The Tidy Econometrics Workbook

An Undergraduate's Guide to Causal Inference, Instrumental Variables, and Time Series with R

Colleen O'Briant

8/14/2022

Table of contents

Preface **5**

Course Schedule **6**

Programming Philosophy **7**
Q: "How does R compare to other languages?" . . . 7
Declarative vs Imperative Programming . . . 7
Things to Avoid when Programming Declaratively in the Tidyverse . . . 8
Your Approach . . . 8
What is "Good Code"? . . . 9

Setting up your workspace **10**
Install R and RStudio . . . 10
Get Acquainted with the RStudio IDE . . . 10
Install the Tidyverse . . . 10
Install gapminder . . . 10
Install a few Packages we'll use for Plots . . . 11
Install `qelp` . . . 11
Install the Tidyverse Koans . . . 11

1 Least Squares **14**
1.1 Deriving OLS Estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ . . . 14
1.2 Numerical Example . . . 16
1.3 Exercises . . . 16
1.4 References . . . 16

2 Exogeneity **17**
2.1 Classwork 1 #1 . . . 17
2.2 Classwork 1 #2 . . . 18
2.3 Classwork 1 #3 . . . 18
2.4 Conditional Expectations . . . 20
2.5 Exogeneity . . . 20
2.6 Standard Errors . . . 20
2.7 Summary . . . 22
2.8 Exercises . . . 23
2.9 References . . . 23

3 Causal Inference **24**
3.1 Effect of Health Insurance on Health . . . 24
3.2 Selection Bias: Rubin Causal Model . . . 28
3.3 Exercises . . . 30
3.4 References . . . 30

4 Consistency **31**
4.1 Exercises . . . 31

5 Heteroskedasticity **32**
5.1 Exercises . . . 32

6 Time Series **33**
6.1 Exercises . . . 33

7 Stationarity **34**
7.1 Exercises . . . 34

8 Instrumental Variables **35**
8.1 Exercises . . . 35

9 Differences-in-differences **36**
9.1 Exercises . . . 36

Classwork **38**
CW1: Deriving OLS Estimators (analytical) . . . 38
CW2: lm and qplot (R) . . . 38
CW3: dplyr murder mystery (R) . . . 38
CW4: hypothesis testing (analytical) . . . 38
CW5: causal inference (analytical) . . . 38
CW6: causal inference (R) . . . 38
CW7: consistency (analytical) . . . 38
CW8: heteroskedasticity (analytical) . . . 38
CW9: heteroskedasticity (R) . . . 38
CW10: simulation (R) . . . 38
CW11: dynamics (analytical) . . . 38
CW12: dynamics (R) . . . 38
CW13: time trends (analytical) . . . 38
CW14: random walks (half analytical, half R) . . . 38
CW15: IV (analtyical) . . . 38
CW16: IV (R) . . . 38
CW17: Diff-in-diff (anayltical) . . . 38

Math Rules and Formulas **39**
Summation Rules 39
Variance and Covariance 39

References **41**

Preface

This book is the companion text to EC 421, Econometrics Part 2, taught by Colleen O'Briant in Fall 2022. In this course, we'll build on the foundation laid in EC 320 (Econometrics Part 1) and MATH 243 (Probability and Statistics). We'll cover a range of econometric tools and techniques that can be used for data analysis with cross-sectional, time series, and panel data. We'll also spend some time on causal inference.

The course will teach you how to write programs in R to solve your problems, with a focus on clarity and readability. You will learn to program in a functional, declarative style, and to think about using layers of abstraction to develop simple solutions to complicated problems.

Course Schedule

Classwork for the week is always due on Wednesdays at 5pm. Homework is always due before the next class period.

Date	Day	Classwork	Homework
Wed 9/28	1	Syllabus	CH1: Least Squares
Mon 10/3	2	CW1: Deriving OLS Estimators (analytical)	Koans 1-3
Wed 10/5	3	CW2: lm and qplot (R)	Koans 4-7
Mon 10/10	4	CW3: dplyr murder mystery (R)	CH2: Exogeneity
Wed 10/12	5	CW4: hypothesis testing (analytical)	CH3: Causal Inference
Mon 10/17	6	CW5: causal inference (analytical)	Koans 8-10
Wed 10/19	7	CW6: causal inference (R)	CH4: Consistency
Mon 10/24	8	CW7: consistency (analytical)	Ch 5: Heteroskedasticity
Wed 10/26	9	CW8: heteroskedasticity (analytical)	Koans 11-14
Mon 10/31	10	CW9: heteroskedasticity (R)	practice midterm
Wed 11/2	11	**Midterm Exam**	Koans 15-16
Mon 11/7	12	CW10: simulation (R)	CH6: Time Series
Wed 11/9	13	CW11: dynamics (analytical)	Koans 17-18
Mon 11/14	14	CW12: dynamics (R)	CH7: Stationarity
Wed 11/16	15	CW13: time trends (analytical)	Koans 19-20
Mon 11/21	16	CW14: random walks (half analytical, half R)	CH8: IV for causal inference
Wed 11/23	17	CW15: IV (analtyical)	CH9: IV for simultaneous equations
Mon 11/28	18	CW16: IV (R)	CH10: Diff-in-diff
Mon 11/30	19	CW17: Diff-in-diff (anayltical)	practice final
TBA	20	**Final Exam**	

Programming Philosophy

Q: "How does R compare to other languages?"

A: R is a statistical programming language, which means it is designed for working with data. Other languages, such as Python and Java, are general-purpose programming languages, which means they can be used for a wider variety of tasks. But R (with the tidyverse especially) is a great choice for us because of its affinities with functional programming languages like Lisp. In R, your focus will be to program using functions and compositions instead of always needing to get into the nitty-gritty with objects and inheritance.

Declarative vs Imperative Programming

Programming in the tidyverse may feel pretty different from your other experiences learning to program. That's because the tidyverse is *declarative*, not *imperative*. What's the difference? Imperative programming is relatively low-level: you think in terms of manipulating values using for loops and if statements. Declarative programming is programming at a higher abstraction level: you make use of handy functions (AKA abstractions) to manipulate large swaths of data at one time instead of going value-by-value.

A good metaphor for the difference between imperative and declarative programming is this: suppose I'm trying to help you drive from your house to school. Imperative programming is when I send you turn-by-turn directions, and declarative programming is when I tell you to just put "University of Oregon" into your GPS. With declarative programming, I can declare *what I want you to do* without telling you exactly *how I want you to do it* like with imperative programming. Telling you to put "University of Oregon" into your GPS has advantages over giving you turn-by-turn directions: the GPS may have information about traffic and road closures that I'm not aware of. And the declarative approach is much easier for me: I could help the whole class get from their houses to the university by telling everyone to put "University of Oregon" into their GPS's, while sending each person their own set of turn-by-turn instructions would be a lot more work.

Likewise, when you use the tidyverse's abstractions like `filter()`, `mutate()`, `map()`, `reduce()`, and all of ggplot2's great plotting functions, you're taking advantage of the fact that the engineers who built those functions know efficiency tricks in R that you may not be aware of. And when you're programming declaratively, you can continue thinking about your problem

at a high level instead of getting weighed down by nitty-gritty details. When it comes to data analysis, declarative programming has a lot of huge benefits.

But under the hood, all these great tidyverse functions are just a few for loops and if statements. Imperative programming certainly has its time and place, and that time and place is when your problems include implementing an *algorithm* by hand. If you're interested, I highly recommend Project Euler for teaching yourself imperative programming. But imperative programming is not something you'll need for this class. You may have mixed declarative with imperative programming in previous classes, but we'll stay strictly in the declarative territory in this class.

Things to Avoid when Programming Declaratively in the Tidyverse

Use these only when you're programming imperatively in base R:

- for loops (we'll use `map()` instead)
- if statements (we'll use the vectorized function from dplyr `if_else()`)
- `matrix()` (our 2d data structure of choice is the `tibble()`)
- `$` syntax for extracting a column vector from a tibble. We avoid this because our workflow goes like this: vectors go into tibbles and we do data analysis on *tibbles*. Going from tibbles to vectors (what `$` lets you do) is the reverse of what we need, so we avoid it in this class. It just causes unnecessary headaches!

One more thing: I often see students using assignment `<-` wayyyy too much. If you're creating a variable for something, and you only use that thing one other time, and naming that thing doesn't help the readability of your code, why are you creating that variable? If you let your default be "no assignment" instead of "always assignment", then your code will be much prettier and your global environment will stay clean.

Your Approach

When you're stuck on a hard problem, here are the steps I recommend:

1. Get crystal clear about the problem you're trying to solve. Write out what you *have* versus what you *want*.
2. Break the problem into small steps and make a plan about how you're going to do each step.

3. Not sure about how to do a certain step? *Don't* just guess wildly and stop googling every problem you're stuck on. And **get the hell off of stack overflow!** The solutions on that site are usually written in base R, they sometimes pre-date the tidyverse, and even if they work, they won't help your understanding. Instead, get in the habit of reading the help docs for functions. I've created a package called `qelp` (quick help) which is just beginner friendly help docs for almost all the functions you'll need in this class. Other helpful resources are the tidyverse cheat sheets from RStudio (especially on ggplot, dplyr, and purrr), and of course office hours.

What is "Good Code"?

What are we trying to do here?

First, come to terms with the fact that there's no such thing as good code. All code is bad code, and it's OK! You can't be a perfectionist with this stuff.

But *really bad* code is code that is unnecessarily complicated. If you want examples, just check out stackoverflow! We should always be striving to write simple, elegant solutions because those are easy for others to read and understand, easy for *ourselves* to read and understand, they're easy for a data engineer at your future company to optimize, and when something is broken, it's easy to debug.

Let's not get ahead of ourselves though! Good code, first and foremost, solves the problem at hand! If your solution works, you can always just leave it there. That is sometimes the best thing you can do for your sanity.

- Good code…
  - Solves the problem.
  - Solves the problem in the simplest way.
  - Solves the problem in the simplest way, that's also clear and readable for others.
  - Solves the problem in the simplest way, that's also clear and readable for others, with comments that tell readers why you're doing what you're doing.

Setting up your workspace

Install R and RStudio

Follow the instructions here if you don't have R or RStudio downloaded. Select the CRAN mirror nearest to you (probably Oregon State University). If you have a new apple silicon macbook, make sure to download the version of R that says "Apple silicon arm64 build".

An alternative: R and RStudio are both already installed on all academic workstations at UO. The downside is the limited hours, especially on weekends.

Get Acquainted with the RStudio IDE

Watch this video from RStudio to learn a little about the RStudio IDE. Don't get overwhelmed, we'll only use a small subset of the things in there and you'll learn very quickly what's useful to you.

Install the Tidyverse

Run these lines of code in your console to make sure you have the tidyverse installed and attached to your current session.

```
install.packages("tidyverse", dependencies = TRUE)
library(tidyverse)
```

Install gapminder

You'll use this package a lot in the koans.

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

Install a few Packages we'll use for Plots

```
install.packages("gganimate", dependencies = TRUE)
install.packages("hexbin")
```

Install `qelp`

`qelp` (quick help) is an alternative set of beginner friendly help docs I created (with contributions from previous EC421 students) for commonly used functions in R and the tidyverse. Once you have the package installed, you can access the help docs from inside RStudio.

```
install.packages("Rcpp", dependencies = TRUE)
install.packages("devtools", dependencies = TRUE)
library(devtools)
install_github("cobriant/qelp")
```

Now run:

```
?qelp::install.packages
```

If everything went right, the help docs I wrote on the function `install.packages` should pop up in the lower right hand pane. Whenever you want to read the qelp docs on a function, you type `?`, `qelp`, two colons `::` which say "I want the help docs on this function which is from the package qelp", and then the name of the function you're wondering about.

Install the Tidyverse Koans

Visit the koans on github.

Click on the green button that says `Code` and then hit `Download ZIP`.

Find the file (probably in your downloads folder) and open it to unzip it. Navigate to the new folder named `tidyverse_koans-main` and double click on the R project `tidyversekoans.Rproj`. RStudio should open. If it doesn't, open RStudio and go to `File > Open Project` and then find `tidyversekoans.Rproj`.

In RStudio, go to the lower righthand panel and hit the folder `R`. This takes you to a list of 20 exercises (koans) you'll complete as homework over the course of the quarter. The first 3 (`K01_vector`, `K02_tibble`, and `K03_pipe`) will be due before class on Wednesday (July 20).

Open the first koan: `K01_vector.R`. Before you start, modify 2 keybindings:

First, make it so that you can hit `Cmd/Ctrl Shift K` to compile a notebook:

`Macs: Tools > Modify keyboard shortcuts > filter for Compile a Notebook > Cmd Shift K`

`Windows: Tools > > Modify keyboard shortcuts > filter for Compile a Notebook > Ctrl Shift K`

Second, make it so that you can hit `Cmd/Ctrl Shift T` to run the test for only the active koan instead of all the koans:

`Macs: Tools > Modify keyboard shortcuts > Run a test file > Cmd Shift T`

`Windows: Tools > Modify keyboard shortcuts > Run a test file > Ctrl Shift T`

Now hit `Cmd/Ctrl Shift T` (Cmd Shift T on a mac; Ctrl Shift T on windows). You've just tested the first koan. You should see:

`[ FAIL 0 | WARN 0 | SKIP 9 | PASS 0 ]`

What does this mean? If there are errors in your R script, the test will not complete. Since it completed, you know there are no errors. Since `FAIL` is 0, you also haven't failed any of the questions yet. But `PASS` is also 0, so you haven't passed the questions either. Since they're blank right now, the test will skip them. That's why `SKIP` is 9.

The tests are meant to help you figure out whether you're on the right track, but they're not perfect: if you keep failing the tests but you think your answer is correct, don't spend too much time worrying about it. The tests are sometimes a little fragile… They're a work in progress!

Go ahead and start working on the koans and learning about the tidyverse! There's no need to wait until they're due to start the koans. I find that the students who end up becoming the strongest programmers spend a lot of time making sure their koans are well done.

When you're finished with a koan, make sure to run the tests one last time (`Ctrl/Cmd Shift T`) and then publish an html verson of the document (`Ctrl/Cmd Shift K`, and if that doesn't do anything, change the keybinding for `File > Compile Report` to be `Ctrl/Cmd Shift K`). You'll upload the html version to Canvas for me to grade.

One last thing: whenever you want to work on the koans, make sure you open RStudio by opening the `tidyverse_koans-main` project, not just the individ ual koan file. If you open the koans in a session that's not associated with the `tidyverse_koans-main` project, the tests will fail to run. You can always see which project your current session is being associated with by looking at the upper right hand corner of RStudio: if you're in the `tidyverse_koans-main` project, you'll see `tidyverse_koans-main` up there. That's good. If you're in no project at all, you'll see `Project: (None)` up there. That's not good, especially if you want the tests to

run. If you see `Project: (None)`, just click that text and you'll be able to switch over to the `tidyverse_koans-main` project.

1 Least Squares

1.1 Deriving OLS Estimators $\hat{\beta}_0$ and $\hat{\beta}_1$

In this chapter, we'll be recalling what the method of least squares is and how it works. We'll be starting from scratch, so if EC 320 is not top of mind, don't worry!

Suppose education (x) has a linear effect on wage (y). If someone has zero years of education, they will earn $5 per hour on average, and every extra year of education a person has results in an extra 50 cents added to their wage. Then a linear model would be the correct specification:

$$wage_i = \beta_0 + \beta_1 education_i + u_i$$

Where $\beta_0 = 5$ and $\beta_1 = 0.50$.

When we take some data on the education and earnings of a bunch of people, we could use OLS to *estimate* β_0 and β_1. I'll put hats on the betas to indicate they are estimates: $\hat{\beta}_0$ and $\hat{\beta}_1$ are our estimates of the true parameters β_0 and β_1. We might get $\hat{\beta}_0 = 4$ and $\hat{\beta}_1 = 0.75$ instead of the true values of the parameters $\beta_0 = 5$ and $\beta_1 = 0.50$.

β_0 is the true value of the intercept: if x takes on a 0, this is the expected value for y to take on. In mathematical terms, this is a conditional expectation: $E[y|x = 0] = \beta_0$, which is pronounced "the expectation of y **given** x takes 0 is β_0". And β_1 is the true effect of x on y: if x increases by one unit, β_1 is the amount by which y is expected to increase. In mathematical terms: $E[y|x = \alpha + 1] - E[y|x = \alpha] = \beta_1$ for any α.

1: Introduction to OLS

2: OLS is a method to combine observations

The method of least squares was first published by Frenchman Adrien Marie Legendre in 1805, but there is controversy about whether he was the first inventor or it was the German mathematician and physicist Carl Friedrich Gauss. The method of least squares founded the study of statistics, which was then called "the combination of observations," because that's what least squares helps you do: combine observations to understand a true underlying process. Least squares helped to solve two huge scientific problems in the beginning of the 1800s:

1. There's a field of science called Geodesy that was, at the time, concerned with measuring the circumference of the globe. They had measurements of distances between cities and angles of the stars at each of the cities, done by different observers through different procedures. But until least squares, they had no way to combine those observations.
2. Ceres (the largest object in the asteroid belt between Mars and Jupiter) was discovered. "Speculation about extra-terrestrial life on other planets was open to debate, and the potential new discovery of such a close neighbour to Earth was the buzz of the scientific community," Lim et al. (2021). Astronomers wanted to figure out the position and orbit of Ceres, but couldn't extrapolate that with only a few noisy observations. Until least squares came along.

The method of least squares quickly became the dominant way to solve this statistical problem and remains dominant today.

One reason the method of least squares is so popular is that it's so simple: the entire procedure can be summed up in one statement: **the method of least squares fits a linear model that minimizes the sum of the squared residuals.**

In the next few videos, we'll see that for a simple regression, we can take that statement of the method of least squares and derive:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_i x_i y_i - \bar{x}\bar{y}n}{\sum_i x_i^2 - \bar{x}^2 n}$$

3: Residuals are vertical distances: $e_i = y_i - \hat{y}_i$

4: OLS as $\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_i e_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$

5: $e_i^2 = y_i^2 - 2\hat{\beta}_0 y_i - 2\hat{\beta}_1 x_i y_i + 2\hat{\beta}_0 \hat{\beta}_1 x_i + \hat{\beta}_0^{\,2} + \hat{\beta}_1^{\,2} x_i^2$

6: Some summation rules

Reference these summation rules in the future here.

7: Taking first order conditions

8: Simplifying the FOC for $\hat{\beta}_0$

9: Simplifying the FOC for $\hat{\beta}_1$

1.2 Numerical Example

10: Calculate $\hat{\beta}_0$ and $\hat{\beta}_1$ for a 3 observation example

11: Calculate fitted values $\hat{y}_i$ and residuals e_i for a 3 observation example

12: u_i versus e_i

1.3 Exercises

Classwork 1: Deriving OLS Estimators

Koans 1-3: Vectors, Tibbles, and Pipes

Classwork 2: lm and qplot

Koans 4-7: dplyr

Classwork 3: dplyr murder mystery

1.4 References

Dougherty, C. 2016. Introduction to Econometrics, Chapter 1: Simple Regression Analysis. Oxford University Press.

Lim, Milton, COVID-19 Mortality Working Group, Mike Callan, Actuaries Institute, James Pyne, Chris Dolman, Kitty Chan, and John Connor. 2021. "Gauss, Least Squares, and the Missing Planet." Actuaries Digital. https://www.actuaries.digital/2021/03/31/gauss-least-squares-and-the-missing-planet/#:~:text=The%20early%20history%20of%20statistics,subject%20to%20random

2 Exogeneity

We'll pick up where we left off from chapter 1 with a formula for $\hat{\beta}_1$ from a simple regression $y_i = \beta_0 + \beta_1 x_i + u_i$:

$$\hat{\beta}_1 = \frac{\sum_i (x_i y_i) - n\bar{x}\bar{y}}{\sum_i (x_i^2) - n\bar{x}^2}$$

2.1 Classwork 1 #1

In classwork 1, I asked you to take the formula above and show that:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Why did we do that? What insight about OLS does this give us?

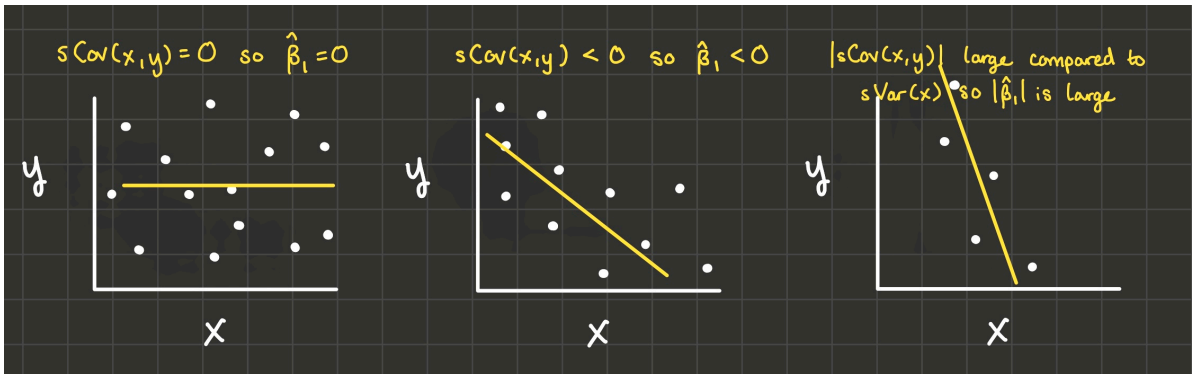
Recall that the sample variance (I'll invent some notation and use $sVar$) of the variable x is: $sVar(x) = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$. And the sample covariance of x and y is $sCov(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$.

So you can see that:

$$\hat{\beta}_1 = \frac{sCov(x, y)}{sVar(x)}$$

A couple of interesting things to point out about this formula:

- If x and y don't covary (their sample covariance is 0), then we'd estimate the slope of the linear model to be 0.
- If they covary negatively (when x is large, y is small and when x is small, y is large), then we'd estimate the slope of the linear model to be negative because the denominator is positive (variances are always positive). And if they covary positively, we'd estimate the slope of the lienar model to be positive.
- The larger in magnitude the covariance of x and y is compared to the variance of x, the steeper the line of best fit is.



2.2 Classwork 1 #2

The next thing we did in classwork 1 was to show:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x}) y_i}{\sum_i (x_i - \bar{x})^2}$$

What intuition does this formula give us?

1: $\hat{\beta}_1$ is a weighted sum of the yi's

2: Numerical Example: calculate w_i

3: Numerical Example: calculate $\hat{\beta}_1$

4: Numerical Example: calculate $\hat{\beta}_1$ with some different values for y_i

2.3 Classwork 1 #3

Finally in question 3, I had you derive a final formula for $\hat{\beta}_1$:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x}) u_i}{\sum_i (x_i - \bar{x})^2}$$

Or if we let $w_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}$, then

$$\hat{\beta}_1 = \beta_1 + \sum_i w_i u_i$$

Note that the $\hat{\beta}_1$ on the left hand side refers to the **estimate** and the β_1 on the right hand side refers to the **true value** of the effect of x on y. So this equation will give us some intuition about when the estimate may not be equal to the true value.

In particular, we'll use this formula to show what assumptions are necessary for $\hat{\beta}_1$ to be an unbiased estimator of β_1: that is, $E[\hat{\beta}_1] = \beta_1$. Taking the expectation of both sides of the equation above and recognizing that the true value of β_1 is a constant:

$$E[\hat{\beta}_1] = \beta_1 + E[\sum_i w_i u_i]$$

And since the expectation of a sum is the same as the sum of the expectations:

$$E[\hat{\beta}_1] = \beta_1 + \sum_i E[w_i u_i]$$

In EC 320, you assumed that explanatory variables x were "predetermined", "nonstochastic", or "randomly assigned" like in a scientific experiment. For instance, x_i would take on 1 if the person was given the medication and x_i would take on 0 if the person was given a placebo. Then u_i absorbs the effect of any unobserved variable like "healthy habits". Because x_i is randomized, we can assume x (medication or placebo) is independent of u (healthy habits). And since w_i is just a function of x, then w would also be independent of u. So by independence,

$$E[w_i u_i] = E[w_i]E[u_i]$$

And if we assume $E[u_i] = 0$ (which is actually a freebie if our model contains an intercept because the intercept will absorb a nonzero expectation for u), then we get:

$$E[\hat{\beta}_1] = \beta_1 + \sum_i E[w_i](0)$$

And $\hat{\beta}_1$ is an unbiased estimator for β_1:

$$E[\hat{\beta}_1] = \beta_1$$

But we don't actually need to make such a strong assumption: x doesn't have to be randomly assigned for OLS to unbiased. A slightly weaker assumption is all that is required: that assumption is called **exogeneity**: $E[u_i|X] = 0$. Exogeneity is that the conditional expectation of u_i given all the explanatory variables across all the observations is zero.

Before we do the proof of the unbiasedness of $\hat{\beta}_1$ under exogeneity, let's talk a little about conditional expectations.

2.4 Conditional Expectations

Proof of the unbiasedness of $\hat{\beta}_1$ under exogeneity:

2.5 Exogeneity

Endogeneity of education in the education-wage model

Exogeneity of treatment in a randomized controlled trial

2.6 Standard Errors

So far, we've established that $\hat{\beta}_1$ is a random variable where $E[\hat{\beta}_1] = \beta_1$ when we have exogeneity: $E[u_i|X] = 0$. What else can we say about the distribution of $\hat{\beta}_1$?

1. $\hat{\beta}_1$ is distributed normally if u_i is distributed normally. Why? $\hat{\beta}_1$ is a weighted sum of u_i:

$$\hat{\beta}_1 = \beta_1 + \sum_i w_i u_i$$

And according to the Central Limit Theorem, that makes $\hat{\beta}_1$ be distributed normally.

2. Under exogeneity, homoskedasticity, and no autocorrelation, the standard error of $\hat{\beta}_1$ (our approximation of the standard deviation of $\hat{\beta}_1$) is $\sqrt{\frac{\sum_i e_i^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$. Here's the proof of that:

$$\hat{\beta}_1 = \beta_1 + \sum_i w_i u_i$$

Take the variance of both sides and recognize that β_1 is a constant that has zero variance:

$$Var(\hat{\beta}_1) = Var\left(\sum_i w_i u_i\right)$$

Recall the definition of the variance of a random variable: $Var(Z) = E\left[(Z - E[Z])^2\right]$.

$$Var(\hat{\beta}_1) = E\left[(\sum_i w_i u_i - E[\sum_i w_i u_i])^2\right]$$

By exogeneity, we've already shown that $E\left[\sum_i w_i u_i\right] = 0$.

$$Var(\hat{\beta}_1) = E\left[(\sum_i w_i u_i)^2\right]$$

Which "foils" to be:

$$Var(\hat{\beta}_1) = E\left[\sum_i w_i^2 u_i^2 + 2\sum_i \sum_j w_i w_j u_i u_j\right]$$

An expected value of a sum is the same as the sum of the expected values:

$$Var(\hat{\beta}_1) = \sum_i E\left[w_i^2 u_i^2\right] + 2\sum_i \sum_j E\left[w_i w_j u_i u_j\right]$$

We're stuck unless we consider the conditional expectations instead of the unconditional ones. If we can show that the conditional expectations are constants, then the unconditional expectations are the same constants:

$$\sum_i E\left[w_i^2 u_i^2 | X\right] = \sum_i w_i^2 E[u_i^2 | X]$$

$$2\sum_i \sum_j E\left[w_i w_j u_i u_j | X\right] = 2\sum_i \sum_j w_i w_j E[u_i u_j | X]$$

Note: $Var(u_i|X) = E\left[(u_i - E(u_i|X))^2 | X\right]$, and since we're assuming exogeneity holds, $Var(u_i|X) = E[u_i^2|X]$. Here we make our next assumption called homoskedasticity: that $Var(u_i|X)$ is a constant.

The same way, note that $Cov(u_i, u_j|X) = E\left[(u_i - E[u_i|X])(u_j - E[u_j|X])|X\right]$, and with exogeneity, $Cov(u_i, u_j|X) = E[u_i u_j]$. If we assume that u_i is not autocorrelated, we can assume $Cov(u_i, u_j|X) = 0$. That will be our next big assumption.

So under these two assumptions of homoskedasticity and no autocorrelation,

$$Var(\hat{\beta}_1) = Var(u)\sum_i w_i^2 + 0$$

Since $w_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}$, we have $\sum_i w_i^2 = \frac{1}{\sum_i (x_i - \bar{x})^2}$.

$$Var(\hat{\beta}_1) = \frac{Var(u)}{\sum_i (x_i - \bar{x})^2}$$

And the standard deviation of $\hat{\beta}_1$ is the square root:

$$sd(\hat{\beta}_1) = \sqrt{\frac{Var(u)}{\sum_i (x_i - \bar{x})^2}}$$

There's just one last problem: u is unobservable, so we can't calculate $Var(u)$ or $sd(\hat{\beta}_1)$ directly. Instead, we *estimate* $sd(\hat{\beta}_1)$ using $sVar(e_i)$ as an approximation for $Var(u)$, and the estimation of the standard deviation of $\hat{\beta}_1$ is what we call the standard error of $\hat{\beta}_1$.

The sample variance of residuals e_i is $sVar(e_i) = \frac{\sum_i (e_i - \bar{e})^2}{n-1}$. Recall that $\bar{e} = 0$. To estimate $Var(u)$ using $sVar(e_i)$, we lose another degree of freedom and divide by $n-2$ instead of $n-1$. So $Var(u)$ is estimated by $\frac{\sum_i e_i^2}{n-2}$. Thus:

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum_i e_i^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$$

2.7 Summary

In this chapter we learned:

- $\hat{\beta}_1 = \frac{sCov(x,y)}{sVar(x)}$
- $\hat{\beta}_1 = \sum_i w_i y_i$: observations far from $\bar{x}$ are the ones that determine the estimate of the effect of x on y.
- $\hat{\beta}_1 = \beta_1 + \sum_i w_i u_i$: exogeneity ($E[u_i|X] = 0$) is the key assumption for $\hat{\beta}_1$ to be unbiased. Exogeneity is met in randomized experiments, but it's violated when there is omitted variable bias.
- Finally, $se(\hat{\beta}_1) = \sqrt{\frac{\sum_i e_i^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$ under exogeneity, homoskedasticity, and no autocorrelation.

Now that we can calculate standard errors, we can do hypothesis tests.

2.8 Exercises

Classwork 4: hypothesis testing

2.9 References

Dougherty (2016) Chapter 1: Simple Regression Analysis

Dougherty (2016) Chapter 8: Stochastic Regressors and Measurement Errors

3 Causal Inference

We know that correlation does not mean causation, but what would it take to convince ourselves that $\hat{\beta}_1$ is the **causal** effect of x on y instead of just describing a correlation?

In this chapter, I'll show you the answer is exogeneity. While it's true that:

$$correlation \neq causation,$$

it's also true that:

$$correlation + exogeneity = causation.$$

3.1 Effect of Health Insurance on Health

But the problem is that $\hat{\beta}_1$ is just a measure of correlation (recall that $\hat{\beta}_1 = \frac{sCov(x,y)}{sVar(x)}$). What would we need to be sure that we've estimated a *causal effect*?

This is what we'd need:

- I'd observe you (say you have health insurance), and I'd measure your health,
- Then I'd need to travel back in time, changing only one thing: your decision to buy health insurance. Then I'd press fast forward and I'd observe your health in this moment, not having health insurance.

To sum it up, I can only figure out how much health insurance has effected you by seeing you in two parallel universes where only one thing has been changed: your decision to buy health insurance. But obviously, we can't observe two parallel universes at once. This is the **fundamental problem of causal inference:** how much a variable truly effects a person is fundamentally unknowable because outcomes in two parallel universes can never be observed at once.

So what's the second-best thing? Instead of trying to identify an *individual* treatment effect, we may be able to identify an *average* treatment effect: the amount that a treatment or a variable x effects a larger population on average.

How? In this chapter we'll explore a couple of different possibilities. Since we've ruled out observing the **same person** at the **same time** with different levels of insurance because of the fundamental problem of causal inference, what if we instead observed:

1) The **same person** at **different times**, where sometimes they have insurance and sometimes they don't?

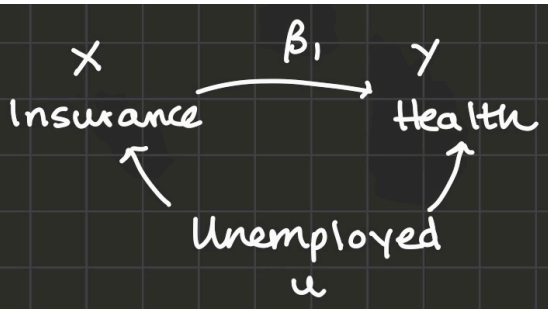
2) **Different people** at the **same time**, where some people have insurance and some people don't?

3) **Twins** at the **same time**, where one twin has health insurance and one twin doesn't?

I'll tackle 1) first. Suppose you have no health insurance between the ages of 26 and 30, and then you have health insurance between the ages of 30 and 34. In your late 20's your average health was a 7 and in your late 30's, your average health was a 8.5. So did having health insurance *cause* the 1.5 point increase in health? Maybe, but maybe not: what if you had no health insurance in your late 20s because you were underemployed? And because you didn't have a great job, you also found yourself anxious and depressed? But then at 30, you finally landed the job of your dreams, you got health insurance because you were employed full time, and you were much happier and healthier? So it could look like health insurance boosted your health, but in reality it was just that you tend to have health insurance at times in your life when your also have steady employment and enjoy better health because of your employment.

For 2) different people at the same time (some insured and some not insured): Can we take the average healths of the insured, subtract the average healths of the uninsured, and consider this a causal effect? Probably not, because just like in the previous paragraph, there's *selection bias*: those who have insurance selected in, so they may be different on unobservables from those who did not. If the uninsured group is more likely to be underemployed (and perhaps more anxious and depressed), again it may look like health insurance makes people much healthier, but actually it's just the effect of steady employment.

You may be wondering: does this have anything to do with exogeneity? Of course it does: what we're really talking about here is exogeneity!



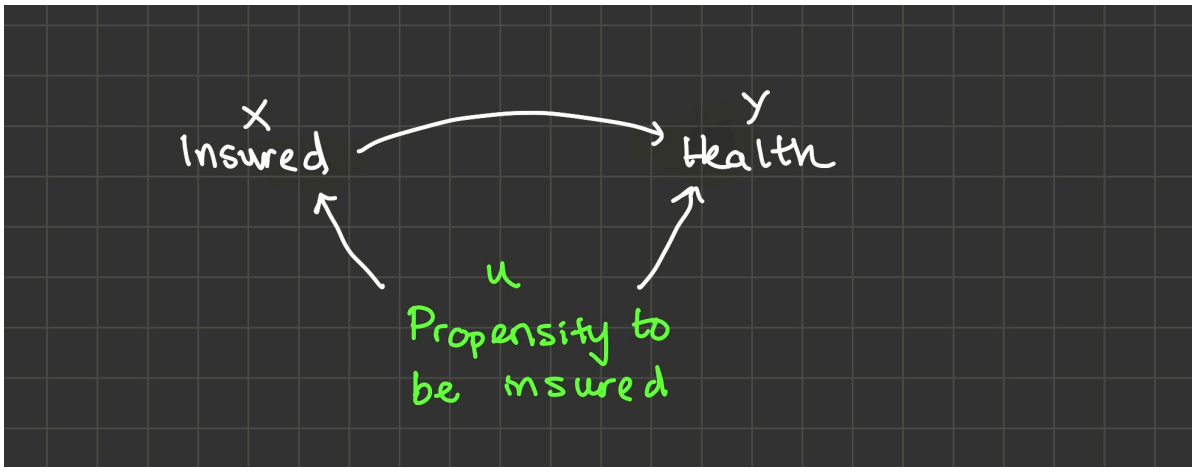
Exogeneity: $E(u_i | X) = 0$

2 people walk in, you only know their X (insured):

Does knowing their X give you any clue about who might be unemployed (u)?

Yes! The uninsured person is more likely to be unemployed.
So $E(u_i | X)$ is not a constant,
$E(u_i | X) \neq 0$,
and insurance is endogenous.

Selection bias is a type of omitted variable bias where the omitted variable is the person's propensity to get treated ("buy health insurance"). A selection bias diagram: if "propensity to be insured" correlates both with "being insured" and someone's "health", then $\hat{\beta}_1$ is biased.



Clearly, someone's propensity to be insured correlates with whether they are insured or not. Does "propensity to be insured" correlate with a person's health? Yes, through multiple channels:

- Stable employment boosts people's propensity to be insured *and* their health, as we've discussed before
- Careful, responsible people are more likely to be insured *and* they're probably healthier because they take care of themselves in other ways as well
- But these variables may be correlated in another way as well: consider a person with a chronic health condition that requires them to frequent the doctor's office or hospital. They would have a higher propensity to be insured because they know they need to visit the doctor frequently. And they also would have a lower health than a person without such a condition.

All of these are reasons why $\hat{\beta}_1$ might be biased due to selection.

Finally, let's consider 3) "Twins at the same time, one of whom has health insurance while the other doesn't". If the twin who has health insurance has a health of 9 while the twin that doesn't has a health of 7, does that mean health insurance boosts people's healths by 2 points? No: we're still worried about selection bias. What other things are different between these twins besides the fact that one has health insurance and one doesn't? **But what if we gave out health insurance randomly to one twin, and not to the other?** That is, what if we did some kind of randomized experiment on these twins, and then observed their healths after a little while? And what if we got a bunch of twins and did the same thing? This would be one way to find the causal effect of health insurance on health because by randomizing who gets health insurance, we're enforcing exogeneity. Why?

Imagine the two twins walk in to the room and you're only told which one has health insurance and which one doesn't. Does that give you any information about which one might have steadier employment, which one might be more responsible, or which one might have a chronic health condition? No! Because we *randomized* which of the twins got the insurance. So

$$E[unemployed,\ responsible,\ chronic\ condition \mid health\ insurance] = 0$$

in a randomized experiment, exogeneity holds and $\hat{\beta}_1$ will be an unbiased estimator of the causal effect of health insurance on health.

And actually we don't need twins after all: we just need a big group of people who we can divide randomly into a treatment and a control group. As long as the treatment and control groups look enough like each other on average, exogeneity will hold. This is why we say *correlation + exogeneity = causation*. And this is why **a randomized controlled experiment (RCT) is the gold standard for causal inference**. At the end of this course, we'll talk about a few second-best approaches for causal inference using instrumental variables and then differences-in-differences, but it's good to keep in mind that if an experiment is ethical and cost-effective, it's the best approach.

So what's the ideal experiment to find the causal effect of some variable X on some variable Y? It's an RCT where you randomize X and compare average differences in Y between treatment and control groups.

3.2 Selection Bias: Rubin Causal Model

The Rubin Causal Model helps us think a little more rigorously about selection bias. Here it is:

There are two types of people: people that choose to get health insurance and people that don't. The people who choose to not get health insurance have some health level which we'll call $health_{0i}$: the 0 indicates that's their health in the universe that they are not insured.

Let's suppose health insurance has some causal effect on a person's health, and we'll call that effect τ_i. Then for the types of people who choose to get health insurance, their health, $health_{1i}$ is equal to their health if they hadn't gotten insured plus the treatment effect: $health_{0i} + \tau_i$. So:

$$health_{1i} = health_{0i} + \tau_i$$

When we estimate the model:

$$health_i = \beta_0 + \beta_1 insurance_i + u_i$$

$\hat{\beta}_1$ will be the average difference in the insured people's healths and the uninsured people's healths:

$$\begin{aligned}\hat{\beta}_1 =& E[health_{1i} \mid type\ of\ people\ who\ get\ insured] - \\ & E[health_{0i} \mid type\ of\ people\ who\ don't\ get\ insured]\end{aligned}$$

$$$$

And since $health_{1i} = \tau_i + health_{0i}$,

$$\begin{aligned}\hat{\beta}_1 =& E[\tau_i + health_{0i} \mid type\ of\ people\ who\ get\ insured] - \\ & E[health_{0i} \mid type\ of\ people\ who\ don't\ get\ insured]\end{aligned}$$

Distributing the expectation across $\tau_i + health_{0i}$ and recognizing $E[\tau_i] = \bar{\tau}$:

$$\begin{aligned}\hat{\beta}_1 =& \bar{\tau} + E[health_{0i} \mid type\ of\ people\ who\ get\ insured] - \\ & E[health_{0i} \mid type\ of\ people\ who\ don't\ get\ insured]\end{aligned}$$

Then define *selection bias* as:

$$\begin{aligned}selection\ bias =& E[health_{0i} \mid type\ of\ people\ who\ get\ insured] - \\ & E[health_{0i} \mid type\ of\ people\ who\ don't\ get\ insured]\end{aligned}$$

That is, **selection bias is the average difference in y for the two types of people (people who will choose x = 1 and people who will choose x = 0), insurance level held constant.** It actually doesn't matter if we hold insurance level constant at 0 or at 1: we'll get the same answer. Finally:

$$\hat{\beta}_1 = \bar{\tau} + selection\ bias$$

Numerical Example: Angrist and Pischke (2014)

3.3 Exercises

Classwork 5: Causal Inference (analytical)

Koans 8-10: ggplot2

Classwork 6: Causal Inference (R)

3.4 References

Angrist and Pischke (2014) Chapter 1

4 Consistency

This page is coming soon!

4.1 Exercises

Classwork 7: Consistency

5 Heteroskedasticity

This page is coming soon!

5.1 Exercises

Classwork 8: Heteroskedasticity (analytical)

Koans 11-14: lm, statistical distributions, and functions

Classwork 9: Heteroskedasticity (R)

Koans 15-16: map

Classwork 10: Simulation (R)

6 Time Series

This page is coming soon!

6.1 Exercises

Classwork 11: Dynamics (analytical)

Koans 17-18: lags and first differences

Classwork 12: Dynamics (R)

7 Stationarity

This page is coming soon!

7.1 Exercises

Classwork 13: Time Trends

Koans 19-20: reduce and accumulate

Classwork 14: random walks

8 Instrumental Variables

This page is coming soon!

8.1 Exercises

Classwork 15: IV part 1

Classwork 16: IV part 2

9 Differences-in-differences

This page is coming soon!

9.1 Exercises

Classwork 17: diff-in-diff

Classwork

CW1: Deriving OLS Estimators (analytical)

CW2: lm and qplot (R)

CW3: dplyr murder mystery (R)

CW4: hypothesis testing (analytical)

CW5: causal inference (analytical)

CW6: causal inference (R)

CW7: consistency (analytical)

CW8: heteroskedasticity (analytical)

CW9: heteroskedasticity (R)

CW10: simulation (R)

CW11: dynamics (analytical)

CW12: dynamics (R)

CW13: time trends (analytical)

CW14: random walks (half analytical, half R)

CW15: IV (analtyical)

CW16: IV (R)

CW17: Diff-in-diff (anayltical)

Math Rules and Formulas

For your convenience, listed below are all the math rules we'll use in this course.

Summation Rules

Let x and y be vectors of length n.

1. Summation definition: $\sum_{i=1}^{n} x_i = x_1 + x_2 + ... + x_n$
2. The sum of x + y is the same as the sum of x + the sum of y: $\sum_i (x_i + y_i) = \sum_i x_i + \sum_i y_i$
3. For any constant c, the sum of c * x is the same as c times the sum of x. $\sum_i cx_i = c \sum_i x_i$
4. In general, the sum of x times y is not equal to the sum of x times the sum of y: $\sum_i x_i y_i \neq \sum_i x_i \sum_i y_i$

Variance and Covariance

- Sample variance:

$$sVar(x) = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$$

  The sample variance measures: on average, how far away is each observation from the mean? By squaring the deviance from the mean, it gets rid of negative numbers and makes it so that a few large deviances translate to a much larger variance than many small deviances. Dividing by n - 1 instead of n is called "Bessel's Correction": since the mean $\bar{x}$ was calculated by looking at the same sample data, the deviances from $\bar{x}$ in the sample will be smaller than if we knew and instead used the true expectation of the random variable x. So to estimate the population variance given a sample, we make the number a little bigger by dividing by n - 1 instead of n.

- Sample covariance of two variables x and y:

$$sCov(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

  Notice that this implies that the sample covariance of x with itself is the same as the sample variance of x: $sCov(x, x) = sVar(x)$.

- Population Variance: on average, what is the square deviance of X from its mean? $Var(X) = E[(X - E[X])^2]$
- Population Covariance: $Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$
- Variance Rules:
  - The variance of a constant is zero: $Var(c) = 0$
  - The variance of a constant times a random variable: $Var(cX) = c^2 Var(X)$
  - The variance of a constant plus a random variable: $Var(c + X) = Var(X)$
  - The variance of the sum of two random variables: $Var(X+Y) = Var(X)+Var(Y)+2Cov(X, Y)$
- Covariance Rules:
  - The covariance of a random variable with a constant is 0: $Cov(X, c) = 0$
  - The covariance of a random variable with itself is its variance: $Cov(X, X) = Var(X)$
  - You can bring constants outside of the covariance: $Cov(X, cY) = cCov(X, Y)$
  - If Z is a third random variable: $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

References

Angrist, J. D., and J. S. Pischke. 2014. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press. https://books.google.com/books?id=dEh-BAAAQBAJ.

Dougherty, C. 2016. *Introduction to Econometrics*. Oxford University Press. https://books.google.com/books?id=Q5cMEAAAQBAJ.

Lim, Milton, COVID-19 Mortality Working Group, Mike Callan, Actuaries Institute, James Pyne, Chris Dolman, Kitty Chan, and John Connor. 2021. "Gauss, Least Squares, and the Missing Planet." *Actuaries Digital*. https://www.actuaries.digital/2021/03/31/gauss-least-squares-and-the-missing-planet/#:~:text=The%20early%20history%20of%20statistics,subject%20to%20random%20measurement%20errors.