beta_0 + \beta_1 TIME + \epsilon$$

which in R can be obtained from:

Figure 9.7: The fundamental problem of causal inference. The causal effect of a treatment can never be observed, because we can't observe both the outcome under *treatment*, and under *no treatment*.




```
dids <- subset(did, STATE == 1)
summary(lm(EMP ~ TIME, data=dids))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.4306     0.5289  38.627   <2e-16 ***
TIME          0.4667     0.7480   0.624    0.533
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.298 on 616 degrees of freedom
Multiple R-squared:  0.0006315, Adjusted R-squared:  -0.0009909
F-statistic: 0.3892 on 1 and 616 DF,  p-value: 0.5329
```

From the above output, we see that there has been a small increase in employment over time, but that the increase is not statistically significant (the p-value is 0.533).

It might be tempting to attribute the change in employment over time to the increase in minimum wage, and call it the causal effect. But this is tough to justify. The problem is, what if employment was increasing (or decreasing) over time anyway?

9.4.2 Estimating the causal effect of a treatment: the fundamental problem of causal inference

The true causal effect of the minimum wage increase on employment is the difference between what did happen in New Jersey, and what would have happened in New Jersey without the minimum wage increase. We cannot observe both of these "potential" outcomes! This is the *fundamental problem of causal inference*. See Figure 9.7.

Typically, we are interested in estimating the *difference* that a treatment makes. That is we want to know:

$$\mathrm{E}\left[y_1 - y_0\right]$$

where y_1 is the potential outcome with treatment and y_0 is the potential outcome without treatment (only one of these outcomes can be observed!). "Treatment" is broadly defined and doesn't necessarily

refer to treatment with a drug. Some examples of treatments (and outcomes) are:

- Health insurance (y_1 and y_0 the number of visits to the doctor with or without insurance).
- Education (y_1 and y_0 the wage with/without an education).
- A job training program (y_1 and y_0 the employment rate with/without the program).
- Monetary policy.
- Student debt forgiveness.
- Information.
- Increase in the minimum wage (y_1 and y_0 the employment rate).

Because an individual can't be in both states (treated and not treated), we can't observe both y_1 and y_0. We can never observe a causal effect!

- One of the two outcomes will occur, and is factual.
- The other outcome is imagined, or counterfactual.
- We only ever observe either y_1 **or** y_0.
- To *estimate* a causal effect, we must *predict* the missing values. See Table 9.1.

What would it take to truly observe a causal effect? Jeffrey Wooldridge has called the fundamental problem of causal inference a problem of "missing data" (we are missing either y_1 or y_0 for each individual). How could we observe the missing data? In the New Jersey minimum wage example, we might travel back in time and prevent the minimum wage from increasing, so that we could observe the employment level without the minimum wage increase (the missing y_0). Alternatively, perhaps we could find a parallel universe where everything is the same except that New Jersey did not increase their minimum wage in 1992. Barring time travel or observing parallel universes, it seems we have to think in counterfactuals and try to find ways to estimate what the unobserved outcome (y_1 or y_0) would have looked like so that we can calculate $y_1 - y_0$.

Table 9.1: To estimate a causal effect, we need the outcomes under "no treatment" to be similar to the outcomes that the treated individuals would have experienced, had they not been treated. Sometimes we can assume that the average y_0 we observe for the "no treatment" group is the same as we would observe for the treated group (if they instead hadn't been treated). This is often possible in a *randomized controlled experiment*, but with economics it usually difficult to justify this link between the treated and not treated groups.

Unit	Treated	Outcome under treatment (y_1)	Outcome under no treatment (y_0)
1	yes	✓	?
2	yes	✓	?
3	no	?	✓
4	no	?	✓

9.4.3 DiD for the New Jersey data

The naive approach for the New Jersey data was to take the difference between New Jersey's employment before and after the wage increase. But for this to be the causal effect, we would need to assume that the level of employment would have stayed constant over the 9 months in which employment was measured! This is not likely. We need to try to estimate what *would* have happened in New Jersey if there had been now wage increase. We can do this using the DiD estimator, which in this case can be calculated from the sample means for the *four* different groups. See Table 9.4.

Table 9.2: Average employment by `STATE` and `TIME`. These average employment levels are also depicted in Figure 9.8.

	`TIME` = 0	`TIME` = 1	Difference
Pennsylvania `STATE` = 0 (no treatment)	23.380	21.096	-2.283
New Jersey `STATE` = 1 (treatment)	20.431	20.897	0.466
Difference	-2.949	-0.199	**2.750**

- We'll make the **parallel trends** assumption: the *difference* in employment that occurred for the control group would have also occurred for the treatment group (if they hadn't have been treated): -2.283. This is our estimated y_0, the employment outcome that would have occurred if the minimum wage increase hadn't have happened (this is the *counterfactual*).
- The *difference* in employment that actually did occur under treatment was 0.466. This is our factual y_1.
- The *difference-in-difference* (DiD) is 0.466 – (-2.283) = **2.750**.

We got the DiD estimator above by differencing sample means between four groups. But often, we want to include other "X" variables in the model in order to avoid OVB. In such cases, we can use LS, and if we estimate the model:

$$EMP = \beta_0 + \beta_1 TIME + \beta_2 STATE + \beta_3 (TIME \times STATE) + \epsilon \tag{9.5}$$

then b_3 is the DiD estimator!

- Other "X" variables can be added to the model.
- $TIME \times STATE$ is an **interaction term**.
- β_1 is the difference in EMP over $TIME$, for the no treatment group.
- β_2 is the difference in EMP between the two states, at $TIME = 0$.
- β_3 is the difference in the effect of $TIME$ between the two states: the difference-in-difference.

Table 9.3: To find the interpretation of β_3, we can "plug" in the four possible combinations for the dummy variables into model 9.5.

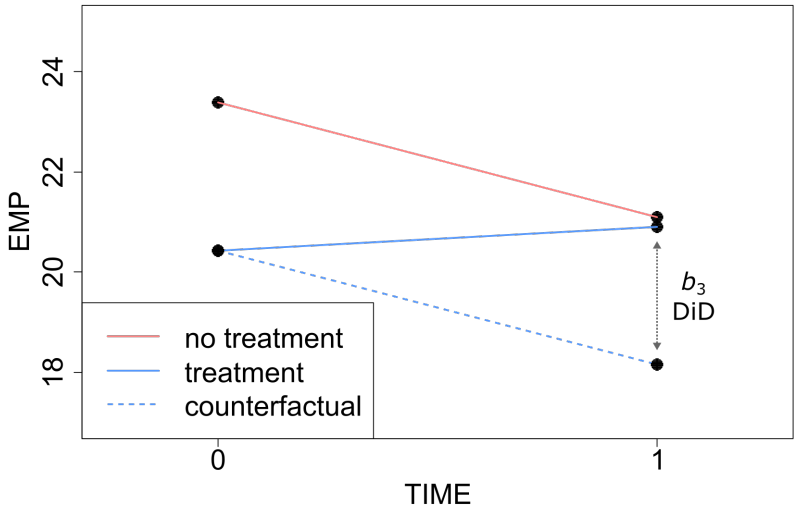
	`TIME` = 0	`TIME` = 1	Difference
Pennsylvania `STATE` = 0 (no treatment)	β_0	$\beta_0 + \beta_1$	β_1
New Jersey `STATE` = 1 (treatment)	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_1 + \beta_3$
Difference	β_2	$\beta_2 + \beta_3$	β_3

From Table 9.3 notice that:

- Difference over time for the no treatment group: β_1
- Difference over time for the treatment group: $\beta_1 + \beta_3$
- The difference-in-difference (DiD): $(\beta_1 + \beta_3) - \beta_1 = \beta_3$

Finally, to estimate model 9.5 using LS and get the DiD estimator in R, we can use:

Figure 9.8: Average employment before and after the minimum wage increase, by state. The DiD estimator is also the LS estimator b_3 from equation 9.5.




```
summary(lm(EMP ~ TIME + STATE + I(TIME * STATE), data = did))
```

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       23.380      1.098  21.288   <2e-16 ***
TIME              -2.283      1.553  -1.470   0.1419
STATE             -2.949      1.224  -2.409   0.0162 *
I(TIME * STATE)    2.750      1.731   1.588   0.1126
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.511 on 764 degrees of freedom
Multiple R-squared:  0.007587,  Adjusted R-squared:  0.00369
F-statistic: 1.947 on 3 and 764 DF,  p-value: 0.1206
```

The DiD estimate is 2.75.

9.5 Review Questions

1. Describe the usefulness of interaction terms.
2. Using the CPS data, determine if there is a different effect of *education* on *wage*, between men and women.
3. Winnipeg is the so-called “Slurpee capital of the world”. In the year 2028, sky-rocketing diabetes rates has led the city to impose a Slurpee tax, in the hopes of reducing consumption. In August 2027, Slurpee consumption in 45 different 7-11 stores is recorded. The Slurpee tax happens in January 2028, and Slurpee consumption for the same 45 stores is recorded again in August 2028. Slurpee consumption for a “control” group is also recorded. Slurpee consumption for Edmonton in 45 different 7-11 stores is recorded for both August 2027 and August 2028, but there was no Slurpee tax in Edmonton. Download the dataset:

   ```
   dat <- read.csv("https://rtgodwin.com/data/slurpee.csv")
   ```

   The variables in the data are:

Variable	Description
`slurpee`	the number of Slurpees sold in a store over the month
`time`	= 1 if after the Slurpee tax (August 2028)
	= 0 if before the Slurpee tax (August 2027)
`winnipeg`	= 1 if in Winnipeg (treatment group)
	= 0 if in Edmonton (no-treatment group)

a) What is the difference in average Slurpee consumption in Winnipeg, before and after the tax?
b) Did the tax *cause* this increase in Slurpee consumption?
c) What is the "fundamental problem of causal inference"?
d) What is a "counterfactual"?
e) What is the difference in Slurpee consumption in Edmonton, over the year?
f) What is the "parallel trends" assumption?
g) What is the average amount of Slurpees that would have been sold in Winnipeg, if there had been no Slurpee tax?
h) What is the estimated difference-in-difference (DiD)?
i) Fill out all of the averages and differences in a table, similar to Table 9.4.
j) Use least-squares and an interaction term to get the DiD estimate.
k) Sketch Slurpee consumption for both cities, before and after the tax. Sketch the counterfactual Slurpee consumption, and label the estimated DiD.

9.6 Answers

1. Interaction terms are useful when we want to allow the effect of X on Y to depend on a different X variable. When one variable in the interaction term is a continuous variable, and the other is a dummy, the interaction term allows for a different marginal effect for the two different groups (as defined by the dummy).
   When both variables in the interaction term are dummies, we are able to estimate a "difference-in-difference". In both cases, interaction terms allow us to estimate, and test for, differences between groups.
2. Load the CPS data:

```
cps <- read.csv("http://rtgodwin.com/data/cps1985.csv")
```

We'll introduce an interaction term into our population model:

$$\log wage = \beta_0 + \beta_1 education + \beta_2 female + \beta_3 age + \beta_4 experience + \beta_5 education \times female + \epsilon$$

To estimate this model in R, we can use:

```
summary(lm(log(wage) ~ education + gender + age + experience
         + gender*education, data = cps))
```

```
Coefficients:
                       Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)             1.23263     0.69231    1.780  0.075576 .
education               0.14950     0.11402    1.311  0.190364
genderfemale           -0.69499     0.20315   -3.421  0.000672 ***
age                    -0.06472     0.11345   -0.570  0.568616
experience              0.07754     0.11355    0.683  0.494959
education:genderfemale  0.03362     0.01531    2.196  0.028545 *
```

The estimated difference is that an additional year of education increases wages by 3.36% more for women than for men (note that the dependent variable is $\log wage$. To test to see if this difference is *insignificant* we test the null hypothesis that the coefficient on the interaction term is equal to zero ($H_0 : \beta_5 = 0$). R has already performed this test for us: the associated p-value is 0.0286. We reject the null hypothesis that there are no differences in the effect of education on wages between men and women, at the 5% significance level.

3. a) What is the difference in average Slurpee consumption in Winnipeg, before and after the tax?

We need to subtract average `slurpee` in Winnipeg at `time = 0` from the average at `time = 1`:

```
mean(dat$slurpee[dat$time==1 & dat$winnipeg==1]
    ) - mean(dat$slurpee[dat$time==0 & dat$winnipeg==1])
```

```
[1] 616.2
```

Slurpee consumption increased by 616.2 on average, even after the tax was introduced!

---

b) Did the tax *cause* this increase in Slurpee consumption?

We cannot say that the tax caused this difference in Slurpee consumption, because other things that effect Slurpee consumption may have happened over time. We would need to know what Slurpee consumption would have been in August 2028 *if there had been no tax*, to estimate the causal effect of the tax. Unless Slurpee consumption would have been the same in August 2028 without the tax, as it was in 2027, we cannot call the difference in part (a) the causal effect.

---

c) Explain the "fundamental problem of causal inference" using this example.

The fundamental problem of causal inference is that we cannot observe Slurpee consumption in Winnipeg under both "treatment" and "no treatment". We cannot observe what would have happened in Winnipeg in 2028 if there had been no tax. Observing the causal effect of the tax is impossible.

---

d) What is the "counterfactual" that we need to estimate the causal effect of the tax?

We observe what happens in Winnipeg after the tax (these are facts). We need to imagine, or *predict*, what would have happened in Winnipeg if there had been no tax. This is called a "counterfactual". Comparing the two outcomes estimates the causal effect.

---

e) What is the difference in Slurpee consumption in Edmonton, over the year?

Similar to before:

```
mean(dat$slurpee[dat$time==1 & dat$winnipeg==0]
    ) - mean(dat$slurpee[dat$time==0 & dat$winnipeg==0])
```

```
[1] 1225.956
```

Slurpee consumption increased by a lot!

---

f) What is the "parallel trends" assumption?

The "parallel trends" assumption is where we assume that the trend observed in the "no treatment" group is the same trend that we would have observed in the "treatment" group, if they hadn't been treated. In this example, we will assume that the Slurpee market in Edmonton is faces a similar situation to Winnipeg, over the year period. We assume that the increase in Slurpee consumption in Edmonton is the same that would have happened in Edmonton, if there had been no tax.

---

g) What is the average amount of Slurpees that would have been sold in Winnipeg, if there had been no Slurpee tax?

Using the *parallel trends* assumption, we can *predict* that Winnipeg's Slurpee consumption would have also increased by 1225.956 on average.

---

h) What is the estimated difference-in-difference (DiD)?

We take the difference between Winnipeg's difference in average Slurpee consumption, and Edmonton's difference in average Slurpee consumption.

$$616.2 - 1225.956 = -609.756$$

This is the estimated effect of the tax! Slurpee consumption decreased by 609.756 on average in August 2028, compared to what it would have been in the absence of the tax.

---

i) Fill out all of the averages and differences in a table, similar to Table 9.4.

Table 9.4: Average Slurpee consumption in Winnipeg and Edmonton, before and after a tax in Winnipeg.

	`time` = 0	`time` = 1	Difference
Edmonton `winnipeg` = 0 (no-treatment)	4103.089	5329.044	1225.956
Winnipeg `winnipeg` = 1 (treatment)	4206.844	4823.044	616.2
Difference	103.755	-506	**-609.756**

---

j) Use least-squares and an interaction term to get the DiD estimate.

We can estimate a model that uses an *interaction* term in order to get the DiD estimator:

$$slurpee = \beta_0 + \beta_1 winnipeg + \beta_2 time + \beta_3 (winnipeg \times time) + \epsilon$$

In R:

```
summary(lm(slurpee ~ winnipeg + time + winnipeg*time, data = dat))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    4103.09      11.18 367.092  < 2e-16 ***
winnipeg        103.76      15.81   6.564 5.67e-10 ***
```

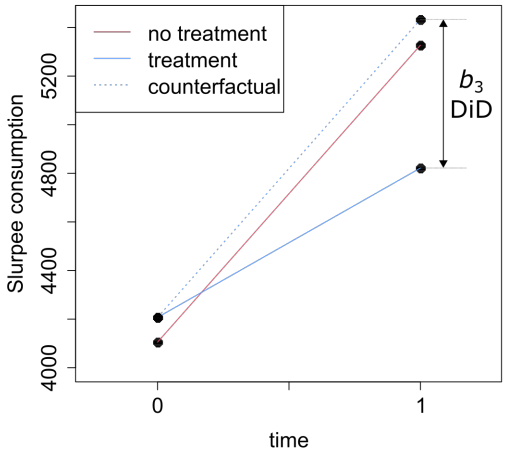
```
time            1225.96      15.81   77.557   < 2e-16 ***
winnipeg:time   -609.76      22.35  -27.277   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 74.98 on 176 degrees of freedom
Multiple R-squared:  0.9781,    Adjusted R-squared:  0.9777
F-statistic:  2620 on 3 and 176 DF,  p-value: < 2.2e-16
```

The DiD estimator is -609.76.

---

k) Sketch Slurpee consumption for both cities, before and after the tax. Sketch the counterfactual Slurpee consumption, and label the estimated DiD.

Figure 9.9: Average Slurpee consumption in Winnipeg and Edmonton.



Chapter 10

Heteroskedasticity

The estimators that we have used so far have good statistical properties provided that the following assumptions hold:

A1 The population model is linear in the βs.
A2 There is no perfect multicollinearity between the X variables.
A3 The random error term, ϵ, has mean zero.
A4 ϵ is identically and independently distributed.
A5 ϵ and X are independent.
A6 ϵ is Normally distributed.

These assumptions ensure that OLS is unbiased, efficient, and consistent, and that hypothesis testing is valid. A violation of one or more of these assumptions might lead us to estimators beyond OLS. OLS is simple, and easy to use, but is often thought of a starting point in econometric modelling since the above assumptions are often unreasonable.

In this section, we will consider that assumption A4 is violated in a particular way. Specifically, we consider what happens where the error term, ϵ, is *not* identically distributed.

10.1 Homoskedasticity

If assumption A4 is satisfied, then ϵ is identically distributed. This means that all of the ϵ_i have the same variance. That is, all of the random effects that determine Y, outside of X, have the same dispersion. The term *homoskedasticity* (same dispersion) refers to this situation of identically distributed error terms.

Stated mathematically, homoskedasticity means:

$$\text{Var}[\epsilon_i|X_i] = \sigma^2 \ , \ \forall i$$

The variance of ϵ is constant, even conditional on knowing the value of X.

Homoskedasticity means that the squared vertical distance of each data point from the (population or estimated) line is, on average, the same. The values of the X variables do not influence this distance (the variance of the random unobservable effects are not determined by any of the values of X). See figure 10.2.

10.2 Heteroskedasticity

Heteroskedasticity refers to the situation where the variance of the error term ϵ is not equal for all observations. The term heteroskedasticity means *differing dispersion*. Mathematically:

$$\text{Var}[\epsilon_i|X_i] \neq \sigma^2 \ , \ \forall i$$

Figure 10.1: Possible heteroskedasticity in the CPS data. The variance in `wage` may be increasing as `education` increases. The reasoning is that individuals who have not completed highschool (or university) are precluded from many high-paying jobs (doctors, lawyers, etc.). However, having many years of education does not preclude individuals from low-paying jobs. The spread in wages is higher for highly educated individuals.

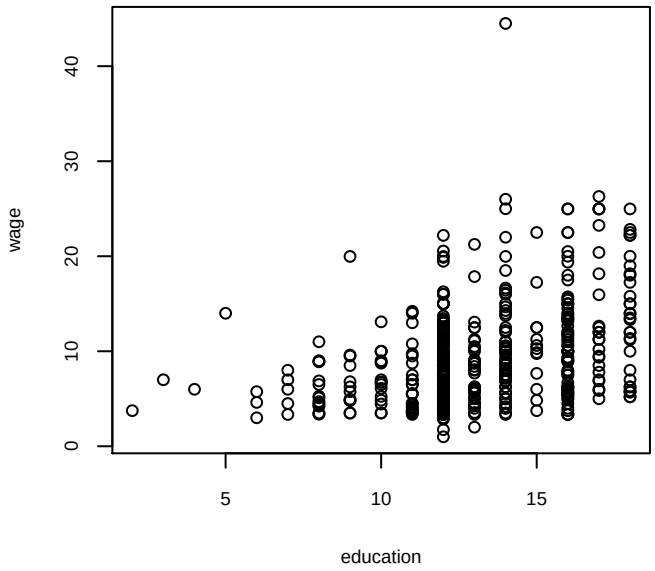


Figure 10.2: Homoskedasticity. The average squared vertical distance from the data points to the OLS estimated line is the same, regardless of the value of X.

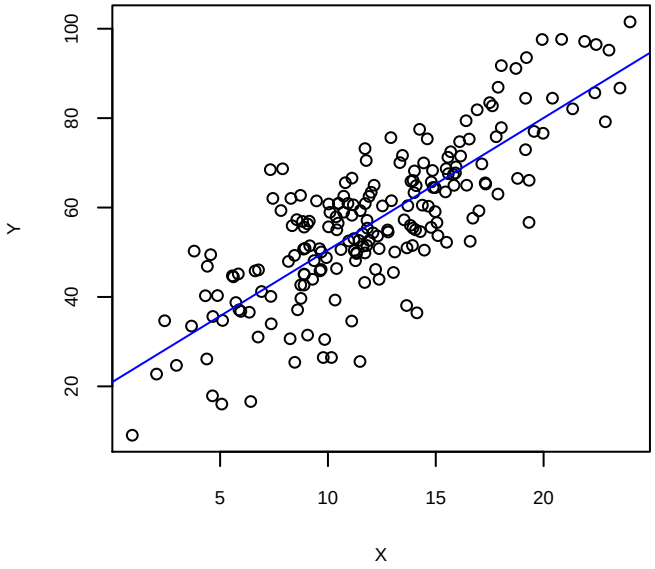
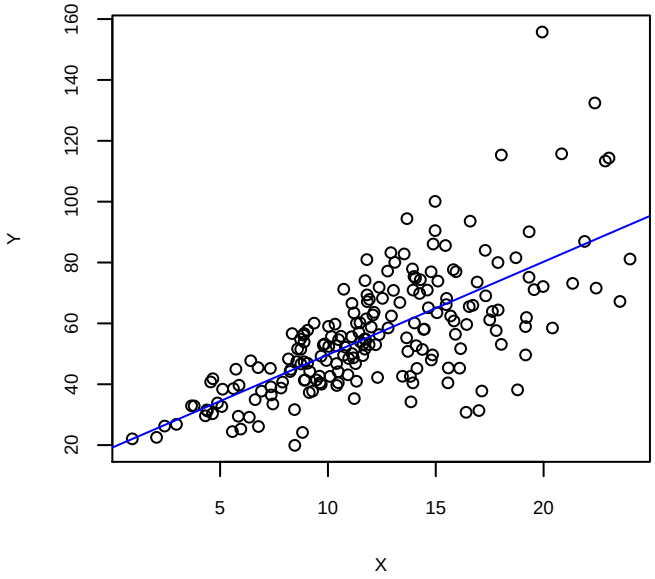


Figure 10.3: Heteroskedasticity. The squared vertical distance of a data point from the OLS estimated line is influenced by X.



or

$$\text{Var}[\epsilon_i|X_i] = \sigma_i^2$$

Each observation can have its own variance, and the value of X may influence this variance.

Heteroskedasticity means that the squared vertical distance of each data point from the estimated regression line is not the same on average, and may be influenced by one or more of the X variables. See figure 10.3, where the larger the value of X is, the larger the variance of ϵ.

10.2.1 The implications of heteroskedasticity

Heteroskedasticity is a violation of A.4, since each ϵ_i is not identically distributed. Heteroskedasticity has two main implications for the estimation procedures we have been using in this book:

(i) The OLS estimator is no longer efficient.
(ii) The estimated standard errors are inconsistent.

The inefficency of OLS is arguably a smaller problem than the inconsistency of the variance estimator. (ii) means that the estimated standard errors in our regression output are wrong, leading to the incorrect t-statistics and confidence intervals. Hypothesis testing, in general, is invalid. The problem arises because the formula that is the basis for estimating the standard errors in OLS (equation 5.7):

$$\text{Var}\left[b_1\right] = \frac{\sigma_\epsilon^2}{\sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}}$$

is only correct under homoskedasticity.

10.2.2 Fixing heteroskedasticity - robust standard errors

To fix problem (i), the inefficiency of OLS, we must use a different estimator, such as Generalized Least Squares (GLS). GLS is not discussed here. To fix (ii), the more important problem of the inconsistency of the standard errors, the formula for $\text{Var}\left[b_1\right]$ must be updated to take into account the possibility of heteroskedasticity.

Updating the formula to allow for heteroskedasticity in the estimation of the standard errors gives what is typically referred to as *robust standard errors*. In R, we will use the code:

```
install.packages("sandwich")
library(sandwich)
install.packages("lmtest")
library(lmtest)
```

to install and load a package that can estimate the robust standard errors, and then use

```
coeftest(my.lm.model, vcov = vcovHC(my.lm.model, "HC1"))
```

to estimate the correct standard errors and updated t-statistcs and p-values, where `my.lm.model` is the least-squares regression that we have estimated using the `lm()` command.

10.2.3 Testing for heteroskedasticity

There are several (approximately) equivalent tests for heteroskedasticity, but we'll focus on the most famous: White's[1] test.

In White's test, the null hypothesis is that there is homoskedasticity, and the alternative is heteroskedasticity. That is:

$$H_0 : \text{var}[\epsilon_i] = \sigma^2$$
$$H_A : \text{var}[\epsilon_i] \neq \sigma^2$$

Take a simple population model with two regressors. Remember that the population model and the estimated model are (respectively):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$
$$y = b_0 + b_1 x_1 + b_2 x_2 + e$$

The residual e is the counterpart to the unobservable error term ϵ! Sometimes, we can use the residuals to test assumptions or properties of the error term. For example, we can look to see if the *residuals* are homoskedastic or heteroskedastic, in order to infer those propertied about the error term. That is, if e looks homoskedastic, we will conclude that so is ϵ.

White's test tries to explain differences in the size of the squared residuals from a least-squares model by regressing them on the original x variables, and the squares and cross products of the x. If the R^2 from this regression is high, then we conclude that there is some pattern to the size of the residuals, and reject the null hypothesis of homoskedasticity.

To test for heteroskedasticity in the population model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

we would estimate it by LS, for example by using `lm(y ~ x1 + x2)`. We then get the squared residuals from this regression, and estimate the following equation by LS:

$$e^2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 (x_1 \times x_2) + \beta_4 x_1^2 + \beta_5 x_2^2 + \varepsilon \tag{10.1}$$

Equation 10.1 is looking for any approximate way to explain variation in the size of the squared residuals. If the estimated model from equation 10.1 fits well (in terms of the R-squared), then there is some explanation for the variance in the error term, and the error term is heteroskedastic. White's test statistic is the nR^2 from this auxiliary regression, and the p-value for the test comes from the Chi-square distribution. As usual, if the p-value is small, we reject the null of homoskedasticity, in favour of heteroskedasticity.

To test for heteroskedasticity in R, we need to install and load a package:

[1] White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica: journal of the Econometric Society, 817-838.

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

and then use:

```
white(my.lm.model, interactions = TRUE)
```

where `my.lm.model` is the model we have estimated by LS. If we find heteroskedasticity, then we need to use heteroskedastic robust standard errors (such as White's standard errors).

10.2.4 Heteroskedasticity in food expenditure data

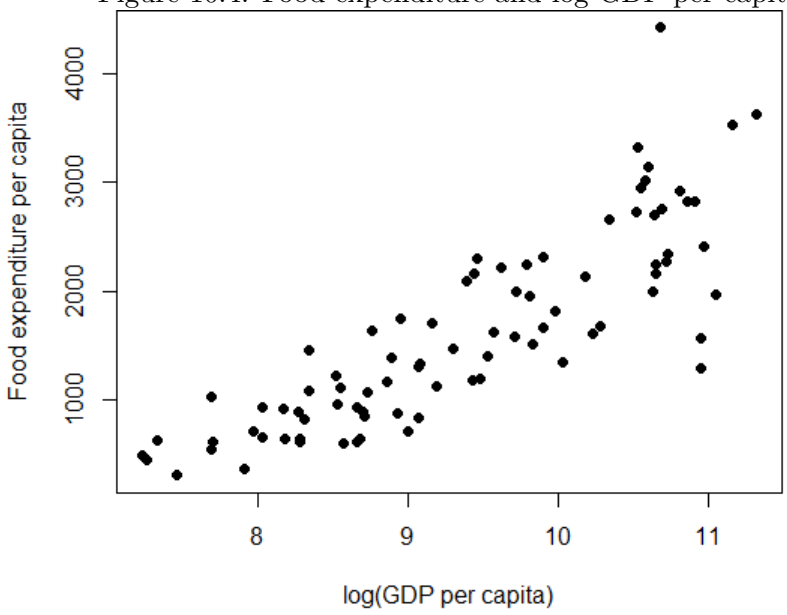
Download a data set on food expenditure by country, in 2016:

```
food <- read.csv("https://rtgodwin.com/data/foodexp.csv")
```

The variables are `foodexp` - food expenditure per capita (in US dollars), and `GDPpercap` - GDP per capita. There are 84 countries in the sample. Plot the data, taking the log of GDP per capita (see Figure 10.4):

```
plot(log(food$GDPpercap), food$foodexp, pch=16, xlab="log(GDP per capita)",
     ylab="Food expenditure per capita")
```

Figure 10.4: Food expenditure and log GDP per capita.



Estimate the population model

The following model for food expenditure:

$$foodexp = \beta_0 + \beta_1 \log(GDPpercap) + \epsilon$$

can be estimated in R using:

```
food.mod <- lm(foodexp ~ log(GDPpercap), data=food)
summary(food.mod)
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    -4737.68     451.38  -10.50   <2e-16 ***
log(GDPpercap)   677.40      47.81   14.17   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 479 on 82 degrees of freedom
Multiple R-squared:   0.71,     Adjusted R-squared:  0.7065
F-statistic: 200.8 on 1 and 82 DF,  p-value: < 2.2e-16
```

Test for heteroskedasticity

If heteroskedasticity is present in this data, then the standard errors, t-statistics, and p-values, are all wrong! Hypothesis testing, and any conclusions we draw, may be incorrect due to the heteroskedasticity. To test for heteroskedasticity, we can use White's test:

```
install.packages("skedastic")
library(skedastic)
white(food.mod)
```

```
  statistic p.value parameter method       alternative
      <dbl>   <dbl>     <dbl> <chr>        <chr>
1      11.6 0.00304         2 White's Test greater
```

The test statistic from the White test is 11.6, with an associated p-value of 0.00304. We reject the null hypothesis of homoskedasticity. To see what the function `white()` is doing, we'll calculate the White test statistic and p-value "by hand":

```
food.resid.sq <- food.mod$residuals ^ 2
summary(lm(food.resid.sq ~ log(GDPpercap) + I(log(GDPpercap) ^ 2), data=food))
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          4703163    3680579   1.278    0.205
log(GDPpercap)      -1121179     795218  -1.410    0.162
I(log(GDPpercap)^2)    67703      42508   1.593    0.115

Residual standard error: 444800 on 81 degrees of freedom
Multiple R-squared:  0.138,     Adjusted R-squared:  0.1167
F-statistic: 6.485 on 2 and 81 DF,  p-value: 0.002442
```

The test statistic is $nR^2 = 84 \times 0.138 = 11.6$ (same as from the `white()` command). The p-value can be found from:

```
1 - pchisq(84 * 0.138, 2)
```

```
0.003039689
```

which is the same from the `white()` command.

White's heteroskedastic consistent standard errors

To recalculate the standard errors, t-statistics, and p-values, we can use the `coeftest()` function:

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

```
install.packages("lmtest")
library(lmtest)
coeftest(food.mod, vcov = vcovHC(food.mod, "HC1"))
```

```
t test of coefficients:

                 Estimate Std. Error t value  Pr(>|t|)
(Intercept)     -4737.680    476.516 -9.9423 9.705e-16 ***
log(GDPpercap)    677.399     54.069 12.5284 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Notice that the estimated βs have not changed, but that the standard errors have changed, t-statistics, and p-values have changed.

Heteroskedastic errors have a pretty severe consequence; hypothesis testing may be invalid. The prevalence of heteroskedasticity in many economics data has led to the common practice of erring on the side of caution. Heteroskedastic robust standard errors are often used, if heteroskedasticity is suspected. Note that homoskedasticity is a special case of heteroskedasticity, so the downside of using the robust estimator when it is not needed, is small.

10.3 Review Questions

1. Explain the difference between homoskedasticity and heteroskedasticity.
   Answer. Under homoskedasticity, the variance of the error term is constant. Under heteroskedasticity, the variance of the error term can differ by observation, and can depend on the x variables, or on other variables. To state the difference mathematically, see Sections 10.1 and 10.2.
2. Provide an example of heteroskedasticity using data from another chapter.
   Answer. There are many examples throughout the book (heteroskedasticity is quite common). You could use White's test on any of the models that we have estimated, or you can point out how the data in a scatterplot seems to be differently dispersed depending on the value of the x-axis.
3. Describe the problem that heteroskedasticity brings to LS estimation.
   Answer. If we assume that the errors are homoskedastic, when they are actually heteroskedastic, the estimators for the standard errors of the b are *biased* and *inconsistent*. This means that t-statistics, p-values, confidence intervals, will be wrong, and hypothesis testing in general will be invalid. The usual computer output from the `summary()` command will be wrong (but the b themselves are unaffected). As a side note, the LS estimator is *inefficient* under heteroskedasticity, but it is still *unbiased* and *consistent*.
4. Briefly explain how to fix the inconsistency of the standard errors in LS estimation, in the presence of heteroskedasticity.
   Answer. We can use *robust standard errors*. We can update the formula for the variance of b, in order to take account of heteroskedasticity. This gives us a *consistent* estimator for the standard errors.
5. Explain the basic idea behind testing for heteroskedasticity. What are the null and alternative hypotheses?
   Answer. The null hypothesis is homoskedasticity, and the alternative is heteroskedasticity. The strategy is to estimate the model as usual. Get the residuals. Then do an auxiliary (extra) regression, using the squared residuals as the y variables, and all the x variables, their squares, and their cross products, as regressors. If the R^2 from this extra regression is high enough, that means that the size of the residuals can be explained using x, and the null is rejected.

Chapter 11

Instrumental Variables

For least-squares to work well, we need to make a very important assumption about the error term ϵ.

The error term ϵ must be independent from the x variables, or else least-squares is biased and inconsistent.

For example, in the simple model:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

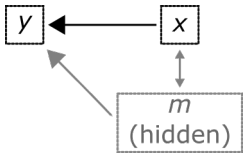
if x is correlated with ϵ then the least-squares estimator for β_1 will be wrong (biased and inconsistent)!

11.1 Missing, lurking, or confounding variables

The error term contains missing variables, that determine y. So, those missing variables need to be uncorrelated with the x variables for LS to work. This is often unreasonable!

A *lurking*, or *confounding* variable is one that threatens our ability to correctly estimate the effect that an x variable has on a y variable. Confounding variables are a major issue in analyses of *causal inference*, and are of tremendous import in many areas, not just economics.

Figure 11.1: A missing m variable that is correlated with x and that determines y will make estimation of the effect of x on y difficult (or impossible).



The situation depicted in Figure 11.1, where m is correlated with both x and y, implies that the effect of x on y cannot be estimated correctly by LS. That is, the estimated β_1 (b_1) is *wrong* in the population model:

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

The reason that b_1 gives the *wrong* answer for the true effect of x on y is that:

- A change in m is associated with a change in both x and y.
- When we "see" x changing, we know m is also changing.
- Attributing changes in y due to changes in x alone becomes impossible, since we don't know how much of the change in y came from m.

The solution to the problem is to include the m variable in the model! If we can't actually observe m (but we can imagine that it is there) then we must use clever strategies and more advanced methods to attempt to estimate the effect of x on y. One of those possible methods is Instrumental Variables (IV) estimation, the focus of this chapter.

11.1.1 House price again

Let's return to the house price data:

```
house <- read.csv("https://rtgodwin.com/data/houseprice.csv")
bad.mod <- lm(Price ~ Fireplaces, data=house)
summary(bad.mod)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    171824       3234   53.13   <2e-16 ***
Fireplaces      66699       3947   16.90   <2e-16 ***
```

This model is suffering from omitted variable bias. The estimated effect of an additional fireplace on house price is wrong (biased and inconsistent). $67,000 is likely not the true effect. This is because there is a missing variable `Living.Area` (the size of the house in square feet), that is correlated with fireplaces and that also determines price. Notice that the missing variable is *inside* the error term (as are all other variables that determine y), but that this missing variable is correlated with x. This means that ϵ and x are correlated, and that least-squares will be biased and inconsistent.

Once we include the missing variable `Living.Area`, the problem is solved:

```
better.mod <- lm(Price ~ Fireplaces + Living.Area, data=house)
summary(better.mod)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 14730.146   5007.563   2.942  0.00331 **
Fireplaces   8962.440   3389.656   2.644  0.00827 **
Living.Area   109.313      3.041  35.951  < 2e-16 ***
```

But what if we can't include the missing variable, because we don't observe it? All hope is not lost. If we can find an *instrument*, then we can still get a consistent estimator for the β.

11.1.2 Endogeneity

- When an x variable is correlated with the error term, that variable is sometimes said to be **endogenous**.
- Simultaneous causality (or just "simultaneity") is another way that we can have endogeneity. We will soon see that this is the case with demand and supply.

11.2 Instrumental variable (IV)

A variable, z, qualifies as an instrument if it satisfies two conditions.

An instrumental variable, z, must be:

1. Correlated with the endogenous variable 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 missing variable m.
    - This is sometimes called the "exclusion" restriction.
    - This restriction cannot be easily tested.

If we can find a valid instrument, then we can use it to extract the "good" or "clean" variation from x. With endogeneity, changes in x are associated with changes in ϵ. But, changes in x **due to** z are not associated with the error term, because z is not correlated with ϵ.

11.3 IV estimation / Two-stage least-squares (2SLS)

Instrumental variables estimation, also called two-stage least-squares, is a statistical method for estimating β_1 in the equation:

$$y = \beta_0 + \beta_1 x + \epsilon,$$

when x is endogenous (correlated to ϵ), but when we have a valid instrument z. The IV estimation gives us a consistent estimator for β_1, whereas LS gives us an inconsistent estimator and should not be used.

11.3.1 1st stage of 2SLS

In the first stage, we estimate an auxiliary regression to *extract* variation from x which is independent from ϵ. The 1st stage regression model is:

$$x = \alpha_0 + \alpha_1 z + u \tag{11.1}$$

After estimating this model by least-squares, we have the estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$. We then use this model to get the least-square predictions for x:

$$\hat{x} = \hat{\alpha}_0 + \hat{\alpha}_1 z \tag{11.2}$$

The LS predicted values $\hat{x}$ from equation 11.2 are independent from the error term! That is, $\hat{x}$ contains changes in x that are due to z only, and since z is uncorrelated with ϵ, so is $\hat{x}$ uncorrelated with ϵ.

11.3.2 2nd stage of 2SLS

In the second stage, we estimate the population model by LS, but instead of using x, we replace it with $\hat{x}$ from the 1st stage. Although x is endogenous, $\hat{x}$ is not! Estimating the following equation by LS gives us the IV estimator:

$$y = \beta_0 + \beta_1 \hat{x} + \epsilon$$

11.3.3 Direct formula for the IV/2SLS estimator

For the model $y = \beta_0 + \beta_1 x + \epsilon$, recall that the formulas for the LS estimators are:

$$b_1 = \frac{\sum \left[(y - \bar{y})(x - \bar{x}) \right]}{\sum (x - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1 \bar{x}$$

Applying these formulas to the 1st stage regression in equation 11.1, the formulas look like:

$$\hat{\alpha}_1 = \frac{\sum \left[(x - \bar{x})(z - \bar{z}) \right]}{\sum (z - \bar{z})^2}$$
$$\hat{\alpha}_0 = \bar{x} - \hat{\alpha}_1 \bar{z}$$

The LS predicted values from the 1st stage are:

$$\hat{x} = \hat{\alpha}_0 + \hat{\alpha}_1 z = \bar{x} - \frac{\sum \left[(x - \bar{x})(z - \bar{z}) \right]}{\sum (z - \bar{z})^2} \bar{z} + \frac{\sum \left[(x - \bar{x})(z - \bar{z}) \right]}{\sum (z - \bar{z})^2} z \tag{11.3}$$

and the LS slope estimator for the model in the 2nd stage is:

$$b_1 = \frac{\sum \left[(y - \bar{y}) \left(\hat{x} - \bar{\hat{x}} \right) \right]}{\sum \left(\hat{x} - \bar{\hat{x}} \right)^2} \tag{11.4}$$

Plugging the predicted values (equation 11.3) into the 2nd stage estimator in equation 11.4) yields the formula for the IV estimator:

$$\hat{\beta}_{IV} = \frac{\sum \left[(y - \bar{y})(z - \bar{z}) \right]}{\sum \left[(x - \bar{x})(z - \bar{z}) \right]} \tag{11.5}$$

11.4 Example of a missing variable: Distance from college

Let's look at data from Card (1993).[1]

- 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 (a missing variable), 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.

The population model that we want to estimate is:

$$wage = \beta_0 + \beta_1 education + \beta_2 urban + \beta_3 gender + \beta_4 ethnicity + \beta_5 unemp + \epsilon \tag{11.6}$$

We are primarily interested in β_1 (the returns to education). The other variables are included as controls, in order to avoid omitted variable bias. The difficulty with estimating equation 11.7 is that education is *endogenous*. From the Card (1993) paper:

> One of the most important "facts" about the labor market is that better-educated workers earn higher wages. Hundreds of studies in virtually every country show earnings gains of 5-15 percent (or more) per additional year of schooling. Despite this evidence, most analysts are reluctant to interpret the earnings gap between more and less educated workers as a reliable estimate of the economic return to schooling. Education levels are not randomly assigned across the population; rather, individuals make their own schooling choices. Depending on how these choices are made, measured earnings differences between workers with different levels of schooling may over-state or under-state the "true" return to education.

11.4.1 LS is the wrong method

First, let's try LS. It is the wrong method to use because it is inconsistent when there is endogeneity. Load the data, and estimate the model:

[1] Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (No. w4483). National Bureau of Economic Research.

```
college <- read.csv("https://rtgodwin.com/data/collegedist.csv")
ls <- lm(wage ~ education + urban + gender + ethnicity + unemp, data = college)
summary(ls)
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.000192   0.156928  50.980   <2e-16 ***
education          0.005369   0.010362   0.518   0.6044
urbanyes           0.070117   0.044727   1.568   0.1170
gendermale         0.085242   0.037069   2.300   0.0215 *
ethnicityhispanic 0.012048   0.062385   0.193   0.8469
ethnicityother     0.556056   0.052167  10.659   <2e-16 ***
unemp              0.133101   0.006711  19.834   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.268 on 4732 degrees of freedom
Multiple R-squared:  0.1098,    Adjusted R-squared:  0.1087
F-statistic: 97.27 on 6 and 4732 DF,  p-value: < 2.2e-16
```

Notice that the returns to education are estimated to be very small (an additional year of education leads to an increase in wage of half of a cent per hour). No point in going to school! But we know that LS is wrong (inconsistent) if *education* is correlated with the error term.

11.4.2 2SLS using distance from college as an IV

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
- *distance from college* is uncorrelated with the missing variables that simultaneously determine education and wage

1st stage

To use this *distance from college* variable in two-stage least-squares, we first regress *education* (the problem endogenous variable) on *distance from college* and all the controls. Then we save the LS predicted values from this regression:

```
first.stage <- lm(education ~ urban + gender + ethnicity + unemp + distance,
                  data = college)
education.hat <- first.stage$fitted.values
```

2nd stage

Now, we estimate the original population model in equation 11.7 using LS, but we replace the *education* variable with 1st stage predicted values $\widehat{education}$. That is, we estimate the equation:

$$wage = \beta_0 + \beta_1 \widehat{education} + \beta_2 urban + \beta_3 gender + \beta_4 ethnicity + \beta_5 unemp + \epsilon \tag{11.7}$$

The R code is:

```
iv <- lm(wage ~ education.hat + urban + gender + ethnicity + unemp,
         data = college)
summary(iv)
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.657025   1.358890  -0.484  0.62876
education.hat       0.647099   0.100592   6.433 1.38e-10 ***
urbanyes            0.046144   0.044691   1.033  0.30188
gendermale          0.070753   0.036978   1.913  0.05576 .
ethnicityhispanic -0.124051   0.065641  -1.890  0.05884 .
ethnicityother      0.227240   0.072984   3.114  0.00186 **
unemp               0.139163   0.006748  20.622  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.263 on 4732 degrees of freedom
Multiple R-squared:  0.1175,    Adjusted R-squared:  0.1163
F-statistic:   105 on 6 and 4732 DF,  p-value: < 2.2e-16
```

The return to education is now positive and significant!

Under LS the estimated returns to education are 0.005, but under IV they are 0.647.

11.4.3 Using the direct formula: `ivreg()`

We can use a direct formula like in equation 11.5 to get the IV estimates (instead of using the two-stage approach). Install and load the `ivreg` package:

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

We need to give the `ivreg()` function:

- the population model that we want to estimate
- the list of instruments that we will use

The population model and list of instruments are separated by `|`:

```
iv <- ivreg(wage ~ education + urban + gender + ethnicity + unemp |
                 distance + urban + gender + ethnicity + unemp,
          data = college)
summary(iv)
```

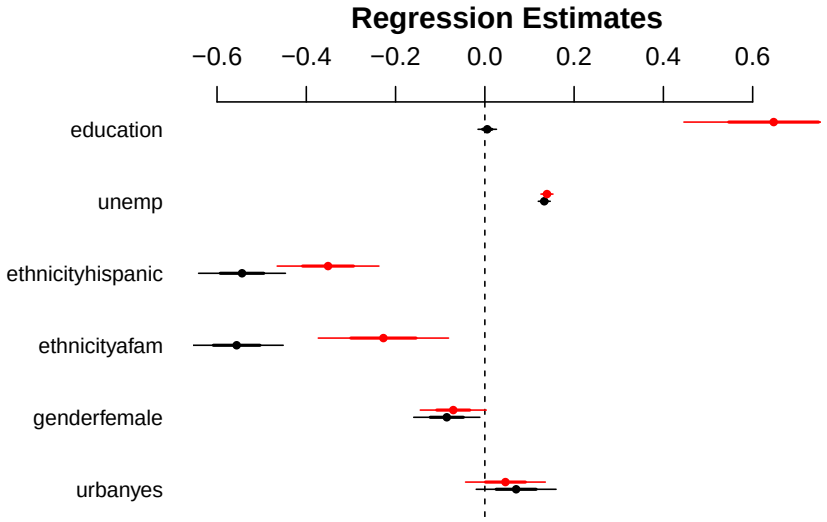
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        -0.65702    1.83641  -0.358   0.7205
education           0.64710    0.13594   4.760 1.99e-06 ***
urbanyes            0.04614    0.06039   0.764   0.4449
gendermale          0.07075    0.04997   1.416   0.1569
ethnicityhispanic -0.12405    0.08871  -1.398   0.1621
ethnicityother      0.22724    0.09863   2.304   0.0213 *
unemp               0.13916    0.00912  15.259  < 2e-16 ***
```

We get the same results as from using the two-stage method!

11.5 Estimating demand with IV

We have tried to estimate a demand curve several times in this course. We have been doing it wrong! This is because the *price* variable that we have been using as a regressor (on the RHS of the model) is endogenous! The price and quantity values that we observe in our data set are actually due to the

Figure 11.2: Results of LS and IV (in red) regression using Card (1993) data. Dependent variable is *wage*; *distance from college* is an instrument for *education*. Horizontal lines are 95% and 99% confidence intervals. Notice that the returns to education are insignificant under LS, but significant under IV.



intersection of demand and supply. The price and quantity values that we observe are due to *two* equations, demand and supply:

$$\begin{array}{rcccccccll} q & = & \alpha_0 & + & \alpha_1 p & + & \alpha_2 s & + & \varepsilon & \text{(supply)} \\ q & = & \beta_0 & + & \beta_1 p & + & \beta_2 d & + & \epsilon & \text{(demand)} \end{array} \tag{11.8}$$

where:

- q is *both* quantity demanded and supplied
- p is price
- d are "demand-shifters" (such as income, prices of complements and substitutes, etc.)
- s are "supply-shifters" (such as prices of inputs, weather, etc.)
- α_1 should be $(+)$ and β_1 should be $(-)$

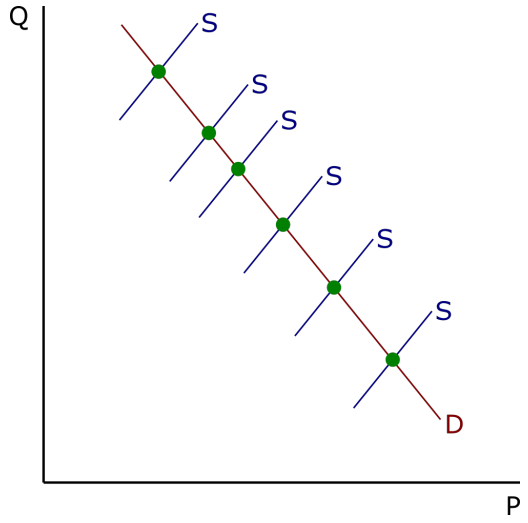
The relationship between q and p is *both* positive and negative (depending on whether we look at the supply or demand curve)! How can we fit a line through price and quantity data, and call it a demand curve? We could be estimating the supply curve, or (most likely) a combination of the two. If we want to estimate the slope of the demand curve, then we need to hold it's position constant. That is, the variation in price would have to come only from shifts in supply, so that we are tracing out points along a demand curve. See figure 11.3.

The problem is, the demand curve is shifting along with the supply curve! The data that we observe is the result of demand and supply intersecting. See figure 11.4.

To estimate the slope of the demand curve, we need to use variation in price that is due to shifts of the *supply curve only*. To do this, we can use the *supply-shifter* variables as instruments for price! In the first stage of 2SLS, we regress the *price* variable on the *supply-shifter* variables[2], and get the LS predicted values from this 1st stage regression ($\hat{price}$). These predicted values are changes in price due to changes in supply only. In the 2nd stage, we estimate the demand equation, but we use $\hat{price}$ from the 1st stage instead of just *price*.

[2]We also need to regress *price* on all of the variables in the demand equation as well.

Figure 11.3: In order to estimate the slope of the demand curve, variation in quantity and price must come from shifts in supply.



11.5.1 Fulton fish market data

Graddy (1995) produces data on the Fulton fish market, and Angrist, Graddy, and Imbens (2000) estimate the demand curve in this market using instrumental variables. The version of the data that we use is from Wooldridge (2020). Download the data:

```
fish <- read.csv("https://rtgodwin.com/data/fish.csv")
```

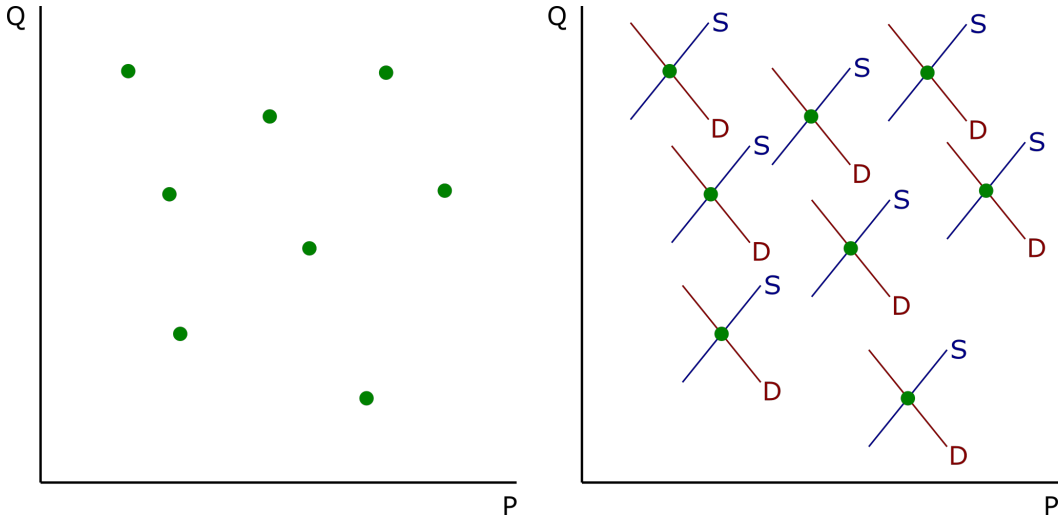
Table 11.1: Description of **some** of the variables in the Graddy (1995) Fulton fish market data. We only use a few variables for this example. In parentheses the variables are labeled as either demand-shifters d or supply-shifters s.

totqty (q)	quantity of fish sold that day
avgprc (p)	price of fish that day
mon (d)	dummy variable equal to 1 if it's Monday
tues (d)	
wed (d)	
thurs (d)	
wave2 (s)	average max last 2 days wave height
wave3 (s)	average max wave heights of 3 and 4 day lagged heights

The variables in the data set are shown in Table 11.1. Demand may change depending on the day: the dummy variables are the demand-shifters. Supply is affected by the weather: if the sea is rough it is harder to fish. Using the variables `wave2` and `wave3` as instruments for price, we can use variations in price that are due to changes in supply only, in order to estimate the slope of the demand curve. Graddy's own description of the 2SLS approach:

> ...first a regression is run with log price as the dependent variable and the storminess of the weather as the explanatory variable. This regression seeks to measure the variation in price that is attributable to stormy weather. The coefficients from this regression are then used to predict log price on each day, and these predicted values for price are inserted back into the regression.

Figure 11.4: Price and quantity data is the result of the intersection of shifting demand and supply curves. We cannot attribute changes in quantity due to changes in price as coming just from the demand curve. Quantity and price are *endogenous* variables.



To estimate the demand equation:

$$\log(totqty) = \beta_0 + \beta_1 \log(avgprc) + \beta_2 mon + \beta_3 tues + \beta_4 wed + \beta_5 thurs + \epsilon$$

using 2SLS/IV, we can use the R code:

```
install.packages("ivreg")
library(ivreg)
iv.fish <- ivreg(log(totqty) ~ log(avgprc) + mon + tues + wed + thurs |
                         wave2 + wave3 + mon + tues + wed + thurs,
              data = fish)
summary(iv.fish)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.16410    0.18171  44.930  < 2e-16 ***
log(avgprc) -0.81582    0.32744  -2.492  0.01453 *
mon         -0.30744    0.22921  -1.341  0.18317
tues        -0.68473    0.22599  -3.030  0.00318 **
wed         -0.52061    0.22357  -2.329  0.02209 *
thurs        0.09476    0.22521   0.421  0.67492
```

Since the variables are in logs, we have estimated the *elasticity* of the demand curve: when price increases by 1%, the quantity demanded is estimated to decrease by 0.81582%. Let's compare this to the LS estimates (as we would have done in previous chapters):

```
ls.fish <- lm(log(totqty) ~ log(avgprc) + mon + tues + wed + thurs,
                          data = fish)
summary(ls.fish)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.24432    0.16281  50.637  < 2e-16 ***
log(avgprc) -0.52466    0.17611  -2.979  0.00371 **
mon         -0.31093    0.22582  -1.377  0.17193
tues        -0.68279    0.22267  -3.066  0.00285 **
```

```
wed          -0.53389     0.21994   -2.427   0.01717 *
thurs         0.06723     0.22042    0.305   0.76107
```

The LS estimate for the elasticity is much lower (0.52466%). The LS estimator is inconsistent because price is an endogenous variable!

11.6 Review Questions

1. What is the *missing variable* problem?
   Answer. When a variable that is both correlated to X and Y is *omitted*, we get *omitted variable bias*. The solution is to include the omitted variable in the model. The *missing variable* problem is similar, except that the missing variable cannot be included in the model (because it's missing!). When a variable that is correlated to both X and Y is missing, the least squares estimator for the effect of X on Y will be biased and inconsistent.
2. What is the general goal of instrumental variable (IV) estimation?
   Answer. The general goal of IV estimation is to solve the missing variable problem: obtain a consistent estimator for the effect of X on Y when there is a missing variable that is correlated to both X and Y.
3. What two properties must an instrumental variable have?
   Answer. An instrument must be (i) uncorrelated with the missing variable, and (ii) correlated with the "problem" X variable (the "problem" being the correlation with the missing variable).
4. What is two-stage-least-squares (2SLS)?
   Answer. 2SLS is another name for "instrumental variables estimation". It is the procedure for using an instrumental variable to get a consistent estimators for the β in our models.
5. What is the 2SLS procedure?
   Answer. In the first stage, the "problem" X variable is regressed on the instrument and other X variables, and the LS predicted values from this regression are obtained (the $\hat{X}$). In the second stage, the $\hat{X}$ are used in place of X in the population model.
6. Why does IV estimation work?
   Answer. IV estimation works because the instrument, which is uncorrelated to the missing variable, is used to predict the "problem" X variables. Those predicted values, $\hat{X}$, are now uncorrelated with the missing variable. Using $\hat{X}$ instead of X gives an estimator that is consistent.

References

Adler, D., D. Murdoch, and others (2018). rgl: 3D Visualization Using OpenGL. R package version 0.99.16. https://CRAN.R-project.org/package=rgl

Angrist, J. D., Graddy, K., & Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. The Review of Economic Studies, 67(3), 499-527.

Croissant, Y. (2016). Ecdat: Data Sets for Econometrics. R package version 0.3-1. URL https://CRAN.R-project.org/package=Ecdat

Graddy, K. (1995). Testing for Imperfect Competition at the Fulton Fish Market. *Rand Journal of Economics*, 26, 75–92.

Kleiber, C., and A. Zeileis (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER

Nunn, N., and Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. Review of Economics and Statistics, 94(1), 20-36.

Prest, A. R. (1949). Some experiments in demand analysis. *The Review of Economics and Statistics*, 33-49.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/

Verbeek, M. (2008). A Guide to Modern Econometrics. John Wiley & Sons.

Wooldridge, J. M. (2020). Introductory Econometrics 7th ed.