Introductory Econometrics

Description, Prediction, and Causality

Third edition

David M. Kaplan

1st edition, January 2019; 2nd edition, June 2020; 3rd edition, June 2021
Updated February 5, 2023

To my past, present, and future students, including NLK and OAK.
—DMK

The chief difficulty Alice found at first was in managing her flamingo: she succeeded in getting its body tucked away, comfortably enough, under her arm, with its legs hanging down, but generally, just as she had got its neck nicely straightened out, and was going to give the hedgehog a blow with its head, it would *twist itself round and look up in her face, with such a puzzled expression that she could not help bursting out laughing: and when she had got its head down, and was going to begin again, it was very provoking to find that the hedgehog had unrolled itself, and was in the act of crawling away. . . Alice soon came to the conclusion that it was a very difficult game indeed.*

Lewis Carroll, *Alice's Adventures in Wonderland*
(An allegory for econometrics)

Brief Contents

Contents

Preface

This textbook was prepared for the 15-week semester Introductory Econometrics course at the University of Missouri. The class focuses on statistical description, prediction, and "causality," including both structural parameters and treatment effects. Description and prediction (forecasting) with time series are also covered. Students learn to think probabilistically, understand prediction and causality, judge whether various assumptions hold true in real-world examples, and apply econometric methods in R.

As usual, this textbook may be used to teach different types of classes. In full, the textbook provides a 15-week semester class that assumes a previous class in probability and statistics. That prerequisite could be skipped if more time is spent on the "review" material in the first few chapters. Calculus is avoided but could be added in the usual places. A shorter class could omit the time series material. Of course, any material may be expanded, condensed, or skipped, as the instructor desires.

Some complementary, complimentary texts and courses deserve mention. Econometrics professor Matt Masten has a "Causal Inference Bootcamp" video series,[1] as well as some "Causal Inference with R" free courses on DataCamp.[2] Relevant videos are linked at the beginning of each chapter in this textbook. Stanford statistics professors Trevor Hastie and Rob Tibshirani created a free introductory machine learning (statistical learning) course, focusing more on prediction and estimation.[3] Their course uses their free textbook (James, Witten, Hastie, and Tibshirani, 2013) that includes R examples.[4] Hastie, Tibshirani, and Friedman (2009) also provide their more advanced statistical learning textbook for free.[5] For econometrics texts focused on prediction and time series, see Diebold (2018a,b,c).[6] The forecasting book by Hyndman and Athanasopoulos (2019) is at `https://otexts.com/fpp2` and uses R. Finally, Hanck, Arnold, Gerber, and Schmelzer (2018) mirror the structure of the (expensive) textbook of Stock and Watson

[1] https://mattmasten.github.io/bootcamp

[2] https://www.datacamp.com/community/open-courses

[3] https://www.edx.org/course/statistical-learning

[4] https://statlearning.com

[5] https://web.stanford.edu/~hastie/ElemStatLearn

[6] http://www.ssc.upenn.edu/~fdiebold/Textbooks.html

(2015), providing many R examples to illustrate the concepts they explain.[7]

One distinguishing feature of this textbook is the development of the ideas of (and distinctions among) statistical description, prediction, causal inference, and structural estimation in the simplest possible settings. Other texts combine these with all the complications of regression from the beginning, often confusing students (like my past self).

A second distinguishing feature is that this text's source files are freely available. Instructors may modify them as desired, or copy and paste LaTeX code into their own lecture notes, subject to the Creative Commons license linked on the copyright page. I wrote the textbook in Overleaf, an online (free) LaTeX environment that includes knitr support, so most of the R code and output is in the same .Rtex files alongside the LaTeX code. Graphs are either generated from code in the .Rtex files or else from a single .R file also provided in the source material. You may see, copy, and download the entire project from Overleaf[8] or from my website.[9]

Third, I provide learning objectives for the overall book and for each chapter. This follows current best practices for course design. Upon request, I can provide a library of multiple choice questions, labeled by learning objective. (Empirical exercises are already at the end of each textbook chapter.)

Fourth, in-class (or online) discussion questions are included along the way. When I teach in person (30–40 students), I prefer to punctuate lectures with such questions every 20–30 minutes, where students first discuss them for a couple minutes in small groups of 2–3 students, and then volunteer to share their group's ideas with the whole class for another couple minutes. This provides an active learning opportunity, a time for students to realize they don't understand the lecture material (so they can ask questions), practice discussing econometrics with peers, and (if nothing else) a few minutes' rest.

Thanks to everyone for their help and support: my past econometrics instructors, my colleagues and collaborators, my students (who have not only inspired me but alerted me to typos and other deficiences in earlier drafts), and my family.

David M. Kaplan
Summer 2018 (edited Summer 2020)
Columbia, Missouri, USA

[7] https://www.econometrics-with-r.org

[8] https://www.overleaf.com/read/fszrgmwzftrk

[9] https://kaplandm.github.io/teach.html

Textbook Learning Objectives

For good reason, it has become standard practice to list learning objectives for a course as well as each unit within the course. Below are the learning objectives corresponding to this textbook overall. Each chapter lists more specific learning objectives that map to one or more of these overall objectives. The accompanying exercises are also classified by learning objectives. I hope you find these helpful guidance, whether you are a solo learner, a class instructor, or a class student.

The textbook learning objectives (TLOs) are the following.

1. Define terms from probability, statistics, and econometrics, both mathematically and intuitively.

2. Describe various econometric methods both mathematically and intuitively, including their objects of interest and assumptions, and the logical relationship between the assumptions and corresponding theorems and properties.

3. Interpret the values that could be estimated with infinite data, in terms of description, prediction, and causality (or economic meaning).

4. Explain the frequentist/classical statistical and asymptotic frameworks, including their benefits and limitations.

5. Provide multiple possible (causal) explanations for any statistical result, distinguishing between statistical and causal relationships.

6. For a given economic question, dataset, and econometric method, judge whether the method is appropriate and judge the economic significance and statistical significance of the results.

7. Using R (or Stata): manipulate and analyze data, interpreting results both economically and statistically.

Notation

Much of the notation below will not make sense until you get to the corresponding point in the textbook. The following is primarily for your reference later.

Variables

Usually, uppercase denotes a random variable, whereas lowercase denotes a non-random (fixed, constant) value. The primary exception is for certain counting variables, where uppercase indicates the maximum value and lowercase indicates a general value; e.g., time period t can be $1, 2, 3, \ldots, T$, or regressor k out of K total regressors. Greek letters like β and θ usually denote non-random (fixed) population parameters.

Estimators usually have a "hat" on them, like $\hat{\theta}$. Since estimators are computed from data, they are random from the frequentist perspective. Thus, even if θ is a non-random population parameter, $\hat{\theta}$ is a random variable.

I try to put "hats" or bars on other quantities computed from the data, too. For example, a t-statistic would be $\hat{t}$ instead of just t (which looks like a non-random scalar). The sample average of $Y_1, \ldots, Y_n$ is $\bar{Y}$.

Estimators and other statistics (i.e., things computed from data) may sometimes have a subscript with the sample size n to remind us that their sampling distribution depends on n. For example, $\hat{\theta}_n$, $\hat{t}_n$, and $\bar{Y}_n$.

Symbols

In addition to the following symbols, vocabulary words and abbreviations (like "regression" or "OLS") can be looked up in the Index in the very back of the textbook.

$\Longrightarrow$	implies; see Section 6.1
$\Longleftarrow$	is implied by; see Section 6.1
$\Longleftrightarrow$	if and only if; see Section 6.1
$\lim\limits_{n\to\infty}$	limit (like in pre-calculus)
$\operatorname*{plim}\limits_{n\to\infty}$	probability limit; see Section 3.6.3

$\equiv$	is defined as
$\approx$	approximately equals
$\sim$	is distributed as
$X \perp\!\!\!\perp Y$	X and Y are statistically independent; see Section 6.2.5
$\mathrm{N}(\mu, \sigma^2)$	normal distribution with mean μ and variance σ^2
$\mathrm{N}(0, 1)$	standard normal distribution
$F_Y(\cdot)$	cumulative distribution function (CDF) of Y; see Section 2.3
$\mathbb{1}\{\cdot\}$	indicator function; see (2.1)
$\mathrm{P}(A)$	probability of event A
$\mathrm{P}(A \mid B)$	conditional probability of A given B; see Section 6.2.3
$\mathrm{E}(Y)$	expectation (mean) of Y; see Section 2.3
$\mathrm{E}(Y \mid X = x)$	CMF (a function of x); see Section 6.3
$\mathrm{E}(Y \mid X)$	conditional expectation of Y given X; this is a random variable
$\sum\limits_{i=1}^{n}$	summation from $i = 1$ to $i = n$
$\mathrm{Var}(Y)$	variance of Y; see (3.20)
$\mathrm{Var}(Y \mid X = x)$	conditional variance (a non-random value); see Section 6.7.1
$\mathrm{Var}(Y \mid X)$	conditional variance (a random variable)
$\mathrm{Cov}(Y, X)$	covariance
$\mathrm{Corr}(Y, X)$	correlation
$\{a, b, \ldots\}$	a set (containing elemnts a, b, etc.)
$i = 1, \ldots, n$	same as $i \in \{1, \ldots, n\}$ (integers from 1 to n)
$j = 1, \ldots, J$	same as $j \in \{1, \ldots, J\}$ (integers from 1 to J)
$s \in \mathcal{S}$	element s is in set $\mathcal{S}$
$\hat{\mathrm{E}}(Y)$	expectation for sample distribution; see Section 3.4.1
$\bar{Y}_n$	$\frac{1}{n}\sum_{i=1}^{n} Y_i$; same as $\hat{\mathrm{E}}(Y)$; see Section 3.4.1
$\hat{\theta}$	estimator of population parameter θ; see Section 3.4
$\arg\min\limits_{g} f(g)$	the value of g that minimizes $f(g)$
$\arg\max\limits_{g} f(g)$	the value of g that maximizes $f(g)$

Chapter 1

Getting Started with R (or Stata)

⟹ Kaplan video: Course Introduction

Depends on: no other chapters

Unit learning objectives for this chapter

1.1. Run statistical software (R/RStudio or Stata) [TLO 7]

1.2. Write code to do basic data manipulation, description, and display [TLO 7]

You will use R (or Stata) for the empirical exercises in this textbook. The code examples in the textbook are all in R.

No previous experience with any statistical software is assumed. Consequently, the primary goal of the empirical exercises is to develop your confidence and experience with statistical software, applying the text's methods and ideas to real datasets. Toward this goal, there are lots of explicit hints about the code you need to write.

If you actually do have previous experience (or above-average interest), then the empirical exercises may feel too boring. You could try figuring out alternative ways to code the solution, or coding alternative analyses, etc. You can also explore other online resources like one of the free DataCamp courses.[1]

Due to the many excellent resources online (see Section 1.4), there are many people who can write R code, but most do not understand how to properly interpret econometric results or judge which method is most appropriate. So, overall, this class/textbook focuses more on understanding econometrics than coding.

[1] https://www.datacamp.com/community/open-courses

1.1 Comparison of R and Stata

I like both R and Stata statistical software, and I have used both professionally. They excel in different ways mentioned below.

For this textbook/class, I focus on R for the following reasons.

1. It's widely used in the private sector, government, and academia alike, in many fields (including economics).
2. It's free to download/use, and can even be used through a web browser.
3. It has many econometric/statistical functions available, and creators of new econometric/statistical methods often provide code in R.
4. There are many online resources for learning R and getting help.

In comparison, Stata:

1. is widely used in economics and certain social sciences, but less so in fields like data science and statistics.
2. is not free, and can't be used in a browser; but is free to use in many campus computer labs.
3. is easier to use for standard econometric methods, and has some new econometric methods (while others take a few years to be implemented).
4. also has good help files (documentation) and online support.

1.2 R

1.2.1 Accessing the Software

There are three ways you could run R: downloaded onto your own computer, through a web browser (in the cloud), or on another computer like in a campus computer lab.

Other computers or web browser versions may have the core R software but lack certain packages needed for the empirical exercises. In some cases, you can simply install the necessary packages with a single command (e.g., in Mizzou computer labs). In other cases, you may be prohibited from installing packages, in which case you won't be able to complete the exercises, so make sure to check this first.

Through a Web Browser

There are many free options for using R through a web browser, and they evolve quickly. This means both new and improved options becoming available, as well as existing options disappearing, even from major companies (e.g., Microsoft Azure Notebooks was "retired").

For Mizzou students, Software Anywhere offers free web browser access to RStudio as well as a variety of other software (including Stata). From the Software Anywhere web page,[2] click the "Getting Started" tab and follow the instructions. Once logged in, it's essentially the same as if you had installed RStudio on your own computer. Technical assistance: https://doit.missouri.edu/tech-support/

Another good option is RStudio Cloud (name changed to Posit Cloud in 2022). It's free, reliable, and the same RStudio interface as if you downloaded RStudio, so you can learn from the latter half of my RStudio video. To get started:

1. Go to https://posit.cloud/ in any web browser
2. Click the GET STARTED FOR FREE button (or else "Log In" if you already have an account)
3. Follow the instructions to sign up for a free account
4. Start using RStudio like it were on your own computer
5. Install necessary packages like usual: see Section 1.2.2
6. After you log out and later log in, click "Untitled Project" (feel free to rename) to get back to where you were

At http://mybinder.org/v2/gh/binder-examples/r/master?urlpath=rstudio you can also use the RStudio interface through a web browser, without even making an account, but 1) it does not run the most current version of R, 2) it cannot save your files from one session to the next, 3) you have to install the packages every time. But these are not critical problems for this class: older R versions are fine, you can save your code/output in a text file on your own computer, and it only takes 1 line of code (and under 1 minute) to install the packages.

The following options are not as good for our class, so details are omitted.

- CoCalc (no account required): https://cocalc.com
- Google CoLab (requires Google account): https://colab.research.google.com/drive/1BYnnbqeyZAlYnxR9IHC8tpW07EpDeyKR
- DataCamp Workspace (requires free account): https://www.datacamp.com/workspace
- Gradient by Paperspace (free account required): https://gradient.paperspace.com/

In a Mizzou Computer Lab

You can check which Mizzou computing sites/labs have your favorite software on the Computing Sites Software web page.[3] Scroll down to RStudio to see where you can use R with RStudio. However, sometimes there are classes or other events in computer labs; you can check the weekly schedule posted near the door to find a free time, or you can check online.[4]

[2] https://doit.missouri.edu/services/software/software-anywhere

[3] https://doit.missouri.edu/services/computing-sites/sites-software

[4] https://doit.missouri.edu/services/computing-sites and click the lab name

After you log into the computer in the computer lab, open RStudio from the Start menu. (RStudio calls R itself in the background; you don't have to open R directly.) Then just start typing commands, and hit Enter to run them.

The computer labs don't currently have the necessary packages pre-installed, but you can easily install them. Note that you'll have to do this every time you log in (because any files you download/save get deleted when you log out), but you can just run the same line of code when you start RStudio each time.

Also, make sure to email yourself your code (or otherwise save it, if you haven't finished and uploaded to Canvas) before you log out, because your files get deleted when you log out.

Downloading Software

⟹ Kaplan video: Getting Started with R/RStudio

You'll download two pieces of software: R itself, and RStudio. Both are free. R has all the functions you need. RStudio makes the interface nicer and makes things easier for you.

On Windows:

- Download the .exe installer file for R: Google "R Windows" or try `https://cran.r-project.org/bin/windows/base` and click the "Download. . . " link near the top.
- Open the downloaded .exe installer and follow the instructions.
- Download the .exe installer file for RStudio Desktop (free version): Google "RStudio download" or try `https://www.rstudio.com/products/rstudio/download/#download`
- Open the downloaded .exe installer and follow the instructions.

On Mac:

- Download the .pkg file for R: Google "R Mac" or try `https://cran.r-project.org/bin/macosx`
- Open the file and follow the usual Mac installation procedure.
- Download the .dmg file for RStudio Desktop (free version): Google "RStudio download" or try `https://www.rstudio.com/products/rstudio/download/#download`
- Open the file and follow the usual Mac installation procedure.

On Linux, etc.: if you can figure out how to run something besides Windows or Mac, you can probably figure out how to download a couple files by yourself, but please let me know if not.

Regardless of OS, after both are installed, you only ever need to open RStudio, never R. Once you open RStudio, just type a command and hit Enter to run it.

1.2.2 Installing Packages

You may need to install certain packages to do the empirical exercises. This can be done with a single command in R. You should double-check the package names required for each exercise, but it would be something like:

```
install.packages(c('wooldridge','lmtest','sandwich','forecast','survey'))
```

With R on your own computer, you only need to run this once (not every time you use your computer), but with a web interface or computer lab, you may need to run this code every time you start a session in R. You can check which packages are already installed with `installed.packages()`

A bit about the packages:

- `wooldridge` (Shea, 2018) has datasets originally collected by Wooldridge (2020) from various sources.
- `lmtest` and `sandwich` (Zeileis, 2004; Zeileis and Hothorn, 2002) help construct confidence intervals (and other things) appropriate for economic data.
- `survey` (Lumley, 2004, 2019) has functions for dealing with complex survey sampling.
- `forecast` (Hyndman, Athanasopoulos, Bergmeir, Caceres, Chhay, O'Hara-Wild, Petropoulos, Razbash, Wang, and Yasmeen, 2020; Hyndman and Khandakar, 2008) has methods for forecasting.

1.3 Stata

1.3.1 Accessing the Software

There are three ways you could run Stata: in a campus computer lab, through Mizzou's Software Anywhere, or (if you purchase your own copy) downloaded onto your own computer.

Empirical exercises only require built-in commands. Stata has additional commands available for download, but none are needed for the exercises, so any (internet-connected) computer with Stata is sufficient.

In a Mizzou Computer Lab

You can check which Mizzou computing sites/labs have your favorite software on the Computing Sites Software web page.[5] Scroll down to Stata to see where it's available.

[5] https://doit.missouri.edu/services/computing-sites/sites-software

However, sometimes there are classes or other events in computer labs; you can check the weekly schedule posted near the door to find a free time, or you can check online.[6]

After you log into the computer in the computer lab, open Stata from the Start menu (the actual name is somewhat longer, like "StataSE 15 (64-bit)"). Ideally, you should open the do-file editor, and save a .do file, but for this class you could just type commands into the short, horizontal space at the bottom labeled "Command." You type a command and hit Enter to run it.

Also, make sure to email yourself your code (or otherwise save it, if you haven't finished and uploaded to Canvas) before you log out, because your files get deleted when you log out.

Purchasing and Downloading Software

Student pricing is shown on the Stata website.[7] Currently (Spring 2020), the cheapest option is the 6-month Stata/IC license. Other, more expensive licenses are fine, too.

The software is delivered via download. Follow instructions for installation, and contact Stata if you have any technical difficulties.

Software Anywhere (Mizzou)

From the Software Anywhere web page,[8] click the "Getting Started" tab and follow the instructions. Once logged in, it's the same as if you were sitting at a computer in a Mizzou computer lab (see above).

Technical assistance: `https://doit.missouri.edu/tech-support/`

1.3.2 Installing Additional Commands

Like in R, there are additional Stata commands that can be easily downloaded and installed. Commonly, this can be done with the command `ssc install` followed by the name of the command.

For the exercise sets, the only additional command you'll need is `bcuse`. You can install this with the command `ssc install bcuse`. If you're in a computer lab, you may need to run this command every time you start Stata; if you have it on your computer, just once is sufficient. This command makes it easy to load the datasets from Wooldridge (2020).[9]

[6] `https://doit.missouri.edu/services/computing-sites` and click the lab name

[7] `https://www.stata.com/order/new/edu/gradplans/student-pricing`

[8] `https://doit.missouri.edu/services/software/software-anywhere`

[9] Descriptions: `http://fmwww.bc.edu/ec-p/data/wooldridge/datasets.list.html`

1.4 Optional Resources

If you only want to learn enough R (or Stata) to do well in this class, then you may skip this section. If you'd like to learn more on your own, these resources might help you get started in the right direction.

1.4.1 R Tutorials

Eventually, you will be able to simply Google questions you have about R. There are lots of people on the internet really excited about helping you figure stuff out in R, which is great.

However, when you are first getting started, it may help to go through a basic tutorial. You are welcome to Google "R basic tutorial" yourself, or you could try one of the following.

1. Section 2.3 ("Lab: Introduction to R") in James, Witten, Hastie, and Tibshirani (2013)
2. Section 1.1 in Hanck et al. (2018)
3. Sections 1.1–1.3 in Heiss (2016)
4. Sections 2.1–2.5 in Kleiber and Zeileis (2008) [Chapter 2 is free on their website]
5. Chapter 2 in Kaplan (2020)
6. Professor Sebastian Wai's introductory R notes
7. No longer free after first chapter: `datacamp.com` courses like Introduction to R[10]

1.4.2 R Quick References

At first, it may help to have some quick reference "cheat sheets."[11,12]

1.4.3 Running Code in This Textbook

If you'd like, you should be able to copy code directly from the textbook .pdf file and paste it into R. Sometimes, you need to install a certain package first. This can be done either manually or with the R function `install.packages()`. For example, to install package `mgcv`, run the command `install.packages('mgcv')` within R.

[10] https://www.datacamp.com/courses/free-introduction-to-r

[11] https://www.rstudio.com/wp-content/uploads/2016/10/r-cheat-sheet-3.pdf

[12] https://cran.r-project.org/doc/contrib/Short-refcard.pdf

1.4.4 Stata Resources

For Stata, helpful cheat sheets (quick references) are available for free[13] as well as various tutorials.[14]

[13] https://www.stata.com/bookstore/statacheatsheets.pdf

[14] https://www.stata.com/links/resources-for-learning-stata

Empirical Exercises

Empirical Exercise EE1.1. In either R or Stata, create a **script** (a sequence of commands, with one command per line) to do the following. The data are from a New York Times article on December 28, 1994.

a. R: load (and install if necessary) package `wooldridge`:

   ```
   if (!require(wooldridge)) {
     install.packages('wooldridge'); library(wooldridge)
   }
   ```

   Stata: run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse wine, clear`

b. View basic dataset info with R command `?wine` or Stata command `describe`

c. View the first few rows of the dataset with R command `head(wine)` or Stata command `list if _n<=5`

d. Rename the `alcohol` column, which measures liters of alcohol from wine (consumed per capita per year).

   R: `names(wine)[2] <- 'wine'`

   Stata: `rename alcohol wine`

e. Add a column named `id` whose value is just 1, 2, 3, 4, 5, etc.

   R: `wine$id <- 1:nrow(wine)`

   Stata: `generate id = _n`

f. Display the countries with fewer than 100 heart disease deaths per 100,000 people.

   R: `wine$country[wine$heart<100]`

   Stata: `list country if heart<100`

g. Display the rows for the countries with the 5 lowest death rates, sorted by death rate.

   R: `wine[order(wine$deaths)[1:5],]`

   Stata: `sort deaths` followed by `list if _n<=5`

h. Add a column with the sum of heart and liver disease deaths per 100,000.

   R: `wine$heart.plus.liver <- wine$heart + wine$liver`

   Stata: `generate heart_plus_liver = heart + liver`

i. Generate a variable with the squared death rate.

   R: `wine$deaths.sq <- wine$deaths^2`

   Stata: `generate deaths_sq = deaths^2`

j. Display the sorted death rates.

   R: `print(sort(wine$deaths))`

   Stata: `sort deaths` followed by `list deaths`

k. R: create a vector with the proportion of total deaths (per 100,000) caused by heart disease with command `heart.prop <- wine$heart/wine$deaths` and then name the entries by country with `names(heart.prop) <- wine$country` and print the named vector of heart disease death proportions, rounded to three decimal places, with `print(round(heart.prop, digits=3))`

   Stata: add a column with the proportion of heart deaths to total deaths with command `generate heart_prop = heart / deaths`

l. Create a histogram of liver deaths.

   R: `hist(wine$liver)`

   Stata: `histogram liver`

m. Create a scatterplot of liver death rates (vertical axis) against wine consumption (horizontal axis).

   R: `plot(x=wine$wine, y=wine$liver)`

   Stata: `scatter liver wine`

n. R only: make the same plot but with axes starting at zero, adding the arguments `xlim=c(0,max(wine$wine))` and `ylim=c(0,max(wine$liver))` to the previous `plot()` command

Part I

Analysis of One Variable

Introduction

This textbook explores methods to answer three types of economic questions, each detailed in Part I:

1. Description (how things are/were: statistical properties and relationships)
2. Prediction (guessing an unknown value, without interfering)
3. Causality (how changing one variable would affect another, all else equal)

For example, imagine you are interested in income. Depending on your job, you may want to answer a different type of question, like:

1. Description: how many adults in the U.S. have an income below $20,000/yr? What's the mean income among U.S. adults? What's the difference in mean income between two socioeconomic or demographic groups, like those with and without a college degree?
2. Prediction: for advertising purposes, what's the best guess of the income of an unknown person visiting your company's website? What's the best prediction if you also know their zip code (where they live)?
3. Causality: for a given individual, how different would her income be if she had a college degree than if she didn't, keeping everything else about her (parents, height, social skills, etc.) identical? How different would her income be if she were a man, all else equal? If she were white?

Description helps us see. It summarizes an incomprehensible mass of numbers into specific, economically important features we can understand. By analogy: knowing the color of each of 40,000 pixels in a photograph is not as valuable as knowing it's a cat.

Prediction aids decisions dependent on unknowns. The example questions above consider the purpose of advertising, where correctly guessing a person's income helps decide which ad is most effective. In other private sector jobs, you may need to predict future demand to know how many self-driving cars to start producing, or predict future oil prices to aid a freight company's decisions. In government or non-profit work, optimal policy may depend on predicting next year's unemployment rate. In each case, as detailed in Section 2.5, the "best" prediction depends on the consequences of the related decision.

Causality also aids decisions. The example question about the causal effect on income of a college degree matters for government policies to subsidize college (or not), as well as individual decisions to attend college. With business decisions, like changes to advertising or website layout, the causal effect on consumer behavior is what matters: does the change itself actually cause consumers to buy more? Among the three types, questions of causality are the most difficult to answer. Learning about causality from data has been a primary focus of the field of econometrics.

Of course, not all important questions concern description, prediction, and causality. Policy questions usually involve tradeoffs that ultimately require value judgments. For example, how much future wellbeing is worth sacrificing to be better off right now? How much GDP is worth sacrificing to decrease inequality? Should a school have honors classes that help the best students at the expense of the other students? Each of these policy questions requires a subjective value judgment that cannot be answered objectively from data.

That said, each policy question also depends on objectively quantified description, prediction, and causality. For example, the policy question about decreasing inequality depends on the current levels of GDP and inequality (description) as well as the causal effect of the policy (e.g., tax) change on GDP and inequality (causality). The future/present wellbeing tradeoff depends on the current level of wellbeing (description) as well as future levels (prediction). The honors class tradeoff depends on the causal effect of honors classes on different types of students (causality) as well as the current mix of student types (description) and future mix (prediction).

Chapter 2

One Variable: Population

⟹ Kaplan video: Chapter Introduction

Chapter 2 studies a single variable by itself. This setting's simplicity helps us focus on the complexity of fundamental concepts in probability, description, and prediction. This fundamental understanding will help you tackle more complex models later, in this class and beyond.

If you've previously had a probability or statistics class, then most of this chapter may be review for you, although the optimal prediction material is probably new. If you haven't, then now is your opportunity to catch up.

Unit learning objectives for this chapter

2.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

2.2. Describe and distinguish among different types of populations, including which is most appropriate for answering a certain question [TLO 3]

2.3. Compute, describe, and interpret the mean for different types of random variables, for description and prediction [TLO 3]

2.4. Assess the most appropriate loss function and prediction in a real-world situation [TLO 6]

2.5. Compute mean loss and the optimal prediction in simple mathematical examples [TLO 2]

2.1 The World is Random

⟹ Kaplan video: "Before" and "After" Perspectives of Data

2.1.1 Before and After: Two Perspectives

Consider a coin flip. The two possible outcomes are heads (h) and tails (t). After the flip, we observe the outcome (h or t). Before the flip, either h or t is possible, with different probabilities.

Let variable W represent the outcome. After the flip, the outcome is known: either $W = h$ or $W = t$. Before the flip, both $W = h$ and $W = t$ are possible. If the coin is "fair," then possible outcome $W = h$ has probability 1/2, as does $W = t$. (Recall it is equivalent to write 1/2, 0.5, or 50%.)

The "after" view sees W as a **realized value** (or **realization**). It is either heads or tails. Even if the actual "value" (heads or tails) is unknown to us, there is just a single value. For example, in physics the variable c represents the speed of light in a vacuum; you may not know the value, but c represents a single value.

Instead, the "before" view sees W as a **random variable**. That is, instead of representing a single (maybe unknown) value like in algebra, W represents a set of possible values, each associated with a probability. In the coin flip example, the possible outcomes are h and t, and the associated probabilities are 0.5 and 0.5.

Other terms for W include a **random draw** (or just **draw**), or more specifically a random draw (or "randomly drawn") from a particular probability distribution. Seeing the population as a probability distribution (see Section 2.2), we could say W is randomly sampled from its population distribution, or if there are multiple random variables $W_1, W_2, \ldots$ (e.g., multiple flips of the same coin), we could say they are randomly sampled from the population or that they collectively form a **random sample**; see Section 3.2 for more about sampling.

Notationally, in this textbook, random variables are usually written uppercase (like W or Y), whereas realized values are usually written lowercase (like w or y). This notation is not unique to this textbook, but beware that other books use different notation. (For more on notation, see the Notation section in the front matter before Chapter 1.)

Example 2.1 (Kaplan video)**.** Let $R = 1$ if it rains in Columbia, MO on Tuesday and $R = 0$ if not. If today is Monday, then either outcome is possible, so we have the "before" view: R is a random variable, with some probability of $R = 0$ and some probability of $R = 1$. If instead today is Wednesday, then what happened Tuesday is already determined, so we have the "after" view. If it rained, then $R = 1$; if not, $R = 0$. There is only a single value, not multiple possible values. Even if we don't know the realized value r, we know it's just a single value.

2.1.2 Before and After Sampling

Extending Section 2.1.1 are the **before sampling** and **after sampling** perspectives, or "before observation" and "after observation." Similar to Section 2.1.1, "before" corresponds to random variables, whereas "after" corresponds to realized values. Before sampling a unit (person, firm, etc.) from a population, we don't know which one we'll get, so there are multiple possible values. After sampling, we can see the specific values we got.

Example 2.2. Imagine you plan to record the age of one person living in your city. You take a blank piece of paper on which you'll write the age. As in Section 2.1.1, after you find a person and write their age ("after sampling"), that number can be seen as a realized value, like w. In contrast, before sampling, there are many possible numbers that could end up on your paper. It's not that peoples' ages are undetermined; they each know their own age. But before you "sample" somebody, it's undetermined whose age will end up on your paper. It could be your neighbor DeMarcus, age 88. It could be your kid's friend Lucia, age 7. It could be your colleague Xiaohong, age 35. The random variable W is like your blank paper: it has many possible values, like $W = 88$, $W = 7$, or $W = 35$.

Discussion Question 2.1 (web traffic)**.** Let $Y = 1$ if you're logged into the course website and $Y = 0$ if not.

a) From what perspective is Y a non-random value?
b) From what perspective is Y a random variable?

There is always a "before" view from which data samples (like ages) can be seen as random variables, although sometimes it requires some additional peculiar thought experiments, like imagining we first "sample" one universe out of many, like with the superpopulation in Section 2.2.

In Sum: Before & After

Before: multiple possible values $\implies$ random variable
After: single observed value $\implies$ realized value (non-random)

2.1.3 Outcomes and Mechanisms

Knowing everything about a coin does not fully determine the outcome of a single coin flip. For example, even if we flip two identical coins (i.e., same probability of heads) at the same time, one may get heads while the other gets tails. Mathematically, with two coins represented by W and Z, even if they are "identical" in that $\mathrm{P}(W = h) = \mathrm{P}(Z = h)$ and $\mathrm{P}(W = t) = \mathrm{P}(Z = t)$, we could still sometimes observe $W = h$ and $Z = t$. More abstractly: knowing everything about random variable W does not fully determine any particular realization w. Even if random variables W and Z have the same properties, specific realizations $W = w$ and $Z = z$ may differ.

Conversely, a single coin flip's outcome does not tell us everything about the coin itself. For example, consider a "fair" coin W with $\mathrm{P}(W = h) = \mathrm{P}(W = t) = 1/2$ (50% chance of either heads or tails), and biased coin Z with $\mathrm{P}(Z = h) = 0.99$ (99% heads). By chance, we may flip both and observe $W = h$ and $Z = h$. But the fact that they both came up heads once does not imply that the coins themselves are identical. More abstractly: observing a single realization $W = w$ does not tell us all the properties of random variable W.

We usually want to learn about the underlying mechanisms, like the coin itself. The "before" view in Section 2.1.1 lets us describe the underlying properties that we want to learn, like a coin's probability of heads, $\text{P}(W = h)$.

The coin flip is a metaphor for more complex mechanisms. In economics, instead of learning how coin flip outcomes are determined, we care about the underlying mechanisms that determine a wide variety of outcomes like unemployment, wages, inflation, trade volume, fertility, and education. The underlying mechanism is often called the **data-generating process** (DGP).

Example 2.3. Let $Y = 1$ if somebody is currently employed. Imagine we are interested specifically in the probability of being employed for a 40-year-old with a master's degree in economics. That is, we want to know $\text{P}(Y = 1)$, from the before view. However, sampling a single person from that population cannot teach us the probability. Even sampling five people does not teach us the probability (although it can help us make a better guess).

2.2 Population Types

⟹ Kaplan video: Population Types

This section describes different **population** types and how to determine which is most appropriate for a particular economic question, which in turn helps determine which econometric method is most appropriate.

In this textbook, the population is modeled mathematically as a probability distribution. This is appropriate for the infinite population or superpopulation below, but not the finite population. Consequently, it is most important to distinguish between the finite population and the other two types.

The finite population cares more about the "after" view: which outcomes actually occurred? The other two population types care more about the "before" view, describing properties of the underlying mechanisms that generated the outcomes (the DGP).

2.2.1 Finite Population

In English, "population" means all the people living in some area, like everybody living in Missouri. In econometrics, this type of population is called a **finite population**. Other examples of finite populations are all employees at a particular firm, all firms in a particular industry, all students in a particular school, or all hospitals of a certain size.

The finite population is appropriate when we only care about the outcomes of the population members, not the mechanisms that determine such outcomes. For example, if we want to know how many individuals in Missouri are currently unemployed, then our interest is in a finite population. That is, we don't care why they're unemployed, and we don't care about the probability that they're unemployed; we only care about whether or not they are currently unemployed.

2.2.2 Infinite Population

Sometimes a finite population is so large compared to the sample size (i.e., the number of population members we observe) that an **infinite population** is a reasonable approximation. For example, if we observe only 600 individuals out of the 6+ million in Missouri, econometric results based on finite and infinite populations are practically identical.

Although "infinite" sounds more complex than "finite," it is actually simpler mathematically. Instead of needing to track every single member of a finite population, an infinite population is succinctly described by a probability distribution or random variable. For example, a finite population would need to consider the employment status of all 6+ million Missourians, because sampling somebody unemployed then reduces the number of unemployed individuals remaining in the population who could be sampled next. In contrast, an infinite population considers realizations of a random variable W with some probability of having value "unemployed." There is no effect of removing one individual from an infinite population because $1/\infty = 0$.

Besides this convenience, sometimes there is no finite population (however large) that answers your question. For example, imagine there's a new manufacturing process for carbon monoxide monitors that should sound an alarm above 50ppm. Most work properly, but some are faulty and never alarm. Specifically, this manufacturing process corresponds to some probability of producing a faulty monitor. This is similar to the probability of the coin flipping process producing a "heads." Mathematically, the manufacturing process can be modeled as random variable W with some probability of the value "faulty." If you want to learn this probability (i.e., this property of the manufacturing process), then there is no finite number of monitors that can exactly answer your question; no finite number of realizations exactly determines $\mathrm{P}(W = \text{faulty})$. This is an infinite population question.

2.2.3 Superpopulation

One variation of the infinite population is the **superpopulation** (coined by Deming and Stephan, 1941). This imagines (infinitely) many possible universes; our actual universe is just one out of infinity. Thus, even if it appears we have a finite population, we could imagine that our universe's finite population is actually a single sample from an infinite number of universes' finite populations. The term "superpopulation" essentially means "population of populations." Our universe's finite population "is only one of the many possible populations that might have resulted from the same underlying system of social and economic causes" (Deming and Stephan, 1941, p. 45).

For example, imagine we want to learn the relationship between U.S. state-level unemployment rates and state minimum wage levels. It may appear we are stuck with a finite population because there are only 50 states, each of which has an observable unemployment rate and minimum wage. However, observing all 50 states still doesn't fully answer our question about the underlying mechanism that relates unemployment and minimum wage, so a finite population seems inappropriate. But we can't just manufacture new states like we can manufacture new carbon monoxide monitors, so an infinite population also seems inappropriate. The superpopulation imagines manufacturing new entire universes, each with 50 states and the same economic and legal systems. Given these underlying systems and mechanisms, the states' unemployment rates can be seen as random variables, with various probabilities of the possible values. To answer our economic question, we need to learn about the properties of these random variables, not merely the actual unemployment in our actual 50 states.

2.2.4 Which Population is Most Appropriate?

Practically, you need to decide which econometric method to use to answer a particular question. This decision depends partly on which population type is most appropriate. Specifically, finite-population methods differ from other methods that are appropriate for either superpopulations or infinite populations. Because they are less commonly used in econometrics, finite-population methods are not covered in this textbook.

Consequently, it is most important to judge whether or not a finite population is more appropriate than the other types. Which is most appropriate depends on your question (i.e., what you want to learn).

The finite population is most appropriate if you could fully answer your question by observing every member of a finite population. If not, then a superpopulation or infinite population is more appropriate.

The distinction is described by Deming and Stephan (1941, p. 45). They say the finite population perspective is more appropriate for "administrative purposes" or "inventory purposes," whereas the superpopulation perspective is more appropriate for "scientific generalizations and decisions for action [policy]," as well as "prediction" (assuming you want to predict values outside the finite population, like in the future).

In Sum: Population Type

Hypothetically, could a finite number of observations fully answer your question?
No $\implies$ superpopulation or infinite population, modeled as probability distribution (as in this textbook)
Yes $\implies$ finite population (use different methods unless sample is much smaller than population)

Example: Coin Flips

Imagine the president flips a coin 20 times and then randomly selects 10 observations to report to you; which population types is most appropriate? It depends on your question.

The finite population is most appropriate if you only care about the outcomes of those 20 flips. For example, this may be true if the president was flipping the coin to make a major military decision that you care about (like, "invade if at least 10/20 heads"). Then, knowing the 20 flip outcomes is enough to learn the decision. Further, the sample size is a fairly large proportion (10/20), so approximating 20 as infinity seems inappropriate.

The infinite population is more appropriate if you care about the properties of the coin. For example, even with a fair coin ($p = 1/2$), maybe only 5 of 20 flips came up heads. You don't care that the finite-population proportion of heads was 1/4; you care about the $p = 1/2$ property of the coin itself. You still have uncertainty about p even after observing all 20 outcomes.

Other Examples

Consider the employment status of individuals in Missouri. A finite population is more appropriate if you want to document the actual percentage of Missouri individuals unemployed last week. A superpopulation is more appropriate if you want to learn about the underlying mechanism that relates education and unemployment. That is, knowing each individual's employment status fully answers the first question, but not the second question.

Consider the productivity of employees at your company (you're the CEO). If you want to know each employee's productivity over the past fiscal quarter, then a finite population is more appropriate. If you want to learn how a particular company policy affects productivity, then a superpopulation is more appropriate. That is, knowing each employee's productivity fully answers the first question, but not the second question.

Discussion Question 2.2 (student data)**.** Imagine you're a high school principal. You have data on every student, including their standardized test scores from last spring.

a) Describe a specific question for which the finite population is most appropriate, and explain why.
b) Describe a specific question for which an infinite population or superpopulation is most appropriate, and explain why.

2.3 Description: Population Mean

Like most econometrics textbooks, this textbook models the population as a probability distribution. Section 2.2 helps you distinguish when this is appropriate.

Description of a population is thus description of a probability distribution. Some distributions are completely described by a single number, like a coin's probability of heads. Others are very complicated, so they are summarized by particular features like the mean and standard deviation.

Specifically, the population mean is discussed here. A random variable's **mean** is a probability-weighted average of its possible values. This is the most important feature for understanding the rest of this textbook (regression, average treatment effects, etc.). Other distributional features are also important, and I hope you can learn about them in another class/book, perhaps someday in "Distributional and Nonparametric Econometrics" (Kaplan, 2020).

Remember: there is no data yet. In practice (and starting in Section 3.4), you use data to learn about the population, to answer questions about description, prediction, or causality. Here, we consider what could possibly be learned, specifically for description.

In Sum: Population Mean for Different Variable Types

Binary: $\mathrm{E}(Y) = \mathrm{P}(Y = 1)$ in (2.3)
Discrete: $\mathrm{E}(Y) = \sum_{j=1}^{J} \mathrm{P}(Y = y_j) y_j$ in (2.4); same units of measure as Y
Categorical: no mean
Continuous: qualitatively similar to discrete; same units as Y
Linearity: $\mathrm{E}(aY + bZ) = a\,\mathrm{E}(Y) + b\,\mathrm{E}(Z)$ as in (2.9)

2.3.1 Binary Variable

A **binary variable** has two possible values. Other terms for a binary variable are **dummy variable**, **indicator variable**, and **Bernoulli random variable**. In economics, "dummy" and "binary" are most common.

Unless otherwise specified, a binary variable's two possible values are 0 and 1. For writing mathematical models, these values are usually more convenient than values like "heads" and "tails." Mathematically, this can be indicated by $Y \in \{0, 1\}$: the value of Y must be in the set that includes only the numbers 0 and 1. (Notationally: the set $\{0, 1\}$ is different than the interval $[0, 1]$ that also contains 0.23 and 0.444 and all other real decimal numbers between 0 and 1.)

Mathematically, binary variables are often defined using the **indicator function**. The indicator function $\mathbb{1}\{\cdot\}$ equals 1 if the argument is true and 0 if false:

$$\mathbb{1}\{A\} = \begin{cases} 1 & \text{if } A \text{ is true} \\ 0 & \text{if } A \text{ is false.} \end{cases} \tag{2.1}$$

Example 2.4. Many important variables are binary:

- whether the economy is in a recession (1) or not (0);
- whether somebody has a college degree (1) or not (0);
- whether a pharmaceutical drug is branded (1) or generic (0);
- whether somebody is employed (1) or not (0);
- whether a retailer is a franchise (1) or not (0).

Example 2.5 (Kaplan video)**.** Consider defining a binary random variable Y based on the coin flip random variable W. Recall that the possible values of the flip are $W = h$ (heads) and $W = t$ (tails). We now want $Y = 1$ to indicate heads, and $Y = 0$ tails. Mathematically,

$$Y = \mathbb{1}\{\text{heads}\} = \mathbb{1}\{W = h\} = \begin{cases} 1 & \text{if } W = h \text{ (heads)} \\ 0 & \text{if } W = t \text{ (tails).} \end{cases} \tag{2.2}$$

Other examples can also be written with an indicator function. For example, $Y = \mathbb{1}\{\text{recession}\}$, $Y = \mathbb{1}\{\text{branded}\}$, or $Y = \mathbb{1}\{\text{franchise}\}$.

For binary Y, the mean $\mathrm{E}(Y)$ is

$$\mathrm{E}(Y) = \sum_{j=0}^{1} (j)\,\mathrm{P}(Y = j) = \overbrace{(0)\,\mathrm{P}(Y = 0)}^{j=0} + \overbrace{(1)\,\mathrm{P}(Y = 1)}^{j=1} = \mathrm{P}(Y = 1). \tag{2.3}$$

Thus, we can interpret the mean as the probability of $Y = 1$.

For terminology, the mean $\mathrm{E}(Y)$ is also called the **expected value** or **expectation**. These names explain the letter E in the mathematical notation.

However, the terms "expectation" and "expected value" cause much confusion. They are technical terms whose meanings differ greatly from their common colloquial English meaning. For example, if you say in plain English, "I expect the value will be 0.5," it means you think there's a high probability that the value will exactly equal 0.5. This is not what $\mathrm{E}(Y) = 0.5$ means. In fact, with a binary Y, it is impossible to have $Y = 0.5$, even if $\mathrm{E}(Y) = \mathrm{P}(Y = 1) = 0.5$. We may expect (colloquially) $Y = 1$ if $\mathrm{P}(Y = 1)$ is high, or we may expect (colloquially) $Y = 0$ if $\mathrm{P}(Y = 0)$ is high, but it is impossible to have $Y = \mathrm{E}(Y)$ (unless $p = 1$ or $p = 0$), which is very confusing. I suggest you say to yourself "mean" every time you see $\mathrm{E}(Y)$ (or "expected value" or "expectation").

2.3.2 Discrete Variable

A binary variable is a special case of a **discrete variable**, which has any countable number of possible values.

Example 2.6. Many important variables are discrete:

- an individual's years of education;
- number of children in a household;
- the number of times a stock has split since its IPO;
- the number of trading partners a country has;
- number of students in a classroom.

The units of measure are important for interpreting a discrete variable and its distribution. For most discrete variables, like number of children, the units are obvious. Sometimes it is not immediately obvious, like "number of students"... per room, or per grade, or per school? Or "number of bills passed"... in one month, or one year, or one term?

Generalizing the binary mean in (2.3), the mean of discrete Y can be written in terms of the J possible values y_j ($j = 1, \ldots, J$) and their probabilities:

$$\mathrm{E}(Y) = \sum_{j=1}^{J} \mathrm{P}(Y = y_j) y_j = \mathrm{P}(Y = y_1) y_1 + \cdots + \mathrm{P}(Y = y_J) y_J. \tag{2.4}$$

If Y is binary, then $J = 2$, $y_1 = 0$, and $y_2 = 1$, in which case (2.4) simplifies to (2.3).

The mean gives a rough sense of whether the distribution has high or low values, weighted by their probability.

Example 2.7 (Kaplan video)**.** Consider $J = 3$, and the possible values are 1, 2, and 3, so $y_1 = 1$, $y_2 = 2$, and $y_3 = 3$. Imagine $\mathrm{P}(Y = y_j) = 1/3$ for all $j = 1, 2, 3$. Plugging all these values into (2.4),

$$\mathrm{E}(Y) = \overbrace{y_1}^{1} \overbrace{\mathrm{P}(Y = y_1)}^{1/3} + \overbrace{y_2}^{2} \overbrace{\mathrm{P}(Y = y_2)}^{1/3} + \overbrace{y_3}^{3} \overbrace{\mathrm{P}(Y = y_3)}^{1/3} = (1/3) + (2/3) + (1) = 2. \tag{2.5}$$

Example 2.8. Consider random variables W and Z, with $\mathrm{P}(W = 0) = \mathrm{P}(W = 2) = 1/2$ and $\mathrm{P}(Z = 2) = \mathrm{P}(Z = 4) = 1/2$. Then,

$$\mathrm{E}(W) = (0)(1/2) + (2)(1/2) = 1, \quad \mathrm{E}(Z) = (2)(1/2) + (4)(1/2) = 3, \tag{2.6}$$

reflecting that Z tends to have higher values.

Example 2.9 (Kaplan video)**.** Imagine W and Z both have possible values $j = 1, 2, 3, 4$, but $\mathrm{P}(W = j) = j/10$, whereas $\mathrm{P}(Z = j) = (5 - j)/10$. Although the possible values are identical, W has higher probability of the higher values, which is reflected by its larger mean:

$$\begin{aligned}
\mathrm{E}(W) &= \sum_{j=1}^{4} (j)(j/10) = (1)(1/10) + (2)(2/10) + (3)(3/10) + (4)(4/10) = 3, \\
\mathrm{E}(Z) &= \sum_{j=1}^{4} (j)(5 - j)/10 = (1)(4/10) + (2)(3/10) + (3)(2/10) + (4)(1/10) = 2.
\end{aligned}$$

Beware: the mean is sensitive to very large values, so it does not reflect the value of the "average member of the population." (Such very large values are often called "outliers.") If you want to learn about the "average person" (or average firm, or average hospital, etc.), then you want the median, not the mean.

Example 2.10. Let Y denote hourly wage ($/hr) for a population with three equally-likely types of individuals. The possible values are $y_1 = 10$, $y_2 = 20$, and $y_3 = 270$. The probabilities are $\mathrm{P}(Y = y_j) = 1/3$ for $j = 1, 2, 3$. The "average person" is in the middle, paid $20/hr. But the mean is, in $/hr,

$$\mathrm{E}(Y) = (10)(1/3) + (20)(1/3) + (270)(1/3) = 300/3 = 100. \tag{2.7}$$

This $100/hr mean wage is way higher than what the "average person" earns ($20/hr). The reason is that the extremely high value $270/hr brings the mean way up.

Example 2.11 (Kaplan video)**.** Let $\mathrm{P}(Y=10)=0.99$ and $\mathrm{P}(Y=3010)=0.01$. The "average person" is one of the 99% who make \$10/hr, but the mean is four times larger, \$40/hr:

$$\mathrm{E}(Y)=(10)(0.99)+(3010)(0.01)=9.9+30.1=40. \tag{2.8}$$

The mean helps capture the aggregate earnings rate of the population as a whole, but it does not capture the typical wage of the average population member.

The mean has a useful property called **linearity**. Formally: the mean of a linear combination of random variables equals the linear combination of their means. For example, given random variables Y and Z, and non-random constants a and b,

$$\mathrm{E}(aY+bZ)=a\,\mathrm{E}(Y)+b\,\mathrm{E}(Z). \tag{2.9}$$

Here, $aY+bZ$ is a linear combination of random variables Y and Z. Thus, the mean of the linear combination of Y and Z equals the linear combination of the means $\mathrm{E}(Y)$ and $\mathrm{E}(Z)$.

2.3.3 Categorical or Ordinal Variable

A **categorical variable**'s possible values are "categories," not numbers. This was true of several examples in Section 2.3.1, like whether a retailer is a franchise or not, or whether a pharmaceutical drug is branded or generic. In Section 2.3.1, such values were coded as 0 or 1 for convenience. Generally, categorical variables can have more than two possible values.

Example 2.12. Many important variables are categorical:

- non-franchise retailers could be categorized further as national chain, regional chain, or independent;
- geographic region (north, south, east, west);
- mode of transportation (car, bike, train, etc.);
- industry (like NAICS industry code);
- college major (economics, English, ecology, electrical engineering, etc.).

The previous examples' categories have no particular order to them, so they constitute **nominal variables** (or **nominal categorical variables**). Often these are simply called categorical variables.

In contrast, an **ordinal variable** (or **ordinal categorical variable**) has categories with a natural order, usually from "low" to "high."

Example 2.13. Many important variables are ordinal:

- bond rating (e.g., D, C, . . . , AA+, AAA);
- self-reported health status (poor, fair, good, excellent);
- teaching evaluation responses (disagree, neutral, agree);
- letter grades (F, C, B, A; although often A is 4.0 and C is 2.0, there is nothing intrinsic in the letter grade system that suggests A is exactly twice as good as C).

Categorical variables, whether nominal or ordinal, do not have a mean. It is not possible to average “good” and “excellent” because we don’t have numeric values for such categories; there’s no such thing as “good and a half.” (Some people assign numbers like 3 and 4, but these are totally arbitrary, so the resulting analysis may also be totally arbitrary; don’t do that.)

However, categorical variables can be used to generate binary variables, which do have a mean. Although laters chapters do not explicitly allow categorical variables, you could incorporate them (into regressions and such) by defining appropriate dummy variables.

Example 2.14 (Kaplan video)**.** Consider a teaching evaluation response with possible values “disagree,” “neutral,” and “agree.” Using the indicator function from (2.1), define $W = \mathbb{1}\{\text{disagree}\}$, $X = \mathbb{1}\{\text{neutral}\}$, $Y = \mathbb{1}\{\text{agree}\}$. Then using (2.3), $\mathrm{E}(W) = \mathrm{P}(W = 1)$ is the probability of “disagree,” $\mathrm{E}(X) = \mathrm{P}(X = 1)$ is the probability of “neutral,” and $\mathrm{E}(Y) = \mathrm{P}(Y = 1)$ is the probability of “agree.”

2.3.4 Continuous Variable

A **continuous variable** differs from a discrete variable in some strange technical ways, but the intuition is the same. This textbook primarily uses discrete variables because the math is simpler. You could imagine a continuous variable as a discrete variable with a very (infinitely) large number of possible values packed very (infinitely) tightly together. Indeed, many variables typically treated as “continuous” are actually discrete, like monetary values that have discrete units (like $0.01). Practically, the difference is negligible.

Example 2.15. Many important variables are modeled as continuous:

- market concentration measures (like market share of largest firm or HHI);
- a country’s per capita annual meat consumption;
- percentage growth of GDP (or sales, or stock price, etc.);
- crime rates (e.g., a city’s number of property crimes per year per 10,000 people).

Always specify **units of measure**. For example, if Y is the distance from an individual’s residence to their workplace, it is meaningless to say $Y = 15$ because 15 is just a number, not a measure of distance. It could be 15 km, but it could also be 15 mi, which is 24 km; or it could even be measured in meters or feet (or parsecs, though unlikely). The mean shares the same units as the variable itself. Units always matter greatly, whether for description, prediction, or causality.

Continuous random variables share the same intuition for the mean and the same linearity property from (2.9). Computing the mean of a continuous random variable by hand requires calculus, so it is not covered in this textbook.

2.4 Prediction: Developing Intuition

⟹ Kaplan video: Intuition for Prediction

What does **prediction** mean? It may seem surprising to discuss prediction without any data, and with a completely known distribution. In English, usually prediction means using what you know now to "predict" what will happen in the future (e.g., "Beware the Ides of March!"). In econometrics and statistics, prediction shares the qualities of guessing something unknown using something known, but the details differ. (Predicting the future is a special case of prediction called forecasting; see Part III.)

To develop intuition, this section introduces prediction concepts through a simple example. The two main goals are

1. to show that there is no single best prediction because "best" depends on the ultimate purpose of the prediction, and
2. to begin translating intuition into formal mathematics.

Toward the first goal, the view of "prediction" here is broad, perhaps verging on "statistical decision theory" (making optimal decisions using statistics).

In the running example, you predict whether it will rain, where

$$\begin{aligned}
Y &: \text{random variable representing rain } (Y = 1) \text{ or no rain } (Y = 0),\\
y &: \text{realized value of } Y,\\
g &: \text{your guess/prediction of rain } (g = 1) \text{ or no rain } (g = 0),\\
L(y, g) &: \textbf{loss function} \text{ quantifying how bad it is to have guessed } g \text{ when it's really } y.
\end{aligned}$$

The loss function is essentially a negative utility function. If you had a utility function $u(y, g)$ that says how good it is to have guessed g when the truth is y, then you could simply use $L(y, g) = -u(y, g)$. Because higher loss is bad, good consequences can be represented by negative values like $L(0, 0) = -10$, or alternatively loss functions can be normalized to have $L(y, y) = 0$ by expressing loss relative to that best-guess case ($g = y$).

Understanding the role of the loss function is crucial. I have seen PhD students puzzled by their results because they did not use an appropriate loss function. Even the fanciest machine learning predictions cannot choose your loss function for you (and they may default to something totally inappropriate for your application).

2.4.1 Easy: "Predict" Current Weather

Imagine you're standing outside, and you want to "predict" whether or not it's currently raining. This is the "after" view of Section 2.1.1: instead of multiple possible values of random variable Y, you see the realized value y, with no uncertainty.

You make a simple \$1 bet with your friend: if you guess right ($g = y = 0$ or $g = y = 1$) then you win \$1, but if you guess wrong ($g \neq y$) then you lose \$1. For simplicity, imagine you have a linear utility function (no risk aversion), so the loss function is just how much money you lose (so negative is winning):

$$L(0, 0) = L(1, 1) = -1, \quad L(0, 1) = L(1, 0) = 1. \tag{2.10}$$

Obviously, you guess $g = y$. You are correct. You win \$1.

To formally show that $g = y$ is optimal, compute $L(y, 0)$ and $L(y, 1)$, and pick g to minimize loss. If $y = 1$, so $L(y, 0) = 1$ and $L(y, 1) = -1$, then "guessing" $g = 1$ minimizes loss because $-1 < 1$. If $y = 0$, so $L(y, 0) = -1$ and $L(y, 1) = 1$, then "guessing" $g = 0$ minimizes loss because $-1 < 1$. Thus, the best "guess" is indeed $g = y$.

2.4.2 Minimizing Mean Loss

Consider predicting tomorrow's weather Y if $\mathrm{P}(Y = 1) = 0.4$ (40% probability of rain) and $\mathrm{P}(Y = 0) = 0.6$. For your bet, should you predict rain?

Analogous to maximizing mean utility in microeconomics, we can choose g to minimize mean loss. This doesn't guarantee winning every time, but over the long-run (if you bet many times), it generates the lowest total loss. (Mean loss is sometimes called "expected loss" or "risk," but those phrases are more confusing due to their different meanings in common English.)

The mean loss is computed separately for each possible guess, here $g = 0$ and $g = 1$. With $g = 0$, $L(Y, 0)$ is a random variable: using (2.10), its possible values are -1 (if $Y = 0$) and 1 (if $Y = 1$). That is, $L(Y, 0)$ is a random variable with

$$\mathrm{P}(L(Y, 0) = -1) = \mathrm{P}(Y = 0) = 0.6, \quad \mathrm{P}(L(Y, 0) = 1) = \mathrm{P}(Y = 1) = 0.4. \tag{2.11}$$

Thus, using the expected value formula (2.4), mean loss when guessing no rain ($g = 0$) is

$$\mathrm{E}[L(Y, 0)] = \mathrm{P}(Y = 0)L(0, 0) + \mathrm{P}(Y = 1)L(1, 0) = (0.6)(-1) + (0.4)(1) = -0.2. \tag{2.12}$$

Similarly, with $g = 1$,

$$\mathrm{E}[L(Y, 1)] = \mathrm{P}(Y = 0)L(0, 1) + \mathrm{P}(Y = 1)L(1, 1) = (0.6)(1) + (0.4)(-1) = 0.2. \tag{2.13}$$

The optimal guess is $g = 0$ because it minimizes mean loss: $\mathrm{E}[L(Y, 0)] < \mathrm{E}[L(Y, 1)]$.

2.4.3 Different Probability

Consider Section 2.4.2 but with $\mathrm{P}(Y = 1) = 0.7$. Intuitively, rain being more likely might change the optimal prediction from "no rain" to "rain."

Analogous to (2.12) but with $\mathrm{P}(Y = 1) = 0.7$, mean loss for $g = 0$ is

$$\mathrm{E}[L(Y, 0)] = \mathrm{P}(Y = 0)L(0, 0) + \mathrm{P}(Y = 1)L(1, 0) = (0.3)(-1) + (0.7)(1) = 0.4. \tag{2.14}$$

Similarly, mean loss for $g = 1$ is

$$\mathrm{E}[L(Y, 1)] = \mathrm{P}(Y = 0)L(0, 1) + \mathrm{P}(Y = 1)L(1, 1) = (0.3)(1) + (0.7)(-1) = -0.4. \tag{2.15}$$

Opposite Section 2.4.2, now $\mathrm{E}[L(Y, 1)] < \mathrm{E}[L(Y, 0)]$: $g = 1$ minimizes mean loss, so $g = 1$ is the best prediction for your bet. This shows how the distribution of Y can affect the optimal prediction.

2.4.4 Different Loss Function

Now consider $\mathrm{P}(Y = 1) = 0.4$ as in Section 2.4.2, but with a different loss function. Specifically, if you correctly predict rain, you win \$10 (i.e., -10 loss), but otherwise $L(y, g)$ is the same as in (2.10):

$$L(0,0) = -1, \quad L(1,1) = -10, \quad L(0,1) = L(1,0) = 1. \tag{2.16}$$

Intuitively, even though rain is less probable than no rain, the much larger payoff for correctly predicting rain might make us want to bet on rain.

Given the loss function in (2.16) and $\mathrm{P}(Y = 1) = 0.4$, the mean loss for $g = 0$ is

$$\mathrm{E}[L(Y,0)] = \mathrm{P}(Y = 0)L(0,0) + \mathrm{P}(Y = 1)L(1,0) = (0.6)(-1) + (0.4)(1) = -0.2. \tag{2.17}$$

Similarly, mean loss for $g = 1$ is

$$\mathrm{E}[L(Y,1)] = \mathrm{P}(Y = 0)L(0,1) + \mathrm{P}(Y = 1)L(1,1) = (0.6)(1) + (0.4)(-10) = -3.4. \tag{2.18}$$

Opposite Section 2.4.2, now $\mathrm{E}[L(Y,1)] < \mathrm{E}[L(Y,0)]$: $g = 1$ minimizes mean loss, so $g = 1$ is the best prediction for your bet. This shows how the loss function can affect the optimal prediction.

2.5 Prediction: Generic Results

⟹ Kaplan video: Optimal Prediction

2.5.1 Optimal Prediction: Generic Example with Two Choices

In Sum: Optimal Prediction

1. Choose appropriate loss function $L(y, g)$: quantifies how bad it is to guess g when the true value is y
2. Optimal prediction: the value of g with smallest mean loss $\mathrm{E}[L(Y, g)]$

The following generalizes the approach of Section 2.4.2. The possible values of Y are now $Y = a$ and $Y = b$ (before, $a = 0$ was no rain, and $b = 1$ was rain). Like before, there are only two choices of g: $g = a$ or $g = b$ (you could only guess rain, or no rain).

Step 1 is to write down the loss function, based on the consequences of correct and incorrect predictions, like in (2.10) and (2.16). That is, write out the numeric values of $L(a, a)$, $L(a, b)$, $L(b, a)$, and $L(b, b)$.

Step 2 is to compute the mean loss for each possible guess g. Generalizing (2.12) and (2.13),

$$\begin{aligned} \mathrm{E}[L(Y,a)] &= \mathrm{P}(Y=a)L(a,a) + \mathrm{P}(Y=b)L(b,a), \\ \mathrm{E}[L(Y,b)] &= \mathrm{P}(Y=a)L(a,b) + \mathrm{P}(Y=b)L(b,b). \end{aligned} \tag{2.19}$$

Step 3 is to pick the g that minimizes $\mathrm{E}[L(Y,g)]$. If $\mathrm{E}[L(Y,a)] < \mathrm{E}[L(Y,b)]$, then $g=a$ is the optimal predictor; if $\mathrm{E}[L(Y,b)] < \mathrm{E}[L(Y,a)]$, then $g=b$ is the optimal predictor; or if $\mathrm{E}[L(Y,a)] = \mathrm{E}[L(Y,b)]$, then $g=a$ and $g=b$ are equally good (or, equally bad!).

Example 2.16. Let $\mathrm{P}(Y=a)=0.7=1-\mathrm{P}(Y=b)$. Step 1: imagine $L(a,a)=L(b,b)=0$, $L(a,b)=5$, $L(b,a)=7$. Using (2.19), Step 2 yields

$$\mathrm{E}[L(Y,a)] = \mathrm{P}(Y=a)L(a,a) + \mathrm{P}(Y=b)L(b,a) = (0.7)(0) + (0.3)(7) = 2.1, \tag{2.20}$$

$$\mathrm{E}[L(Y,b)] = \mathrm{P}(Y=a)L(a,b) + \mathrm{P}(Y=b)L(b,b) = (0.7)(5) + (0.3)(0) = 3.5. \tag{2.21}$$

Step 3 says $g=a$ is the optimal prediction because $\mathrm{E}[L(Y,a)] < \mathrm{E}[L(Y,b)]$.

Example 2.17. Imagine you work at a carnival where people pay five tickets to see if you can guess their age. If you guess correctly, they win nothing; if incorrect, they win a big stuffed animal. Because they pay five tickets regardless of your guess, that does not enter the loss function. For simplicity, imagine everyone is either 20 or 25 years old, with $\mathrm{P}(Y=20)=0.6$ and $\mathrm{P}(Y=25)=0.4$. Step 1: let s be the value of the stuffed animal that's "lost" if you're wrong, so $L(20,25)=L(25,20)=s$, whereas $L(20,20)=L(25,25)=0$. Step 2: mean losses are

$$\begin{aligned} \mathrm{E}[L(Y,20)] &= (0.6)L(20,20) + (0.4)L(25,20) = (0.6)(0) + (0.4)(s) = 0.4s, \\ \mathrm{E}[L(Y,25)] &= (0.6)L(20,25) + (0.4)L(25,25) = (0.6)(s) + (0.4)(0) = 0.6s. \end{aligned}$$

Step 3: because $0.4s < 0.6s$, it's better to guess $g=20$.

2.5.2 Quadratic Loss and the Mean

Just as the population mean is useful for description (Section 2.3), it is also useful for prediction. Specifically, it is the optimal predictor when a particular loss function is used: quadratic loss.

Define **quadratic loss** as

$$L_2(y,g) = (y-g)^2. \tag{2.22}$$

This is zero when the guess is perfect ($g=y$) and larger when g is farther from y. Thus, quadratic loss distinguishes between a slightly-wrong guess and a really-wrong guess.

Quadratic loss is most useful with continuous or discrete Y, rather than the binary special case of Section 2.5.1. If there are very many or infinite possible Y values, then it may be impractical or impossible to write down every possible $L(y,g)$ value. A formula like $(y-g)^2$ is convenient and can be used if it's a reasonable approximation of the true loss in each case. (Sometimes we may not even know the true loss because we do not know how our prediction will eventually affect decisions.)

Example 2.18 (Kaplan video)**.** Let the true $y = 100$. Consider quadratic loss for different g. The guess $g = y = 100$ is best because $L_2(100, 100) = 0$ is the smallest possible loss (because $(y - g)^2 < 0$ is impossible). The guess $g = 99$ is worse: loss is $L_2(100, 99) = (100 - 99)^2 = 1$. The guess $g = 90$ is even worse: $L_2(100, 90) = (100 - 90)^2 = 100$. Further, even though 90 is only 10 times farther from y than 99, the loss is 100 times as big. Also, the guess $g = 110$ is just as bad as 90 because they are both wrong by 10 (higher or lower doesn't matter): $L_2(100, 110) = (100 - 110)^2 = 100$.

There are some cases when quadratic loss is inappropriate. For example, sometimes it may be much worse to over-predict ($g > y$) than under-predict ($g < y$), or vice-versa. However, quadratic loss does not distinguish between over-prediction and under-prediction because $(y - g)^2 = (g - y)^2$. As another example, sometimes it may be twice as bad to over-predict by 20 units ($g - y = 20$) than 10 units ($g - y = 10$), but quadratic loss says it's four times worse because $10^2 = 100$ and $20^2 = 400$. As another example, quadratic loss is inappropriate when it only matters if you are correct or not, like in the rain bet example (you can't guess $g = 0.5$ and be half-wrong).

Discussion Question 2.3 (banana loss function)**.** Imagine you run a small banana shop. You buy bananas wholesale for 2 cents each (\$0.02) and sell each for 40 cents (\$0.40). The wholesaler delivers every Monday. Any bananas not sold by the next Monday spoil; you cannot sell them (they just go in the compost). Let y be the actual number of bananas that customers want to buy in some week. Let g be your guess, i.e., how many you bought wholesale on Monday.

a) Consider the loss function $L(y, g) = 0$ if $y = g$, otherwise $L(y, g) = 1$ if $y \neq g$; that is, zero loss for a correct guess, loss of 1 for incorrect guess. Why isn't that loss function appropriate?
b) Why isn't quadratic loss appropriate?
c) What might the loss function look like, if you only care about maximizing profit? Try to be as specific and mathematical as you can. In particular, consider the different consequences of over-buying ($g > y$) versus under-buying ($g < y$).

Under quadratic loss, the mean is the optimal predictor that minimizes mean loss. Although the details are beyond our scope,

$$g_2^* \equiv \arg\min_g \mathrm{E}[(Y - g)^2] = \mathrm{E}(Y). \tag{2.23}$$

That is, g_2^* is defined as the value of g that minimizes $\mathrm{E}[(Y - g)^2]$, and it happens to equal the population mean $\mathrm{E}(Y)$.

This says the population mean $\mathrm{E}(Y)$ has two interpretations. For description (Section 2.3), the mean helps summarize the "center" of the distribution. For prediction, the mean is the "best" guess of an unknown value of Y, given quadratic loss.

Discussion Question 2.4 (optimal banana prediction)**.** Consider the same setup as in DQ 2.3, and again assume you want to maximize (mean) profit. Imagine you know the distribution of Y (banana quantity demanded in one week).

a) Do you think the mean $\mathrm{E}(Y)$ is a good "predicted" number of bananas to buy wholesale? Explain why or why not; if not, also explain why you think $\mathrm{E}(Y)$ is too high or too low.

b) What if the retail price were \$99 per banana, and the wholesale cost is still \$0.02 per banana; would $\mathrm{E}(Y)$ be good, or too high, or too low, and why?

In Sum: Quadratic Loss and the Mean

Quadratic loss: $L(y, g) = (y - g)^2$
Population mean: $\mathrm{E}(Y)$ is the best guess of Y given quadratic loss

Optional Resources

Optional resources for this chapter

- Basic probability: the Khan Academy AP Statistics unit includes instructional material and practice questions
- Mean (expected value) (Lambert video)
- Probability distribution basics on Wikipedia (more than you need to know for this class)
- Optimal prediction: Hastie, Tibshirani, and Friedman (2009, §2.4)
- Section 2.1 ("Random Variables and Probability Distributions") in Hanck et al. (2018)

Chapter 3

One Variable: Sample

⟹ Kaplan video: Chapter Introduction

Sections 2.3 and 2.5 considered only the population distribution, whereas Chapter 3 considers data sampled from that distribution. The words **data**, **dataset**, **sample values**, and **sample** all refer to the same thing: the set of values that the researcher actually sees. But, as in Chapter 2, this could be seen either from the "before" perspective as random variables, or from the "after" perspective as non-random realized values. Section 2.1 gave the general idea of seeing observations as random variables (the "before" view); here, specific details are provided on estimation and uncertainty.

Although long, this chapter is mostly review of material you should have seen already in an introductory statistics class.

Unit learning objectives for this chapter

3.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

3.2. Describe and distinguish Bayesian and frequentist perspectives [TLO 4]

3.3. Identify and interpret properties of a sampling procedure or estimator [TLO 4]

3.4. Judge which estimator is better based on its properties [TLO 6]

3.5. Interpret different measures of statistical uncertainty [TLOs 6 and 7]

3.6. Assess the economic significance of empirical results [TLO 6]

3.7. In R (or Stata): compute estimates of a population mean along with measures of uncertainty [TLO 7]

3.1 Bayesian and Frequentist Perspectives

⟹ Kaplan video: Bayesian and Frequentist Perspectives

Two frameworks constitute econometrics and statistics: **Bayesian** and **frequentist** (or **classical**). These are cynically deemed "sects" by some, but outside the vocal extremes (and amusing webcomics: `xkcd.com/1132`), most econometricians appreciate and respect both frameworks (and the people who use them), sometimes working with both in turn.

This textbook uses the frequentist framework. Why? Mostly, that's just how I wrote it; I'll spare you post hoc rationalization.

There is little disagreement about the population and what we want to learn. Generally, both Bayesian and frequentist perspectives agree on everything in Chapter 2 about the population and how data are generated.

The disagreements are about how to use the sampled data to learn about the population, as briefly described in the remainder of Section 3.1. At minimum, I hope you get a sense of these two different ways of quantifying uncertainty, and the different types of questions they can (and cannot) answer.

3.1.1 Very Brief Overview: Bayesian Approach

The Bayesian approach models your beliefs about an unknown population value θ, like the mean $\theta = \mathrm{E}(Y)$. Your **prior** (or prior belief) is what you believe about θ before seeing the data. Your **posterior** (or posterior belief) is what you believe about θ after seeing the data. The Bayesian approach describes how to update your prior using the observed data, to get your posterior.

Mathematically, "belief" is a probability distribution. For example, let random variable B represent your belief about the population mean. If you think there's a 50% chance the mean is negative, then $\mathrm{P}(B < 0) = 50\%$. If you think there's a $1/4$ probability that B is below -1, then $\mathrm{P}(B < -1) = 1/4$. (Elsewhere, you may see this written more confusingly as $\mathrm{P}(\theta < 0)$ and $\mathrm{P}(\theta < -1)$.)

For example, imagine you see a bird flying in your backyard, and you grab your binoculars to try to identify it. Let θ represent the true species, while B is your belief. Imagine (for simplicity) you only ever see three types of bird in your backyard, all woodpeckers: downy, hairy, and red-bellied, written $\theta = d$, $\theta = h$, and $\theta = r$. Based on the location and habitat, you know hairy is somewhat less likely in general, so your prior is $\mathrm{P}(B = d) = \mathrm{P}(B = r) = 0.4$, $\mathrm{P}(B = h) = 0.2$. Looking through your binoculars (looking at the data), you're pretty sure it's not the red-bellied, but it's too far to distinguish downy from hairy, so your updated posterior belief has $\mathrm{P}(B = d) = 0.6$, $\mathrm{P}(B = h) = 0.3$, $\mathrm{P}(B = r) = 0.1$. The low probability of red-bellied comes from the data, whereas the higher probability of downy than hairy comes from your prior.

The posterior distribution is the Bayesian way of quantifying uncertainty. It is relatively intuitive, similar to how people talk about uncertainty in daily life. The posterior distribution is often summarized by a **credible interval**, i.e., a range of values that you're pretty sure (like 90% sure) contains the true θ. Or in the above example with categorical θ, the **credible set** $\{d, h\}$ has 90% posterior belief: you'd say, "I'm 90% sure it's a downy or hair woodpecker, although I think there's a 10% chance I'm wrong and it's a red-bellied woodpecker."

3.1.2 Very Brief Overview: Frequentist Approach

The core of the frequentist approach is the “before” perspective, which can also be described in terms of **repeated sampling**. Instead of the belief probabilities of a Bayesian posterior, frequentist probabilities are from the “before” view of what dataset (and thus value of estimator and such) could be randomly sampled. Equivalently, as a thought experiment, we can imagine many different random samples drawn from the same population; the “before” probabilities are then how often certain values occur in these many random datasets.

For intuition, imagine you could randomly sample 100 datasets from the same population. Then, the frequentist probability of an event says approximately how many times that event occurs among the 100 samples. For example, we could compute the sample mean $\bar{Y}$ in all 100 samples; because the datasets are all different, the sample means $\bar{Y}$ are also all different. If $\bar{Y} \leq 0$ in 50 of the 100 hypothetical samples, then $\mathrm{P}(\bar{Y}_n \leq 0) \approx 50/100 = 50\%$. Or, if $\bar{Y}$ is in the interval $[-0.4, 0.4]$ in 70 of 100 samples, then $\mathrm{P}(-0.4 \leq \bar{Y} \leq 0.4) = \mathrm{P}(\bar{Y} \in [-0.4, 0.4]) \approx 70\%$. A similar example is in Table 3.1.

3.1.3 Bayesian and Frequentist Differences

The following makes explicit some of the differences between the Bayesian and frequentist approaches described above.

First, the frameworks treat different variables as random or non-random. The frequentist framework treats the population mean and other population features as non-random values, whereas it treats the data as random. For example, the population mean $\mu = \mathrm{E}(Y)$ is a non-random value, whereas an observation Y is a random variable. In contrast, the Bayesian framework treats population features as random (to reflect your beliefs), whereas it treats the data as non-random values (the “after” view).

Second, due to this different treatment, the frameworks answer different types of questions, especially when quantifying uncertainty. The Bayesian framework answers questions about our beliefs after seeing the data. The frequentist framework answers questions about probabilities of seeing different features in the data, given the true population values.

Example 3.1 (Kaplan video)**.** Consider the question, “Given the observed data, what do I believe is the probability that the population mean is above 1/2?” This is a Bayesian question. Mathematically, if y is the “observed data,” this question is commonly written as $\mathrm{P}(\mu > 1/2 \mid y)$, noting the conventional but confusing notation where μ represents beliefs. This question makes no sense from the frequentist perspective: either $\mu > 1/2$ or not; it cannot be “maybe,” with some probability.

Example 3.2 (Kaplan video)**.** Consider the question, “Given the value of $\mu = \mathrm{E}(Y)$, what’s the probability that the sample mean is above 1/2?” This is a frequentist question. Mathematically, this is usually written $\mathrm{P}(\bar{Y} > 1/2)$, or $\mathrm{P}_\mu(\bar{Y} > 1/2)$ to be explicit about the dependence on μ. The sample mean $\bar{Y}$ is a function of data, so it is treated as a random variable. This question makes no sense from the Bayesian perspective: we can see the data, so we can see either $\bar{Y} > 1/2$ or not; it cannot be "maybe," with some probability.

Interestingly, both frameworks can answer questions like $\mathrm{P}(\bar{Y} < \mu)$, but with different interpretations. The Bayesian answer interprets $\bar{Y}$ as a number (that we see in the data) and μ as a random variable representing our beliefs. The frequentist answer interprets $\bar{Y}$ as the random variable (from the "before" view) and μ as the non-random population value.

Third, frequentist methods use only the data, whereas Bayesian methods can formally incorporate additional knowledge. In practice, though, even frequentist results should be interpreted in light of other knowledge. The difference is that this process is not formalized within the frequentist methodology itself. Unfortunately, many people do not combine frequentist results with other knowledge, instead interpreting frequentist results as if one single dataset contains the full, absolute truth of the universe; please do not do this!

In Sum: Bayesian & Frequentist

Frequentist: "before" view of data (random variables); assess methods' performance across repeated random samples from same population
Bayesian: "after" view of data (non-random); model beliefs (about population features) as random variables

3.2 Types of Sampling

⟹ Kaplan video: Types of Sampling

In practice, judging which econometric method is most appropriate requires understanding different types of sampling procedures and sampling properties. Such judgment is mostly left to another textbook, but this section hopes to help your understanding.

Notationally, we observe the values from n **units**, which could be individuals, firms, countries, etc. Let $i = 1$ refer to the first unit, $i = 2$ to the second, etc., up to $i = n$, where n is the **sample size**. The corresponding values are $Y_1, Y_2, \ldots, Y_n$, with Y_i more generally denoting the observation for unit i. A particular dataset may have specific values like $Y_1 = 5$, $Y_2 = 8$, etc., but to analyze statistical properties, each Y_i is seen as a random variable as in Section 2.1.

In this section, two important sampling properties are considered: "independent" and "identically distributed." If both hold, then the Y_i are called **independent and identically distributed** (iid) random variables (or "sampled iid"), and "sampling is iid." Sometimes the vague phrase **random sample** refers to iid sampling.

This iid sampling is mathematically simplest but not always realistic. Although iid sampling is the focus here (like other introductory textbooks), weights are briefly mentioned, and Part III considers dependent (i.e., not independent) data.

Notationally, iid sampling is indicated by $\overset{iid}{\sim}$. For example, with population CDF $F_Y(\cdot)$,

$$Y_i \overset{iid}{\sim} F_Y, \quad i = 1, \ldots, n. \tag{3.1}$$

The F_Y can be replaced by another distribution function or name.

There are other sampling properties not considered in this section, like **sampling bias**. This is about whether we observe a "representative sample" of the population we want to learn about (the population of interest). Sometimes sampling bias is our fault (for using the wrong dataset for our economic question), but sometimes we try to get the right data and people refuse to answer our survey, or we can't get access to certain confidential data, etc. This is discussed more in Chapter 12, in terms of "missing data" and "sample selection."

After introducing "independent" and "identically distributed" sampling, examples are discussed in Section 3.2.3.

3.2.1 Independent

Qualitatively, in the context of sampling, **independence** (or independent sampling) means that from the "before" view, any two observations are unrelated. For example, the value of Y_2 is unrelated to Y_1: we are not any more likely to see a high Y_2 if we see a high Y_1 in the sample.

Mathematically, independence means

$$Y_i \perp\!\!\!\perp Y_k \text{ for any } i \neq k, \tag{3.2}$$

where $\perp\!\!\!\perp$ denotes statistical independence. That is, $Y_1 \perp\!\!\!\perp Y_2$, $Y_1 \perp\!\!\!\perp Y_8$, $Y_6 \perp\!\!\!\perp Y_4$, etc. For any $i \neq k$, independent sampling implies (but is not implied by), among other properties,

$$\mathrm{Cov}(Y_i, Y_k) = 0, \quad \mathrm{Var}(Y_i + Y_k) = \mathrm{Var}(Y_i) + \mathrm{Var}(Y_k), \quad \mathrm{E}(Y_i \mid Y_k) = \mathrm{E}(Y_i). \tag{3.3}$$

Example 3.3 (Kaplan video)**.** You plan to flip a coin and record $Y_1 = 1$ if heads and $Y_1 = 0$ if tails. You plan flip the same coin again and record $Y_2 = 1$ if heads and $Y_2 = 0$ if tails. These are independent: $Y_1 \perp\!\!\!\perp Y_2$. Although the probabilities are very closely related (actually identical), the realization of the first flip (heads or tails) has no relationship with the second flip. For example, even if we know the first flip is heads, this does not change the probability of heads for the second flip: $\mathrm{P}(Y_2 = 1 \mid Y_1 = 1) = \mathrm{P}(Y_2 = 1)$.

Example 3.4 (Kaplan video)**.** You plan to pick a random person in the world and record how many years of formal education they've had as Y_1. You plan to then pick another random person and record their years of education in Y_2. The way you sample Y_2 has no relation to the first sampled person or their Y_1 value, so there is independence: $Y_1 \perp\!\!\!\perp Y_2$. Among other implications, this means Y_1 and Y_2 have zero correlation (uncorrelated) and zero covariance, $\mathrm{Cov}(Y_1, Y_2) = 0$.

3.2.2 Identically Distributed

The **identically distributed** property means that from the "before" view, the distribution of Y_i is the same for any i. Qualitatively, all units are sampled from the same population. Mathematically, given shared population CDF $F_Y(\cdot)$, $Y_i \sim F_Y$ for all $i = 1, \ldots, n$. This means that for any i and k, Y_i and Y_k have the same distribution and thus the same properties like mean and variance, $\text{E}(Y_i) = \text{E}(Y_k)$ and $\text{Var}(Y_i) = \text{Var}(Y_k)$.

Mathematically, identically distributed Y_i means that for any i and k, Y_i and Y_k have the same distribution. Thus, any feature of their distributions is also identical. For example, $\text{E}(Y_i) = \text{E}(Y_k)$ and $\text{Var}(Y_i) = \text{Var}(Y_k)$.

Example 3.5 (Kaplan video)**.** The Y_1 and Y_2 in Example 3.3 are identically distributed because they are from the same coin, so the probability of heads is the same each time. (Unless you cheat or flip it differently or something, but those are nuances for physics class, not econometrics.)

Practice 3.1 (i/id sampling)**.** You are planning to sample values Y_1 and Y_2, but you have not yet sampled them. Each of the following four statements implies one of the four sampling properties: 1) independent, 2) not independent (i.e., dependent), 3) identically distributed, 4) not identically distributed. Which is which?

a) You are just as likely to get $Y_1 = 3$ as $Y_2 = 3$, and similarly for any other value besides 3.
b) If you get a negative Y_1, then you'll probably get a negative Y_2; but if you get a positive Y_1, then you'll probably get a positive Y_2.
c) Separately and simultaneously, you will randomly sample Y_1 while your friend samples Y_2.
d) For Y_1 you are going to get the salary of somebody with an economics degree, and Y_2 will be the salary of somebody with an art history degree.

3.2.3 More Examples

Consider the following sampling procedures and their properties. Each example has 4 observations of Mizzou students. You can imagine 4 buckets (or pieces of paper), initially empty, that will eventually contain information from 4 observations. The sampling procedure does not determine the specific numeric values that end up in the buckets, but it determines how the buckets get filled. Again, the goal for this class is to understand why sampling is iid or not.

Example 3.6 (Kaplan video)**.** Imagine randomly picking a Mizzou student ID number, then randomly picking a 2nd, then 3rd, then 4th. The corresponding Y_i are both independent and identically distributed (iid). They are independent because each ID number is randomly drawn without any consideration of how the other numbers are drawn, and without any consideration of the other observed Y_i values. They are identically distributed because each ID number is drawn from the same population (anyone who has a Mizzou student ID).

Example 3.7 (Kaplan video)**.** Each Mizzou student is classified as either a resident of Missouri (“in-state”) or not (“non-resident”). Imagine buckets 1 and 2 say “in-state,” while buckets 3 and 4 say “non-resident”: observations Y_1 and Y_2 are from in-state students, while Y_3 and Y_4 are from non-resident students. (This is “stratified sampling”: assigning buckets to different strata before sampling.) For most variables, the in-state distribution differs from the non-resident distribution, so the distribution of Y_1 and Y_2 (in-state) differs from the distribution of Y_3 and Y_4 (non-resident). That is, sampling is not identically distributed. Thus, even if the samples are all independent, sampling is not iid.

Example 3.8 (Kaplan video)**.** Imagine randomly picking a class (like my econometrics class) at Mizzou, and filling the first two buckets (Y_1 and Y_2) with two random students from that class; then randomly picking another class, and another two students for the other buckets (Y_3 and Y_4). (This is an example of “clustered sampling,” where each class is a “cluster”; this differs from “clustering” in cluster analysis.) Observations are identically distributed (because each Y_i has the same probability of getting any particular student) but probably not independent. For example, dependence may come from students in the same class being similarly affected by their shared experience. Here, buckets 1 and 2 are correlated, and 3 and 4 are correlated, but not 1 and 3, nor 2 and 4, etc. Thus, sampling is not iid.

Example 3.9 (Kaplan video)**.** Imagine randomly picking 2 Mizzou students (like with random ID numbers), then observing them this semester and next semester. For example, imagine bucket 1 contains the first student’s GPA this semester, bucket 2 contains the same student’s GPA next semester, and buckets 3 and 4 contain the other student’s GPAs from this semester and next semester. Buckets 1 and 2 (Y_1 and Y_2) are probably both high or both low, rather than one high and one low, and similarly for buckets 3 and 4 (Y_3 and Y_4). That is, buckets 1 and 2 are correlated, and 3 and 4 are correlated. Further, observations may not even be identically distributed if fall GPA and spring GPA do not have the same distribution. Thus, sampling is not iid.

Example 3.10 (Kaplan video)**.** If you randomly pick one student and observe the same student over four consecutive semesters, there is probably dependence, in which case sampling is not iid. For example, because it’s the same student, we are more likely to see four relatively high values, or four relatively low values, or four mediocre values, than to see high-low-low-high or low-high-low-high; there is positive correlation among Y_1, Y_2, Y_3, and Y_4. (Imagine Y is height, or hours of sleep, or GPA; these may change over time, but not as much as they differ among different students.) This is time series data; see Part III.

Practice 3.2 (rural household sampling)**.** You want to learn about household consumption in rural Indonesia. In an area with 100 villages, you either i) pick 5 villages at random, then survey every household in each of the 5 villages; or ii) make a list of all households in all 100 villages, then randomly pick 5% of them. Explain why each approach is or isn’t iid.

3.3 The Empirical Distribution

⟹ Kaplan video: The Empirical Distribution

The **empirical distribution** is a probability distribution that reflects the sample data. It can be confusing at first, but it unifies many approaches in this class and beyond, helping them seem less ad hoc and mysterious. Qualitatively, the empirical distribution treats the sample as if it were the population.

Mathematically, first consider a binary variable. The population is represented by binary random variable Y with $\mathrm{P}(Y=1)=p$. A sample of size n can be represented by binary random variable S with

$$\mathrm{P}(S=1)=\hat{p}=\frac{\text{how many } Y_i=1}{n}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i=1\}, \tag{3.4}$$

the sample proportion of observations with $Y_i=1$. The distribution of S is the empirical distribution.

The **plug-in principle** or **analogy principle** suggests that we use S to compute whatever features we want to learn about Y. For example, if we want to learn $\mathrm{E}(Y)$, then compute $\mathrm{E}(S)$. With enough data, S is usually very similar to Y, so features of S should usually be very similar to those of Y.

Mathematically, consider now a categorical or discrete variable. The population is represented by random variable Y with possible values $(v_1,\dots,v_J)$. The sample is represented by random variable S with

$$\mathrm{P}(S=v_j)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{Y_i=v_j\},\quad j=1,\dots,J. \tag{3.5}$$

That is, $\mathrm{P}(S=v_j)$ is the sample proportion of observations with value v_j.

Mathematically, consider finally a continuous variable. Because the sampled Y_i values are all unique, the sample is represented by random variable S with

$$\mathrm{P}(S=Y_i)=1/n,\quad i=1,\dots,n. \tag{3.6}$$

Even though Y is continuous, S is discrete, with $1/n$ probability on each observed Y_i value.

Notationally, a **hat** (circumflex) often denotes a **sample analog**, a feature of S analogous to a population feature of Y. (More generally, a “hat” just denotes anything computed from the sample data.)

Example 3.11 (Kaplan video)**.** For the population $\mathrm{P}(Y=y)$, the sample analog is $\hat{\mathrm{P}}(Y=y)=\mathrm{P}(S=y)$, which is also the proportion of observed Y_i equal to y.

Example 3.12 (Kaplan video)**.** For the population mean $\mathrm{E}(Y)$, the sample analog is $\hat{\mathrm{E}}(Y)=\mathrm{E}(S)$, which is also the sample average of the Y_i.

3.4 Estimation of the Population Mean

Sections 2.3 and 2.5 helped us think about which features of the population are useful for description and prediction. Such a population feature is called the **estimand** or **object of interest**. In practice, it must be estimated using data.

This section specifically considers estimating the population mean from iid data. The same concepts appear in later chapters.

3.4.1 "Description": Sample Mean

The population mean can be estimated by its sample analog, the mean of the empirical distribution (Section 3.3). This is called the **sample mean**. It is also called the **sample average** because it averages the sample Y_i values. Notationally, the sample average is usually $\bar{Y}$ (or $\bar{Y}_n$). Mathematically, using Section 3.3 notation,

$$\bar{Y} = \hat{\mathrm{E}}(Y) = \mathrm{E}(S) = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{3.7}$$

These expressions are equivalent, just emphasizing different interpretations. (The last equality is not obvious for discrete Y, but it's derivation is beyond our scope.)

3.4.2 "Prediction": Least Squares

Section 2.5.2 showed that the population mean $\mathrm{E}(Y)$ also solves an optimal prediction problem: $\mathrm{E}(Y) = g_2^* \equiv \arg\min_g \mathrm{E}[(Y-g)^2]$.

The analogy principle (Section 3.3) suggests solving the same optimal prediction problem for the empirical distribution (S replacing Y). Skipping the calculus,

$$\hat{g}_2^* \equiv \arg\min_g \mathrm{E}[(S-g)^2] = \arg\min_g \frac{1}{n}\sum_{i=1}^{n}(Y_i - g)^2 \implies \hat{g}_2^* = \bar{Y}. \tag{3.8}$$

The prediction-motivated estimator equals the description-motivated estimator! As with any random variable, the mean of S equals the best predictor of S (with quadratic loss).

Rewriting (3.8) allows the introduction of some terms and concepts used in later chapters. In (3.8), the $1/n$ has no effect on the minimization problem because it is unaffected by g. Consequently, it is equivalent to write

$$\hat{g}_2^* = \arg\min_g \sum_{i=1}^{n}(Y_i - g)^2. \tag{3.9}$$

To dissect the right-hand side of (3.9), imagine any estimate $\hat{g}$. Because $\hat{g}$ can be seen as trying to predict Y, sometimes $\hat{g}$ is called the **predicted value** of Y_i. However, the observed value of Y_i is used to compute $\hat{g}$, so it seems misleading to say Y_i was "predicted": usually we assume the true value is not known when we discuss prediction. Instead, calling

$\hat{g}$ the **fitted value** is more appropriate. Either way, $\hat{U}_i = Y_i - \hat{g}$ is called the **residual** for observation i. The squared residuals are $\hat{U}_i^2 = (Y_i - \hat{g})^2$. The **sum of squared residuals** (SSR) is then

$$\sum_{i=1}^{n} \hat{U}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{g})^2. \tag{3.10}$$

Consequently, (3.8)–(3.10) together say that $\bar{Y}$ minimizes the SSR. For this reason, $\bar{Y}$ is a **least squares** estimator: "least" referring to minimization, and "squares" referring to the second S in SSR.

3.4.3 Non-iid Sampling: Weights

If your dataset has weights, then you should probably use them. Weights help adjust the sample to be more representative of the population. Conversely, ignoring weights can produce misleading results because the sample is not representative of the population. Details are beyond our scope.

3.5 Sampling Distribution of an Estimator

$\Longrightarrow$ Kaplan video: Sampling Distribution of an Estimator

Our goal in this section is to understand what it means for an estimator to have a probability distribution. (There are many interesting (to me) details of approximating sampling distributions, but they are beyond our scope.)

An estimator's **sampling distribution** is simply its probability distribution, treating the estimator as a random variable from the "before" view. Equivalently, from the repeated sampling perspective, the sampling distribution imagines computing the estimator in a large number of randomly sampled datasets from the same population, and seeing which values occur with what probability.

Consider the sample mean as an estimator of the population mean, with iid sampling. Here, the n subscript is added to $\bar{Y}_n$ because the sampling distribution depends on n. For example, the sampling distribution of $\bar{Y}_1 = Y_1$ differs from that of $\bar{Y}_2 = (Y_1 + Y_2)/2$.

From the "before" view, the sample mean $\bar{Y}_n$ is a random variable. The Y_i are all random variables, so their average is also a random variable. That is, the Y_i have multiple possible values, so the sample mean also has multiple possible values.

Example 3.13. Let $n = 1$, so $\bar{Y}_1 = Y_1$. The sampling distribution of estimator $\bar{Y}_1$ is the same as the population distribution of Y_1.

Example 3.14. Imagine binary Y with population mean $\mathrm{E}(Y) = \mathrm{P}(Y = 1) = p$. Let $n = 2$ with iid sampling. Despite the simplicity, it takes some work to derive the sampling distribution of $\bar{Y}_2$. There are four possible values of (Y_1, Y_2): $(0, 0)$, $(0, 1)$, $(1, 0)$, $(1, 1)$.

This makes three possible values of $\bar{Y}_n = (Y_1 + Y_2)/2$: 0, 1/2, or 1. Thus,

$$\begin{aligned}
\mathrm{P}(\bar{Y}_n = 0) &= \overbrace{\mathrm{P}(Y_1 = 0 \text{ and } Y_2 = 0)}^{\text{use } Y_1 \perp\!\!\!\perp Y_2} = \overbrace{\mathrm{P}(Y_1 = 0)}^{=1-p}\,\overbrace{\mathrm{P}(Y_2 = 0)}^{=1-p} = (1-p)^2, \\
\mathrm{P}(\bar{Y}_n = 1) &= \overbrace{\mathrm{P}(Y_1 = 1 \text{ and } Y_2 = 1)}^{\text{use } Y_1 \perp\!\!\!\perp Y_2} = \overbrace{\mathrm{P}(Y_1 = 1)}^{=p}\,\overbrace{\mathrm{P}(Y_2 = 1)}^{=p} = p^2, \\
\mathrm{P}(\bar{Y}_n = 1/2) &= 1 - \mathrm{P}(\bar{Y}_n = 0) - \mathrm{P}(\bar{Y}_n = 1) = 1 - (1-p)^2 - p^2 = 2p(1-p).
\end{aligned} \tag{3.11}$$

This can be interpreted from the "before" view, or in terms of repeated sampling; for example, if we randomly sample 100 datasets, then around $100p^2$ datasets should have $\bar{Y}_n = 1$.

Discussion Question 3.1 (probability of positive mean)**.** After seeing the data, you want to know the probability that the true mean is strictly positive, $\mathrm{E}(Y) > 0$. Does the frequentist sampling distribution help? If yes, explain how; if no, explain why not. Hint: recall Section 3.1.

3.5.1 Example: Values in Repeated Samples

Table 3.1 records values and events across 100 datasets randomly sampled from the same population. The population is discrete, with $\mathrm{P}(Y = j) = 1/5$ for $j = -2, -1, 0, 1, 2$, so the population mean is $\mathrm{E}(Y) = 0$. Sampling is iid, so each Y_i has the same distribution as the population Y, and all Y_i are mutually independent. Let $n = 10$.

Table 3.1: Example estimates and event probabilities.

Sample	$\bar{Y}_n$	$\mathbb{1}\{\bar{Y}_n \le 0\}$	$\mathbb{1}\{\bar{Y}_n - 0.4 \le 0 \le \bar{Y}_n + 0.4\}$
#1	0.50	0	0
#2	0.20	0	1
#3	0.00	1	1
#4	−0.10	1	1
#5	−0.50	1	0
⋮	⋮	⋮	⋮
#100	0.30	0	1
Average	0.01	52/100	67/100

Note: $\mathrm{P}(Y = j) = 0.2$ for $j = -2, -1, 0, 1, 2$, iid, $n = 10$.

Table 3.1 shows the value of $\bar{Y}_n$ computed from each sample (dataset). It shows that $\bar{Y}_n = 0.5$ in the first sample, $\bar{Y}_n = 0.2$ in the second sample, etc. This reflects the sampling distribution.

Table 3.1 also shows for each sample whether or not the sample mean $\bar{Y}_n$ is less than or equal to the population mean $\mathrm{E}(Y) = 0$, in the column labeled with $\mathbb{1}\{\bar{Y}_n \le 0\}$. That is, 1 indicates that it does, 0 indicates that it doesn't. For example, in Sample #1, $\bar{Y}_n = 0.5$, which is not negative, so $\mathbb{1}\{\bar{Y}_n \leq 0\} = 0$. In Sample #4, $\bar{Y}_n = -0.1$, which is negative, so $\mathbb{1}\{\bar{Y}_n \leq 0\} = 1$. From the frequentist view, the event $\bar{Y}_n \leq \mathrm{E}(Y)$ is "random" in that it could occur or not occur, with some probability for each possibility. The $\mathrm{E}(Y)$ is non-random, but $\bar{Y}_n$ is random, hence the event is random. The event's probability is the probability of randomly sampling a dataset in which the event occurs. The bottom row of the table says the event occurred 52 times out of 100 samples (52% of the time). Because there are only 100 samples and not ∞, this is not the exact probability, but it reflects that the event occurs slightly more than half the time.

Table 3.1 also shows for each sample whether or not the random interval $[\bar{Y}_n - 0.4, \bar{Y}_n + 0.4]$ contains the population mean $\mathrm{E}(Y) = 0$, i.e., whether or not $\bar{Y}_n - 0.4 \leq \mathrm{E}(Y) \leq \bar{Y}_n + 0.4$. The interval is "random" in the frequentist sense that it has different possible values in different datasets (because it depends on $\bar{Y}_n$). In Sample #1, the interval does not contain $\mathrm{E}(Y)$: $\bar{Y}_n = 0.5$, so the interval is $[0.5 - 0.4, 0.5 + 0.4] = [0.1, 0.9]$, which does not contain $\mathrm{E}(Y) = 0$. In Sample #2, the interval does contain $\mathrm{E}(Y)$: $\bar{Y}_n = 0.2$, so the interval is $[-0.2, 0.6]$, which contains $\mathrm{E}(Y) = 0$. The bottom row of the table says this event occurred 67 times out of 100 samples (67% of the time). This is essentially the "coverage probability" of a "confidence interval," described in Section 3.7.1.

3.6 Quantifying Accuracy of an Estimator

From the frequentist perspective, an estimator's accuracy can be quantified by comparing features of its sampling distribution to the true population value. Bias is an important, commonly mentioned property, but it is not sufficient to quantify accuracy. Mean squared error better quantifies accuracy.

Throughout, let θ be the population parameter estimated by $\hat{\theta}_n$; for example, $\theta = \mathrm{E}(Y)$ and $\hat{\theta}_n = \bar{Y}_n$.

3.6.1 Bias

Definitions

The **bias** of $\hat{\theta}_n$ compares the mean of its sampling distribution to the true population θ. Mathematically,

$$\mathrm{Bias}(\hat{\theta}_n) \equiv \mathrm{E}(\hat{\theta}_n) - \theta. \tag{3.12}$$

The bias captures if the estimator systematically differs from θ in a particular direction, i.e., how wrong the average $\hat{\theta}_n$ is.

There are four types of bias:

$$\begin{aligned}\textbf{upward bias (positive bias)}\text{:}\ & \mathrm{E}(\hat{\theta}_n) > \theta, \\ \textbf{downward bias (negative bias)}\text{:}\ & \mathrm{E}(\hat{\theta}_n) < \theta, \\ \textbf{attenuation bias (bias toward zero)}\text{:}\ & 0 < \frac{\mathrm{E}(\hat{\theta}_n)}{\theta} < 1, \text{so } |\mathrm{E}(\hat{\theta}_n)| < |\theta|, \\ \text{bias away from zero:}\ & \frac{\mathrm{E}(\hat{\theta}_n)}{\theta} > 1, \text{so } |\mathrm{E}(\hat{\theta}_n)| > |\theta|.\end{aligned}$$

An estimator is **unbiased** if its bias is zero. Using (3.12),

$$\mathrm{Bias}(\hat{\theta}) = 0 \iff \mathrm{E}(\hat{\theta}) = \theta, \tag{3.13}$$

where symbol $\iff$ can be read as "is equivalent to" (see Section 6.1).

Example 3.15 (Kaplan video)**.** With iid sampling, the sample mean is an unbiased estimator of the population mean. The estimator is $\hat{\theta}_n = \bar{Y}_n$, and the population parameter is $\theta = \mathrm{E}(Y)$. With $n = 1$, $\bar{Y}_1 = Y_1$, so $\mathrm{E}(\bar{Y}_1) = \mathrm{E}(Y_1) = \mathrm{E}(Y)$. With $n = 2$,

$$\mathrm{E}[\bar{Y}_2] = \mathrm{E}[(1/2)Y_1 + (1/2)Y_2] = \overbrace{(1/2)\,\mathrm{E}(Y_1)}^{\mathrm{E}(Y)/2} + \overbrace{(1/2)\,\mathrm{E}(Y_2)}^{\mathrm{E}(Y)/2} = \mathrm{E}(Y), \tag{3.14}$$

using the linearity property of $\mathrm{E}(\cdot)$ from (2.9). Similar derivations hold for any n, so $\mathrm{E}(\bar{Y}_n) = \mathrm{E}(Y)$, thus the bias is zero given (3.13).

Example 3.16 (Kaplan video)**.** The estimator $\hat{\theta}_n = \bar{Y}_n + 1$ has positive bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(\bar{Y}_n + 1) = \mathrm{E}(\bar{Y}_n) + 1 = \mathrm{E}(Y) + 1 > \mathrm{E}(Y)$.

Example 3.17 (Kaplan video)**.** The estimator $\hat{\theta}_n = \bar{Y}_n - 2$ has negative bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(\bar{Y}_n - 2) = \mathrm{E}(\bar{Y}_n) - 2 = \mathrm{E}(Y) - 2 < \mathrm{E}(Y)$.

Example 3.18 (Kaplan video)**.** The estimator $\hat{\theta}_n = 0.5\bar{Y}_n$ has attenuation bias for the mean $\mathrm{E}(Y)$: $\mathrm{E}(\hat{\theta}_n) = \mathrm{E}(0.5\bar{Y}_n) = 0.5\,\mathrm{E}(\bar{Y}_n) = 0.5\,\mathrm{E}(Y)$, so $0 < [\mathrm{E}(\hat{\theta}_n)/\,\mathrm{E}(Y)] = 0.5 < 1$.

Insufficiency of Bias to Quantify Accuracy

Bias alone does not fully quantify accuracy. That is, if you only consider bias when choosing between two possible estimators, then you may be fooled into choosing the worse estimator.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two different estimators of the same unknown parameter θ. Here, the subscripts 1 and 2 do not indicate n but just that the estimators are different. For simplicity, let $\theta = 0$. The first estimator's distribution is

$$\mathrm{P}(\hat{\theta}_1 = -100) = \mathrm{P}(\hat{\theta}_1 = 100) = 1/2. \tag{3.15}$$

The second estimator's distribution is

$$\mathrm{P}(\hat{\theta}_2 = 1) = 1. \tag{3.16}$$

The first estimator has smaller bias. The estimators' means are

$$\mathrm{E}(\hat{\theta}_1) = (1/2)(-100) + (1/2)(100) = 0, \quad \mathrm{E}(\hat{\theta}_2) = (1)(1) = 1. \tag{3.17}$$

Thus, recalling $\theta = 0$, the bias of each estimator is

$$\mathrm{Bias}(\hat{\theta}_1) = \mathrm{E}(\hat{\theta}_1) - \theta = 0 - 0 = 0, \quad \mathrm{Bias}(\hat{\theta}_2) = \mathrm{E}(\hat{\theta}_2) - \theta = 1 - 0 = 1. \tag{3.18}$$

Estimator $\hat{\theta}_1$ is unbiased, whereas $\hat{\theta}_2$ has upward bias.

But intuitively, $\hat{\theta}_2$ is much better. It always differs from the true θ by only 1, whereas $\hat{\theta}_1$ always differs by 100, which is much worse. That is, regardless of the dataset, $\hat{\theta}_2$ is always 100 times closer than θ_1 to the true $\theta = 0$. This illustrates how bias alone does not properly quantify our preferences: it tells us to prefer $\hat{\theta}_1$ (lower bias) when in fact we strongly prefer $\hat{\theta}_2$ (always much closer to θ).

3.6.2 Mean Squared Error

⟹ Kaplan video: MSE Examples

The **mean squared error** (MSE) is a more complete measure of "how bad" an estimator is. The idea is analogous to using quadratic loss for prediction (e.g., Section 2.5.2). Among other possible loss functions, this is most common and generally reasonable. MSE is mean quadratic loss:

$$\mathrm{MSE}(\hat{\theta}) \equiv \mathrm{E}[L_2(\hat{\theta}, \theta)] = \mathrm{E}[(\hat{\theta} - \theta)^2]. \tag{3.19}$$

Continuing the example, our intuitive preference for $\hat{\theta}_2$ over $\hat{\theta}_1$ is supported by MSE. Because MSE measures "how bad" an estimator is, $\hat{\theta}_2$ being "better" means it has lower MSE. Specifically,

$$\begin{aligned}
\mathrm{MSE}(\hat{\theta}_1) &= \mathrm{E}[(\hat{\theta}_1 - \theta)^2] = (1/2)(-100 - 0)^2 + (1/2)(100 - 0)^2 = 10{,}000, \\
\mathrm{MSE}(\hat{\theta}_2) &= \mathrm{E}[(\hat{\theta}_2 - \theta)^2] = (1)(1 - 0)^2 = 1.
\end{aligned}$$

This matches our intuition: $\hat{\theta}_2$ is much better than $\hat{\theta}_1$ because it has much lower MSE.

MSE can also be decomposed into variance plus squared bias. The variance is

$$\mathrm{Var}(\hat{\theta}) \equiv \mathrm{E}[(\hat{\theta} - \mathrm{E}(\hat{\theta}))^2]. \tag{3.20}$$

(The square root of this is the standard deviation, also called the "standard error" of the estimator $\hat{\theta}$.) Skipping the math, using the bias and variance definitions in (3.12) and (3.20),

$$\mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + [\mathrm{Bias}(\hat{\theta})]^2. \tag{3.21}$$

All else equal, larger bias is bad, but it's also bad to have very high and very low estimates across datasets (large variance and "standard error") even if they happen to average to θ.

Example 3.19 (Kaplan video)**.** Continue the previous example, but instead of assuming $\theta = 0$, let

$$\mathrm{P}(\hat{\theta}_1 = \theta - 100) = \mathrm{P}(\hat{\theta}_1 = \theta + 100) = 1/2, \quad \mathrm{P}(\hat{\theta}_2 = \theta + 1) = 1. \tag{3.22}$$

The MSEs are the same as before because the θ cancels out:

$$\begin{aligned} \mathrm{MSE}(\hat{\theta}_1) &= \mathrm{E}[(\hat{\theta}_1 - \theta)^2] = (1/2)(\theta - 100 - \theta)^2 + (1/2)(\theta + 100 - \theta)^2 = 10{,}000, \\ \mathrm{MSE}(\hat{\theta}_2) &= \mathrm{E}[(\hat{\theta}_2 - \theta)^2] = (1)(\theta + 1 - \theta)^2 = 1. \end{aligned} \tag{3.23}$$

Example 3.20 (Kaplan video)**.** Imagine we know the bias and variance of two estimators, but not the full sampling distributions. This is still sufficient to compute MSE using (3.21). For example, let

$$\mathrm{Bias}(\hat{\beta}_1) = 1, \mathrm{Var}(\hat{\beta}_1) = 16, \quad \mathrm{Bias}(\hat{\beta}_2) = 10, \mathrm{Var}(\hat{\beta}_2) = 9. \tag{3.24}$$

Plugging these into (3.21),

$$\mathrm{MSE}(\hat{\beta}_1) = 1^2 + 16 = 17, \quad \mathrm{MSE}(\hat{\beta}_2) = 10^2 + 9 = 109. \tag{3.25}$$

According to MSE, $\hat{\beta}_1$ is better because it has lower MSE (“less bad”) than $\hat{\beta}_2$. In this case, although $\hat{\beta}_1$ has larger variance, its bias is enough smaller than its overall MSE is also smaller.

Practice 3.3 (estimator MSE)**.** Consider three estimators of the population mean $\mu = \mathrm{E}(Y)$, and their three sampling distributions: $\hat{\mu}_1 \sim \mathrm{N}(\mu, 25)$, $\hat{\mu}_2 \sim \mathrm{N}(\mu + 3, 16)$, and $\hat{\mu}_3 \sim \mathrm{N}(\mu + 2, 9)$, i.e., the sampling distributions of the three estimators are all normal distributions with respective means μ, $\mu + 3$, and $\mu + 2$, and respective variances 25, 16, and 9. (Hint: for MSE, does it matter that the distributions are normal?)

a) Compute the MSE of each estimator.

b) Rank the three estimators from best to worst, in terms of MSE.

3.6.3 Consistency and Asymptotic MSE

Without getting into technical details, an estimator is **consistent** if in “large” samples (large n), there is a “high” probability of the estimator being “close” to the true value. This is similar to the idea of “probably approximately correct” in computer science: estimator $\hat{\theta}_n$ is “consistent” if with large n it is “probably approximately correct.” Unfortunately, there are usually no precise quantitative definitions of “large,” “high,” and “close.”

If $\hat{\theta}_n$ is not consistent, then it has **asymptotic bias**: even with infinite data, the estimator would still be biased. One way to formally define asymptotic bias is

$$\mathrm{AsyBias}(\hat{\theta}_n) \equiv \operatorname*{plim}_{n\to\infty} \hat{\theta}_n - \theta, \tag{3.26}$$

where plim is a probabilistic limit (details omitted), meaning the value that estimates tend to be very close to when n is large. Analogous to “unbiasedness” being “zero bias,” here "consistency" is "zero asymptotic bias": roughly speaking, with a large dataset, there is very little bias. There are the same four types of asymptotic bias as bias (upward/positive, downward/negative, attenuation, and away from zero).

It is also possible to compare approximate (asymptotic) mean squared error; although the details are beyond our scope the intuition is the same (lower is better; depends on both bias and variance components).

As an alternative to (3.26), asymptotic bias can be defined as the approximate bias when the sample size n is large. Formally, $\mathrm{AsyBias}(\hat{\theta}_n) \equiv \lim_{n\to\infty} \mathrm{Bias}(\hat{\theta}_n) - \theta$. This definition is not equivalent to (3.26), nor is it as commonly used, but requires less math to understand, so it is used in the following examples.

Example 3.21 (Kaplan video)**.** Imagine the sampling distribution of $\hat{\theta}_n$ has $\mathrm{E}(\hat{\theta}_n) = \theta + (1/n)$. The estimator has positive bias because $\mathrm{E}(\hat{\theta}_n) - \theta = 1/n > 0$. But with large n, $1/n$ is very close to zero (and would disappear completely with infinite data), so the estimator is asymptotically unbiased (by the alternative definition above).

Example 3.22 (Kaplan video)**.** Imagine the sampling distribution of $\hat{\theta}_n$ has $\mathrm{E}(\hat{\theta}_n) = [0.5 + (1/n)]\theta$. Assuming $n > 2$, the estimator has attentuation bias because $\mathrm{E}(\hat{\theta}_n)/\theta = 0.5 + (1/n)$ is between 0 and 1. Even with very large n, the estimator's value is on average half of the true value. So, the estimator also has asymptotic attenuation bias, which means it's a problem even if the dataset is very large. Alternatively, we can see this by writing $\mathrm{E}(\hat{\theta}_n) = 0.5\theta + (1/n)\theta$, and noting $(1/n)\theta$ is very close to zero for large n, so $\mathrm{E}(\hat{\theta}_n) - \theta$ is approximately -0.5θ. This asymptotic bias is negative when θ is positive, but it is positive when θ is negative, so the asymptotic bias is toward zero (attenuation bias).

3.7 Quantifying Uncertainty

The point estimates in Section 3.4 provide our best guesses about unknown population values, but they offer no sense of our uncertainty. Here, we consider only statistical uncertainty (or sampling uncertainty), meaning the uncertainty due to observing only a random sample of data instead of knowing the true population distribution. (This is only for convenience; other sources of uncertainty may be more important to consider in practice, even if R cannot automatically incorporate them.) Although the term is ambiguous, **inference** often refers to methods that quantify uncertainty.

This section focuses on confidence intervals because econometricians and statisticians generally agree that confidence intervals are more informative and easier to interpret than p-values and hypothesis tests. That is, you should use confidence intervals (not p-values and hypothesis tests) whenever possible.

Complementing this section, Section 3.8 provides warnings about misinterpretation and misuse of conventional frequentist inference methods.

3.7.1 Confidence Intervals

Instead of presenting formulas and critical values for you to memorize (which R computes for you anyway), this section's goal is for you to actually understand the interpretation of a confidence interval.

A **confidence interval** (CI) is computed from data to help quantify statistical uncertainty. The CI $[\hat{L}, \hat{U}]$ ranges from lower endpoint $\hat{L}$ to upper endpoint $\hat{U}$, where both $\hat{L}$ and $\hat{U}$ are computed from data. These endpoints are random variables from the frequentist perspective (before sampling).

A CI should contain the true population value with high probability. If the true value is θ, then the CI "contains" θ when $\hat{L} \leq \theta \leq \hat{U}$, or in other notation $\theta \in [\hat{L}, \hat{U}]$. This happens in some datasets but not others. The probability of randomly sampling a dataset in which the CI contains the true value is

$$\mathrm{P}(\hat{L} \leq \theta \leq \hat{U}). \tag{3.27}$$

Given the same probability, a longer CI indicates more statistical uncertainty. That is, with more uncertainty, the $\hat{L}$ and $\hat{U}$ vary more across random samples, so they need to be farther apart (longer CI) in order to contain θ with the same probability; with less uncertainty, the $\hat{L}$ and $\hat{U}$ are more stable across samples, so they can be closer together (shorter CI) and still contain θ with the same probability.

However, a CI only captures the statistical uncertainty from random sampling, not from any other source. Thus, short intervals can be misleading if there is still uncertainty about certain assumptions or methodological choices.

Example 3.23 (Kaplan video)**.** Recall the last column in Table 3.1. In each of 100 random samples, it showed whether or not the interval $[\bar{Y}_n - 0.4, \bar{Y}_n + 0.4]$ contained the true mean $\mathrm{E}(Y) = 0$, i.e., whether or not $\bar{Y}_n - 0.4 \leq \mathrm{E}(Y) \leq \bar{Y}_n + 0.4$. This CI contained the true population mean in 67 of the 100 datasets. From the "before" view, the probability of randomly sampling a dataset in which the CI contains the true value is around 67%.

A 90% CI does not mean, "I believe there's a 90% chance that the true value is in this range." That is the interpretation of a Bayesian credible interval; see Section 3.1. The difference is subtle and can be confusing; the frequentist "90%" is not about beliefs, but rather that before you sample your dataset, there's a 90% chance of sampling a dataset for which the 90% CI contains the true value. Happily, in many models, with enough data (or weak enough prior belief), frequentist and Bayesian intervals are very similar, though the interpretation still differs.

Example 3.24 (Kaplan video)**.** You have a large dataset from which you want to learn about the population mean age (in years). R tells you that the frequentist 90% confidence interval is $[38.7, 39.4]$, which is also a Bayesian 90% credible interval (rounded to the same precision) given your prior belief. The Bayesian interpretation is: given your prior belief and the data, you have a new belief about the true population mean (which you still don't know exactly), in which you think there's a 90% chance the true mean is between 38.7 and 39.4 years old. The frequentist interpretation is: the CI was computed by a procedure such that before the dataset was sampled ("before sampling"), there was a 90% probability of sampling a dataset whose CI would contain the true mean age; or, if we go randomly sample another 99 datasets, then we'll get a different CI in each dataset, and around 90 of the 100 CIs should contain the true mean age.

Unfortunately, the actual probability that a CI contains the true value often differs from the desired probability. In practice, when you ask R to compute a CI, you specify your desired probability (like 90% or 95%), called the **confidence level** or **nominal coverage probability** (or "nominal level" or other variations). The actual probability is the **coverage probability**, as in (3.27). There are three possibilities.

1. Ideally, a CI's coverage probability is close to the nominal level.
2. Sometimes, a CI is too long and has coverage probability above what you requested. This is bad because it does not help you narrow down the possible values of the population parameter well (because the CI is longer than necessary).
3. Sometimes, a CI is too short and has coverage probability below what you requested, as low as 80%, 50%, or even close to 0%. This is bad because you think the true value is inside the CI, but actually in many datasets (more than you realized) the CI does not contain the true value.

Example 3.25 (Kaplan video)**.** Consider the CI $(-\infty, \infty)$, which ignores the data and is infinitely long. Regardless of the true value of parameter θ, $-\infty < \theta < \infty$, meaning the CI always contains the true value. The true coverage probability is thus 100%. But we do not learn anything from this CI: it does not incoporate any information from the data, and it includes every possible value.

Example 3.26 (Kaplan video)**.** Consider a CI for mean wage (dollars per hour) that is $[10.11, 10.13]$ regardless of the data. If by chance the true mean is between \$10.11/hr and \$10.13/hr, then the CI actually has 100% coverage probability and is very short/precise. Otherwise, however, the CI has 0% coverage probability. Further, clearly this is bad because it does not incorporate any information from the data.

Example 3.27 (Kaplan video)**.** Consider a CI for the employment probability p of a particular subpopulation. Using the estimator $\hat{p}$ (the proportion of employed individuals in the sample) and sample size n, a (usually) reasonable CI is $[\hat{p} - 2\sqrt{\hat{p}(1-\hat{p})/n}, \hat{p} + 2\sqrt{\hat{p}(1-\hat{p})/n}]$. Without worrying about the details, we can at least see that the CI will differ across random samples because it uses $\hat{p}$. We can also see that given the same $\hat{p}$, its length will be smaller with larger n. That is, with more data, we have less uncertainty, which is reflected by a shorter CI. With more data, the true coverage probability will be close to 95%, too. However, we can also see that with a small enough sample, this CI will have very low coverage probability. In the extreme with $n = 1$, then either $\hat{p} = 0$ or $\hat{p} = 1$, so the CI is either $[0, 0]$ or $[1, 1]$, both of which fail to include the true p (assuming

$0 < p < 1$), so the true coverage probability is zero! (And this is still true even if we increase the 2 in the formula to a 4, or even 99, because that term is zero regardless.)

The levels 90% and 95% are most common, but sometimes you may desire 99% or even higher, if it is particularly important that the true value be in the interval (or if you have a very large sample with very short CIs). The convention of 95% is related to the convention of 5% "significance level," which has come under attack for being arbitrary and inappropriate in many situations. The 5% convention seems to have originated from Ronald Fisher, who wrote in 1926(!), "We shall not often be astray if we draw a conventional line at 0.05."

Practice 3.4 (CI interpretation)**.** Imagine you have a CI with 95% nominal coverage probability for the true θ, $[1.4, 2.9]$.

a) Explain why this does *not* mean, "I think there's a 95% chance that $1.4 \leq \theta \leq 2.9$."
b) Explain why it's still possible that the true value is $\theta = 0$.
c) Explain why if the true coverage probability is also 95% and you had 99 other randomly sampled datasets, then around 95 of the 100 total datasets would have a CI containing the true θ.

R Code

The following R example constructs two-sided 95% confidence intervals for the mean, from simulated iid standard normal data (so the true population mean is zero). One CI uses `t.test()`, a standard t-test; the other CIs use nonparametric bootstrap methodology from the `boot` package, though details are beyond our scope.

```
library(boot)
set.seed(112358) #for replicability
Y <- rnorm(n=50, mean=0, sd=1) # iid N(0,1)
CIttest <- t.test(x=Y, conf.level=0.95,
                  alternative='two.sided')$conf.int
ret <- boot(data=Y, statistic=function(x,i) mean(x[i]), R=100)
tmp <- boot.ci(boot.out=ret, conf=0.95, type=c('basic','bca'))
out.table <- rbind(CIttest,tmp$basic[4:5],tmp$bca[4:5])
rownames(out.table) <- c('Normality','Boot.basic','Boot.BCa')
colnames(out.table) <- c('Lower','Upper')
print(round(out.table,digits=3))

##              Lower Upper
## Normality  -0.213 0.370
## Boot.basic -0.248 0.322
## Boot.BCa   -0.168 0.382
```

3.7.2 Statistical Significance

If a CI does not contain zero, then a result is called **statistically significant**, or having **statistical significance**. These terms are usually used when trying to estimate an effect (or difference) that could possibly be zero. A statistically significant result means that if the true effect/value were zero, then there would be a low probability of observing a dataset with such a large estimated effect.

Conceptually, statistical significance is not a yes/no property, but a continuum; i.e., not "if" but "how much?" Results can be somewhat statistically significant, or extremely statistically significant, or lacking statistical significance, etc.

In practice, often people say a result is statistically significant at a particular level. For example, if a 95% CI does not contain zero (i.e., the CI contains only positive values, or only negative values), then the result is "statistically significant at a 95% confidence level," or sometimes people say "5% level" (where $5 = 100 - 95$) because this is equivalent to a p-value being below 5%. Generally, there is statistical significance at a $C\%$ confidence level (or the $(100 - C)\%$ level) if the $C\%$ CI does not contain zero

Why is 95% most common? Indeed, 95% is arbitrary. Its origin seems to be from Ronald Fisher, who wrote in 1926, "We shall not often be astray if we draw a conventional line at 0.05," referring to the p-value threshold that's analogous to 95% confidence level. Recently, 72 prominent researchers from many fields (including statistics, econometrics, and economics) wrote a piece simply titled, "Redefine statistical significance" (Benjamin, Berger, Johannesson, Nosek, Wagenmakers, Berk, Bollen, Brembs, Brown, Camerer, Cesarini, Chambers, Clyde, Cook, De Boeck, Dienes, Dreber, Easwaran, Efferson, Fehr, Fidler, Field, Forster, George, Gonzalez, Goodman, Green, Green, Greenwald, Hadfield, Hedges, Held, Ho, Hoijtink, Hruschka, Imai, Imbens, Ioannidis, Jeon, Jones, Kirchler, Laibson, List, Little, Lupia, Machery, Maxwell, McCarthy, Moore, Morgan, Munafó, Nakagawa, Nyhan, Parker, Pericchi, Perugini, Rouder, Rousseau, Savalei, Schönbrodt, Sellke, Sinclair, Tingley, Van Zandt, Vazire, Watts, Winship, Wolpert, Xie, Young, Zinman, and Johnson, 2018). One high-level message was to not treat statistical significance (at any level) as completely definitive in either direction. They also note that (as in this textbook) it is better to focus on confidence intervals than statistical significance.

3.8 Quantifying Uncertainty: Misinterpretation and Misuse

This section addresses misinterpretations and misuse of frequentist inference. Some of the most common problems are discussed below, as well as on the (pretty good) Wikipedia page devoted to the topic.[1]

3.8.1 Multiple Testing (Multiple Comparisons)

⟹ Kaplan video: Multiple Testing

[1] https://en.wikipedia.org/wiki/Misunderstandings_of_p-values

An insightful comic (xkcd.com/882) illustrates the **multiple testing problem** (or **multiple comparisons problem**). Essentially, the scientists keep testing whether a different color jelly bean (a candy) causes acne (a skin condition), until they finally find $p < 0.05$ and reject the null hypothesis of "no effect" at a 5% level. This is essentially the same as if they keep computing a 95% CI for the effect of each color jelly bean, until they find a CI that does not contain zero. For a 95% CI, this happens roughly 5% of the time, or 1 in 20 datasets; knowing this, the comic shows them testing 20 different colors. The multiple testing problem is essentially that if you keep trying enough times, eventually you'll get a "false positive": the data show a non-zero effect, even though the true effect is zero. As an analogy: even though there's a full moon less than 5% of all nights, as long as you keep looking up at the sky every night, eventually you'll see a full moon.

Practice 3.5 (research assistants)**.** Imagine you're a powerful professor with a cadre of 100 research assistants (post-docs, grad students, undergrads, your neighbor's precocious high-schooler, etc.). You assign each research assistant (RA) one of 100 variables characterizing different counties in the U.S.: number of tennis courts, average temperature, per capita income, etc. Each RA collects a dataset with their particular variable and computes the correlation with county-level May 2020 COVID-19 rates. Each RA then computes a 95% CI for the correlation. Of the 100 RAs, 5 report a CI that does not include zero, including a CI with only positive correlation values for the tennis courts variable, and a CI with only negative values for temperature. In light of the multiple comparisons problem, how do you interpret these results?

Discussion Question 3.2 (jellybean solution?)**.** Consider the jelly bean comic from xkcd.com/882, discussed above.

a) Would it help to use a 99% CI instead of 95%? Explain why, why not, or how much it might help.
b) Would it be even better to use a 100% CI? Explain why or why not.

3.8.2 Publication Bias and Science

The jelly bean comic's final panel illustrates **publication bias**: the newspaper only reports the exciting positive result, omitting the 19 negative results for the other 19 colors. The underlying problem is the same as with multiple testing, but the reader of the publication has no way to know about the other 19 results. Not only popular media but even academic journals are more likely to publish "positive" results (especially if surprising), so reading only published results gives a biased perspective.

The jelly bean experiments also illustrate the importance of remembering what "science" means. The result of a single study (even a good one) by itself is not science. The scientific method is a process of replication and repeated testing of hypotheses. If you ever hear, "There was this one new study that found [crazy result]!" you can ignore it and wait till it gets replicated at least a few times.[2]

[2] This is related to the "replication crisis": https://en.wikipedia.org/wiki/Replication_crisis.

3.8.3 Ignoring Point Estimates (Economic Significance)

Sometimes people focus too much on whether or not the CI contains zero ("statistical significance"), without looking at the magnitude of the point estimate or values in the CI, i.e., the economic significance.

Specifically, **economic significance** assesses if the effect is "economically" distinguishable from zero. "Economically" just means "for real-world purposes," like whether it is important to consider for policy purposes. One way to think about this is: would you personally care about the difference? For example, imagine $\hat{\theta}$ estimates the effect on your final exam score of studying an additional hour per week. Would you care about having a final exam score that's $\hat{\theta}$ percentage points higher? If $\hat{\theta} = 0.01$, then no; if $\hat{\theta} = 50$, then yes. Of course, it's a continuum, so somewhere between "yes" and "no" are varying degrees of "maybe," corresponding to varying degrees of "moderate" economic significance (between "high" and "low").

Example 3.28. Would you care if you had $\hat{\theta} = 2$ additional years of education? This is a lot, like an entire master's degree, so presumably you would indeed care.

Practice 3.6 (salary increase significance)**.** Imagine you compute a 95% CI of $[4.1, 5.9]$ around your estimated annual salary effect of $\hat{\theta} = 5$ dollars per year. Are these results statistically significant (at 95% confidence level)? Are they economically significant? Hint: would you care if your annual salary increased by $\hat{\theta} = 5$ dollars per year?

It is important to consider units of measure. For example, imagine the estimated effect on income is $\hat{\theta} = 10$; is that economically significant? If the units are dollars per hour, then yes; if it's dollars per year, then no; if it's thousands of dollars per month, then yes; etc.

It is also important to consider realistic policy changes. For example, imagine your estimated $\hat{\theta}$ is the effect of a one-unit increase in the proportion of the state budget allocated to higher education. If the current proportion is 0.08 (meaning 8%), then a realistic policy change would be something like 0.02 units. A one-unit increase would mean changing from 0% to 100% of the budget spent on higher education. Even if $\hat{\theta}$ looks economically significant, maybe $0.02\hat{\theta}$ does not.

Practice 3.7 (significance: distance and education)**.** You observe a sample of married couples; for each, you observe the difference in their years of education, divided by the difference in the distance between their childhood homes and the nearest college or university. That is, if E_1 and E_2 are the years of education, and D_1 and D_2 are the distances, you observe $Y = (E_2 - E_1)/(D_2 - D_1)$. Distance is measured in kilometers ($1\,\text{km} = 0.6\,\text{mi}$). You estimate $\bar{Y} = -0.03$. You compute a 95% CI of $[-0.05, -0.01]$.

a) How economically significant is the point estimate of -0.03? Hint: consider the units.
b) Is this statistically significant at a 95% confidence level?

3.8.4 Other Issues

Unfortunately, there are yet more way to misinterpret or misuse methods that quantify uncertainty. Here are a few more examples.

- Non-iid sampling: if you use a method that only works with iid sampling, but your data's sampling was not iid, then you may get misleading results. More generally, methods often require other "assumptions" beyond iid sampling; if any is false, then the results can be misleading.

- Bayesian: frequentist results are often misinterpreted as Bayesian results; for example, p-values are often misinterpreted as the probability of the null hypothesis being true.

- Unlikely events happen: remember that even a 99.9% CI fails to contain the true value in around one of every 1000 datasets; you may be that unlucky one. (Like how winning the lottery is very unlikely, yet somebody somewhere wins the lottery every day; this is the "lottery" of drawing a really unrepresentative dataset.)

Example 3.29 (Kaplan video)**.** Your friend claims to have magical powers. You have a deck of playing cards; you repeatedly draw a card (without showing it) and ask your friend to guess whether the card is black or red. You record the data and compute a 90% CI for your friend's probability p of guessing correctly. Random guessing would yield $p = 0.5$, but your CI is $[0.52, 0.61]$, all values above 0.5. Your friend's interpretation is that statistics have now proved true the claim of magical powers. However, you think it was just luck and ask to gather more data. Indeed, the new dataset's 90% CI is $[0.44, 0.51]$. You try another few datasets, and those CIs also contain 0.5. It seems the first result was simply luck, not magic.

Practice 3.8 (frequentist or Bayesian?)**.** For each of the following, say whether it is a frequentist question, Bayesian question, neither, or both; if both, explain the two possible interpretations. Hint: use Section 3.1 as well as Section 3.7.

a) What's the probability that the current natural unemployment rate in the U.S. is between 4.5% and 7.5%?

b) Can we create a diagnostic tool for our company's daily website traffic data to identify whether it's normal or has been hacked, limiting the rate of falsely reporting "hacked" on normal days to only 1% of normal days?

c) What is the probability that the true unemployment rate is within 1 percentage point of the estimated unemployment rate?

d) Is the positive estimate $\hat{\theta} > 0$ primarily due to the income effect or substitution effect?

Optional Resources

Optional resources for this chapter

- Basic statistics: the Khan Academy AP Statistics unit includes instructional material and practice questions
- Quantifying uncertainty and statistical significance (Masten video)
- Estimator properties (Lambert video)
- Unbiasedness and consistency (Lambert video 1 of 2)
- Unbiasedness and consistency (Lambert video 2 of 2)
- iid sampling (Lambert video)
- Bayesian vs. frequentist cookie inference example (StackExchange)
- Section 2.8 ("Exploratory Data Analysis with R") in Kleiber and Zeileis (2008) [Chapter 2 is available free on their website]
- Section 2.2 ("Random Sampling and the Distribution of Sample Averages") and Chapter 3 ("A Review of Statistics Using R") in Hanck et al. (2018)
- Sections 1.5.4 ("Fundamental Statistics") and 1.9.3 ("Simulation of Confidence Intervals and t Tests") in Heiss (2016)
- R package `boot` (Canty and Ripley, 2019; Davison and Hinkley, 1997)

Empirical Exercises

Empirical Exercise EE3.1. The data are originally from Card (1995), with individual-level observations of wages, years of education, and other variables.

a. R only: run `install.packages(c('wooldridge','survey'))` to download and install those packages (if you have not already)

b. Load the `card` dataset.

   R: load package `wooldridge` with command `library(wooldridge)` and a `data.frame` variable named `card` becomes available; the command `?card` then shows you details about the dataset.

   Stata: run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse card , clear`

c. Compute the sample average of variable `wage`.

   R: `mean(card$wage)`

   Stata: `mean wage` (which also computes a 95% confidence interval)

d. Estimate the population mean accounting for the sampling weights.

   R: `weighted.mean(x=card$wage, w=card$weight)`

   Stata: `mean wage [pweight=weight]` (also computes a 95% CI)

e. R only (because Stata reported this already): compute a two-sided 95% CI for the mean ignoring weights with `t.test(x=card$wage, conf.level=0.95)`

f. R only (because Stata reported this already): compute a two-sided 95% confidence interval for the mean accounting for weights, first loading the `survey` package with `library(survey)` and then with commands

```
carddes <- svydesign(data=card, weights = ~weight, id = ~1)
svyret <- svymean(x = ~wage, design=carddes)
c(w.mean=coef(svyret), SE=SE(svyret),
  CI=confint(svyret, level=0.95))
```

g. Compute a weighted, 90% confidence interval for wage.

   R: replace `level=0.95` with `level=0.90`

   Stata: add “option” `level(90)` to get `mean wage [pweight=weight] , level(90)`

h. Optional: repeat computation of a point estimate and 95% confidence interval (without and with weights) for the mean of a different variable in the dataset.

   R: part (c) computes the unweighted point estimate, part (d) computes the weighted point estimate, part (e) computes the unweighted CI, and part (f) computes the weighted CI.

Stata: part (c) computes both the unweighted point estimate and unweighted CI, and part (d) computes both the weighted point estimate and weighted CI.

Chapter 4

One Variable, Two Populations

⟹ Kaplan video: Chapter Introduction

With two populations, we can discuss not only description and prediction, but also causality. Foundational ideas introduced here are extended to regression in Part II.

Unit learning objectives for this chapter

4.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

4.2. Describe and distinguish among descriptive, predictive, and causal questions, and among different approaches to learning about causality from data in economics [TLOs 3, 5, and 6]

4.3. Describe and interpret the elements of a common statistical framework for understanding causality [TLO 3]

4.4. Assess whether a mean difference can be interpreted with causal meaning in a real-world example [TLO 6]

4.5. In R (or Stata): compute estimates of mean differences, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

4.1 Description

4.1.1 Interpretation of Population Mean Difference

Let Y^A and Y^B be random variables representing Y for two populations (labeled A and B). For example, if Y is income, A is the population of individuals without a high-school degree, and B is the population of individuals with a high-school degree, then Y^A is income for individuals who do not have a high-school degree, and Y^B is income for those who do.

The difference of means is $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$. It describes how much higher (or lower, if negative) is the mean in population B than in population A.

Example 4.1. Let $Y \in \{0, 1, 2\}$ be the number of kids per family. Let the distributions in populations A and B be, respectively,

$$\begin{aligned} &\mathrm{P}(Y^A = 0) = 0.8,\ \mathrm{P}(Y^A = 1) = 0.2,\ \mathrm{P}(Y^A = 2) = 0, \\ &\mathrm{P}(Y^B = 0) = \mathrm{P}(Y^B = 1) = \mathrm{P}(Y^B = 2) = 1/3, \end{aligned} \tag{4.1}$$

where Y^A represents the number of kids per family in population A, and Y^B represents the number of kids per family in population B. Then,

$$\begin{aligned} \mathrm{E}(Y^B) - \mathrm{E}(Y^A) &= \left[\sum_{y=0}^{2} y\,\mathrm{P}(Y^B = y)\right] - \left[\sum_{y=0}^{2} y\,\mathrm{P}(Y^A = y)\right] \\ &= [(0)(1/3) + (1)(1/3) + (2)(1/3)] - [(0)(0.8) + (1)(0.2) + (2)(0)] \\ &= [(1/3) + (2/3)] - 0.2 = 0.8. \end{aligned}$$

That is, the mean kids per family is 0.8 higher in population B than in population A.

Always clarify whether you are subtracting the mean of populaton A from that of B, or B from A. Saying, “The difference in mean number of children between the populations is 0.8,” it is unclear which population’s mean is larger. Instead say, “The mean number of children in population B is 0.8 higher than the mean in A.”

The difference of means is also the mean of the differences. “Mean difference” could mean either; they’re equal anyway. Because of the linearity of the expectation operator as in (2.9),

$$\mathrm{E}(Y^B - Y^A) = \mathrm{E}(Y^B) - \mathrm{E}(Y^A). \tag{4.2}$$

Despite mathematical equality, the interpretation differs. For example, the expression $Y^B - Y^A$ is the number of children difference between a family from population B and a family from population A. Seeing Y^B and Y^A as random variables, the difference $Y^B - Y^A$ is itself a random variable. Thus, $\mathrm{E}(Y^B - Y^A)$ is the population mean of the child number difference $Y^B - Y^A$, whereas $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ is the difference between the mean number of children in B and the mean number of children in A. Generally, due to (4.2), either interpretation of the mean difference is correct; the same population value has two interpretations. It’s like if one person says, “The glass is half full of water,” and a second person says, “The glass is half empty”; both are correct interpretations of the same thing.

Example 4.2. Imagine Y^A is a student’s GPA in fall semester last year, and Y^B is their GPA in spring semester. Then, $\mathrm{E}(Y^A)$ is the mean GPA (over all students in the population) in fall, $\mathrm{E}(Y^B)$ is the mean GPA in spring, and $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ is the change in the mean GPA from fall to spring. Also, $Y^B - Y^A$ is an individual student’s GPA change from fall to spring, so $\mathrm{E}(Y^B - Y^A)$ is the mean fall-to-spring GPA change. From (4.2), the change in mean GPA $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ equals the mean GPA change $\mathrm{E}(Y^B - Y^A)$.

4.1.2 Estimation and Inference

Separately estimate each mean (Section 3.4) and take the difference, like $\bar{Y}^B - \bar{Y}^A$ with iid data. If each individual estimator is consistent, then this is a consistent estimator of $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$, and thus a consistent estimator of $\mathrm{E}(Y^B - Y^A)$ due to (4.2).

The following R code shows an estimate and 95% confidence interval for the mean hourly wage difference between individuals who at age 14 lived with their mom and dad and individuals who did not. It uses an old but well-known dataset. Notes: treating as iid for simplicity, not because it is; dividing by 100 to turn cents into dollars (per hour).

```
library('wooldridge')
YA <- card$wage[card$momdad14==0]/100
YB <- card$wage[card$momdad14==1]/100
# estimate mean wage difference
round(mean(YB) - mean(YA), digits=2)

## [1] 0.84

# 95% CI for mean diff
round(t.test(x=YB, y=YA, alternative='two.sided',
             mu=0, conf.level=0.95)$conf.int[1:2], digits=2)

## [1] 0.63 1.05
```

4.2 Prediction

Prediction is essentially the same as with one population. Given a loss function, an optimal predictor can be defined to minimize mean loss in the population, and this optimal predictor can be estimated from data. For example, mean quadratic loss is minimized by the population mean, and the means $\mathrm{E}(Y^A)$ and $\mathrm{E}(Y^B)$ can be estimated by (weighted) sample means.

Prediction accuracy improves by distinguishing between individuals (or firms, etc.) from population A and those from population B. For example, at your carnival job, imagine you now guess people's height instead of age. In Chapter 2, you make the same guess for everybody. Now, we consider two populations, like child and adult, and we can make a different prediction for each population, like 165 cm for adults and 105 cm for children. Naturally, this performs better than guessing 135 cm for every individual.

Part II extends this idea, exploring how regression models can incorporate additional information to improve prediction accuracy.

Discussion Question 4.1 (DPC with two populations)**.** Let Y denote the hourly wage of an individual in the U.S. Let Y^A be the wage of an individual without a college degree in the U.S., and Y^B the wage of an individual with a college degree.

a) How are means $\mathrm{E}(Y^A)$ and $\mathrm{E}(Y^B)$ more helpful for description than only $\mathrm{E}(Y)$?
b) How could $\mathrm{E}(Y^A)$ and $\mathrm{E}(Y^B)$ be used to make better predictions than only $\mathrm{E}(Y)$?
c) Why can't we interpret $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ as the causal effect of a college degree on wage? Hint: what other factors might make $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ large, even if the effect of a college degree itself is small?

4.3 Causality: Overview

The concepts in the remainder of this chapter appear often in later chapters.

First: in practice, when is causality important, rather than description or prediction? We have an innate sense of cause and effect, although trying to articulate it sometimes creates more confusion than understanding.[1] For example, start reading the Wikipedia page on causality and see how you feel in 10 minutes. Unlike description and prediction, causality is about "why." A "cause" is the "because" of the effect. Description helps us see which variables tend to have high or low values together. Prediction helps us guess one variable's value based on other information. But only causality concerns why. Why do these two variables tend to have similar values? Causality (not description or prediction) helps us evaluate policy decisions: we want to know how a policy change itself influences other variables, causing them to change.

Example 4.3 (Kaplan video)**.** Consider the relationship between an individual's employment status and mental health, specifically anxiety. A descriptive question is: what's the proportion of employed individuals who have generalized anxiety disorder (GAD), and how much higher or lower is that proportion among unemployed individuals? A predictive question is: given somebody's employment status, what's the "best" guess of their score on the GAD-7 anxiety measure? A causal question is: how does being employed (instead of unemployed) affect an individual's level of anxiety as quantified by the GAD-7?

Discussion Question 4.2 (description, prediction, causality)**.** Which type of question (description, prediction, causality) is each of the following? Explain why. Hint: there's one of each.

a) If you only know whether an individual is from Canada or the U.S., what is your best guess of their income?
b) You are currently working in the U.S. but considering moving to Canada. How will your income change if you do?
c) Which country's population has higher income: Canada or the U.S.?

4.3.1 Correlation Does Not Imply Causation

⟹ Kaplan video: Correlation Does Not Imply Causation

[1] Some of my failed attempts include: "causality is about what will happen if a policy changes" (but isn't "what will happen" prediction?) whereas "description is seeing how things are" (but aren't causal relationships also "how things are"?).

Generally, imagine $\mathrm{E}(Y^B) > \mathrm{E}(Y^A)$. This shows a clear descriptive relationship: population B has a higher mean. The implication for prediction is clear: under quadratic loss, the optimal prediction is higher for population B than A. In contrast, the implication for causality is not clear. It's possible that being in population B has a positive causal effect on the outcome variable. But it's also possible that people with large Y choose to join population B. Or maybe there is something else altogether that separately causes people to join population B and have high Y. Or maybe all of these. The causal interpretation of $\mathrm{E}(Y^B) > \mathrm{E}(Y^A)$ is ambiguous.

Example 4.4 (Kaplan video)**.** Consider rainfall and umbrellas. Let Y^A denote rainfall when nobody is carrying an umbrella, and Y^B rainfall when everybody is carrying an umbrella. For description, it rains more on days when everyone carries an umbrella than on days when nobody does; e.g., $\mathrm{E}(Y^B) > \mathrm{E}(Y^A)$. For prediction, it's better to predict a higher rainfall value if you see everyone carrying an umbrella than if you see no umbrellas; e.g., under quadratic loss, the optimal predictions are $\mathrm{E}(Y^B)$ and $\mathrm{E}(Y^A)$. For causality, if there's a drought and we want rain, should we all walk around with umbrellas to cause it to rain? No: rain causes umbrella-carrying, not vice-versa.

Example 4.5 (Kaplan video)**.** Let Y^A be my commute time when nobody is carrying umbrellas, and let Y^B be my commute time when everyone is carrying umbrellas. Descriptively, $\mathrm{E}(Y^B) > \mathrm{E}(Y^A)$, and you should predict a longer commute time if you see everybody has an umbrella. But causally, this doesn't mean that you can make me late for class by opening lots of umbrellas.

In Example 4.5, rain is a **confounder** that has a causal effect on both umbrella-carrying and commute time, as depicted in Figure 4.1.

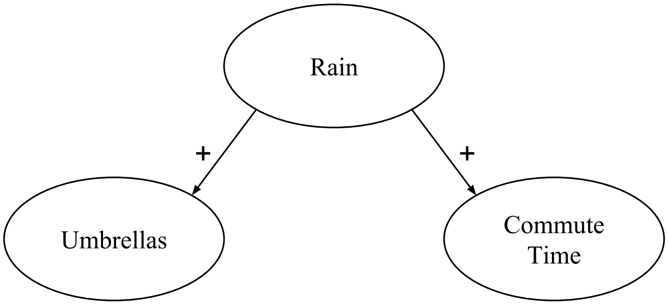


Figure 4.1: Causal relationships among rain, umbrella-carrying, and commute time.

Examples 4.4 and 4.5 illustrate the famous saying, "correlation does not imply causation."[2] The saying is a bit imprecise: correlation does indeed imply some sort of causal relationship, just not any one particular type of causal relationship. In Example 4.4, "correlation does not imply causation" means that "higher rainfall when people carry umbrellas" (rain is correlated with umbrellas) does not imply "carrying umbrellas causes

[2] https://en.wikipedia.org/wiki/Correlation_does_not_imply_causation

rain." But, the correlation is ultimately driven by a causal relationship: rain causes umbrella-carrying. In Example 4.5, "correlation does not imply causation" means that "longer commute when people carry umbrellas" (commute time is correlated with umbrellas) does not imply "carrying umbrellas causes longer commutes." But, the correlation is ultimately driven by causal relationships: rain causes both umbrella-carrying and longer commutes.

Example 4.6. In the August 2018 election in Missouri, a "right-to-work" proposition appeared on the ballot. To clarify upfront: whether such laws are "good" or "bad" is irrelevant here; we are only interested in an econometric question of causality. One ad opposing right-to-work said something like, "Do you want $8000 less in your pocket each year?" The ad's footnote said this $8000/yr was computed as the difference in workers' mean annual income between states that had a right-to-work law and those that did not, like $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$. Recall there was a non-zero mean difference in the example with umbrellas and commute time, too, but we did not conclude that umbrellas have a causal effect on commute time. For example, maybe having lower income causes states to pass such laws, i.e., causality is in the opposite direction (reverse causality). Or maybe there is a third, unobserved characteristic that causes states to pass such laws and causes lower income, i.e., a confounder, like rain in the commute example. Of course, it's also possible that $8000/yr really is the causal effect. The point is not that the number is right or wrong (or that the law is good or bad), but that the econometric argument is incomplete. Additional assumptions are required to interpret a mean difference as a causal effect, as discussed more in Section 4.6.

4.3.2 Structural and Reduced Form Approaches

There are two general econometric approaches to learning about causality: the reduced form approach, and the structural approach. Confusingly, the reduced form approach is sometimes called **causal inference** even though the structural approach also aims to learn about causality.

Both approaches consider **counterfactual** analysis, but in different ways. Broadly, a counterfactual is a universe that's different than our actual universe. Usually, the counterfactual universe is nearly identical to our actual universe except for one particular policy whose effect we want to learn.

The **reduced form** approach tries to isolate causal effects by using comparisons that are either randomized or "as good as randomized." In our current context of populations A and B, **randomized** would mean that units (e.g., individuals, firms, hospitals) are randomly assigned to a population, without regard to the units' characteristics. The "treated" population would receive some special treatment that the "untreated" ("control") population does not. Hopefully, it is then appropriate to interpret the mean difference as the effect of the treatment. "As good as randomized" means that although we did not explicitly randomly assign units to each population, the actual assignment mechanism did not depend on units' characteristics anyway.

In contrast, the **structural approach** tries to explicitly model the inner workings of causal systems. Structural models often come from economic theory, like decision-making or market equilibria models. The goal is to estimate such models' parameters, like elasticities, discount factors, risk aversion, and demand curves.

The structural and reduced form approaches have complementary advantages, and often both are helpful; e.g., see the survey by Lewbel (2019). Structural models often require stronger (less realistic) assumptions, but in return they can analyze a wider variety of possible policies.

Example 4.7 (Kaplan video)**.** Imagine trying to learn how a retirement pension formula (i.e., how much money somebody gets paid after retiring, based on their years of experience, age, and salary history) affects the age at which a teacher decides to retire. A reduced-form analysis might compare the mean retirement age of teachers who joined a school in the year 1998 with the mean retirement age of teachers who joined in 1999, just after the formula was changed, hoping that the two groups of teachers are otherwise "as good as randomized." A structural analysis might explicitly model a teacher's retirement decision within an expected utility framework that "discounts" the value of future periods (like net present value). The structural analysis requires strong (maybe unrealistic) assumptions about things like the utility function and the distribution of unobserved variables. However, it can then evaluate the effect of hypothetical pension changes that may have never been implemented before, rather than only estimating the effect of the historical 1999 pension change.

Example 4.8. Imagine trying to learn about the effect of free public childcare on how much mothers work in the formal sector. A reduced-form analysis might estimate how much mothers work in cities that just opened such childcare centers last year compared to mothers in cities that plan to open them next year. The hope is that whether a city opens the childcare centers last year or next year is "as good as randomized," so that the mean difference in hours worked can be interpreted as the effect of the childcare (rather than the effect of something else that's different). A structural analysis might try to estimate an economic model of a mother's decision to work in the formal sector, including variables like the price of childcare, wages, and utility from different activities. Such a model requires strong assumptions (although "as good as randomized" may also be unrealistic!), but can then be used to evaluate the effects of a wide variety of hypothetical policies, not only the effect of the childcare centers that opened last year.

In Sum: Structural & Reduced Form Approaches

Reduced form: randomized or "as good as randomized" comparisons to isolate causality
Structural: more explicit economic models of causal relationships

4.3.3 General Equilibrium and Partial Equilibrium

Besides structural vs. reduced form, another dichotomy is between **general equilibrium** (GE) and **partial equilibrium** (PE) analysis. GE more ambitiously tries to model entire markets, sometimes multiple markets, whereas PE takes current market equilibria as given. Similar to the tradeoff between the structural and reduced form approaches, the tradeoff is that the GE framework can analyze policies that change equilibria (i.e., that have **general equilibrium effects**), but it requires stronger assumptions to do so.

Example 4.9 (Kaplan video)**.** Imagine you were analyzing the impact of free public childcare on mothers' employment. A PE analysis would consider how mothers might respond to different childcare policies given the current prices of private childcare, current wages, etc. A GE analysis might further model the childcare and labor markets, to allow for the possible general equilibrium effects of public childcare policy on the prices in those markets. If there is a big expansion of free public childcare, then private childcares may indeed change their prices. If the expansion allows many mothers to enter the workforce, then the labor supply curve shifts out, which could lower wages. However, if the proposed changes to childcare policy are relatively small, then such GE effects may be negligible, and PE analysis may suffice.

In Sum: General & Partial Equilibrium Models

Partial equilibrium models treat prices and other market equilibria as fixed, whereas **general equilibrium models** allow markets to change.

4.4 Causality: Potential Outcomes Framework

⟹ Kaplan video: Potential Outcomes and the ATE

The reduced form approach uses the **potential outcomes framework**, also called the **Neyman–Rubin causal model** after its two earliest contributors (although sometimes Neyman's name is dropped). It is popular not only in economics, but statistics, medicine, political science, and other fields.

The terms **treatment** and **treatment effect** just refer to any variable and its causal effect on another variable. In English, usually "treatment" makes us think narrowly about medicine (or lumber. . . and facials?), but it can be anything. For example, the "treatment" could be a job training program, and the "treatment effect" is the causal effect of the program on a person's wage. Or, a treatment could be going to a charter school (instead of public school). Another treatment could be a policy or law, like a higher sales tax, or a certain labor law.

This section says "individual" to be concrete, but you can also imagine a firm, county, school, etc.

4.4.1 Potential Outcomes

Imagine two parallel universes. The universes are identical except for one difference: whether or not an individual is treated. The individual's outcome in the universe without treatment is their **untreated potential outcome**, and the individual's outcome in the universe with treatment is their **treated potential outcome**.

Notationally, in this chapter, Y^T represents the treated potential outcome and Y^U the untreated potential outcome. Elsewhere, often Y_1 and Y_0 represent the treated and untreated potential outcomes, or $Y(1)$ and $Y(0)$.

Unlike in Section 4.1, potential outcomes Y^U and Y^T are not always observable. Often, if an individual is untreated in our universe, then we can observe her untreated potential outcome Y^U, but not her Y^T; conversely, if she is treated, then we observe Y^T but not Y^U. This partial observability makes causal inference more difficult than description or prediction.

Example 4.10. Consider parallel universes identical except for whether a particular student takes Introductory Econometrics or Applied Statistical Models I (STAT 4510/7510). Literally everything else in each universe is identical: the student's parents, her other classes, her height, her DNA, the weather on October 14, etc. (For now, some difficulties with "everything" are glossed over; e.g., what if econometrics is required for her degree?) The "treatment" is taking econometrics (instead of statistics). The outcome variable is the student's annual income five years after graduation, in thousands of U.S. dollars per year (e.g., $Y = 70$ is \$70,000/yr). Let Y^U denote her outcome in the universe without treatment (statistics class), and Y^T her outcome in the universe with treatment (econometrics class). That is, Y^T is her treated potential outcome, and Y^U is her untreated potential outcome.

Example 4.11. In the right-to-work example (Example 4.6), Y^T is an individual's income in the universe where the individual's state has a right-to-work law, and Y^U is their income in the universe that's identical except there is no such law. In our universe, either the individual's state does or does not currently have such a law; it cannot be both, so we cannot observe both potential outcomes. (Perhaps the state did not have the law last year and does this year, but the universe "last year" was different in many ways than the universe "this year"; much more than one single law has changed.)

Example 4.12 (Kaplan video)**.** Imagine one universe where a student wins the lottery to enter a popular charter school, and another universe where the student remains in the conventional public school. Potential outcomes Y^T and Y^U are dummy (binary) variables for whether or not the student eventually graduated from college in each respective universe. Again, in our universe, we can observe Y^T if the student wins the lottery and Y^U if not, but we cannot observe both.

4.4.2 Treatment Effects

The difference $Y^T - Y^U$ between an individual's two potential outcomes is that individual's **treatment effect**. Just as different individuals can have different (Y^U, Y^T), individuals can have different treatment effects $Y^T - Y^U$; i.e., individuals can be affected differently by the same treatment. The fancy term for people being different is **heterogeneity**, more specifically here "treatment effect heterogeneity."

Example 4.13. In the intro econometrics example (Example 4.10), the student's treatment effect $Y^T - Y^U$ has the following interpretation. Recall Y^T is their income after taking econometrics and Y^U after instead taking STAT 3500. Thus, that particular student's treatment effect is how much higher (or lower, if negative) their income is in the parallel universe that is identical other than taking econometrics instead of STAT 3500.

Example 4.14. In the right-to-work example (Examples 4.6 and 4.11), $Y^T - Y^U$ is the treatment effect of the law on an individual's income. The interpretation now is the difference between their income in the universe with the law and the universe without the law, with everything else held constant. The treatment effect can be big or small, positive or negative (or zero). A numerical example is shown later in Table 4.2.

Example 4.15 (Kaplan video)**.** In the charter school example (Example 4.12), $Y^T - Y^U$ is the treatment effect of the charter school on college graduation. That is, it is the difference between the college graduation outcomes in the charter school universe and the public school universe. Because the outcome is binary (1 if graduate college, 0 if don't), there are only four possible values of (Y^U, Y^T) (student types) and only three possible treatment effect values: $Y^T - Y^U = 1$ if the student graduates in the charter school universe ($Y^T = 1$) but not the public school universe ($Y^U = 0$); $Y^T - Y^U = -1$ if they only graduate in the public school universe ($Y^U = 1$) but not the charter school universe ($Y^T = 0$); and $Y^T - Y^U = 0$ if they graduate either in both universes ($Y^T = Y^U = 1$) or neither ($Y^T = Y^U = 0$). This is seen in the later example of Table 4.1.

In economics, where many systems are interrelated, sometimes it's difficult just to specify which "effect" we care about. For example, consider racial differences in salary. In the parallel universe that's "identical" except for the individual's race, does "identical" include having the same job at the same firm? Or does it allow for an effect of race on hiring? Does it allow for an effect on educational opportunities, or an effect on family background (parents' education, wealth, etc.)? There is no "right" or "wrong" specification, but each answers a different question.

In Sum: Causality in Potential Outcomes Framework

Treatment effect: the difference in outcomes between parallel universes identical except for treatment

4.4.3 SUTVA

SUTVA Definition

The potential outcomes definition of causality relies critically on the **stable unit treatment value assumption** (SUTVA), which has two parts.

The first part of SUTVA is that every treated individual receives the same treatment. This seems obvious, and is easily satisfied in other fields like medicine, but it often requires thought in economics.

Example 4.16. In the right-to-work example, the same law applies (or doesn't) to everybody equally. However, different states may have different implementations of such a law.

Example 4.17. In the charter school example, it seems every student has the same treatment: going to the same school. Still, it's worth remembering that this "same treatment" may actually consist of different teachers, different classmates, and different extra-curricular activities. We would expect a lot of treatment effect heterogeneity, and expect that the effect may change over time as the school gets new teachers, students, and activities.

Example 4.18. As another ambiguous example, imagine a one-on-one mentoring program to help teen parents. Of course, there are many different mentors. Is every teen parent receiving the "same treatment"?

The second part of SUTVA is the **no interference** assumption. This assumes that one person's treatment (or non-treatment) does not affect the potential outcomes of any other person. This often makes sense for medical treatments (e.g., my knee surgery doesn't affect your health), but it requires careful thought in economics, where often individuals interact personally or through markets.

Example 4.19. In the charter school example, if a student's success depends on being surrounded by other highly motivated students, then SUTVA (specifically no interference) is violated. That is, one student's outcome depends on whether the other motivated students are in the same school (whether charter or not), i.e., depends on the other students' "treatment."

SUTVA Violations

As alluded to above, SUTVA can be violated in many ways, especially in economics. This is not about sampling, or randomization, or data; it is about the potential outcomes framework itself. Without SUTVA, it's unclear what "treatment effect" even means.

One common violation of SUTVA is from **spillover effects** that benefit even untreated individuals. That is, the treatment's benefit "spills over" into untreated individuals. Perhaps the treated individuals can share the treatment itself with others, or perhaps others benefit from the improved outcomes of treated individuals.

Example 4.20. Consider a treatment that provides treated individuals with helpful information about financial planning. Treated individuals might share such information with their untreated friends and family. Thus, an untreated individual's outcome may depend on whether or not their friend is treated. This spillover effect violates the "no interference" part of SUTVA.

Example 4.21. Consider a "treatment" that leads to less binge drinking among treated individuals. Even if the treatment itself is not shared, the reduction in binge drinking may reduce social pressure and result in less binge drinking among untreated individuals. Here, untreated individuals are affected by the treatment through the changed behavior of treated individuals. This spillover effect violates SUTVA.

Another common violation of SUTVA is from **general equilibrium effects** (Section 4.3.3), such as changing market prices.

Example 4.22 (Kaplan video)**.** Consider a new agricultural technology hoping to increase cacao farmers' earnings (through increased productivity). If only one farmer gets this treatment (technology), then she benefits from increased production, selling more cacao at the current global price. But if all farmers in the world get the technology, then the global cacao supply curve shifts and the price drops. Thus, each farmer's untreated and treated potential outcomes (earnings) are affected by all other farmers' treatment status, which affects the market equilibrium price. This violates SUTVA.

Example 4.23. Consider the "treatment" that provides a subsidy for buying a house. This increases demand, which increases prices. This general equilibrium effect violates SUTVA.

Discussion Question 4.3 (cash transfer spillovers)**.** Consider the effect of income on food consumption (Y) in a rural village. Consider an "unconditional cash transfer" program (like GiveDirectly) that (potentially) gives the equivalent of \$1000 to a treated individual. Describe different possible spillover effects that would violate SUTVA.

4.5 Average Treatment Effect

⟹ Kaplan video: Potential Outcomes and the ATE (again)

Although the full distribution of potential outcomes (Y^U, Y^T) contains the most information, usually only certain summary features are studied. Although summary features like standard deviations and percentiles are interesting, we'll focus on means.

4.5.1 Definition and Interpretation

The **average treatment effect** (ATE) is $\mathrm{E}(Y^T - Y^U)$. "Average" refers to the population mean, while "treatment effect" refers to $Y^T - Y^U$. Thus, the ATE may be interpreted as the probability-weighted average (mean) of all possible individual treatment effects in the population. Another name for the ATE is the **average causal effect** (ACE), but I use ATE to emphasize that this concept is from the potential outcomes framework.

The ATE has another interpretation, analogous to the two "mean difference" interpretations in (4.2). Using linearity as in (2.9),

$$\text{ATE} \equiv \text{E}(Y^T - Y^U) = \text{E}(Y^T) - \text{E}(Y^U). \tag{4.3}$$

Here, $\text{E}(Y^T)$ is the mean treated potential outcome, and $\text{E}(Y^U)$ is the mean untreated potential outcome. This could be interpreted as "the treatment effect on the mean outcome": treatment causes the mean outcome to change from $\text{E}(Y^U)$ to $\text{E}(Y^T)$.

Example 4.24 (Kaplan video)**.** Table 4.1 shows a numerical version of the charter school example. The four student "types" refer to the four possible values of (Y^U, Y^T), and each type has its own probability. Given the probabilities, the mean untreated outcome $\text{E}(Y^U)$, mean treated outcome $\text{E}(Y^T)$, and ATE $\text{E}(Y^T - Y^U)$ are computed using (2.4):

$$\text{E}(Y^U) = (0.3)(0) + (0.3)(0) + (0.1)(1) + (0.3)(1) = 0.4, \tag{4.4}$$
$$\text{E}(Y^T) = (0.3)(0) + (0.3)(1) + (0.1)(0) + (0.3)(1) = 0.6, \tag{4.5}$$
$$\text{E}(Y^T - Y^U) = (0.3)(0) + (0.3)(1) + (0.1)(-1) + (0.3)(0) = 0.2. \tag{4.6}$$

To verify (4.3),

$$\text{E}(Y^T - Y^U) = 0.2 = 0.6 - 0.4 = \text{E}(Y^T) - \text{E}(Y^U). \tag{4.7}$$

Table 4.1: Charter school example population of potential outcomes and ATE.

Student type	Probability	Y^U	Y^T	$Y^T - Y^U$
1	0.3	0	0	0
2	0.3	0	1	1
3	0.1	1	0	−1
4	0.3	1	1	0
Mean		0.4	0.6	0.2

Example 4.25. Table 4.2 shows a numerical version of the right-to-work example. Each worker "type" corresponds to a different value of (Y^U, Y^T), each type with its own probability. Given the probabilities, the mean untreated outcome $\text{E}(Y^U)$, mean treated outcome $\text{E}(Y^T)$, and ATE $\text{E}(Y^T - Y^U)$ are, in dollars per year,

$$\text{E}(Y^U) = (0.5)(40{,}000) + (0.2)(40{,}000) + (0.2)(50{,}000) + (0.1)(50{,}000) = 43{,}000, \tag{4.8}$$
$$\text{E}(Y^T) = (0.5)(41{,}000) + (0.2)(38{,}000) + (0.2)(51{,}000) + (0.1)(47{,}000) = 43{,}000, \tag{4.9}$$
$$\text{E}(Y^T - Y^U) = (0.5)(1000) + (0.2)(-2000) + (0.2)(1000) + (0.1)(-3000) = 0. \tag{4.10}$$

Table 4.2: Right-to-work example population of potential outcomes and ATE.

Worker type	Probability	Y^U ($/yr)	Y^T ($/yr)	$Y^T - Y^U$ ($/yr)
1	0.5	40,000	41,000	1000
2	0.2	40,000	38,000	−2000
3	0.2	50,000	51,000	1000
4	0.1	50,000	47,000	−3000
Mean		43,000	43,000	0

Again, to verify (4.3),

$$\mathrm{E}(Y^T - Y^U) = \$0/\text{yr} = \$43{,}000/\text{yr} - \$43{,}000/\text{yr} = \mathrm{E}(Y^T) - \mathrm{E}(Y^U). \tag{4.11}$$

4.5.2 Limitations of ATE

Zero ATE does not mean zero effect. The logical implication is one way: $Y_T - Y_U = 0 \implies \mathrm{E}(Y_T - Y_U) = 0$, but not $\impliedby$. Thus, if we focus only on the mean, then we may miss other important effects of the treatment, like increased standard deviation or skewness. This idea is retold in joke form by Hansen (2020, p. 29):

> An economist was standing with one foot in a bucket of boiling water and the other foot in a bucket of ice. When asked how he felt, he replied, "On average I feel just fine."

Example 4.26 (Kaplan video)**.** Imagine a professional skills workshop that increases wages by $3/hr for half the population but actually decreases wages by $3/hr for the other half of the population. The workshop "treatment" clearly has an effect, but it averages to zero ATE. Mathematically, let $\mathrm{P}(Y_T - Y_U = 3) = \mathrm{P}(Y_T - Y_U = -3) = 1/2$. The ATE is $\mathrm{E}(Y_T - Y_U) = (1/2)(3) + (1/2)(-3) = 0$, even though everybody is affected by the treatment.

Example 4.27 (Kaplan video)**.** Let $Y_U = 3$ for everybody, so $\mathrm{E}(Y_U) = \mathrm{E}(3) = 3$. Let $\mathrm{P}(Y_T = 1) = 0.9$ and $\mathrm{P}(Y_T = 21) = 0.1$, so $\mathrm{E}(Y_T) = (0.9)(1) + (0.1)(21) = 3$. The ATE is zero: $\mathrm{E}(Y_T) - \mathrm{E}(Y_U) = 3 - 3 = 0$. However, the treatment hurts 90% of individuals.

Example 4.28 (Kaplan video)**.** Let $\mathrm{P}(Y_U = -9) = \mathrm{P}(Y_U = 9) = 1/2$, so $\mathrm{E}(Y_U) = (1/2)(-9) + (1/2)(9) = 0$, and $\mathrm{P}(Y_T = -2) = \mathrm{P}(Y_T = 2) = 1/2$, so $\mathrm{E}(Y_T) = (1/2)(-2) + (1/2)(2) = 0$, too. The ATE is zero: $\mathrm{E}(Y_T) - \mathrm{E}(Y_U) = 0 - 0 = 0$. However, the treatment greatly reduces dispersion/spread.

Complementing ATE, another approach examines effects on percentiles ("quantile treatment effects"), but these are beyond our scope.

Practice 4.1 (unrepresentative ATE)**.** Describe a population in which the ATE is zero but every individual is affected by the treatment (i.e., all treatment effects are non-zero). For simplicity, assume there are only two types of individual. For each type, state the probability, potential outcomes Y^U and Y^T, and causal effect $Y^T - Y^U$, which must be non-zero. Then compute the ATE to verify it's zero.

Another limitation is that ATE compares a universe where everybody is treated to a universe where nobody is treated, which may be unrealistic. We may instead be interested in a smaller policy change that encourages treatment of some additional individuals "on the margin." ATE does not help us learn about the effect on those marginal individuals, nor does it help us learn about how individuals decide to get treated (if it is a choice).

Example 4.29 (Kaplan video)**.** Imagine Missouri is considering increasing the income threshold to qualify for Medicaid (health insurance for "low-income" individuals), so more people would be eligible. The marginal individuals are those who are currently not eligible (because income is too high) but would become eligible if the threshold decreases. The effect of the policy change is only the effect on these individuals on the margin. In contrast, the ATE would be the average effect of changing from nobody having Medicaid to everyone being eligible for Medicaid, which is not relevant here. Further, not everyone eligible actually gets Medicaid. The ATE does not help us learn about an eligible individual's decision to use Medicaid or not, which is also relevant to the effect of a policy that changes eligibility.

4.6 ATE: Identification

⟹ Kaplan video: ATE Identification

Generally, **identification** is a concept central to econometrics that appears throughout this textbook. A parameter is **identified** if it equals a summary feature of the population distribution of observable variables.

Here, the ATE is "identified" when it equals the mean difference. That is, we can interpret the mean difference as the ATE. This is helpful because we already know how to learn about the mean difference (Section 4.1.2).

Generally, identification results have the logical form of: if certain **identifying assumptions** are true, then some parameter is identified. Both logic and identifying assumptions for the ATE are covered formally in Section 6.1 and Section 6.4, respectively.

4.6.1 Setup and Identification Question

For each individual, a single value is observed. If the individual was actually treated (in our universe), then treated potential outcome Y^T is observed; otherwise, Y^U is observed.

Consider actually treated individuals to be population B, and consider actually untreated individuals to be population A. The two populations are represented by random variables Y^B and Y^A, respectively. For a population B individual, Y^B is always observable, with $Y^B = Y^T$. Similarly, for a population A individual, Y^A is always observable, with $Y^A = Y^U$. A random sample of Y^B can be taken from actually treated individuals, and a random sample of Y^A can be taken from actually untreated individuals.

Example 4.30 (Kaplan video)**.** Consider the outcome of college graduation in the charter school example. Then, Y^B represents the graduation outcome for a student who actually attended the charter school (in our universe), while Y^A is the outcome for a student who did not. We observe Y^T (but not Y^U) for all students who attended the charter school, whereas we observe Y^U (but not Y^T) for all students who did not. The mean difference $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ is the graduation rate of charter school attendees minus the graduation rate of non-attendees.

The **identification** question is whether the ATE equals the mean difference. Mathematically, using the $\mathrm{E}(Y^T) - \mathrm{E}(Y^U)$ form of the ATE from (4.3), the identification question is whether or not

$$\mathrm{E}(Y^T) - \mathrm{E}(Y^U) = \mathrm{E}(Y^B) - \mathrm{E}(Y^A). \tag{4.12}$$

Example 4.31 (Kaplan video)**.** Continuing the charter school example (Example 4.30), recall the mean difference $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$ is the graduation rate of charter school attendees minus the graduation rate of non-attendees. The ATE is identified when this mean difference equals the ATE $\mathrm{E}(Y^T) - \mathrm{E}(Y^U)$, which is the graduation rate in the universe where everyone is treated (everyone in charter school) minus the graduation rate in the universe where nobody is treated (no charter school). That is, if the ATE is identified, then we can interpret the difference in graduation rate between charter and non-charter schools as the causal effect of the charter school itself.

Example 4.32. For the right-to-work example, if the ATE is identified, then it equals mean income in right-to-work states minus mean income in other states, $\mathrm{E}(Y^B) - \mathrm{E}(Y^A)$. That is, we can interpret the mean income difference between treated and untreated states as the causal effect of the treatment (right-to-work law).

4.6.2 Randomization

If an individual's treatment status is independent of that individual's potential outcomes (Y^U, Y^T), then the ATE is identified and equal to the mean difference (between actually treated individuals' outcomes and actually untreated individuals' outcomes). One way to make treatment independent is to randomize treatment, assuming we have full control of every individual's treatment status. Section 6.4 contains more formal arguments for why randomization can help identify the ATE.

For this reason, randomized experiments are often used to estimate the ATE. In a **randomized experiment**, also called a **randomized controlled trial** (RCT), ideally the experimenter can fully control who is treated and who is not (but see Section 4.6.3 for examples of how this can fail). Mathematically, the experimenter gets to decide whether to observe Y^U or Y^T for each individual. "Randomized" means this decision is made without regard to the individual's characteristics (like by flipping a coin).

For intuition, consider the following experimental strategy. First, imagine we only want to estimate $\mathrm{E}(Y^T)$. We could take a random sample of individuals from the population and treat each one, allowing us to observe their Y^T. That is, we have a random sample from the population distribution of Y^T. As in Chapter 3, we can estimate $\mathrm{E}(Y^T)$ by the sample mean. Second, we can repeat the process for a second random sample but force everyone to be untreated. The key is the ability to force anyone to be either treated or untreated, and assigning treatment independently of the individual's characteristics; this allows us to take random samples of Y^T and Y^U. (Mathematically, it's equivalent to randomly order individuals and treat the first half or to randomly assign half to treatment.)

Example 4.33. You have a random sample of people newly diagnosed with COVID (in 2021), and randomly give molnupiravir to half of them and a placebo to the other half. For the treated half, you observe potential treated outcome $Y^T = 1$ if they're hospitalized (and $Y^T = 0$ if not). Similarly, for the untreated half, you observe untreated potential outcome Y^U. Because of randomization (and SUTVA seems satisfied), the ATE is identified and equal to the mean difference: the mean difference between treatment and control group hospitalization rates is equal to the ATE $\mathrm{E}(Y^T) - \mathrm{E}(Y^U)$, recalling $\mathrm{E}(Y^T) = \mathrm{P}(Y^T = 1)$ and $\mathrm{E}(Y^U) = \mathrm{P}(Y^U = 1)$ are the hospitalization rates in the treated universe and untreated universe, respectively. Thus, we can interpret the lower hospitalization rate of the treatment group as a causal effect (ATE) of the drug molnupiravir, and interpret a 95% CI for the mean difference as a 95% CI for the ATE.

Example 4.34 (Kaplan video)**.** Imagine you have a random sample of students and that you can force half of them (chosen at random) to attend the charter school, and force the other half not to attend the charter school. Then (ignoring possible SUTVA violations), the ATE is identified. However, clearly this is not ethical/legal (a common difficulty). (To address the ethical challenge: if we have 200 students who want to join the charter school, but there are only 100 openings, then we could randomize which 100 are admitted, with the caveat that these 200 students are not representative of the full student population: they're the ones whose families wanted to apply to the charter school.)

Hypothetically, there could be a treatment that happens to be "as good as randomized," in which case the ATE would be identified, but this is rare; see Section 4.6.3.

4.6.3 Reasons for Identification Failure

ATE identification can fail due to violations of SUTVA (as discussed in Section 4.4.3) or if the treatment assignment is related to the potential outcomes, which is the focus here.

Outside of experiments, random or "as good as random" treatment is rare. This is often because the treatment is chosen based on its costs and benefits, as economic theory suggets. That is, there is **self-selection** into treatment. For example, if the treatment has a net benefit for some people but not others, and people are free to choose treatment, then we observe the treated potential outcomes of individuals who (on net) benefit from the treatment, and we observe untreated potential outcomes of individuals who don't. Similarly, although government decisions are not always based on good reasons, they are not completely randomized; they usually take into account what's going on economically or otherwise.

Example 4.35. In the right-to-work example, treatment is almost certainly not random. For example, optimistically, legislatures may consider the distribution of Y^U when deciding whether or not to pass the law (which would switch everyone's annual earnings from their Y^U to their Y^T), to see if they think the law would be helpful for their particular state (with their particular mix of industries, skills, etc.). More realistically, just looking at a map, it is notable that (as of 2019) zero U.S. states in the Northeast census region have right-to-work laws, whereas almost all states in the South census region have right-to-work laws (the exceptions being Delaware and Maryland, which are not particularly "Southern" and indeed border the Northeast region). If the potential outcome annual earnings distributions already differ between South and Northeast for other reasons (besides right-to-work laws), then interpreting the entire mean difference as due solely to the right-to-work effect is not correct. Whatever effect the law has, we cannot isolate it from the mix of other effects on earnings due to other Notheast/South differences.

Example 4.36 (Kaplan video)**.** If charter school attendance is not randomized, then maybe families who apply tend to be those who value education most highly (and are organized enough to submit an application on time). If so, they may help their children in other ways. Even if the charter school graduation rate is higher, we can't isolate the effect of the school itself from the effect of everything else these families do to help their children. In sum: charter school attendance is not "as good as random" but related to family characteristics, so the ATE is not identified, meaning we can't interpret the mean difference as only the effect of the charter school.

Example 4.37 (Kaplan video)**.** Consider some other brief examples of treatment not being random but rather based on factors that are in turn related to outcomes. Whether or not a road is widened (treatment) is related to the existing traffic level (outcome). Whether or not somebody gets a college degree (treatment) is related to factors like family wealth that themselves affect wages (outcome). Whether or not Walmart builds a store in a particular town (treatment) is related to existing economic conditions (outcome). Somebody's choice of health insurance plan (treatment) depends on their anticipated utilization based on their anticipated health level (outcome).

Even with randomized treatment *assignment*, treatment itself may not be random if it's not ethical or legal to force somebody to be (un)treated. That is, if individuals can still ultimately choose their treatment status, then there is still a self-selection problem. Included in this is if individuals can choose to not participate (**attrition**) after learning their treatment assignment.

Example 4.38 (Kaplan video)**.** Imagine you randomly assign people to attend a job training program, but some do not. People who skip the program may also skip work regularly, which results in lower income. Thus, many low-income individuals who should have been in the treatment group (if we could force them) are now in the control group. This decreases the control group's average income and raises the treatment group's average income, which falsely makes the treatment seem more effective than it is.

One way to avoid this incorrect conclusion is to change perspective: compare groups based on treatment assignment rather than actual treatment. The resulting ATE is called the **intention-to-treat** effect because it measures the mean change in Y corresponding to the intention to treat (i.e., assignment to treatment, or offer of treatment). Sometimes this is more directly relevant for policy anyway, if the actual policy would not force people to be treated.

Example 4.39. In the job training example, we can compare mean earnings for the treatment group that was invited to the training with the control group that was not invited. Even if some invited individuals do not attend, this lets us estimate the effect of being invited, if not the effect of attending.

Other concerns are introduced later, especially in Section 12.3.

Discussion Question 4.4 (breakfast effect?)**.** Schools with a high enough percentage of low-income students are eligible for a federally-funded free breakfast program for all students. Although the program is not mandatory, all eligible schools choose to have it. You compute a 95% CI for the mean math test score of the "breakfast" schools minus the mean of the other schools, and it is $[-32, -17]$ points. (The test is out of 100 points; most scores are in the 60 to 100 range.) How do you interpret this result? Think about ATE identification, statistical uncertainty, and frequentist vs. Bayesian perspectives.

4.7 ATE: Estimation and Inference

If the ATE is identified, then estimates and CIs for the ATE are identical to those for the mean difference (Section 4.1.2) because the ATE equals the mean difference. However, a CI does not incorporate any uncertainty about identification. For example, if the ATE is actually not identified, then a 95% CI for the mean difference may only contain the ATE with 80% probability, or even 50% or near 0%.

Optional Resources

Optional resources for this chapter

- Structural and reduced form approaches: Lewbel (2019)
- Potential outcomes and SUTVA (Wikipedia)

- Causal inference intro (Masten video)
- Correlation vs. causation (Masten video)
- ATE (Masten video)
- Individual causal effects (Masten video)
- Potential outcomes example (Masten video)
- Counterfactuals (Masten video)
- Randomized experiments (Masten video)
- SUTVA and spillovers (Masten video)
- Empirical example: property rights effect (Masten video)
- Structural modeling advantages (Masten video)
- Potential outcomes and confounding (Lambert video)

Empirical Exercises

Empirical Exercise EE4.1. You will analyze the effects of being assigned to a job training program, where assignment was randomized. The specific program was the National Supported Work Demonstration in the 1970s in the U.S. Data are originally from LaLonde (1986), via Wooldridge (2020). You will look at effects on earnings (`re78`) and unemployment (`unem78`), both overall and for different subgroups (e.g., married or not). The `train` variable indicates (randomized) assignment to job training if it equals 1, and 0 otherwise. For now, we focus on computing various estimates; in later chapters we'll think more critically about what could go wrong even with randomized assignment.

a. R only: run `install.packages('wooldridge')` to download and install that package (if you have not already)

b. Load the `jtrain2` dataset.

   R: load package `wooldridge` with command `library(wooldridge)` and a `data.frame` variable named `jtrain2` becomes available; the command `?jtrain2` then shows you details about the dataset.

   Stata: run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse jtrain2 , clear`

c. R only: separate the data into "treatment" and "control" groups (depending on the value of `train`, the job training variable) with

```
trt <- jtrain2[jtrain2$train==1 , ]
ctl <- jtrain2[jtrain2$train==0 , ]
```

d. Estimate the mean 1978 earnings (in thousands of dollars) for the treatment group minus that of the control group, along with a 95% CI for the mean difference.

   R:

```
mean(trt$re78) - mean(ctl$re78)
t.test(x=trt$re78, y=ctl$re78)
```

   Stata: `ttest re78 , by(train) unequal` (also estimates the mean difference)

e. R only: separate out the data for treated, married individuals and untreated, married individuals, with

```
trt.mar1 <- trt[trt$married==1 , ]
ctl.mar1 <- ctl[ctl$married==1 , ]
```

f. Compute the mean difference estimate and 95% CI for the 1978 earnings outcome variable, comparing treated and untreated married individuals.

   R:

```
mean(trt.mar1$re78) - mean(ctl.mar1$re78)
t.test(x=trt.mar1$re78, y=ctl.mar1$re78)
```

Stata: `ttest re78 if married==1 , by(train) unequal` or alternatively `bysort married : ttest re78 , by(train) unequal`

g. Repeat your above analysis in parts (c)–(f), but first create a variable where earnings are in dollars (instead of thousands of dollars).

R: `jtrain2$re78USD <- 1000*jtrain2$re78`

Stata: `generate re78USD = re78*1000`

h. Optional: repeat your analysis in parts (e) and (f) for unmarried (instead of married) individuals.

i. Optional: repeat your analysis in parts (d)–(f) but for unemployment (`unem78`) instead of earnings. For interpretation: note that `unem78` equals 1 if unemployed all of 1978, and equals 0 otherwise, so the population mean is the probability of being unemployed all year (a value between $0 = 0\%$ and $1 = 100\%$), and the sample average is the fraction of the sample thus unemployed. So, a value like 0.14 means 14%, and a difference of $0.14 - 0.11 = 0.03$ is a difference of 3 percentage points, etc.

Empirical Exercise EE4.2. You will analyze data from an "audit study" that attempts to measure the effect of race on receiving a job offer. The Urban Institute found pairs of seemingly equally qualified individuals (one black, one white) and had them interview for a variety of entry-level jobs in Washington, DC in 1988. See Siegelman and Heckman (1993) for details and critique, and the raw data in their Table 5.1 (p. 195). In the data, each row (observation) corresponds to one job, to which one pair applied. Value `w=1` indicates that the white applicant in the pair got a job offer, while `b=1` if the black applicant got an offer.

a. R only: run `install.packages('wooldridge')` to download and install that package (if you have not already)

b. Load the `audit` dataset.

R: load package `wooldridge` with command `library(wooldridge)` and a `data.frame` variable named `audit` becomes available; the command `?audit` then shows you details about the dataset.

Stata: run `ssc install bcuse` to ensure command `bcuse` is installed, and then load the dataset with `bcuse audit , clear`

c. Compute the difference (white minus black) in the sample fraction of job offers.

R: `mean(audit$w) - mean(audit$b)`

Stata: `ttest w==b` (which also computes a 95% CI)

d. Compute the sample mean of all the pairs' white-minus-black difference. Note that `w-b` equals 1 if the white individual got a job offer but the black individual did not, equals -1 if the black but not white individual got an offer, and equals 0 if both or neither of the pair got an offer.

R: `mean(audit$w - audit$b)`

Stata: `generate wminusb = w-b` then `ttest wminusb==0` (also computes 95% CI; see row labeled `diff` for both).

e. R only (because Stata already reported this in the row labeled `diff`): compute a 95% CI for the population mean difference with either `t.test(x=audit$w, y=audit$b, paired=TRUE)` or `t.test(x=audit$w-audit$b)`

Chapter 5

Midterm Exam #1

⟹ Kaplan video: Chapter Introduction

When I teach this class, the first midterm exam is this week. This "chapter" makes the chapter numbers match the week of the semester. The midterm covers Chapters 2–4, i.e., everything up till now except R/Stata coding.

Part II

Regression

Introduction

Part II concerns regression. Regression is the workhorse of empirical economics (and many other fields), for description, prediction, and causality alike.

Part II extends the concepts and methods of Part I to the regression setting. In the population, the concepts of description, prediction, and causality from Part I are extended to regression models. In the data, estimation and inference methods extend those of Part I.

More flexible regression is also considered, including different models, interpretation, and a glimpse of nonparametric regression and machine learning.

Chapter 6

Comparing Two Distributions by Regression

⟹ Kaplan video: Chapter Introduction

Chapter 6 revisits Chapter 4 from the perspective of regression, with a single binary regressor X. The concepts of description, prediction, and causality are translated into regression language and regression models in the population. Estimation and quantifying uncertainty are also discussed.

The term **regression** has different meanings in different contexts (and by different people). In the population, it usually refers to how the mean of a random variable Y depends on the value of another random variable(s), as in Section 6.3. In the sample, as in Section 6.6, it usually refers to a particular estimation technique. But, beware of other (or ambiguous) uses of the word “regression,” especially in online resources.

Unit learning objectives for this chapter

6.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

6.2. Describe different ways of thinking about two distributions, both mathematically and intuitively [TLO 3]

6.3. Describe, interpret, identify, and distinguish among different population models and their parameters and estimators [TLO 3]

6.4. Judge which interpretation of a regression slope is most appropriate in a real-world example [TLO 6]

6.5. Interpret logical relationships and form appropriate logical conclusions [TLO 2]

6.6. In R (or Stata): estimate the parameters in a simple regression model, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

6.1 Logic

$\Longrightarrow$ Kaplan video: Logic Terms Example

Some basic logic is useful for understanding certain parts of econometrics. Theoretically, logic helps you understand the relationships among different conditions, like assumptions for theorems. Practically, logic helps you interpret results.

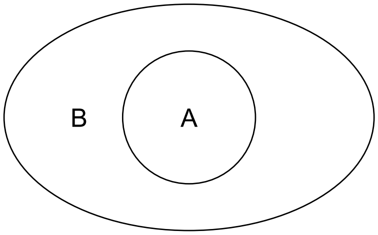
6.1.1 Terminology

Many words and notations can refer to the same logical relationship. Let A and B be two statements that can be either true or false. For example, maybe A is "$Y \geq 10$" and B is "$Y \geq 0$." Or, A is "this animal is a cat," and B is "this animal is a mammal." The following ways of describing the logical relationship between A and B all have the same meaning.

1. If A is true, then B is true (often shortened: "if A, then B")
2. $A \implies B$
3. A **implies** B
4. $B \impliedby A$
5. B is **implied by** A
6. B is true **if** A is true
7. A is true **only if** B is true
8. A is a sufficient condition for B (shorter: "A is **sufficient** for B")
9. B is a necessary condition for A (shorter: "B is **necessary** for A")
10. A is **stronger** than B
11. B is **weaker** than A
12. It is impossible for B to be false when A is true (but it is fine if both are true, or both are false, or A is false and B is true)
13. The truth table (T=true, F=false):

A	B	$A \implies B$
T	T	T
T	F	F
F	T	T
F	F	T

14. The diagram (everything in A is also in B):



To state equivalence of A and B, opposite statements can be combined. Specifically, any of the following have the same meaning.

1. $A \iff B$ (meaning both $A \implies B$ and $A \impliedby B$)
2. A is true **if and only if** B is true (meaning A is true if B is true *and* A is true only if B is true)
3. B is true if and only if A is true
4. A is necessary and sufficient for B
5. B is necessary and sufficient for A
6. A and B are equivalent
7. It is impossible for A to be false when B is true, and impossible for A to be true when B is false.
8. The truth table (T=true, F=false):

A	B	$A \iff B$
T	T	T
T	F	F
F	T	F
F	F	T

Variations of $A \implies B$ have the following names. Read $\neg A$ as "not A": $\neg A$ is false when A is true, and $\neg A$ is true when A is false.

- $\neg A \implies \neg B$ is the **inverse** of $A \implies B$.
- $B \implies A$ is the **converse** of $A \implies B$.
- $\neg B \implies \neg A$ is the **contrapositive** of $A \implies B$.

The statement $A \implies B$ is logically equivalent to its contrapositive. That is, statements "$A \implies B$" and "$\neg B \implies \neg A$" can be both true or both false, but it's impossible for one to be true and the other false.

The statement $A \implies B$ is not logically equivalent to either its inverse or converse. (The inverse and converse are equivalent to each other because the inverse is the contrapositive of the converse.)

Example 6.1 (Kaplan video)**.** Let A be "$X \leq 0$" and let B be "$X \leq 10$."

- $A \implies B$: any number below 0 is also below 10.
- We could equivalently say "A implies B" or "B is true if A is true" or "A is stronger than B" or "A is sufficient for B."
- The contrapositive is $X > 10 \implies X > 0$, which is also true: any number above 10 is also above 0.
- The inverse is $X > 0 \implies X > 10$, which is false: e.g., if $X = 5$, then $X > 0$ but not $X > 10$.
- The converse is $X \leq 10 \implies X \leq 0$, also false: again if $X = 5$, then $X \leq 10$ but not $X \leq 0$.

6.1.2 Theorems

Theorems all have the same logical structure: if assumption A is true, then conclusion B is true. Sometimes A and B have multiple parts, like A is really "A_1 and A_2." (Like, "If SUTVA holds and treatment is randomized, then the ATE is identified and equals the mean difference.") The theorem's practical use is: if we can verify that A is true, then we know B is also true.

What if we think A is false? Then, B could be false, or it could be true. This may be seen most readily from the picture version of the A and B relationship in Section 6.1.1: we could be somewhere inside B but outside A (i.e., B true, A false); or we could be outside both (both false). That is, as in Section 6.1.1, the theorem $A \implies B$ is not equivalent to its inverse.

Also from Section 6.1.1, a theorem is equivalent to its contrapositive. That is, if the theorem's conclusion is false, then we know at least one of its assumptions is false.

Example 6.2 (Kaplan video)**.** Consider three line segments, x, y, and z. Let A be "x, y, and z form a triangle"; let B be "the length of z is less than or equal to the sum of the lengths of x and y."

- In Euclidean geometry, if assumption A is true, then conclusion B is true; the theorem "$A \implies B$" is correct (known as the triangle inequality).
- The inverse is, "if A is false, then B is false," or: "if x, y, and z do not form a triangle, then the length of z is greater than the sum of the lengths of x and y." This statement is incorrect: if the segments do not form a triangle, then they can be any lengths.
- The contrapositive is, "if B is false, then A is false," or: "if the length of z is greater than the sum of the lengths of x and y, then the three segments do not form a triangle." The contrapositive is true; if you have three such segments, it's impossible to arrange them into a triangle.

6.1.3 Comparing Assumptions

To compare assumptions, the terms "stronger" and "weaker" are most commonly used. Let A_1 and A_2 denote different assumptions. Per Section 6.1.1, "A_1 is stronger than A_2" is equivalent to $A_1 \implies A_2$, which is also equivalent to "A_2 is weaker than A_1."

All else equal, it is more useful to have a theorem with weaker assumptions because it applies to more settings. That is, if $A_1 \implies A_2$, then we prefer a theorem based on A_2, the weaker assumption. A theorem based on A_1 can only be used when A_1 is true. In contrast, a theorem based on A_2 can be used not only when A_1 is true (because $A_1 \implies A_2$), but also sometimes when A_1 is false (but A_2 is still true).

Example 6.3 (Kaplan video)**.** Let assumption A_1 be, "a city is in Missouri," and let assumption A_2 be, "a city is in the United States." Consider the theorems $A_1 \implies B$ and $A_2 \implies B$. (The conclusion is irrelevant here, but to be concrete you could imagine B is "the city is in the northern hemisphere.") Because Missouri is part of the United

States, $A_1 \implies A_2$, i.e., A_1 is the stronger assumption and A_2 is the weaker assumption. We prefer the theorem based on the weaker assumption because it applies to more cities. For example, only the theorem $A_2 \implies B$ applies to Houston; A_1 is false, but A_2 is true. (And recall that when A_1 is false, the theorem $A_1 \implies B$ does not conclude that B is false; it just says, "I don't know if B is true or false," i.e., it is useless.)

Practice 6.1 (median theorem logic)**.** Consider the theorem, "If sampling is iid, then the sample median consistently estimates the population median." Hint: draw a picture and/or write it as $A \implies B$.

a) What does this tell us about consistency of the sample median when sampling is not iid?

b) What does this tell us about sampling when the sample median is not consistent?

Practice 6.2 (mean theorem logic)**.** Consider the theorem, "If sampling is iid and the population mean is well-defined, then the sample mean consistently estimates the population mean." Hint: there may be multiple possible pictures that show this relationship among A_1 (iid), A_2 (well-defined), and B (consistency).

a) What does this tell us about consistency of the sample mean when sampling is not iid?

b) What does this tell us about sampling when the sample mean is not consistent?

Discussion Question 6.1 (logic with feathers)**.** Consider two theorems. Theorem 1 says, "If X is an adult eagle, then it has feathers." Theorem 2 says, "If X is an adult bird, then it has feathers."

a) Describe each theorem logically: what's the assumption (A), what's the conclusion (B), what's the relationship?

b) State Theorem 1's contrapositive; is it true?

c) Compare: does Theorem 1 or Theorem 2 have a stronger assumption? Why?

d) Compare: which theorem is more useful? (Which applies to more situations?)

6.2 Preliminaries

⟹ Kaplan video: Joint, Marginal, and Conditional Distributions

This section reviews some material particularly useful for understanding regression. If it is not familiar to you from a previous statistics class, then you may want to consult additional resources for a deeper understanding; or you may not. In Section 6.2, there is no data; only the population is considered.

6.2.1 Population Mean Model in Error Form

The population mean $\mathrm{E}(Y)$ can be "modeled" in two equivalent ways. Both look silly and over-complicated, but they help bridge Chapter 4 to Chapter 6. Both use notation $\mu_Y \equiv \mathrm{E}(Y)$.

First, the mean can be written directly:

$$\mathrm{E}(Y) = \mu_Y. \tag{6.1}$$

Second, in terms of an **error term** $U \equiv Y - \mu_Y$, the **error form** of this model is

$$Y = \mu_Y + U, \quad \mathrm{E}(U) = 0. \tag{6.2}$$

Here $\mathrm{E}(U)$ is not "assumed" but simply follows from the definition of U and linearity:

$$\mathrm{E}(U) = \mathrm{E}(Y - \mu_Y) = \mathrm{E}(Y) - \mathrm{E}(\mu_Y) = \mu_Y - \mu_Y = 0. \tag{6.3}$$

This error term has a precise statistical meaning, but no causal or economic meaning.

To see the equivalence of the two models, take the mean of both sides of (6.2):

$$\mathrm{E}(Y) = \mathrm{E}(\mu_Y + U) = \mathrm{E}(\mu_Y) + \mathrm{E}(U) = \mu_Y + 0 = \mu_Y. \tag{6.4}$$

The error form often facilitates theoretical analysis, but the direct model is easier to interpret.

6.2.2 Joint and Marginal Distributions

The **joint distribution** describes the probabilities of possible (X, Y) pair values, like $\mathrm{P}(X = 4, Y = 1)$ or $\mathrm{P}(X = \text{cat}, Y = \text{dog})$. (With continuous X or Y, there are added technical complications, but the intuition is the same.) For regression, the focus is on numeric (discrete or continuous) X and Y. Implicitly, this also applies to categorical variables that have been turned into dummy variables with the indicator function, like $X = \mathbb{1}\{\text{cat}\}$ or $Y = \mathbb{1}\{\text{employed}\}$.

Each **joint probability** can be written multiple equivalent ways:

$$\mathrm{P}((X, Y) = (x, y)) = \mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x \text{ and } Y = y). \tag{6.5}$$

Example 6.4. Let X be years of education and Y type of pet. The joint distribution of (X, Y) consists of probabilities like $\mathrm{P}((X, Y) = (12, \text{rabbit})) = 0.02$, meaning 2% of the population has both 12 years of education and a pet rabbit. Or, $\mathrm{P}(X = 16, Y = \text{dog}) = 0.11$ means that 11% of the population has both 16 years of education and a pet dog.

Example 6.5 (Kaplan video)**.** Let $Y = 1$ if somebody is employed, and $Y = 0$ if not. Let $X = 1$ if somebody is married, and $X = 0$ if not. The joint distribution of employment and marital status describes the probabilities of each possible value of (X, Y). There are four possible values: not married and not employed $(0, 0)$; not married and employed $(0, 1)$; married and not employed $(1, 0)$; and married and employed $(1, 1)$. Table 6.1 shows an example, where $\mathrm{P}(X = 0, Y = 0) = 0.1$, $\mathrm{P}(X = 0, Y = 1) = 0.1$, $\mathrm{P}(X = 1, Y = 0) = 0.2$, and $\mathrm{P}(X = 1, Y = 1) = 0.6$.

Table 6.1: Joint distribution of marital status (X) and employment status (Y).

	$Y=0$	$Y=1$	Marginal for X (row sum)
$X=0$	0.10	0.10	0.20
$X=1$	0.20	0.60	0.80
Marginal for Y (column sum)	0.30	0.70	1.00

A **marginal probability** (or **unconditional probability**) considers just one of the random variables, ignoring the other. Whereas $\mathrm{P}(X=x, Y=y)$ is a joint probability, $\mathrm{P}(X=x)$ is a marginal probability, as is $\mathrm{P}(Y=y)$.

Example 6.6 (Kaplan video)**.** In Table 6.1, the outer values show the marginal probabilities: $\mathrm{P}(X=0)=0.20$ (at the right end of the $X=0$ row), $\mathrm{P}(X=1)=0.80$, $\mathrm{P}(Y=0)=0.30$ (at the bottom of the $Y=0$ column), and $\mathrm{P}(Y=1)=0.70$. That is, X by itself is a random variable with $\mathrm{P}(X=0)=0.20$ and $\mathrm{P}(X=1)=0.80$: the population probability of an individual being married is 0.8 (80%). Similarly, by itself, Y is a random variable with $\mathrm{P}(Y=0)=0.30$ and $\mathrm{P}(Y=1)=0.70$.

6.2.3 Conditional Distributions

The **conditional distribution** of Y given $X=x$ refers to the distribution of Y among individuals in the population with $X=x$.

The **conditional probability** of one event (like $Y=1$) given another event (like $X=0$) considers only the times when the conditioning event (like $X=0$) occurs, and then takes the proportion of *those* times that the first event (like $Y=1$) occurs. Mathematically, the conditional probability $\mathrm{P}(Y=1 \mid X=0)$ can be read as "the probability that Y equals one conditional on X equal to zero" or "the probability of Y being one given X equals zero" or other variations. More generally, $\mathrm{P}(Y=y \mid X=x)$ is "the probability that Y equals y conditional on X equal to x."

Example 6.7 (Kaplan video)**.** Let X be the type of pet and Y its age (years). The joint distribution of (X,Y) consists of probabilities like $\mathrm{P}(X=\text{cat}, Y=5)$. The marginal distribution of Y is the distribution of age for all pets in the population, consisting of probabilities like $\mathrm{P}(Y=5)$. The conditional distribution of Y given $X=\text{cat}$ is the distribution of age for cats, consisting of probabilities like $\mathrm{P}(Y=5 \mid X=\text{cat})$.

For non-continuous variables, a conditional probability can be written in terms of joint and marginal probabilities. Specifically,

$$\mathrm{P}(Y=y \mid X=x)=\frac{\mathrm{P}(Y=y, X=x)}{\mathrm{P}(X=x)}. \tag{6.6}$$

(This doesn't apply directly to continuous X because the denominator would be $\mathrm{P}(X=x)=0$, but there is an analogous result with similar intuition.)

Example 6.8 (Kaplan video)**.** Continuing Example 6.7, applying (6.6) yields $\mathrm{P}(Y = 5 \mid X = \text{cat}) = \mathrm{P}(Y = 5, X = \text{cat}) / \mathrm{P}(X = \text{cat})$, the probability of a 5-year-old cat divided by the probability of a cat.

Example 6.9 (Kaplan video)**.** For intuition, it can help to write probabilities as percentages and imagine each percent is a person. For example, in Table 6.1, $\mathrm{P}(X = 1, Y = 1) = 0.6 = 60\%$, so we can imagine 60 people who are both married and employed. Similarly, $\mathrm{P}(X = 1, Y = 0) = 0.2 = 20\%$, so we can imagine 20 people who are married and not employed. The conditional probability of employment given being married is then the proportion of married individuals who are employed: there are $60 + 20 = 80$ total married individuals, of whom 60 are employed, so $60/80 = 0.75$. This matches (6.6), which says $\mathrm{P}(Y = 1 \mid X = 1) = \mathrm{P}(X = 1, Y = 1) / \mathrm{P}(X = 1) = 0.6/0.8 = 0.75$. Similary, $20/80 = 0.25$ is the proportion of married individuals who are not employed, which matches $\mathrm{P}(X = 1, Y = 0) / \mathrm{P}(X = 1) = 0.2/0.8 = 0.25$.

6.2.4 Conditional Mean

The **conditional mean** is the mean of a conditional distribution. Notationally,

$$\mathrm{E}(Y \mid X = x) \tag{6.7}$$

is "the conditional mean of Y given $X = x$" or "the mean of Y conditional on $X = x$." This is the mean of the conditional distribution of Y given $X = x$. Extending the unconditional mean formula (2.4), the conditional mean is an average of possible values y_j weighted by their conditional probability $\mathrm{P}(Y = y_j \mid X = x)$,

$$\mathrm{E}(Y \mid X = x) = \sum_{j=1}^{J} \mathrm{P}(Y = y_j \mid X = x) y_j . \tag{6.8}$$

Example 6.10 (Kaplan video)**.** In Example 6.9, we computed the conditional distribution of employment status (Y) given being married ($X = 1$): $\mathrm{P}(Y = 1 \mid X = 1) = 0.75$ and $\mathrm{P}(Y = 0 \mid X = 1) = 0.25$. The mean of that conditional distribution is written $\mathrm{E}(Y \mid X = 1)$. We can use the usual expected value formula, plugging in conditional probabilities. For comparison, the unconditional and conditional (on $X = 1$) means of Y are, respectively,

$$\mathrm{E}(Y) = (0)\,\mathrm{P}(Y = 0) + (1)\,\mathrm{P}(Y = 1) = (0)(0.3) + (1)(0.7) = 0 + 0.7 = 0.7, \tag{6.9}$$

$$\begin{aligned} \mathrm{E}(Y \mid X = 1) &= (0)\,\mathrm{P}(Y = 0 \mid X = 1) + (1)\,\mathrm{P}(Y = 1 \mid X = 1) \\ &= (0)(0.25) + (1)(0.75) = 0 + 0.75 = 0.75. \end{aligned} \tag{6.10}$$

Because Y is binary (0 or 1), the (conditional) mean is the (conditional) probability of $Y = 1$: $\mathrm{E}(Y) = \mathrm{P}(Y = 1) = 0.7$ and $\mathrm{E}(Y \mid X = 1) = \mathrm{P}(Y = 1 \mid X = 1) = 0.75$. Compared to the overall population employment rate 0.7, the employment rate for married individuals is modestly higher, 0.75.

Table 6.2: Joint distribution of education (X) and weekly hours worked (Y).

	$Y=0$	$Y=20$	$Y=40$
$X=11$	0.10	0.05	0.05
$X=12$	0.05	0.10	0.15
$X=16$	0.10	0.10	0.30

Example 6.11 (Kaplan video)**.** Imagine Y is hours worked per week, which is either 0, 20, or 40, and X is years of education, which is either 11, 12, or 16. Using (6.8), the conditional mean is

$$\begin{aligned}\mathrm{E}(Y \mid X=x) = (0)\,\mathrm{P}(Y=0 \mid X=x) + (20)\,\mathrm{P}(Y=20 \mid X=x)\\ + (40)\,\mathrm{P}(Y=40 \mid X=x).\end{aligned} \tag{6.11}$$

Table 6.2 shows example joint probabilities. Consider the conditional mean $\mathrm{E}(Y \mid X = 16)$, the mean hours worked for individuals with 16 years of education. First, the marginal probability $\mathrm{P}(X = 16)$ sums all entries in the $X = 16$ row:

$$\mathrm{P}(X=16) = 0.10 + 0.10 + 0.30 = 0.5. \tag{6.12}$$

Second, plugging this into (6.6),

$$\begin{aligned}\mathrm{P}(Y=20 \mid X=16) &= \frac{\mathrm{P}(Y=20, X=16)}{\mathrm{P}(X=16)} = \frac{0.10}{0.50} = 0.2,\\ \mathrm{P}(Y=40 \mid X=16) &= \frac{\mathrm{P}(Y=40, X=16)}{\mathrm{P}(X=16)} = \frac{0.30}{0.50} = 0.6.\end{aligned} \tag{6.13}$$

Third, plugging these into (6.11),

$$\mathrm{E}(Y \mid X=16) = 0 + (20)(0.2) + (40)(0.6) = 4 + 24 = 28. \tag{6.14}$$

As a sanity check, the probability of $Y = 40$ is higher than that of $Y = 0$, so it makes sense that the conditional mean is above 20.

6.2.5 Independence and Dependence

If random variables X and Y are **independent**, then they are completely unrelated, statistically speaking. Notationally, independence is usually written as $X \perp\!\!\!\perp Y$, which is equivalent to $Y \perp\!\!\!\perp X$.

Independence implies equality of marginal and conditional distributions. Mathematically, the marginal (unconditional) distribution of Y is the same as the conditional distribution of Y given $X = x$, for any x. Intuitively, if X is unrelated to Y, then knowing the value of X has no information about the value of Y. Mathematically,

$$Y \perp\!\!\!\perp X \implies \mathrm{P}(Y=y) = \mathrm{P}(Y=y \mid X=x). \tag{6.15}$$

If the right side holds for *all* possible y and x, then it is equivalent to independence:

$$Y \perp\!\!\!\perp X \iff \mathrm{P}(Y=y)=\mathrm{P}(Y=y \mid X=x) \text{ for all possible } y \text{ and } x. \tag{6.16}$$

For continuous variables, the technicalities differ, but the intuition is the same.

Independence implies equality of marginal and conditional means, known as **mean independence**. That is, for any possible x value,

$$Y \perp\!\!\!\perp X \implies \mathrm{E}(Y)=\mathrm{E}(Y \mid X=x). \tag{6.17}$$

Independence implies many other properties, too, like $\mathrm{Cov}(X,Y)=\mathrm{Corr}(X,Y)=0$ and $\mathrm{P}(X=x,Y=y)=\mathrm{P}(X=x)\,\mathrm{P}(Y=y)$.

The opposite of independence is **dependence**. If any condition implied by independence does not hold, then the variables are dependent, written $X \not\perp\!\!\!\perp Y$. That is, using Section 6.1, if A is "$X \perp\!\!\!\perp Y$" and B is such that $A \implies B$, then the contrapositive is true: if B is false, then A is false.

Example 6.12. Here are a few examples, using definitions, implications, and contrapositives.

- Independence implies zero correlation; thus, if X and Y are correlated ($\mathrm{Corr}(X,Y) \neq 0$), then $X \not\perp\!\!\!\perp Y$.
- Mean independence is defined as $\mathrm{E}(Y)=\mathrm{E}(Y \mid X=x)$ for all x; thus, if $\mathrm{E}(Y \mid X=1) \neq \mathrm{E}(Y \mid X=0)$, then X and Y are not mean independent.
- Independence implies mean independence; thus (the contrapositive), if X and Y are not mean independent, then they are not independent.
- Consider binary X and Y with $\mathrm{P}(Y=1 \mid X=0)=0.3$ and $\mathrm{P}(Y=1 \mid X=1)=0.3$, which implies $\mathrm{P}(Y=0 \mid X=0)=\mathrm{P}(Y=0 \mid X=1)=0.7$. Thus, $X \perp\!\!\!\perp Y$.
- Consider binary X with Y taking values 1, 2, or 3. Let $\mathrm{P}(Y=2 \mid X=0)=1$ and $\mathrm{P}(Y=j \mid X=1)=1/3$ for $j=1,2,3$. Here, X and Y are mean independent because $\mathrm{E}(Y \mid X=0)=(1)(2)=2$ and $\mathrm{E}(Y \mid X=1)=(1/3)(1)+(1/3)(2)+(1/3)(3)=2$. However, they are not fully independent because (among other reasons) $\mathrm{P}(Y=1 \mid X=0)=0$ but $\mathrm{P}(Y=1 \mid X=1)=1/3$, not equal.

6.3 Conditional Mean Function: Description and Prediction

$\Longrightarrow$ Kaplan video: CMF (Binary X)

This and the following sections consider what we want to learn about the population, and how we can write it mathematically. There is no data, no estimation, no uncertainty.

For simplicity, X is binary ($X=0$ or $X=1$) in this chapter.

A **model** describes the relationship between two (or more) variables, like education and income. If it describes how income changes with education, then income is the usually written as Y and called the **outcome variable**, **regressand**, **dependent variable**,

left-hand side variable, or **response variable**, while education is written as X and called the **regressor**, **independent variable**, **right-hand side variable**, **predictor**, **covariate**, or **conditioning variable**.

Like before, these variables are treated mathematically as random variables, and the "population" is a joint probability distribution of the observable variables, (Y, X).

There are different models for different types of relationships between two variables. Section 6.3 models a statistical relationship with interpretations for description or prediction, whereas Sections 6.4 and 6.5 model causal relationships. Sometimes the descriptive and causal models coincide, but generally they differ.

In Sum: Conditional Mean Function

CMF: $m(x) \equiv \mathrm{E}(Y \mid X = x)$
Description: mean Y for subpopulation with same $X = x$
Prediction: with quadratic loss, optimal prediction of Y given $X = x$
Causality: CMF difference/slope sometimes has causal interpretation (Sections 6.4 and 6.5)

6.3.1 Conditional Mean Function

Using (6.7), let $m(\cdot)$ be the **conditional mean function** (CMF) of Y given X:

$$m(x) \equiv \mathrm{E}(Y \mid X = x). \tag{6.18}$$

That is, the CMF $m(\cdot)$ takes a value of x as input, like $x = 1$, and tells us the corresponding conditional mean of Y, like $\mathrm{E}(Y \mid X = 1) = 7$. Other names for the CMF are conditional mean response (CMR) and conditional expectation function (CEF).

Example 6.13 (Kaplan video)**.** For the example in Table 6.1, the CMF is $m(0) = 0.5$ and $m(1) = 0.75$, as shown here. From (6.10), $m(1) \equiv \mathrm{E}(Y \mid X = 1) = 0.75$. Also,

$$\begin{aligned} m(0) \equiv \mathrm{E}(Y \mid X = 0) &= (0)\,\mathrm{P}(Y = 0 \mid X = 0) + (1)\,\mathrm{P}(Y = 1 \mid X = 0) \\ &= \mathrm{P}(Y = 1 \mid X = 0) = \frac{\mathrm{P}(Y = 1, X = 0)}{\mathrm{P}(X = 0)} = \frac{0.1}{0.2} = 0.5. \end{aligned}$$

The two conditional means could be studied directly, as in Chapter 4. That is, Y^A has the distribution of Y for the $X = 0$ subpopulation, and Y^B has the distribution of Y for the $X = 1$ subpopulation; then, $\mathrm{E}(Y^A) = m(0)$ and $\mathrm{E}(Y^B) = m(1)$. However, the CMF regression model extends more readily to more complex settings.

It helps to remember what's random and what's non-random. The CMF $m(\cdot)$ is a non-random function, just as $\mathrm{E}(Y)$ is non-random. For any $X = x$, Y has a conditional distribution whose mean is $m(x)$, a non-random value. You can draw a graph of a CMF just like you graphed any other (non-random) function in high school. In contrast, $m(X)$

is a random variable. That is, there are multiple possible values of $m(X)$ because there are multiple possible values of X.

Example 6.14 (Kaplan video)**.** Continuing Example 6.13, consider $m(X)$ as a random variable. From Table 6.1, the marginal distribution of X is $\mathrm{P}(X=0)=0.2$, $\mathrm{P}(X=1)=0.8$. Thus, $m(X)$ is a random variable with

$$\mathrm{P}(m(X)=0.5)=\mathrm{P}(X=0)=0.2, \quad \mathrm{P}(m(X)=0.75)=\mathrm{P}(X=1)=0.8. \tag{6.19}$$

6.3.2 Linear CMF Model

With binary X, the model in (6.26) is equivalent to

$$\begin{aligned} Y &= m(0)\,\mathbb{1}\{X=0\}+m(1)\,\mathbb{1}\{X=1\}+V && (6.20)\\ &= m(0)(1-X)+m(1)(X)+V \\ &= m(0)+[m(1)-m(0)]X+V. && (6.21)\end{aligned}$$

Substituting $\beta_0 \equiv m(0)$ and $\beta_1 \equiv m(1)-m(0)$ yields the conventional linear CMF form,

$$Y=\beta_0+\beta_1 X+V, \quad \mathrm{E}(V \mid X)=0. \tag{6.22}$$

In (6.22), β_0 and β_1 are called the **parameters**. Greek letters like β are commonly used to denote unknown parameters in a population model. In the frequentist framework, these are seen as unknown but fixed (non-random) values, whereas Y, X, and V are random variables. In (6.22) specifically, β_0 is the intercept, and β_1 is the slope. Sometimes regression parameters are called **coefficients**; β_1 is the **slope coefficient** or the **coefficient on X**.

Model (6.22) is a linear CMF model. It is a "CMF" model because $\mathrm{E}(Y \mid X=x)=\beta_0+\beta_1 x$, or $\mathrm{E}(Y \mid X)=\beta_0+\beta_1 X$. The "linear" part is explained in Section 8.2.1; for now, it suffices to recall that a graph of $\beta_0+\beta_1 x$ is a straight line.

Because X is binary, no assumptions were required to write (6.22) given (6.26). However, when X has more than two possible values, it is more complicated, as in Chapter 7.

In Sum: Linear CMF with Binary X

$m(x)=\mathrm{E}(Y \mid X=x)=\beta_0+\beta_1 x$
$\beta_0=\mathrm{E}(Y \mid X=0)$
$\beta_1=\mathrm{E}(Y \mid X=1)-\mathrm{E}(Y \mid X=0)$

6.3.3 CMF Interpretation: Units of Measure

To interpret (6.22), first consider the units of measure of the parameters. Use the equivalent form $\mathrm{E}(Y \mid X)=\beta_0+\beta_1 X$. If Y is numeric, then the left-hand side $\mathrm{E}(Y \mid X)$

has the same units as Y. If Y is binary, then $\mathrm{E}(Y \mid X) = \mathrm{P}(Y = 1 \mid X)$, in probability units (like 0.43 for 43%). (Some variables may happen to only have values 0 and 1 in the data, but are numeric, like "number of siblings" or "number of master's degrees," whereas others are not numeric, like $Y = 1$ if it rains, or $Y = 1$ if it's a cat.) Because they are equal, the right-hand side must have the same units as $\mathrm{E}(Y \mid X)$. If X is numeric, then

1. β_0 has the same units as Y if Y is numeric, otherwise β_0 has probability units if Y is binary;
2. $\beta_1 X$ has the same units as $\mathrm{E}(Y \mid X)$, so the units of β_1 are the units of Y divided by units of X if Y is numeric, otherwise probability units divided by units of X (if Y is binary).

If X is not numeric, then β_1 is interpreted as $m(1) - m(0)$ and has the same units as $\mathrm{E}(Y \mid X)$. For example, if $X = 1$ means cat and $X = 0$ means dog, then β_1 is the mean Y difference between cats and dogs.

Example 6.15. If Y is salary measured in \$/yr, and X is the number of college degrees, then the units of β_0 are \$/yr and the units of β_1 are (\$/yr)/(#degrees).

Example 6.16. If $Y = 1$ if employed (and $Y = 0$ otherwise) and X is the number of pets, then β_0 has probability units, and β_1 is probability units divided by number of pets. For example, $\beta_0 = 0.80$ is 80 percentage points, and $\beta_1 = 0.11$ would be 11 percentage points per pet.

Practice 6.3 (regression parameter units)**.** Let Y be a country's annual GDP in \$/yr, and let X be how many oceans it borders. In (6.22), what are the units of measure for β_0 and β_1, respectively?

6.3.4 CMF Interpretation: Description

For description, the CMF is a summary of the conditional distribution. As seen in (6.21), $\beta_0 = m(0) = \mathrm{E}(Y \mid X = 0)$, the mean outcome among all individuals with $X = 0$, while $\beta_1 = m(1) - m(0)$, the mean Y difference between the $X = 1$ and $X = 0$ subpopulations. A common phrase to describe such statistical (but maybe not causal) differences is **associated with**.

Example 6.17 (Kaplan video)**.** If individuals who attended college ($X = 1$) have a mean annual income that is \$20,000/yr higher than the mean annual income of non-college individuals ($X = 0$), then $\beta_1 =$ \$20,000/yr, and you could say, "On average, having a college degree is associated with having a \$20,000/yr higher annual income." This does not claim that attending college has such a causal effect on income, only a statistical association.

6.3.5 CMF Interpretation: Prediction

For prediction, with a quadratic loss function, the CMF provides the optimal prediction of Y given $X = x$. Section 2.5.2 says the mean is the best predictor of Y (unconditionally) if the loss function is quadratic. This continues to be true conditionally: the conditional mean of Y given $X = x$ is the best predictor given quadratic loss. Intuitively, if you know somebody is from the $X = x$ subpopulation, then you want to use the subpopulation mean, which is $\mathrm{E}(Y \mid X = x)$.

Formally, a slightly different statement of prediction optimality follows, again assuming quadratic loss. Let $g(\cdot)$ denote any possible guess of Y, as a function of X. For an individual with (Y, X), our prediction error $Y - g(X)$ shows how wrong our guess was, and applying quadratic loss gives $(Y - g(X))^2$. Mean loss is thus $\mathrm{E}[(Y - g(X))^2]$. The CMF $m(\cdot)$ is the best possible guess in that it achieves the minimum possible mean loss:

$$m(\cdot) = \operatorname*{arg\,min}_{g(\cdot)} \mathrm{E}[(Y - g(X))^2]. \tag{6.23}$$

In terms of the model parameters in (6.22), the best predictor of Y given $X = 0$ is $m(0) = \beta_0$, the best predictor of Y given $X = 1$ is $m(1) = \beta_0 + \beta_1$, and the best predictor of Y given X is $m(X) = \beta_0 + \beta_1 X$.

As before, if quadratic loss is not appropriate in a particular situation, then the CMF may not provide a good prediction.

Example 6.18 (Kaplan video)**.** You know whether somebody is currently employed ($X = 1$) or not ($X = 0$), and want to predict their total net wealth (Y). With quadratic loss, the optimal prediction is $\mathrm{E}(Y \mid X)$, the mean wealth given their employment status.

Example 6.19 (Kaplan video)**.** You know whether somebody is currently employed ($X = 1$) or not ($X = 0$), and want to predict whether next week they will be employed ($Y = 1$) or not ($Y = 0$). The conditional mean gives you a decimal number (probability) between 0 and 1, like 0.86, which cannot possibly equal next week's employment status ($Y = 1$ or $Y = 0$), so you will always be wrong. Here, quadratic loss is not appropriate, so the conditional mean is not the best prediction.

6.3.6 CMF Model in Error Form

The CMF model is more confusing in error form, but it is commonly presented that way, so it is helpful to understand what it really means.

Extending the population mean error term in Section 6.2.1, the **CMF error term** is

$$V \equiv Y - m(X), \tag{6.24}$$

the difference between an individual's actual outcome Y and the CMF evaluated at their X value, $m(X)$. (Other letters could be used besides V, like U or W; in other textbooks, you may see u or e or ϵ.) Because Y and X are random variables, so is V.

Example 6.20 (Kaplan video)**.** Consider binary X and Y, with $\mathrm{E}(Y \mid X = 0) = 0.2$ and $\mathrm{E}(Y \mid X = 1) = 0.3$, and $\mathrm{P}(X = 1) = 0.6$. Because $m(X)$ has two possible values, and Y has two possible values, the CMF error term $V \equiv Y - m(X)$ has four possible values:

- $V = 1 - 0.2 = 0.8$ if $Y = 1$ and $X = 0$ (so $m(X) = 0.2$);
- $V = 0 - 0.2 = -0.2$ if $Y = 0$ and $X = 0$ (so $m(X) = 0.2$);
- $V = 1 - 0.3 = 0.7$ if $Y = 1$ and $X = 1$ (so $m(X) = 0.3$);
- $V = 0 - 0.3 = -0.3$ if $Y = 0$ and $X = 1$ (so $m(X) = 0.3$).

Thus, V is a random variable, with $\mathrm{P}(V = 0.8) = \mathrm{P}(Y = 1, X = 0)$, $\mathrm{P}(V = -0.2) = \mathrm{P}(Y = 0, X = 0)$, $\mathrm{P}(V = 0.7) = \mathrm{P}(Y = 1, X = 1)$, and $\mathrm{P}(V = -0.3) = \mathrm{P}(Y = 0, X = 1)$.

The CMF error always has conditional mean zero. Again, this is not an "assumption" but follows directly from its definition. Extending (6.3), for any $X = x$,

$$\begin{aligned}\mathrm{E}(V \mid X = x) &= \mathrm{E}(Y - m(X) \mid X = x) = \mathrm{E}(Y \mid X = x) - \mathrm{E}(m(X) \mid X = x) \\ &= m(x) - m(x) = 0.\end{aligned}$$

Equivalently,

$$\mathrm{E}(V \mid X) = 0. \tag{6.25}$$

That is, $\mathrm{E}(V \mid X)$ is a random variable depending on X, but it equals zero for every possible value of X; or, just imagine "$\mathrm{E}(V \mid X = x) = 0$ for all x" every time you see "$\mathrm{E}(V \mid X) = 0$."

Given (6.24), extending (6.2), the CMF model in error form is

$$Y = m(X) + V, \quad \mathrm{E}(V \mid X) = 0. \tag{6.26}$$

As shown, this is mathematically equivalent to saying explicitly that $m(\cdot)$ is the CMF and defining $V \equiv Y - m(X)$, which is much more clear but less commonly seen (on the internet, etc.).

6.4 Causality: Potential Outcomes and ATE

⟹ Kaplan video: Identification of College Effect on Earnings

Sections 4.4 and 4.6 are extended this chapter's notation, including more formal results. As before, Y^U and Y^T denote the untreated and treated potential outcomes, respectively.

Updating (4.12) to this chapter's notation, the ATE is identified when

$$\mathrm{E}(Y^T) - \mathrm{E}(Y^U) = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0). \tag{6.27}$$

We can directly learn about the right-hand side because we observe (Y, X) for all individuals, and it always has a descriptive and predictive interpretation. We cannot directly learn about the left-hand side because Y^T and Y^U are not both observable for all individuals. However, if (6.27) is true, then the conditional mean difference (right-hand side) also has a causal interpretation as the ATE (left-hand side).

6.4.1 Identifying Assumptions

Assumption A6.1 is SUTVA, as discussed in Section 4.4.

Assumption A6.2 is related to the discussion of randomized treatment in Section 4.3. Mathematically, the key is that randomization satisfies statistical independence between the treatment assignment and the individual's pair of potential outcomes: $X \perp\!\!\!\perp (Y^U, Y^T)$. (Technically, this could be weakened to "mean independence," but the intuition is the same.)

Assumption A6.3 was not discussed before, but it is intuitive: if everybody (or nobody) is treated, then it's impossible to compare treated and untreated outcomes. For example, if $\mathrm{P}(X = 1) = 0$, then nobody is treated, so it's impossible to learn about $\mathrm{E}(Y^T)$ because Y^T is never observed.

The following identifying assumptions combined together are sufficient, but not necessary. That is, if they are all true, then the ATE is identified, but there may be other ways to identify the ATE even if they are violated.

The assumptions have various names. Assumption A6.1 is usually just called SUTVA, but the main part of it is often called **no interference** (or **non-interference**). Assumption A6.2 has many names: **independence**, **ignorability**, or **unconfoundedness**. The combination of A6.2 and A6.3 is called **strong ignorability**. For more detail, history, and discussion, see Imbens and Wooldridge (2007).

Assumption A6.1 (SUTVA)**.** Everyone with $X = 1$ receives the same treatment, and one individual's treatment does not affect any other individual's potential outcomes.

Assumption A6.2 (unconfoundedness)**.** Treatment is independent of the potential outcomes: $X \perp\!\!\!\perp (Y^U, Y^T)$.

Assumption A6.3 (overlap)**.** There is strictly positive probability of both treatment and non-treatment: $0 < \mathrm{P}(X = 1) < 1$.

6.4.2 Identification Results

Theorem 6.1 formally states the ATE identification result. Intuitively, the key is that A6.2 allows us to observe representative samples of both Y^U and Y^T; treatment cannot be chosen or assigned based on an individual's potential outcomes. Mathematically, A6.2 implies that the means of the potential outcomes do not statistically depend on the treatment X:

$$\mathrm{E}(Y^T) = \mathrm{E}(Y^T \mid X = 1), \quad \mathrm{E}(Y^U) = \mathrm{E}(Y^U \mid X = 0). \tag{6.28}$$

We observe $Y = Y^T$ when $X = 1$ and $Y = Y^U$ when $X = 0$, so

$$\mathrm{E}(Y^T \mid X = 1) = \mathrm{E}(Y \mid X = 1), \quad \mathrm{E}(Y^U \mid X = 1) = \mathrm{E}(Y \mid X = 0). \tag{6.29}$$

Combining (6.28) and (6.29), this says that the population mean of the treated *potential* outcome, $\mathrm{E}(Y^T)$, equals the mean of the *observed* outcome in the treated population,

$\mathrm{E}(Y \mid X = 1)$. Thus, $\mathrm{E}(Y^T) = \mathrm{E}(Y \mid X = 1)$ is identified. Similarly, $\mathrm{E}(Y^U) = \mathrm{E}(Y \mid X = 0)$ is identified, so $\mathrm{E}(Y^T) - \mathrm{E}(Y^U)$ is identified.

Theorem 6.1 (ATE identification)**.** *Under A6.1–A6.3, the ATE is identified:*

$$\mathrm{E}(Y^T - Y^U) = \mathrm{E}(Y^T) - \mathrm{E}(Y^U) = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0),$$

which is the slope β_1 in the linear CMF model in (6.22)*.*

Proof. Using the above,

$$\begin{aligned} \mathrm{ATE} \equiv \overbrace{\mathrm{E}(Y^T - Y^U)}^{\text{use linearity, (4.3)}} &= \overbrace{\mathrm{E}(Y^T) - \mathrm{E}(Y^U)}^{\text{use (6.28)}} \\ &= \overbrace{\mathrm{E}(Y^T \mid X = 1)}^{\text{use (6.29)}} - \overbrace{\mathrm{E}(Y^U \mid X = 0)}^{\text{use (6.29)}} \\ &= \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0). \end{aligned}$$

□

Example 6.21. Imagine a knee surgery treatment (X) to help arthritis, where Y is knee-specific pain (between 0 and 100). For each individual, we can imagine two parallel universes, identical except for whether the individual gets the treatment (surgery) or not. It is the same surgery for everybody, and naturally one person's surgery cannot affect another person's pain, so SUTVA is satisfied. Half of patients are randomly assigned the treatment, so $X \perp\!\!\!\perp (Y^U, Y^T)$ and $0 < \mathrm{P}(X = 1) < 1$. Thus, Assumptions A6.1–A6.3 are all satisfied, and Theorem 6.1 says the ATE equals the CMF slope.

Example 6.22 (Kaplan video)**.** Consider Theorem 6.1 when X is rain and Y is commute time. In Columbia, MO, there is much less traffic in the "summer" (mid-May to mid-August) when most students are gone, meaning both Y^T and Y^U are lower. There is also more rain ($X = 1$). That is, X and (Y^U, Y^T) are related, violating Assumption A6.2. Intuitively, the problem is that we see more short rainy commutes in the summer and more long dry commutes during the academic year, which makes it seem like rain causes short commutes; but correlation does not imply causation.

Practice 6.4. Discuss the right-to-work example from Sections 4.3–4.5 in terms of Assumptions A6.1–A6.3.

6.5 Causality: Structural Model

A **structural model** also captures causal relationships. The assumption is that the model itself does not change even when variable values and policies change. ("Policy" has a broad meaning here: policies of countries, firms, schools, etc., or even just personal decisions.) More specifically, if we want to assess the causal effect of a certain policy change, then the structural model should be **invariant** to that particular policy change. That is, the policy may change the population distribution of variables, but it cannot change the structural model itself, otherwise the model is not useful.

6.5.1 Linear Structural Model

Consider the linear structural model

$$Y = \beta_0 + \beta_1 X + U. \tag{6.30}$$

Unlike in a CMF, the structural model's β_1 and U have economic and/or causal meaning by definition. In (6.30), β_1 is called a **structural parameter** (as is β_0). It has some economic or causal interpretation, like an elasticity or demand curve slope. Similarly, U is called the **structural error term**. This U can be interpreted as the aggregation of all other variables that causally determine Y. It's possible $\mathrm{E}(U \mid X) = 0$, but usually not. In contrast, the CMF error $Y - m(X)$ usually does not have causal or economic meaning.

Example 6.23. Let Y be an individual's income and $X = 1$ if they have a college degree. In the structural model (6.30), β_1 is the causal effect (on income) of getting a college degree, and U contains everything else that helps determine a person's income: their occupation, their different skill levels (human capital), where they live, etc.

The coefficient β_1 is perhaps best interpreted as an average effect of X on Y; we'll call it the **average structural effect** (ASE). Superficially, (6.30) seems to state that the effect of X on Y is the same for everybody, but it is possible that X secretly appears inside the structural error term U, too. For example, using potential outcomes notation, $Y = Y^U + X(Y^T - Y^U)$: we observe $Y = Y^U$ if $X = 0$, or we observe $Y = Y^T$ if $X = 1$. If we define $\beta_0 \equiv \mathrm{E}(Y^U)$ and $\beta_1 \equiv \mathrm{E}(Y^T - Y^U) = \mathrm{ATE}$, then

$$Y = \beta_0 + \beta_1 X + U, \quad U \equiv Y^U - \beta_0 + X(Y^T - Y^U - \beta_1). \tag{6.31}$$

Here, the structural error term U includes $Y^T - Y^U - \beta_1$, which is how much the individual's treatment effect $Y^T - Y^U$ differs from the average treatment effect β_1. In other settings, the ASE may not be directly related to the ATE and potential outcomes, but the interpretation is qualitatively similar.

Warning(!): if you see a model $Y = \beta_0 + \beta_1 X + U$, make sure you know whether it's a CMF model or a structural model, or yet another type of model (like in Chapter 7). The equation by itself only shows a linear relationship; it does not tell us the meaning of the parameters or the error term U. This is something to be very wary of when you look at econometric resources online or in other books; they may have models that look identical but are interpreted very differently.

Practice 6.5. Let $X = 1$ if an individual's body mass index (BMI) is 30 or greater (the technical definition of obesity) and $X = 0$ otherwise, and let Y denote hourly wage. Consider the model $Y = \delta_0 + \delta_1 X + W$, where δ_0 and δ_1 are unknown, non-random parameters, and W is the unobserved error term. What is the interpretation of δ_1 and W? Explain. (Hint: yes, this is a "trick" question with a very short answer, related to the above warning.)

Structural models do not need to be linear. More general models like $Y = h(X, U)$ are more realistic, but also more complex, so they are not considered here.

6.5.2 Identifying Assumptions

Qualitatively, the structural slope is identified if X and U are "unrelated." That is, the regressor X must be unrelated to the unobserved determinants of Y (that comprise U); U cannot be systematically higher or lower for certain X values. If they are indeed unrelated, then X is called **exogenous** (link to pronunciation). If not, then X is called **endogenous** (link to pronunciation). The precise mathematical condition for a regressor's exogeneity (or endogeneity) depends on the model.

Assumption A6.4 (mean independent error)**.** U is mean independent of X: $\mathrm{E}(U \mid X) = \mathrm{E}(U)$. For binary X, equivalently, $\mathrm{E}(U \mid X = 0) = \mathrm{E}(U \mid X = 1)$.

Assumption A6.4 is one way to quantify "exogeneity" of X here, because it is sufficient for the identification result in Theorem 6.2. Other ways to quantify exogeneity here are $X \perp\!\!\!\perp U$ (independent) and $\mathrm{Corr}(U, X) = 0$ (uncorrelated). Generally, from strongest to weakest,

$$X \perp\!\!\!\perp U \implies \mathrm{E}(U \mid X) = \mathrm{E}(U) \implies \mathrm{Corr}(U, X) = 0, \tag{6.32}$$

although with binary X the last two are equivalent ($\iff$).

6.5.3 Formal Results

Theorem 6.2 formally states the identification theorem. You do not need to write (or even fully understand) proofs for this class, but the proof may help deepen understanding and appreciation for some students.

Theorem 6.2 (linear structural identification)**.** *Consider the linear structural model in* (6.30) *with binary X. If A6.4 holds, then the structural slope β_1 is identified and equals the CMF slope. If additionally* $\mathrm{E}(U) = 0$*, then the structural intercept β_0 is also identified and equals the CMF intercept.*

Proof. Starting from the structural model,

$$Y = \beta_0 + \beta_1 X + U = \beta_0 + \beta_1 X + U + \overbrace{\mathrm{E}(U) - \mathrm{E}(U)}^{=0} = \overbrace{\beta_0 + \mathrm{E}(U)}^{\gamma_0} + \beta_1 X + \overbrace{U - \mathrm{E}(U)}^{V}.$$

The CMF intercept is $\gamma_0 = \beta_0 + \mathrm{E}(U)$ (which equals β_0 if $\mathrm{E}(U) = 0$) and the CMF slope is β_1 because $V \equiv U - \mathrm{E}(U)$ is a CMF error:

$$\mathrm{E}[U - \mathrm{E}(U) \mid X] = \overbrace{\mathrm{E}[U \mid X]}^{=\mathrm{E}(U) \text{ by A6.4}} - \mathrm{E}[\mathrm{E}(U) \mid X] = \mathrm{E}(U) - \mathrm{E}(U) = 0. \qquad \square$$

Example 6.24 (Kaplan video)**.** Let Y be commute time and $X = 1$ if people are carrying umbrellas (otherwise $X = 0$). Because the umbrellas themselves have no effect on Y, the structural $\beta_1 = 0$. Because rain affects Y, rain is part of U, although U may also include traffic conditions and such. When $X = 0$, there is probably no rain, whereas when $X = 1$, there probably is rain; thus, $\mathrm{E}(U \mid X=1) > \mathrm{E}(U \mid X=0)$. This structural error U is clearly not a CMF error. Consequently, the CMF slope has only statistical meaning, not causal meaning. If we could also observe weather conditions, then it might be plausible that the remaining parts of U are unrelated to X; this strategy is considered in Chapters 9 and 10.

Discussion Question 6.2 (ES habits and final scores)**.** Let Y be a student's final semester score in this class, $0 \leq Y \leq 100$, and $X=1$ if the student starts each exercise set well ahead of the due date (and $X=0$ if not). Consider the structural model $Y=a+bX+U$ and the CMF model $Y=c+dX+V$.

a) What does U represent? Give some specific examples of what U includes here. (Hint: imagine two students with the same X but different Y; what causes them to have different Y?)
b) Do you think $\mathrm{E}(U \mid X=0)=\mathrm{E}(U \mid X=1)$? Why/not?
c) Do you think $b=d$, $b<d$, or $b>d$? Why?

Practice 6.6 (ES habits: parameters)**.** In DQ 6.2, what would you guess are reasonable possible values of the parameters a, b, c, and d? Explain.

Discussion Question 6.3 (marriage and salary)**.** Let $X=1$ if married and otherwise $X=0$. Let Y be annual salary. Consider the structural model $Y=\beta_0+\beta_1 X+U$.

a) Explain why probably $\mathrm{E}(U \mid X=1) \neq \mathrm{E}(U \mid X=0)$, and say which you think is higher. (Hint: first think about what else is in U, i.e., what determines someone's salary; or think about variables that differ on average between married and unmarried individuals, and whether any of those help determine salary.)
b) Does the average salary difference between married and unmarried individuals have a structural meaning? Why/not?

6.6 Estimation: OLS

⟹ Kaplan video: OLS in R

This section considers estimation of the CMF model (6.22) when X is binary. The interpretation (description, prediction, causality) does not matter for estimation.

6.6.1 The Least Squares Approach

One approach is to define Y^A as the $X=0$ subpopulation and Y^B as the $X=1$ subpopulation. Then $\beta_0=\mathrm{E}(Y^A)$ and $\beta_1=\mathrm{E}(Y^B)-\mathrm{E}(Y^A)$, so Section 4.1.2 can be used to estimate $\mathrm{E}(Y^A)$ and $\mathrm{E}(Y^B)$.

Though simple, that approach does not generalize as well as **ordinary least squares** (OLS). The least squares intuition comes from the characterization of the CMF as minimizing the mean quadratic loss. This idea extends (3.9) for estimating the unconditional mean of Y. In the population, from (6.23) with $\mathrm{E}(Y \mid X) = \beta_0 + \beta_1 X$,

$$(\beta_0, \beta_1) = \operatorname*{arg\,min}_{b_0, b_1} \mathrm{E}[(Y - b_0 - b_1 X)^2]. \tag{6.33}$$

The analogous minimization problem in the sample is

$$\text{OLS:} \quad (\hat\beta_0, \hat\beta_1) = \operatorname*{arg\,min}_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2. \tag{6.34}$$

The estimated CMF is thus

$$\hat m(x) = \hat\beta_0 + \hat\beta_1 x. \tag{6.35}$$

Equation (6.34) can be described with the terms introduced around (3.10). Given any estimates $(\hat\beta_0, \hat\beta_1)$, the **fitted values** are

$$\hat Y_i \equiv \hat\beta_0 + \hat\beta_1 X_i = \hat m(X_i). \tag{6.36}$$

Given $\hat Y_i$, the **residual** is defined as

$$\hat U_i \equiv Y_i - \hat Y_i = Y_i - \hat\beta_0 - \hat\beta_1 X_i. \tag{6.37}$$

Consequently, (6.34) can be interpreted as saying that the OLS estimates $(\hat\beta_0, \hat\beta_1)$ make the sum of squared residuals $\sum_{i=1}^{n} \hat U_i^2$ as small as possible, hence “least” (smallest) “squares.”

6.6.2 Code

The following is the same empirical example from Section 4.1.2 but now with the `lm()` function to run OLS estimation. The outcome variable is wage; it is divided by 100 to get dollars instead of cents. The regressor is $X_i = 1$ if individual i at age 14 lived with their mother and father, and $X_i = 0$ if not. In the output below, the $X_i = 0$ sample mean equals the estimated regression intercept $(\hat\beta_0)$, and the $X_i = 1$ sample mean equals the sum of the estimated regression intercept and slope $(\hat\beta_0 + \hat\beta_1)$, so the sample mean difference equals the estimated slope $(\hat\beta_1)$.

```r
library('wooldridge')
# run OLS with lm()
ret <- lm(formula=wage/100~momdad14, data=card)
# show estimated intercept and slope
print(coef(ret), digits=2)
```

```
## (Intercept)    momdad14
##        5.11        0.84
```

```
# compute subsample means
means <- c( mean(card$wage[card$momdad14==0]),
            mean(card$wage[card$momdad14==1]) )
# show subsample means and mean difference
round( c(means, means[2]-means[1]) / 100 , digits=2)

## [1] 5.11 5.95 0.84
```

6.7 Quantifying Uncertainty

⟹ Kaplan video: OLS in R (again)

The ways to quantify uncertainty in Section 3.7 also apply to β_0 and β_1 in the linear CMF model (6.22). The same interpretations and misinterpretations apply. In particular, these methods do not reflect uncertainty about identifying assumptions.

One new consideration is discussed in Section 6.7.1, followed by sample code in Section 6.7.2.

6.7.1 Heteroskedasticity

Different methods for quantifying uncertainty make different assumptions about the conditional variance. Whereas the conditional mean $\mathrm{E}(Y \mid X = x)$ is the mean of the conditional distribution of Y given $X = x$, the conditional variance $\mathrm{Var}(Y \mid X = x)$ is the variance of the conditional distribution of Y given $X = x$. The term **homoskedasticity** means $\mathrm{Var}(Y \mid X = x)$ is a constant that does not depend on x, whereas **heteroskedasticity** means $\mathrm{Var}(Y \mid X = x)$ is not constant but instead varies with x. Equivalently, we could write $Y = \beta_0 + \beta_1 X + U$ and consider the conditional variance of U because $\mathrm{Var}(Y \mid X) = \mathrm{Var}(U \mid X)$, so often homoskedasticity and heteroskedasticity are thought of as properties of the error term.

Always use methods that are **robust to heteroskedasticity** (or **heteroskedasticity-robust**). This means they're valid with homoskedasticity or heteroskedasticity, whereas other methods only work with homoskedasticity. Logically, the heteroskedasticity-robust methods have weaker assumptions, so they work more often. Besides, heteroskedasticity is very common in real economic data.

(The term "robust" by itself is ambiguous. You should always ask: robust to what? Methods can be robust to heteroskedasticity, robust to clustered sampling, robust to measurement error, robust to infinite variance, etc.)

Example 6.25 (Kaplan video)**.** Consider a population of pets including birds, cats, dogs, and horses. Let Y be weight, and $X = 1$ if cat (otherwise $X = 0$). There is heteroskedasticity because there is much more variance in weight among non-cats (including very light birds and very heavy horses) than among cats. Mathematically, $\mathrm{Var}(Y \mid X = 0) > \mathrm{Var}(Y \mid X = 1)$.

Practice 6.7 (heteroskedasticity)**.** Let $Y = 1$ if employed (and $Y = 0$ if not), and let $X = 1$ if female (and $X = 0$ if not). Explain why there is probably heteroskedasticity. (Hint: if $p = \mathrm{P}(Y = 1)$, then $\mathrm{Var}(Y) = p(1 - p)$. If $p_x = \mathrm{P}(Y = 1 \mid X = x)$, then what's $\mathrm{Var}(Y \mid X = x)$?)

6.7.2 Code

Unfortunately, the default in R is to use homoskedasticity-based methods, so you have to make an extra effort to get heteroskedasticity-robust results. The below code does this. Because X is binary, the same results can be obtained with a two-sample unpaired t-test with "unequal variances," as shown.

The below code quantifies uncertainty about the CMF slope in a regression with a single, binary regressor. Using a variety of methods, the code computes a 95% confidence interval.

In the table of output at the very end, the first two rows assume homoskedasticity, whereas the remaining four rows do not. The first row uses a two-sample t-test assuming equal variances; the second row shows the default results based on `lm()` output. The third row uses a two-sample t-test allowing for unequal variances. The remaining rows use more general, regression-based methodology allowing for heteroskedasticity, based on the `lmtest` and `sandwich` packages in R (Zeileis, 2004; Zeileis and Hothorn, 2002).

Overall, the first two output rows are identical, and the following four rows are very similar to each other, but there is a big difference between the first two rows and the next four rows. This shows the (potentially) big difference between assuming homoskedasticity (as in the first two rows) and allowing for heteroskedasticity (as in the last four). There are multiple ways to allow for heteroskedasticity, like the HC0, HC1, and HC3 shown in the table. The differences are beyond our scope, but as the table suggests, the differences are often very small in practical terms.

```r
library(lmtest); library(sandwich)
set.seed(112358)
n <- 1000
df <- data.frame(Y=c(rnorm(n=n/4,mean=0,sd=1),
                     rnorm(n=3*n/4,mean=0.2,sd=2)),
                 X=c(rep(0,n/4), rep(1,3*n/4)))
ret <- lm(formula=Y~X, data=df)
# Store results for slope in sl.out
rn <- c('ttest.eq','Homosk.','ttest.uneq','HC0','HC1','HC3')
sl.out <- data.frame(row.names=rn, CI.lower=rep(NA,6), CI.upper=NA)
# HC0: original from Hal White (1980)
retVC0 <- vcovHC(ret, type="HC0")
```

```
# HC1: matches Stata default, and two-sample t.test below
retVC1 <- vcovHC(ret, type="HC1")
# HC3: recommended/default (and larger SE than HC0, HC1)
retVC3 <- vcovHC(ret, type="HC3")
# Heteroskedasticity-robust CIs (shortest to longest)
sl.out['HC0',] <- coefci(ret, vcov. = retVC0)['X',]
sl.out['HC1',] <- coefci(ret, vcov. = retVC1)['X',]
sl.out['HC3',] <- coefci(ret, vcov. = retVC3)['X',]
sl.out['Homosk.',] <- confint(ret, level=0.95)['X',]

# For comparison: t.test() results for slope
t.sl <- t.test(x=df$Y[df$X==1], y=df$Y[df$X==0], mu=0, conf.level=0.95,
               alternative='two.sided', paired=FALSE, var.equal=FALSE)
sl.out['ttest.uneq',] <- t.sl$conf.int
# For comparison: var.equal=TRUE
t2 <- t.test(x=df$Y[df$X==1], y=df$Y[df$X==0], mu=0, conf.level=0.95,
             alternative='two.sided', paired=FALSE, var.equal=TRUE)
sl.out['ttest.eq',] <- t2$conf.int
print(round(sl.out, digits=4))
```

```
##            CI.lower CI.upper
## ttest.eq    -0.0259    0.476
## Homosk.     -0.0259    0.476
## ttest.uneq   0.0382    0.412
## HC0          0.0385    0.412
## HC1          0.0383    0.412
## HC3          0.0381    0.412
```

Practice 6.8 (regression significance)**.** Consider the setup of the "audit study" from Bertrand and Mullainathan (2004). Resumes were fabricated that were identical except for the name: Emily (suggesting a white female), Greg (white male), Lakisha (black female), or Jamal (black male). The resumes were then submitted to job openings, and it was recorded whether or not an in-person interview for the job was then offered. Here, let $Y = 1$ if an interview was offered and $Y = 0$ if not; let $X = 1$ if the name is "black" and $X = 0$ if not. Note that $\mathrm{E}(Y \mid X = x) = \mathrm{P}(Y = 1 \mid X = x)$, i.e., the conditional probability of an interview. A regression of Y on X (including an intercept, as always) is run, and heteroskedasticity-robust 95% CIs are computed. Consider both economic significance and statistical significance in the following possible results.

a) Slope estimate $\hat{\beta}_1 = 0.00001$, CI $[0.000008, 0.000012]$.
b) $\hat{\beta}_1 = -0.1$, CI $[-0.3, 0.1]$.
c) $\hat{\beta}_1 = -0.2$, CI $[-0.24, -0.16]$.
d) $\hat{\beta}_1 = -0.01$, CI $[-0.03, 0.01]$.

Optional Resources

Optional resources for this chapter

- Conditional probability (Khan Academy)
- Basic joint, marginal, and conditional distributions (Khan Academy)
- James et al. (2013, §3.1)
- Covariance and correlation (Lambert video)
- Overlap assumption (Masten video)
- Correlation vs. causation (Masten video)
- Assumptions for randomized experiment validity (Masten video)
- Structural vs. causal/reduced form approach (Masten video)
- OLS computation (Masten video)
- Sections 2.1 ("Simple OLS Regression") and 2.2 ("Coefficients, Fitted Values, and Residuals") in Heiss (2016)
- Section 5.3 ("Regression When X is a Binary Variable") in Hanck et al. (2018)
- R packages `lmtest` and `sandwich` (Zeileis, 2004; Zeileis and Hothorn, 2002)

Empirical Exercises

Empirical Exercise EE6.1. You will essentially replicate EE4.1 but with regression commands.

a. R only: load the needed packages and look at a description of the dataset:

   ```
   library(wooldridge); library(sandwich); library(lmtest)
   ?jtrain2
   ```

b. Stata only: run `ssc install bcuse` if necessary, then load the data with

   ```
   bcuse jtrain2 , nodesc clear
   ```

c. Run a regression of 1978 earnings (`re78`) on the job training assignment indicator (`train`).

   R: `ret <- lm(re78~train, data=jtrain2)`

   Stata: `regress re78 train , vce(robust)` in which `vce(robust)` requests heteroskedasticity-robust standard errors

d. R only (because already reported in Stata): output the estimates along with heteroskedasticity-robust standard errors and two-sided 95% confidence intervals with the code

   ```
   coeftest(ret, vcov.=vcovHC(ret, type='HC1'))
   coefci(  ret, vcov.=vcovHC(ret, type='HC1'))
   ```

   where argument `type='HC1'` refers to one specific type (among multiple) of heteroskedasticity-robust standard error estimator (HC stands for "heteroskedasticity-consistent")

e. R only: create a subset of the data including only married individuals, with code `jt2.mar1 <- jtrain2[jtrain2$married==1 , ]`

f. Run your previous analysis for the subset of married individuals.

   R: replace `data=jtrain2` with `data=jt2.mar1`

   Stata: `regress re78 train if married==1 , vce(robust)`

g. Repeat your analysis, but for unmarried individuals

h. Repeat your analysis on the full sample of individuals, but for the outcome variable `unem78` (1978 unemployment indicator) instead of `re78` (and remember unemployment is bad, so negative coefficient is good).

Chapter 7

Simple Linear Regression

⟹ Kaplan video: Chapter Introduction

Surprisingly, many critical issues arise with three (instead of two) possible X values. With two, the regression modeled conditional means, useful for description, prediction, and (sometimes) causality. However, with three (or more) X values, we may fail to model the conditional means. In simple cases, this can be solved with a more flexible model; in other cases, we need to reinterpret what OLS actually estimates in practice.

Generally, OLS estimates something called a linear projection. This can also be interpreted as a "best" linear approximation of the CMF (for description) or a "best" linear predictor of Y given X (for prediction). These interpretations are discussed along with statistical properties of OLS as an estimator of the linear projection (not CMF).

Unit learning objectives for this chapter

7.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

7.2. Interpret what a linear regression estimates, in multiple ways, mathematically and intuitively [TLOs 2 and 3]

7.3. Assess whether certain assumptions for linear regression seem true or not in real-world examples [TLOs 2 and 6]

7.4. In R (or Stata): estimate a simple linear regression, along with measures of statistical uncertainty, and judge economic and statistical significance [TLO 7]

7.1 Misspecification

⟹ Kaplan video: Misspecification of Linear CMF

Consider the linear population model

$$Y = \beta_0 + \beta_1 X + U, \tag{7.1}$$

where supposedly $\mathrm{E}(U \mid X) = 0$, and this time X has three possible values: 0, 1, and 2.

Intuitively, you should worry already: there are now three conditional means, but still only two parameters. That is, we want to learn the three values

$$m(0) \equiv \mathrm{E}(Y \mid X = 0), \quad m(1) \equiv \mathrm{E}(Y \mid X = 1), \quad m(2) \equiv \mathrm{E}(Y \mid X = 2),$$

but (7.1) has only two parameters, β_0 and β_1. That's like trying to put three babies into only two car seats.

Mathematically, the question is whether the true $m(x)$ is indeed a straight line,

$$m(x) = \beta_0 + \beta_1 x, \quad x = 0, 1, 2.$$

Note: when economists (like me) call such functions "linear," they really mean "affine."

A wrong model is euphemistically termed **misspecified**. That is, the model assumes something that is not actually true. For the linear CMF model, it is misspecified when the true CMF is not linear, or equivalently when $m(1) - m(0) \neq m(2) - m(1)$. This type of misspecification is called **functional form misspecification** because it is the linear functional form that is wrong. That is, even though any values of (β_0, β_1) are allowed, $\beta_0 + \beta_1 x$ is always a straight-line function of x, so it has a linear **functional form** (the general "shape" of the function). If a model happens to be correct, then it is called **properly specified** (or **correctly specified**).

Example 7.1 (Kaplan video)**.** Consider Figure 7.1, in which Y is income and X is number of siblings. The three points represent the true CMF (in thousands of \$/yr): $m(0) = 60$ and $m(1) = m(2) = 40$. Qualitatively, there is a big income gap between only children ($X = 0$) and individuals with one sibling ($X = 1$), but having a second sibling ($X = 2$) is on average the same as having just one. From $m(1)$ and $m(2)$ alone, the CMF appears flat (zero slope); the line in Figure 7.1 with $\beta_0 = m(1)$ and $\beta_1 = 0$ fits these two points, but not the first point. But from $m(0)$ and $m(1)$ alone, the slope appears negative; the line in Figure 7.1 with $\beta_0 = m(0)$ and $\beta_1 = m(1) - m(0) < 0$ fits the first two CMF points, but not the third. Evidently, it is impossible to draw a straight line $(\beta_0 + \beta_1 x)$ through all three points on this CMF, as Euclid could tell us. Thus, the linear CMF is misspecified.

Practice 7.1 (misspecification)**.** Investigate whether the problem with the sibling example was that $X = 0$ was a possible value (so that the intercept had to be $\beta_0 = m(0)$), as follows. Consider the same example but with $X = 1, 2, 3$ instead of $X = 0, 1, 2$, so $m(1) = 60$, $m(2) = m(3) = 40$. Is it possible to write $m(x) = \beta_0 + \beta_1 x$ now? Why or why not?

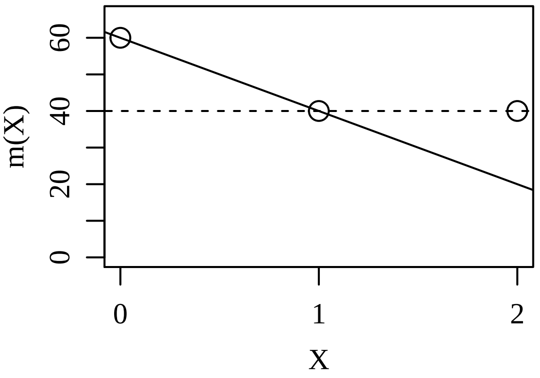


Figure 7.1: Misspecification of linear CMF.

7.2 Coping with Misspecification

There are two ways to cope with misspecification: change the model, or reinterpret it. The first way is now discussed for (7.1), while reinterpretation is detailed in Sections 7.3–7.5.

7.2.1 Model of Three Values

To fix the misspecification, the model needs to be more flexible. Continuing with $X = 0, 1, 2$ for simplicity, there are three conditional means, so the model should have three parameters to be flexible enough to avoid misspecification.

One way to add another parameter is to use a dummy variable (Section 2.3.1) for each possible value of X. Here,

$$\mathbb{1}\{X = j\} = \begin{cases} 1 & \text{if } X = j \\ 0 & \text{otherwise} \end{cases}, \quad j = 0, 1, 2. \tag{7.2}$$

Because only three values of X are possible, $\mathbb{1}\{X = 0\} = 1 - \mathbb{1}\{X = 1\} - \mathbb{1}\{X = 2\}$. Thus, extending (6.20),

$$\begin{aligned} m(x) &= m(0)\,\mathbb{1}\{x = 0\} + m(1)\,\mathbb{1}\{x = 1\} + m(2)\,\mathbb{1}\{x = 2\} &(7.3)\\ &= m(0)[1 - \mathbb{1}\{x = 1\} - \mathbb{1}\{x = 2\}] + m(1)\,\mathbb{1}\{x = 1\} + m(2)\,\mathbb{1}\{x = 2\} \\ &= m(0) + [m(1) - m(0)]\,\mathbb{1}\{x = 1\} + [m(2) - m(0)]\,\mathbb{1}\{x = 2\} \\ &= \beta_0 + \beta_1\,\mathbb{1}\{x = 1\} + \beta_2\,\mathbb{1}\{x = 2\}, \\ \beta_0 &\equiv m(0),\ \beta_1 \equiv m(1) - m(0),\ \beta_2 \equiv m(2) - m(0). &(7.4) \end{aligned}$$

Although the structure of (7.3) is easier to interpret, the structure of (7.4) is more common and can be interpreted as follows. The parameter $\beta_0 = m(0)$ is the conditional mean for the **base category** $X = 0$. The other parameters show how other conditional means differ from this base category. Specifically, $\beta_1 = m(1) - m(0)$ is the conditional mean difference between the $X = 1$ and $X = 0$ subpopulations, and $\beta_2 = m(2) - m(0)$ is the conditional mean difference between the $X = 2$ and $X = 0$ subpopulations.

Example 7.2 (Kaplan video)**.** Continue from Example 7.1, where Y is income (in thousands of \$/yr), and X is number of siblings. The parameter β_0 is the population mean income among individuals with zero siblings (the base category). Then, β_1 is the difference in mean income between the 1-sibling and 0-sibling subpopulations. Earlier, $m(0) = 60$ and $m(1) = 40$, so $\beta_1 = m(1) - m(0) = -20$. Finally, β_2 is the mean income difference between the 2-sibling and 0-sibling (not 1-sibling) subpopulations, $m(2) - m(0) = 40 - 60 = -20$.

Discussion Question 7.1 (Facebook)**.** Let $X = 0, 1, 2$ be the number of Facebook accounts somebody has, and Y is hours of social media consumption per week.

a) Explain what it means for a CMF model $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$ to be misspecified.
b) Describe a specific real-world reason to suspect misspecification in this example.
c) Consider the CMF model in (7.4). Guess whether β_1 is negative, zero, or positve, and explain why (using real-world reasons). Do the same for β_2.

7.2.2 More Than Three Values

More generally, even if X has more than three possible values, dummy variables could be used similarly to avoid CMF misspecification. Extending (7.3), there can be a dummy variable for each possible value of X, and a corresponding parameter for each. Any such model allowing an arbitrarily different conditional mean of Y for each possible value of X is called **fully saturated**. A fully saturated CMF model cannot be misspecified. (But, it may not have any causal meaning and may be practically impossible to estimate.)

In more complex settings, it is impossible to fix misspecification completely. For example, if X could be any real number between 0 and 1, then an infinite number of parameters is required to model the conditional expectations for the infinite number of X values; this is impossible in practice.

In such settings where misspecification is unavoidable, how can we interpret the model and its parameters? There are three interpretations of a more general linear model that includes the linear CMF model as a special case. These are discussed next.

In Sum: Interpretations of What OLS Estimates

1. *Linear projection (LP)*: gets $\beta_0 + \beta_1 X$ "closest to" Y, probabilistically (Section 7.3)
2. *"Best" linear approximation (BLA) of CMF*: "best" (smallest mean quadratic loss) approximation of $\mathrm{E}(Y \mid X)$ with linear form $\beta_0 + \beta_1 X$ (Section 7.4)
3. *"Best" linear predictor (BLP)*: "best" (smallest mean quadratic loss) prediction of Y given X with linear form $\beta_0 + \beta_1 X$ (Section 7.5)

7.3 Linear Projection

⟹ Kaplan video: Linear Projection and "Best" vs. "Good"

The linear projection model is important because it is what OLS actually estimates. Two additional interpretations are described in Sections 7.4 and 7.5.

7.3.1 Geometric Intuition

You may have seen orthogonal **projection** in geometry or linear algebra. There is some shape (or vector space), and there is a point outside it. Projecting the point onto the shape consists of finding the point within the shape that is closest to the outside point.

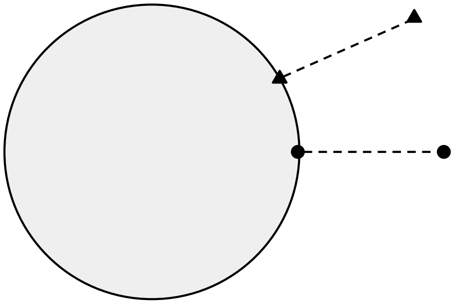


Figure 7.2: Orthogonal projection

Figure 7.2 illustrates projection. There is a large gray circle shape, and two points outside of it (small triangle, dot). The small triangle on the border of the large circle is the "closest" point to the outside small triangle, as measured by Euclidean distance. That is, the dashed line connecting the small triangles is just barely long enough to reach the gray circle from the outside triangle point; if it were any shorter, it could not reach any point in the gray circle. Similarly, the dot on the border of the gray shape is the projection of the outside dot onto the shape: of all the points in the gray space, it is closest to the outside dot (by Euclidean distance).

This idea can be written mathematically. Let $d_E(w, z)$ denote the Euclidean distance between points w and z. Let $\mathcal{S}$ denote a shape, which is a set of points. Let y denote the outside point, and p the projection. In Figure 7.2, the gray circle is $\mathcal{S}$, the outside small triangle (or dot) is y, and the small triangle (or dot) on the circle's border is p. The projection of point y onto shape $\mathcal{S}$ is the point inside $\mathcal{S}$ that's closest to y, i.e., that minimizes the distance to y. Mathematically,

$$p = \arg\min_{s \in \mathcal{S}} d_E(y, s). \tag{7.5}$$

7.3.2 Probabilistic Projection

Linear projection with random variables is the same idea, but with a different definition of distance and a different "shape" to search over.

Notationally, let $\mathrm{LP}(Y \mid 1, X)$ denote the **linear projection** (LP) of Y onto $(1, X)$. The $(1, X)$ specifies the "shape" that we search over: random variables that can be written as $a + bX$ for constants a and b, i.e., linear combinations of $(1, X)$. (Linear combinations and linearity are detailed in Section 8.2.1.) Without the 1, $\mathrm{LP}(Y \mid X)$ would only consider bX with no intercept.

The closest "point" inside the "shape" is usually written $\beta_0 + \beta_1 X$. Mathematically, parallel to (7.5),

$$\mathrm{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \underset{a+bX}{\arg\min}\, d(Y, a + bX) = \underset{a+bX}{\arg\min} \sqrt{\mathrm{E}[(Y - a - bX)^2]}, \quad (7.6)$$

where Euclidean distance $d_E(\cdot, \cdot)$ has been replaced by a probabilistic "distance" measure

$$d(A, B) \equiv \sqrt{\mathrm{E}[(A - B)^2]}. \quad (7.7)$$

Linear projection gets $\beta_0 + \beta_1 X$ as "close" to Y as possible, in a probabilistic sense.

7.3.3 Formulas and Interpretation

Some calculus (omitted) yields a formula for each **linear projection coefficient** (LPC), β_0 and β_1. In this special case with a single regressor X and an intercept,

$$\beta_1 = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}, \quad \beta_0 = \mathrm{E}(Y) - \beta_1 \mathrm{E}(X). \quad (7.8)$$

The slope β_1 can be rewritten in terms of correlation:

$$\beta_1 = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(Y, X)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}} \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}} = \mathrm{Corr}(Y, X) \sqrt{\frac{\mathrm{Var}(Y)}{\mathrm{Var}(X)}}. \quad (7.9)$$

Either version of the formula shows how the linear projection slope β_1 is related to the linear dependence (covariance or correlation) between Y and X. Once the slope is determined, the intercept β_0 simply moves the linear projection line up or down so that $\mathrm{E}(Y) = \beta_0 + \beta_1 \mathrm{E}(X)$. That is, the linear projection always goes exactly through the point $(x, y) = (\mathrm{E}(X), \mathrm{E}(Y))$.

People often interpret the linear projection coefficients less precisely. For the slope, a common phrase is, "A one-unit increase in X is **associated with** a β_1 change in Y." The intercept is often not mentioned because $\beta_0 = \mathrm{E}(Y) - \beta_1 \mathrm{E}(X)$ is not easy to interpret, except when the regressor has been demeaned so that $\mathrm{E}(X) = 0$, in which case $\beta_0 = \mathrm{E}(Y)$. In this case, β_0 is called the "centercept" instead of intercept; but despite the better interpretation, it is rarely seen in economics.

For description, (7.8) shows that the LPCs summarize the joint probability distribution of (Y, X). The joint distribution of (Y, X) determines $\mathrm{E}(Y)$, $\mathrm{E}(X)$, $\mathrm{Cov}(Y, X)$, and $\mathrm{Var}(X)$, which then determine β_0 and β_1. Although a two-number summary of a complicated joint distribution is very convenient, clearly much information is lost in such a summary. Methods like quantile regression complement the LPCs in describing (Y, X), but such are beyond our scope.

7.3.4 Linear Projection Model in Error Form

Analogous to (6.26) for the CMF, the linear projection model can be written in error form. Define the LP error as

$$U \equiv Y - \mathrm{LP}(Y \mid 1, X) = Y - (\beta_0 + \beta_1 X). \tag{7.10}$$

This implies $\mathrm{E}(U) = \mathrm{Cov}(X, U) = 0$. Thus, the model

$$Y = \beta_0 + \beta_1 X + U, \quad \mathrm{E}(U) = \mathrm{Cov}(X, U) = 0 \tag{7.11}$$

is equivalent to $\mathrm{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X$.

7.4 Description: "Best" Linear Approximation

⟹ Kaplan video: "Best" Linear Approximation

⟹ Kaplan video: Linear Projection and "Best" vs. "Good" (again)

7.4.1 Definition and Interpretation

For description, the linear projection can be interpreted as the **best linear approximation** (BLA) of the true CMF. "Best" here assumes quadratic loss, similar to how the mean $\mathrm{E}(Y)$ is the "best" predictor of Y with quadratic loss. "Linear" refers to a function of the form $a + bX$ (details in Section 8.2.1). Mathematically,

$$\mathrm{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \overbrace{\operatorname*{arg\,min}_{a+bX} \mathrm{E}\{[m(X) - (a + bX)]^2\}}^{\text{BLA}}, \quad m(X) \equiv \mathrm{E}(Y \mid X). \tag{7.12}$$

That is, among all possible $a + bX$, the linear projection $\beta_0 + \beta_1 X$ is the function of X that best approximates $\mathrm{E}(Y \mid X)$.

The linear projection equals the CMF if the CMF is linear, but otherwise the BLA treats more probable X as more important when trying to get the linear approximation "close" to the true CMF. The mean $\mathrm{E}\{\cdot\}$ in (7.12) is a weighted average with more weight on more probable X, so it is more important to make $m(X) - (a + bX)$ close to zero for such X values.

7.4.2 Limitations

Unfortunately, “best” does not always mean “good.” Sometimes, the CMF is so highly nonlinear that even the best linear approximation is still a very poor approximation. By analogy: “Among all cities in Missouri, St. Louis is closest to Kuwait” does not mean “St. Louis is close to Kuwait.” (Kuwait is the true CMF, Missouri is the set of all functions linear in X, and St. Louis is the BLA.) Sometimes the best (closest) is still not good (not close).

Example 7.3 (Kaplan video)**.** The following example of a “bad” BLA is from Hansen (2020, §2.28). Let $Y = X + X^2$, with no error term, so $m(x) = x + x^2$, too. If $X \sim \mathrm{N}(0,1)$, then the BLA/LP turns out to be $\mathrm{LP}(Y \mid 1, X = x) = 1 + x$. The function $1 + x$ is a bad approximation of $x + x^2$ (try graphing it).

Further, the distribution of X can greatly affect the BLA of a nonlinear CMF.

Example 7.4 (Kaplan video)**.** Figure 7.1 showed two possible BLA lines for the same nonlinear CMF. One line is the BLA when the distribution of X satisfies $\mathrm{P}(X = 2) = 0$. The other line is the BLA when $\mathrm{P}(X = 0) = 0$. The two lines are very different.

7.5 Prediction: “Best” Linear Predictor

For prediction, the linear projection can be interpreted as the **best linear predictor** (BLP) of Y given X. As with the BLA, “best” assumes quadratic loss, and “linear” refers to the form $a + bX$. As in (2.23), the optimal predictor minimizes mean quadratic loss. Mathematically,

$$\mathrm{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X = \overbrace{\operatorname*{arg\,min}_{a+bX} \mathrm{E}\{[Y - (a + bX)]^2\}}^{\text{BLP}} . \tag{7.13}$$

That is, among all possible $a + bX$, the linear projection $\beta_0 + \beta_1 X$ is precisely the function of X that “best” predicts Y given knowledge of X.

Mathematically, (7.13) is the same as (7.6) but without the $\sqrt{\cdot}$. Although phrased differently, the linear projection goal of getting $\beta_0 + \beta_1 X$ “closest” to Y is essentially the same as prediction: we want a predictor $\beta_0 + \beta_1 X$ that is “closest” to Y.

Unfortunately, as with BLA, “best” does not mean “good.” However, as with BLA, this means the CMF does not need to be exactly linear in order for the linear projection to make good predictions.

As in Section 2.5, “prediction” here is defined entirely within the population. It does not refer to using data to guess the future; there is no data here. Instead, the BLP is an ideal predictor; it is the (linear) predictor we would use if we fully knew everything about the population. The BLP is something we wish to learn. Fortunately, the BLP (and BLA and LP) is precisely what OLS estimates.

Discussion Question 7.2 (BLP)**.** Let Y be income (thousands of dollars per year) and X be number of siblings. When $X = 0$, the mean Y is 60 and $50 \leq Y \leq 70$. When $X = 1$, the mean Y is 40 and $30 \leq Y \leq 50$. When $X = 2$, it's the same as when $X = 1$: the mean Y is 40 and $30 \leq Y \leq 50$. In a population with mostly $X = 1$ and $X = 2$, the BLP is $\mathrm{LP}(Y \mid 1, X) = 43 - 2X$.

a) What Y does the BLP predict when $X = 0$?
b) Is the prediction from (a) good? Why/not?

7.6 Causality Under Misspecification

Some things can be said about causality under misspecification, but none as pleasing as the BLP for prediction or BLA for description. For example, if the structural error U satisfies the CMF error property $\mathrm{E}(U \mid X) = 0$, then the structural function is the CMF, so the linear projection is also the best linear approximation of the structural function. Alternatively, if the structural model is linear, $Y = \beta_0 + \beta_1 X + U$, and if $\mathrm{Cov}(X, U) = 0$, then β_1 equals the linear projection slope coefficient (regardless of whether the CMF is linear). However, the linear structural model may be misspecified, too. This is one motivation for "nonparametric" CMF estimation (Section 8.3).

7.7 OLS Estimation and Inference

⟹ Kaplan video: OLS in R

OLS estimation was initially discussed in Section 6.6, along with important terms like fitted values and residuals. Here are additional insights, statistical properties, and code.

7.7.1 OLS Estimator Insights

OLS does not try to estimate the CMF, but rather the BLP. The "least squares" formulation of the OLS estimator in (6.34) mirrors the BLP definition in (7.13). That is, following the analogy principle (Section 3.3), replacing the population mean (E) in (7.13) with the sample mean ($\frac{1}{n}\sum_{i=1}^{n}$) yields the OLS estimator, (6.34). This reinforces that OLS fundamentally estimates the BLP (or equivalently LP or BLA), *not* the CMF.

7.7.2 Statistical Properties

The following statistical properties consider OLS as an estimator of the linear projection coefficients. These properties hold true under relatively general assumptions.

Assumptions

The following assumptions combined are sufficient for Theorems 7.1 and 7.2 but not necessary (using logical terms from Section 6.1).

Assumption A7.1 (iid sampling)**.** Sampling of (Y_i, X_i) is iid.

Assumption A7.2 (non-constant regressor)**.** The regressor X is not a constant, i.e., there is no single value x such that $\mathrm{P}(X = x) = 1$.

Assumption A7.3 (finite variances)**.** The variances of Y and X are finite: $\mathrm{Var}(Y) < \infty$, $\mathrm{Var}(X) < \infty$. Or, equivalently, the expected values of Y^2 and X^2 (the **second moments** of Y and X) are finite: $\mathrm{E}(Y^2) < \infty$, $\mathrm{E}(X^2) < \infty$.

Assumption A7.4 (finite fourth moments)**.** The expected values of Y^4 and X^4 (**fourth moments**) are finite: $\mathrm{E}(Y^4) < \infty$, $\mathrm{E}(X^4) < \infty$.

Assumption A7.1 was discussed in Section 3.2 for Y_i by itself. If we let vector $\boldsymbol{W}_i \equiv (Y_i, X_i)$ be what's observed about individual i, and vector $\boldsymbol{W}_k \equiv (Y_k, X_k)$ be the observation for individual k, then the iid assumption is essentially the same as before: $\boldsymbol{W}_i \perp\!\!\!\perp \boldsymbol{W}_k$ for $i \neq k$ ("independent"), and $\boldsymbol{W}_i$ and $\boldsymbol{W}_k$ have the same distribution ("identically distributed"). That is, "independent" means $(Y_i, X_i) \perp\!\!\!\perp (Y_k, X_k)$ for $i \neq k$, which implies $Y_i \perp\!\!\!\perp Y_k$, $X_i \perp\!\!\!\perp X_k$, $Y_i \perp\!\!\!\perp X_k$, and $X_i \perp\!\!\!\perp Y_k$, but implies nothing about (in)dependence between X_i and Y_i (or X_k and Y_k). "Identically distributed" says (Y_i, X_i) and (Y_k, X_k) have the same joint distribution, which implies the conditional and marginal distributions (and their features) are also identical. For example, $\mathrm{E}(Y_i) = \mathrm{E}(Y_k)$, $\mathrm{Var}(X_i) = \mathrm{Var}(X_k)$, $\mathrm{E}(Y_i \mid X_i = x) = \mathrm{E}(Y_k \mid X_k = x)$, $\mathrm{P}(Y_i \leq 0 \mid X_i = x) = \mathrm{P}(Y_k \leq 0 \mid X_k = x)$, etc. All this readily generalizes to multiple regressors, just redefining $\boldsymbol{W}_i \equiv (Y_i, X_{1i}, X_{2i}, \ldots)$.

Assumption A7.2 is qualitatively similar to the overlap assumption (A6.3). They both say we must see different values of X in order to learn about a relationship involving X. Conveniently, if A7.2 seems false in the data, then your statistical software will report an error or warning.

Assumptions A7.3 and A7.4 are similar, but A7.4 is stronger: A7.4 $\Longrightarrow$ A7.3.

Assumptions A7.3 and A7.4 are usually true with economic data, but there are some exceptions. For example, bounded variables like age or education have $|Y| \leq b$ for some (finite) maximum possible value b, in which case $\mathrm{E}(Y^4) \leq \mathrm{E}(b^4) = b^4 < \infty$, satisfying A7.3 and A7.4. However, stock returns (or other asset returns) may not have finite fourth moment or even variance. (Or, maybe technically they do, but it is not a good theoretical approximation because they can have large outliers.) Whether to model such financial returns with finite or infinite variance is a matter of ongoing debate (e.g., Grabchak and Samorodnitsky, 2010).

Theoretical Results

Theorem 7.1 (OLS consistency, 1 regressor)**.** *If A7.1–A7.3 are true, then the OLS intercept and slope estimators are consistent for the population linear projection intercept and slope.*

Theorem 7.1 says that with enough data, the OLS coefficient estimators should be close to the true linear projection coefficients with high probability. However, with a small dataset, the OLS estimates may not be close to the true values; and even with a large dataset, without further assumptions, the OLS estimates may not tell us anything about causality or even the CMF.

Logically, Theorem 7.1 does not say that OLS is a bad estimator if sampling is not iid (the "inverse"), as discussed in Section 6.1. In fact, the iid assumption can be relaxed in certain ways; for example, even with some "dependence" (like with time series data as in Part III), OLS can still consistently estimate the population linear projection coefficients.

Theorem 7.2 (coverage probability, 1 regressor)**.** *If A7.1, A7.2, and A7.4 are true, then the heteroskedasticity-robust confidence intervals in Section 7.7.3 are asymptotically correct. That is, with a large enough sample size, the coverage probability is approximately equal to the desired confidence level.*

The accuracy of the confidence intervals mentioned in Theorem 7.2 comes from being able to accurately approximate the sampling distribution of the OLS estimator, but the technical details are not helpful in practice. A good approximation of the sampling distribution is also possible with non-iid sampling, but the confidence intervals must be constructed differently (like with a different R function).

7.7.3 Code

The following code is based on the example from Section 7.1. Each row in the final output shows the estimate $\hat{\beta}_1$ (in the column labeled `X`) along with a heteroskedasticity-robust 95% CI for β_1 (lower endpoint in column `2.5 \%`, upper endpoint in column `97.5 \%`).

There are three randomly simulated datasets. All have $X \in \{0, 1, 2\}$. The first dataset has linear CMF $m(x) = 60 - 20x$, and $\text{P}(X = j) = 1/3$ for $j = 0, 1, 2$. The next two datasets have nonlinear CMF $m(0) = 60$, $m(1) = m(2) = 40$, but different distributions of X: for $j = 0, 1, 2$, the first distribution has $\text{P}(X = j) = (3 - j)/6$ while the second has $\text{P}(X = j) = (j + 1)/6$. As seen, the distribution of X affects the linear projection slope when the CMF is nonlinear, as discussed in Section 7.4.

Finally, dummy variables are used to estimate a properly specified nonlinear CMF, as in (7.4). Only the estimated coefficients are displayed below, using the `coefficients()` function. Specifically, the number under `(Intercept)` is the estimated intercept (the conditional mean for the $X = 0$ base category), the number under `D1` is the estimated coefficient on `D1` (the dummy for $X = 1$), and the number under `D2` is the estimated coefficient on `D2` (the dummy for $X = 2$).

```
library(lmtest); library(sandwich)
set.seed(112358)
n <- 500  # sample size
m012 <- c(60,40,20)  # m(0),m(1),m(2) (linear CMF)
df <- data.frame(X=sample(x=0:2, size=n, prob=c(1,1,1)/3, replace=TRUE),
                 U=rnorm(n))
```

```
df$Y <- rnorm(n=n, mean=m012[1+df$X]) + df$U
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
CMF <- c(coef(ret)['X'], coefci(ret, vcov. = retVC1)['X',])
#
# Now: nonlinear CMF; LPC depends on X dist
set.seed(112358)
n <- 500;  m012 <- c(60,40,40)
df <- data.frame(X=sample(x=0:2, size=n, prob=3:1/6, replace=TRUE),
                 U=rnorm(n))
df$Y <- rnorm(n=n, mean=m012[1+df$X]) + df$U
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
LP1 <- c(coef(ret)['X'], coefci(ret, vcov. = retVC1)['X',])
#
set.seed(112358)
n <- 500;  m012 <- c(60,40,40)
df <- data.frame(X=sample(x=0:2, size=n, prob=1:3/6, replace=TRUE),
                 U=rnorm(n))
df$Y <- rnorm(n=n, mean=m012[1+df$X]) + df$U
ret <- lm(formula=Y~X, data=df)
retVC1 <- vcovHC(ret, type="HC1")
LP2 <- c(coef(ret)['X'], coefci(ret, vcov. = retVC1)['X',])
tmp <- rbind(CMF, LP1, LP2)
round(x=tmp, digits=3)
```

```
##         X  2.5 % 97.5 %
## CMF -19.8 -19.98 -19.67
## LP1 -12.3 -12.91 -11.69
## LP2  -7.7  -8.31  -7.09
```

```
#
# Use dummies to estimate nonlinear CMF
df$D0 <- (df$X==0)  # not used
df$D1 <- as.integer(df$X==1)  # D1=1 iff X=1
df$D2 <- as.integer(df$X==2)  # D2=1 iff X=1
ret <- lm(formula=Y~D1+D2, data=df)
coefficients(ret)
```

```
## (Intercept)          D1          D2
##        59.8       -19.8       -19.8
```

7.8 Simple Linear Regression

⟹ Kaplan video: OLS in R

The prior results are essentially the same when X has more than three possible values, too. Misspecification is likely. The linear projection, best linear approximation, and best linear predictor interpretations all still apply. OLS estimation and heteroskedasticity-robust confidence intervals are computed the same way. As in Section 6.3.3, the units of measure of β_1 are the units of Y divided by the units of X, if both Y and X are numeric.

The main difference is that using dummy variables to avoid misspecification is more difficult or impossible when X has many possible values. Chapter 8 addresses alternative ways to model a CMF that is not linear in X.

Although we won't go into detail (but see some commands in EE7.1), it is very helpful to visualize your data. Humans are good at seeing visual patterns and anomalies. Plot histograms (or boxplots) of each variable individually. Then make a scatter plot of (X_i, Y_i).

Practice 7.2 (linear fit)**.** For each scatterplot in Figure 7.3, guess what the OLS estimated regression line looks like, i.e., the line $\hat{\beta}_0 + \hat{\beta}_1 x$. (Hint: remember OLS minimizes the sum of the squares of the vertical distances from each point to the fit line.) You can also make your own puzzles in R: first make a scatterplot like

```
Y <- c(1,2,3,4,13); X <- c(1,2,3,4,5); plot(X,Y)
```

and then (after guessing) plot the OLS fit with `abline(lm(Y~X))`

Practice 7.3 (regression units)**.** Consider a regression of wage Y (\$/hr) on "distance to nearest university" X. Let γ_1 be the estimated slope when X is measured in miles, and let δ_1 be the estimated slope when X is measured in kilometers, where $1\,\text{mi} = 1.6\,\text{km}$.

a) What are the units of γ_1? δ_1?

b) Do you think $\gamma_1 = \delta_1$, $\gamma_1 > \delta_1$, or $\gamma_1 < \delta_1$?

c) Can you come up with a formula relating γ_1 and δ_1? (Hint: what change in Y is associated with a 1.6 km increase in X, in terms of γ_1? In terms of δ_1?)

Discussion Question 7.3 (student-teacher ratio simple regression)**.** Let Y be the average math standardized test score (in units of points) for a school's 5th-grade students. Let X be the 5th-grade student-teacher ratio (total number of 5th-grade students divided by total number of 5th-grade teachers; like the average class size), generally around $15 \leq X \leq 25$. For schools $i = 1, \ldots, n$, the values (Y_i, X_i) are recorded. A linear regression is run to estimate β_0 and β_1 in the CMF model $Y = \beta_0 + \beta_1 X + V$, $\mathrm{E}(V \mid X) = 0$. Respond to any three of the following (for example, parts a, c, and e; or b, c, f; or d, e, f; etc.).

a) What are the units of β_0 and β_1?

b) What's the interpretation of β_0? What is it useful for?

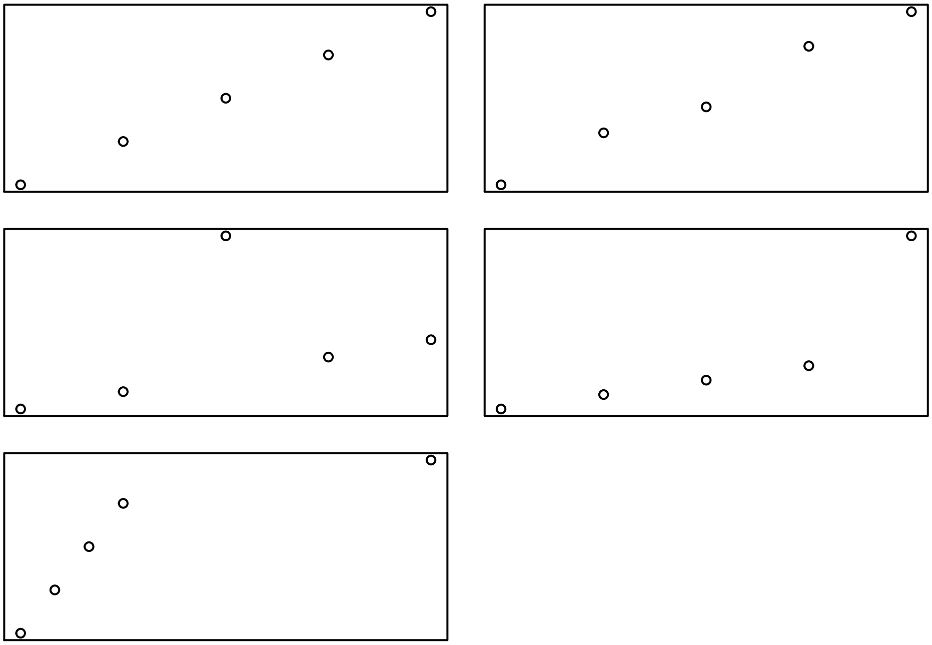


Figure 7.3: Scatterplots for Practice 7.2.

c) Consider the estimate $\hat{\beta}_1 = -2.28$. What does this imply about the average score difference between 15-student classes and 25-student classes? Is it economically significant (Section 3.8.3)? (Hint: make additional assumptions about the scoring system/scale if you need to.)

d) Consider further that $\hat{\beta}_1$ has heteroskedasticity-robust standard error 0.8, so a 95% CI is $[-3.88, -0.068]$. Describe our statistical uncertainty about $\hat{\beta}_1$.

e) Describe one reason you doubt $\hat{\beta}_1$ has a causal interpretation.

f) Describe one reason you think the linear CMF model is misspecified.

Optional Resources

Optional resources for this chapter

- Regression as description (Masten video)
- James et al. (2013, §3.1)
- Sections 4.1–4.2 ("Simple Linear Regression" and "Estimating the Coefficients of the Linear Regression Model") in Hanck et al. (2018)
- Sections 2.1 ("Simple OLS Regression") and 2.2 ("Coefficients, Fitted Values, and Residuals") in Heiss (2016) [repeated from Chapter 6]

Empirical Exercises

Empirical Exercise EE7.1. You will analyze data on colleges' athletic success and number of applications. The data were collected by Patrick Tulloch for an economics term project, from various college and sports data records. As the R description says, "The 'athletic success' variables are for the year prior to the enrollment and academic data."

a. Load the data (assuming you've already installed the R package or Stata command).

   R: `library(wooldridge)`

   Stata: `bcuse athlet1 , nodesc clear`

b. Keep only data from 1993.

   R: `dat <- athlet1[athlet1$year==1993 , ]`

   Stata: `keep if year==1993`

c. Create a new variable equal to the sum of `bowl` (football bowl game) and `finfour` (men's basketball Final Four).

   R: `dat$bowl4 <- dat$bowl + dat$finfour`

   Stata: `generate bowl4 = bowl + finfour`

d. Display the number of observations with each possible value of `bowl4` (0, 1, or 2).

   R: `table(dat$bowl4)`

   Stata: `tabulate bowl4`

e. Regress the number of applications (for admission) on the prior year's athletic success.

   R: `ret <- lm(apps~bowl4, data=dat)`

   Stata: `regress apps bowl4 , vce(robust)`

f. R only: save the fitted OLS values of $\hat{Y}$ for the three possible values of X (`bowl4`) with `fit012 <- predict(ret, newdata=data.frame(bowl4=0:2))` and optionally add helpful labels with `names(fit012) <- c('X=0','X=1','X=2')`

g. Estimate and store the three CMF values.

   R: `mean(dat$apps[dat$bowl4==0])` to estimate $m(0)$, and replace `0` with `1` to estimate $m(1)$ and with `2` to estimate $m(2)$; store these into a vector named `m012` with `m012 <- c( m0 , m1 , m2 )` where `m0` is your code for estimating $m(0)$ and similarly for `m1` and `m2`.

   Stata: `bysort bowl4 : egen CMF = mean(apps)` to compute the sample mean of `apps` within each group of observations with the same value of `bowl4`, storing it into a new variable named `CMF`

h. Display the numerical values of the OLS fit and the estimated CMF.

   R: `rbind(m012, fit012)`

   Stata: `collapse (mean) meanapps=apps , by(bowl4)` followed by `predict OLSfit , xb` and `list`

i. Plot the fitted OLS line against the estimated CMF points.

   R: `plot(x=0:2, y=m012)` (to plot estimated CMF points) followed by `abline(ret)` (to plot the OLS fit line)

   Stata: `twoway scatter CMF bowl4 || lfit apps bowl4`

j. Optional: make the same plot, but adjust the line color and style, the title, the axis labels, and whatever else you'd like to adjust.

   R: inside the `plot()` command, add argument `main='...'` to set the title and similarly for `xlab='...'` and `ylab='...'` to set the x-axis and y-axis labels (where you replace all the `...` with whatever names you want); inside the `abline()` function, add arguments `col=2` to change the line's color, `lty=2` to change the line style, and `lwd=3` to change the line width; again, you can set whatever values you like.

   Stata: `twoway scatter CMF bowl4 || lfit apps bowl4 , XXX` but replace the `XXX` with options to change the graph's appearance (all separated by spaces, not any more commas), like `title("...")` `xtitle("...")` `ytitle("...")` for the title and axis labels, and `lcolor(red) lpattern(dash)` for the line color and style; use whatever values you'd like.

Chapter 8

Nonlinear and Nonparametric Regression

⟹ Kaplan video: Chapter Introduction

Having mastered regression with a linear functional form, we now consider nonlinear functions. First nonlinear functions of X are allowed, and then nonparametric estimation and machine learning are introduced.

Unit learning objectives for this chapter

8.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

8.2. Interpret the coefficients in various nonlinear regression models [TLOs 3 and 5]

8.3. Judge which model seems most appropriate, using both economic reasoning and statistical insights [TLO 6]

8.4. In R (or Stata): estimate nonlinear and nonparametric regression models, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

8.1 Log Transformation

Sometimes a simple regression model improves greatly by transforming Y or X or both. The most common transformation in economics is the natural logarithm function, which economists just call "log."

Three different log models are discussed below. A model with the familiar form $Y = \beta_0 + \beta_1 X + U$ could be called a "linear-linear" model (although it's just called a **linear model**), meaning both Y and X are in their original units, i.e., **in levels**. If Y is replaced by its log, $\ln(Y)$, it's called a log-linear model; if instead we have Y and $\ln(X)$, then it's linear-log; and if both are **in logs**, then log-log.

Here in Section 8.1, the distinction among causal, CMF, and linear projection models is unimportant. The interpretation of U is left ambiguous intentionally. Instead, emphasis is on the interpretation of β_1 in terms of units of measure.

8.1.1 Properties of the Natural Log Function

Basic Shape and Properties

The natural **log function** is peculiar, especially if you haven't taken calculus. It is written $\ln(\cdot)$, although often people will simply say "log" (without "natural") and write $\log(\cdot)$, because the natural log is the only one commonly used in economics; in R, the function is `log()`.

The log function is the inverse of the exponential function: $\ln(\exp(x)) = x$, where $\exp(x)$ is the same as e^x. Consequently, if $e^x = M$, then $\ln(M) = \ln(e^x) = x$.

Figure 8.1 shows the log function, giving a general idea of its shape. However, two important features are unclear. First, as x gets closer and closer to 0, $\ln(x)$ decreases toward $-\infty$. Second, $\ln(x)$ keeps increasing to ∞ as x increases to ∞.

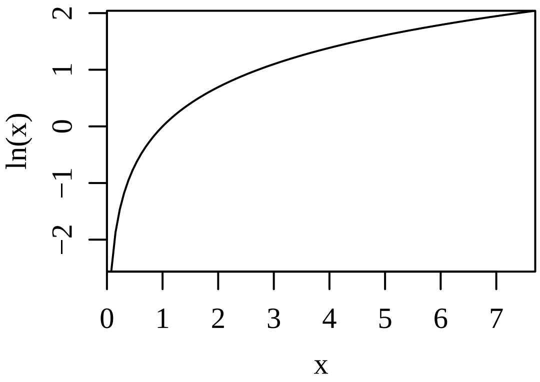


Figure 8.1: The (natural) log function, $\ln(\cdot)$.

The log function has many properties, including the following.

1. $\ln(x)$ is only defined for $x > 0$
2. $\ln(x)$ is strictly increasing: for any $x_2 > x_1 > 0$, $\ln(x_2) > \ln(x_1)$
3. $\ln(x)$ increases more slowly with larger x; it is very steep for x near zero, but less and less steep (i.e., flatter) as x increases
4. For any $x > 0$ and any b, $\ln(x^b) = b\ln(x)$
5. For any $x_1 > 0$ and $x_2 > 0$, $\ln(x_1/x_2) = \ln(x_1) - \ln(x_2)$ and $\ln(x_1 x_2) = \ln(x_1) + \ln(x_2)$
6. $\lim_{x\downarrow 0} \ln(x) = -\infty$ and $\lim_{x\to\infty} \ln(x) = \infty$

Percentage Approximation

Near $w = 1$, $\ln(w)$ is approximately the same as the linear function $f(w) = w - 1$: $\ln(w) \approx w - 1$. At $w = 1$ exactly, $\ln(1) = 1 - 1 = 0$ exactly. However, the approximation is worse when w is farther from one.

Example 8.1.

- For $w = 1.01$, $\ln(1.01) = 0.00995$, very close to $w - 1 = 0.01$.
- For $w = 0.99$, $\ln(0.99) = -0.01005$, very close to $w - 1 = -0.01$.
- For $w = 1.1$, $\ln(1.1) = 0.0953$, close to $w - 1 = 0.1$.
- For $w = 1.5$, $\ln(1.5) = 0.405$, not close to $w - 1 = 0.5$.

A difference of p in log units is approximately a $100p\%$ difference. That is, for values y_2 and y_1, if $p = \ln(y_2) - \ln(y_1)$, then the difference between y_2 and y_1 is approximately $100p\%$, meaning $y_2/y_1 \approx 1 + p$. Mathematically, assuming y_2 is near y_1 (so y_2/y_1 is near 1), the reason is

$$\ln(y_2) - \ln(y_1) = \ln(y_2/y_1) \approx (y_2/y_1) - 1 = (y_2 - y_1)/y_1. \tag{8.1}$$

Again, this approximation is exact with $p = 0$ but worse for p farther from zero.

Example 8.2. Let $y_1 = 100$ and $y_2 = 105.2$, a percentage difference of 5.2%: $(y_2 - y_1)/y_1 = 5.2/100 = 0.052$. The log difference is $\ln(105.2) - \ln(100) = 4.656 - 4.605 = 0.051$, a good approximation of the true 0.052. The approximation becomes poor if instead $y_2 = 152$: $\ln(152) - \ln(100) = 0.42$, not close to the true percentage difference of $(y_2 - y_1)/y_1 = 0.52$ (52%).

Recall the difference between a percentage change and a percentage point change. "Percentage point" only applies when the units are already percentages. For example, a 1 percentage point increase is changing from 10% to 11%, or from 67% to 68%. A percentage change can apply to any numeric variable and equals $100[(y_2 - y_1)/y_1]\%$.

Example 8.3. Let W be hourly wage (\$/hr). Consider a change from $w_1 = \$12.50$/hr to $w_2 = \$13.75$/hr. Then, $(w_2 - w_1)/w_1 = (13.75 - 12.50)/12.50 = 0.10$, meaning a $100[(w_2 - w_1)/w_1]\% = 100[0.10]\% = 10\%$ increase.

Example 8.4. Let R be the one-year recidivism rate for individuals convicted of a felony whose sentence does not include prison time. If the initial rate $r_1 = 0.08$ (meaning 8%) changes to $r_2 = 0.06$ (6%), then we could say that the recidivism rate decreased by two percentage points $(8 - 6 = 2)$, or we could say that the rate decreased by 25%, because $(r_2 - r_1)/r_1 = (0.08 - 0.06)/0.08 = -0.25$. Both statements are mathematically true, although one may sound to you (and others) like a bigger decrease.

8.1.2 The Log-Linear Model

Interpretation

A **log-linear model** specifies

$$\ln(Y) = \beta_0 + \beta_1 X + U. \tag{8.2}$$

Because X is in levels, the coefficient β_1 tells us about a one unit increase in X. Specifically, a one unit increase in X is associated with a β_1 change in $\ln(Y)$ (increase if $\beta_1 > 0$, decrease if $\beta_1 < 0$). Sometimes, people call this a β_1 change in Y in **log units**.

Instead of log units, we can interpret β_1 in terms of a percentage change in Y. Specifically, a one-unit increase in X is associated with a change from Y to rY, which is a $100(r-1)\%$ change in Y. Using some properties of log and rearranging, $\beta_1 = \ln(rY) - \ln(Y) = \ln(rY/Y) = \ln(r)$, so $r = e^{\beta_1}$, meaning a $100(e^{\beta_1} - 1)\%$ change in Y.

If β_1 is close to zero, then (8.1) offers a simpler but approximate interpretation: a one-unit increase in X is associated with an approximate $100\beta_1\%$ change in Y. Mathematically, if $r \approx 1$, then $\ln(r) \approx r - 1$, so (from above) $\beta_1 \approx r - 1$; thus, the $100(r-1)\%$ change is approximately a $100\beta_1\%$ change. However, as usual, the approximation may be poor for large β_1, or when considering changes in X larger than one unit.

Example 8.5 (Kaplan video)**.** Let $\beta_1 = 0.02$. A one-unit increase in X is associated with a 0.02 increase in $\ln(Y)$, meaning a 0.02 "log point" increase in Y. Alternatively, a one-unit increase in X is associated with a $100(e^{\beta_1} - 1)\% = 2.02\%$ increase in Y. This is approximately a $100\beta_1\% = 2\%$ increase in Y.

Example 8.6. Again let $\beta_1 = 0.02$. Now consider a 50-unit increase in X. This is associated with a $50\beta_1 = 1$ log point increase in Y, or a $100(e^{50\beta_1} - 1)\% = 172\%$ increase in Y. However, this increase is poorly approximated by $100(50\beta_1)\% = 100\%$.

When to Use It

When does a log-linear model make sense? Sometimes, scatterplots of the raw Y and X data suggest it. For example, maybe the relationship between Y and X looks nonlinear, but the relationship between $\ln(Y)$ and X looks approximately linear.

Sometimes the log-linear model makes more sense economically or intuitively. For example, with Y variables like income, it may seem more natural to model effects as (approximate) percentage changes in Y, like a 1% higher income instead of a \$500/yr higher income. Further, the log-linear form derives from economic models of human capital where there is a multiplicative effect on wage. The most famous of these is the "Mincer equation" for earnings as a function of education (schooling) and experience, named after the log-linear model in Mincer (1974, Ch. 5, p. 84).

Issue with Prediction

Unfortunately, the log-linear model is not optimal for predicting Y, even if $\mathrm{E}(U \mid X) = 0$. From (8.2), the CMF is

$$\mathrm{E}(Y \mid X = x) = e^{\beta_0 + \beta_1 x}\,\mathrm{E}(e^U \mid X = x).$$

It is easy to plug in $\hat{\beta}_0$ and $\hat{\beta}_1$, but difficult to estimate $\mathrm{E}(e^U \mid X = x)$. We could simply ignore the difficult term, but $e^{\beta_0 + \beta_1 x}$ is generally not the best predictor of Y given $X = x$. There are alternatives, but they are beyond our scope.

8.1.3 The Linear-Log Model

Interpretation

A **linear-log model** specifies

$$Y = \beta_0 + \beta_1 \ln(X) + U. \tag{8.3}$$

When X increases by one log unit, the corresponding change in Y is β_1; but one log unit is a very big change (more than doubling). To use the percentage approximation, a smaller change in X must be used. Specifically, an increase of X by 1% is associated with a change in Y of approximately $\beta_1/100$ units.

For larger changes in X, use the exact change in Y. Consider X increases by $100p\%$, from x to $(1+p)x$. The corresponding log difference is $\ln((1+p)x) - \ln(x) = \ln((1+p)x/x) = \ln(1+p)$, so Y changes by $\beta_1 \ln(1+p)$ units:

$$[\beta_0 + \beta_1 \ln((1+p)x)] - [\beta_0 + \beta_1 \ln(x)] = \beta_1[\ln((1+p)x) - \ln(x)] = \beta_1 \ln(1+p).$$

Example 8.7 (Kaplan video)**.** Consider a change from $X = 40$ to $X = 60$. This is a 50% increase ($p = 0.50$), so the corresponding change in Y is $\beta_1 \ln(1.5) = 0.41\beta_1$. More directly,

$$[\beta_0 + \beta_1 \ln(60)] - [\beta_0 + \beta_1 \ln(40)] = \beta_1[\ln(60) - \ln(40)] = \beta_1 \ln(60/40) = 0.41\beta_1.$$

The approximation $p\beta_1 = 0.5\beta_1$ is not good. If instead $p = 0.01$ for a change from $X = 40$ to $X = 40.4$, then the exact change is $\beta_1 \ln(40.4/40) = 0.00995\beta_1$, very close to the approximation $p\beta_1 = 0.01\beta_1$. Note the accuracy of the approximation does not depend on β_1.

Example 8.8. Let $\beta_1 = 23$ and consider a 1% increase in X. This is associated with an approximate $\beta_1/100 = 0.23$-unit change in Y, and this is a very good approximation (the exact change is 0.229 units).

When to Use It

When does a linear-log model make sense? Sometimes, the scatterplot of Y and X reveals a shape that looks like a log function: increasing steeply at first, then getting less and less steep, but without ever decreasing. (Or: switch "increasing" and "decreasing," if $\beta_1 < 0$.) That is, the relationship between Y and X looks nonlinear, but maybe plotting Y against $\ln(X)$ looks closer to linear. The log function's shape also helps model diminishing marginal benefits: the first unit of X helps increase Y a lot, but each additional unit of X helps less and less.

8.1.4 The Log-Log Model

Interpretation

A **log-log model** specifies

$$\ln(Y) = \beta_0 + \beta_1 \ln(X) + U. \tag{8.4}$$

A 1% increase in X is associated with an approximate β_1% change in Y. This percentage interpretation is particularly nice: β_1 represents an elasticity of Y with respect to X. But, if the percentages are too large, then the approximation is poor. To be exact, a p% increase in X is associated with a $100[(1+p)^{\beta_1} - 1]$% change in Y, which is approximately $p\beta_1$% for small $p\beta_1$.

Example 8.9 (Kaplan video)**.** Let $\beta_1 = 3.2$. A 1% increase in X ($p = 0.01$) is associated with a $100[(1+p)^{\beta_1} - 1]\% = 3.24\%$ increase in Y, which indeed is approximately $100p\beta_1\% = 3.2\%$. A 20% increase in X ($p = 0.20$) is associated with a $100[(1+0.2)^{3.2} - 1]\% = 79.2\%$ increase in Y, which is not well approximated by $100p\beta_1\% = 64\%$.

When to Use It

When does a log-log model make sense? First, it's a simple way to get an elasticity interpretation. Second, a scatterplot of $\ln(Y)$ against $\ln(X)$ may look roughly linear. Third, if you suspect a power law type of relationship between Y and X, exponentiating both sides of (8.4) yields

$$\exp\{\ln(Y)\} = \exp\{\beta_0 + \beta_1 \ln(X) + U\}, \implies Y = e^{\beta_0} \exp\{\ln(X^{\beta_1})\} e^U = e^{\beta_0} X^{\beta_1} e^U.$$

Issue with Prediction

As with the log-linear model, $e^{\beta_0} X^{\beta_1}$ is generally not the CMF because $\mathrm{E}(e^U \mid X) = 1$ is not implied by $\mathrm{E}(U \mid X) = 0$. Consequently, predicting Y as $e^{\hat\beta_0} X^{\hat\beta_1}$ is generally not optimal.

In Sum: Regression Models with Log Transformations

Log-linear: 1-unit ↑ X associated with $100(e^{\beta_1}-1)\%$ change in Y, approximately $100\beta_1\%$ if β_1 near zero; d-unit ↑ X associated with $100(e^{d\beta_1}-1)\%$ change in Y, approximately $100d\beta_1\%$ for small $d\beta_1$
Linear-log: 1% ↑ X associated with $\beta_1 \ln(1.01)$-unit change in Y, approximately $\beta_1/100$-unit change; $100p\%$ ↑ X associated with $\beta_1 \ln(1+p)$-unit change in Y, approximately $p\beta_1$-unit change for small p
Log-log: 1% ↑ X associated with $100(1.01^{\beta_1}-1)\%$ change in Y, approximately $\beta_1\%$ change in Y (elasticity); $100p\%$ ↑ X associated with $100((1+p)^{\beta_1}-1)\%$ change in Y, approximately $100p\beta_1\%$ for small $p\beta_1$

Discussion Question 8.1 (pollution and house price)**.** Consider the relationship between the price of a house and the concentration of air pollution. Explain which type of model (linear, log-linear, linear-log, or log-log) you think would best fit, and why. (Hint: think especially about changes in levels vs. in logs.)

8.1.5 Warning: Model-Driven Results

⟹ Kaplan video: Warnings About Model-Driven Results

When choosing a model, beware self-fulfilling prophecy. Empirical results are driven by data, but also by your model's structure. For example, the function $\beta_0+\beta_1 X$ specifies a constant (β_1) change for every unit increase in X; different datasets can lead to different estimated slopes ($\hat{\beta}_1$), but the slope will always be constant, regardless of the data. The log-linear model may seem more flexible than a linear model, but it is not: it still only has two parameters. It is just different, not more flexible. Consequently, the fitted log-linear model always shows a diminishing effect of X on Y as X increases. This pattern does not come from the data, but from the model itself, regardless of the data.

Figure 8.2, based on the comic at `https://xkcd.com/2048`, illustrates such self-fulfilling prophecy. Each graph shows the same scatterplot from the same data (the dots), but with a very different fitted model in each (the line). Clearly, the differences do not come from the data because it's the exact same data. All differences are entirely due to the model. The top-left shows the linear model, which by construction imposes a constant slope β_1. Below that is a log-linear model; the constant percentage increase of Y with each unit of X leads to exponential growth (hence the "exponential" label in the comic). The top-right shows the "tapering off" of the linear-log model. Although mostly beyond our scope, some comments on "model selection" are in Sections 8.3 and 15.2.

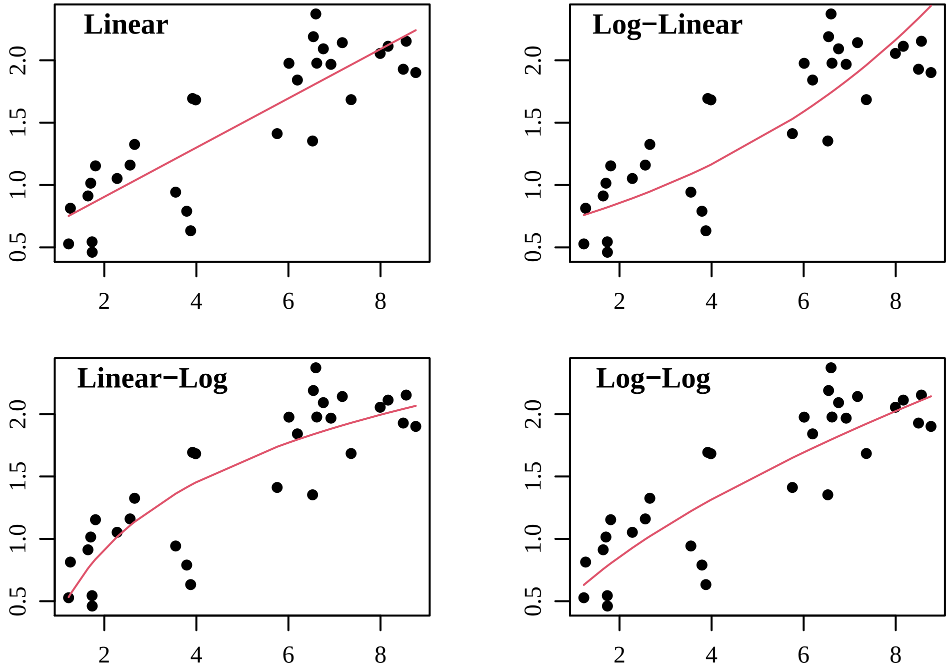


Figure 8.2: Same data, different models.

8.1.6 Code

Figure 8.2 is generated by the following code that compares linear, log-linear, linear-log, and log-log estimation given the same dataset. The four fitted functions are plotted on four copies of the same scatterplot in Figure 8.2, in homage to https://xkcd.com/2048. The results illustrate the concerns of Section 8.1.5.

```
par(family='serif', mar=c(3,3,1,1), mgp=c(2.1,0.8,0), mfrow=c(2,2))
set.seed(112358)
n <- 31
X <- sort(runif(n=n, min=1, max=9))
Y <- 1 + pnorm(q=X, mean=5, sd=1.5) +
     2*( rbeta(n=n, shape1=10-X, shape2=X) - (10-X)/10 )
df <- data.frame(X=X, Y=Y)
ret.linlin <- lm(Y~X, data=df)
ret.loglin <- lm(log(Y)~X, data=df)
ret.linlog <- lm(Y~log(X), data=df)
ret.loglog <- lm(log(Y)~log(X), data=df)
#
XL <- '';  YL <- ''
plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
```

```
lines(predict(ret.linlin)~df$X, col=2)
title("Linear", line=-1, adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(predict(ret.linlog)~df$X, col=2)
title("Linear-Log", line=-1, adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(exp(predict(ret.loglin))~df$X, col=2)
title("Log-Linear", line=-1, adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, main='', xlab=XL, ylab=YL)
lines(exp(predict(ret.loglog))~df$X, col=2)
title("Log-Log", line=-1, adj=0.1)
```

8.2 Nonlinear-in-Variables Regression

Discussion Question 8.2 (nonlinear OVB)**.** Imagine a structural model $Y = \beta_0 + \beta_1 X + \beta_2 X^2$, with no error term: X completely determines Y. To be more concrete, imagine $Y = 1 + X^2$ (i.e., $\beta_0 = 1$, $\beta_1 = 0$, $\beta_2 = 1$), with $0 \leq X \leq 5$. You run a linear-in-variables regression; OLS estimates the function $\hat{\gamma}_0 + \hat{\gamma}_1 X$.

a) Approximately what value would you expect $\hat{\gamma}_1$ to be? (Hint: recall Sections 7.3–7.5.)
b) What does $\hat{\gamma}_0 + \hat{\gamma}_1 X$ suggest about the relationship between X and Y? What features are similar or different compared to the true $1 + X^2$? (Hint: draw a picture.)

Beyond replacing X with a single transformation of X like $\ln(X)$, we can replace X with a more complicated nonlinear function involving multiple terms and multiple parameters. OLS can still be used for estimation as long as the function is "linear-in-parameters" (Section 8.2.1). Again, the distinctions among causal, CMF, and linear projection models are not emphasized here.

There are two types of (non)linearity. They are often confused. Further, people often say "linear model" or "nonlinear model" without clarifying which type they mean.

8.2.1 Linearity

The root of "linearity" is **linear combination**. A linear combination is like a weighted sum. For example, a linear combination of A and B is anything with the form

$$w_1 A + w_2 B, \tag{8.5}$$

where w_1 and w_2 are weights that may take any value, including zero or even negative numbers. Linear combinations may involve more than two terms, like $w_1A+w_2B+w_3C+w_4D$. In some cases, instead of A, B, C, and D, we have something like Y_1, Y_2, Y_3, and Y_4, in which case the linear combination may be written in summation notation:

$$w_1Y_1 + w_2Y_2 + w_3Y_3 + w_4Y_4 = \sum_{i=1}^{4} w_iY_i. \tag{8.6}$$

Example 8.10. The expected value formula $\mathrm{P}(Y = y_1)y_1 + \mathrm{P}(Y = y_2)y_2 + \cdots$ for discrete random variables in (2.4) is a linear combination of the possible values y_j, where the linear combination weights are the probabilities, $w_j = \mathrm{P}(Y = y_j)$.

Example 8.11. The sample mean is a linear combination of observed Y_i values, with weights $w_i = 1/n$:

$$\bar{Y}_n = \sum_{i=1}^{n} w_iY_i = \sum_{i=1}^{n}(1/n)Y_i = (1/n)\sum_{i=1}^{n} Y_i.$$

A function is **linear-in-parameters** if it is a linear combination of the parameters.

Example 8.12 (Kaplan video)**.** The function $\beta_0 + \beta_1 x$ is linear-in-parameters because it is a linear combination of the parameters β_0 and β_1 with weights $w_1 = 1$ and $w_2 = x$:

$$w_1\beta_0 + w_2\beta_1 = (1)(\beta_0) + (x)(\beta_1) = \beta_0 + \beta_1 x.$$

A function is **linear-in-variables** if it is a linear combination of the regressors. However, here the intercept is interpreted as the "coefficient" on a secret regressor that's a always equals one, the constant $X_0 = 1$.

Example 8.13 (Kaplan video)**.** The function $\beta_0+\beta_1 x$ is linear-in-variables. Using $x_0 \equiv 1$, the linear combination of x_0 and x has weights $w_1 = \beta_0$ and $w_2 = \beta_1$:

$$(w_1)(x_0) + (w_2)(x) = (\beta_0)(1) + (\beta_1)(x) = \beta_0 + \beta_1 x.$$

For this reason, in economics, people often call $\beta_0 + \beta_1 x$ "linear in x" even though technically it is "affine in x" and "linear in x_0 and x."

These two types of linearity can apply to CMFs or linear projections. For example, if the CMF is $\mathrm{E}(Y \mid X = x) = \beta_0 + \beta_1 x$, then the CMF is linear-in-parameters and linear-in-variables. Regardless of the CMF, the linear projection of Y onto $(1, X)$ is $\mathrm{LP}(Y \mid 1, X) = \beta_0 + \beta_1 X$, which is always linear-in-parameters and linear-in-variables by definition.

Confusingly, people often refer to models themselves as linear. For example, $Y = \beta_0 + \beta_1 X + U$ is often called a "linear model" even though $\beta_0 + \beta_1 X + U$ is neither a linear combination of (β_0, β_1) nor of $(1, X)$, due to the $+U$.

8.2.2 Nonlinearity

To increase flexibility, economists often use functions that are nonlinear-in-variables but still linear-in-parameters. That is, they can be written as a linear combination of the parameters β_j, but may include nonlinear functions of x in the linear combination weights:

$$\sum_{j=0}^{J} \beta_j f_j(x). \tag{8.7}$$

Example 8.14 (Kaplan video)**.** A **quadratic function** is a special case of (8.7) with $J = 2$, $f_0(x) = 1$, $f_1(x) = x$, and $f_2(x) = x^2$, yielding $\beta_0 + \beta_1 x + \beta_2 x^2$. This is nonlinear-in-variables because it cannot be written as a linear combination of $(1, x)$. This is linear-in-parameters because it is a linear combination of $(\beta_0, \beta_1, \beta_2)$, with weights $(1, x, x^2)$.

Example 8.15. The function $\beta_0+\beta_1 x+\beta_2 x^2+\beta_3 x^3+\beta_4 x^4$ is (8.7) with $J = 4$, $f_j(x) = x^j$. The function $\beta_0 + \beta_1 \sin(x) + \beta_2 \cos(x)$ instead has $J = 2$, $f_0(x) = 1$, $f_1(x) = \sin(x)$, and $f_2(x) = \cos(x)$. The function $\beta_0 + \beta_1 \ln(x) + \beta_2 \sqrt{x} + \beta_3 x^{1/3}$ has $J = 3$, $f_0(x) = 1$, $f_1(x) = \ln(x)$, $f_2(x) = \sqrt{x}$, and $f_3(x) = x^{1/3}$. All of these are nonlinear-in-variables and linear-in-parameters.

A **nonlinear-in-parameters** model cannot be written as a linear combination of the parameters. These are also used in economics, but less commonly; they are not discussed further here.

Example 8.16 (Kaplan video)**.** In the power law model

$$Y = \beta_0 X^{\beta_1} + U, \tag{8.8}$$

the term $\beta_0 X^{\beta_1}$ cannot be written as a linear combination of β_0 and β_1.

8.2.3 Estimation and Inference

OLS can estimate nonlinear-in-variables models as long as they are linear-in-parameters. As always, the OLS estimates are the parameter values that minimize the sum of squared residuals, solving the empirical analog of the optimal prediction problem (minimizing mean quadratic loss).

Inference on parameters is also the same. For example, the same R code to compute a confidence interval for β_1 earlier still works, and a confidence interval for β_2 can be computed the same way. The underlying code/math is very similar, too, although confidence intervals for predicted values now involve multiple coefficients.

8.2.4 Parameter Interpretation

Unlike estimation and inference, which remain similar, interpretation of parameters changes greatly with nonlinear-in-variables models.

Insufficiency of Linear Coefficient

With a nonlinear-in-variables function $\beta_0+\beta_1 x+\cdots$, we cannot learn anything by looking at the coefficient β_1 alone. Not even its sign (+ or −) has meaning.

Example 8.17 (Kaplan video)**.** Consider the quadratic function $\beta_0+\beta_1 x+\beta_2 x^2$. Specifically, let $\beta_0=0$, $\beta_1=5$, and $\beta_2=-1$, so the function is $5x-x^2$. Going from $x=0$ to $x=1$, the change is

$$[(5)(1)-1^2]-[(5)(0)-0^2]=4-0=4.$$

From $x=1$ to $x=2$, the change is

$$[(5)(2)-2^2]-[(5)(1)-1^2]=6-4=2,$$

still positive, but smaller. From $x=2$ to $x=3$,

$$[(5)(3)-3^2]-[(5)(2)-2^2]=6-6=0,$$

no change at all. And from $x=3$ to $x=4$,

$$[(5)(4)-4^2]-[(5)(3)-3^2]=4-6=-2,$$

a negative change (decrease). Even though $\beta_1=5$ is positive, sometimes the function decreases as x increases.

Summarizing Nonlinear Functions

There are two general approaches to summarizing an estimated function $\hat{f}(\cdot)$ from the model $Y=f(X)+U$.

First, with only one X, the best summary is to plot the function (along with a scatterplot of data), like in Figure 8.2. As the saying goes, "A picture is worth a thousand words [or numbers]." However, with many different regressors (as in later chapters), pictures get confusing (trying to show slices of many-dimensional manifolds...).

Second, we can plug in changes of X that are relevant to policy or a particular economic question. For a change from $X=x_1$ to $X=x_2$, the (estimated) associated change in Y is

$$\hat{f}(x_2)-\hat{f}(x_1).$$

There are variations of this approach, like letting x_1 be the average X value, or averaging these changes over many (x_1,x_2) pairs, etc.

Example 8.18 (Kaplan video)**.** Let Y be income and X education. We can compute $\hat{f}(x)$ for $x=8,9,10,\ldots,21$ and plot the estimated function in a graph. If we specifically want to understand the value of the 12th year of education, then we can compute $\hat{f}(12)-\hat{f}(11)$. With a quadratic model, $\hat{f}(x)=\hat{\beta}_0+\hat{\beta}_1 x+\hat{\beta}_2 x^2$, the estimated $\hat{f}(12)-\hat{f}(11)$ is

$$\hat{\beta}_0+\hat{\beta}_1(12)+\hat{\beta}_2(12)^2-[\hat{\beta}_0+\hat{\beta}_1(11)+\hat{\beta}_2(11)^2]=\hat{\beta}_1(12-11)+\hat{\beta}_2(12^2-11^2)=\hat{\beta}_1+23\hat{\beta}_2.$$

In Sum: Interpreting and Summarizing Nonlinear Models

The $\beta_1 X$ term alone has no meaning.
Given nonlinear model $Y = f(X) + U$, a change from $X = x_1$ to $X = x_2$ is associated with a change in Y of $f(x_2) - f(x_1)$, estimated by $\hat{f}(x_2) - \hat{f}(x_1)$.

Practice 8.1 (quadratic example)**.** You regress Y on X and X^2 and get the fitted function $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ with $\hat{\beta}_0 = 2$, $\hat{\beta}_1 = 4$, and $\hat{\beta}_2 = -2$.

a) What's the predicted value of Y when $X = 0$? $X = 1$? $X = 2$?
b) What's the predicted change in Y when X changes from 0 to 1? from 1 to 2?

Discussion Question 8.3 (nonlinear wage model interpretation)**.** Let Y be wage ($/hr) and X years of education. Given a sample of data, you estimate $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ with $\hat{\beta}_0 = 14.4$, $\hat{\beta}_1 = -1.6$, and $\hat{\beta}_2 = 0.1$.

a) Does $\hat{\beta}_1 < 0$ mean that more education is associated with lower wage? Why/not?
b) What does this estimated function suggest about the (descriptive) relationship between wage and education? (Hint: try plugging in salient values like $X = 12$ [high school] or $X = 16$ [college], or graph the whole function.)

8.2.5 Description, Prediction, and Causality

The interpretation of a nonlinear-in-variables model as causal, CMF, or linear projection is similar to linear-in-variables models. The main difference is that we may wish to clarify the word "linear" in linear projection, best linear approximation, and best linear predictor.

Description and Prediction

Consider a quadratic model when the true CMF is not quadratic. Then, the "linear" projection of Y onto $X_0 = 1$, X, and X^2 is defined the same way as in (7.6) before:

$$\begin{aligned}\mathrm{LP}(Y \mid 1, X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2 &= \operatorname*{arg\,min}_{a,b,c} d(Y, a + bX + cX^2) \\ &= \operatorname*{arg\,min}_{a,b,c} \sqrt{\mathrm{E}[(Y - a - bX - cX^2)^2]}. \qquad (8.9)\end{aligned}$$

These linear projection coefficients are what OLS estimates. This same function of X is again a "best" CMF approximation and "best" predictor of Y. Specifically, mirroring (7.12) and (7.13),

$$\begin{aligned}\mathrm{LP}(Y \mid 1, X, X^2) = \beta_0 + \beta_1 X + \beta_2 X^2 &= \overbrace{\operatorname*{arg\,min}_{a,b,c} \mathrm{E}\{[\mathrm{E}(Y \mid X) - (a, b, c)]^2\}}^{\text{BLA}} \\ &= \overbrace{\operatorname*{arg\,min}_{a,b,c} \mathrm{E}\{[Y - (a + bX + cX^2)]^2\}}^{\text{BLP}}. \qquad (8.10)\end{aligned}$$

As before, if the true CMF actually is quadratic, then these all equal the true CMF.

Structural Identification

When the structural model $Y = f(X)+U$ satisfies $\mathrm{E}(U \mid X) = 0$, changes in $f(X)$ (which is also the CMF) can be interpreted as average structural effects (ASEs) of X on Y. The ASE of changing $X = x_1$ to $X = x_2$ is then $f(x_2) - f(x_1)$. If we somehow correctly guess the functional form of $f(\cdot)$, then OLS can estimate it, and then $\hat{f}(x_2) - \hat{f}(x_1)$ is the estimated ASE.

Example 8.19 (Kaplan video)**.** Consider the quadratic structural model $Y = m(X) + U = \beta_0 + \beta_1 X + \beta_2 X^2 + U$, where the structural error U happens to satisfy the CMF error property $\mathrm{E}(U \mid X) = 0$. Then, running OLS can estimate the coefficients, and the estimated ASE of changing from x_1 to x_2 is

$$\hat{m}(x_2) - \hat{m}(x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_2 + \hat{\beta}_2 x_2^2 - (\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_1^2) = \hat{\beta}_1(x_2 - x_1) + \hat{\beta}_2(x_2^2 - x_1^2).$$

8.2.6 Code

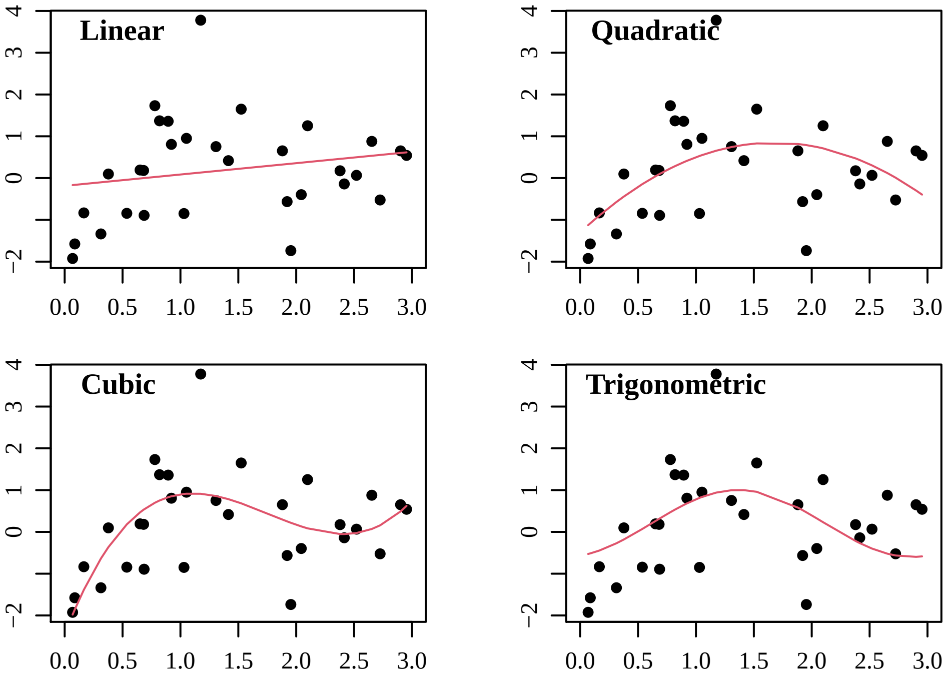


Figure 8.3: Same data, different models.

Figure 8.3 is generated by the following code that fits the same data with four models: linear, quadratic, and cubic polynomials, and a trigonometric model with a sine and cosine term. Figure 8.3 shows four identical scatterplots with the four different fitted models.

Note how the four fitted lines have very different qualitative features, even though they use the same data. This illustrates the same concerns about model-driven results and "self-fulfilling prophecy" as in Section 8.1.5 and Figure 8.2.

```
par(family='serif', mar=c(3,3,1,1), mgp=c(2.1, 0.8, 0), mfrow=c(2,2))
set.seed(112358)
n <- 31
X <- sort(3*rbeta(n=n,shape1=1,shape2=1))
df <- data.frame(X=X, Y=1+10*(X/2-0.5)^2*(X/2-0.5-1) + rnorm(n=n))
ret.poly1 <- lm(Y~X, data=df)
ret.poly2 <- lm(Y~X+I(X^2), data=df)
ret.poly3 <- lm(Y~X+I(X^2)+I(X^3), data=df)
ret.trig <- lm(Y~I(cos(2*pi*(X-0)/3))+I(sin(2*pi*(X-0)/3)), data=df)
#
XL <- '';  YL <- ''
plot(x=df$X, y=df$Y, type='p', pch=16, xlab=XL, ylab=YL,
     main='', xlim=c(0,3))
lines(predict(ret.poly1)~df$X, col=2)
title("Linear",line=-1,adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, xlab=XL, ylab=YL,
     main='', xlim=c(0,3))
lines(predict(ret.poly2)~df$X, col=2)
title("Quadratic",line=-1,adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, xlab=XL, ylab=YL,
     main='', xlim=c(0,3))
lines(predict(ret.poly3)~df$X, col=2)
title("Cubic",line=-1,adj=0.1)
#
plot(x=df$X, y=df$Y, type='p', pch=16, xlab=XL, ylab=YL,
     main='', xlim=c(0,3))
lines(predict(ret.trig)~df$X, col=2,)
title("Trigonometric",line=-1,adj=0.1)
```

8.3 Nonparametric Regression

⟹ Kaplan video: Model Flexibility in Nonparametric Regression

In **nonparametric regression**, the functional form of the CMF $m(\cdot)$ is unknown. This is more general than nonlinear-in-variables regression, where $m(\cdot)$ is nonlinear but has a known functional form, like a cubic polynomial or log-linear model, in which only the coefficient values are unknown.

In principle, this allows a very **flexible** model for $m(\cdot)$, although in practice the (hopefully) optimal level of flexibility must be chosen somehow. There is no universal quantitative definition of "flexible," but the qualitative meaning is the same as the physical flexibility of a hose or cable: can it bend around sharply in many places to take whatever shape you wish (flexible), or can it only take on particular shapes? The number of parameters (terms) in a model is a general guide to how flexible the model is. For example, a model with 20 paramters is more flexible than a model with only 2 parameters.

Example 8.20. The CMF $m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ is more flexible than $m(x) = \beta_0 + \beta_1 x$ because it allows both straight and curved lines, whereas the latter allows only straight lines. Also, the CMFs are essentially the same except that the less flexible one implicitly sets $\beta_2 = 0$, whereas the first CMF more flexibly allows non-zero β_2.

Example 8.21. Consider the CMFs $m(x) = \ln(x)$ and $m(x) = \beta_0 + \beta_1 x$. The log function is curved, whereas the "linear" function is straight. However, here the linear function is more flexible. In fact, the log function is not flexible at all: there is only one possibility because there are no parameters! In contrast, the linear function has two parameters; β_0 allows the intercept to move up and down, while β_1 allows the slope to change. In general, the number of parameters is a better gauge of "flexibility" than curviness.

Many **machine learning** methods are nonparametric CMF estimators. In machine learning, often prediction is emphasized over description and causality, but recall that the CMF is the best predictor of Y given X (under quadratic loss).

8.3.1 Model Selection

One view of nonparametric regression is that it is like nonlinear regression, but choosing the model with a formal statistical procedure instead of guessing. Even if we know our chosen model will be wrong, we may hope to find the least-wrong model, and hope that it is a good enough approximation to be useful. As Box (1979, p. 2) famously wrote, "All models are wrong but some are useful."[1] The steps are basically:

1. Choose a group of possible regression models.
2. Choose a way to evaluate models.
3. Evaluate the quality of each model, given the data.
4. Select the best (least bad) model.
5. Use the estimates from the selected model.

Steps 1–4 describe **model selection**, i.e., choosing which model to use for estimation. This is unavoidable. Sometimes model selection is informal; e.g., somebody just

[1] See https://en.wikipedia.org/wiki/All_models_are_wrong for additional discussion.

feels like using a quadratic model today. With nonparametric regression, usually Step 1 is done informally (but thoughtfully). For Step 2, there are many formal, statistical evaluation procedures to choose from; this choice (of procedure) is also done informally but thoughtfully. Steps 3 and 4 are done by the chosen statistical procedure using the data.

In R, usually Steps 1 and 2 require you to pick a particular R function (and certain arguments), and then the function computes Steps 3 and 4 (and Step 5) for you. Depending on the chosen model, Step 5 may be identical to Section 8.2.

Some intuitive ways to evaluate models are really bad. First, maximizing R^2 is bad. Whenever you add a term to your model, R^2 always increases, even if the model is worse (i.e., yields worse CMF estimates and predictions). Adjusted R^2 is better but still not designed for optimal model selection. Second, hypothesis testing is bad. Different significance levels yield different chosen models, and the answer to "which model is best?" never starts with "I controlled the type I error rate. . . ."

Example 8.22. You estimate $m(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ and test $H_0\colon \beta_2 = 0$ to see if a linear model would be better; the p-value is $p = 0.08$. You say this means you accept $\beta_2 = 0$ because $p > 0.05$, so a linear model is better. Your sister argues you should actually reject linearity because $p < 0.1$. Your grandmother says you should both quiet down and never use hypothesis testing for model selection anyway.

The first difficulty in selecting a good CMF model is that $m(\cdot)$ could be very nonlinear. Imagine $Y = m(X)$ exactly. Even without any error term, we could get a bad estimate if we specify $m(x) = \beta_0 + \beta_1 x$ when really $m(\cdot)$ is not linear-in-variables. So, our model must be flexible enough to approximate the true $m(\cdot)$ well.

The second difficulty is distinguishing $m(X_i)$ from the CMF error $V_i \equiv Y_i - m(X_i)$ in the data. If we knew $Y_i = m(X_i)$ exactly ($V_i = 0$), then we could learn $m(x)$ perfectly for all $x = X_i$. But in reality, we observe $Y_i = m(X_i) + V_i$. If Y_i is big, we don't know if $m(X_i)$ is big or V_i is big. You can think of $m(X_i)$ as the "signal" and V_i as the "noise"; we want to distinguish the signal from the noise. If our model is too flexible, we risk **overfitting**, mistaking noise for signal. For example, perhaps the true $m(\cdot)$ is linear, but we estimate a very nonlinear function.

In practice, the key is balancing the two difficulties described above. If the model is too simple, it may fail to approximate the true CMF. If the model is too complex, it may lead to overfitting. The CMF estimate is bad in either case.

Example 8.23. Imagine the true CMF is $m(x) = x^2$, and $-1 \le X \le 1$. First, you estimate $\beta_0 + \beta_1 x$. However, because there is no x^2 term, this only estimates the population linear projection, which is (let's say) $\text{LP}(Y \mid 1, X) = 0.4$, a flat line that is very different than the CMF x^2. The linear model is not flexible enough here, so the approximation error is very large. Second, you estimate $\sum_{j=0}^{10} \beta_j x^j$. You estimate noticeably non-zero values for all coefficients: $\hat{\beta}_j \neq 0$ for all j. The resulting estimated function goes up and down a lot, often far above or far below the true $m(x) = x^2$. This is because the observations are $Y_i = m(X_i) + V_i$, and some V_i are very high and some are very negative.

In this case, we get a bad estimate due to overfitting: our model is too flexible and fits all the V_i "noise."

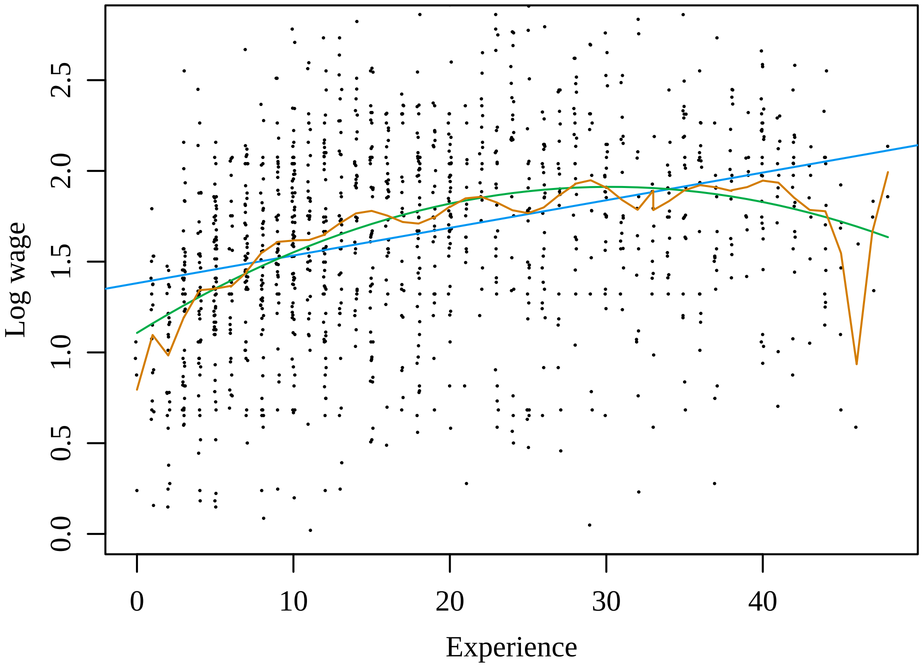


Figure 8.4: Estimates of relationship between wage and experience for Example 8.24.

Example 8.24. Consider the relationship between log wage and years of experience. Figure 8.4 shows three different estimates using a real dataset, using the code below. The linear model is too simple: it shows that an additional year of experience is associated with the same increase in log wage, regardless of the initial experience level. This does not reflect the pattern seen in the scatter plot: log wage increases more steeply with experience at lower experience levels than at higher experience levels. At the other extreme, the 22-degree polynomial is too flexible; its many ups and downs are very likely from overfitting. The quadratic function is at least not obviously wrong, but a more sophisticated nonparametric estimate may be even better.

```
library(wooldridge)
df <- data.frame(Y=beauty$lwage, X=beauty$exper)
lm1 <- lm(Y~X, data=df)
lm2 <- lm(Y~X+I(X^2), data=df)
lm22 <- lm(Y~poly(X,22), data=df)
plot(df$X+runif(length(df$X),-0.1,0.1), df$Y, type='p', pch=19,
     cex=0.1, cex.lab=CEXLAB, cex.axis=CEXAXIS,
     main='', xlab='Experience', ylab='Wage ($/hr)')
abline(lm1, col=2)
lines(sort(df$X), predict(lm2)[order(df$X)], col=3)
lines(sort(df$X), predict(lm22)[order(df$X)], col=4)
```

In more complex models, optimal model selection for prediction may not be optimal for causality. Historically, model selection has focused on prediction. Model selection for causal estimation is a cutting edge area of econometrics research.

8.3.2 Code

The following code shows a particular example of nonparametric regression. Specifically, it uses something called a smoothing spline estimator, implemented in function `smooth.spline()` in R. The different estimates shown (thick red lines) correspond to different levels of flexibility of the model. The plots labeled "GCV" and "LOOCV" refer to formal model selection procedures, provided through the `smooth.spline()` function automatically. The others show intentionally bad fits: one model is "Too flexible," the other is "Not flexible enough." Note that the same data is used for each estimate, as seen in the scatter plots. The thin black line is the true CMF.

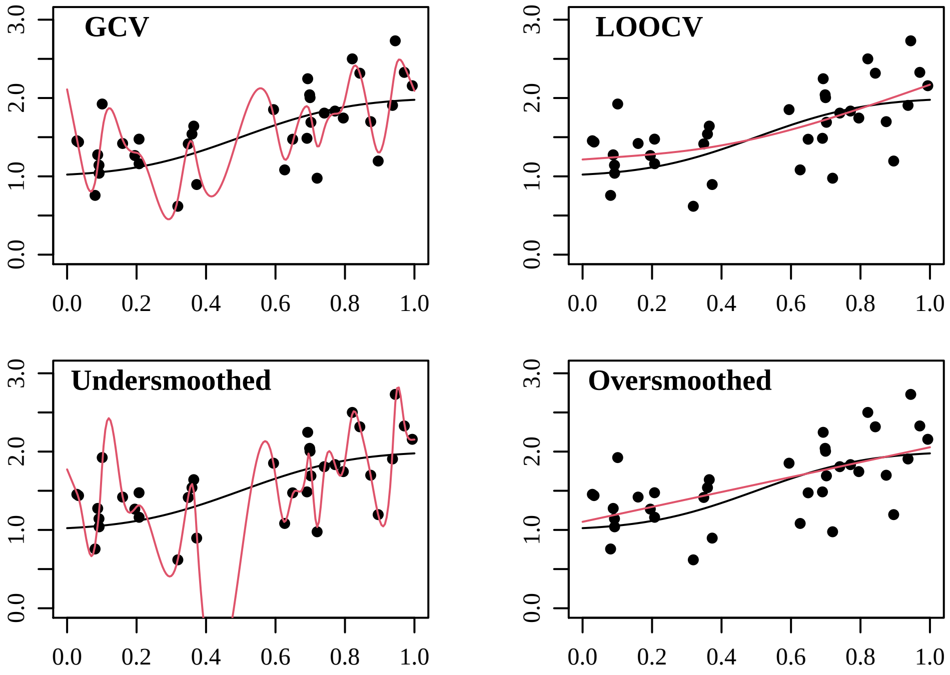


Figure 8.5: Smoothing spline estimates: same data, different amounts of flexibility.

Figure 8.5 shows the results from the following code.

```r
par(family='serif', mar=c(3,3,1,1), mgp=c(2.1,0.8,0), mfrow=c(2,2))
set.seed(112358)
n <- 48;  CMF <- function(x) { 1 + pnorm(12*(x-1/2)) }
df <- data.frame(X=sort(runif(n)))
df$Y <- CMF(df$X) + rbeta(n=n,shape1=2,shape2=2)*2-1
rets <- list()
titles <- c('GCV','LOOCV','Too flexible', 'Not flexible enough')
rets[[1]] <- smooth.spline(x=df$X, y=df$Y, cv=FALSE) #GCV
rets[[2]] <- smooth.spline(x=df$X, y=df$Y, cv=TRUE) #LOOCV
rets[[3]] <- smooth.spline(x=df$X, y=df$Y, df=n)
rets[[4]] <- smooth.spline(x=df$X, y=df$Y, df=2)
xx <- seq(from=0, to=1, by=0.005)
for (ifig in 1:4) {
  plot(x=df$X, y=df$Y, type='p', pch=16, xlab='', ylab='',
       main='', xlim=0:1, ylim=0:1*3.04)
  lines(x=xx, y=CMF(xx), col=1)
  lines(predict(rets[[ifig]], x=xx), col=2)
  title(main=titles[ifig], line=-1, adj=0.1)
}
```

Discussion Question 8.4 (model evaluation)**.** In practice, why don't we just make graphs like in Figure 8.5 and see which fitted function looks best? (Hint: can we make such graphs in practice? If so, how can we agree on which "looks best"? What does "best" mean?)

Optional Resources

Optional resources for this chapter

- Functional form misspecification (Lambert video)
- Log-log example (Lambert video)
- Overfitting (Lambert video)
- Sections 2.4 ("Nonlinearities," including log models), 6.1.3 ("Logarithms"), and 6.1.4 ("Quadratics and Polynomials") in Heiss (2016)
- Section 8.2 ("Nonlinear Functions of a Single Independent Variable") in Hanck et al. (2018)
- Nonparametric regression: Chapter 7 ("Moving Beyond Linearity") in James et al. (2013), including §7.5 ("Smoothing Splines"); and Chapter 5 ("Basis Expansions and Regularization") in Hastie, Tibshirani, and Friedman (2009), including §5.4 ("Smoothing Splines")

- Model selection: Chapter 7 ("Model Assessment and Selection") in Hastie, Tibshirani, and Friedman (2009)
- Bias–variance tradeoff: James et al. (2013, §2.2.2), Hastie, Tibshirani, and Friedman (2009, §§2.9,5.5.2,7.2,7.3)
- Part V ("Nonparametric Regression") in Kaplan (2020)
- R package `splines`

Empirical Exercises

Empirical Exercise EE8.1. You will analyze data on law schools and their student outcomes, originally collected by Kelly Barnett for an economics term project. The idea is to compare median starting salaries of graduates from each law school with the school's cost. Of course, these are not causal estimates: does a Harvard Law graduate make a lot of money because Harvard is expensive, or because she's very skilled (enough to get into Harvard)? Because school cost is essentially a continuous variable, you will explore possible nonlinearity in the (statistical) relationship between cost and salary.

a. Load the data (assuming you've already installed that R package or Stata command).

   R: `library(wooldridge)`

   Stata: `bcuse lawsch85 , nodesc clear`

b. Stata only: make a graph with a local linear nonparametric CMF estimate (of salary given cost), a linear fit, and a quadratic fit, with command `lpoly salary cost , degree(1) n(100) addplot(lfit salary cost || qfit salary cost)` where `n(100)` simply specifies the number of CMF values to estimate and plot, and `lfit` and `qfit` stand for linear fit and quadratic fit, and model selection is done with a "rule-of-thumb" formula that attempts to optimally balance variance and squared bias.

c. R only: make a data frame named `df` with only salary and cost variables, and only when both are observed, with

   ```
   df <- data.frame(Y=lawsch85$salary, X=lawsch85$cost)
   df <- df[!(is.na(df$Y) | is.na(df$X)) , ]
   ```

   where `is.na()` is `TRUE` if the entry is missing and `FALSE` if not.

d. R only: compute and store linear and quadratic (in variables) regressions with `retlm <- lm(Y~X, data=df)` and `retnl <- lm(Y~X+I(X^2), data=df)`

e. R only: compute and store a nonparametric smoothing spline CMF estimate with GCV model selection with command `retss <- smooth.spline(x=df$X, y=df$Y , cv=FALSE)`

f. R only: specify a sequence of X values and compute CMF estimates at each value from each of the three models (linear, quadratic, nonparametric). Store the sequence as `xx` with `xx <- seq(from=min(df$X), to=max(df$X), length.out=100)` and then compute the estimates as

   ```
   fitlm <- predict(retlm, newdata=data.frame(X=xx))
   fitnl <- predict(retnl, newdata=data.frame(X=xx))
   fitss <- predict(retss, newdata=data.frame(X=xx))
   ```

g. R only: make a scatterplot of raw data with

```
plot(x=df$X, y=df$Y, xlab='Cost', ylab='Starting Salary')
```

h. R only: plot the three estimated CMFs as lines over the scatterplot with

```
lines(x=xx, y=fitlm, col=1, lty=1)
lines(x=xx, y=fitnl, col=2, lty=5)
lines(fitss, col=4, lty=3)
```

i. Optional: repeat your analysis but with the school's rank (variable `rank`) instead of cost.

j. Optional: repeat again but with log salary and log rank. Log salary is already in the dataset as variable `lsalary` (that's a lowercase L before salary).

   R: `df <- data.frame(Y=lawsch85$lsalary, X=log(lawsch85$rank))`

   Stata: `generate lrank = log(rank)` then use `lrank` and `lsalary`

Empirical Exercise EE8.2. You will analyze data on sleep and wages, originally from Biddle and Hamermesh (1990). Specifically, you'll estimate the CMF of daily hours of sleep conditional on hourly wage. For now, just drop missing values without worry, and focus on the linear, quadratic, and nonparametric estimation.

a. Load the data (assuming you've already installed that R package or Stata command).

   R: `library(wooldridge)`

   Stata: `bcuse sleep75 , nodesc clear`

b. R only: follow the same steps (identical code) as in EE8.1 through part (h), after setting up the data frame named `df`. Specifically, replace EE8.1(c) with

```
df <- data.frame(Y=sleep75$slpnaps/7/60, X=sleep75$hrwage)
df <- df[!(is.na(df$Y) | is.na(df$X)) , ]
```

   and then use the same code for all subsequent steps

c. Stata only: generate a new variable that translates the total weekly minutes of sleep into average daily hours of sleep with `generate sleephrsdaily = slpnaps/7/60`

d. Stata only: graph linear, quadratic, and nonparametric (local linear) CMF estimates similar to EE8.1(b), with command `lpoly sleephrsdaily hrwage , degree (1) n(100) addplot(lfit sleephrsdaily hrwage || qfit sleephrsdaily hrwage )`

e. Optional: repeat your analysis, but instead of `hrwage` use `totwrk` as the conditioning variable (regressor); this is total minutes of work per week. (You could also adjust it to be average daily hours of work, to make it more comparable to the sleep variable you use.)

Empirical Exercise EE8.3. You will analyze data from the 1994–1995 men's college basketball season scores and Las Vegas betting "spreads," originally collected by Scott

Resnick. Before each game, people can bet on whether the score difference will be "over" or "under" the spread set by bookmakers in Las Vegas. (In the data, the "difference" is the favored team's score minus the other team's score; so the variable `spread` is always positive, but the actual score difference `scrdiff` can be negative if the favored team loses.) Basically, the bookmaker adjusts the spread so that half the bets are "over" and half "under," so regardless of the actual score outcome, half win and half lose (and the bookmaker always profits): the losers pay the winners, and the bookmaker keeps the transaction fees. (It's a little complicated because bets can be placed at different times, and the spread can change over time, but we can imagine a simplified version where everyone bets at once and the spread is set so that half bet "over" and half "under.") See the Wikipedia entry at https://en.wikipedia.org/wiki/Spread_betting for more on spread betting. Consequently, the spread does not reflect the bookmaker's belief, but rather the aggregate beliefs of everybody betting on the game. The accuracy of such aggregate wisdom has spurred the creation of "prediction markets" for events beyond sports, like presidential elections, although there have been notable failures (e.g., 2016 U.S. presidential election).[2] You will check whether the Las Vegas spread is indeed a good predictor of the actual score difference.

Technically, the above arguments suggest that given the spread, the *median* score difference should equal the spread, not the mean. But, such an investigation would require "median regression" (a type of "quantile regression"), which is beyond our scope. Instead, you will investigate whether the spread is still a good predictor of the actual score difference with quadratic loss. Specifically, you can check if the OLS fit has intercept close to 0 and slope close to 1 (and whether those values are in the respective confidence intervals).

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

   ```
   library(wooldridge); library(sandwich); library(lmtest)
   ?pntsprd
   ```

b. Stata only: load the data with `bcuse pntsprd , nodesc clear` (assuming `bcuse` already installed)

c. For each observation (each game), compute whether the actual score difference was over, under, or equal to the spread. In math and in the code below, the "sign" function (not to be confused with "sine") equals $+1$ for strictly positive values, -1 for strictly negative values, and 0 for zero

   R: `overunder <- sign(pntsprd$scrdiff-pntsprd$spread)`

   Stata: `generate overunder = sign(scrdiff - spread)`

d. Display the frequency of over, under, and equal.

   R: `table(overunder, useNA='ifany')`

[2] See https://en.wikipedia.org/wiki/Prediction_market for more.

Stata: `tabulate overunder , missing`

e. Regress the score difference on the spread.

   R: `ret <- lm(scrdiff~spread, data=pntsprd)`

   Stata: `regress scrdiff spread , vce(robust)`

f. R only (because already reported by Stata): display the point estimates and heteroskedasticity-robust 95% confidence intervals for the intercept and slope with

```
cbind(coeftest(ret, vcov.=vcovHC(ret, type='HC1'))[,1:2],
      coefci(  ret, vcov.=vcovHC(ret, type='HC1')) )
```

g. Plot nonparametric CMF fitted values against the line $Y = X$ (intercept zero, slope one).

   R: `plot(smooth.spline(x=pntsprd$spread, y=pntsprd$scrdiff))` then `abline(a=0, b=1, col=2)`

   Stata: `lpoly scrdiff spread , degree(1) addplot(function y=x , range(spread)) noscatter`

h. Optional: repeat your analysis in parts (e)–(g) but with the reverse regression: regress the spread on the score difference. (Is the slope still close to 1? Are you surprised? Consider games with the biggest possible score difference; should the spread be even bigger half the time?)

Chapter 9

Regression with Two Binary Regressors

⟹ Kaplan video: Chapter Introduction

Perhaps surprisingly, there is a lot to think about with even just two binary regressors. Topics include (mis)specification of a CMF model, interaction between regressors as a type of nonlinearity, interpretation of regression coefficients, causality, estimation, and more.

Unit learning objectives for this chapter

9.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

9.2. Assess whether there is bias from omitting a variable in a real-world example, including the direction of bias [TLOs 5 and 6]

9.3. Interpret (appropriately) the coefficients of a regression with two binary variables, mathematically and intuitively, for description, prediction, and causality [TLO 3]

9.4. Assess whether comparing changes in two groups over time can be interpreted causally, and interpret such differences appropriately [TLOs 2, 3, and 6]

9.5. In R (or Stata): estimate regression models with two binary variables, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

9.1 Causality: Omitted Variable Bias

⟹ Kaplan video: Omitted Variable Bias

For causality, **omitted variable bias** (OVB) is a common problem in economics. More broadly, it is a common problem in any field that uses observational (non-experimental) data and has many variables interact in complex ways. Generally, OVB arises because a variable outside our model is moving with X and causing Y to change, but our model assumes these changes are entirely from X.

9.1.1 An Allegory

Imagine a ghost (Q) that often accompanies a child (X), i.e., the ghost and child are often in the same place at the same time. The ghost always makes a huge mess (Y): spilling flour, knocking over chairs, drawing on walls, etc. The child's parents only observe the child and the mess; they do not observe the ghost. The parents note that when the child is in the kitchen, then there is often a mess in the kitchen, and when the child is in the bathroom, then there is often a mess in the bathroom, etc. Thus, they infer that the child (X) causes the mess (Y). However, we know that it only appears that way because

GHOST.1 the ghost (Q) often accompanies the child (X) and

GHOST.2 the ghost (Q) causes a mess (Y).

The child is the regressor. The ghost is the omitted variable. The parents are economists who over-estimate how much mess the child causes. This phenomenon is OVB.

9.1.2 Formal Conditions

The ghost of OVB can be formalized as follows. Consider the structural model

$$Y = \beta_0 + \beta_1 X + \beta_2 Q + V, \tag{9.1}$$

where $\text{Cov}(X, V) = 0$. If we don't observe Q, then instead we have the structural model

$$Y = \beta_0 + \beta_1 X + U, \quad U \equiv \beta_2 Q + V. \tag{9.2}$$

Here, X is sometimes called the **included regressor** (included in the model; not omitted). If X is binary, then for OLS to estimate β_1 requires $\text{E}(U \mid X = 0) = \text{E}(U \mid X = 1)$: the average effect of the structural error term U must be the same for both X groups. For simplicity, imagine Q is also binary.

Condition GHOST.1, "the ghost follows the child," means that we usually see $Q = 1$ when $X = 1$, and $Q = 0$ when $X = 0$. More generally, it means Q is correlated with X. This correlation does not need to have a causal interpretation. It does not matter why the ghost follows the child: maybe the ghost likes the child's company (or vice-versa), or maybe they just get hungry at the same time. It only matters that they tend to be in the same place: Q and X tend to have the same value. OVB can also occur if there is a negative correlation, e.g., if usually $Q = 1$ when $X = 0$, and $Q = 0$ when $X = 1$.

Condition GHOST.2, "the ghost causes a mess," means that Q is a causal determinant of Y. In (9.1), this means $\beta_2 \neq 0$. Although in the example $\beta_2 > 0$ (more mess), OVB can occur with $\beta_2 < 0$, too. For example, maybe the child is really messy, but the ghost cleans everything up; then the parents would incorrectly think the child is not messy.

To summarize: for variable Q that is not included as a regressor (it is omitted from the model), it will cause OVB if both of the following conditions hold.

OVB.1 $\text{Corr}(Q, X) \neq 0$: the omitted variable is correlated with the included regressor.

OVB.2 The omitted variable Q is a causal determinant of Y (not only through X).

The terms for Q can be confusing. The variable Q may be called an "omitted variable," although that can sound ambiguous. The term "confounder" is more precise but usually is defined to require Q to have a causal effect on X, whereas here only correlation is required. Sometimes people (including me) say "confounder" for any variable causing OVB.

Assessing OVB Conditions Empirically

If Q is observed in the data, then you can compare $\hat{\beta}_1$ (the estimated coefficient on X) when Q is included as a regressor to $\hat{\beta}_1$ when Q is omitted. If the estimates are meaningfully different (economically), then it may be best to include Q to reduce OVB. However, there are other types of variables that would also lead to a different $\hat{\beta}_1$ but are actually worse to include, so careful thought is required; see Section 9.6.

If Q is not observed in the data, then even $\text{Corr}(Q, X)$ in OVB.1 cannot be assessed empirically (i.e., using data).

Beware of "omitted variable" tests that are not concerned with this type of OVB. For example, Stata's `ovtest` implements the Ramsey test (RESET). Although the `ov` in `ovtest` indeed stands for "omitted variables," the Ramsey test only looks for (certain types of) nonlinearity, to see whether a polynomial model might be better than a linear model. That is, it is about nonlinearity in X (Section 8.2), not about a separate Q variable. Besides, as you learned in Section 8.3, hypothesis testing is a bad way to do model selection.

Example

For example, imagine we want to learn the effect of kindergarten classroom size on earnings as an adult. (This is inspired by Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2011), who actually have randomized experimental data to answer this question.) Let Y denote the annual earnings of the individual at age 30. Let $X = 1$ if (as a child) the individual was in a kindergarten classroom with more than 24 students and $X = 0$ otherwise. Imagine X is not randomized. We are curious whether we can just regress Y on X, or if there is OVB. Consider the following possible omitted variables.

First, consider Q to be somebody's first grade class size. (First grade is the year after kindergarten in the U.S.) As with X, $Q = 1$ if it is above 24 students and $Q = 0$ otherwise. Since it seems like kindergarten class size has an effect on adult earnings (Y) according to Chetty et al. (2011), probably first grade class size does, too, satisfying OVB.2. If all students in the population are completely randomly assigned to classes each year, $\text{Corr}(X, Q) = 0$; then, OVB.1 does not hold, so this Q would not cause OVB. However, students tend to stay in the same school, and some schools tend to have smaller class sizes than others, so OVB.1 probably does hold. Because both OVB.1 and OVB.2 are true, there is OVB.

Second, consider Q as the number of cubbies (places to put clothes, backpacks, etc.) in somebody's kindergarten classroom. Presumably larger classes ($X = 1$) require more cubbies because there are more students, so $\text{Corr}(Q, X) > 0$, satisfying OVB.1. However, I'd guess the number of cubbies does not have a causal effect on future earnings Y. That is, if we simply went into classrooms and added a few cubbies (without adding students), I don't think it would affect students' future earnings. Thus, OVB.2 does not hold, and this Q does not cause OVB.

Third, consider $Q = 1$ if the kindergarten is in a high-income area and $Q = 0$ otherwise. Areas with higher income are more likely to be able to afford more teachers to keep class sizes small. That is, it's more likely to see $Q = 1$ and $X = 0$, or $Q = 0$ and $X = 1$, so $\text{Corr}(Q, X) < 0$, satisfying OVB.1. Also, Chetty, Hendren, and Katz (2016) provide evidence that growing up in a higher-income area has a positive causal effect on earnings as an adult (not only because of smaller kindergarten classes), meaning Q is a causal determinant of Y, satisfying OVB.2. Thus, omitting this Q causes OVB.

In Sum: Possible Omitted Variables (Q) in Kindergarten Example

First grade class size: affects earnings (OVB.2), and probably correlated with kindergarten class size (OVB.1) if population includes multiple schools $\implies$ OVB
Cubbies: more if more students (OVB.1), but no causal effect on earnings (no OVB.2) $\implies$ no OVB
Neighborhood income: smaller classes if higher income (OVB.1), and affects earnings (OVB.2) $\implies$ OVB

Discussion Question 9.1 (assessing OVB)**.** Among public elementary schools (students mostly 5–11 years old) in California, let Y be the average standardized math test score among a school's 5th-graders, and let X be the school's student-teacher ratio for 5th-graders (like average number of students per class). Consider a simple regression of Y on X. For *any two of the following* variables, assess each OVB condition separately, and then decide whether you think it's a source of OVB.

a) School's parking lot area per student. (Remember 5–11-year-olds don't have cars to park.)
b) Time of day of the test.
c) School's total spending per student (including books, facilities, etc.).
d) Percentage of English learners (non-native speakers) among a school's 5th-grade students.

9.1.3 Consequences

The practical problem of OVB is that we systematically over-estimate or under-estimate the true structural parameter. This consequence is quantified below.

Formulas

The following results are much more general than OVB with binary regressors. Beyond OVB, they quantify the consequences of any source of endogeneity that causes correlation between the regressor X and structural error term U. Other sources of endogeneity are discussed in Section 12.3. The results also apply to any discrete and continuous X.

Given structural model $Y = \beta_0 + \beta_1 X + U$, the OLS estimator of β_1 has the property

$$\operatorname*{plim}_{n\to\infty} \hat{\beta}_1 = \beta_1 + \frac{\mathrm{Cov}(X,U)}{\mathrm{Var}(X)}. \tag{9.3}$$

That is (from Section 3.6.3), for large samples (large n), the estimator $\hat{\beta}_1$ is close to the right-hand side expression in most randomly sampled datasets.

Equation (9.3) shows OVB is not solved by having lots of data. Unless $\mathrm{Corr}(X,U) = 0$, the OLS estimator is not consistent for the structural β_1.

Rearranging (9.3), the asymptotic bias (as in (3.26)) is

$$\mathrm{AsyBias}(\hat{\beta}_1) \equiv \operatorname*{plim}_{n\to\infty} \hat{\beta}_1 - \beta_1 = \frac{\mathrm{Cov}(X,U)}{\mathrm{Var}(X)} = \text{slope coefficient in } \mathrm{LP}(U \mid 1, X). \tag{9.4}$$

The characterization as a linear projection slope coefficient comes from replacing Y with U in (7.8). This can't be computed from data because U is unobserved, but it is helpful for thinking about the direction and magnitude of asymptotic bias.

Although technically this is "asymptotic bias" rather than "bias" (Section 3.6.1), the practical implication is the same. Although very different mathematically, we won't worry about such technicalities.

Direction of Asymptotic Bias

The direction ($+$ or $-$) of the asymptotic bias in (9.4) depends on the sign ($+$ or $-$) of the slope in $\mathrm{LP}(U \mid 1, X)$. Ths sign of this slope is equivalent to the sign of $\mathrm{Corr}(X,U)$.

If $\mathrm{Corr}(X,U) > 0$, then $\mathrm{AsyBias}(\hat{\beta}_1) > 0$. This is positive (upward) asymptotic bias, meaning we systematically estimate a value "above" the true β_1. "Above" does not mean "bigger in magnitude": it could be that $\beta_1 = -9$ and positive asymptotic bias causes $\mathrm{plim}_{n\to\infty} \hat{\beta}_1 = 0$. This is "positive" because $0 - (-9) > 0$ (positive), but we might also say that we're estimating a "smaller" effect (in fact zero effect) in the sense that $|0| < |-9|$. This can be confusing.

If $\mathrm{Corr}(X,U) < 0$, then $\mathrm{AsyBias}(\hat{\beta}_1) < 0$, meaning negative (downward) asymptotic bias. Again confusing, negative asymptotic bias can actually make effects look bigger, e.g., if $\beta_1 = 0$ and $\mathrm{plim}_{n\to\infty} \hat{\beta}_1 = -9$: the true effect is zero, but the negative asymptotic bias makes it appear like there is an effect.

Results in Terms of Q

For OVB specifically, the general results in terms of U can be translated to Q. As in (9.2), let $U = \beta_2 Q + V$, with $\mathrm{Cov}(X, V) = 0$. Then, using a linearity property of covariance,

$$\mathrm{Cov}(X,U) = \mathrm{Cov}(X, \beta_2 Q + V) = \beta_2 \mathrm{Cov}(X,Q) + \overbrace{\mathrm{Cov}(X,V)}^{=0}. \tag{9.5}$$

Plugging this into (9.3),

$$\operatorname*{plim}_{n\to\infty} \hat\beta_1 = \beta_1 + \frac{\mathrm{Cov}(X,U)}{\mathrm{Var}(X)} = \beta_1 + \beta_2 \frac{\mathrm{Cov}(X,Q)}{\mathrm{Var}(X)} = \beta_1 + \beta_2 \mathrm{Corr}(X,Q)\sqrt{\frac{\mathrm{Var}(Q)}{\mathrm{Var}(X)}}. \tag{9.6}$$

Interestingly, similar to (9.4), $\mathrm{Cov}(X,Q)/\mathrm{Var}(X)$ is the slope of the population linear projection of Q onto X (and an intercept), $\mathrm{LP}(Q \mid 1, X)$. So, the asymptotic bias is the product $\beta_2\gamma_1$, where β_2 is the structural slope coefficient on Q in (9.1), and γ_1 is the linear projection slope coefficient in $\mathrm{LP}(Q \mid 1, X) = \gamma_0 + \gamma_1 X$.

Equation (9.6) shows why both Conditions OVB.1 and OVB.2 are required for OVB. Condition OVB.1 says $\mathrm{Corr}(X,Q) \neq 0$, while OVB.2 says $\beta_2 \neq 0$. If either $\beta_2 = 0$ or $\mathrm{Corr}(X,Q) = 0$ in (9.6), then $\beta_2 \mathrm{Corr}(X,Q) = 0$, and the asymptotic bias disappears, $\mathrm{AsyBias}(\hat\beta_1) = 0$.

The direction of asymptotic bias can also be interpreted in terms of Q. Using (9.6), the sign of the asymptotic bias is the sign of $\beta_2 \mathrm{Corr}(X,Q)$. That is, if $\beta_2 \mathrm{Corr}(X,Q) > 0$, then there is positive (upward) asymptotic bias; if $\beta_2 \mathrm{Corr}(X,Q) < 0$, then there is negative (downward) asymptotic bias.

Example

Consider the asymptotic bias direction in the example where $X = 1$ if the kindergarten class size is large and $Q = 1$ if the neighborhood income is high. Earlier, we thought probably $\mathrm{Corr}(X,Q) < 0$ and $\beta_2 > 0$. Thus, there is negative OVB because $\beta_2 \mathrm{Corr}(X,Q) < 0$. That is, if the true effect of class size on earnings is β_1, then we systematically estimate something below β_1.

Does this make the effect size (absolute value) appear bigger or smaller? Because smaller classes are better, average earnings (Y) are higher when $X = 0$ than when $X = 1$. This means a negative slope: $\beta_1 < 0$. That is, the effect of changing from a smaller class ($X = 0$) to a larger class ($X = 1$) is lower future earnings ($\beta_1 < 0$). Negative asymptotic bias means we estimate something even more negative: $\mathrm{plim}_{n\to\infty} \hat\beta_1 < \beta_1 < 0$. This makes the size of the effect appear larger than it really is: we estimate something farther away from zero.

Intuitively, this OVB direction makes sense. Individuals who had a small kindergarten class tend to have grown up in wealthier areas with lots of other advantages that also cause higher earnings. If we ascribe the entire mean earnings difference to kindergarten, then it falsely appears that kindergarten alone cause the big difference, when in reality many different forces were all working together in the same direction.

In Sum: OVB Assessment

1. Think of a specific variable Q
2. Assess OVB.1: correlated with X?
3. Assess OVB.2: causal effect on Y? (separate from X effect)
4. If both OVB.1 and OVB.2 $\implies$ OVB
5. OVB direction: positive bias if $\mathrm{Corr}(X, Q)$ and effect of Q on Y are either both $+$ or both $-$; otherwise, negative bias
6. OVB magnitude: all else equal, larger (in absolute value) if i) larger effect of Q on Y, ii) larger $\mathrm{Corr}(X, Q)$, iii) larger $\mathrm{Var}(Q)/\mathrm{Var}(X)$.

Practice 9.1 (OVB: kindergarten)**.** Consider the OVB example with earnings as an adult (Y), kindergarten classroom size (X), and childhood neighborhood income (Q). But, reverse the definition of X: let $X = 1$ for smaller classrooms (24 or fewer students) and $X = 0$ for larger classrooms. Say whether you think each of the following is positive or negative, and explain why: a) β_1; b) $\mathrm{Corr}(X, Q)$; c) β_2; and d) OVB. Also discuss: e) will our estimated effect $\hat{\beta}_1$ tend to be larger or smaller than the true effect β_1, and why?

Discussion Question 9.2 (OVB: ES habits)**.** Recall from DQ 6.2 the example with Y as a student's final semester score ($0 \leq Y \leq 100$) and $X = 1$ if a student starts the exercise sets well ahead of the deadline (and $X = 0$ otherwise).

a) What's one variable that might cause OVB? Explain why you think both OVB conditions are satisfied.

b) Which direction of asymptotic bias would your omitted variable cause? Explain.

9.1.4 OVB in Linear Projection

For linear projection (without causal interpretation), the OVB formula is actually the same as (9.6), just with β_1 and β_2 interpreted as linear projection coefficients rather than structural coefficients. Similar results for larger linear projection models are in Hansen (2020, §2.24), for example.

However, if we are interested in prediction, we don't care whether our $\hat{\beta}_1$ estimates a particular linear projection coefficient; we only care whether we can predict Y well. Of course, we don't want to omit Q if it's helpful for prediction, but we don't care about OVB itself. That is, OVB is only a problem for causality, not prediction.

9.2 Linear-in-Variables Model

The simplest CMF model with two binary variables is linear-in-variables (Section 8.2.1),

$$\mathrm{E}(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{9.7}$$

Misspecification

Unfortunately, (9.7) may be misspecified. Recall from Section 7.1 that misspecification arose when X had three values but the CMF model $\beta_0 + \beta_1 X$ had only two parameters. The case here is similar: (9.7) has only 3 parameters, but there are 4 possible values of (X_1, X_2). Specifically, (X_1, X_2) could equal $(0,0)$, $(0,1)$, $(1,0)$, or $(1,1)$. Consequently, there are four CMF values:

$$\begin{aligned} m(0,0) &= \mathrm{E}(Y \mid X_1 = 0, X_2 = 0), \quad & m(0,1) &= \mathrm{E}(Y \mid X_1 = 0, X_2 = 1), \\ m(1,0) &= \mathrm{E}(Y \mid X_1 = 1, X_2 = 0), \quad & m(1,1) &= \mathrm{E}(Y \mid X_1 = 1, X_2 = 1). \end{aligned} \tag{9.8}$$

To see the possible misspecification, we can write the β_j regression coefficients in terms of the CMF values $m(x_1, x_2)$. If (9.7) were true, then

$$m(0,0) = \beta_0 + (\beta_1)(0) + (\beta_2)(0) = \beta_0, \tag{9.9}$$

$$m(0,1) = \beta_0 + (\beta_1)(0) + (\beta_2)(1) = \beta_0 + \beta_2, \tag{9.10}$$

$$m(1,0) = \beta_0 + (\beta_1)(1) + (\beta_2)(0) = \beta_0 + \beta_1, \tag{9.11}$$

$$m(1,1) = \beta_0 + (\beta_1)(1) + (\beta_2)(1) = \beta_0 + \beta_1 + \beta_2. \tag{9.12}$$

Consequently, β_1 has two interpretations. It equals either (9.12) minus (9.10), or (9.11) minus (9.9):

$$\begin{aligned} m(1,1) - m(0,1) &= (\beta_0 + \beta_1 + \beta_2) - (\beta_0 + \beta_2) = \beta_1, \\ m(1,0) - m(0,0) &= (\beta_0 + \beta_1) - \beta_0 = \beta_1. \end{aligned}$$

Thus, the model implicitly assumes $m(1,1) - m(0,1) = m(1,0) - m(0,0)$, which may not be true of the real CMF. For example,

$$\begin{aligned} &m(0,0) = 0, m(1,0) = 1, m(0,1) = 2, m(1,1) = 4 \\ &\qquad\qquad \Longrightarrow m(1,1) - m(0,1) = 2, m(1,0) - m(0,0) = 1. \end{aligned}$$

Because $m(1,1) - m(0,1) \neq m(1,0) - m(0,0)$, the CMF model in (9.7) is misspecified (wrong). That is, there are no possible $(\beta_0, \beta_1, \beta_2)$ such that $m(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

As discussed in Chapter 7, if the CMF model is wrong, then OLS estimates the linear projection. Here, OLS estimates $\mathrm{LP}(Y \mid 1, X_1, X_2)$. However, this is not useful for causality, and the misspecification is easily fixed.

More Consideration

Before we fix the misspecification, consider more carefully why (9.7) is usually misspecified. To be concrete, imagine Y is wage, $X_1 = 1$ if an individual has a college degree (and $X_1 = 0$ if not), and $X_2 = 1$ if an individual has at least 10 years of work experience (and $X_2 = 0$ if not). For simplicity, we'll call X_1 "education" and X_2 "experience." The quantity $m(1,1) - m(0,1)$ compares the mean wage in the high-education, high-experience group (subpopulation) with the mean wage in the low-education, high-experience group. That is, within the high-experience subpopulation, it compares the mean wage of the high-education and low-education sub-sub-populations. The quantity $m(1,0) - m(0,0)$ also compares mean wages across high and low education, but within the low-experience subpopulation. Thus, assuming $m(1,1)-m(0,1) = m(1,0)-m(0,0)$ can be interpreted as assuming that the mean wage difference between high-education and low-education groups is identical within the high-experience subpopulation and within the low-experience subpopulation. This is a strong assumption that is probably not true in this example (or in most examples).

9.3 Fully Saturated Model

⟹ Kaplan video: Fully Saturated Model Interpretation

Misspecification is avoided by adding the **interaction term** X_1X_2:

$$\mathrm{E}(Y \mid X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{9.13}$$

Mathematically, interaction terms often involve the product of two regressors, like X_1X_2 here. Economically, the interaction term allows the mean Y difference associated with X_1 to depend on the value of X_2. Similarly, it allows the mean Y difference associated with X_2 to depend on the value of X_1. For example, the mean wage difference associated with education can depend on the value of experience. More generally, interaction terms allow the change in Y associated with a unit increase in one regressor to depend on the value of another regressor.

The CMF model in (9.13) is also called **fully saturated** (Section 7.2.2) because it is flexible enough to allow a different CMF value for each value of (X_1, X_2). Logically, having the same number (four) of possible values of (X_1, X_2) as β_j parameters is necessary but not sufficient for the model to be fully saturated.

Interpretation of the coefficients requires writing them in terms of different CMF values. First, similar to (9.9)–(9.12), each CMF value can be written in terms of the β_j:

$$m(x_1, x_2) = \beta_0 + (\beta_1)(x_1) + (\beta_2)(x_2) + (\beta_3)(x_1)(x_2),$$

$$m(0,0) = \beta_0 + (\beta_1)(0) + (\beta_2)(0) + (\beta_3)(0)(0) = \beta_0, \tag{9.14}$$

$$m(0,1) = \beta_0 + (\beta_1)(0) + (\beta_2)(1) + (\beta_3)(0)(1) = \beta_0 + \beta_2, \tag{9.15}$$

$$m(1,0) = \beta_0 + (\beta_1)(1) + (\beta_2)(0) + (\beta_3)(1)(0) = \beta_0 + \beta_1, \tag{9.16}$$

$$m(1,1) = \beta_0 + (\beta_1)(1) + (\beta_2)(1) + (\beta_3)(1)(1) = \beta_0 + \beta_1 + \beta_2 + \beta_3. \tag{9.17}$$

From (9.14)–(9.17) and their differences,

$$\overbrace{\beta_0 = m(0,0)}^{(9.14)}, \tag{9.18}$$

$$\beta_1 = \overbrace{(\beta_0 + \beta_1) - \beta_0 = m(1,0) - m(0,0)}^{(9.16)\text{ minus }(9.14)}, \tag{9.19}$$

$$\beta_2 = \overbrace{(\beta_0 + \beta_2) - \beta_0 = m(0,1) - m(0,0)}^{(9.15)\text{ minus }(9.14)}, \tag{9.20}$$

$$\beta_3 = [\beta_2 + \beta_3] - [\beta_2] = \overbrace{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_1)]}^{(9.17)\text{ minus }(9.16)} - \overbrace{[(\beta_0 + \beta_2) - (\beta_0)]}^{(9.15)\text{ minus }(9.14)}$$

$$= \overbrace{\overbrace{[m(1,1) - m(1,0)]}^{\text{difference}} - \overbrace{[m(0,1) - m(0,0)]}^{\text{difference}}}^{\text{difference-in-differences}} \tag{9.21}$$

$$= [m(1,1) - m(0,1)] - [m(1,0) - m(0,0)] \tag{9.22}$$

$$= \overbrace{[(\beta_0 + \beta_1 + \beta_2 + \beta_3) - (\beta_0 + \beta_2)]}^{(9.17)\text{ minus }(9.15)} - \overbrace{[(\beta_0 + \beta_1) - (\beta_0)]}^{(9.16)\text{ minus }(9.14)}.$$

Because of the **difference-in-differences** structure seen in (9.21) and (9.22), this model is sometimes called a difference-in-differences model, particularly when X_2 represents time and X_1 represents a "treatment" (see Section 9.7).

Using (9.18)–(9.22), the four β_j in (9.13) have the following interpretations, both in terms of the wage example (Y wage, X_1 education, X_2 experience) and more generally.

- $\beta_0 = m(0,0)$ is the mean wage among low-education, low-experience individuals.

  More generally, β_0 is the mean Y in the subpopulation with $X_1 = 0$ and $X_2 = 0$.

  Caution: generally $\beta_0 \neq \mathrm{E}(Y)$.

- $\beta_1 = m(1,0) - m(0,0)$ is the mean wage difference between high-education and low-education individuals within the low-experience subpopulation.

  More generally, β_1 is the mean Y difference between $X_1 = 1$ and $X_1 = 0$ individuals within the $X_2 = 0$ subpopulation.

  Caution: generally $\beta_1 \neq \mathrm{E}(Y \mid X_1 = 1) - \mathrm{E}(Y \mid X_1 = 0)$; it additionally conditions on $X_2 = 0$.

- $\beta_2 = m(0,1) - m(0,0)$ is the mean wage difference between high-experience and low-experience individuals within the low-education subpopulation.

  More generally, β_2 is the mean Y difference between $X_2 = 1$ and $X_2 = 0$ individuals within the $X_1 = 0$ subpopulation.

  Caution: generally $\beta_2 \neq \mathrm{E}(Y \mid X_2 = 1) - \mathrm{E}(Y \mid X_2 = 0)$; it additionally conditions on $X_1 = 0$.

- $\beta_3 = [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)]$ is the mean wage difference associated with experience in the high-education subpopulation *minus* the mean wage difference associated with experience in the low-education subpopulation.

  More generally, β_3 is the mean Y difference associated with X_2 in the $X_1 = 1$ subpopulation *minus* the mean Y difference associated with X_2 in the $X_1 = 0$ subpopulation.

- $\beta_3 = [m(1,1) - m(0,1)] - [m(1,0) - m(0,0)]$ is also the mean wage difference associated with education in the high-experience subpopulation *minus* the mean wage difference associated with education in the low-experience subpopulation.

  More generally, β_3 is the mean Y difference associated with X_1 in the $X_2 = 1$ subpopulation *minus* the mean Y difference associated with X_1 in the $X_2 = 0$ subpopulation.

The β_j interpretations can also be seen by considering the regression of Y on X_1 when $X_2 = 0$ and separately when $X_2 = 1$. That is, plugging in $x_2 = 0$ first and then $x_2 = 1$ second,

$$m(x_1, 0) = \beta_0 + \beta_1 x_1 + (\beta_2)(0) + (\beta_3)(x_1)(0) = \beta_0 + \beta_1 x_1, \tag{9.23}$$

$$m(x_1, 1) = \beta_0 + \beta_1 x_1 + (\beta_2)(1) + (\beta_3)(x_1)(1) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1. \tag{9.24}$$

That is, when changing from $X_2 = 0$ to $X_2 = 1$, the intercept changes by β_2 and the slope changes by β_3. These changes could be positive or negative, or zero. The interaction coefficient β_3 describes how the slope with respect to X_1 differs when $X_2 = 1$ versus $X_2 = 0$.

Equivalently, we could switch all the X_1 and X_2 and interpret β_3 as the difference between the slope with respect to X_2 when $X_1 = 1$ versus when $X_1 = 0$:

$$m(0, x_2) = \beta_0 + (\beta_1)(0) + \beta_2 x_2 + (\beta_3)(0)(x_2) = \beta_0 + \beta_2 x_2, \tag{9.25}$$

$$m(1, x_2) = \beta_0 + (\beta_1)(1) + \beta_2 x_2 + (\beta_3)(1)(x_2) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_2. \tag{9.26}$$

Example 9.1 (Kaplan video)**.** Let Y be wage (\$/hr), $D_1 = 1$ if an individual has a college degree ($D_1 = 0$ if not), and $D_2 = 1$ if an individual has more than 15 years of experience (and $D_2 = 0$ if not). You have a sample of data and run OLS on the fully saturated model, yielding $\hat{m}(d_1, d_2) = 10 + 5d_1 + d_2 + 2d_1 d_2$. For the college-educated subpopulation ($d_1 = 1$), the estimated change in mean wage associated with changing from low to high experience ($d_2 = 0$ to $d_2 = 1$) is

$$\overbrace{10 + 5 + 1 + 2}^{\hat{m}(1,1)} - \overbrace{10 + 5 + 0 + 0}^{\hat{m}(1,0)} = 3.$$

Within the low-experience subpopulation ($d_2 = 0$), the estimated difference in mean wage between the college ($d_1 = 1$) and no-college ($d_1 = 0$) subpopulations is

$$\overbrace{10 + 5 + 0 + 0}^{\hat{m}(1,0)} - \overbrace{10 + 0 + 0 + 0}^{\hat{m}(0,0)} = 5.$$

Within the high-experience subpopulation ($d_2 = 1$), the estimated difference in mean wage between the college ($d_1 = 1$) and no-college ($d_1 = 0$) subpopulations is

$$\overbrace{10 + 5 + 1 + 2}^{\hat{m}(1,1)} - \overbrace{10 + 0 + 1 + 0}^{\hat{m}(0,1)} = 7.$$

The interaction term coefficient 2 (in $2d_1d_2$) represents the mean wage difference associated with higher education in the high-experience subpopulation minus the mean wage difference associated with higher education in the low-experience subpopulation. This is the difference between the last two results above ($7 - 5 = 2$). Compared with the low-experience group, the high-experience group has a larger mean wage gap between college and no-college individuals.

9.4 Causality: Structural Identification by Exogeneity

Imagine Y is determined by the structural model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + U. \tag{9.27}$$

The qualitative condition for identification is the same as in Section 6.5. Specifically, if U (which contains other causal determinants of Y) is unrelated to the regressors, then the structural parameters are identified.

Mathematically, one sufficient definition of “unrelated” here is “uncorrelated.” If

$$\text{Cov}(U, X_1) = \text{Cov}(U, X_2) = \text{Cov}(U, X_1X_2) = 0, \tag{9.28}$$

then β_1, β_2, and β_3 are the linear projection slope coefficients from $\text{LP}(Y \mid 1, X_1, X_2, X_1X_2)$. Other mathematical definitions of “unrelated” imply (9.28) and are thus sufficient for identification. For example, $U \perp\!\!\!\perp (X_1, X_2)$ logically implies (9.28), as does mean independence $\text{E}(U \mid X_1, X_2) = \text{E}(U)$.

If the structural β_1, β_2, and β_3 are also linear projection coefficients, then they can be estimated by OLS. That is, we can interpret the OLS-estimated slope coefficients as the structural parameters in (9.27).

9.5 Causality: Identification by Conditional Independence

By extending the independence assumption (A6.2), variants of the ASE and ATE can be identified. The ASE argument applies to more realistic (more complex) models than in Section 9.4, but we focus on the ATE here. (Note: more details and examples are in the Spring 2020 edition.)

Consider the subpopulation with $X_2 = 1$, and whether the mean difference $\text{E}(Y \mid X_1 = 1, X_2 = 1) - \text{E}(Y \mid X_1 = 0, X_2 = 1)$ has a causal interpretation. This is equivalent to redefining the population as everybody with $X_2 = 1$ and asking if the mean difference

$\mathrm{E}(Y \mid X_1 = 1) - \mathrm{E}(Y \mid X_1 = 0)$ has a causal interpretation. This question was studied in Sections 6.4 and 6.5, for both structural and potential outcomes models.

Extending the independence assumption from Sections 6.4 and 6.5, **conditional independence** assumes independence within each subpopulation ($X_2 = 1$ and $X_2 = 0$). The conditional independence assumption has other names like **unconfoundedness**, **selection on observables**, and **ignorability**; see Imbens and Wooldridge (2007, p. 6) and references therein. Mathematically, both structural and potential outcomes versions of conditional independence are stated in Assumption A9.1.

Assumption A9.1 (conditional independence assumption)**.** The binary treatment X_1 is independent of the potential outcomes Y^T and Y^C, conditional on the control variable X_2: $(Y^T, Y^C) \perp\!\!\!\perp X_1 \mid X_2$. More generally, X_2 may be replaced by multiple control variables, $X_2, X_3, X_4, \ldots$.

Given A9.1, the ATE within subpopulation $X_2 = 1$ is identified and equal to the conditional mean difference $\mathrm{E}(Y \mid X_1 = 1, X_2 = 1) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 1)$. Similarly, the ATE within subpopulation $X_2 = 0$ is identified and equal to the conditional mean difference $\mathrm{E}(Y \mid X_1 = 1, X_2 = 0) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 0)$. The ATE for the full population averages these two conditional ATEs, weighted by $\mathrm{P}(X_2 = 1)$:

$$\begin{aligned}\mathrm{ATE} = {} & \mathrm{P}(X_2 = 1)[\mathrm{E}(Y \mid X_1 = 1, X_2 = 1) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 1)] \\ & + \mathrm{P}(X_2 = 0)[\mathrm{E}(Y \mid X_1 = 1, X_2 = 0) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 0)].\end{aligned}$$

9.6 Collider Bias

Although OVB shows the risk of omitting certain types of variables, other types of variables actually *should* be omitted, otherwise they cause a different type of (asymptotic) bias.

A **collider** or **common outcome** is a variable on which both X and Y have a causal effect. For example, imagine you want to learn the effect of a firm's ownership structure (say $X = 1$ for family-owned, $X = 0$ otherwise) on its research and development expenditure Y. Both X and Y affect the firm's performance Z, so Z is a collider.

Including a collider as a regressor causes **collider bias** when estimating a causal relationship. This is not as intuitive as OVB, but consider the following example.[1]

Imagine you're interested in the causal effect of eating falafel or salad on having the flu (which is zero effect), and you have a sample of 200 individuals. You randomly assigned 100 people to eat falafel for lunch, and 100 salad; a few hours later, you test each for flu (assume there is no testing error). Let $Y = 1$ if somebody has the flu (otherwise $Y = 0$), and $X = 1$ if somebody ate falafel for lunch ($X = 0$ if salad). Let $Z = 1$ if the individual has a fever (otherwise $Z = 0$). Sadly, the salad had some romaine contaminated with E. coli, so 40% of those who ate salad got a fever from the E. coli, unrelated to whether or not they had the flu. Among individuals with flu, 90% have a fever, but 10% don't.

[1] Modified from `https://doi.org/10.1093/ije/dyp334`

Table 9.1: Counts in falafel/salad/flu example.

			Fever		No fever	
	Flu	No flu	Flu	No flu	Flu	No flu
Falafel	50	50	45	0	5	50
Salad	50	50	47	20	3	30

Table 9.1 shows the number of individuals in different categories. Overall, there is no relationship between lunch and flu, so the flu rate is the same in the falafel and salad groups. To make the numbers easier, the overall flu rate is 50% (100/200 overall, 50/100 in each group). Because nobody who ate falafel got E. coli, the only reason for fever is the flu, which has a 90% fever rate. Thus, among the 50 with flu who at falafel, $(50)(0.9) = 45$ have a fever and 5 do not. This entirely explains the Falafel row. In the salad row, given the statistical independence of flu (probability 0.5) and E. coli (probability 0.4), the probability of having neither is

$$\begin{aligned}\mathrm{P}(\text{not flu and not E. coli}) &= \mathrm{P}(\text{not flu})\,\mathrm{P}(\text{not E. coli}) = [1 - \overbrace{\mathrm{P}(\text{flu})}^{0.5}][1 - \overbrace{\mathrm{P}(\text{E. coli})}^{0.4}] \\ &= (0.5)(0.6) = 0.3,\end{aligned}$$

hence $(100)(0.3) = 30$ salad-eaters who have neither flu nor E. coli, and thus no fever. This explains the No fever / No flu entry of 30 in the Salad row. Similarly,

$$\begin{aligned}\mathrm{P}(\text{flu, not E. coli}) &= (0.5)(0.6) = 0.3 \quad \text{(30 people)},\\ \mathrm{P}(\text{flu, E. coli}) &= (0.5)(0.4) = 0.2 \quad \text{(20 people)},\\ \mathrm{P}(\text{not flu, E. coli}) &= (0.5)(0.4) = 0.2 \quad \text{(20 people)}.\end{aligned}$$

The "not flu and E. coli" are the 20 individuals who have a fever (from the E. coli) but not flu. The 20 with both flu and E. coli all have a fever, due to E. coli. Among the 30 with flu but not E. coli, 90% have a fever, i.e., $(30)(0.9) = 27$ have a fever, so 3 do not. This 3 is the No fever / Flu entry in the Salad row. The 27 combine with the 20 who had both illnesses to make 47 who have both flu and a fever in the Salad row.

If we regress Y (flu) on X (food), then we correctly estimate zero effect, but if we also use Z (fever), then we incorrectly estimate a non-zero effect. If we only look at the "no fever" group, then there is (appropriately) zero difference: the flu rate for the falafel eaters is $5/55 = 1/11$, identical to the $3/33 = 1/11$ for the salad eaters. Mathematically, these "rates" are estimates of the conditional mean of the binary Y flu variable; e.g., $5/55 = \hat{\mathrm{E}}(Y \mid \text{falafel, no fever})$, recalling $\mathrm{E}(Y) = \mathrm{P}(Y = 1)$ for binary Y. However, if we also look at the "fever" group, the flu rate is much higher in the falafel group. In fact, the falafel group's flu rate is $45/45 = 100\%$, whereas the salad group's flu rate is only

$47/(47+20) = 70\%$, substantially lower. Mathematically,

$$\begin{aligned}&\overbrace{\hat{\mathrm{E}}(Y \mid X=1, Z=0)}^{5/55} - \overbrace{\hat{\mathrm{E}}(Y \mid X=0, Z=0)}^{3/33} = 0,\\ &\overbrace{\hat{\mathrm{E}}(Y \mid X=1, Z=1)}^{45/45} - \overbrace{\hat{\mathrm{E}}(Y \mid X=0, Z=1)}^{47/67} = 0.30.\end{aligned} \tag{9.29}$$

This suggests eating falafel causes flu, but this incorrect conclusion is entirely collider bias.

9.7 Causal Identification: Difference-in-Differences

⟹ Kaplan video: Diff-in-Diff Intuition

If X_1 is a treatment indicator and X_2 is a time period indicator, then the fully saturated model with two binary regressors is called a **difference-in-differences** (diff-in-diff) model. This is a special case of (9.13), whose coefficients were interpreted in Section 9.3.

Below, the parameter β_3 from (9.13) is shown to have a certain causal interpretation under certain conditions.

The general setup is that some individuals (or firms, or cities, etc.) were exposed to some "treatment," like a training program or law or other policy. The treatment wasn't randomized, but there's a group of untreated individuals whose outcomes can be used to form a **counterfactual**: what's the mean outcome of treated individuals in the parallel universe where they weren't treated?

Such setups are sometimes called **natural experiments** or **quasi-experiments** (see also Section 4.3.2). Because they weren't fully randomized experiments, it's invalid to simply compare treated and untreated outcomes, as seen in Section 9.7.1. However, there is enough randomness that a valid comparison can be found, with some additional work (like diff-in-diff).

Example 9.2 (Kaplan video)**.** Let Y be annual labor income, and we are interested in the effect of minimum wage. Imagine our city recently implemented a large minimum wage increase. The goal is to learn the effect of this particular minimum wage increase on Y (income), for individuals in our city. Notationally, $X_1 = 1$ if the individual lives in our city (and $X_1 = 0$ otherwise), and $X_2 = 1$ if the observation is from the year after the minimum wage increase (and $X_2 = 0$ if before the increase).

Notationally, $X_1 = 1$ is the "treated group" and $X_1 = 0$ the "untreated group"; $X_2 = 0$ is the time period "before" treatment and $X_2 = 1$ is "after."

9.7.1 Bad Approaches

One bad approach is to use only data from the treated group, comparing before and after. That is, we could try to estimate $\mathrm{E}(Y \mid X_2 = 1, X_1 = 1) - \mathrm{E}(Y \mid X_2 = 0, X_1 = 1)$. Part of this mean difference is indeed due to the effect of the treatment. However, there are almost always other important determinants of Y that change over time. In that case, there is omitted variable bias: the mean difference is a combination of the treatment effect plus many other effects, so it is wrong to interpret the mean difference as only the effect of the treatment.

Example 9.3 (Kaplan video)**.** Continuing the minimum wage example (Example 9.2), one bad approach is to use only data from our city, before and after the minimum wage increase. Coincidentally, there may have been a national (or global) recession right after the minimum wage law was passed. This may make everybody's income lower in the year after. It would look like the minimum wage hurt incomes, but really it was the recession. Alternatively, there may have been great national (macroeconomic) conditions that made incomes go up, which would make us incorrectly conclude that the law increased incomes greatly.

Another bad approach is to use only data from the "after" period, comparing the treated group to an untreated group. That is, we could try to estimate $\mathrm{E}(Y \mid X_2 = 1, X_1 = 1) - \mathrm{E}(Y \mid X_2 = 1, X_1 = 0)$. Part of this mean difference is indeed due to the effect of the treatment. However, there are almost always other important determinants of Y that differ between the treated and untreated groups. In that case, there is again omitted variable bias.

Example 9.4 (Kaplan video)**.** Again continuing the minimum wage example (Example 9.2), this bad approach compares incomes in our city and another city, in the year after our law passed. By using the other city as a sort of control group, we avoid the problem of misinterpreting macroeconomic changes as treatment effects. However, it's hard to know which other city to pick. We could pick one that has the same population, for example, but our city may still have much higher (or lower) income for reasons other than our minimum wage. For example, San Francisco and Columbus, OH have very similar populations, but they have (and have for a while had) very different incomes. If San Francisco happens to have a higher minimum wage, it is wrong to attribute the entire mean difference in income as a causal effect of their higher minimum wage. There may indeed be a minimum wage effect, but it's mixed with the effects of education, industry types, geography, etc.

Discussion Question 9.3 (bad panel approach #1, for Mariel boatlift)**.** Consider the basic setup from Card (1990). Due to a seemingly random/exogenous political decision, Cubans were temporarily permitted to immigrate to the U.S. for a few months in 1980. About half settled in Miami, FL, while the other half went to live in other cities around the U.S. We could compare wages of native-born workers in Miami in 1979 (before boatlift) and 1981 (after). Explain why this change in average wage would not be a good estimate of the average treatment effect of the Mariel boatlift on native worker wage. (Hint: are 1979 Miami and 1981 Miami the same except for how many Cubans live there, or might something else have changed?)

Discussion Question 9.4 (bad panel approach #2, for Mariel boatlift)**.** Consider the same setup as in DQ 9.3. But now compare 1981 wages of native workers in Miami and Houston, TX, a city that did not receive a large influx of Cuban immigrants in 1980. Explain why this difference (Miami minus Houston) in average wage would not be a good estimate of the average treatment effect of the Mariel boatlift on native worker wage. (Hint: are 1981 Miami and Houston the same except for how many Cubans live there, or might there be other differences between the cities that might cause omitted variable bias?)

Practice 9.2 (bad panel approach #1, for fracking)**.** Practices 9.2 and 9.3 are based loosely on the setting of Street (2018), who uses much better approaches. For counties in North Dakota, let Y denote crime rate. Consider the average crime rate in counties that started fracking activity, before and after the fracking started. (Fracking was a new technology that allowed extraction of certain underground oil and natural gas reserves that were previously infeasible or unprofitable to extract.) Explain why this change in average crime rate would not be a good estimate of the average treatment effect of the fracking activity on crime rate.

Practice 9.3 (bad panel approach #2, for fracking)**.** Consider the same setup as in Practice 9.2, but now compare the "after" crime rates in North Dakota counties with fracking to those without fracking. Explain why this difference (fracking minus non-fracking) in average crime rate would not be a good estimate of the average treatment effect of fracking on crime rate.

9.7.2 Counterfactuals and Parallel Trends

The difference-in-differences idea is to combine the before vs. after comparison with the treated vs. untreated comparison.

Conceptually, the goal is to construct a **counterfactual** (link to pronunciation), like what our city's mean income would have been if there were not a minimum wage increase. Thinking of the potential outcomes framework, the counterfactual is the parallel universe where the treatment never happened.

The key identifying assumption is called **parallel trends**. Conceptually, in the running example, parallel trends says that without the minimum wage law, our city's mean income would have increased by exactly the same amount as the other city's mean income. Mathematically, with $m(x_1, x_2) \equiv \mathrm{E}(Y \mid X_1 = x_1, X_2 = x_2)$, the other city's mean income increase (i.e., "after" minus "before") is

$$m(0,1) - m(0,0) = \mathrm{E}(Y \mid X_1 = 0, X_2 = 1) - \mathrm{E}(Y \mid X_1 = 0, X_2 = 0). \tag{9.30}$$

Parallel trends assumes that adding this increase to the "before" mean income in our city, $m(1,0) = \mathrm{E}(Y \mid X_1 = 1, X_2 = 0)$, gives us the counterfactual income for our city in the "after" time period.

Given parallel trends, we can learn about causality by comparing

$$\begin{aligned}&\overbrace{\mathrm{E}(Y \mid X_1=1, X_2=1)}^{\text{actual (our city, after)}} \quad \text{vs.} \\ &\overbrace{\underbrace{\mathrm{E}(Y \mid X_1=1, X_2=0)}_{\text{our city, before}} + \underbrace{\mathrm{E}(Y \mid X_1=0, X_2=1) - \mathrm{E}(Y \mid X_1=0, X_2=0)}_{\text{increase in other city over time}}}^{\text{counterfactual}}.\end{aligned} \tag{9.31}$$

$$\overbrace{m(1,1)}^{\text{actual}} - \overbrace{\{m(1,0) + [m(0,1) - m(0,0)]\}}^{\text{counterfactual}} = \overbrace{[m(1,1) - m(1,0)] - [m(0,1) - m(0,0)]}^{\beta_3 \text{ in (9.13)}}.$$

Figure 9.1 visualizes this effect. We can think of constructing the counterfactual outcome, and then subtracting it from the actual outcome $m(1,1)$, or we can think of taking the before/after difference for our city, $m(1,1) - m(1,0)$, and subtracting off the before/after difference in the other city, $m(0,1) - m(0,0)$.

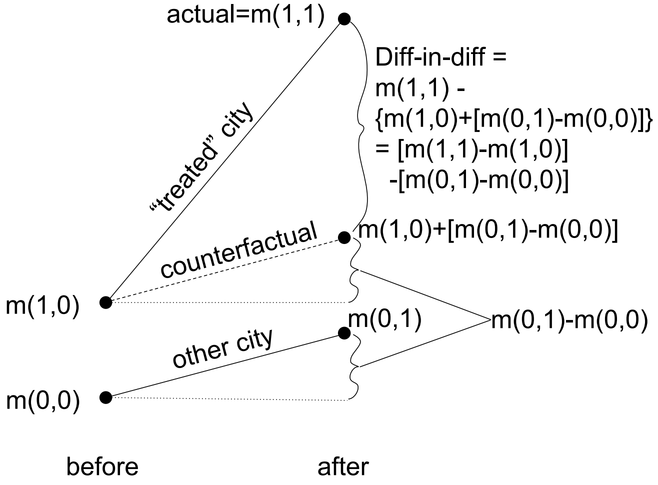


Figure 9.1: Difference-in-differences.

9.7.3 Identification

Population Object of Interest: ATT

Most fundamentally, the difference-in-differences approach only learns the average treatment effect for the group that was actually treated (in our universe). This is called the **average treatment effect on the treated** (ATT) (or sometimes ATTE or ATET). Mathematically, ATE meant $\mathrm{E}(Y^1 - Y^0)$, where Y^1 and Y^0 are the treated and untreated potential outcomes, respectively (previously Y^T and Y^C). ATT is the same, but for the subpopulation who was actually treated in our universe. Because $X_1 = 1$ if somebody is actually treated, the ATT is

$$\text{ATT} \equiv \text{E}(Y^1 - Y^0 \mid X_1 = 1). \tag{9.32}$$

It's possible but uncommon that ATT = ATE. For example, maybe there are different demographics in our city than the comparison city, or different levels of unionization, or different other labor laws, or different industry mix, so the minimum wage effect is different in our city ($X_1 = 1$) than elsewhere. (This is essentially a question of external validity; see Chapter 12.) Self-selection is another major reason the ATT may differ from the ATE. Whether it's a city choosing which laws to adopt, or an individual choosing which exercise routine to use, the choice is often made based on the anticipated benefit of the "treatment," so often those who get treated are those who most benefit, making ATT > ATE.

Example 9.5 (Kaplan video)**.** Consider the "treatment" of a small business receiving a loan, and the outcome of monthly sales revenue. Although there are certainly other factors, economic theory suggests that the small businesses applying for loans tend to be the ones that would most benefit from loans. Thus, we'd guess that the ATT (the effect on small businesses who in reality got a loan) is probably higher than the overall ATE that includes businesses who did not apply for loans. That is, because small businesses can (partially) self-select into treatment depending on their benefit from the treatment, the benefit is probably higher among the actually-treated businesses.

Identification of ATT

Parallel trends is sufficient to identify the counterfactual. In potential outcomes notation, "parallel trends" is

$$\begin{aligned}&\text{E}(Y^0 \mid X_1 = 1, X_2 = 1) - \text{E}(Y^0 \mid X_1 = 1, X_2 = 0)\\ &= \text{E}(Y^0 \mid X_1 = 0, X_2 = 1) - \text{E}(Y^0 \mid X_1 = 0, X_2 = 0).\end{aligned} \tag{9.33}$$

That is, the mean untreated potential outcome changes over time ($X_2 = 0$ to $X_2 = 1$) by the same amount in the treated ($X_1 = 1$) and untreated ($X_1 = 0$) groups. The term $\text{E}(Y^0 \mid X_1 = 1, X_2 = 1)$ is the counterfactual, like our city's mean wage in the "after" period in the parallel universe where minimum wage never increased. In the other three terms, $Y^0 = Y$, i.e., the untreated Y^0 is the observed Y. Only when $X_1 = X_2 = 1$ is the treated Y^1 observed, $Y = Y^1$. Thus, the counterfactual can be written uniquely in terms of the joint distribution of (Y, X_1, X_2):

$$\begin{aligned}&\text{E}(Y^0 \mid X_1 = 1, X_2 = 1)\\ &= \text{E}(Y^0 \mid X_1 = 1, X_2 = 0) + [\text{E}(Y^0 \mid X_1 = 0, X_2 = 1) - \text{E}(Y^0 \mid X_1 = 0, X_2 = 0)]\\ &= \text{E}(Y \mid X_1 = 1, X_2 = 0) + [\text{E}(Y \mid X_1 = 0, X_2 = 1) - \text{E}(Y \mid X_1 = 0, X_2 = 0)]\\ &= m(1, 0) + [m(0, 1) - m(0, 0)].\end{aligned} \tag{9.34}$$

Because the counterfactual is identified, so is the ATT. Specifically, the ATT equals β_3 in the fully saturated CMF model (9.13),

$$\begin{aligned}
\text{ATT} &= \mathrm{E}(Y^1 - Y^0 \mid X_1 = 1, X_2 = 1) \\
&= \overbrace{\mathrm{E}(Y^1 \mid X_1 = 1, X_2 = 1)}^{Y^1 = Y \text{ since } X_1=1, X_2=1} - \overbrace{\mathrm{E}(Y^0 \mid X_1 = 1, X_2 = 1)}^{\text{use counterfactual from (9.34)}} \\
&= \mathrm{E}(Y \mid X_1 = 1, X_2 = 1) - \overbrace{\{m(1,0) + [m(0,1) - m(0,0)]\}}^{\text{counterfactual}} \\
&= m(1,1) - \{m(1,0) + [m(0,1) - m(0,0)]\} \\
&= [m(1,1) - m(1,0)] - [m(0,1) - m(0,0)] \\
&= \beta_3.
\end{aligned}$$

Skepticism About Parallel Trends

In practice, the parallel trends condition may not hold for various reasons. For example, maybe our city was experiencing fast wage growth, whereas the comparison city was declining (maybe due to reliance on different industries). Maybe our city passed the minimum wage law partly because everybody's wages were increasing anyway. In that case, we can't tell whether our city's wages grew more than the other city's wages because of the minimum wage, or because of other factors (our industries were growing, theirs were declining, etc.).

Parallel trends is also a bit fragile because nonlinear functions of Y change whether it's true or not. For example, if there are parallel trends when Y is wage, then there are not parallel trends for log-wage $\ln(Y)$. Similarly, if there are parallel log-wage trends, then the wage trends cannot be parallel.

In the data, you can try to see if parallel trends seems plausible, but it is not directly testable. Specifically, "pre-trend analysis" compares trends for a few periods before the treatment takes place. But even if the trends were parallel before, it does not mean for sure that the trends would have remained parallel after the treatment year. We can never know because the "trend" refers to the treated group's untreated potential outcomes, which by definition are not observed. So, there is no empirical test that can replace careful critical thought.

Discussion Question 9.5 (parallel trends skepticism)**.** Consider U.S. state traffic fatality (i.e., car accident death) rates (Y), where the year 1980 is "before" ($X_2 = 0$) and 1990 is "after" ($X_2 = 1$). Consider states that adopt a 0.08 blood alcohol content (BAC) limit law sometime between 1980 and 1990 ($X_1 = 1$) and states that never have such a law ($X_1 = 0$). Explain why you might doubt the parallel trends assumption. Hint #1: is a BAC law the only way states try to reduce fatal accidents? Hint #2: this is more difficult than simply thinking of an omitted variable that would cause OVB in a cross-sectional regression, because parallel trends allows certain types of such omitted variables.

9.7.4 Extensions

There are many interesting extensions of the basic diff-in-diff idea, although all are beyond our scope. For example, there are related models that allow additional regressors, or more time periods, or quantile treatment effects.

9.8 Estimation and Inference

⟹ Kaplan video: Difference-in-Differences Example

Because (9.13) is just a special case of a regression model, standard regression techniques and R functions can be used. For estimation, OLS consistently estimates each β_j under fairly general conditions; remember to use sample/survey weights if they are available in the data. The same heteroskedasticity-robust methods from earlier (like Section 7.7.3) can be used to compute confidence intervals if sampling is iid.

The following code shows different R syntax to get the same coefficient estimates, with simulated data. The notation `X1:X2` is the interaction term (or in the output, its coefficient). Heteroskedasticity-robust CIs are also reported.

```
library(sandwich); library(lmtest)
n <- 4*8
set.seed(112358)
m00 <- 10; m10 <- 15; m01 <- 16; m11 <- 25
df <- data.frame(X1=c(rep(0,n/2),rep(1,n/2)),
                 X2=rep(rep(0:1,each=n/4),times=2))
df$Y <- c(rep(m00,n/4),rep(m01,n/4),
          rep(m10,n/4),rep(m11,n/4) ) + rnorm(n)
# Three equivalent estimates
ret1 <- lm(Y~X1*X2, data=df)
ret2 <- lm(Y~X1+X2+X1:X2, data=df)
df$Xint <- df$X1*df$X2
ret3 <- lm(Y~X1+X2+Xint, data=df)
TrueBetas <- c(m00,m10-m00,m01-m00,(m11-m01)-(m10-m00))
retmat <- rbind(coef(ret1),coef(ret2),coef(ret3),TrueBetas)
rownames(retmat) <- c('est1','est2','est3','true')
print(round(retmat, digits=2))

##      (Intercept)   X1   X2 X1:X2
## est1          10 5.17 6.12  3.73
## est2          10 5.17 6.12  3.73
## est3          10 5.17 6.12  3.73
## true          10 5.00 6.00  4.00
```

```
round(coefci(ret1, vcov.=vcovHC(ret1,type='HC1')),digits=2)

##             2.5 % 97.5 %
## (Intercept)  9.30  10.77
## X1           4.31   6.04
## X2           4.97   7.27
## X1:X2        2.20   5.27
```

Optional Resources

Optional resources for this chapter

- ATT (Masten video)
- Potential outcomes and CATE (Masten video)
- OVB/confounders (Masten video)
- conditional independence/unconfoundedness (Masten video)
- ATE/conditional independence example (Masten video)
- Difference-in-differences (Masten video)
- Parallel trends (Masten video)
- Diff-in-diff example: immigration and unemployment (Masten videos)
- Parallel trends example: immigration and unemployment (Masten videos)
- Diff-in-diff example: minimum wage (Masten video)
- Diff-in-diff example: posting calorie counts (Masten video)
- OVB example: test score and class size (Lambert video)
- OVB example: wages and education (Lambert video)
- Sections 3.3 ("Ceteris Paribus Interpretation and Omitted Variable Bias") and 6.1.5 ("Interaction Terms") in Heiss (2016)
- Section 13.2 ("Difference-in-Differences") in Heiss (2016)
- Collider bias examples: `https://doi.org/10.1093/ije/dyp334`
- Collider bias review (very detailed): `https://doi.org/10.1146/annurev-soc-071913-043455`
- Sections 6.1 ("Omitted Variable Bias") and 8.3 ("Interactions Between Independent Variables") in Hanck et al. (2018)

Empirical Exercises

Empirical Exercise EE9.1. You will analyze data on driving laws and fatal accident rates, originally from Freeman (2007). In particular, you'll compare weekend driving fatality (death) rates for states that adopted a 0.08 blood alcohol content (BAC) law and states that didn't, comparing rates before and after the law adoption. Standard errors can be smaller if the full dataset is used, but such methods are beyond our scope. Either way, the difference-in-differences approach is probably not identifying a treatment effect: probably states that adopted such laws also adopted other ways to discourage drunk driving, whether official laws or just changing cultural norms. This violates the parallel trends assumption.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

   ```
   library(wooldridge); library(sandwich); library(lmtest)
   ?driving
   ```

b. Stata only: load the data with

   ```
   use https://raw.githubusercontent.com/kaplandm/stata/main/data/
       driving.dta , clear
   ```

c. Keep only years 1980 and 1990.

   R: `df <- driving[driving$year==1980 | driving$year==1990 , ]`

   Stata: `keep if year==1980 | year==1990`

d. Create a dummy variable for the "after" period (year 1990).

   R: `df$after <- (df$year==1990)`

   Stata: `generate after = (year==1990)`

e. Create variable `bac` equal to `1` (or `TRUE`) if there's any BAC law that year.

   R: `df$bac <- (df$bac08+df$bac10>=1)`

   Stata: `generate bac = (bac08 + bac10 >= 1)`

f. Drop states that already had a BAC law in the "before" period (1980), leaving only states that never had the law or adopted it between 1980 and 1990.

   R: `dropst <- unique(df$state[(!df$after) & df$bac])` to get a list of the states to drop, and then remove them with `df <- df[!(df$state %in% dropst) , ]`

   Stata:

   ```
   generate dropflag = ((!after) & bac)
   bysort state : egen dropst = max(dropflag)
   drop if dropst
   ```

g. Create a treatment dummy equal to 1 for states that adopted a BAC law by 1990.

   R: `treatst <- unique(df$state[df$bac])` followed by `df$treat <- (df$state %in% treatst)`

   Stata: `bysort state : egen treat = max(bac)`

h. Run a difference-in-differences regression with the intercept, "after" dummy, treatment dummy, and interaction term. Below, the `*` in R and the `##` in Stata automatically generate the desired interaction term.

   R:

   ```
   ret <- lm(wkndfatrte~treat*after, data=df)
   coeftest(ret, vcov.=vcovHC(ret, type='HC1'))
   coefci(  ret, vcov.=vcovHC(ret, type='HC1'))
   ```

   Stata: `regress wkndfatrte treat##after , vce(robust)`

i. To see how the OLS coefficient estimates relate to the conditional means (CMF estimates), compute the sample mean weekend driving fatality rate within each of the four groups defined by the time period and "treatment" status.

   R: `(agg <- aggregate(wkndfatrte~treat*after, data=df, FUN=mean))`

   Stata: `tabulate treat after , summarize(wkndfatrte) means missing`

j. Display the CMF-based replication of the OLS estimates.

   R: `c(agg[1,3], agg[2,3]-agg[1,3], agg[3,3]-agg[1,3])` for the first three coefficient estimates and `c((agg[4,3]-agg[3,3])-(agg[2,3]-agg[1,3]), (agg[4,3]-agg[2,3])-(agg[3,3]-agg[1,3]))` to show both (equivalent) ways to compute the interaction coefficient estimate.

   Stata:

   ```
   collapse (mean) wkndfatrte , by(treat after)
   display wkndfatrte[1]
   display wkndfatrte[3]-wkndfatrte[1]
   display wkndfatrte[2]-wkndfatrte[1]
   display (wkndfatrte[4]-wkndfatrte[3])-(wkndfatrte[2]-wkndfatrte[1])
   ```

k. Optional: repeat part (h) but with a different outcome variable to replace `wkndfatrte`, like the weekend fatalities per 100 million miles driven (instead of population), or the total fatality rate (not just weekends), etc.

l. Optional: repeat parts (e)–(h) but replacing your `bac` treatment variable created in part (e) with a treatment dummy equal to 1 if `perse` (a different driving law) equals 1 (and equal to 0 otherwise).

Empirical Exercise EE9.2. You will analyze wage data for different types of individuals from the 1976 Current Population Survey (conducted by the U.S. Census Bureau). Specifically, you'll look at dummy variables for `nonwhite` (race) and `female`, as well as their interaction. The results are clearly not causal, but the interaction term shows (descriptively) the difference in the white/nonwhite wage gap for females compared to non-females, or (equivalently) the difference in the female/non-female wage gap for non-whites compared to whites.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```
library(wooldridge); library(sandwich); library(lmtest)
?wage1
```

b. Stata only: load the data with `bcuse wage1 , nodesc clear` (assuming `bcuse` is already installed)

c. Display the group mean wage for the four groups defined by the `nonwhite` and `female` dummy variables.

   R: `(agg <- aggregate(wage~nonwhite*female, data=wage1, FUN=mean))`

   Stata: `tabulate female nonwhite , summarize(wage) means missing`

d. Run a “difference-in-differences” type of regression with the intercept, non-white dummy, female dummy, and interaction term.

   R:

```
ret <- lm(wage~nonwhite*female, data=wage1)
coeftest(ret, vcov.=vcovHC(ret, type='HC1'))
coefci(  ret, vcov.=vcovHC(ret, type='HC1'))
```

   Stata: `regress wage female##nonwhite , vce(robust)`

e. Compute the OLS coefficient estimates manually from the four conditional means.

   R: store the conditional means with `m00 <- agg$wage[1]; m10 <- agg$wage[2]; m01 <- agg$wage[3]; m11 <- agg$wage[4]` and show that you can replicate the OLS estimates with `rbind(coef(ret), c(m00, m10-m00, m01-m00, (m11-m01) -(m10-m00)) )` and also note that `c( (m11-m01) - (m10-m00) , (m11-m10) - (m01-m00) )` shows the equivalence of the two interpretations of the interaction term coefficient.

   Stata: collapse the dataset to just the four conditional means with `collapse ( mean) wage , by(female nonwhite)` and then display the manually calculated coefficient estimates with

```
display wage[1]
display wage[3]-wage[1]
display wage[2]-wage[1]
display (wage[4]-wage[3])-(wage[2]-wage[1])
display (wage[4]-wage[2])-(wage[3]-wage[1])
```

f. Optional: repeat part (d) but using `south` instead of `female`

g. Optional: repeat part (d) again with any two dummy variables of your choice; you may use one from a previous analysis as long as it is combined with a different dummy. The dataset comes with many dummy variables already, like `nonwhite`, `female`, `south` (and other regions), `servocc` (and other occupational fields and industries), and `married`, or you can create your own. For example, you can generate a "more than high school education" dummy with R code `wage1$gtHS <- (wage1$educ>12)` or Stata command `generate gtHS = (educ>12)`

Chapter 10

Regression with Multiple Regressors

⟹ Kaplan video: Chapter Introduction

Allowing multiple regressors opens a multitude of combinations, especially when combined with nonlinear functions like in Chapter 8. Most of Chapter 10 focuses on the different functional forms themselves, with the different types of flexibility they do (and don't) allow. These discussions apply equally to descriptive, predictive, and causal models.

Unit learning objectives for this chapter

10.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

10.2. Assess in a real-world example whether there is bias from omitted variables and whether a linear model seems realistic [TLOs 2 and 6]

10.3. Describe and interpret models with multiple regressors, including those in which two variables interact [TLO 3]

10.4. Judge which assumptions seem true and which interpretation seems most appropriate for real-world regressions [TLOs 2 and 6]

10.5. In R (or Stata): estimate a regression with multiple variables, along with measures of uncertainty, and judge economic and statistical significance [TLO 7]

10.1 Causality: Omitted Variable Bias

One motivation for this chapter is that omitted variable bias (OVB, Section 9.1) can still be a problem even if we include two regressors. We may need to include three, or even 10 or 100 regressors to avoid OVB. But even with 100 regressors, OVB can still be a big problem.

Consider OVB with the linear structural model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + V. \tag{10.1}$$

For OLS to consistently estimate β_j for $j = 1, 2, 3$ (the slope coefficients) requires $\mathrm{Cov}(X_j, V) = 0$ for $j = 1, 2, 3$. Imagine this is true, but X_3 is omitted, so

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U, \quad U \equiv \beta_3 X_3 + V. \tag{10.2}$$

In (10.2), OLS consistency for β_1 and β_2 requires $\mathrm{Cov}(X_1, U) = \mathrm{Cov}(X_2, U) = 0$. Because

$$\mathrm{Cov}(X_j, U) = \beta_3 \mathrm{Cov}(X_j, X_3) + \mathrm{Cov}(X_j, V), \tag{10.3}$$

this requires either $\beta_3 = 0$ (i.e., X_3 is not a causal determinant of Y) or else $\mathrm{Cov}(X_1, X_3) = \mathrm{Cov}(X_2, X_3) = 0$.

There are other mathematical formulations, but they all make the point that even including 100 regressors is not sufficient to avoid OVB if there is still an important omitted variable. That is, even if (10.2) becomes

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + U, \quad U \equiv \gamma Q + V, \tag{10.4}$$

then we still have OVB if $\gamma \neq 0$ and any $\mathrm{Cov}(X_j, Q) \neq 0$.

That is, there is OVB if both of the following conditions hold.

OVB.1′ The omitted variable is correlated with an included regressor; in (10.4), $\mathrm{Corr}(X_j, Q) \neq 0$ for some j.

OVB.2′ The omitted variable Q is a causal determinant of Y; in (10.4), $\gamma \neq 0$.

Discussion Question 10.1 (OVB with multiple regressors)**.** Consider the example of California schools where Y is a school's average standardized math test score for 5th-graders, X_1 is the 5th-grade student-teacher ratio, and X_2 is the percentage of 5th-graders who are English learners (non-native speakers). Judge whether a school's total expenditures per student satisfies each of Conditions OVB.1′ and OVB.2′ for OVB, explaining why you came to that conclusions.

10.2 Linear-in-Variables Model

⟹ Kaplan video: Wage Regression Example

10.2.1 Model and Coefficient Interpretation

The linear-in-variables model and discussion from Section 9.2 naturally generalize to non-binary and/or more than two regressors. With J regressors $X_1, X_2, \dots, X_J$,

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J + U = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + U \equiv g(X_1, \dots, X_J) + U. \tag{10.5}$$

If U is a CMF error, then $g(\cdot)$ represents the CMF. However, the following discussion is essentially the same if U is a linear projection error and $g(\cdot)$ is the linear projection, or if the β_j have a causal interpretation.

Regardless of interpretation, the coefficient β_j shows how the function $g(\cdot)$ changes when X_j increases by one unit. This is true whether X_j is binary, discrete, or continuous. For example, X_1 only appears in the $\beta_1 X_1$ term, so if we change from $X_1 = x_1$ to $X_1 = x_1 + 1$ (unit increase), that term changes from $\beta_1 x_1$ to $\beta_1(x_1 + 1) = \beta_1 x_1 + \beta_1$, a change of β_1. That is, for any starting values $X_1 = x_1$, $X_2 = x_2$, etc., a unit increase in X_1 changes the function by

$$\begin{aligned}
&g(x_1 + 1, x_2, \dots, x_J) - g(x_1, x_2, \dots, x_J) \\
&= [\beta_0 + \beta_1(x_1 + 1) + \sum_{j=2}^{J} \beta_j x_j] - [\beta_0 + \beta_1 x_1 + \sum_{j=2}^{J} \beta_j x_j] = \beta_1(x_1 + 1 - x_1) = \beta_1.
\end{aligned} \tag{10.6}$$

Example 10.1 (Kaplan video)**.** Given a linear-in-variables model, if Y is wage in \$/hr, and X_1 is years of education, and $\beta_1 = (\$5/\text{hr})/\text{yr}$, then each additional year of education is associated with a (\$5/hr)/yr change, regardless of the initial education level or other variables like experience.

More generally, if X_1 changes by Δ_1 units, then the function's value changes by $\beta_1 \Delta_1$. Regardless of the starting values, if X_1 changes from x_1 to $x_1 + \Delta_1$, then similar to (10.6),

$$\begin{aligned}
&g(x_1 + \Delta_1, x_2, \dots, x_J) - g(x_1, x_2, \dots, x_J) \\
&= [\beta_0 + \beta_1(x_1 + \Delta_1) + \sum_{j=2}^{J} \beta_j x_j] - [\beta_0 + \beta_1 x_1 + \sum_{j=2}^{J} \beta_j x_j] = \beta_1(x_1 + \Delta_1 - x_1) = \beta_1 \Delta_1.
\end{aligned} \tag{10.7}$$

10.2.2 Limitations

While pleasingly simple, these formulas may not be realistic. That is, the change in Y may depend on not only Δ_1, but the starting value x_1, or other x_j.

Example 10.2 (Kaplan video)**.** Let Y be wage, X_1 years of experience, and X_2 years of education. Due to diminishing marginal benefits, perhaps the first years of experience are associated with bigger increases in mean wage than later years of experience. The wage increase associated with the change from $X_1 = 0$ to $X_1 = 1$ is probably larger than the increase from $X_1 = 40$ to $X_1 = 41$, even though $\Delta_1 = 1$ in both cases. Further, the change from $X_1 = 0$ to $X_1 = 1$ may be associated with a larger wage increase for highly educated individuals (large X_2) than for less-educated individuals. Mathematically, the change depending on the starting value of X_1 implies some nonlinearity in X_1, and the dependence on the value of X_2 implies some sort of interaction term(s).

Nonlinear and nonparametric functions of a single variable are discussed in Sections 8.2 and 8.3; interactions are discussed in Sections 9.3 and 10.3. Nonparametric models with multiple regressors are beyond our scope.

10.2.3 Code

The following code shows a simple linear-in-variables regression with real data. It should be interpreted as estimating the linear projection (or BLA/BLP), not anything causal (nor even the CMF). The outcome variable is log wage, and the three regressors are years of work experience, years of education, and a dummy for living in a "city" (metropolitan area). In the output, the row labeled `exper` is for the coefficient on experience, showing the OLS estimate (`Estimate`) and corresponding heteroskedasticity-robust 95% confidence interval (lower endpoint under `2.5 %`, upper endpoint under `97.5 %`). Similarly for the `(Intercept)` and the other regressors' coefficients. Because this is a log-linear regression, the coefficients can be interpreted as approximate percentages. Approximately: a one-year increase in experience is associated with a 1.6% increase in wage; a one-year increase in education is associated with a 10.7% increase in wage; and living in a metropolitan area is associated with having a 6% higher wage than living in a more rural area. The confidence intervals show there is some statistical uncertainty about each of these estimates, especially for `city`.

```
library(sandwich); library(lmtest); library(wooldridge)
ret <- lm(log(wage)~exper+educ+city, data=mroz)
retVC1 <- vcovHC(ret, type="HC1")
round(cbind(Estimate=coef(ret), coefci(ret, vcov. = retVC1)), digits=3)

##             Estimate  2.5 % 97.5 %
## (Intercept)   -0.412 -0.776 -0.048
## exper          0.016  0.008  0.024
## educ           0.107  0.082  0.133
## city           0.060 -0.068  0.188
```

10.3 Interaction Terms

⟹ Kaplan video: Interaction Model

⟹ Kaplan video: Wage Regression Example (again)

To start, imagine there are two regressors, one of which is binary. To help us remember which is which, let D (for "dummy") be the binary regressor ($D = 1$ or $D = 0$) and X the other regressor. Assume X is the regressor of interest.

10.3.1 Limitation of Linear-in-Variables Model

With a linear-in-variables model,

$$Y = g(X, D) + U, \quad g(X, D) = \beta_0 + \beta_1 X + \beta_2 D. \tag{10.8}$$

A unit increase in X always changes the function $g(X, D)$ by β_1 units, regardless of the starting value of X or the value of D. As discussed in Section 10.2, this is often unrealistic.

Because D has only two possible values, we can plug them each into $g(X, D)$:

$$g(X, 0) = \beta_0 + \beta_1 X, \tag{10.9}$$

$$g(X, 1) = \beta_0 + \beta_1 X + (\beta_2)(1) = \overbrace{(\beta_0 + \beta_2)}^{\text{intercept}} + \beta_1 X. \tag{10.10}$$

These are two functions of X: one when $D = 0$, one when $D = 1$. They have the same slope (β_1) but different intercepts (β_0 and $\beta_0 + \beta_2$).

10.3.2 Interpretation of Interaction Term

To allow both the intercept and slope to differ between $g(X, 0)$ and $g(X, 1)$, an **interaction term** can be used, specifically the product DX. Mathematically, adding this term to (10.8),

$$g(X, D) = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX. \tag{10.11}$$

The function in (10.11) is more general because setting $\beta_3 = 0$ yields (10.8). Given (10.11), instead of (10.9) and (10.10),

$$g(X, 0) = \beta_0 + \beta_1 X + \overbrace{(\beta_2)(0) + (\beta_3)(0)(X)}^{=0} = \beta_0 + \beta_1 X, \tag{10.12}$$

$$g(X, 1) = \beta_0 + \beta_1 X + (\beta_2)(1) + (\beta_3)(1)(X) = \overbrace{(\beta_0 + \beta_2)}^{\text{intercept}} + \overbrace{(\beta_1 + \beta_3)}^{\text{slope}} X. \tag{10.13}$$

Now, the slope differs (by β_3), too. Just as $\beta_2 > 0$, $\beta_2 < 0$, and $\beta_2 = 0$ are all possible, so are $\beta_3 > 0$, $\beta_3 < 0$, and $\beta_3 = 0$.

Figure 10.1 illustrates the interpretation of the function from (10.11). In the figure's example, $\beta_2 > 0$ and $\beta_3 > 0$. Omitting the interaction term is equivalent to assuming $\beta_3 = 0$, in which case the two lines would be parallel (same slope).

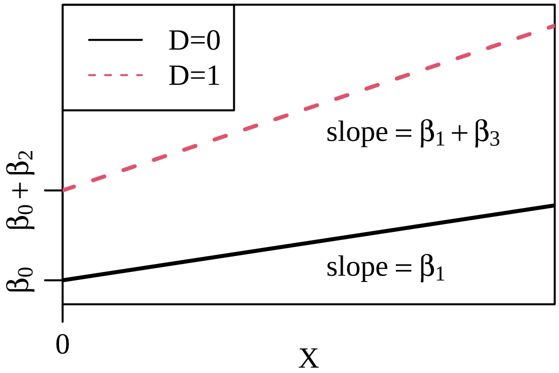


Figure 10.1: Visualization of $\beta_0 + \beta_1 X + \beta_2 D + \beta_3 DX$.

If you're interested in D, don't only look at β_2. Rearranging (10.11),

$$g(X, D) = (\beta_0 + \beta_1 X) + D(\beta_2 + \beta_3 X), \tag{10.14}$$

so the slope coefficient on D is $\beta_2 + \beta_3 X$. For example, even if $\beta_2 = -2$, the slope $\beta_2 + \beta_3 X$ is positive if $\beta_3 X > 2$. The opposite is also possible, e.g., if $\beta_2 = 5$, $\beta_3 = -1$, and $X > 5$: then $\beta_2 > 0$, but the slope is negative, $\beta_2 + \beta_3 X < 0$.

Example 10.3 (Kaplan video)**.** Let Y be commute time in minutes, $D = 1$ if somebody walks to work (and $D = 0$ if drive), and X the distance in kilometers from the person's home to their work. Consider a linear projection with the form of (10.11). Imagine a suburban setting (not crazy traffic). For the slope: if X (distance) increases by one unit (1 km), then Y (commute time) increases more when walking ($D = 1$) than driving ($D = 0$). Thus, the slope $\beta_1 + \beta_3$ should be higher than the slope β_1, and both are positive (more distance X, more time Y), so $\beta_1 > 0$ and $\beta_3 > 0$. For the intercept: walking is always slower than driving, so we might think the $D = 1$ line is shifted up from the $D = 0$ line, meaning $\beta_2 > 0$; however, if $X = 0$ (work at home), then $Y = 0$ regardless of walking or driving, so it also seems that both intercepts should be zero ($\beta_0 = 0$, $\beta_0 + \beta_2 = 0$). For the CMF, certainly the intercept is zero. However, the linear projection is only an approximation of the CMF, which is probably nonlinear for the $D = 0$ (driving) subpopulation: driving longer distances (X) usually allows you to drive on more major roads/freeways with higher speeds, so the CMF for $D = 0$ should be steeper near $X = 0$ and flatter at larger X. This could make the linear projection intercept $\beta_0 > 0$, even if the walking linear projection intercept is zero. Or if walking farther distances means slower average speed, then the $D = 1$ CMF is increasing and convex, and its linear projection intercept would actually be negative.

Discussion Question 10.2 (sleep and interactions)**.** Let Y be a person's hours of sleep per night, X the person's age, and $D = 1$ if the person lives in the same house as children under 8 years old (and $D = 0$ if not). Consider a linear projection with the form of (10.11). Answer any two of the following parts.

a) What would you guess for the sign (+, −, or zero) of β_1? Explain why.
b) What would you guess for the sign (+, −, or zero) of β_2? Explain why.
c) What would you guess for the sign (+, −, or zero) of β_3? Explain why.
d) Given the same set of regressors (X, D), describe another nonlinear term (i.e., another function of X and/or D, besides X, D, and DX) that would improve the CMF estimate, and why you think that term would help.

10.3.3 Non-Binary Interaction

Discussion Question 10.3 (linear-in-variables?)**.** Let Y be log wage, X_1 years of education, and X_2 years of experience. Consider possible linear-in-variables CMF model $\mathrm{E}(Y \mid X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
a) Explain one reason you think this CMF model is misspecified (wrong).
b) How do you think the true CMF (not the misspecified linear-in-variables CMF) slope with respect to experience might differ for different values of education? (Hint: draw a graph with different lines like $\mathrm{E}(Y \mid X_1 = 12, X_2 = x_2)$, where you fix the X_1 value and then graph the CMF as a function of only x_2.)

Even if neither regressor were binary, an interaction term allows the slopes to depend on the other regressor's value. Replacing $X = X_1$ and $D = X_2$ in (10.11),

$$g(X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2. \tag{10.15}$$

Consider the slope of $g(X_1, X_2)$ with respect to X_1, at different values of X_2. Generally, rearranging (10.15) as a function of X_1,

$$g(X_1, X_2) = \overbrace{(\beta_0 + \beta_2 X_2)}^{\text{intercept}} + \overbrace{(\beta_1 + X_2\beta_3)}^{\text{slope}} X_1. \tag{10.16}$$

Plugging values $X_2 = a$ and $X_2 = b$ into (10.16), similar to (10.12) and (10.13),

$$g(X_1, a) = \overbrace{(\beta_0 + a\beta_2)}^{\text{intercept}} + \overbrace{(\beta_1 + a\beta_3)}^{\text{slope}} X_1, \tag{10.17}$$

$$g(X_1, b) = \overbrace{(\beta_0 + b\beta_2)}^{\text{intercept}} + \overbrace{(\beta_1 + b\beta_3)}^{\text{slope}} X_1. \tag{10.18}$$

Changing X_2 from a to b changes the intercept from $\beta_0 + a\beta_2$ to $\beta_0 + b\beta_2$, and it changes the slope from $\beta_1 + a\beta_3$ to $\beta_1 + b\beta_3$. Alternatively, we could plug in $X_1 = a$ and $X_1 = b$ and consider $g(a, X_2)$ and $g(b, X_2)$ as functions of X_2, where again both the intercept and slope may change.

As in Example 8.17 for the quadratic model, we cannot learn anything from β_1 alone.

Example 10.4 (Kaplan video)**.** Imagine X_1 is experience and X_2 is years of education, and Y is wage ($/hr). Imagine

$$\hat{Y} = \hat{g}(X_1, X_2) = 5 - 15X_1 + 2X_2 + 2X_1X_2, \tag{10.19}$$

i.e., $\hat{\beta}_0 = 5$, $\hat{\beta}_1 = -15$, $\hat{\beta}_2 = 2$, and $\hat{\beta}_3 = 2$. Superficially, $\hat{\beta}_1 = -15$ seems like a negative relationship between experience (X_1) and wage: it looks like more experience is associated with much lower wage. However, the interaction term affects the slope with respect to X_1. Using (10.16), that slope is $\beta_1 + \beta_3 X_2 = 2X_2 - 15$. If everyone in the data has at least 10 years of education, then $X_2 \geq 10$, so $2X_2 - 15 \geq (2)(10) - 15 = 5$: the slope with respect to X_1 is always positive. Even though $\hat{\beta}_1 < 0$, $\hat{g}(X_1, X_2)$ is always increasing in X_1, for any possible $X_2 \geq 10$.

This interaction model is more general than the linear-in-variables model, but not fully general. For example, imagine Y is wage, X_1 is education, and X_2 is experience. Maybe the slope with respect to X_2 should be increasing a lot with X_1 when X_1 is around 12 or 16, but less so around $X_1 = 20$ (or maybe more so?). This type of nonlinearity in the interaction is not allowed by simply including $X_1 X_2$. There are nonlinear and nonparametric models to address such situations, but details are beyond our scope.

10.3.4 Code

The following code illustrates estimation, heteroskedasticity-robust inference, and prediction with a model including an interaction term, using real data. The outcome variable is total minutes worked per week. The regressor like D is a dummy for male, and the regressor like X is hourly wage (in 1975 U.S. dollars). The model includes an interaction term, like DX. In R formula syntax, the term `D:X` is the same as including the interaction term DX like in (10.11). Alternatively, `D*X` includes both linear and interaction terms, equivalent to `D+X+D:X`. So, both estimation models below are identical.

```
library(sandwich); library(lmtest); library(wooldridge)
# Equivalent estimates (ret = ret2)
ret  <- lm(formula=totwrk~male*hrwage, data=sleep75)
ret2 <- lm(formula=totwrk~male+hrwage+male:hrwage, data=sleep75)
retVC <- vcovHC(ret, type="HC1")
round(cbind(Estimate=coef(ret), coefci(ret, vcov. = retVC)), digits=0)

##             Estimate 2.5 % 97.5 %
## (Intercept)     1536  1215   1856
## male            1023   655   1391
## hrwage            58   -29    145
## male:hrwage      -67  -157     22

# Predict totwrk for (male,hrwage)=(1,4), (1,6), and (0,4)
predict(ret, newdata=data.frame(male=c(1,1,0),hrwage=c(4,6,4)))

##    1    2    3
## 2521 2503 1767
```

The results can be interpreted as follows. The rows correspond to the coefficients in (10.11): `(Intercept)` is for β_0, `hrwage` for β_1, `male` for β_2, and `male:hrwage` for β_3. The intercept does not have a useful interpretation because $X = 0$ means zero wage. The estimate $\hat{\beta}_1 = 58$ say a one-unit (\$1/hr) increase in wage is associated with working 58 minutes (around one hour) more per week for females; however, for males, $\hat{\beta}_1 + \hat{\beta}_3 = -9$ says that a \$1/hr increase in wage is associated with working 9 minutes less per week. The linear coefficient β_2 cannot be interpreted by itself. Instead, we could consider somebody with median wage $X = 4.3$ (it was 1975): being male is associated with $\beta_2 + 4.3\beta_3$ more minutes worked per week, which is estimated to be $\hat{\beta}_2 + 4.3\hat{\beta}_3 = 1023 + (4.3)(-67) = 735$ (around 12 hours). The CIs show much statistical uncertainty. For example, the 95% CIs for both `hrwage` and the interaction term include both positive and negative values, and the CI for the `male` coefficient is around 12 hours wide. The predictions show how predicted minutes worked is not sensitive to `hrwage` for males, but changing `male` changes the predicted value a lot.

Even just for the linear projection interpretation, there are still issues to think about like the fact that non-employed individuals do not have an hourly wage (and thus by default are quietly dropped by R); see Section 12.3.5.

10.4 Other Examples

⟹ Kaplan video: Wage Regression Example (again again)

Models can get very complex with multiple regressors. We could have more than 2 regressors; we could have many nonlinear functions of each regressor by itself; and we could have many interactions. For example, even if we only have 5 regressors, there are 10 pairs of regressors (like X_1 and X_4, X_2 and X_3, etc.), and each pair may have multiple interaction terms (i.e., not just $X_1 X_4$, but also $X_1 X_4^2$ or something). With each regressor by itself, we may have multiple nonlinear terms. There could be 40 or 50 terms in our regression just from 5 original regressors. Even if all 5 are binary, the fully saturated model requires $2^5 = 32$ parameters.

With such complicated models, it is better to look at predicted changes using the full model instead of looking at individual coefficients. This is done in R with the `predict()` function.

10.5 Assumptions for Linear Projection

Below are formal assumptions sufficient for good performance (with enough data) of the OLS estimator and heteroskedasticity-robust CIs. These are relatively weak assumptions. As usual, stronger assumptions are required to interpret the linear projection as a CMF or structural model.

The assumptions are basically the same as in Section 7.7.2, with one exception (perfect multicollinearity). Like before, iid sampling is sufficient but not necessary; the same OLS estimator can work well even with non-iid data, and there are alternative CIs that can work well in such settings, too.

10.5.1 Multicollinearity (Two Types)

The one new assumption is that there cannot be **perfect multicollinearity**. This essentially says redundant regressors are not allowed. Remember that the intercept term can be seen as the coefficient on regressor $X_0 = 1$. More formally, perfect multicollinearity means that one regressor is a linear combination (see Section 8.2.1) of other regressors.

Example 10.5 (Kaplan video)**.** If X_1 is distance in km and $X_2 = 1.6X_1$ is distance in miles, then there is perfect multicollinearity. Once we have X_1, X_2 is redundant (has the exact same information). (Formally, $1.6X_1$ is a linear "combination" of X_1.)

Example 10.6 (Kaplan video)**.** If $X_3 = X_1 + X_2$, then X_3 is a linear combination of other regressors X_1 and X_2, so there is perfect multicollinearity. If we include X_1 and X_2, then we cannot include X_3.

Example 10.7 (Kaplan video)**.** If $X_1 = 1$ for females and $X_2 = 1$ for non-females, then $X_1 + X_2 = 1 = 1 = X_0$ (the secret constant regressor). That is, the regressor X_0 is a linear combination of regressors X_1 and X_2, which means perfect multicollinearity. (Or if we omit the intercept, then there is no X_0, so no perfect multicollinearity.)

Something nice about perfect multicollinearity is that computers can check it for us. If you try to run a regression with perfect multicollinearity, R will simply report `NA` for coefficients of the "redundant" regressors (without warning or error). Other statistical packages may give you a warning or error.

For prediction, redundant variables don't help, so dropping them is fine.

For causality, we are unable to distinguish the separate effects among redundant variables. But if they are merely "control variables," then we do not care.

A related concept is **imperfect multicollinearity**. This refers to regressors being strongly correlated, but not perfectly correlated (i.e., not completely redundant).

This makes it more difficult to learn about the slope coefficients on the highly correlated regressors, but it does not invalidate any results on identification, estimation, or inference. "More difficult" means confidence intervals can be large. This makes sense: if regressors X_1 and X_2 are highly correlated, and we observe that Y is high when X_1 and X_2 are high, it's unclear whether Y is high because X_1 is high or because X_2 is high. Because they are highly correlated, there are few observations where only X_1 or X_2 (not both) is high to help distinguish the effect of X_1 from that of X_2. This is similar to the logic behind omitted variable bias, except we can see the ghost. With prediction, it may be best to include only X_1 or X_2 (not both), but standard model selection procedures can handle this without any special consideration. (But: if you have a job interview and sense that your interviewer thinks imperfect multicollinearity is really important for some reason, just go with it.)

Example 10.8. The following example shows with real data how imperfect multicollinearity can lead to wide CIs for the coefficients of correlated regressors, but does not cause any other problems. The outcome is a binary variable equal to 1 if the individual reports being in good health. The two regressors are total sleep (adjusted to hours per day) and sleep at night (excluding naps). They have a very high correlation (0.89) because naps are a small fraction of total sleep for most individuals. The estimated coefficients have different signs (one positive, one negative) but very wide 95% CIs that contain both positive and negative values; note the CI when only `slpnaps` is a regressor is less than half as wide as after `sleep` is added as a second regressor. (Recall the interpretation that, for example, a coefficient of 0.01 means a one-hour increase in sleep is associated with a one percentage point higher probability of good health.) The predictions do not change much when using only `slpnaps` instead of both sleep variables; this is reflected by the extremely similar R^2 values.

```
library(sandwich); library(lmtest); library(wooldridge)
c(corr=cor(sleep75$sleep, sleep75$slpnaps))

##  corr
## 0.893

ret <- lm(formula=gdhlth~I(slpnaps/60/7), data=sleep75)
retVC <- vcovHC(ret, type="HC1")
round(cbind(Estimate=coef(ret), coefci(ret, vcov. = retVC)), digits=3)

##                  Estimate  2.5 % 97.5 %
## (Intercept)         1.149  0.975   1.32
## I(slpnaps/60/7)    -0.032 -0.054  -0.01

ret <- lm(formula=gdhlth~I(slpnaps/60/7)+I(sleep/60/7), data=sleep75)
retVC <- vcovHC(ret, type="HC1")
round(cbind(Estimate=coef(ret), coefci(ret, vcov. = retVC)), digits=3)

##                  Estimate  2.5 % 97.5 %
## (Intercept)         1.137  0.951  1.322
## I(slpnaps/60/7)    -0.039 -0.092  0.014
## I(sleep/60/7)       0.009 -0.049  0.067

c(R2a=summary(lm(formula=gdhlth~slpnaps, data=sleep75))$r.squared,
  R2b=summary(lm(formula=gdhlth~slpnaps+sleep, data=sleep75))$r.squared)

##    R2a    R2b
## 0.0148 0.0150
```

10.5.2 Formal Assumptions and Results

The assumptions and results refer to the linear projection model

$$\mathrm{LP}(Y \mid X_1, \ldots, X_J) = \beta_0 + \beta_1 X_1 + \cdots + \beta_J X_J. \tag{10.20}$$

The X_j may include nonlinear functions of an original set of regressors. For example, if X and D are observed regressors, then the model could have $X_1 = D$, $X_2 = X$, and $X_3 = DX$. It could also include $X_4 = X^2$, etc.

Assumption A10.1. Sampling of $(Y_i, X_{1i}, \ldots, X_{Ji})$ is iid from the population joint distribution of $(Y, X_1, \ldots, X_J)$.

Assumption A10.2. There is no perfect multicollinearity.

Assumption A10.3. The variances of Y and all X_j are finite: $\mathrm{Var}(Y) < \infty$, $\mathrm{Var}(X_j) < \infty$ for $j = 1, \ldots, J$.

Assumption A10.4. The fourth moments are finite: $\mathrm{E}(Y^4) < \infty$, $\mathrm{E}(X_j^4) < \infty$ for $j = 1, \ldots, J$.

The following theorems extend Theorems 7.1 and 7.2 to multiple regressors.

Theorem 10.1 (OLS consistency). *If A10.1–A10.3 are true, then the OLS intercept and slope estimators are consistent for the population linear projection intercept and slope in* (10.20).

Theorem 10.2 (coverage probability, multiple regressors). *If A10.1, A10.2, and A10.4 are true, then heteroskedasticity-robust confidence intervals are asymptotically correct. That is, with a large enough sample size, the coverage probability is approximately equal to the desired confidence level.*

10.6 Causality: Identification

There are many identification results in which there is a causal interpretation for something OLS can estimate. Here are a couple.

10.6.1 Linear Structural Model

Consider the structural model

$$Y = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + U. \tag{10.21}$$

Some of the X_j are allowed to be nonlinear functions of regressors, including interaction terms. If $\mathrm{Cov}(U, X_j) = 0$ for all $j = 1, \ldots, J$, then the structural β_j are also linear projection coefficients (which OLS can estimate).

10.6.2 Conditional ATE

As alluded to in Section 9.5, a **conditional average treatment effect** (CATE) can be identified under conditional independence (A9.1). Here, X_1 is the binary treatment variable. Given Assumption A9.1 (and SUTVA and overlap), the CATE equals a CMF difference:

$$\begin{aligned}
&\mathrm{E}[Y^T - Y^U \mid X_2 = x_2, \ldots, X_J = x_J] \qquad (10.22)\\
&= \mathrm{E}[Y \mid X_1 = 1, X_2 = x_2, \ldots, X_J = x_J] - \mathrm{E}[Y \mid X_1 = 0, X_2 = x_2, \ldots, X_J = x_J].
\end{aligned}$$

Intuitively, conditional independence says that within any subpopulation defined by having the same $(X_2, \ldots, X_J)$ values, treatment is "as good as random," so comparing mean treated and untreated observed outcomes (within the subpopulation) has a causal interpretation.

For estimation, we either need to know the CMF's functional form (and use OLS) or use nonparametric estimation techniques.

Optional Resources

Optional resources for this chapter

- James et al. (2013, §3.2)
- Hastie, Tibshirani, and Friedman (2009, §§2.3.1,2.4,3.1–3.2)
- Linear projection (theory): Hansen (2020, §7)
- Average structural effects and their identification: Hansen (2020, §2.30)
- Regression example (Masten video)
- Perfect multicollinearity (Lambert video)
- Imperfect multicollinearity example (Lambert video)
- Dummy coefficients (Lambert video)
- Dummy interactions (Lambert video)
- Continuous interactions (Lambert video)
- Sections 3.1 ("Multiple Regression in Practice") and 6.1.5 ("Interaction Terms") in Heiss (2016)
- Section 4.4 ("Reporting Regression Results") in Heiss (2016)
- Section 8.3 ("Interactions Between Independent Variables") in Hanck et al. (2018)

Empirical Exercises

Empirical Exercise EE10.1. You will analyze data collected from Botswana's 1988 Demographic and Health Survey by James Heakins for an economics term project. In particular, you'll see how the number of living children a woman (in Botswana) has relates to various other variables, with particular interest in the woman's years of education. You'll start with a simple regression of `children` on `educ` that shows an economically significant negative coefficient. Then, you'll see how this coefficient changes (generally moving toward zero) as you add other regressors as control variables, like the husband's education (`heduc`) and the woman's age (`age`). These changes in the estimated coefficient suggest omitted variable bias in the original simple regression. But, even with a large number of control variable regressors, there is probably still omitted variable bias.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```
library(wooldridge); library(sandwich); library(lmtest)
?fertil2
```

b. Stata only: load the data with `bcuse fertil2 , nodesc clear` (assuming `bcuse` is already installed)

c. Run a simple regression of `children` on `educ`.

   R: `ret1 <- lm(children~educ, data=fertil2)`

   Stata: `regress children educ , vce(robust)`

d. Repeat but adding `heduc` as a control variable regressor.

   R: `ret2 <- lm(children~educ+heduc, data=fertil2)`

   Stata: `regress children educ heduc , vce(robust)`

e. Repeat but adding yet another regressor (woman's age).

   R: `ret3 <- lm(children~educ+heduc+age, data=fertil2)`

   Stata: `regress children educ heduc age , vce(robust)`

f. Repeat but add even more regressors (in addition to `educ`, `heduc`, and `age`): `agesq`, `knowmeth`, `usemeth`, `electric`, `urban`, and `catholic`, as well as interactions between `age` and `knowmeth` and between age and `usemeth`.

   R: store the result as `ret4`, and you can simply write `knowmeth:age` and `usemeth :age` in the regression formula to generate the interactions.

   Stata: first create the two interaction variables like with `generate know_age = knowmeth*age` and then run another regression with your two new variables added to your list of regressors.

g. R only (because already displayed by Stata): output the four sets of estimated regression coefficients with

```
coef(ret1)
coef(ret2)
coef(ret3)
coef(ret4)
```

h. Optional: repeat one more time, with whichever regressors (in addition to `educ`) you think appropriate; feel free to create additional interaction terms and/or nonlinear terms (like `age^3`, etc.).

Empirical Exercise EE10.2. You will analyze data originally from Harrison and Rubinfeld (1978), including housing prices and pollution measures. The data are not for individual houses, but instead small areas (census tracts, I'd guess), within which the median housing price is computed along with other characteristics that may affect housing prices, including pollution. You'll start with a simple regression of log `price` on log `nox` (the pollution measure). The coefficient is around -1, meaning a 1% increase in pollution is associated with (approximately) a 1% decrease in price. Then, you'll add other regressors to try to reduce omitted variable bias. By adding just a couple variables, the pollution coefficient estimate's magnitude is cut in half, suggesting that there was indeed much OVB. However, even with a large number of regressors, serious OVB may remain.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```
library(wooldridge); library(sandwich); library(lmtest)
?hprice2
```

b. Stata only: load the data with `bcuse hprice2 , nodesc clear` (assuming `bcuse` is already installed)

c. Run a simple log-log regression of `price` on `nox`.

   R: `ret1 <- lm(log(price)~log(nox), data=hprice2)`

   Stata: `regress lprice lnox , vce(robust)`

d. Repeat but adding `rooms` as a control variable regressor.

   R: `ret2 <- lm(log(price)~log(nox)+rooms, data=hprice2)`

   Stata: `regress lprice lnox rooms , vce(robust)`

e. Repeat but adding yet another regressor (`crime` rate per capita).

   R: `ret3 <- lm(log(price)~log(nox)+rooms+crime, data=hprice2)`

   Stata: `regress lprice lnox rooms crime , vce(robust)`

f. Repeat but add even more regressors: `dist`, `radial`, `stratio`, and `lowstat`. Store the result as `ret4` in R.

g. R only (because already displayed by Stata): output the four sets of estimated regression coefficients with

```
coef(ret1)
coef(ret2)
coef(ret3)
coef(ret4)
```

h. Optional: repeat one more time, with whichever regressors you think appropriate; try to use interaction terms and/or nonlinear terms (like `rooms^2`, etc.).

Chapter 11

Midterm Exam #2

⟹ Kaplan video: Chapter Introduction

When I teach this class, the second midterm exam is this week. This “chapter” makes the chapter numbers match the week of the semester. This midterm covers all chapters between the first midterm and now. It does not explicitly include questions about the material before the first midterm exam, but of course that materials was foundational for the material covered on the new exam, so it may (or may not) still help to review it.

Chapter 12

Internal and External Validity

⟹ Kaplan video: Chapter Introduction

This chapter discusses many reasons to worry about the validity of econometric results and their application to decisions. Like statistical and economic significance, "validity" is better thought of as a continuum rather than a yes/no property. To tweak Box's aphorism, "All results are invalid, but some are useful."

Unit learning objectives for this chapter

12.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

12.2. Assess possible problems with regression results and their application to real-world questions (of description, prediction, and causality), and the likely direction of bias [TLOs 5 and 6]

12.3. In R (or Stata): check datasets for possible issues like missing data [TLO 7]

12.1 Terminology

An econometric study has **internal validity** if the methods are appropriate for the study's setting and sample, i.e., if all identifying assumptions and other assumptions hold.

An econometric study has **external validity** for a different setting if the results can be used to learn something about the new setting. This does not necessarily mean the values or overall effects are identical, but that at least the prediction model or structural model is the same.

The **population studied** refers to the population from which the data was sampled, whereas the **population of interest** is the one that you (as the researcher, policy maker, or decision maker) want to learn about.

Example 12.1. Imagine you see an econometric study of the causal effect of a minimum wage change from \$4.25/hr to \$5.05/hr in New Jersey in 1992, but your job is to advise Missouri about a possible minimum wage increase next year. The study is internally valid if it properly estimates the causal effect of the New Jersey minimum wage increase on people in New Jersey in 1992, i.e., for the population studied. It is externally valid if the estimates can be used to learn about the (potential) policy effects in Missouri, your population of interest. Again, this doesn't necessarily mean the effect itself next year in Missouri must be identical to the effect in 1992 New Jersey, but that the model estimated with the 1992 New Jersey data can be applied to current Missouri data to learn the potential policy effect.

This chapter briefly discusses many **threats to validity**, i.e., reasons an analysis may not be internally or externally valid.

12.2 Threats to External Validity

Threats to external validity are generally more obvious than threats to internal validity, but they harm evidence-based decisions just as much. For example, consider the descriptive task of estimating the median house price in Missouri. Obviously, a sample of house prices from California (which is much more expensive) does not help. Even with the price of every house in California, we learn little about Missouri. We can try to learn about relationships between price and house features (size, land area, etc.) in California, but probably even such relationships themselves differ in Missouri. This problem is a lack of external validity.

The Lucas critique (Lucas, 1976, also Section 4.3.3) can also be interpreted in terms of external validity. When macroeconomic policy changes, that fundamentally changes the setting. Even if our estimates from historical data have internal validity, they might not be accurate in the new setting under the new policy, i.e., they might not have external validity.

A few common threats to external validity are now discussed. Although these don't automatically imply lack of validity, they are reasons for skepticism.

12.2.1 Different Place

Different places have different legal, political, cultural, and economic settings. The house price example highlights just one of many important differences between California and Missouri. Ideally, you can always find an empirical study from the same place you're interested in. If not, you have to decide whether you think the other place is similar enough to still help you make a good decision.

Example 12.2. Imagine you need to quantify costs and benefits of expanding public bus systems in Missouri. The neighboring states of Oklahoma and Illinois recently collected data during their (hypothetical) bus system expansions, with different results. Although both states are very close geographically to Missouri, other characteristics matter, too. Illinois has almost double the state gas tax of Missouri, and its urban population share is over 15 percentage points greater than Missouri's; both of these may be important for both people's decision to ride the bus (versus drive) and the cost of bus operation. In contrast, Oklahoma's gas tax and urban population share are very similar to Missouri's, so there is probably greater external validity. Still, there may be other important differences between Missouri and Oklahoma, some of which may be difficult to measure accurately or quantify, like cultural attitudes.

12.2.2 Different Time

Even in the same place, there may be important changes over time in the legal, political, cultural, and economic setting. If you are making a decision today based on analysis of historical data, then external validity depends on how much has changed between then and now. In certain places for certain variables, maybe not much has changed in the past five years. But for other places or variables, maybe there are important changes every year, or even every day; or just coincidentally, a new law went into effect yesterday. Assessing external validity requires critical thinking about how much has changed between the time of the dataset and now. Alternatively, sometimes we can model how variables change over time, in order to predict how they have changed since the data sample was collected; see Chapters 14 and 15.

Example 12.3. Consider again the median house price in Missouri. Having learned not to use California data, we get Missouri data—from the year 1975. This is also bad because house prices were much lower in 1975 than today. Adjusting for inflation would help some, but the housing market supply and demand have both changed substantially, even basics like how many people live in Missouri and houses' size, age, and quality. We could try to use all these variables in a model, but some may not have data available, and the model itself may have changed since 1975, due to changes in factors like consumer preferences, regulations, and materials costs. If we instead had historical data from last month, then our analysis should have high external validity, barring any sudden dramatic change in the past month (like a pandemic).

12.2.3 Different Population

Even in the same place, at the same time, the population studied may differ from your population of interest. There is a threat to external validity if the variables and their relationship differ between these two populations. That is, even if we correctly learn about the population studied, it may not help us learn about the population of interest.

Example 12.4. To guide tax incentives for first-time homebuyers in Missouri this year, you want to estimate the median first-year mortgage payment for first-time homebuyers. If you find a study estimating the median mortgage payment among all home owners in Missouri this year, then your number will be much too big because the studied population (all owners) differs greatly from the population of interest (first-time owners), even though they're in the same place (Missouri) at the same time (this year).

Example 12.5 (Kaplan video)**.** Imagine you're estimating the benefits of expanding government subsidies for college in the U.S. You (amazingly) find an internally valid estimate of the mean wage increase from college, in the same place (U.S.), from just a few months ago. However, the estimate is for the whole U.S. population (the population studied), including individuals who already got college degrees even without the additional subsidy. Instead, your population of interest is individuals who currently do not (or cannot) choose to graduate from college, but who would with the additional subsidy. Such individuals may not have the same causal effect of college on their wages.

Discussion Question 12.1 (external validity: minimum wage)**.** You're deciding whether to vote for a minimum wage increase in your state or country (yes, you! wherever you live or vote right now), from $10/hr to $15/hr (or an equivalent increase in your country's currency). You find a study (Card and Krueger, 1994) of effects of a minimum wage increase from $4.25/hr to $5.05/hr in New Jersey in 1992. Explain your specific concerns about external validity. (Note: this is only a question about external validity; arguments about whether minimum wage should be lower or higher are completely irrelevant.)

12.3 Threats to Internal Validity

For description and prediction, see Items 1–5 in the list below.

For causality, the following common threats to internal validity are described below.

1. Functional form misspecification (Section 12.3.1)
2. Measurement error (Sections 12.3.2 and 12.3.3)
3. Non-iid sampling and weights (Section 12.3.4)
4. Missing data (Section 12.3.5)
5. Sample selection (Section 12.3.6)
6. Omitted variables (Section 12.3.7)
7. Simultaneity and reverse causality (Section 12.3.8)

Additionally, violation of SUTVA (as discussed earlier) is another threat to internal validity for treatment effect analysis.

12.3.1 Functional Form Misspecification

Misspecifying the functional form leads to inconsistent estimates of the CMF. This is bad for description, prediction, and causality alike. Details are in Chapters 7–10, including reasons for misspecification, ways to address it, and interpretations of what OLS estimates when it's not a CMF.

12.3.2 Measurement Error in the Outcome Variable

⟹ Kaplan video: Measurement Error

Without good data, it's hard to get valid econometric results. As they say, "Garbage in, garbage out." But, it is not as simple as "good" and "bad" data. Certain data problems can safely be ignored; others can't be ignored, but they can be fixed; and yet other problems cannot be fixed by any amount of econometrics magic.

Sometimes the true value of a variable is not the value seen in the data. This is especially true in survey data, where individuals (or firms, schools, etc.) report their own information ("self-reported"), and with macroeconomic variables that are difficult to measure accurately. With survey data, people may simply forget the exact value, or they may intentionally lie in some cases.

Notation and Terminology

To define some notation and terms, consider the example of exercise. In a survey, people are asked how many minutes of exercise they did last week, and their responses Y are recorded in the data. Let Y^* be how much exercise somebody truly did last week. This Y^* is the **latent** (unobserved) true value. In contrast, the observed value is $Y = Y^* + M$, where M is the **measurement error**. That is, $M = Y - Y^*$ is the difference between the observed and true values.

All of (Y^*, Y, M) are uppercase to show they're random variables. For example, one individual could have true $Y^* = 98.52$ and report $Y = 100$, so $M = 100 - 98.52 = 1.48$, whereas another individual could have $Y^* = 271$, report $Y = 250$, and have $M = 250 - 271 = -21$, where all values are in units of "exercise minutes per week." There are different values of (Y^*, Y, M) for different individuals; the population distribution describes the probabilities of these different possible values.

Discussion Question 12.2 (exercise error)**.** Consider the example where Y^* is true exercise minutes last week and Y is the value somebody reports. Explain one reason (each) why an individual could have a) $M = 0$, b) $M < 0$, or c) $M > 0$. Overall, would you guess $\mathrm{E}(M) = 0$, $\mathrm{E}(M) < 0$, or $\mathrm{E}(M) > 0$? Why? (Hint: you'll probably need to make additional assumptions and definitions; e.g., what does "exercise" mean?)

Regression

Imagine the true linear projection in error form is

$$Y^* = \beta_0 + \beta_1 X + V, \quad \mathrm{E}(V) = \mathrm{Cov}(X, V) = 0. \tag{12.1}$$

We want to learn β_1. Substituting in $Y^* = Y - M$,

$$\begin{aligned} Y - M &= \beta_0 + \beta_1 X + V, \\ Y &= \beta_0 + \beta_1 X + \overbrace{(V + M)}^{U} = \beta_0 + \beta_1 X + U. \end{aligned} \tag{12.2}$$

The OLS estimator $\hat{\beta}_1$ may be asymptotically biased if X and M are related because then X and U are related. From (9.4), the asymptotic bias is

$$\mathrm{AsyBias}(\hat{\beta}_1) = \frac{\mathrm{Cov}(X,U)}{\mathrm{Var}(X)} = \frac{\mathrm{Cov}(X,V+M)}{\mathrm{Var}(X)} = \frac{\overbrace{\mathrm{Cov}(X,V)}^{=0} + \mathrm{Cov}(X,M)}{\mathrm{Var}(X)}. \tag{12.3}$$

This is the slope coefficient in $\mathrm{LP}(M \mid 1, X)$, the linear projection of the measurement error onto the regressor (and an intercept). That is, writing $\mathrm{LP}(M \mid 1, X) = \gamma_0 + \gamma_1 X$, then the asymptotic bias is $\mathrm{AsyBias}(\hat{\beta}_1) = \gamma_1$.

With binary X, γ_1 is a mean difference as in (6.21), so (12.3) is equivalent to

$$\mathrm{AsyBias}(\hat{\beta}_1) = \mathrm{E}(M \mid X = 1) - \mathrm{E}(M \mid X = 0). \tag{12.4}$$

This helps us think about both the sign (direction) and magnitude of asymptotic bias. For example, if there tends to be more positive measurement error when $X = 1$ than when $X = 0$, then $\gamma_1 > 0$, so the OLS estimator $\hat{\beta}_1$ has positive asymptotic bias.

In Sum: Measurement Error in the Outcome

Y observed; Y^* latent/true; M measurement error
$Y = Y^* + M \iff M = Y - Y^*$
Binary X: $\hat{\beta}_1$ asymptotic bias is $\mathrm{E}(M \mid X = 1) - \mathrm{E}(M \mid X = 0)$; see (12.4)
General X: $\hat{\beta}_1$ asymptotic bias is γ_1 in $\mathrm{LP}(M \mid 1, X) = \gamma_0 + \gamma_1 X$; see (12.3).

Example

Continuing with Y^* as weekly exercise, let $X = 1$ if somebody has a gym membership and $X = 0$ otherwise. The goal is to learn β_1 in $\mathrm{LP}(Y^* \mid 1, X) = \beta_0 + \beta_1 X$. Because X is binary, β_1 is also the mean difference $\mathrm{E}(Y^* \mid X = 1) - \mathrm{E}(Y^* \mid X = 0)$. Also due to binary X, the slope in $\mathrm{LP}(M \mid 1, X) = \gamma_0 + \gamma_1 X$ is the same as the mean difference, $\gamma_1 = \mathrm{E}(M \mid X = 1) - \mathrm{E}(M \mid X = 0)$.

There's no asymptotic bias in a few cases. Obviously, if $M = 0$ for everybody, then $Y = Y^*$, so regressing Y on X is identical to regressing Y^* on X. Even if everyone overreports ($\mathrm{E}(M) > 0$) or underreports ($\mathrm{E}(M) < 0$), as long as it's the same for both gym members and non-members, then $\gamma_1 = 0$, so there is no asymptotic bias. It's also fine if $\mathrm{E}(M \mid X = 0) = \mathrm{E}(M \mid X = 1) = 0$ but $\mathrm{Var}(M \mid X = 1) < \mathrm{Var}(M \mid X = 0)$, i.e., the gym members report more accurately (smaller variance of M; in the extreme, even $M = 0$), but both groups are accurate on average.

However, there is asymptotic bias if there's systematic overreporting by only gym members. Maybe gym members are more likely to feel guilty about not exercising and not using their membership, which may cause them to report going to the gym and exercising more than they actually do. Or, conversely, perhaps individuals who think they exercise more than they do (and thus have large M) are more likely to become gym members because they think it'll be worth it. Either way, more positive M (overreporting) is associated with $X = 1$ compared to $X = 0$, i.e., $\gamma_1 > 0$. This leads to positive (upward) asymptotic bias of $\hat{\beta}_1$.

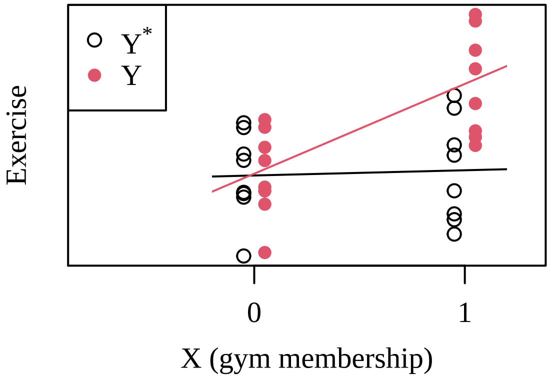


Figure 12.1: Bias from measurement error in Y.

Figure 12.1 illustrates the upward bias of $\hat{\beta}_1$ in the gym/exercise example. The $X = 0$ group does not report perfectly, but there is no systematic reporting bias. The $X = 1$ group systematically overreports exercise. Consequently, the red line's slope (using observed Y) is much larger than the black line's slope (using true but unobserved Y^*). That is, if we could observe Y^*, we would estimate the black line; but we can't, and using the observed Y yields a very different (biased) estimate of the slope β_1.

Alternatively, maybe non-gym members tend to have larger M. Maybe gym members only report gym time, whereas non-members include walking the dog, lifting groceries, etc. In that case, $\mathrm{E}(M \mid X = 0) > \mathrm{E}(M \mid X = 1)$, so $\gamma_1 < 0$ and there's negative asymptotic bias.

Discussion Question 12.3 (measurement error: scrap rate)**.** Imagine the government wants to help increase the efficiency of chalk manufacturing firms. Specifically, Y^* is a firm's "scrap rate": what proportion of their output has to be "scrapped" (trashed/not sold) due to manufacturing defects? For example, $Y^* = 0.04$ means 4% scrap rate. The government randomly assigns firms to a control group and treatment group, to run an experiment. On January 1, the treated firms receive grant money, which they are supposed to use to improve efficiency. All firms self-report their scrap rates on December 31; this is Y.

a) Describe a reason why treated firms might systematically overreport ($M > 0$) or underreport ($M < 0$) their scrap rates.
b) In that case, and assuming untreated firms report accurately ($M = 0$), would we overestimate or underestimate the treatment effect of a grant? Why?
c) If the government uses these incorrect estimates to decide whether or not to continue the program, what incorrect decision might they make? Why?

Methods to Address Measurement Error

In some cases, there are methods to reduce or eliminate the bias from measurement error. However, such methods often have additional requirements, like a second measurement of the same variable, and they are beyond our scope.

12.3.3 Measurement Error in the Regressors

There are similarities between measurement error in X and measurement error in Y. Much of the math is similar. The causes of measurement error are the same, because a variable may be the Y variable in one model but the X variable in another.

To see how measurement error might cause asymptotic bias, equations like (12.1) and (12.2) can be derived. The true LP with latent X^* is

$$Y = \beta_0 + \beta_1 X^* + R, \quad \mathrm{E}(R) = \mathrm{Cov}(X^*, R) = 0. \tag{12.5}$$

Because the observed X is $X = X^* + M$, substituting in $X^* = X - M$,

$$Y = \beta_0 + \beta_1 (X - M) + R = \beta_0 + \beta_1 X + (R - \beta_1 M). \tag{12.6}$$

Like (12.3), the asymptotic bias is

$$\mathrm{AsyBias}(\hat{\beta}_1) = \frac{\mathrm{Cov}(X, R - \beta_1 M)}{\mathrm{Var}(X)},$$

so the asymptotic bias is zero if and only if $\mathrm{Cov}(X, R - \beta_1 M) = 0$, i.e., if the observed X is uncorrelated with the unobserved "error term" $R - \beta_1 M$. Using (12.5) and linearity,

$$\begin{aligned}
\mathrm{Cov}(X, R - \beta_1 M) &= \mathrm{Cov}(X, R) - \mathrm{Cov}(X, \beta_1 M) \\
&= \mathrm{Cov}(X^* + M, R) - \beta_1 \mathrm{Cov}(X, M) \\
&= \overbrace{\mathrm{Cov}(X^*, R)}^{=0} + \mathrm{Cov}(M, R) - \beta_1 \mathrm{Cov}(X, M).
\end{aligned}$$

If M is uncorrelated with the LP error $R = Y - \beta_0 - \beta_1 X^*$, and if $\beta_1 = 0$ (which means Y and the true X^* are not correlated), then this is zero. Otherwise, there is almost certainly asymptotic bias, in particular when $\mathrm{Cov}(X, M) \neq 0$.

Attenuation Bias: Assumptions and Result

Unfortunately, $\mathrm{Cov}(X, M) = 0$ is very unlikely. Consider what seems to be the best-case scenario: M is just random noise unrelated to the true value X^*, so $\mathrm{Cov}(X^*, M) = 0$. Unfortunately, using $\mathrm{Cov}(X^*, M) = 0$,

$$\mathrm{Cov}(X, M) = \mathrm{Cov}(X^* + M, M) = \overbrace{\mathrm{Cov}(X^*, M)}^{=0} + \overbrace{\mathrm{Cov}(M, M)}^{=\mathrm{Var}(M)} = \mathrm{Var}(M). \tag{12.7}$$

Assuming not everybody has $M = 0$, then $\text{Var}(M) > 0$, so $\text{Cov}(X, M) > 0$. Thus, even if $\text{Cov}(M, R) = 0$, the asymptotic bias is not zero, $-\beta_1 \text{Cov}(X, M) \neq 0$.

In this case with $\text{Cov}(X, M) > 0$ and $\text{Cov}(M, R) = 0$, the resulting bias is called **attenuation bias**. This means that the estimates $\hat{\beta}_1$ tend to be in between 0 and β_1: $0 < \text{plim}\,\hat{\beta}_1/\beta_1 < 1$, implying $|\text{plim}\,\hat{\beta}_1| < |\beta_1|$. That is, the estimates are systematically pushed closer to zero by the measurement error. This is different than positive (upward) bias, which tends to make $\hat{\beta}_1 > \beta_1$, or negative (downward) bias, which tends to make $\hat{\beta}_1 < \beta_1$. With attenuation bias, if $\beta_1 > 0$, then generally $0 < \hat{\beta}_1 < \beta_1$, whereas if $\beta_1 < 0$, then generally $0 > \hat{\beta}_1 > \beta_1$.

Even if we cannot fix the attenuation bias, it is helpful to know the direction of the bias. For example, if we estimated $\hat{\beta}_1 = 7$, and we suspect attenuation bias, then we may think β_1 might be even larger, but probably not smaller.

Attenuation Bias: Example

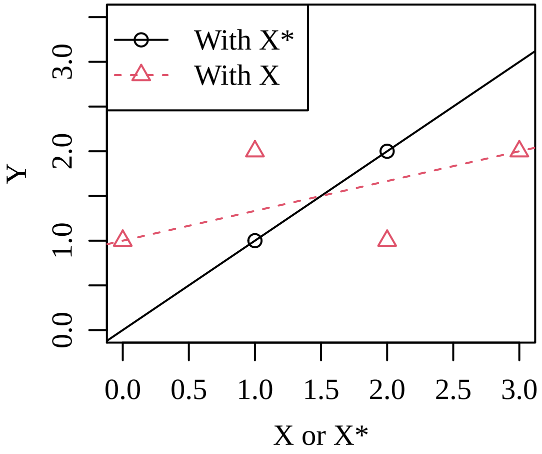


Figure 12.2: Bias from measurement error in X.

Figure 12.2 illustrates attenuation bias. It shows a simple example where $P(X^* = 1) = P(X^* = 2) = 0.5$, and $Y = X^*$ (no error term). The linear projection is just the line through $(X^*, Y) = (1, 1)$ and $(2, 2)$, which has $\beta_0 = 0$ and $\beta_1 = 1$ (intercept zero, slope one). Then, imagine adding error: $P(M = -1) = P(M = 1) = 0.5$, regardless of X^* or Y. Then the $X^* = 1$ values become $X = X^* + M$: either $X = 1 - 1 = 0$ or $X = 1 + 1 = 2$. Similarly, the $X^* = 2$ values become either $X = 2 - 1 = 1$ or $X = 2 + 1 = 3$. Now we have four possible values of (X, Y), each with equal 0.25 probability: $(0, 1)$, $(2, 1)$, $(1, 2)$, and $(3, 2)$, forming a parallelogram. The result is $\text{LP}(Y \mid 1, X) = 1 + X/3$ (slope is 1/3), very different than $\text{LP}(Y \mid 1, X^*) = X^*$ (slope is 1). That is, when we add horizontal noise (errors in X), the slope of the linear projection $\text{LP}(Y \mid 1, X)$ is flatter (closer to zero) than the slope of $\text{LP}(Y \mid 1, X^*)$.

General Bias

Unfortunately, outside this very special case, the type of bias may differ. It is not necessarily attenuation bias.

In particular, if $\mathrm{Cov}(M, R) \neq 0$ and $|\mathrm{Cov}(M, R)| > |\beta_1 \mathrm{Cov}(X, M)|$, then the sign of the bias is the sign of $\mathrm{Cov}(M, R)$, i.e., positive bias if $\mathrm{Cov}(M, R) > 0$ or negative bias if $\mathrm{Cov}(M, R) < 0$. So, generally, any type of asymptotic bias is possible, depending how the measurement error is related to other variables.

There are methods that address measurement error in X, but these are beyond our scope.

12.3.4 Non-iid Sampling and Survey Weights

As advised in Section 3.4.3, if your dataset has survey weights (a.k.a. sampling weights), then you should probably use them. Most statistical estimation functions in R allow such weights. It's true that in some cases you don't actually need to use weights, but it's safer to just always use them.

Section 3.5 discussed how sampling may be non-iid for other reasons like clustered and/or stratified sampling. Time series data also usually lack iid sampling; see Part III. Generally, with these types of non-iid sampling, estimators are consistent but confidence intervals require different formulas to be accurate. For now, just be aware that you need something besides a heteroskedasticity-robust CI.

12.3.5 Missing Data

⟹ Kaplan video: Bias from Non-Ignorable Missing Data

Like with measurement error in Y (Section 12.3.2), the reason why there is missing data determines whether or not it's a problem. As we saw with measurement error in Y, if the error is completely random (independent of X), then it will not bias linear projection slope estimates. Similarly, if data is missing completely at random (like, a cat walked across your computer keyboard or something), then it's fine to just drop observations with missing data and proceed as usual. This is called **complete case analysis**, where a **complete case** is an observation in which no values are missing (i.e., all values are observed). For example, if the dataset is (Y_i, X_i) for $i = 1, \ldots, n$, then the complete cases are the i for which both Y_i and X_i are observed (not missing).

In other cases, we can't ignore the missing data problem, but there are methods that can fix the problem and avoid asymptotic bias.

In yet other cases, it is very difficult to address the missing data problem. In particular, when the value of Y affects whether or not data are missing, it is very difficult. For example, if Y is income and people with high (or low) income tend not to report their income on a survey, then regression estimates will be biased.

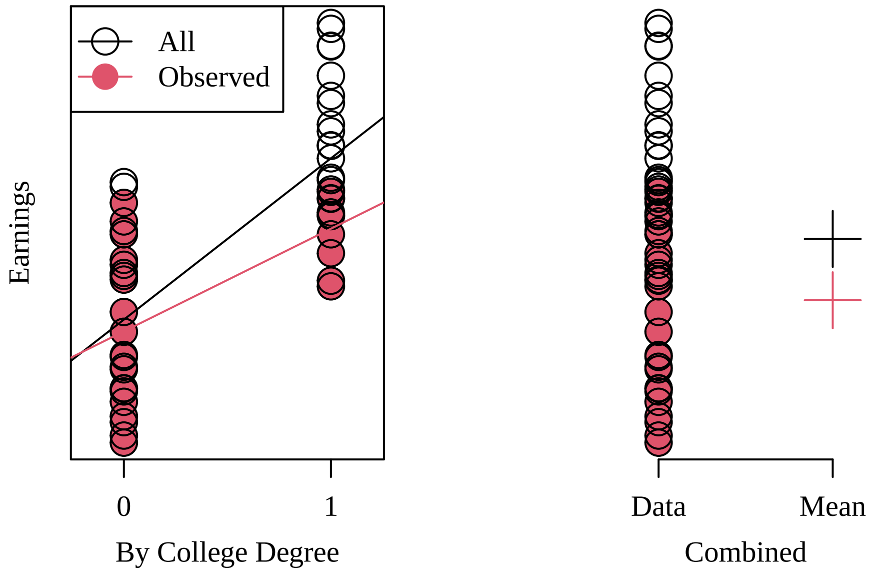


Figure 12.3: Non-ignorable missing data: bias of both OLS and sample mean.

Example 12.6 (Kaplan video)**.** Figure 12.3 shows an example of missingness related to Y. Here, Y is income and $X = 1$ if an individual has a college degree, $X = 0$ if not. In the example, the highest-income individuals do not report Y_i but everyone else does. This mostly affects $X_i = 1$ individuals, but also the very highest Y_i in the no-college group. If we just run OLS on observations with both Y_i and X_i observed, then both the OLS slope and sample mean are biased downward. The OLS intercept is very slightly downward biased, too, because the top Y_i when $X_i = 0$ are missing.

Practice 12.1 (program attrition)**.** Consider a job training program like the federally funded Job Training Partnership Act (JTPA) of 1982. Each eligible individual was randomly assigned to either take the job training or not. You want to estimate the average treatment effect on annual income (Y) of being assigned to the training (the "intention-to-treat" effect from Section 4.6.3). However, some individuals' data is missing because they moved to a different state to take a high-paying job. Explain why this could be a threat to internal validity, and in which direction you think the resulting bias might be.

Discussion Question 12.4 (missing salary data)**.** You get data on a sample of professors from research universities in the U.S., which is the population of interest. However, you only find salary data for public universities, not private.

a) How/does this bias your estimate of the population mean salary? Why?
b) How/does this bias your regression of salary on a dummy for being a professor in a STEM field? Why? (Hint: consider the intercept and slope separately.)

Caution: by default, most commands in Stata and functions in R drop all observations (rows in your dataset) with any missing variable(s) automatically, without any error or warning message. That is, they assume you want complete case analysis. You can still figure out whether or not any observations were dropped. You can also tell R to behave differently if it encounters `NA` values. You can either do this through `options()` to change the default, or for a specific `lm` (or whatever function) call through the `na.omit` argument. See the code below.

```
n <- 5;  set.seed(112358);  options(digits=3)
Y <- rnorm(n); X <- rnorm(n)
Y[2] <- X[3] <- NA #missing values
r <- lm(Y~X) #no hint of missing/dropped obs
coef(r) #still no hint:
## (Intercept)           X
##       0.591       0.704

nrow(r$model) #aha: not n rows!
## [1] 3

#summary(r) #"(2 observations deleted due to missingness)"
options("na.action") #print current default (usually na.omit)
## $na.action
## [1] "na.omit"

predict(lm(Y~X, na.action=na.omit)) # complete case
##       1       4       5
## -0.7037 -0.0139  1.1147

predict(lm(Y~X, na.action=na.exclude)) #fill in NA
##       1       2       3       4       5
## -0.7037      NA      NA -0.0139  1.1147

lm(Y~X, na.action=na.fail) # give an error if NAs in data
## Error in na.fail.default(list(Y = c(-0.471, NA, 0.530,  :
##  missing values in object

options(na.action=na.fail) #set default to na.fail
lm(Y~X) #now gives error as default (if NA values)
## Error in na.fail.default(list(Y = c(-0.471, NA, 0.530,  :
##  missing values in object
```

12.3.6 Sample Selection

⟹ Kaplan video: Sample Selection Bias

Whereas missing data means some values are missing in the dataset, **sample selection** means entire individuals (observations) are missing. Whereas missing values are indicated by `NA` in R, there may be no indication that entire individuals are missing. The number of missing individuals may be unknown.

Terminology caveat: sometimes people use the phrase "sample selection" to refer to missing data, especially missing Y data. In that context, there are methods to try to reduce "sample selection bias" by using the observed regressor values for individuals with missing Y. If instead those individuals are missing completely from the data, then it is generally not possible to correct for sample selection bias, although knowing how many individuals are missing can sometimes help us calculate an upper bound for the bias.

As with missing data, the reason behind the sample selection is crucial for whether it results in sample selection bias. For example, if individuals are "selected" into the sample at random (unrelated to their Y_i or X_i), then it's just like we're taking a random sample of a random sample, so we can just proceed as normal. However, if individuals are selected into the sample based on their Y_i, then OLS (and other estimators) can be very biased. A common problem for surveys is **non-response bias**: people who respond to the survey (the population studied) are not representative of the population of interest, differing in important ways compared to people who do not respond to the survey.

Example 12.7. Similar to Figure 12.3, imagine Y is wage, and individuals with high wage are less likely to take a survey at all. If our dataset only shows individuals who did take the survey, then sample selection bias is likely. The picture is basically the same as Figure 12.3, just that the "missing" data points are now entirely unobserved (those i are not even in our sample).

Example 12.8 (Kaplan video)**.** Perhaps the most famous (Nobel Prize-winning) economic example of "sample selection" is from Heckman (1979), although it is actually the "missing Y" meaning: wages are only observed for currently employed individuals, but our data usually includes unemployed individuals, too. Imagine we want to learn what determines the wage that an individual is offered by a firm. However, if the wage a firm is willing to pay is below the individual's reservation wage or a legal minimum wage, then the individual won't or can't take the offer. But if they don't work, then we can't observe that hypothetical wage. Further, the population of individuals who are working (and thus have an observable wage) is clearly not just a random sample from the full population of interest that includes non-employed individuals.

Methods to address sample selection bias are beyond our scope, but you can at least try to think critically about whether sample selection bias might be an issue in real-world examples.

12.3.7 Omitted Variable Bias and Collider Bias

Omitted variable bias is discussed in Sections 9.1 and 10.1. It is very common with observational economic data: many variables are (cor)related in economics, and many important ones are difficult to measure (human capital, technology, marginal cost, etc.). If they are actually observed in the data, then they can just be included, although recall that including colliders actually makes bias worse (Section 9.6). If not, then other methods can be used under certain specific conditions. For example, difference-in-differences (Section 9.7) allows certain types of omitted variables. Other estimators with panel data (observations for the same unit i over multiple time periods) also allow certain types of omitted variables, like those that do not change over time. However, these and yet other estimators that address omitted variable bias are beyond our scope.

12.3.8 Simultaneity and Reverse Causality

When we regress Y on X, we often (perhaps subconsciously) assume that X may have a causal effect on Y, but that Y does not have an effect on X. However, sometimes in reality Y affects X, too. This is called **reverse causality** or **simultaneous causality**.

The issue of **simultaneity** is basically the same (and often synonymous), but emphasizes that it is not necessarily a direct causal effect of Y on X, just that X and Y are determined by the same system at the same time (simultaneously). Economic systems are often complex, where conditions "determine" the values of multiple variables at the same time. For example, supply and demand curves simultaneously determine the equilibrium market price and quantity. Rather than trying to say price affects quantity and quantity affects price (simultaneous causality), it's more precise to say that price and quantity are determined simultaneously by the same economic system (simultaneity).

Because economists often study systems with complex interactions among many variables, and with observational data, simultaneity and reverse causality are common.

Example 12.9 (Kaplan video)**.** One question economists have studied is the effect of police officers per capita X on crime rate Y in a city. (Note: as with other examples like minimum wage and right-to-work laws, this has nothing to do with "good" or "bad," but only how simplistic econometric analysis can fail to have a causal interpretation.) Of course, it is possible that the density of police has a causal effect on crime rate. But it is also possible that crime rate Y has a causal effect on X, through policy decisions. That is, all else equal, cities with very low crime rate tend to decide to hire fewer police officers (per capita) than cities with higher crime rates. The decision about X (officers per capita) is determined partly by Y (crime rate). Even if X has zero effect on Y, we would see a positive correlation in the data if higher Y tends to cause cities to choose higher X.

With simultaneity or reverse causality, OLS regression of Y on X does not consistently estimate structural or treatment effects. In the police example, even if there were zero effect of X on Y, the response of X to Y would cause positive correlation between X and Y (cities with more crime would have more police), i.e., OLS estimates a positive slope that falsely suggests a positive effect.

There are methods like instrumental variables that can (sometimes) solve the problem of simultaneity or reverse causality, but they are beyond our scope. For now, you can just try to think critically about whether or not simultaneity or reverse causality is a problem in real-world examples.

Discussion Question 12.5 (health and medical expenditure)**.** You want to learn the causal effect of how much an individual spends on medical insurance and care (X, dollars per year) on health (Y, higher value means healthier).

a) Explain why a regression of Y on X would not estimate this causal effect.
b) Would the regression slope be higher or lower than the causal effect? Why?

Optional Resources

Optional resources for this chapter

- Sample selection from survey non-response (Masten video)
- External validity (Masten video)
- Missing data approaches (Masten video)
- Reverse causality and simultaneity (Masten video)
- Reverse causality example: violence (Lambert video)
- Reverse causality example: HDI (Lambert video)
- Greater external validity for "structural" results (Masten video)
- Sections 9.2 ("Measurement Error") and 9.3 ("Missing Data and Nonrandom Samples") in Heiss (2016)
- Chapter 22 ("Missing Data") in Kaplan (2020)
- Chapter 9 ("Assessing Studies Based on Multiple Regression") and Section 13.2 ("Threats to Validity of Experiments") in Hanck et al. (2018)

Empirical Exercises

Empirical Exercise EE12.1. You will analyze data from Rouse (1998) on a "school voucher" program in Milwaukee, Wisconsin. As Rouse (1998) explains, "In 1990 Wisconsin began providing vouchers to a small number of low-income students to attend nonsectarian private schools." Wooldridge notes that many observations with missing data have already been dropped, so there is sample selection. He also notes you can use variable `mnce90` to try to control for this, but `mnce90` is missing for 2/3 students, so then there's a missing data problem, too. If everything were perfect, the estimated ATE of eligibility (binary variable `select`) shouldn't depend too much on the control variables or the subsample of individuals; but clearly it does.

a. Load and see a description of the data.

   R: `library(wooldridge)` and `?voucher`

   Stata:

   ```
   use https://raw.githubusercontent.com/kaplandm/stata/main/data/
      voucher.dta , clear
   describe
   ```

b. R only: copy the dataset into data frame `df` with `df <- voucher`

c. Display the total number of observations (rows) in the dataset.

   R: `nrow(df)`

   Stata: `count`

d. Display summary statistics of `mnce90` and `mnce`, including the number of missing observations.

   R: `summary(df[,c('mnce','mnce90')])`

   Stata: `count if missing(mnce90)` and `summarize mnce mnce90`

e. Run a simple regression of `mnce` (the 1994 math test score) on `select` (the dummy variable for whether a child was ever allowed to use a voucher).

   R: `(ret1 <- lm(mnce~select, data=df))`

   Stata: `regress mnce select , vce(robust)`

f. Repeat but adding the 1990 math test score `mnce90` as a regressor. Also, compare the number of observations used in the regression to the total number of observations in the dataset.

   R: `(ret2 <- lm(mnce~select+mnce90, data=df))` and then `length(ret2$residuals)` or `summary(ret2)` to see the number of observations actually used.

   Stata: `regress mnce select mnce90 , vce(robust)` noting that observations with missing `mnce90` are automatically (and silently) omitted from the regression, but the output shows the number of observations actually used, which you can compare to the number in the full dataset.

g. To try to see how much of the estimate's change is due to controlling for mnce90 versus sample selection bias, re-run your first simple regression but with only the observations used in the second regression, i.e., only observations with non-missing mnce90.

   R: `(ret2b <- lm(mnce~select, data=df[!is.na(df$mnce90),]))`

   Stata: `regress mnce select if !missing(mnce90) , vce(robust)`

h. Optional: repeat the above three regressions but with selectyrs (number of years eligible for voucher program) instead of the binary select

i. Optional: repeat the first three regressions but with additional regressors like female to see if they further change the coefficient on select

Empirical Exercise EE12.2. You will analyze data from Card (1995), first seen in EE3.1, with individual-level observations of wages, years of education, and other variables. You'll focus on the relationship between wage and education. The variable IQ seems like a helpful control variable, but it is not observed for all individuals, which may cause bias depending on why it is missing. You'll estimate the coefficient on education with different sets of regressors and different subsets of data. You'll also look at the difference it makes using the sampling weights (as you should).

a. Load and see a description of the data.

   R: `library(wooldridge)` and `?card`

   Stata: `bcuse card , clear`

b. R only: copy the dataset into data frame df with `df <- card`

c. Display the total number of observations (rows) in the dataset.

   R: `nrow(df)`

   Stata: `count`

d. Show how many observations are missing IQ.

   R: `table(is.na(df$IQ))`

   Stata: `count if missing(IQ)`

e. Run a simple regression of log wage on years of education.

   R: `(ret1u <- lm(log(wage)~educ, data=df))`

   Stata: `regress lwage educ , vce(robust)`

f. Run the same regression but with the provided weights.

   R: `(ret1w <- lm(log(wage)~educ, data=df, weights=weight))`

   Stata: `regress lwage educ [pweight=weight] , vce(robust)`

g. Run the same simple weighted regression but with the subset of observations for which IQ is observed.

   R: replace `df` with `df[!is.na(df$IQ),]`

   Stata: add `if !missing(IQ)` after `educ` (with a space on either side)

h. Regress log wage on education and IQ (which automatically uses only observations where IQ is non-missing).

   R: `(ret2w <- lm(log(wage)~educ+IQ, data=df, weights=weight))`

   Stata: `regress lwage educ IQ [pweight=weight] , vce(robust)`

i. Optional: repeat parts (f)–(h) but with additional regressors of your choice.

Part III

Time Series

Introduction

Part III concerns time series data and models. The focus is on **forecasting**: prediction of future values or events. Also, foundational concepts like (non)stationarity and autocorrelation are introduced.

Related (free) material is from Diebold (2018b) and Hanck et al. (2018, Ch. 14). Chapter 1 in the DataCamp intro time series course is also free.

Chapter 13

Time Series: One Variable

⟹ Kaplan video: Chapter Introduction

Chapter 13 extends Chapter 2 to the time series setting. New concepts like stationarity and autocorrelation are introduced. There are even new complications just with estimating a variable's mean and computing a confidence interval.

Unit learning objectives for this chapter

13.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

13.2. Identify and describe different components and properties of a time series [TLOs 2 and 3]

13.3. Interpret transformed and decomposed time series [TLOs 2 and 3]

13.4. In R (or Stata): estimate basic descriptions of a time series [TLO 7]

13.5. In R (or Stata): decompose a time series into different components [TLO 7]

13.1 Terms and Notation

A **time series** of a single variable is written as Y_t for time periods $t = 1, \ldots, T$. For example, Y could be annual GDP of the U.S., with $t = 1$ indicating the year 2001 and $T = 10$ indicating a total of ten years of data (here 2001, 2002, ..., 2010). Or, Y could be quarterly GDP from 2001Q1 (year 2001, quarter 1) through 2010Q4, a total of $T = 40$ periods where $t = 1$ is 2001Q1, $t = 2$ is 2001Q2, $t = 9$ is 2003Q1, etc. Or, Y could be the weekly return on a certain stock observed over a single calendar year, $t = 1, \ldots, 52$.

In practice, there are many possible complications with timing and measurement, although details are beyond our scope. First, instead of "discrete time" periods $t = 1, \ldots, T$, "continuous time" models let t be any real (decimal) number, not just integers. Second, even with discrete time, the periods may be of different lengths. Third, even with equal discrete periods, it is important to know precisely when and how the "time t" observation is measured. For example, imagine annual data, where t represents an entire year. Is Y_t measured on January 1 of year t? Or December 31? Or is Y_t the average value across the entire year? Such timing is particularly important when analyzing multiple time series. For example, if Y_t is measured on January 1 of year t, but X_t is measured on December 31, then X_t is measured 364 days after Y_t but only 1 day before Y_{t+1}.

Similar to physics, the **sampling frequency** is the inverse of the length of each time period. For example, if each period is one year, then there is one observation per year, so the sampling frequency is yearly (or "annual"). If each period is one quarter, then the sampling frequency is quarterly. Similarly, time series can be monthly, weekly, daily, or even hourly or higher frequency (like for stock prices, website traffic, energy use, etc.).

The following terms describe relationships among observations. Relative to Y_t, the **first lag** (or first lagged value) is Y_{t-1}, i.e., the value from the immediately prior period. Similarly, the second lag is Y_{t-2}, and generally the jth lag is Y_{t-j}. The **first difference** is

$$\Delta Y_t \equiv Y_t - Y_{t-1}. \tag{13.1}$$

(But "second difference" does not refer to $Y_t - Y_{t-2}$.) Looking to the future, Y_{t+1} is the **first lead** (of Y_t), and Y_{t+j} is the jth lead. In many cases, modeling the relationship between Y_{t+1} and Y_t is equivalent to modeling Y_t and Y_{t-1}, for example. If we use observations $Y_1, \ldots, Y_T$ for estimation, then anything in the period $t = 1, \ldots, T$ is called **in-sample**, as opposed to $t = T+1, T+2, \ldots$, which is **out-of-sample**. Sometimes, fewer than T observations are used for estimation, and the definitions are adjusted accordingly (Section 15.2).

Example 13.1 (Kaplan video)**.** You have data on the daily electricity consumption of one household for one year. Specifically, period $t = 1$ is January 1, 2021; $t = 2$ is January 2, 2021; up to $t = 365$ is December 31, 2021, so the sample size is $T = 365$ observations. Thus, the full sample is Y_t for $t = 1, \ldots, 365$; Y_1 is the household's electricity consumption on January 1, 2021, and generally Y_t is the household's electricity consumption on day t. Observations Y_t and Y_{t+7} are the same day of the week (like Wednesday), one week apart. The out-of-sample value Y_{T+1} is for January 1, 2022. The first difference is the change in electricity consumption from day $t-1$ to day t, $\Delta Y_t = Y_t - Y_{t-1}$; a positive value indicates an increase in electricity consumption, whereas a negative value indicates a decrease.

13.2 Populations, Randomness, and Sampling

⟹ Kaplan video: Time Series Populations

We continue the perspective of Y_t as a random variable, just as Y_i was earlier (Sections 2.1 and 2.3). Earlier, Y_i was "random" because we could have sampled a different value from the population. But, what is the "population" for a time series?

One view is like the superpopulation from Section 2.2. That is, we can imagine many (infinite) possible universes. In each, there are the same mechanisms underlying how the time series values are generated, but the actual numerical values differ across universes. Like before, $\mathrm{E}(Y_t)$ is the average of the Y_t values across all the different universes. Similarly, $\mathrm{Var}(Y_t)$ is the variance across universes. Measures like $\mathrm{Corr}(Y_t, Y_{t+1})$ show whether Y_t and Y_{t+1} tend to both be high (or low), or opposite, or unrelated. For example, maybe GDP growth is high in both 2018 and 2019 in many universes, and low in both in other universes, but very few universes have high growth in 2018 and low in 2019, or low and then high. Then, in the (super)population, $\mathrm{Corr}(Y_{2018}, Y_{2019}) > 0$.

Another view is that we observe a sequence of T values within an infinitely long sequence of Y_t. We could think about what the sample average would be if we had a very long sequence, or other "asymptotic" properties.

13.3 Stationarity

⟹ Kaplan video: Stationarity

Will the future be like the past? This question arose in Section 12.2, on external validity. Here, "be like" is formalized in terms of probability distributions.

A time series Y_t is **stationary** if its future is like its past, probabilistically. A necessary (but not sufficient) aspect of this is $\mathrm{E}(Y_t) = \mathrm{E}(Y_s)$ for any time periods t and s: the mean never changes. Likewise, the median never changes, nor the standard deviation; the entire distribution of Y_t is identical to that of Y_s. Further, the relationship between this time period and next period must be stable over time, i.e., the joint distribution of (Y_t, Y_{t+1}) is identical for all t. Similarly, the joint distribution of the previous, current, and next periods' values, (Y_{t-1}, Y_t, Y_{t+1}), never changes. In full, stationarity is defined as the joint distribution of $(Y_{t-J}, \dots, Y_t)$ not depending on t, for any J.

The foregoing describes **strict stationarity** (also called **strong stationarity**); a "weaker" concept called **covariance stationarity** (also called **wide-sense stationarity** or **weak-sense stationarity**) requires only the means and autocovariances (Section 13.4) to be the same at all t, not the full joint distributions. Technically, it is not "weaker" in the logical sense (Section 6.1.1) because of weird distributions whose mean is undefined (e.g., Cauchy), but if you assume Y_t has finite variance, then strict (strong) stationarity implies covariance (weak) stationarity. That is, given finite variance, all strictly stationary series are also covariance stationary, but some covariance stationary series are not strictly stationary.

Example 13.2 (Kaplan video)**.** In every minute $t = 1, 2, 3, \dots$, a coin is flipped to generate $Y_t = 1$ if heads and $Y_t = 0$ if tails, with $\mathrm{P}(Y_t = 1) = 0.5$. This series is strictly stationary: $(Y_{t-J}, \dots, Y_t)$ consists of independent binary random variables all with the same probability of equaling 1, regardless of t, for any J.

Example 13.3 (Kaplan video)**.** Consider independent coin flips Y_t with $\mathrm{P}(Y_t = 1) = 0.5$, and then define the time series $Z_t = Y_t + Y_{t-1}$. For any t, using the fact that the coin flips are independent and have the same 0.5 probability of $Y_t = 1$: $\mathrm{P}(Z_t = 2) = \mathrm{P}(Y_t = Y_{t-1} = 1) = (0.5)(0.5) = 0.25$ and $\mathrm{P}(Z_t = 0) = \mathrm{P}(Y_t = Y_{t-1} = 0) = (0.5)(0.5) = 0.25$, so $\mathrm{P}(Z_t = 1) = 1 - \mathrm{P}(Z_t = 2) - \mathrm{P}(Z_t = 0) = 1 - 0.25 - 0.25 = 0.5$. That is, the distribution of Z_t is the same for any t. The joint distribution of (Z_t, Z_{t-1}) is also the same for any t, as is the joint distribution of $(Z_t, \ldots, Z_{t-J})$ for any J; for example, $\mathrm{P}(Z_t = Z_{t-1} = 2) = \mathrm{P}(Y_{t-2} = Y_{t-1} = Y_t = 1) = (0.5)(0.5)(0.5) = 0.125$, regardless of t, and other calculations follow similarly.

With either type of stationarity, an estimate of $\mathrm{E}(Y_t)$ from historical data can be interpreted as an estimate of the future $\mathrm{E}(Y_{T+1})$, which is the (unconditional) best prediction of Y_{T+1} under quadratic loss. Stationarity essentially assumes external validity over time, allowing us to extrapolate the past into the future. In Chapters 14 and 15, we'll improve upon the unconditional forecast by incorporating other information, but stationarity (and its variations) remain important considerations for external validity.

In practice, you should not blindly assume stationarity, but examine it empirically and economically. That is, you can look at the data to see if it appears stationary, and you can also think about what is happening in the world now that may change the future behavior. A previously stationary time series may no longer be stationary if there is a sudden law change or other event with permanent effect.

Section 13.6 contains more on data that's nonstationary, i.e., not stationary.

13.4 Autocovariance and Autocorrelation

⟹ Kaplan video: Autocorrelation

An important feature of a time series is the correlation between this period's value and last period's value, i.e., between Y_t and Y_{t-1}. This correlation is called the first **autocorrelation** or **serial correlation**.

The first autocorrelation can be positive, negative, or zero. For example, if today's price change is not systematically related to yesterday's price change, then the time series of price changes has zero autocorrelation. If high quarterly GDP growth follows high growth, and low follows low, rather than jumping around randomly each quarter, then GDP growth has a positive autocorrelation. Conversely, negative first autocorrelation implies high values are followed by low values, and low by high, more often than high following high or low following low. In economics, positive autocorrelation is most common.

The sampling frequency affects the first autocorrelation. Usually, especially after adjusting for seasonality (Section 13.6.2), first autocorrelations are closer to positive one with high frequency and closer to zero with low frequency. For example, today's U.S. unemployment rate will be extremely close to yesterday's rate, so the first autocorrelation is near one with daily data. However, with yearly data (lower frequency), the first autocorrelation is lower. If each period is one decade (even lower frequency), then the first autocorrelation may be near zero.

Generally, for a stationary series, the jth autocorrelation (or jth **autocorrelation coefficient**) ρ_j describes the relationship between Y_t and Y_{t-j}, as does the related jth **autocovariance** γ_j. Stationarity implies these values do not vary with t, only j (the lag). Consequently, it is the same (statistically) if we look j periods in the past or j periods in the future, because period $t-j$ is j periods before t just as t is j periods before $t+j$, and $\mathrm{Cov}(W,Z)=\mathrm{Cov}(Z,W)$. Mathematically,

$$\gamma_j \equiv \mathrm{Cov}(Y_t, Y_{t-j}) = \mathrm{Cov}(Y_{t+j}, Y_t) = \gamma_{-j}, \tag{13.2}$$

$$\rho_j \equiv \mathrm{Corr}(Y_t, Y_{t-j}) = \mathrm{Corr}(Y_{t+j}, Y_t) = \rho_{-j}, \tag{13.3}$$

$$\sigma_Y^2 \equiv \mathrm{Var}(Y_t), \tag{13.4}$$

$$\gamma_0 \equiv \mathrm{Cov}(Y_t, Y_t) = \mathrm{Var}(Y_t) = \sigma_Y^2, \quad \rho_0 = \mathrm{Corr}(Y_t, Y_t) = 1, \tag{13.5}$$

$$\rho_j \equiv \mathrm{Corr}(Y_t, Y_{t-j}) = \frac{\mathrm{Cov}(Y_t, Y_{t-j})}{\sqrt{\mathrm{Var}(Y_t)\,\mathrm{Var}(Y_{t-j})}} = \frac{\gamma_j}{\sigma_Y^2} = \frac{\gamma_j}{\gamma_0}. \tag{13.6}$$

In (13.6), the denominator simplifies because stationarity implies $\mathrm{Var}(Y_{t-j}) = \mathrm{Var}(Y_t) = \sigma_Y^2$, and $\sigma_Y^2 = \gamma_0$ from (13.5):

$$\sqrt{\mathrm{Var}(Y_t)\,\mathrm{Var}(Y_{t-j})} = \sqrt{\sigma_Y^2 \sigma_Y^2} = \sigma_Y^2 = \gamma_0.$$

Although sometimes autocovariances are more convenient mathematically, autocorrelations are easier to interpret. The units of autocovariance are the square of the units of Y_t (like "squared dollars"), which is difficult to interpret. The autocorrelation does not depend on the units of Y_t and has the same interpretation as a correlation, where possible values are between -1 (perfect negative linear correlation) and $+1$ (perfect positive linear correlation). The usual caveats about interpreting correlation (nonlinearity, causality, magnitude of change, etc.) apply equally to autocorrelation.[1]

Discussion Question 13.1 (autocorrelation)**.** For each of the following, explain why you think $\rho_1 > 0$, $\rho_1 \approx 0$, or $\rho_1 < 0$.

a) An individual's employment status ($Y_t = 1$ if employed at time t, otherwise $Y_t = 0$), observed weekly.
b) GDP growth, annual.
c) GDP growth, quarterly.
d) Seasonally-adjusted GDP growth, quarterly.

13.5 Estimation

In R, you can estimate the mean with `mean()` and estimate autocorrelations with `acf()`. There are certain conditions required for these to be good estimators, but the most

[1] E.g., https://en.wikipedia.org/wiki/Correlation_and_dependence

important is that the series is indeed stationary. Otherwise, if $\mathrm{E}(Y_t)$ changes with t, then we cannot hope to estimate "the" mean; or if $\mathrm{Corr}(Y_t, Y_{t-1})$ changes with t, then we cannot estimate "the" first autocorrelation; etc.

Example 13.4. The following code estimates autocorrelations for monthly international airline passenger data. The result $\hat{\rho}_{12} > \hat{\rho}_6$ seems surprising at first: Y_t is more strongly correlated with Y_{t-12} than Y_{t-6}, even though Y_{t-6} is closer in time. However, these are monthly data with strong seasonality (Section 13.6.2), so the fact that $t-12$ is the same calendar month as t causes stronger correlation than with $t-6$, which is a very different season; for example, $t-6$ is summer if t is winter. (Because of the seasonality, clearly $\mathrm{E}(Y_t)$ changes with t; do the autocorrelations also change with t? It might be better to remove the seasonality first, but that's later in Section 13.7.) To see a graph, run the code yourself with `plot=TRUE` instead of `FALSE`.

```
retcorr <- acf(AirPassengers, lag.max=12, type='correlation',
               plot=FALSE, ci.type='ma')
retcov <- acf(AirPassengers, lag.max=12, type='covariance',
              plot=FALSE, ci.type='ma')
print(data.frame(lagmonth=0:12, rho.j=round(retcorr$acf,digits=2),
                 gamma.j=round(retcov$acf,digits=0)), row.names=F)
```

```
##  lagmonth rho.j gamma.j
##         0  1.00   14292
##         1  0.95   13549
##         2  0.88   12514
##         3  0.81   11529
##         4  0.75   10757
##         5  0.71   10201
##         6  0.68    9743
##         7  0.66    9474
##         8  0.66    9370
##         9  0.67    9589
##        10  0.70   10043
##        11  0.74   10622
##        12  0.76   10868
```

There are limits to what we can learn from data. For an extreme example, consider ρ_j for $j = T$: because we do not observe any two observations T periods apart (the earliest observation is Y_1, but we don't observe Y_{1+T}), we cannot learn anything about ρ_T. More generally, it is difficult to learn about γ_j for large j (near T).

13.6 Nonstationarity

⟹ Kaplan video: Nonstationarity

This section describes the most common reasons a time series is **nonstationary**, i.e., not stationary.

In Sum: Reasons for Nonstationarity

Stochastic trend (unit root; e.g., random walk): variance increasing over time
Deterministic trend: mean changing over time
Seasonality: mean changing over time (repeating up-and-down pattern)
Cycles: up-and-down patterns without fixed frequency
Breaks: permanent changes

13.6.1 Trends

Stochastic Trends

A **random walk** as in (13.7) generates nonstationary Y_t. This is a special case of a more general **unit root** process, which all share qualitatively similar properties (including nonstationarity). It is also sometimes called a **stochastic trend**. Let Y_0 be the initial value. Let

$$Y_t = Y_{t-1} + \epsilon_t, \tag{13.7}$$

where the increments ϵ_t are iid, mean zero, and independent of all past values Y_s for $s \leq t-1$; i.e., the ϵ_t are **independent white noise** (Section 14.1).

One way to see the nonstationarity is that

$$\mathrm{Var}(Y_t) = \mathrm{Var}(Y_{t-1} + \epsilon_t) = \mathrm{Var}(Y_{t-1}) + \overbrace{\mathrm{Var}(\epsilon_t)}^{>0} + 2\overbrace{\mathrm{Cov}(Y_{t-1}, \epsilon_t)}^{=0 \text{ since } Y_{t-1} \perp\!\!\!\perp \epsilon_t} > \mathrm{Var}(Y_{t-1}), \tag{13.8}$$

violating the property of stationarity that the variance is the same at all t. Logically, stationarity implies same variance at all t, so by the contrapositive, different variance at difference t implies nonstationarity.

For prediction, given (13.7), the "best" guess (under quadratic loss) of next period's Y_{t+1} is the current period's Y_t. Here, Y_t contains all the relevant historical information about the future Y_{t+1}; additionally knowing Y_{t-1} or other past values does not help.

Although nonstationary, the random walk can be transformed into a stationary process by taking a first difference (Section 13.1). Subtracting Y_{t-1} from both sides of (13.7),

$$Y_t - Y_{t-1} = Y_{t-1} + \epsilon_t - Y_{t-1} = \epsilon_t, \tag{13.9}$$

and ϵ_t is iid, which implies stationarity. Because the first difference is stationary, the original time series Y_t is called **difference stationary**.

Deterministic Trends

With a **deterministic trend**, the time series goes up (or down, or up and down, etc.) in a non-random pattern. Analogous to difference stationarity, a time series is **trend stationary** if removing its deterministic trend produces a stationary series.

Example 13.5 (Kaplan video)**.** If $Y_t = t + \epsilon_t$ with $\mathrm{E}(\epsilon_t) = 0$, then

$$\mathrm{E}(Y_t) = \mathrm{E}(t + \epsilon_t) = t + \mathrm{E}(\epsilon_t) = t + 0 = t,$$

which changes with t, violating stationarity. The detrended series is $Y_t - t = \epsilon_t$, so Y_t is trend stationary if ϵ_t is stationary.

Distinguishing Trend Types

Despite their seeming so different, in practice it can be difficult to distinguish a stochastic trend from a deterministic trend. For example, in climate econometrics,[2] there is ongoing debate about whether the earth's temperature currently has a stochastic trend or a deterministic trend that changed at some point in the past; e.g., see Kaufmann, Kauppi, and Stock (2010), Chang, Kaufmann, Kim, Miller, Park, and Park (2020), and references therein.

However difficult, it is important to distinguish stochastic and deterministic trends because they affect forecasts. Roughly, a trend stationary time series is expected to return to its deterministic trend line relatively quickly, whereas the stochastic trend makes deviations more persistent.

Further details of optimal forecasting with trends, unit root testing, and other topics are interesting but beyond our scope.

13.6.2 Seasonality

A time series with **seasonality** tends to have higher values in certain time periods ("seasons") than in others. The seasons could be certain months during a calendar year, days of the week, hours of the day, or other periods within a repeating cycle. Seasonality can be due to human-imposed seasons (holidays, school schedules, elections, etc.) or natural seasons (weather, crops, sunlight, etc.).

Example 13.6. Here are a few brief examples.

- Retail sales are highest near the Christmas holiday season.
- Some agricultural crops are only harvested in one season of the year.
- Restaurant dinner sales are higher on Friday and Saturday than other days.
- Crime rates fluctuate with the day of the week and with the hour of the day.

[2] Although not exactly climate econometrics, half the 2018 Nobel Prize was awarded to William Nordhaus "for integrating climate change into long-run macroeconomic analysis"; see https://www.nobelprize.org/prizes/economic-sciences/2018/press-release/

The presence of seasonality depends on the length of time period for each time series observation. If each observation aggregates all seasons within a cycle, then the time series will not show seasonality.

Example 13.7 (Kaplan video)**.** Let Y_t be retail sales in time period t. If Y_t is quarterly (each t is one quarter of the year), then seasonality appears: Y_t always jumps up during the fourth quarter (October, November, December). That is, if Y_1 is the first quarter of some year, then we expect generally higher retail sales for Y_4, Y_8, Y_{12}, etc., than for Y_1, Y_2, Y_3, Y_5, Y_6, etc. However, if we aggregate over each year, then we do not expect $Y_1 + Y_2 + Y_3 + Y_4$ to be any higher or lower than $Y_5 + Y_6 + Y_7 + Y_8$ due to seasonality: both sums contain one "Christmas season" along with one of every other type of season. Thus, if instead we observe annual Y_t (each t is one full year), then there is no seasonality. If instead t is divided into periods shorter than a quarter, then seasonality is still seen: with monthly data, Y_t jumps up in November and especially December, or with weekly data, Y_t jumps up for several weeks leading up to Christmas.

Some "seasons" are not actually seasons with a fixed frequency, so they must be handled differently. For example, the calendar date of Easter differs from year to year. For forecasting regression models, you can add dummy variables for such events. For Easter specifically, the function `easter()` in the `forecast` package is helpful.

Example 13.8. Figure 13.1 illustrates how seasonality can be seen in plots of Y_t over t that show an up-and-down pattern that repeats every year (or other period). The left graph is from `plot(AirPassengers)` and shows monthly numbers of international airline passengers (in thousands). There is a clear up-and-down seasonal pattern that repeats every year. You can also try using `seasonplot(AirPassengers)`, a function in the `forecast` package (Hyndman et al., 2020; Hyndman and Khandakar, 2008). The right graph of Figure 13.1 is from `plot(log(AirPassengers))` and shows $\ln(Y_t)$ against t. Although both show seasonality, the peak-to-trough magnitude (height) of the seasonal variation is more constant every year for $\ln(Y_t)$; see Section 13.7.

13.6.3 Cycles

What about up-and-down patterns caused by macroeconomic business cycles, or El Niño–Southern Oscillation cycles? Cycles are often important but more difficult to understand. One added difficulty is the unknown and changing length of cycles; e.g., El Niño does not come precisely every five years, nor is there a recession every five years. Here, like in Hyndman and Athanasopoulos (2019, §6), the "trend" is actually a **trend–cycle component** that includes cycles, too. Though beyond our scope, it can be helpful to explicitly split out cycles; for more on cycles, see for example Diebold (2018b, §§6–7).

13.6.4 Structural Breaks

Sometimes there are big, permanent changes in the world, and the properties of a time series also change permanently. This is often called a **structural break**. For example,

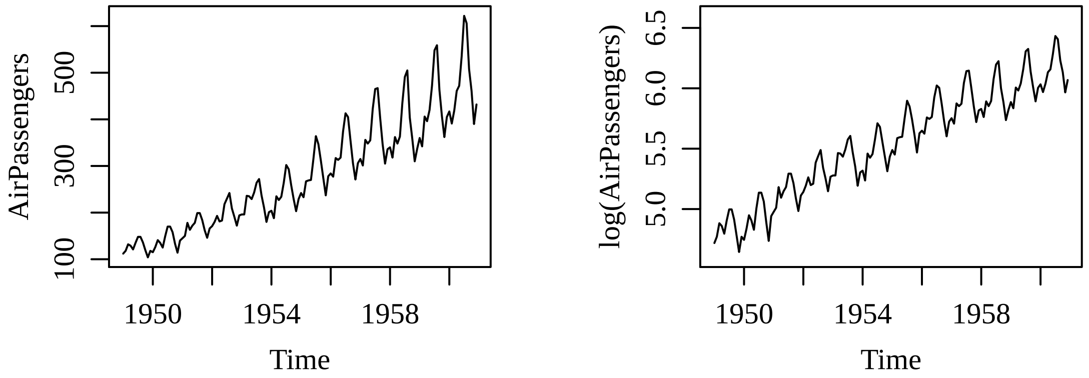


Figure 13.1: Seasonality in international air travel.

in the U.S., many macroeconomic time series look very different before and after 1985; in particular, the reduction in volatility led to the term "Great Moderation."[3] Dealing with breaks is beyond our scope, but they are important to be aware of; see also Section 14.7.

13.7 Decomposition

⟹ Kaplan video: Decomposition

The observed time series Y_t can be written in terms of unobserved components of "trend" (really trend–cycle), seasonality, and a remainder (Diebold, 2018b, §2.10). The **remainder**, also called the random or irregular or residual or noise component, is what remains of Y_t after removing the trend and seasonality.

Notationally, following Hyndman and Athanasopoulos (2019, §6), let T_t denote trend, S_t seasonality, and R_t remainder. Then,

$$R_t \equiv Y_t - T_t - S_t \implies Y_t = T_t + S_t + R_t. \tag{13.10}$$

This is an additive decomposition: Y_t is "decomposed" into additive trend, seasonality, and remainder components, which all have the same units as Y_t.

Alternatively, a multiplicative decomposition is

$$Y_t = T_t \times S_t \times R_t. \tag{13.11}$$

Now, T_t still has the same units as Y_t, but S_t and R_t represent percentage deviations from the trend. For example, $S_t = 1.05$ means 5% higher, or $R_t = 0.85$ means 15% lower. (Often a percentage seasonal component makes more sense.) Actually, taking the log of both sides of (13.11) yields an additive model:

$$\ln(Y_t) = \ln(T_t) + \ln(S_t) + \ln(R_t). \tag{13.12}$$

[3] See https://en.wikipedia.org/wiki/Great_Moderation

Finally, sometimes the decomposition is a mix: $Y_t = T_t \times S_t + R_t$.

There are R functions to decompose time series into trend, seasonal, and remainder components. To choose the right method, you must decide whether the seasonality is additive or multiplicative. For example, compared to sales on July 1, are sales on December 1 usually higher by \$500 (additive), or by 30% (multiplicative)? In other words, is (13.10) or (13.11) more sensible?

For intuition, the following roughly describes a **classical additive decomposition** (Hyndman and Athanasopoulos, 2019, §6.3). First, the trend is estimated, usually by some nonparametric smoother, yielding the estimated trend $\hat{T}_t$. Second, the "seasonal" averages of $Y_t - \hat{T}_t$ (the **detrended** data) are computed. For example, with monthly data, all January values of $Y_t - \hat{T}_t$ are averaged to estimate $\hat{S}_t$ when t is in January, and then all February values are averaged to get $\hat{S}_t$ for February t, etc. Third, $\hat{R}_t = Y_t - \hat{T}_t - \hat{S}_t$. There are many variations, with different estimators of $\hat{T}_t$, or allowing $\hat{S}_t$ to change over time. For **multiplicative decomposition**, either apply the above to $\ln(Y_t)$, or replace subtraction with division: use $Y_t/\hat{T}_t$ in the second step, and $Y_t/(\hat{T}_t\hat{S}_t)$ in the third step.

Example 13.9. Figure 13.2 shows an additive decomposition produced by the following R code that uses `decompose()` (in the built-in `stats` package).

```
par(family='serif', mgp=c(2.1,0.8,0))
ret <- decompose(co2, type='additive')
plot(ret)
```

Example 13.10. Figure 13.3 shows a multiplicative decomposition generated by the following R code. When seasonality is multiplicative instead of additive, specify `type='multiplicative'` as below.

```
par(family='serif', mgp=c(2.1,0.8,0))
ret <- decompose(AirPassengers, type='multiplicative')
plot(ret)
```

Other R decomposition functions to try (or Google) include `stl()`, `HoltWinters()`, and the `forecast` package's `mstl()` (multiple seasonal).

Discussion Question 13.2 (nonstationarity)**.** For each of the following time series, explain specifically why you doubt its strict stationarity: a) GDP, annual; b) stock market index, annual; c) world population, annual; and d) U.S. residential water usage, monthly (hint for non-US students: it's much hotter in summer, and many houses have yards/-gardens that require watering).

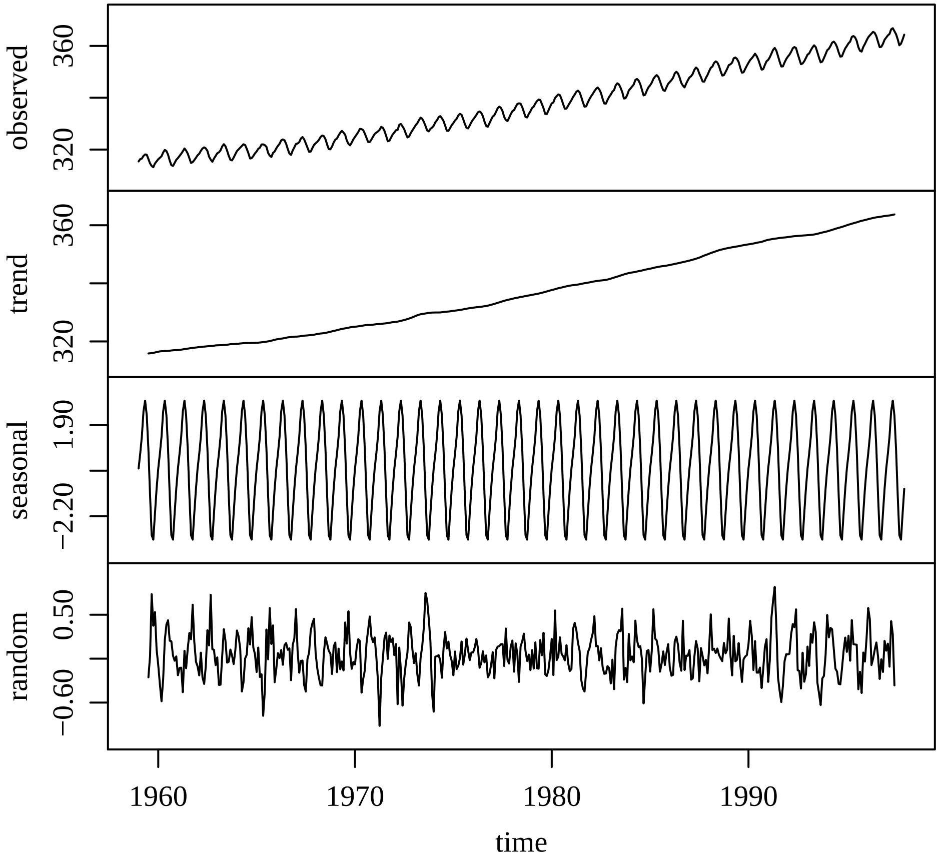


Figure 13.2: Additive decomposition, monthly atmospheric CO_2 (ppm).

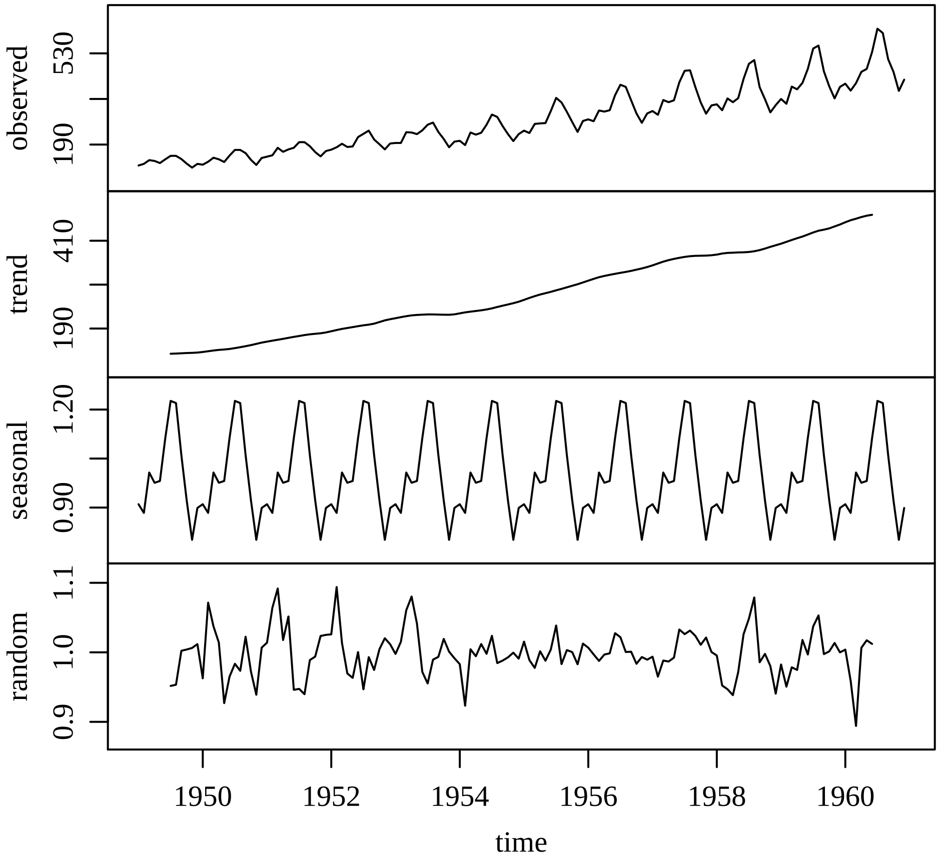


Figure 13.3: Multiplicative decomposition, monthly airline passengers (1000s).

13.8 Transformations

To improve interpretation or statistical properties, it may help to transform a time series before analyzing it. Three common transformations are now briefly discussed.

First, the first difference looks at changes in Y_t, defined in (13.1) as $\Delta Y_t \equiv Y_t - Y_{t-1}$. One motivation is Section 13.6: some nonstationary Y_t are difference stationary, so ΔY_t is stationary. For example, if $Y_t = Y_{t-1} + U_t$, where U_t is iid, then Y_t is a random walk and thus nonstationary. However, $\Delta Y_t = U_t$ is iid, which is stationary. Methods that only work with stationary data could be applied to ΔY_t but not Y_t.

Second, log transformations sometimes help, like in (13.12) where a multiplicative model becomes additive. That is, instead of Y_t, we analyze $Z_t = \ln(Y_t)$.

Third, taking a log difference $\ln(Y_t) - \ln(Y_{t-1})$ yields the compound growth rate. This is the first difference of the log-transformed series: letting $Z_t = \ln(Y_t)$, then $\Delta Z_t = Z_t - Z_{t-1} = \ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_t/Y_{t-1})$. For example, the formula for the final level A after continuously compounded growth at effective annual rate r for t years, starting at initial level P, is $A = Pe^{rt}$, the "Pert" formula you may have learned in high-school for computing compound interest rates. For a single year ($t = 1$ in the formula), the rate r is then solved by $A = Pe^r$ implying $e^r = A/P$ and thus $r = \ln(A/P) = \ln(A) - \ln(P)$, using a log property (from Section 8.1.1) for the last equality. Thus, with annual data, the log difference $\ln(Y_t) - \ln(Y_{t-1})$ represents the effective annual rate.

Optional Resources

Optional resources for this chapter

- Deterministic and stochastic trends (Lambert video)
- Chapter 14 ("Time Series") in Hansen (2020)
- Transformations: Section 3.2 ("Transformations and adjustments") in Hyndman and Athanasopoulos (2019)
- Seasonality and holidays: Section 5.4 ("Some useful predictors") in Hyndman and Athanasopoulos (2019)
- Trends, seasonality, and/or decomposition: Sections 8.1 ("Random Walks...") and 8.2 ("Stochastic vs. Deterministic Trend") in Diebold (2018c), Section 9.4 ("Stochastic and deterministic trends") in Hyndman and Athanasopoulos (2019), Section 14.7 ("Nonstationarity I: Trends") in Hanck et al. (2018), Chapter 5 ("Trend and Seasonality") in Diebold (2018b), Chapter 12 ("Trend and Seasonality") in Diebold (2018a), Chapter 6 ("Time series decomposition") in Hyndman and Athanasopoulos (2019), Section 3.6 ("Classical decomposition") in Holmes, Scheuerell, and Ward (2019), Sections 10.3.3–10.3.4 ("Trends" and "Seasonality") in Heiss (2016)
- Stationarity and random walk: Section 8.1 ("Stationarity and differencing") in Hyndman and Athanasopoulos (2019), Section 11.2 ("The Nature of Highly Persistent

Time Series," i.e., random walks) in Heiss (2016)

- Estimation of mean and autocovariances: Section 13.3 ("Estimation and Inference for the Mean, Autocorrelation and Partial Autocorrelation Functions") in Diebold (2018a)
- HAC standard errors: Section 15.4 ("HAC Standard Errors") in Hanck et al. (2018)
- Section 10.2 ("Time Series Data Types in R") in Heiss (2016)

Empirical Exercises

Empirical Exercise EE13.1. You will analyze monthly U.S. unemployment data. You'll notice that the unemployment rate is not very seasonal (by month), but it is very persistent (positively autocorrelated). Note that `urate` is in percent units, so 5.2 means 5.2%, etc.

a. Load and see a description of the data.

   R: `library(wooldridge)` and `?beveridge`

   Stata: `bcuse beveridge , clear`

b. Tell your software that you have monthly time series data.

   R: `tsdat <- ts(data=beveridge$urate, frequency=12, start=c(2000,12))` creates a time series variable named `tsdat` that's a time series (`ts`) with the unemployment rate data (`urate`) starting in year 2000 month 12 (the first value of `beveridge$month`). Argument `frequency=12` says there are 12 "seasons" before getting back to the first one; in this case, 12 different months per year. (Daily data could use `frequency=7` to allow day-of-week "seasonality.")

   Stata: `tsset ym , monthly`

c. R only: decompose (additively) the unemployment rate time series into trend, seasonal, and remainder components with `tsdec <- decompose(tsdat)` to compute and `plot(tsdec)` to plot. You can also see that the magnitude of the seasonal component is relatively small with `max(abs(tsdec$seasonal))`

d. Stata only: to additively decompose the time series, first estimate the trend component with a nonparametric "moving average smoother" with command

   ```
   tssmooth ma furate=urate , weights(1 2 2 2 2 2 <2> 2 2 2 2 2 1)
   ```

   and plot this smoothed trend against the raw time series with

   ```
   tsline urate furate , name(furate) ylabel(#3)
   ```

e. Stata only: compute the seasonal effects by averaging the difference between the data and the trend within each month (e.g., average among all January values, then separately among all February values, etc.). Generate the month variable with `generate month = month(dofm(ym))` and compute the within-month averages with `bysort month : egen seasadd  = mean(urate-furate)`

f. Stata only: normalize the seasonal effects to average to zero. Compute the average of the raw seasonal effects, and then subtract that value from the seasonal effects (to make them average to zero) with commands (note: the broken-up "line" `scalar normadd ...` should all be on the same line of code)

   ```
   sort ym
   scalar normadd = (seasadd[1]+seasadd[2]+seasadd[3]+seasadd[4]+
      seasadd[5]+seasadd[6]+seasadd[7]+seasadd[8]+seasadd[9]+
   ```

```
   seasadd[10]+seasadd[11]+seasadd[12])/12
replace seasadd = seasadd - normadd
```

g. Stata only: see how big (or small) the seasonal effects are with commands

```
list month seasadd if year(dofm(ym))==year(dofm(ym[1]))+1
summarize seas , detail
```

h. Stata only: generate the remainder term as the raw data minus trend minus seasonality, with command `generate remadd = urate - furate - seasadd`

i. Stata only: plot the seasonal and remainder series, and then make a combined graph with everything (similar to what R shows):

```
tsline seasadd , name(seasadd) ylabel(#3)
tsline remadd , name(remadd) ylabel(#3)
graph combine furate seasadd remadd , cols(1) name(decompurateadd)
```

j. Plot the autocorrelation function (ACF) up to 48 months lag.

   R: `acf(tsdat, lag.max=48, ci=0)`

   Stata: `ac urate , level(95) lags(48)`

k. Display the autocorrelation values up to 24 months.

   R: `acf(tsdat, lag.max=24, type='correlation', plot=FALSE)`

   Stata: `corrgram urate , lags(24) noplot`

l. Optional: repeat the decomposition plot and ACF plot for the vacancy rate variable `vrate`

Empirical Exercise EE13.2. You will analyze monthly data on industrial cement production from Shea (1993). If you're curious, you can view and download more recent cement data from the Federal Reserve Bank of St. Louis.[4] You'll notice that seasonality is very important. You'll also notice that the autocorrelations of the raw data reflect the up-and-down seasonality, whereas the autocorrelations of the seasonally-adjusted data show more consistently positive autocorrelation (up to two years lag or so).

a. Load and see a description of the data.

   R: `library(wooldridge)` and `?cement`

   Stata: `bcuse cement , clear`

b. Tell your software that you have monthly time series data.

   R: use

```
tsdat <- ts(data=cement$ipcem, frequency=12,
            start=c(cement$year[1],cement$month[1]))
```

[4] https://fred.stlouisfed.org/series/IPG3273N

to create a time series variable named `tsdat` that's a time series (`ts`) with the industrial cement production index data (`ipcem`).

Stata:

```
generate yrmo = ym(year, month)
format yrmo %tm
tsset yrmo
```

c. R only: compute, store, and plot a multiplicative decomposition, to see how important seasonality is for industrial cement production:

```
tsdec <- decompose(tsdat, type='mult')
plot(tsdec)
window(tsdec$seasonal, start=c(1964,1), end=c(1964,12))
```

The last line above prints the numerical values for the seasonality plot (which are the same for each year; e.g., 1964 could be replaced by 1971).

d. Stata only: estimate the trend and plot it against the raw data:

```
tssmooth ma fipcem1=ipcem , weights(1 2 2 2 2 2 <2> 2 2 2 2 2 1)
tsline ipcem fipcem1 , name(fipcem1) ylabel(#3)
```

e. Stata only: compute multiplicative seasonal effects with

```
bysort month : egen seasmult = mean(ipcem/fipcem1)
```

(but don't worry about normalizing these to average to 1 like is sometimes done)

f. Stata only: compute the multiplicative remainder as the observed value divided by the trend value, divided yet again by the seasonal effect:

```
generate rem1mult = ipcem/fipcem1/seasmult
```

g. Stata only: plot the seasonal and remainder series, and then all series together (similar to the R plot):

```
tsline seasmult , name(seasmult) ylabel(#3)
tsline rem1mult , name(rem1mult) ylabel(#3)
graph combine fipcem1 seasmult rem1mult , cols(1) name(decompmult)
```

h. Plot the autocorrelation function (ACF) of the raw data up to 48 months lag.

R: `acf(tsdat, lag.max=48, ci=0, na.action=na.omit)`

Stata: `ac ipcem , level(95) lags(48)`

i. Plot the ACF of the seasonally-adjusted data.

R: `acf(tsdat/tsdec$seasonal, lag.max=48, ci=0, na.action=na.omit)`

Stata:

```
generate saipcem = ipcem / seasmult
ac saipcem , level(95) lags(48)
```

j. Optional: repeat the decomposition plot and ACF plots for a different variable in the dataset.

Chapter 14

First-Order Autoregression

⟹ Kaplan video: Chapter Introduction

Decent forecasts are often achieved by simply regressing Y_t on Y_{t-1} (perhaps after detrending and/or seasonal adjustment). Chapter 14 explores this model, which is also useful for description (if not causal inference). Some extensions are discussed, with additional extensions in Chapter 15.

Unit learning objectives for this chapter

14.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

14.2. Describe the first-order autoregressive model and its features, including interpretation for description and prediction [TLOs 2 and 3]

14.3. Interpret and evaluate forecasts, including multi-step and interval forecasts [TLOs 2 and 3]

14.4. In R (or Stata): estimate the parameters of a first-order autoregression [TLO 7]

14.5. In R (or Stata): generate interval and multi-step forecasts [TLO 7]

14.1 Model

The **first-order autoregressive model**, or **AR(1) model**, is essentially a simple linear regression in which the regressor is the first lag of the outcome variable:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t, \tag{14.1}$$

where ϕ_0 and ϕ_1 are constant coefficients, with ϕ_1 called the **autoregressive parameter** (or **autoregressive coefficient**), and the unobserved ϵ_t is something called **white noise**.

A special case called **independent white noise** is if the ϵ_t are iid, with mean zero and finite variance, and independent of all past Y_s values for $s < t$:

$$\epsilon_t \sim \text{iid}, \quad \mathrm{E}(\epsilon_t) = 0, \quad \sigma_\epsilon^2 \equiv \mathrm{Var}(\epsilon_t) < \infty, \quad \epsilon_t \perp\!\!\!\perp Y_{t-1}, Y_{t-2}, \ldots, \text{ for all } t. \tag{14.2}$$

Diebold (2018a, §13.6) and Diebold (2018b, §6.2) have many more details on white noise that are beyond our scope. Vocabulary terms for the unobserved ϵ_t include **error term**, **shock**, and **innovation**.

Given (14.1) and (14.2), stationarity (either type) of Y_t depends on the parameter values. Specifically,

$$Y_t \text{ is stationary} \iff |\phi_1| < 1. \tag{14.3}$$

If instead $|\phi_1|$, then Y_t has a unit root (Section 13.6.1). For example, with $\phi_0 = 0$ and $\phi_1 = 1$, (14.1) becomes the random walk in (13.7). Nonstationary "explosive processes" with $|\phi_1| > 1$ are sometimes considered to model stock market bubbles (or other bubbles) but are beyond our scope.

Assuming stationarity (either type), the mean of Y_t can be solved for in terms of parameters ϕ_0 and ϕ_1. Let $\mu \equiv \mathrm{E}(Y_t)$, which is the same for all t if Y_t is stationary. Using (14.1),

$$\mu = \mathrm{E}(Y_t) = \overbrace{\mathrm{E}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t)}^{\text{use linearity of } \mathrm{E}(\cdot)} = \phi_0 + \phi_1 \overbrace{\mathrm{E}(Y_{t-1})}^{=\mu} + \overbrace{\mathrm{E}(\epsilon_t)}^{=0} = \phi_0 + \phi_1 \mu. \tag{14.4}$$

Solving $\mu = \phi_0 + \phi_1 \mu$ for ϕ_0 and then μ,

$$\phi_0 = \mu(1 - \phi_1), \quad \mu = \frac{\phi_0}{1 - \phi_1}. \tag{14.5}$$

The AR(1) model in (14.1) can be written equivalently in terms of demeaned values. Generally, a **demeaned** random variable has had its mean subtracted (like a "deboned" fish has had its bones removed), so it has mean zero, like the population mean model's error term in Section 6.2.1. Here, $Y_t - \mu$ is demeaned because $\mathrm{E}(Y_t) = \mu$, so

$$\mathrm{E}(Y_t - \mu) = \mathrm{E}(Y_t) - \mu = 0.$$

Similarly, because $\mu = \mathrm{E}(Y_{t-1})$, then $\mathrm{E}(Y_{t-1} - \mu) = \mathrm{E}(Y_{t-1}) - \mu = 0$, and similarly $\mathrm{E}(Y_{t-j} - \mu) = 0$ for all j because all $\mathrm{E}(Y_{t-j}) = \mu$ due to stationarity.

The demeaned AR(1) model is

$$Y_t - \mu = \phi_1 (Y_{t-1} - \mu) + \epsilon_t. \tag{14.6}$$

This is equivalent to (14.1). After adding μ to both sides of (14.6),

$$Y_t = \mu + \phi_1(Y_{t-1} - \mu) + \epsilon_t = \mu + \phi_1 Y_{t-1} - \phi_1 \mu + \epsilon_t = \overbrace{\mu(1 - \phi_1)}^{=\phi_0 \text{ by } (14.5)} + \phi_1 Y_{t-1} + \epsilon_t. \tag{14.7}$$

In Sum: AR(1), Two Equivalent Models

The AR(1) model (14.1) is $Y_t = \phi_0 + \phi_1 Y_{t-1} + \epsilon_t$, like a regression of Y_t on Y_{t-1}
Given stationarity ($|\phi_1| < 1$),
(14.6): demeaned AR(1) model, $Y_t - \mu = \phi_1(Y_{t-1} - \mu) + \epsilon_t$, with $\mu \equiv \text{E}(Y_t)$
(14.5): translate between models with $\mu = \dfrac{\phi_0}{1-\phi_1}$ or $\phi_0 = \mu(1-\phi_1)$

14.2 Description

Certain properties of Y_t are implied by (14.1) and (14.2). Here, we look at the mean, variance, autocovariances, and autocorrelations of Y_t in terms of the model parameters. You are not expected to understand the derivations, but they are provided in case it helps your understanding of the formulas.

The mean is $\mu = \phi_0/(1-\phi_1)$ given covariance stationarity, as shown in (14.5).

The variance is derived by taking the variance of each side of (14.1). Assuming covariance stationarity, let $\sigma_Y^2 \equiv \text{Var}(Y_t) = \text{Var}(Y_{t-1})$. Using variance identities and $\text{Cov}(Y_{t-1}, \epsilon_t) = 0$ (because $\epsilon_t \perp\!\!\!\perp Y_{t-1}$),

$$\begin{aligned}
\overbrace{\text{Var}(Y_t)}^{=\sigma_Y^2} &= \overbrace{\text{Var}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t)}^{\text{can remove constant } \phi_0} = \overbrace{\text{Var}(\phi_1 Y_{t-1} + \epsilon_t)}^{\text{use Var}(V+W)=\text{Var}(V)+\text{Var}(W)+2\,\text{Cov}(V,W)} \\
&= \overbrace{\text{Var}(\phi_1 Y_{t-1})}^{\text{use Var}(aW)=a^2\,\text{Var}(W)} + \overbrace{\text{Var}(\epsilon_t)}^{\sigma_\epsilon^2 \text{ in } (14.2)} + 2\,\overbrace{\text{Cov}(\phi_1 Y_{t-1}, \epsilon_t)}^{\text{use linearity}} \\
&= \phi_1^2 \overbrace{\text{Var}(Y_{t-1})}^{\sigma_Y^2} + \sigma_\epsilon^2 + 2\phi_1 \overbrace{\text{Cov}(Y_{t-1}, \epsilon_t)}^{=0 \text{ since } \epsilon_t \perp\!\!\!\perp Y_{t-1}} \\
&= \phi_1^2 \sigma_Y^2 + \sigma_\epsilon^2.
\end{aligned}$$

Rearranging to solve for σ_Y^2,

$$\sigma_Y^2 = \phi_1^2 \sigma_Y^2 + \sigma_\epsilon^2 \implies \sigma_Y^2(1 - \phi_1^2) = \sigma_\epsilon^2 \implies \sigma_Y^2 = \frac{\sigma_\epsilon^2}{1-\phi_1^2}. \tag{14.8}$$

The autocovariances can also be calculated given covariance stationarity. Substituting for Y_t using (14.1), and using the same properties from above,

$$\begin{aligned}
\gamma_1 &\equiv \text{Cov}(Y_t, Y_{t-1}) = \text{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-1}) \\
&= \overbrace{\text{Cov}(\phi_0, Y_{t-1})}^{=0 \text{ since } \phi_0 = \text{const}} + \text{Cov}(\phi_1 Y_{t-1}, Y_{t-1}) + \overbrace{\text{Cov}(\epsilon_t, Y_{t-1})}^{=0 \text{ by } \epsilon_t \perp\!\!\!\perp Y_{t-1}} = \phi_1 \overbrace{\text{Cov}(Y_{t-1}, Y_{t-1})}^{=\text{Var}(Y_{t-1})} \\
&= \phi_1 \sigma_Y^2.
\end{aligned} \tag{14.9}$$

Using (14.9) recursively,

$$\begin{aligned}
\gamma_2 &\equiv \mathrm{Cov}(Y_t, Y_{t-2}) = \mathrm{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-2}) \\
&= \overbrace{\mathrm{Cov}(\phi_0, Y_{t-2})}^{=0 \text{ since } \phi_0 = \text{const}} + \phi_1 \mathrm{Cov}(Y_{t-1}, Y_{t-2}) + \overbrace{\mathrm{Cov}(\epsilon_t, Y_{t-2})}^{=0 \text{ by } \epsilon_t \perp\!\!\!\perp Y_{t-2}} \\
&= \phi_1 \gamma_1 = \phi_1 \overbrace{\phi_1 \sigma_Y^2}^{(14.9)} \\
&= \phi_1^2 \sigma_Y^2. \qquad (14.10)
\end{aligned}$$

More generally, by induction, if $\gamma_{j-1} = \phi^{j-1}\sigma_Y^2$, then

$$\begin{aligned}
\gamma_j &\equiv \mathrm{Cov}(Y_t, Y_{t-j}) = \mathrm{Cov}(\phi_0 + \phi_1 Y_{t-1} + \epsilon_t, Y_{t-j}) \\
&= \overbrace{\mathrm{Cov}(\phi_0, Y_{t-j})}^{=0} + \phi_1 \overbrace{\mathrm{Cov}(Y_{t-1}, Y_{t-j})}^{=\gamma_{j-1}} + \overbrace{\mathrm{Cov}(\epsilon_t, Y_{t-j})}^{=0} \\
&= \phi_1 \phi_1^{j-1} \sigma_Y^2 \\
&= \phi_1^j \sigma_Y^2, \qquad (14.11)
\end{aligned}$$

which holds for all $j \geq 0$.

The autocorrelations combine (14.11) with (13.6):

$$\rho_j \equiv \mathrm{Corr}(Y_t, Y_{t-j}) = \gamma_j / \sigma_Y^2 = (\phi_1^j \sigma_Y^2)/\sigma_Y^2 = \phi_1^j. \qquad (14.12)$$

With $j = 1$, the first autocorrelation is $\rho_1 = \phi_1$, the autoregressive coefficient in (14.1).

In Sum: AR(1) for Description

Given (14.1) and (14.2) with $|\phi_1| < 1$ ($\implies$ stationary),
mean: $\mu = \mathrm{E}(Y_t) = \phi_0/(1 - \phi_1)$
variance: $\sigma_Y^2 = \mathrm{Var}(Y_t) = \sigma_\epsilon^2/(1 - \phi_1^2)$
jth autocovariance: $\gamma_j = \phi_1^j \sigma_Y^2$
jth autocorrelation: $\rho_j = \phi_1^j$

14.3 Prediction (Forecasting)

For time series, "prediction" usually means **forecasting** future values of Y_t given the current and past values. As in Section 2.5, given a loss function, the optimal forecast (prediction) minimizes mean loss. This optimal forecast is defined in the population (without data) and can be estimated with data. In practice, given the observed Y_t for $t = 1, \ldots, T$, the goal is to forecast Y_{T+1}, or to forecast Y_{T+h} for another $h \geq 1$.

As in Part II, the focus here is on the CMF, the best forecast given quadratic loss. Given (14.2), (14.1) is a CMF model, so $\phi_0+\phi_1 Y_{t-1}$ is the "best" forecast of Y_t given Y_{t-1}. Thus, given sample $Y_1,\ldots,Y_T$ and corresponding OLS estimates $\hat{\phi}_0$ and $\hat{\phi}_1$, a reasonable forecast of Y_{T+1} is

$$\hat{Y}_{T+1}=\hat{\phi}_0+\hat{\phi}_1 Y_T. \tag{14.13}$$

Even if the AR(1) model is wrong (but Y_t is covariance stationary), OLS still estimates the best linear predictor (Section 7.5) of Y_t given Y_{t-1}.

However, "best" does not mean "good" (Section 7.4.2); forecast accuracy may be improved by using additional lags, other variables, and/or nonlinearity (Chapter 15).

Discussion Question 14.1 (forecast and reality)**.** Given sample $Y_1,\ldots,Y_T$, you construct forecast $\hat{Y}_{T+1}=\hat{\phi}_0+\hat{\phi}_1 Y_T$. Then you wait one period and observe the actual Y_{T+1}.

a) Will you be surprised if $Y_{T+1}>\hat{Y}_{T+1}$? Or if $Y_{T+1}<\hat{Y}_{T+1}$? Why/not?
b) How often do you expect to see $Y_{T+1}=\hat{Y}_{T+1}$? Why?
c) Is it usually true that $Y_{T+1}=\phi_0+\phi_1 Y_T$? Why/not? Hint: for any random variable W, how often does $W=\mathrm{E}(W)$, i.e., what's $\mathrm{P}(W=\mathrm{E}(W))$?

14.4 Estimation

In the AR(1), skipping technical details, the OLS estimators $\hat{\phi}_0$ and $\hat{\phi}_1$ are consistent in many cases. If $|\phi_1|<1$, then OLS is consistent. Technical details for a more general version of this result may be found in Case 4 on pages 215–217 of Hamilton (1994, §8.2). In fact, the slope estimator is consistent even if $\phi_1=1$ (Hamilton, 1994, §17.4).

There are other consistent estimators, too, and some research has tried to compare the small-sample properties of these, but such comparison is beyond our scope.

Using the estimated coefficients, a **point forecast** (our single, best guess) $\hat{Y}_{T+1}$ is computed as in (14.13). For the demeaned model in (14.6),

$$\hat{Y}_{T+1}=\hat{\mu}+\hat{\phi}_1(Y_T-\hat{\mu}). \tag{14.14}$$

14.4.1 Code

The following code shows an example. The data `Y` are simulated from an AR(1) model with $\phi_0=0$ (so $\mu=0$) and $\phi_1=0.25$, using `arima.sim()`. The argument `n.ahead` tells `predict()` how many time periods past the end of the sample to make predictions for. In R, `ar()` by default estimates the demeaned model. In the code, $\hat{\mu}$ is `ret$x.mean`, $\hat{\phi}_1$ is `ret$ar`, and $\hat{\phi}_0$ is `ret$x.mean*(1-ret$ar)`. The predicted value in `pr$pred[1]` is shown to be equivalent to (14.13) and (14.14). (Alternatively, with argument `method='ols'`, you can estimate ϕ_0 and ϕ_1 directly, by OLS.) The 95% CI for ϕ_1 is computed using a formula based on asymptotic normality.

```
set.seed(112358)
RHO <- 0.25;  n <- 100
Y <- arima.sim(n=n, model=list(ar=RHO), sd=1)
ret <- ar(x=Y, aic=FALSE, order.max=1)
cat(sprintf("PhiHat0=%5.3f, PhiHat1=%5.3f\n",
            ret$x.mean*(1-ret$ar), ret$ar))

## PhiHat0=0.024, PhiHat1=0.143

# 95% CI for slope PhiHat1
c(CI.low =ret$ar-1.96*sqrt(ret$asy.var.coef),
  CI.high=ret$ar+1.96*sqrt(ret$asy.var.coef) )

##  CI.low CI.high
## -0.0529  0.3390

pr <- predict(ret, n.ahead=1)
# Point forecast
round(pr$pred[1], digits=3)

## [1] 0.196

# Sanity check: same as from formulas:
c(ret$x.mean + ret$ar*(Y[n]-ret$x.mean),
  ret$x.mean*(1-ret$ar) + Y[n]*ret$ar )

## [1] 0.196 0.196
```

14.5 Multi-Step Forecast

Instead of forecasting Y_{t+1} given Y_t, you may need to forecast Y_{t+h} given Y_t for a particular $h > 1$. This is called the ***h*-step-ahead forecast**.

Example 14.1. Imagine you must make a decision that affects your business or government policy for the next year. You have monthly data. Specifically, you want to predict Y_{t+12} given Y_t, i.e., predict the value 12 months in the future. In fact, you want to predict Y_{t+h} for all $h = 1, \ldots, 12$, i.e., predict each of the next 12 months.

There are two common approaches to h-step-ahead forecasting. One approach uses the AR(1) model to derive the best forecast of Y_{T+h} given Y_T, in terms of the model's parameters. A second approach is simply to regress Y_{t+h} on Y_t (and an intercept). Given covariance stationarity, such a regression estimates the best linear predictor of Y_{t+h} given

Y_t. This regression estimates the parameters in

$$Y_{t+h} = \phi_0 + \phi_1 Y_t + \epsilon_{t+h}. \tag{14.15}$$

The forecast of Y_{T+h} is then

$$\hat{Y}_{T+h} = \hat{\phi}_0 + \hat{\phi}_1 Y_T. \tag{14.16}$$

The forecast in (14.13) showed the special case with $h = 1$.

Example 14.2 (Kaplan video)**.** If Y_t is quarterly GDP growth, and we want to predict GDP growth four quarters (i.e., one year) in the future, then $h = 4$. We regress Y_{t+4} on an intercept and Y_t in our quarterly data, predicting $\hat{Y}_{T+4} = \hat{\phi}_0 + \hat{\phi}_1 Y_T$.

There are functions in R that do multi-step forecasts automatically, like the `forecast` function in the `forecast` package (Hyndman et al., 2020; Hyndman and Khandakar, 2008), which also does multi-step interval forecasts; see Section 14.8.

14.6 Interval Forecasts

Like a confidence interval, an **interval forecast** (or **forecast interval**) incorporates uncertainty and tries to contain the true value with high probability (like 95%). Unlike a confidence interval, the true value is a random variable (like Y_{T+1} or Y_{T+h}) rather than a non-random parameter (like β).

Example 14.3 (Kaplan video)**.** Imagine your job is to create 95% interval forecasts, and you make one every day for 1000 days. That is, on each day t, you make an interval forecast for the next day's value Y_{t+1}; then the next day you check whether or not the true value was inside your interval. If you're doing your job well, then you should find that approximately 950 days out of 1000 (95% of the days) your interval contained the true value, and the other 50 days it didn't.

There are two sources of uncertainty in forecasting. The first source of uncertainty is the same as in a confidence interval: parameter uncertainty. That is, we only have estimated parameter values $\hat{\phi}_0$ and $\hat{\phi}_1$; we do not know the true population parameters ϕ_0 and ϕ_1. The second (and usually larger) source of uncertainty is the error term ϵ_{T+1}. Even if we knew ϕ_0 and ϕ_1, we'd still have uncertainty about $Y_{T+1} = \phi_0 + \phi_1 Y_T + \epsilon_{T+1}$.

There are different ways to construct forecast intervals, but details are beyond our scope. The main differences are in how to capture uncertainty about ϵ_{T+1}. For example, we could assume ϵ_{T+1} has a normal distribution (e.g., Diebold, 2018b, §7.3.3), but the corresponding interval may be bad if the true distribution is far from normal.

14.6.1 Code

The following code shows basic interval forecasts using the `forecast` package (which also shows the point forecasts). The argument `h=12` specifies forecasting values for the next

12 time periods, so the results include multi-step interval forecasts. Argument `level=c(80,95)` specifies both 80% and 95% prediction intervals. Although the code is easy to run, an AR(1) is not always appropriate, so critical thought is required; see DQ 14.2.

```
library(forecast)
ret <- ar(AirPassengers, aic=FALSE, order.max=1)
forecast(ret, h=12)
```

```
##    Period Point.Forecast Lo.80 Hi.80 Lo.95 Hi.95
##  Jan-1961            424   375   473   349   499
##  Feb-1961            417   349   484   313   520
##  Mar-1961            410   329   490   286   533
##  Apr-1961            403   312   494   264   542
##  May-1961            396   297   496   245   548
##  Jun-1961            390   284   496   228   553
##  Jul-1961            385   273   497   214   556
##  Aug-1961            379   262   496   200   558
##  Sep-1961            374   253   495   189   560
##  Oct-1961            369   244   494   178   560
##  Nov-1961            365   236   493   168   561
##  Dec-1961            360   229   491   160   561
```

Discussion Question 14.2 (forecast sanity check)**.** Do the point forecasts shown above pass a sanity check? That is, they show steadily decreasing values from January to December 1961; does this seem reasonable given Figure 13.1? Why/not?

14.7 Parameter Stability

Parameter stability pertains to external validity, as in Section 12.2: is ϕ_1 truly a constant, or has it changed over time, and might it change in the future? This is also related to structural breaks (Section 13.6.4). With enough data, we could form multiple historical datasets and see if the estimates $\hat{\phi}_1$ change much over time. But either way, this does not tell us what will happen in the future. Historical data cannot predict a future **black swan**, something new not seen in the past. As usual, purely statistical analysis may fall short; a combination of your statistical and economic expertise (and critical thinking) yields better results.

Parameter instability relates to the Lucas critique (Lucas, 1976): if there is a new macroeconomic policy with general equilibrium effects (Section 4.3.3), then the time series model's parameters may change, so it is not accurate for description or prediction (forecasting).

To address this, there are models allowing time-varying coefficients, and methods to estimate when a parameter changes, but all are beyond our scope.

Discussion Question 14.3 (recession-affected coefficient)**.** Name a variable you think might have different ϕ_1 during an extended recession (than not during a recession), not including the switch from non-recession to recession. For example, if there is a recession from $t = 11$ to $t = 20$, then consider the ϕ_1 for Y_t for $t = 1, \ldots, 10$ compared to the ϕ_1 for $t = 11, \ldots, 20$. As usual, most importantly, explain why you think so. Hint: this is not simply asking which variables are higher or lower in a recession, because that's not what ϕ_1 describes; e.g., the time series $Z_t = Y_t + 10$ would have the exact same ϕ_1 as Y_t, just a different ϕ_0 or μ.

14.8 More R Examples

14.8.1 AR(1) Multi-Step Forecast Intervals

The following code simulates data from an AR(1) model, and then computes (and outputs and plots) various estimates and forecasts. Note that $T = 100$ (`n <- 100`), $\phi_1 = 0.8$ (`RHO`), $\mu = \mathrm{E}(Y_t) = 5$, and $\sigma_\epsilon = 1$ (from the `sd=1` option). The estimated $\hat{\phi}_1$ is not particularly good, although the true value is within two standard errors (there is just a lot of uncertainty).

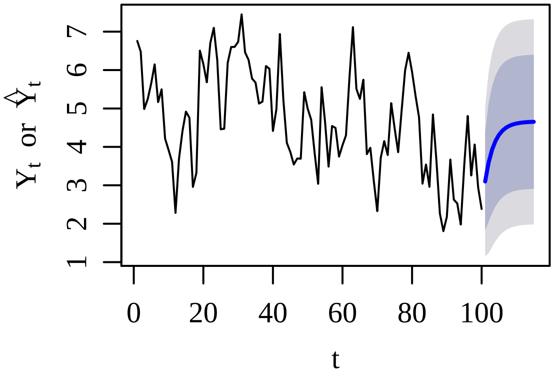


Figure 14.1: Point and interval forecasts.

```
set.seed(112358)
RHO <- 0.80;  n <- 100
Y <- 5 + arima.sim(n=n, model=list(ar=RHO), sd=1)
ret <- ar(x=Y, aic=FALSE, order.max=1)
cat( sprintf("PhiHat1=%5.3f\n", ret$ar) )

## PhiHat1=0.685

# 95% CI for slope PhiHat1
```

```
c(CI.low =ret$ar-1.96*sqrt(ret$asy.var.coef),
  CI.high=ret$ar+1.96*sqrt(ret$asy.var.coef) )

##  CI.low CI.high
##   0.541   0.830
```

```
(fc <- forecast(ret, h=15, level=c(80,95)))
plot(fc)
```

```
##  Period Point.Forecast Lo.80 Hi.80 Lo.95 Hi.95
##     101           3.10  1.83  4.37  1.15  5.04
##     102           3.59  2.05  5.13  1.23  5.95
##     103           3.93  2.27  5.58  1.40  6.46
##     104           4.16  2.45  5.86  1.55  6.76
##     105           4.32  2.59  6.04  1.68  6.96
##     106           4.42  2.69  6.16  1.77  7.08
##     107           4.50  2.76  6.24  1.83  7.16
##     108           4.55  2.81  6.29  1.88  7.22
##     109           4.58  2.84  6.33  1.92  7.25
##     110           4.61  2.86  6.35  1.94  7.28
##     111           4.63  2.88  6.37  1.95  7.30
##     112           4.64  2.89  6.38  1.97  7.31
##     113           4.64  2.90  6.39  1.97  7.32
##     114           4.65  2.90  6.40  1.98  7.32
##     115           4.65  2.91  6.40  1.98  7.32
```

Figure 14.1 was generated by the above code and shows some patterns. The graph essentially plots the table of results (point and interval forecasts) after plotting the original time series. First, in the data itself, we can see some persistence (high values tend to be followed by high values, and low by low), but the values never get too far from the mean $\mathrm{E}(Y_t) = 5$. Second, the point forecasts $\hat{Y}_{t+h}$ get closer and closer to the sample average $\bar{Y} = \frac{1}{T}\sum_{t=1}^{T} Y_t$ as h increases. This is because we chose an AR(1) forecasting model; even if the data were not generated by an AR(1), the forecasts would show the same pattern (so as in Section 8.1.5, beware model-driven forecasts). Third, the forecast intervals get wider and wider as h increases. This makes intuitive sense: the farther in the future, the less certainty we have.

14.8.2 General R Forecast Allowing Seasonality and Trend

⟹ Kaplan video: Forecasting in R

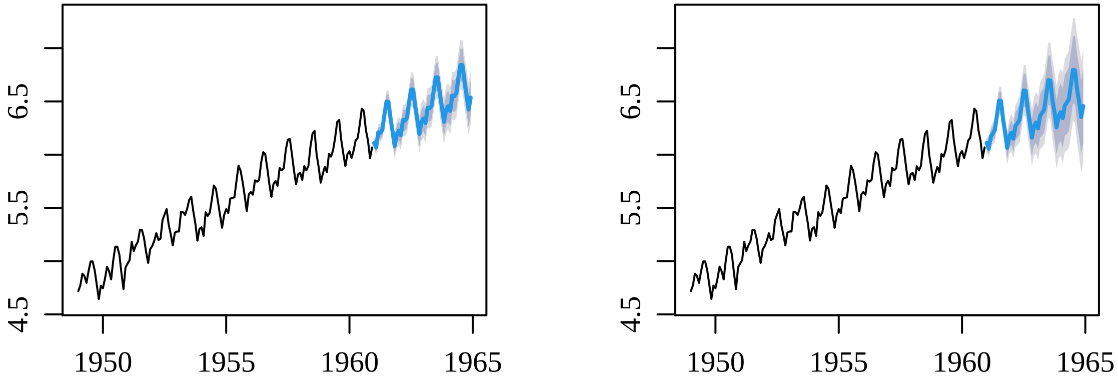


Figure 14.2: Air travel forecasts from stlf (left) and auto.arima (right).

Figure 14.2 uses the stlf() and auto.arima() functions from the forecast package to compute point forecasts and interval forecasts of log passengers from the monthly air travel data. They do much better than the earlier forecast in Section 14.6.1 that ignored seasonality and trend. This general application of stlf() or auto.arima() can sometimes be improved by more carefully considering the type of trend, the properties of the remainder, the type of seasonality, etc., but clearly for series where the trend and/or seasonality is important, it is much better to use these functions that incorporate trend and seasonality than a model that does not allow for trend and seasonality, like the basic AR(1). But, AR models are still very useful: they (or more general ARIMA models) are used by stlf() and auto.arima() to fit the detrended, seasonally-adjusted data.

Either way, it is always good to "sanity check" your forecasts visually. In this case, the point forecasts in Figure 14.2 look reasonable, unlike the earlier basic AR(1). It is also reasonable that the interval forecasts get longer (taller) farther into the future, appropriately reflecting greater uncertainty. However, the stlf() interval forecasts seem too narrow; even multiple years in the future, the interval is relatively short. Reading the ?stlf help file suggests one reason why: it says, "Note that the prediction intervals ignore the uncertainty associated with the seasonal component." That is, it assumes the estimated seasonality is actually the true seasonality, with no uncertainty. Even the auto.arima() intervals may be "too short" because (as usual) they do not account for uncertainty about the true model itself changing over time (i.e., structural breaks), only uncertainty about the parameter values.

Figure 14.2 was generated by the following code.

```r
library(forecast)
par(family='serif', mar=c(1.8,1.8,0.3,0.6), mgp=c(2.1,0.8,0))
ret1 <- stlf(y=log(AirPassengers), h=48)
plot(ret1)
ret2 <- auto.arima(y=log(AirPassengers))
```

```
plot(forecast(ret2, h=48))
```

Optional Resources

Optional resources for this chapter

- AR(1) (Lambert video)
- AR(1) series with different autocorrelations (Lambert video)
- Chapter 12 ("Serial Correlation") in Diebold (2018a)
- Parameter stability: Hanck et al. (2018, §14.8), Diebold (2018a, §12.4–5)
- AR(1) model and properties: Hamilton (1994, §3.4)
- AR(1): Hyndman and Athanasopoulos (2019, §8.3), Hanck et al. (2018, §14.3)
- Asymptotic theory: Hamilton (1994, §§8.2,17.4)
- Forecast/prediction interval: Hyndman and Athanasopoulos (2019, §3.5), Diebold (2018b, §§7.3.3,7.4.3)
- Multi-step forecasting: Diebold (2018b, §6.7.3)
- R package `forecast`: Hyndman and Athanasopoulos (2019, 3.6), Hyndman et al. (2020), Hyndman and Khandakar (2008)

Empirical Exercises

Empirical Exercise EE14.1. You will analyze the New York Stock Exchange (NYSE) value-weighted price index, specifically the weekly close prices every Wednesday. (Unfortunately, the dataset does not note the dates or data source.) You'll consider forecasting price as well as the price change, using an AR(1) model, with both point and interval forecasts. In practice, if you could reliably predict the price change, then you could make a lot of money; so you should be (very) skeptical that you can forecast the price change. (This is related to the "efficient market hypothesis.") Related: if stock prices are a random walk, then the optimal forecast should just be the most recently observed value; you can see if this matches your code's forecasts.

Mathematically, assume the price change $U_t = Y_t - Y_{t-1}$ is indeed unrelated to Y_t and Y_{t-1} (and other past values), and let $\phi_0 = \mathrm{E}(U_t)$ and $V_t = U_t - \mathrm{E}(U_t)$, so $\mathrm{E}(V_t) = 0$. Then $Y_t = Y_{t-1} + U_t = \phi_0 + Y_{t-1} + V_t$ is an AR(1) with $\phi_1 = 1$, in which case $Y_T + \phi_0$ is the best forecast of Y_{T+1}. You will check if $\hat{\phi}_1 \approx 1$ and estimate the value of ϕ_0, among other computations.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

```
library(wooldridge); library(forecast)
?nyse
```

b. Stata only: load the data with `bcuse nyse , nodesc clear` (assuming `bcuse` is already installed)

c. Tell your software that you have weekly time series data.

   R: `tsdat <- ts(data=nyse$price, frequency=52.18)`

   Stata: `tsset t , weekly`

d. Define a variable `holdout` for how many time periods at the end of the sample to "hold out" when fitting your model.

   R: `holdout <- 20`

   Stata: `scalar holdout = 20`

e. R only: using `holdout`, define the time period at the end of the "training" data (just before the "testing" data) as `midpt <- length(tsdat)-holdout` and use it to define the training and testing data respectively:

```
tsdattrain <- subset(tsdat, start=1, end=midpt)
tsdattest <- subset(tsdat, start=midpt+1, end=length(tsdat))
```

f. Estimate an AR(1) model to produce "dynamic" forecasts, i.e., what would be forecast if we were living at the end of the training data.

   R: `ret <- ar(x=tsdattrain, aic=FALSE, order.max=1, method='yw')`

   Stata: `arima price if _n<=_N-holdout , arima(1,0,0)`

g. Pretend you travel back in time to the very end of the training data, and produce dynamic forecasts for the next 20 periods (weeks).

   R: `(fc <- forecast(ret, h=holdout, level=c(80,95)))`

   Stata: `predict fmulti , y dyn(t[_N-holdout+1])` where `fmulti` is the name for a newly created variable and `dyn` tells it to make dynamic forecasts

h. Plot the forecasts against the actual historical data.

   R: `plot(fc)` and `lines(window(tsdattest), col=1)`

   Stata:

   `twoway tsline price || tsline fmulti if _n>_N-holdout , lcolor(red)`

i. Optional: repeat your analysis, but with an AR(1) model of the first-differenced price ($\Delta Y_t = Y_t - Y_{t-1}$), which is already in the dataset as the variable `cprice` ("c" for "change").

   R: when you create `tsdat`, use `data=nyse$cprice[-1]` to exclude the first value of `cprice` (which is missing); otherwise the code should be the same; you may also like to draw a line with `abline(h=0)` at the very end for reference.

   Stata: just use `cprice` and make sure to name a different new variable in your `predict` command, which you'll reference in your graphing command. Note also that instead of `arima cprice`, you could use OLS estimation with `regress cprice cprice_1`, or equivalently `regress D.price L.D.price` where `D.price` means "take the first difference of the variable `price`" and `L.D.price` means "lag of first difference of `price`"

Chapter 15

Higher-Order Autoregression and Autoregressive Distributed Lag Regression

⟹ Kaplan video: Chapter Introduction

Sometimes, accuracy improves by forecasting Y_{t+1} using not only Y_t but also Y_{t-1}. And why stop at Y_{t-1}? Maybe Y_{t-2} contains additional information not found in Y_t and Y_{t-1}; or maybe Y_{t-3} does, or even longer lags of Y_t. Additionally, other variables and possibly their lags may further improve forecasting accuracy. However, as in Section 8.3, too many regressors can worsen performance, so model selection is crucial to good performance.

Note: this chapter is intentionally short, to allow students more time to start preparing for the final exam (in this class or other classes).

Unit learning objectives for this chapter

15.1. Define new vocabulary words (in **bold**), both mathematically and intuitively [TLO 1]

15.2. Explain the problem of choosing the best model both mathematically and intuitively, along with possible solutions [TLO 2]

15.3. Implement and compare different ways to select the best forecasting model [TLOs 2 and 6]

15.4. In R (or Stata): estimate more general time series regression models for the purpose of forecasting future values [TLO 7]

15.1 The AR(p) Model

The **AR(p) model** generalizes the AR(1) model in (14.1):

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \epsilon_t = \phi_0 + \sum_{j=1}^{p} \phi_j Y_{t-j} + \epsilon_t. \tag{15.1}$$

Again ϵ_t is white noise, with properties as in (14.2). Coefficient ϕ_j is called the jth **partial autocorrelation**, for $j = 1, \ldots, p$. (This can be confusing because $\phi_j \neq \mathrm{Corr}(Y_t, Y_{t-j})$, the jth autocorrelation.)

Theoretical details and properties are mostly omitted here, but there are concepts similar to the AR(1). For example, there is the concept of a unit root, which generates nonstationary Y_t, but its mathematical characterization is more complicated than just $\phi_1 = 1$. The autocovariances and autocorrelations can be derived from the coefficients and properties of ϵ_t, but the derivations and formulas are again more complicated.

Instead, the next sections focus on good forecasts.

15.2 Model Selection: How Many Lags?

⟹ Kaplan video: Model Selection for Forecasting

In practice, which p should we use? This is a question of model selection (Section 8.3). Choosing p is equivalent to setting $\hat{\phi}_j = 0$ for $j > p$, instead of estimating those ϕ_j from data.

15.2.1 Difficulties and Intuition

Recall the intuition from Section 8.3. If p is too small, then the model is not flexible enough; implicitly, this sets $\hat{\phi}_j = 0$ for some important $\phi_j \neq 0$. Even if the ϕ_j are estimated perfectly for $j = 0, 1, \ldots, p$, the estimated model may not forecast very well because $\hat{\phi}_j = 0 \neq \phi_j$ for some $j > p$. However, if p is too big, then the model can be too flexible, overfitting the data. This also causes poor forecasts. We want the "just right" p that balances these two sources of error.

Only looking at in-sample fit leads to overfitting (Section 8.3). For example, minimizing the sum of squared residuals (SSR), or equivalently maximizing the R^2, always picks the largest possible p, regardless of the dataset and which model is actually best. The "adjusted R^2" is better, but still not designed for picking the best forecasting model. Similarly, hypothesis testing is not designed to pick the best forecasting model.

With time series, large p additionally limits the amount of usable data. For example, if we observe Y_t for $t = 1, \ldots, T$, and we regress Y_t on lags up to Y_{t-50} ($p = 50$), then we can only use t for which both Y_t and Y_{t-50} are observed. If $t > T$, then Y_t isn't observed; if $t \leq p$, then Y_{t-p} isn't observed. If $T = 51$, then there is only one usable data point:

regressing Y_{51} on $Y_{50}, Y_{49}, \ldots, Y_1$. Because it's impossible to estimate 51 parameters from 1 data point, p must be (much) smaller. Even with $p = 25$, there are $p+1 = 26$ parameters and $T - p = 26$ usable data points; estimates could be computed but certainly suffer from overfitting. With T total observations, you can only estimate an AR(p) with $p < T/2$, and p must be even smaller for reliable estimation.

The most common model selection methods for AR(p) models use information criteria. Basically, an **information criterion** tries to quantify how bad a model is for prediction, so lower values are better (less bad). The two most common are the **Akaike information criterion** (AIC), proposed by Akaike (1974), and the **Bayesian information criterion** (BIC) (or sometimes SIC, SBC, or SBIC) of Schwarz (1978). There is also a "corrected" AICc; e.g., see Hyndman and Athanasopoulos (2019, §8.6).

As seen below, both AIC and BIC try to avoid overfitting by adding a penalty to the in-sample fit. The penalty is larger when the model is larger (more flexible). AIC and BIC can also be used for model selection with other types of models beyond autoregression.

Instead of picking a single "best" forecasting model, averaging multiple forecasts ("forecast averaging," or more generally "model averaging") often performs even better but is beyond our scope.

In Sum: Model Selection for Forecasting

After you think critically about which variables and lags might help forecast future values, AIC (and AICc) and BIC can help you pick which model produces the best forecasts.

15.2.2 AIC and BIC Formulas

There are many different but equivalent formulas for AIC and BIC. This is because the selected model is the one whose value is lower than any other model's value, so only the relative values matter, not the numeric values themselves. Thus, we could add 5 to all values, or multiply by T, or take the log, etc., because this would not change which value of p (number of lags) minimizes the AIC or BIC.

The AIC can be written in terms of the sum of squared residuals (SSR) and a penalty based on p. Specifically,

$$\mathrm{AIC}(p) = \overbrace{T \ln(\mathrm{SSR})}^{\text{in-sample fit}} + \overbrace{2(p+1)}^{\text{penalty}} . \tag{15.2}$$

Intuitively, we'd like our models to fit the data well (small SSR), but given the same fit we prefer less flexible models (small penalty). The penalty prevents overfitting, where a model fits the data sample "too well" because it fits all the noise, which in turn makes its out-of-sample forecasts poor.

The BIC also involves the SSR and a penalty. Specifically,

$$\mathrm{BIC}(p) = \overbrace{T\ln(\mathrm{SSR})}^{\text{in-sample fit}} + \overbrace{(p+1)\ln(T)}^{\text{penalty}}. \tag{15.3}$$

When comparing models with different lag lengths, due to the different number of usable data points, some care is required to ensure a fair comparison. For now, you can try to use built-in functions and hope that they were implemented carefully; e.g., in the `forecast` package, `auto.arima()` does automatic model selection using the AICc (which you can change to AIC or BIC with the `ic` argument).

15.2.3 Comparison of AIC and BIC

Compared to the AIC, the BIC has a larger penalty for large models because $\ln(T) > 2$ if $T > 7$. (And if $T \leq 7$, you should collect more data.) That is, the BIC is more likely to pick smaller p, i.e., shorter lag lengths (smaller models).

Related to this difference, whether AIC or BIC is best depends on what you think about the true model. BIC is better than AIC if the true model is small but worse if the true model is large (Shao, 1997, p. 235). For example, if the true model is an AR(1), and you're selecting among AR(p) models for $p = 0, 1, \ldots, 24$, then BIC is more likely to pick the true model than AIC. However, if the true model is AR(100) and $T = 50$ (in which case picking the true model is impossible), then AIC is more likely than BIC to pick the best feasible model. Generally, AIC is better if you only consider lag length up to p, but the true lag length is even larger.

Example 15.1 (Kaplan video)**.** Imagine choosing from either one or two lags. The AR(2) model always fits the data better (lower SSR) than the AR(1) model. To be concrete, imagine $T\ln(\mathrm{SSR}) = 11$ with $p = 1$, and $T\ln(\mathrm{SSR}) = 8$ with $p = 2$. With AIC, the penalty term equals 4 for $p = 1$ and equals 6 for $p = 2$; the AIC penalty depends only on p, not the data or even T. For BIC, the penalty terms for $p = 1$ and $p = 2$ are $2\ln(T)$ and $3\ln(T)$, respectively; e.g., if $T = 50$, then these are approximately 7.8 and 11.7. Thus, plugging these values into (15.2) and (15.3),

$$\mathrm{AIC}(1) = \overbrace{T\ln(\mathrm{SSR})}^{11} + \overbrace{2(p+1)}^{4} = 15, \quad \mathrm{BIC}(1) = \overbrace{T\ln(\mathrm{SSR})}^{11} + \overbrace{(p+1)\ln(T)}^{7.8} = 18.8,$$
$$\mathrm{AIC}(2) = \overbrace{T\ln(\mathrm{SSR})}^{8} + \overbrace{2(p+1)}^{6} = 14, \quad \mathrm{BIC}(2) = \overbrace{T\ln(\mathrm{SSR})}^{8} + \overbrace{(p+1)\ln(T)}^{11.7} = 19.7.$$

Because $\mathrm{AIC}(2) < \mathrm{AIC}(1)$, $p = 2$ is better according to AIC. However, $\mathrm{BIC}(1) < \mathrm{BIC}(2)$, so $p = 1$ is better according to BIC. If we use AIC, we then fit an AR(2) model and use its estimates to forecast Y_{T+1}. If instead we had used BIC for model selection, we'd estimate an AR(1) model and use it to forecast Y_{T+1}.

Discussion Question 15.1 (lag choice for forecasting)**.** Imagine $Y_t = 50 + 0.5Y_{t-1} + 0.00001Y_{t-2} + \epsilon_t$, where the ϵ_t are independent of past values $Y_{t-1}, Y_{t-2}, \ldots$ and are iid and mean-zero. Do you think an estimated AR(0), AR(1), AR(2), or AR(3) would produce the best forecasts? Explain why you think your estimated model would produce better forecasts than each of the other three estimated models. Hint #1: if you need to make assumptions about things like the value of T, please feel free as long as you say so explicitly. Hint #2: thinking about extreme situations is sometimes helpful; e.g., what if $\epsilon_t = 0$ for all t, or what if $T = 8$, etc. Hint #3: yes, this is a very difficult question.

15.2.4 Code

The following code uses the AIC to choose p, then makes a forecast of Y_{T+1} using an AR(p) model. The AIC-chosen p is shown along with the p used to generate the data. Finally, the BIC is computed for the AIC-chosen p and that $p-1$; the BIC is lower for the latter value, so it prefers a smaller model (smaller p) than AIC in this case.

```
set.seed(112358)
MAXP <- 15 #max lag length for AR(p)
ARCOEFFS <- c(0.6, -0.4, 0.4, 0.1)
TRUEP <- length(ARCOEFFS) #p in true AR(p) DGP
# simulate data
Y <- arima.sim(n=60, model=list(ar=ARCOEFFS), sd=1)
# fit AR(p), using AIC to choose best p
ret <- ar(x=Y, aic=TRUE, order.max=MAXP)
# output optimal p
cat(sprintf("true p=%d; AIC-chosen p=%d\n", TRUEP, ret$order))

## true p=4; AIC-chosen p=7

pr <- predict(ret, n.ahead=1) #compute point forecast
c(round(pr$pred, digits=3)) #output

## [1] -0.434

# check BIC for AIC-chosen p and one smaller
# probably BIC prefers smaller (ret2)
ret1 <- arima(Y, order=c(ret$order,0,0))
ret2 <- arima(Y, order=c(ret$order-1,0,0))
c(BIC(ret1),BIC(ret2))

## [1] 186 185
```

15.3 Autoregressive Distributed Lag Regression

The **autoregressive distributed lag** (ADL) model (or "dynamic distributed lag" model) adds other variables and their lags to the AR(p) model. That is, instead of forecasting Y_{t+1} using only Y_t, Y_{t-1}, and other lags of Y, we could also use X_t, X_{t-1}, etc. Because X_{t+1} is not available at time t, it should not be included as an explanatory variable if we are interested in forecasting. Equivalently, if we regress Y_t on Y_{t-1}, Y_{t-2}, and other lags, we could add X_{t-1}, X_{t-2}, etc., but not X_t. If the goal is not forecasting but rather understanding the economic relationship between Y_t and X_t, then this comment does not apply.

The same ideas from before apply to the ADL model. For example, it could be used for multi-step forecasting by replacing Y_{t+1} with Y_{t+h}, or used for interval forecasts, and forecasts may be evaluated and compared as in Section 15.2.

To handle seasonality, decomposition or **seasonal dummies** can be used. The first option is to "seasonally adjust" your data by removing the seasonal component, and then fit the ADL model (and add back the seasonality into the forecast $\hat{Y}_{T+1}$). The second option is to use the raw data but replace the intercept term with dummies for each possible season. For example, with quarterly data, let $D_{1t} = 1$ if time period t is in quarter 1 of some year (and $D_{1t} = 0$ otherwise), and similarly $D_{2t} = 1$ if t is in quarter 2, $D_{3t} = 1$ for quarter 3, and $D_{4t} = 1$ for quarter 4. All four dummies can be included as regressors if the intercept is removed; alternatively, you can keep the intercept and just add D_{2t}, D_{3t}, and D_{4t} as regressors. So an AR(2) model $Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t$ (for example) would change to either

$$Y_t = \beta_1 D_{1t} + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t \tag{15.4}$$

or

$$Y_t = \phi_0 + \beta_2 D_{2t} + \beta_3 D_{3t} + \beta_4 D_{4t} + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \epsilon_t. \tag{15.5}$$

As another example, for monthly data, let D_{2t} be the dummy for February, D_{3t} for March, up to D_{12t} for December. Then, you can either include the intercept along with $D_{2t}, \ldots, D_{12t}$ as regressors, or else remove the intercept and include all $D_{1t}, \ldots, D_{12t}$ as regressors.

The following code uses ADL models to forecast quarterly GDP growth. First, quarterly GDP G_t is transformed to $Y_t = \ln(G_t) - \ln(G_{t-1})$ and stored in variable `GDPgr` ("gr" for "growth"). Second, lags of T-bill rates are generated. Third, various ADL models are fit and their AIC (actually AICc) calculated. Fourth, the best ADL model is used to forecast Y_{T+1}; the output at the end shows the point forecast along with forecast intervals. Note that `auto.arima()` automatically chooses the best lag length for Y_t, but the best T-bill lag is determined "manually," by calling `auto.arima()` once for each possible T-bill lag.

```
library(AER);  library(forecast);  data('USMacroSWQ')
GDPgr <- diff(x=log(USMacroSWQ[,'gdp'])) # GDP growth
Tblags <- cbind(Tblag1=lag(USMacroSWQ[,'tbill'],-1),
                Tblag2=lag(USMacroSWQ[,'tbill'],-2),
                Tblag3=lag(USMacroSWQ[,'tbill'],-3),
                Tblag4=lag(USMacroSWQ[,'tbill'],-4))
Tblags <- subset(Tblags,end=NROW(GDPgr))
fit1 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:1],start=4))
fit2 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:2],start=4))
fit3 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:3],start=4))
fit4 <- auto.arima(y=subset(GDPgr,start=4),
                   xreg=subset(Tblags[,1:4],start=4))
AICcs <- c(fit1[["aicc"]],fit2[["aicc"]], fit3[["aicc"]],fit4[["aicc"]])
best <- which.min(AICcs)
# fit3 has lowest AIC/AICc; fit1 lowest BIC
# Now fit w/ all available data
fit <- auto.arima(y=GDPgr, xreg=Tblags[,1:best])
tnow <- NROW(USMacroSWQ)
xr <- cbind(Tblag1=USMacroSWQ[tnow-0,'tbill'],
            Tblag2=USMacroSWQ[tnow-1,'tbill'],
            Tblag3=USMacroSWQ[tnow-2,'tbill'],
            Tblag4=USMacroSWQ[tnow-3,'tbill'])
xr <- matrix(xr[,1:best], nrow=1)
(fc <- forecast(fit, h=1, xreg=xr))
```

```
##         Point.Forecast   Lo.80  Hi.80   Lo.95  Hi.95
## 2005 Q1         0.0103 -0.0014 0.0219 -0.0076 0.0281
```

Optional Resources

Optional resources for this chapter

- AIC and BIC: Hanck et al. (2018, §14.6)
- Forecast model evaluation and selection: Hyndman and Athanasopoulos (2019, §§3.4,5.5) and function `forecast::CV()`
- Autoregression: Hyndman and Athanasopoulos (2019, §8.3)
- Lagged predictors: Hyndman and Athanasopoulos (2019, §9.6)

- Example data: `fpp2` package in R (Hyndman, 2018)

Empirical Exercises

Empirical Exercise EE15.1. You will analyze annual U.S. unemployment and inflation data from the 2004 Economic Report of the President, Tables B-42 and B-64. The goal is to forecast the unemployment rate. We'll use the first $T-1$ observations to build a forecast, then compare our forecast to the actual observation in time T.

a. R only: load the needed packages (and install them before that if necessary) and look at a description of the dataset:

   ```
   library(wooldridge); library(forecast)
   ?phillips
   ```

b. Stata only: load the data with `bcuse phillips , nodesc clear` (assuming `bcuse` is already installed)

c. R only: define `thisyr <- 1995` because the Stata dataset only has through year 1996, so that we can get comparable results. Also define `yr1 <- min(phillips$year)`

d. Tell your software that you have annual (yearly) time series data.

   R: `tsdat <- ts(phillips[phillips$year<=thisyr, ], frequency=1, start=yr1)`

   Stata: `tsset year , yearly`

e. Stata only: define `scalar holdout = 1` and `scalar endyr = year[_N]`

f. Plot the unemployment and inflation time series.

   R: `plot(tsdat[,c('unem','inf')])`

   Stata: `tsline unem inf`

g. Considering AR(p) models with $p = 0, 1, 2, 3, 4$, use the AIC to choose the best model, and estimate such a model.

   R: `ret <- ar(tsdat[,'unem'], aic=TRUE, order.max=4)`

   Stata: `varsoc unem , maxlag(4)` and then `arima unem if year<=endyr-holdout , arima(p,0,0)` but replacing the `p` in `arima(p,0,0)` with whatever lag length the previous `varsoc` command said is optimal. (It's possible to do this programmatically, but it gets complicated.)

h. R only (because Stata displayed this already): compute the BIC values for $p = 0, 1, 2, 3, 4$ with `ret$aic+(log(ret$n.used)-2)*1:length(ret$aic)` which adjusts the AIC values to reflect the BIC's different penalty

i. Using the estimates based on data years up to 1995, compute (dynamic) forecasts for the next ten years, 1996–2005, and plot them.

   R: `(fcARp <- forecast(ret, h=10))` and `plot(forecast(ret, h=10))`

   Stata:

```
tsappend , add(9)
predict fcARp if year>endyr-holdout , y
order year unem fcARp
list year unem fcARp if year>=endyr-holdout
twoway tsline unem || tsline fcARp
```

j. Stata only: delete the previously added rows with `drop if year>endyr`

k. Optional: now consider autoregressive distributed lag (ADL) models with up to 2 lags of unemployment and up to 2 lags of inflation. Compute all the AIC values.

   R:

```
unem <- ts(phillips[, 'unem'], frequency=1, start=yr1)
inf  <- ts(phillips[, 'inf'],  frequency=1, start=yr1)
dat <- cbind(Y=unem, L1Y=lag(unem,-1),
             L2Y=lag(unem,-2), L1X=lag(inf,-1), L2X=lag(inf,-2))
dat1 <- window(dat, start=yr1+2, end=thisyr)
r00 <- lm(Y~1, data=dat1)
r01 <- lm(Y~L1X, data=dat1)
r02 <- lm(Y~L1X+L2X, data=dat1)
r10 <- lm(Y~L1Y, data=dat1)
r11 <- lm(Y~L1Y+L1X, data=dat1)
r12 <- lm(Y~L1Y+L1X+L2X, data=dat1)
r20 <- lm(Y~L1Y+L2Y, data=dat1)
r21 <- lm(Y~L1Y+L2Y+L1X, data=dat1)
r22 <- lm(Y~L1Y+L2Y+L1X+L2X, data=dat1)
AICs <- data.frame(L0.inf=c(AIC(r00),AIC(r10),AIC(r20)),
                   L1.inf=c(AIC(r01),AIC(r11),AIC(r21)),
                   L2.inf=c(AIC(r02),AIC(r12),AIC(r22)) )
rownames(AICs) <- c("L0.unem","L1.unem","L2.unem")
print(AICs, digits=4)
```

   Stata:

```
varsoc unem , maxlag(2) exog()
varsoc unem , maxlag(2) exog(L.inf)
varsoc unem , maxlag(2) exog(L.inf L2.inf)
```

l. Optional: estimate the ADL model with the smallest AIC. For example, if the AIC is smallest with one lag of each variable, then use R command `(ret <- lm(Y~L1Y +L1X, data=window(dat,end=thisyr)))` or Stata command `arima unem L.inf if year<=endyr-holdout , arima(1,0,0)`

m. Optional: compute the ADL forecast for unemployment rate in 1996 and compare it with the AR(p) forecast and actual 1996 value.

   R:

```
newdat <- window(dat, start=thisyr+1, end=thisyr+1)
fcADL <- predict(ret, newdata=newdat)
res <- rbind(fcARp$mean[1], fcADL,
             window(unem,start=thisyr+1,end=thisyr+1))
rownames(res) <- c('AR(p)','ADL','Actual')
colnames(res) <- thisyr+1
print(res)
```

Stata:

```
predict fcADL if year>endyr-holdout , y
order year unem fcARp fcADL
list year unem fcARp fcADL if year>=endyr-holdout
```

Chapter 16

Final Exam

⟹ Kaplan video: Good Luck!

When I teach this class, Week 16 is final exams week. There is no new material this week (because there are no classes). My final exam is cumulative: questions may be about any material from any time during the semester. The exception is that there are no questions about coding in R, although there may be some questions showing statistical results in R.

Bibliography

Akaike, Hirotugu. 1974. "A new look at the statistical model identification." *IEEE Transactions on Automatic Control* 19 (6):716–723. URL https://doi.org/10.1109/TAC.1974.1100705. [257]

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E.-J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel J. Hruschka, Kosuke Imai, Guido Imbens, John P. A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson. 2018. "Redefine statistical significance." *Nature Human Behaviour* 2 (1):6–10. URL https://doi.org/10.1038/s41562-017-0189-z. [52]

Bertrand, Marianne and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4):991–1013. URL https://www.jstor.org/stable/3592802. [112]

Biddle, Jeff E. and Daniel S. Hamermesh. 1990. "Sleep and the Allocation of Time." *Journal of Political Economy* 98 (5.1):922–943. URL https://doi.org/10.1086/261713. [153]

Box, G. E. P. 1979. "Robustness in the Strategy of Scientific Model Building." Tech.

Rep. 1954, Mathematics Research Center, University of Wisconsin–Madison. URL http://www.dtic.mil/docs/citations/ADA070213. [146]

Canty, Angelo and B. D. Ripley. 2019. *boot: Bootstrap R (S-Plus) Functions*. URL https://cran.r-project.org/web/packages/boot. R package version 1.3-23. [56]

Card, David. 1990. “The impact of the Mariel boatlift on the Miami labor market.” *Industrial & Labor Relations Review* 43 (2):245–257. [172]

———. 1995. “Using Geographic Variation in College Proximity to Estimate the Return to Schooling.” In *Aspects of Labour Market Behavior: Essays in Honour of John Vanderkamp*, edited by Louis N. Christophides, E. Kenneth Grant, and Robert Swidinsky. University of Toronto Press, 201–222. [57, 217]

Card, David and Alan B. Krueger. 1994. “Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania.” *American Economic Review* 84 (4):772–793. URL https://www.jstor.org/stable/2118030. [204]

Chang, Yoosoon, Robert K. Kaufmann, Chang Sik Kim, J. Isaac Miller, Joon Y. Park, and Sungkeun Park. 2020. “Evaluating trends in time series of distributions: A spatial fingerprint of human effects on climate.” *Journal of Econometrics* 214 (1):274–294. URL https://doi.org/10.1016/j.jeconom.2019.05.014. [230]

Chetty, Raj, John N Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Whitmore Schanzenbach, and Danny Yagan. 2011. “How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR.” *Quarterly Journal of Economics* 126 (4):1593–1660. URL https://doi.org/10.1093/qje/qjr041. [159]

Chetty, Raj, Nathaniel Hendren, and Lawrence F. Katz. 2016. “The Effects of Exposure to Better Neighborhoods on Children: New Evidence from the Moving to Opportunity Experiment.” *American Economic Review* 106 (4):855–902. URL https://doi.org/10.1257/aer.20150572. [160]

Davison, A. C. and D. V. Hinkley. 1997. *Bootstrap Methods and their Applications*. Cambridge: Cambridge University Press. URL http://statwww.epfl.ch/davison/BMA. [56]

Deming, W. Edwards and Frederick F. Stephan. 1941. “On the Interpretation of Censuses as Samples.” *Journal of the American Statistical Association* 36 (213):45–49. URL https://www.jstor.org/stable/2278811. [19, 20]

Diebold, Francis X. 2018a. “Econometric Data Science.” Department of Economics, University of Pennsylvania. http://www.ssc.upenn.edu/~fdiebold/Textbooks.html. [xv, 236, 237, 242, 252]

———. 2018b. “Forecasting.” Department of Economics, University of Pennsylvania. http://www.ssc.upenn.edu/~fdiebold/Textbooks.html. [xv, 221, 231, 232, 236, 242, 247, 252]

———. 2018c. “Time Series Econometrics.” Department of Economics, University of Pennsylvania. http://www.ssc.upenn.edu/~fdiebold/Textbooks.html. [xv, 236]

Freeman, Donald G. 2007. “Drunk Driving Legislation and Traffic Fatalities: New Evidence on BAC 08 Laws.” *Contemporary Economic Policy* 25 (3):293–308. URL https://doi.org/10.1111/j.1465-7287.2007.00039.x. [179]

Grabchak, Michael and Gennady Samorodnitsky. 2010. “Do financial returns have finite or infinite variance? A paradox and an explanation.” *Quantitative Finance* 10 (8):883–893. URL https://doi.org/10.1080/14697680903540381. [124]

Hamilton, James D. 1994. *Time Series Analysis*. Princeton, NJ: Princeton University Press. [245, 252]

Hanck, Christoph, Martin Arnold, Alexander Gerber, and Martin Schmelzer. 2018. “Introduction to Econometrics in R.” URL https://www.econometrics-with-r.org. Department of Business Administration and Economics, University of Duisburg-Essen. [xv, 7, 32, 56, 113, 128, 150, 178, 195, 215, 221, 236, 237, 252, 261]

Hansen, Bruce E. 2020. “Econometrics.” URL https://www.ssc.wisc.edu/~bhansen/econometrics. Textbook draft. [72, 122, 163, 195, 236]

Harrison, David, Jr. and Daniel L. Rubinfeld. 1978. “Hedonic housing prices and the demand for clean air.” *Journal of Environmental Economics and Management* 5 (1):81–102. URL https://doi.org/10.1016/0095-0696(78)90006-2. [197]

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd ed. URL https://web.stanford.edu/~hastie/ElemStatLearn. Corrected 12th printing, January 13, 2017. [xv, 32, 150, 151, 195]

Heckman, James J. 1979. “Sample Selection Bias as a Specification Error.” *Econometrica* 47 (1):153–161. URL https://www.jstor.org/stable/1912352. [213]

Heiss, Florian. 2016. *Using R for Introductory Econometrics*. CreateSpace. URL http://www.urfie.net/read.html. [7, 56, 113, 128, 150, 178, 195, 215, 236, 237]

Holmes, E. E., M. D. Scheuerell, and E. J. Ward. 2019. “Applied Time Series Analysis for Fisheries and Environmental Data.” URL https://nwfsc-timeseries.github.io/atsa-labs. NOAA Fisheries, Northwest Fisheries Science Center, 2725 Montlake Blvd E., Seattle, WA 98112. [236]

Hyndman, Rob. 2018. *fpp2: Data for "Forecasting: Principles and Practice" (2nd Edition)*. URL https://CRAN.R-project.org/package=fpp2. R package version 2.3. [262]

Hyndman, Rob, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. 2020. *forecast: Forecasting functions for time series and linear models*. URL http://pkg.robjhyndman.com/forecast. R package version 8.11. [5, 231, 247, 252]

Hyndman, Rob J. and George Athanasopoulos. 2019. *Forecasting: Principles and Practice*. OTexts. URL https://otexts.com/fpp2. [xv, 231, 232, 233, 236, 252, 257, 261]

Hyndman, Rob J. and Yeasmin Khandakar. 2008. "Automatic time series forecasting: the forecast package for R." *Journal of Statistical Software* 26 (3):1–22. URL http://www.jstatsoft.org/article/view/v027i03. [5, 231, 247, 252]

Imbens, Guido and Jeffrey M. Wooldridge. 2007. "What's New in Econometrics: Estimation of Average Treatment Effects Under Unconfoundedness." NBER summer lecture notes, available at http://www.nber.org/WNE/lect_1_match_fig.pdf. [104, 169]

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning*. Springer Texts in Statistics. Springer, 1st ed. URL http://faculty.marshall.usc.edu/gareth-james/ISL/. Corrected 8th printing, 2017. [xv, 7, 113, 128, 150, 151, 195]

Kaplan, David M. 2020. "Distributional and Nonparametric Econometrics." URL http://faculty.missouri.edu/kaplandm/teach.html. Textbook draft. [7, 21, 151, 215]

Kaufmann, Robert K., Heikki Kauppi, and James H. Stock. 2010. "Does temperature contain a stochastic trend? Evaluating conflicting statistical results." *Climatic Change* 101 (3–4):395–405. URL https://doi.org/10.1007/s10584-009-9711-2. [230]

Kleiber, Christian and Achim Zeileis. 2008. *Applied Econometrics with R*. New York: Springer. URL https://eeecon.uibk.ac.at/~zeileis/teaching/AER. [7, 56]

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76 (4):604–620. URL https://www.jstor.org/stable/1806062. [79]

Lewbel, Arthur. 2019. "The Identification Zoo: Meanings of Identification in Econometrics." *Journal of Economic Literature* 57 (4):835–903. URL https://doi.org/10.1257/jel.20181361. [65, 77]

Lucas, Robert E., Jr. 1976. "Econometric policy evaluation: A critique." In *Carnegie–Rochester Conference Series on Public Policy*, vol. 1. North-Holland, 19–46. [202, 248]

Lumley, Thomas. 2004. “Analysis of Complex Survey Samples.” *Journal of Statistical Software* 9 (1):1–19. R package verson 2.2. [5]

———. 2019. “survey: analysis of complex survey samples.” R package version 3.35-1. [5]

Mincer, Jacob. 1974. *Schooling, Experience, and Earnings*. National Bureau of Economic Research. URL http://www.nber.org/books/minc74-1. [134]

Rouse, Cecilia Elena. 1998. “Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program.” *Quarterly Journal of Economics* 113 (2):553–602. URL https://doi.org/10.1162/003355398555685. [216]

Schwarz, Gideon. 1978. “Estimating the Dimension of a Model.” *Annals of Statistics* 6 (2):461–464. URL https://projecteuclid.org/euclid.aos/1176344136. [257]

Shao, Jun. 1997. “An Asymptotic Theory for Linear Model Selection.” *Statistica Sinica* 7 (2):221–242. URL https://www.jstor.org/stable/24306073. [258]

Shea, John. 1993. “The Input-Output Approach to Instrument Selection.” *Journal of Business & Economic Statistics* 11 (2):145–155. URL https://doi.org/10.1080/07350015.1993.10509943. [239]

Shea, Justin M. 2018. *wooldridge: 111 Data Sets from “Introductory Econometrics: A Modern Approach, 6e” by Jeffrey M. Wooldridge*. URL https://CRAN.R-project.org/package=wooldridge. R package version 1.3.1. [5]

Siegelman, Peter and James J. Heckman. 1993. “The Urban Institute Audit Studies: Their Methods and Findings.” In *Clear and Convincing Evidence: Measurement of Discrimination in America*, edited by Michael E. Fix and Raymond J. Struyk. Washington, DC: Urban Institute Press, 187–258. URL http://webarchive.urban.org/publications/105136.html. [80]

Stock, James H. and Mark W. Watson. 2015. *Introduction to Econometrics*. Pearson, 3rd updated ed. URL https://www.pearson.com/us/higher-education/product/Stock-Introduction-to-Econometrics-Update-3rd-Edition/9780133486872.html. [xv]

Street, Brittany. 2018. “The Impact of Economic Opportunity on Criminal Behavior: Evidence from the Fracking Boom.” Working Paper, available at https://sites.google.com/site/brittanyrstreet/research. [173]

Wooldridge, Jeffrey M. 2020. *Introductory Econometrics: A Modern Approach*. Cengage, 7th ed. [5, 6, 79]

Zeileis, Achim. 2004. “Econometric Computing with HC and HAC Covariance Matrix Estimators.” *Journal of Statistical Software* 11 (10):1–17. [5, 111, 113]

Zeileis, Achim and Torsten Hothorn. 2002. “Diagnostic Checking in Regression Relationships.” *R News* 2 (3):7–10. URL https://CRAN.R-project.org/doc/Rnews. [5, 111, 113]

Index