Econometrics with the Tidyverse

A Flipped Classroom Experience

Colleen O'Briant

8/14/2022

Table of contents

Part I

Introduction

Welcome to *Econometrics with the Tidyverse*! I'm so happy you're here!

Who am I? My name is Colleen O'Briant. I'm an Economics PhD student at the University of Oregon. My research is on developing the Econometrics of machine learning and AI algorithms. I'm also very passionate about teaching, especially teaching this course.

Who you might be: You might be a student in an Econometrics class. I'll assume you have some familiarity with topics in economics, calculus, probability, and statistics. You don't need to be a PhD student. You also don't need to know how to program: I'll teach you, starting from zero.

By the end of this course, you will have excellent command over R (especially the Tidyverse ecosystem). You'll also solidify your understanding of statitics. And most importantly, you'll be able to think like an applied economist: to evaluate when a regression might be able to answer your research question, and when to expect biased or inconsistent results.

Programming Goals: 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.

Econometrics Goals:

- The basics of deriving the least-squares estimators for a simple regression (Chapter 1)
- How the crucial assumption of exogeneity affects the estimators (Chapter 2)
- How exogeneity allows estimators to have a causal interpretation (Chapter 3)

- The property of consistency and how it can be used to sign the bias of estimators suffering from omitted variable bias (Chapter 4)
- Different model specifications and their interpretations (Chapter 5)
- Heteroskedasticity and how it can be caused by a misspecified model (Chapter 6)
- Topics in Time series models (Chapters 7 and 8)
- Strategies for causal inference when exogeneity cannot be assumed, including instrumental variables and the differences-in-differences estimator (Chapters 9, 10, 11)

Course Materials

The chapters of this workbook: introduce you to the econometrics topics, give you some examples and quiz questions in videos, and point you toward references for further reading.

Koans: short programming exercises to learn the tidyverse. There are 20 of them in total, which you'll complete over the course of the term as homework. Once you've downloaded the project from github, there's a way to check that your answers are correct.

Classwork: will be completed in class in small groups. The classwork alternates between analytic exercises (math proofs and discussion questions) and programming exercises (an applied econometrics project using the tidyverse).

Course Schedule

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

Date	Classwork	Homework
W 9/28	Syllabus	CH1: Least Squares
M 10/3	CW1: Deriving OLS Estimators (analytical)	Koans 1-3
W 10/5	CW2: lm and qplot (R)	Koans 4-7
M 10/10	CW3: dplyr murder mystery (R)	CH2: Exogeneity
W 10/12	CW4: hypothesis testing (analytical)	CH3: Causal Inference
M 10/17	CW5: causal inference (analytical)	Koans 8-10
W 10/19	CW6: causal inference (R)	CH4: Consistency and CH5: Model Specification
M 10/24	CW7: consistency (analytical)	Ch 6: Heteroskedasticity

Date	Classwork	Homework
W 10/26	CW8: heteroskedasticity (analytical)	Koans 11-14
M 10/31	CW9: heteroskedasticity (R)	practice midterm
W 11/2	**Midterm Exam**	Koans 15-16
M 11/7	CW10: simulation (R)	CH7: Time Series
W 11/9	CW11: dynamics (analytical)	Koans 17-18
M 11/14	CW12: dynamics (R)	CH8: Stationarity
W 11/16	CW13: time trends (analytical)	Koans 19-20
M 11/21	CW14: random walks (half analytical, half R)	CH9: IV for causal inference
W 11/23	CW15: IV (analtyical)	CH10: IV for simultaneous equations
M 11/28	CW16: IV (R)	CH11: Diff-in-diff
W 11/30	CW17: Practice Final (anayltical)	
R 12/8	**Final Exam**	

Programming Philosophy

We learn to program (same as we learn anything) by forming a mental model of how the framework works, and then trying to use the framework to solve our own problems (practicing syntax). Sadly, there are lots of ways you can go wrong here: wrong syntax is pretty obvious because you'll get errors, but wrong mental models can stick around for a lifetime and they make programming a real uphill battle.

I've spent a lot of time over the past 3 years (and I've taken a lot of time from the smartest programmers and R people I know) to get this part of the course right. Hopefully you'll find that solving problems using the tidyverse is simple, easy, and natural. But if you skim through the koans and do them as fast as possible, things will not work out! Remember, the koans aren't busywork to be done for speed. They will help you build your mental model and learn the syntax, but only if you take your time, read them carefully, and reflect.

Declarative vs Imperative Programming

One wrong way of programming in the tidyverse is to mix paradigms (declarative and imperative). The tidyverse is declarative. But you'll see a lot of R code online that's imperative, which is written in base R. Mixing the two paradigms makes for confusing, complicated code.

Declarative vs 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 tricks in R that you may not be aware of to make things run smoothly. 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 in this workbook. 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, which prevents lots of confusion.

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 imperatively 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.

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). On Macs, opening the file will unzip it. On Windows, you'll right-click and hit "extract". Then 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.

Part II

OLS Basics

1 Least Squares

1.1 Overview

In this chapter, we'll discuss what the method of least squares is and how it works.

In section 1.2, I'll introduce the method of least squares as the method to **combine observations** in order to make a guess about a linear relationship. In section 1.3, I'll derive OLS estimators from scratch by using the definition of the method of least squares. Finally in section 1.4, I'll do a numerical example where you'll find $\hat{\beta}_0$ and $\hat{\beta}_1$ with 3 observations.

1.1.1 Key Terms and Notation

Symbol	Meaning	Example
β_0	Intercept parameter in a linear model	$y_i = \beta_0 + \beta_1 x_i + u_i$
β_1	Slope parameter in a linear model	see above
y_i	dependent variable, outcome variable	see above
x_i	explanatory variable	see above
u_i	unobservable term, disturbance, shock	see above
$\hat{\beta}_0$	Estimate of the intercept	$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$
$\hat{\beta}_1$	Estimate of the slope	see above
$\hat{y}_i$	Fitted value, prediction	$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$
e_i	residual	$y_i - \hat{y}_i$

1.2 Least Squares as the Combination of Observations

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.

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 and mathematically tractable: 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:

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

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

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

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 $\hat{\beta}_1$

1.4 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

1.5 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

1.6 References

Dougherty (2016) Chapter 1: Simple Regression Analysis

Lim et al. (2021)

2 Exogeneity

2.1 Overview

What to expect in this chapter:

- We'll build more intuition about what OLS does in sections 2.2 and 2.3. In 2.2, we'll see that $\hat{\beta}_1$ is just the sample covariance of x and y divided by the sample variance of x. In 2.3 we'll see that $\hat{\beta}_1$ is also a weighted sum of the y_i's, where observations far from $\bar{x}$ have the largest weights.
- In sections 2.4-2.7, we'll discuss the key assumption for estimates $\hat{\beta}$ to be unbiased (the assumption is called "exogeneity").
- OLS standard error derivation in section 2.8.

Definition. **Exogeneity** - the assumption that $E[u_i|X] = 0$, where u_i is the unobserved term and X is all the explanatory variables in all observations. Exogeneity is not as strong an assumption as independence, but it's stronger than zero covariance. The intuition is that when exogeneity holds, u is as good as random, conditioned on observables X.

Definition. **Standard Error** - our estimate of the standard deviation of $\hat{\beta}$. Under exogeneity, homoskedasticity, and no autocorrelation, the standard error for the slope parameter of a simple regression is: $se(\hat{\beta}_1) = \sqrt{\frac{\sum_i e_i^2}{(n-2)\sum_i (x_i - \bar{x})^2}}$.

2.2 Classwork 1 #1

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

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

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?

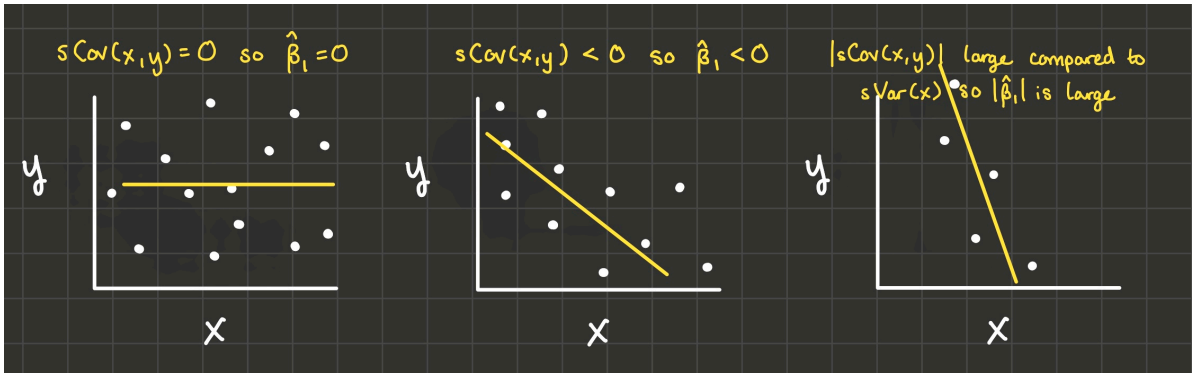
Recall that the sample variance (I'll invent some notation and use $sVar$) of the variable x is: $sVar(x_i) = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$. And the sample covariance of x and y is $sCov(x_i, y_i) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$.

So you can see that:

$$\hat{\beta}_1 = \frac{sCov(x_i, y_i)}{sVar(x_i)}$$

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

- If x and y don't covary (that is, their sample covariance is 0), then we'd estimate the slope of the linear model to be 0 (see the drawing on the left in the image below).
- 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. See the drawing in the middle in the image below.
- 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. See the drawing on the right in the image below.



2.3 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

2.4 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 because $E[A + B] = E[A] + E[B]$:

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

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. The intuition is that when exogeneity holds, u is as good as random, conditioned on observables X.

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

2.5 Conditional Expectations

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

2.7 Exogeneity

Endogeneity of education in the education-wage model

Exogeneity of treatment in a randomized controlled trial

2.8 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 $\hat{\beta}_1$ 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:

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

2.9 Summary

In this chapter we learned:

- $\hat{\beta}_1 = \frac{sCov(x_i, y_i)}{sVar(x_i)}$
- $\hat{\beta}_1 = \sum_i w_i y_i$: 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.

2.10 Exercises

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

Classwork 4: hypothesis testing

2.11 References

Dougherty (2016) Chapter 1: Simple Regression Analysis

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

3 Causal Inference

3.1 Overview

In this chapter, we'll learn how exogeneity makes a causal interpretation of $\hat{\beta}_1$ possible. We know that correlation does not equal causation, but in this chapter, I'll explain why correlation + exogeneity equals causation.

In section 3.2, I'll introduce the fundamental problem of causal inference.

Definition. **The Fundamental Problem of Causal Inference:** an *individual treatment effect* can never be identified because two parallel universes can never be observed at the same time. Example: you'll never know how much your poor sleep effected your performance on an exam you took today because you can never observe your score on the exam in a universe where you got better sleep. The best we can do is to approximate that individual treatment effect with an *average treatment effect* by running an experiment.

In the absence of parallel universes, you might think that looking at the same person at different times will yield an estimate of the average treatment effect. That is, can you look at all the exams you've taken in the past year and the sleep you got the night before those exams, run the regression `score ~ sleep`, and interpret $\hat{\beta}_1$ as causal? In section 3.3, I argue that no, omitted variable bias could be a big problem (maybe you got better sleep *and* scored higher when you were more confident about the material).

In section 3.4, I ask: could you instead compare different people at the same time? That is, take your whole class, ask each person what their score was on the exam, and how much sleep they got the night before. Then run the regression `score ~ sleep`. Does $\hat{\beta}_1$ have a causal interpretation then? The answer is still no, there could be selection bias (the students who decided to get more sleep were the ones who were confident they'd get a good score).

Definition. **Selection Bias:** a type of omitted variable bias where the omitted variable is the person's propensity to have certain values for X.

In section 3.5, I explain that even if you have data about twins at the exact same time (one twin got poor sleep before the exam; the other twin got good sleep), the interpretation is still not causal because of selection bias. The only way to make sure your $\hat{\beta}_1$ is causal is to run a randomized controlled trial (RCT).

Definition. **Randomized Controlled Trial:** a study that assigns participants randomly between control and treatment groups, administers the treatment X, and observes outcome Y.

The chapter ends with section 3.6, which develops the theory of selection bias a little bit more with the Rubin Causal Model.

3.2 Effect of Health Insurance on Health

In the previous video, you calculated $\hat{\beta}_1$, a measure of the correlation between x and y (recall that $\hat{\beta}_1 = \frac{sCov(x,y)}{sVar(x)}$).

To be sure that we've estimated a causal effect with $\hat{\beta}_1$, we would need to observe you (with health insurance), and measure your health. Then we would need to travel back in time, changing only one thing - your decision to buy health insurance. We would then press fast forward and observe your health in this moment, without health insurance.

We can only see the effects of health insurance by observing people in parallel universes. In one universe, people have decided to buy health insurance. In the other universe, they have not. But 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 the same time.

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 X effects some outcome Y for a larger population on average.

How? In this chapter we'll explore a couple of different possibilities. 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.

Let's explore whether $\hat{\beta}_1$ has a causal interpretation in each of these scenarios:

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.

3.3 Same person at different times

I'll tackle 1) first. Suppose you don't have health insurance between the ages of 26 and 30, and then you do 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?

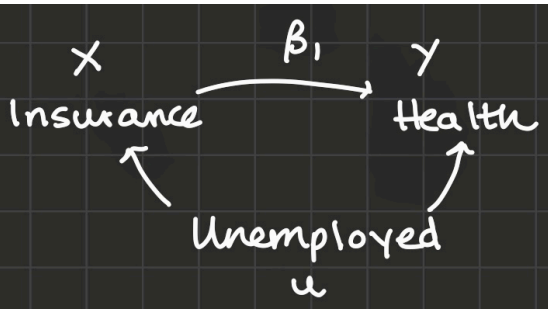
Maybe, but maybe not: what if you had no health insurance in your late 20s because you were underemployed? And because you didn't have a fulfilling job, you also found yourself anxious and depressed? But then at 30, you finally landed the job of your dreams, you got health insurance because you were employed full time, and you were much happier and healthier? 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 you also have steady employment, and you enjoy better health because of your employment.

3.4 Different people at the same time

Can we instead take the average healths of the insured, subtract the average healths of the uninsured, and consider this a causal effect?

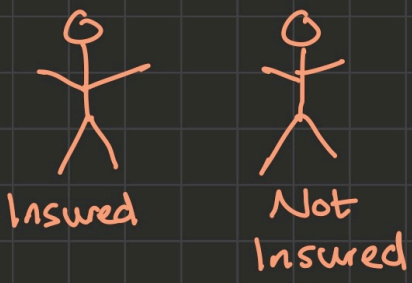
Probably not, because just like in the previous paragraph, there's selection bias: those who have insurance may be different on unobservables from those who don't. If the uninsured group is more likely to be underemployed (and perhaps more anxious and depressed), again it may look like health insurance makes people 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!



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

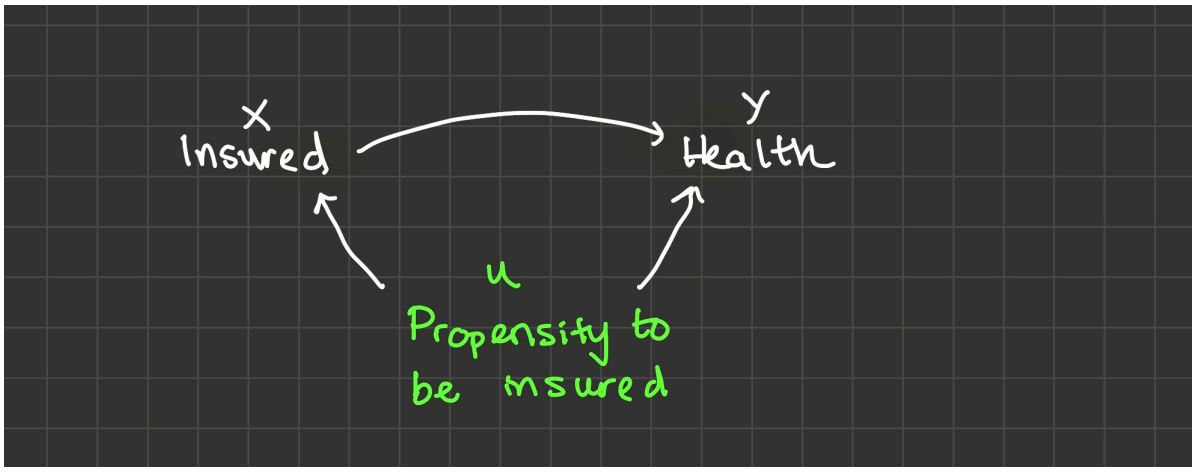
2 people walk in, you only know their X (insured):



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.

3.5 Twins at the same time

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 trial (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.

3.6 Quantifying Selection Bias with the 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: J. D. Angrist and Pischke (2014)

3.7 Exercises

Classwork 5: Causal Inference (analytical)

Koans 8-10: ggplot2

Classwork 6: Causal Inference (R)

3.8 References

J. D. Angrist and Pischke (2014) Chapter 1

4 Consistency

4.1 Overview

What to expect in this chapter:

- Section 4.2 builds some motivation for when you'd need to use the concept of consistency.

Definition. An estimator $\hat{\beta}_1$ is **consistent** if $plim(\hat{\beta}_1) = \beta_1$. That is, $\hat{\beta}_1$ is consistent if it converges in probability to the true value β_1 as n, the number of data points, goes to infinity.

- In section 4.3, I discuss the differences between the concepts of biasedness and consistency.
- In section 4.4, I provide some rules for the plim operator and I prove that $\hat{\beta}_1$ is consistent when $Cov(x_i, u_i) = 0$.

Definition: Probability Limit **plim**. For a sequence of random variables x_n and some value x, $plim(x_n) = x$ if, as n goes to infinity, the probability distribution of x_n collapses to a spike on the value x.

- Section 4.5 explains how the key consistency assumption of 0 covariance between x and u does not hold under omitted variable bias.
- Finally in 4.6, I show how the discussion of consistency also gives us the tools to sign the bias under omitted variable bias.

4.2 Motivation

Let's fast forward a few years. You're at your future job in a brand new data science department at a fast growing company. You're in a meeting and you decide to bring up some concerns you have about selection bias in a model you're developing. Your coworker is dismissive though: they say, "don't worry about selection bias, we'll just get twice the amount of data! How much data do you want? 4 times the amount of data? 10 times?"

You'll have to remember back to econometrics: does selection bias disappear when we let n go to infinity? That is, under selection bias, is OLS consistent? That is the research question

for today. *(spoiler: the answer is no: no amount of data will help if there is selection bias or omitted variable bias. The only solution is to use causal inference techniques like an RCT, instrumental variables (Ch 9), of differences-in-differences (Ch 10).)*

4.3 Bias versus Consistency

Recall that $\hat{\beta}_1$ is **unbiased** iff $E[\hat{\beta}_1] = \beta_1$, and that the key assumption for unbiasedness is exogeneity: $E[u_i|X] = 0$.

For $\hat{\beta}_1$ to be **consistent** however, we need $plim(\hat{\beta}_1) = \beta_1$. That is, as the number of data points n goes to infinity, the probability density function for $\hat{\beta}_1$ must collapse to a spike on the true value β_1 for $\hat{\beta}_1$ to be consistent. "Collapse to a spike" more formally means that $Var(\hat{\beta}_1)$ goes to 0 as n goes to infinity *and* if $\hat{\beta}_1$ is biased, its bias goes to 0 as n goes to infinity. I'll show at the end of the chapter that the key assumption required for $\hat{\beta}_1$ to be consistent is that $Cov(x, u) = 0$.

Bias versus Consistency

Estimators that are consistent and inconsistent; biased and unbiased

Quiz: bias and consistency

4.4 Proof: $\hat{\beta}_1$ is consistent if $Cov(x, u) = 0$

Consistency is defined as $plim(\hat{\beta}_1) = 0$, so to do this proof, first we need some rules about probability limits *plim*.

4.4.1 *plim* rules

Let c be a constant. Let x_n and y_n be sequences of random variables where $plim(x_n) = x$ and $plim(y_n) = y$. That is, for large x, the probability density function of x_n collapses to a spike on the value x and the same for y_n and y. Then:

1) The probability limit of a constant is the constant: $plim(c) = c$
2) $plim(x_n + y_n) = x + y$
3) $plim(x_n y_n) = xy$
4) $plim(\frac{x_n}{y_n}) = \frac{x}{y}$
5) $plim(g(x_n, y_n)) = g(x, y)$ for any function g.

4.4.2 Proof

We'd like to show that $plim(\hat{\beta}_1) = \beta_1$ if $Cov(x, u) = 0$.

I'll start with this formula for $\hat{\beta}_1$, where $sCov$ and $sVar$ refer to the sample covariance and sample variance:

$$\hat{\beta}_1 = \frac{sCov(x_i, y_i)}{sVar(x_i)}$$

If $y_i = \beta_0 + \beta_1 x_i + u_i$, we can substitute in for y_i:

$$\hat{\beta}_1 = \frac{sCov(x_i, \beta_0 + \beta_1 x_i + u_i)}{sVar(x_i)}$$

And use some covariance rules to simplify:

$$\hat{\beta}_1 = \beta_1 + \frac{sCov(x_i, u_i)}{sVar(x_i)}$$

Take the probability limit of both sides and recognize that the probability limit of a constant is the constant:

$$plim(\hat{\beta}_1) = \beta_1 + plim\left(\frac{sCov(x_i, u_i)}{sVar(x_i)}\right)$$

Since $plim(\frac{x_n}{y_n}) = \frac{x}{y}$:

$$plim(\hat{\beta}_1) = \beta_1 + \frac{plim(sCov(x_i, u_i))}{plim(sVar(x_i))}$$

As n increases, a sample variance collapses to the population variance, and the same for covariance:

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(x_i, u_i)}{Var(x_i)}$$

So for $plim(\hat{\beta}_1)$ to be equal to β_1, we just need $Cov(x_i, u_i)$ to be 0.

4.5 $Cov(x_i, u_i) \neq 0$ under omitted variable bias

Suppose the true data generating process is this:

$$wage_i = \alpha_0 + \alpha_1 education_i + \alpha_2 ability_i + v_i$$

But we have to omit *ability*, so we fit this model instead:

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

Then u absorbs v and $\alpha_2 ability_i$, so that $u_i = \alpha_2 ability_i + v_i$. So:

$$\hat{\beta_1} = \beta_1 + \frac{sCov(education_i, u_i)}{sVar(education_i)}$$

And taking probability limits of both sides while substituting in for u_i:

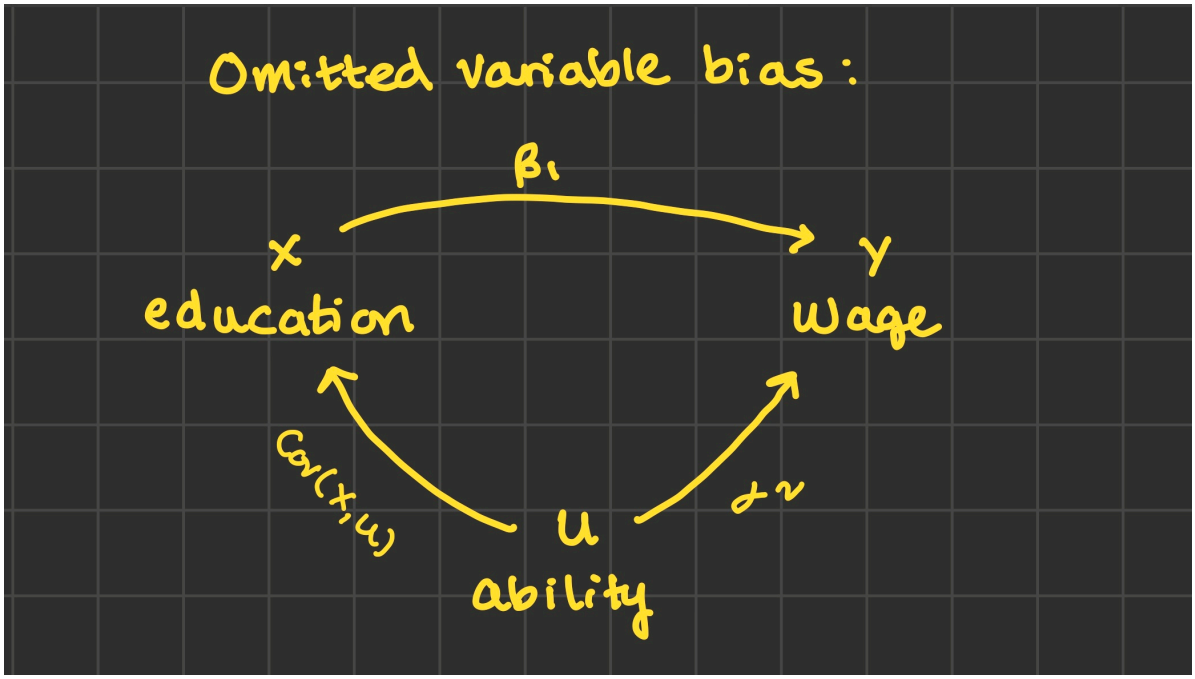
$$plim(\hat{\beta_1}) = \beta_1 + \frac{Cov(education_i, \alpha_2 ability_i + v_i)}{sVar(education_i)}$$

$$plim(\hat{\beta_1}) = \beta_1 + \frac{\alpha_2 Cov(education_i, ability_i) + Cov(education_i, v_i)}{Var(education_i)}$$

Assuming $Cov(education_i, v_i) = 0$:

$$plim(\hat{\beta_1}) = \beta_1 + \frac{\alpha_2 Cov(education_i, ability_i)}{Var(education_i)} \tag{4.1}$$

Compare this to the omitted variable bias diagram:



In the last chapter we learned that if you can draw both lines going out from u (if you can tell stories about why x and u are likely related *and* why u and y are likely related), then u confounds the relationship you're trying to detect between x and y, and $\hat{\beta}_1$ is likely biased.

Now I've added one more detail to the diagram: we can label those lines. Specifically, the relationship between x and u is $Cov(x_i, u_i)$ and the relationship between y and u is α_2 from Equation 4.1. And if both $Cov(x_i, u_i)$ and α_2 are nonzero, both the diagram and Equation 4.1 verify that $\hat{\beta}_1$ will be biased.

4.6 Signing the bias

Labeling the lines in the diagram above does one more thing for us: it lets us sign the bias.

4.7 Exercises

Classwork 7: Consistency

4.8 References

Dougherty (2016) pages 68-75

Rubin (2022)

5 Model Specification

When you're reading papers in applied economics, you'll often see models with transformations of variables (squared, interacted with other variables, logs of variables). This chapter offers some explanation about why you'll see those things. All of these models can be estimated using OLS because while they're not necessarily linear in variables, they're linear in the parameters β.

Models that are linear in parameters (and can be estimated with OLS):

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i \tag{5.1}$$
$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i \tag{5.2}$$
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i \tag{5.3}$$
$$log(y_i) = \beta_0 + \beta_1 log(x_i) + u_i \tag{5.4}$$
$$log(y_i) = \beta_0 + \beta_1 x_i + u_i \tag{5.5}$$

And an example of a model that's not linear in parameters (and can't be estimated with OLS, because β_1 and β_2 can't be separately identified here):

$$y_i = \beta_0 + \beta_1 \beta_2 x_i + u_i$$

5.1 Linear

$y_i = \beta_0 + \beta_1 x_i + u_i$

- Intercept: $\beta_0 = E[y|x = 0]$.
- Slope: β_1 is the expected change in y given an increase in x of one unit.

For example:

$weight_i = -80 + 40 height_i + u_i$

If $weight_i$ is measured in lbs and $height_i$ is measured in feet, then we'd interpret -80 as: "Someone 0 feet tall is expected to weigh -80 lbs". And we'd interpret 40 as "If you're told that a person is 1 foot taller than average, you'd expect them to be 40 lbs heavier than average".

5.2 Squared terms

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + u_i$

5.3 Interactions

An interaction is two variables multiplied. They would usually appear in the model alone as well:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + u_i$

5.4 Log-linear

$log(y_i) = \beta_0 + \beta_1 x_i + u_i$

The formula for exponential growth or decay:

$y = (initial\ amount)\ e^{rt}$

Where r is the rate of change and t is the time (perhaps measured in hours, days, months, etc). The interpretation is that when t increases by 1, y increases by r%.

Let's take the log of both sides. Recalling that $log(ab) = log(a) + log(b)$, and $log(a^b) = b\ log(a)$:

$log(y) = log(initial\ amount) + rtlog(e)$

And since $log(e) = 1$:

$log(y) = log(initial\ amount) + rt$

If we let $\beta_0 = log(initial\ amount)$, $r = \beta_1$, and $t = x$, then we get the log-linear simple regression:

$log(y_i) = \beta_0 + \beta_1 x_i + u_i$

And since $r = \beta_1$, the interpretation of β_1 is the same as the interpretation for r: when t increases by 1, y increases by r%.

5.5 Log-log

$log(y_i) = \beta_0 + \beta_1 log(x_i) + u_i$

Consider a constant elasticity demand curve, where the elasticity ε is the percent change in Q_d corresponding to a 1 percent change in price:

$$Q_d = \beta_0 P^{\beta_1} \tag{5.6}$$

Which parameter represents the elasticity ε?

$$\begin{aligned}
\varepsilon &= \frac{\%\Delta Q_d}{\%\Delta P} \\
&= \frac{\frac{\partial Q}{Q}}{\frac{\partial P}{P}} \\
&= \frac{\partial Q}{\partial P}\frac{P}{Q} \\
&= \frac{\partial(\beta_0 P^{\beta_1})}{\partial P}\frac{P}{Q} \\
&= \beta_0\beta_1 P^{\beta_1 - 1}\frac{P}{Q} \\
&= \beta_0\beta_1 P^{\beta_1 - 1}\frac{P}{\beta_0 P^{\beta_1}} \\
&= \beta_1
\end{aligned}$$

So if we take logs of both sides of Equation 9.1 and change Q to y and P to x:

$$log(y) = log(\beta_0) + \beta_1 log(x)$$

Then we can estimate this model using OLS because it's linear in parameters. β_1 has the same interpretation as an elasticity: it's the expected percent change in y corresponding to a 1 percent change in x. So if we estimate the model and get:

$$log(y) = .25 + .72 log(x)$$

We'd say that a 1 percent increase in x is associated with a .72 percent increase in y.

6 Heteroskedasticity

6.1 Overview

Definition. **Homoskedasticity**: $Var(u_i|X)$ is a constant.
Definition. **Heteroskedasticity**: $Var(u_i|X)$ is some non-constant function of X.

Under heteroskedasticity:

1. OLS is unbiased,
2. But OLS standard errors will not be correct. They could be too small or too large.
3. OLS is no longer BLUE because weighted least squares (WLS) is more efficient.

Look for heteroskedasticity by visual inspection of your data. There are also two formal statistical tests for heteroskedasticity: the Goldfeld-Quandt test and the White test.

6.2 Gauss-Markov Assumptions

OLS is BLUE (the best linear unbiased estimator) if:

1) The data generating process is linear in parameters with an additive disturbance

2) Explanatory variables X are exogeneous: $E(u_i|X) = 0$

3) Explanatory variables X have variation and are not perfectly collinear

4) u_i is iid (independently and identically distributed) $N(0, \sigma^2)$

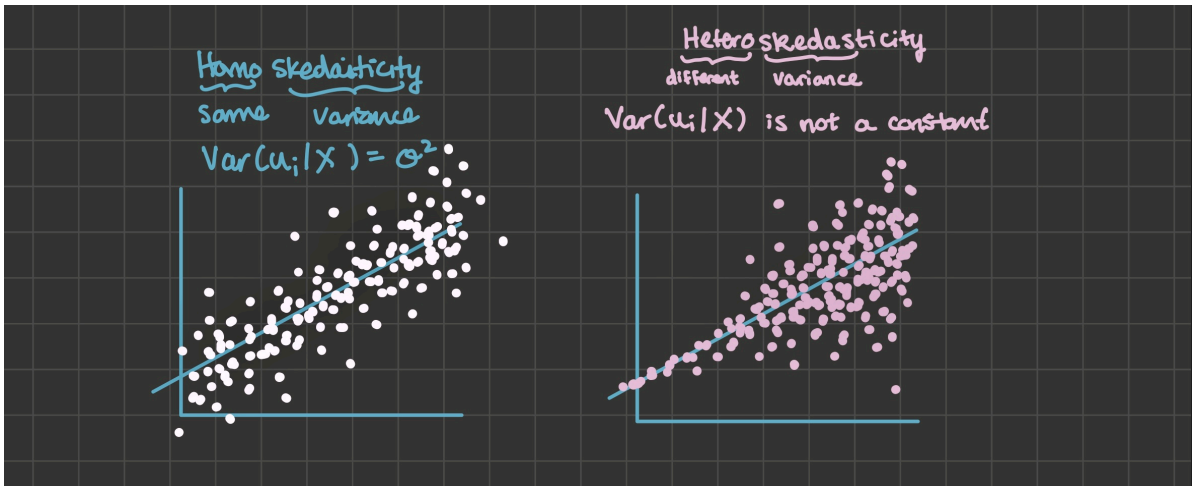
   - If the model has an intercept, the intercept de-means u_i, so $E[u_i] = 0$ is a freebie
   - Homoskedasticity: $Var(u_i) = \sigma^2$, a constant
   - No autocorrelation: $E[u_i, u_j] = 0 \forall i \neq j$: an assumption we'll discuss in the chapter on time series

We needed assumptions 1-3 in the proof of the unbiasedness of OLS, so those assumptions are required for OLS to be unbiased.

We did not need assumption 4 in that proof, so if only assumption 4 is violated, OLS will remain unbiased. But we used assumption 4 when we derived OLS standard errors, so when assumption 4 is violated, OLS standard errors will be incorrect.

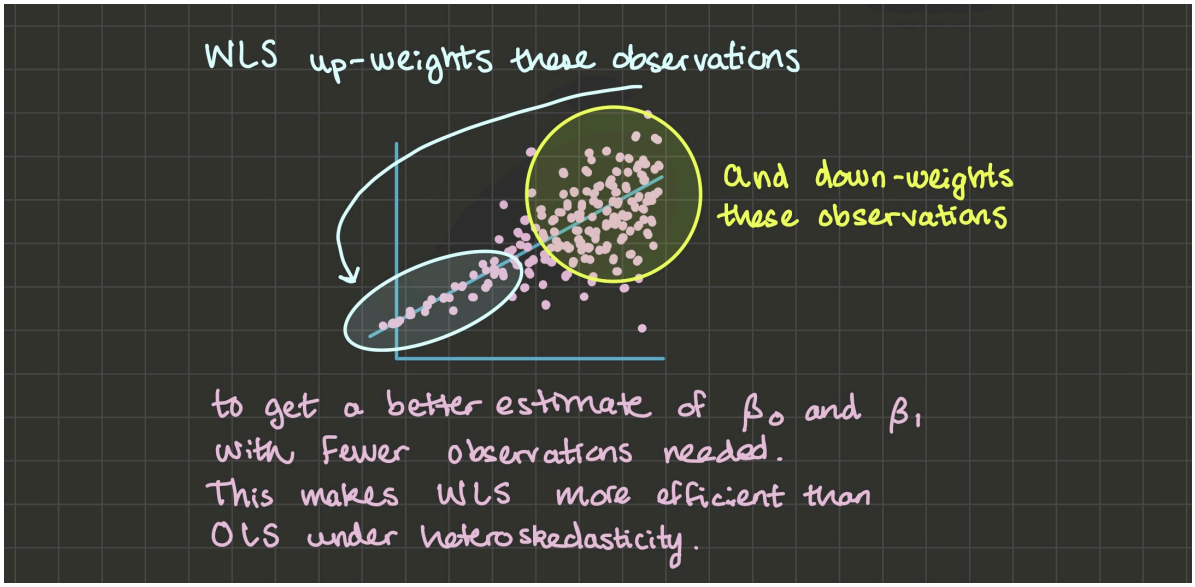
6.3 Detecting Heteroskedasticity through Visual Inspection

Heteroskedasticity tends to be obvious when you plot some explanatory variable on the x-axis and your dependent variable on the y-axis. In the image below, the left hand side illustrates an example of homoskedasticity, where the variance of the u's seem constant across X. The right hand side illustrates an example of heteroskedasticity, where the variance of the u's seems to start small and then increase with X.



6.4 Weighted Least Squares

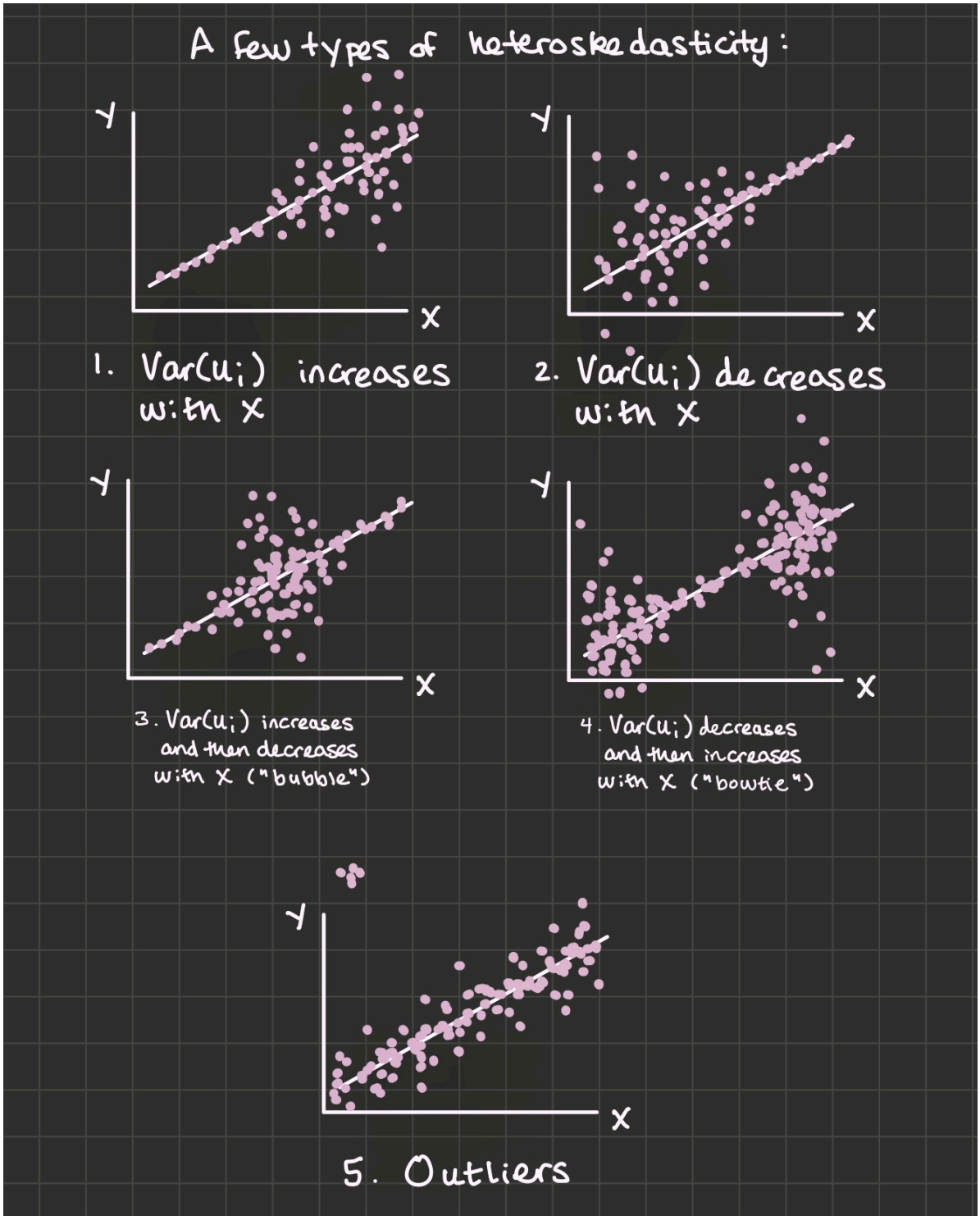
Under heteroskedasticity, OLS is no longer the *best* linear unbiased estimator because weighted least squares (WLS) is more efficient. WLS is very similar to OLS except that you can use it to re-weight observations according to the variance of the u_i's:



Note: heteroskedasticity is the reason you'd see GDP/capita instead of GDP as the dependent variable in a model.

6.5 General types of heteroskedasticity

We'll refer to five general types of heteroskedasticity: "increasing $var(u_i)$", "decreasing $var(u_i)$", "bubble", "bowtie", and heteroskedasticity due to outliers.


A few types of heteroskedasticity:
Y
X
1. Var(u_i) increases with X
2. Var(u_i) decreases with X
3. Var(u_i) increases and then decreases with X ("bubble")
4. Var(u_i) decreases and then increases with X ("bowtie")
5. Outliers

6.6 Tests for heteroskedasticity

6.6.1 Goldfeld-Quandt

1. Arrange the dataset by the explanatory variable you think is associated with the heteroskedasticity.
2. Estimate your model using only the first 3/8 of the data (that is, only low values for x). Then do the same thing for the last 3/8 of the data (only large values for x).
3. Calculate the SSR's (sum of squared residuals $\sum_i e_i^2$) for each of the regressions in step 2. The test statistic is $\frac{SSR_2}{SSR_1}$ where the larger SSR is in the numerator. The idea is that under homoskedasticity, both sides will have similar SSR's and $\frac{SSR_2}{SSR_1}$ will be near 1. But under heteroskedasticity, $\frac{SSR_2}{SSR_1}$ would be much larger than 1.
4. Compare the test statistic to the critical value: $F_{.999,\ df1=df2=(\frac{3}{8}n)-k}$, where k is the number of explanatory variables in the model. If the test statistic is larger than the critical value, the evidence points toward rejecting the null hypothesis of homoskedasticity.

   Which types of heteroskedasticity will the Goldfeld-Quandt test detect?

6.6.2 White Test

1) Estimate the model to get OLS residuals e_i. Square it to get e_i^2. The intuition for the White test is: Does x have explanatory power over e_i^2? If so, that's evidence of possible heteroskedasticity.
2) The test statistic is n * R^2 where the R^2 is from this regression: `lm(e^2 ~ x + x^2)`. Or for a multiple regression: `lm(e^2 ~ x1 + x2 + x1:x2 + x1^2 + x2^2)`.
3) Compare the test statistic to the critical value: $\chi^2_{.999,\ df=k}$, where k is the number of explanatory variables in step 2. Just like in the Goldfeld-Quandt test, a test statistic that's larger than the critical value points to rejecting the null hypothesis of homoskedasticity.

   Which types of heteroskedasticity will the White test detect?

6.7 Heteroskedasticity-Consistent Standard Errors

In chapter 2, we saw that under gauss-markov assumption 4, OLS standard errors are:

$$se(\hat{\beta}_1) = \sqrt{\frac{\sum_i e_i^2}{(n-1)\sum_i (x_i - \bar{x})^2}}$$

Without the homoskedasticity assumption, I'll add an "HC" to indicate they are heteroskedasticity-consistent standard errors:

$$\begin{aligned} Var(\hat{\beta}_1|X)^{HC} &= \sum_i w_i^2 Var(u_i|X) \\ &= \frac{\sum_i (x_i - \bar{x})^2 Var(u_i|X)}{(\sum_i (x_i - \bar{x})^2)^2} \end{aligned}$$

Which White (1980) showed can be estimated by:

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

When you account for heteroskedasticity by using HC standard errors instead of conventional standard errors, you may see that, depending on the type of heteroskedasticity, sometimes your standard errors will increase and sometimes they will decrease. Let's explore this phenomenon to understand why:

$$\begin{aligned} &HC \text{ } standard \text{ } errors > Conv \text{ } standard \text{ } errors \text{ } when: \\ &\sqrt{\frac{\sum_i (x_i - \bar{x})^2 e_i^2}{(\sum_i (x_i - \bar{x})^2)^2}} > \sqrt{\frac{\sum_i e_i^2}{(n-1)\sum_i (x_i - \bar{x})^2}} \\ &\frac{\sum_i (x_i - \bar{x})^2 e_i^2}{(\sum_i (x_i - \bar{x})^2)^2} > \frac{\sum_i e_i^2}{(n-1)\sum_i (x_i - \bar{x})^2} \\ &\frac{\sum_i (x_i - \bar{x})^2 e_i^2}{\sum_i (x_i - \bar{x})^2} > \frac{\sum_i e_i^2}{n-1} \\ &\sum_i (x_i - \bar{x})^2 e_i^2 > \frac{\sum_i e_i^2 \sum_i (x_i - \bar{x})^2}{n-1} \end{aligned}$$

Multiply both sides by $\frac{1}{n-1}$:

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

And if we take the equation above and apply probability limits, we've found that HC standard errors > Conv standard errors when:

$$E[(x_i - \bar{x})^2 e_i^2] > E[(x_i - \bar{x})^2]E[e_i^2]$$

Or, subtracting the right hand side from both sides:

$$E[(x_i - \bar{x})^2 e_i^2] - E[(x_i - \bar{x})^2]E[e_i^2] > 0$$

Finally, recall that you showed in your classwork that $Cov(X, Y) = E[XY] - E[X]E[Y]$, so $Cov(e_i^2, (x_i - \bar{x})^2) = E[e_i^2(x_i - \bar{x})^2] - E[e_i^2]E[(x_i - \bar{x})^2]$.

So HC standard errors > Conv standard errors when:

$$Cov(e_i^2, (x_i - \bar{x})^2) > 0$$

This formula has interesting intuition about heteroskedasticity:

6.8 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)

6.9 References

Dougherty (2016) Chapter 7: Heteroskedasticity

J. Angrist and Pischke (2010)

Part III

Topics in Time Series

7 Time Series

7.1 Introduction

So far, we've been thinking in terms of having cross-sectional data.

Definition. **Cross-sectional Data**: a type of data where a population is sampled at the same time. For example, 1,000 high schoolers across the country are selected and sent a survey. Or 100 small businesses in the state of Oregon are selected and asked questions about how many people they employed this year. Or 100 counties in the US are asked how many cases of covid they had in the last month.

In this chapter and the next, we'll switch focuses toward time series data.

Definition. **Time Series Data**: a type of data where a single subject is observed at different points in time. For example, one student's mathematics standardized test score is observed every year. Or one small business is asked how many people they employ each year for ten years consecutively. Or Lane County is asked how many cases of covid they had in the last month for 30 months in a row.

Then in chapter 11, we'll discuss a causal inference strategy called diff-in-diff that can be used on a third data type, panel data.

Definition. **Panel Data**: a type of data where several subjects are observed at different points in time. For example, 100 students' mathematics standardized test scores are observed every year. We've actually already become familiar with one example of panel data: the gapminder dataset is panel data because it has observations from different countries across different years.

7.2 Overview

What to expect in this chapter:

- 7.3 Time-series models can be **static** or **dynamic**. A simple example of a dynamic model is a model with lags (previous values) of variables as explanatory variables.

Definition. **Static Model**: a time series model where the current value of a variable is modeled by current values of explanatory variables. Example: inflation this month is a function of unemployment this month, *and not* unemployment last month or inflation last month.

Definition. **Dynamic Model**: a time series model where the current value of a variable is modeled by previous values of variables (or even expectations of future values of variables). Example: inflation from the last few months may impact inflation this month, along with unemployment from the last few months.

Definition. **Lag**: a value that a variable took at a previous time period. For example, the lag of GDP per capita in 2022 is the GDP per capita in 2021. There can also be second, third, fourth lags: the second lag of GDP per capita in 2022 is GDP per capita in 2020.

- If you omit a lag of a variable when you shouldn't have, you'll get omitted variable bias. But when you include many lags, you'll get multicollinearity. There are 2 solutions to this problem: first differencing the data 7.4 and including a lag of the dependent variable 7.5.

Definition. **First Difference**: a variable minus the lag of the variable. The first difference is the amount by which the variable has changed going from one period to the next. Example: the first difference of the variable `y` is `y - lag(y)`. If y was 50 last month and is 60 this month, the first difference of `y` this month would be 10 becaue it went up by 10 this month.

- Including a lag of the dependent variable yields biased but consistent estimates as long as u_t is not autocorrelated.
- 7.6 When u_t is autocorrelated, consequences are similar to heteroskedasticity consequences: conventional standard errors are wrong, OLS is no longer BLUE (FGLS is more efficient), but estimates $\hat{\beta}$ are unbiased as long as exogeneity holds. To detect autocorrelation in u_t, use the Breusch-Godfrey test.

7.3 Lags

Consider data for an individual's monthly consumer expenditure and disposable income, from when the person is age 25 to 35. We might design a static model like this:

$$consumption_t = \beta_0 + \beta_1 income_t + u_t$$

The first thing you may notice that's different between time series models and cross-sectional models is the subscript t. This model is saying that the person's consumption *in a certain month* is a linear function of their income *in that same month*. β_0 is still the intercept, or the expectation of the person's consumption if their income that month was zero. β_1 is the

person's marginal propensity to consume: out of every dollar income, the person is expected to spend β_1 dollars.

But if we think about people *consumption smoothing*, or having some habit persistence, the model above doesn't make sense. If there's habit persistence, consumption wouldn't change quickly when income changes. Instead, it might rise or fall slowly over a few months until it reaches a new level corresponding to the new income level. A model that incoorporates some consumption smoothing is a dynamic model with a lag:

$$consumption_t = \beta_0 + \beta_1 income_t + \beta_2 income_{t-1} + u_t$$

Or even multiple lags:

$$consumption_t = \beta_0 + \beta_1 income_t + \beta_2 income_{t-1} + \beta_3 income_{t-2} + \beta_4 income_{t-3} + u_t$$

Suppose the data generating process for $consumption_t$ is:

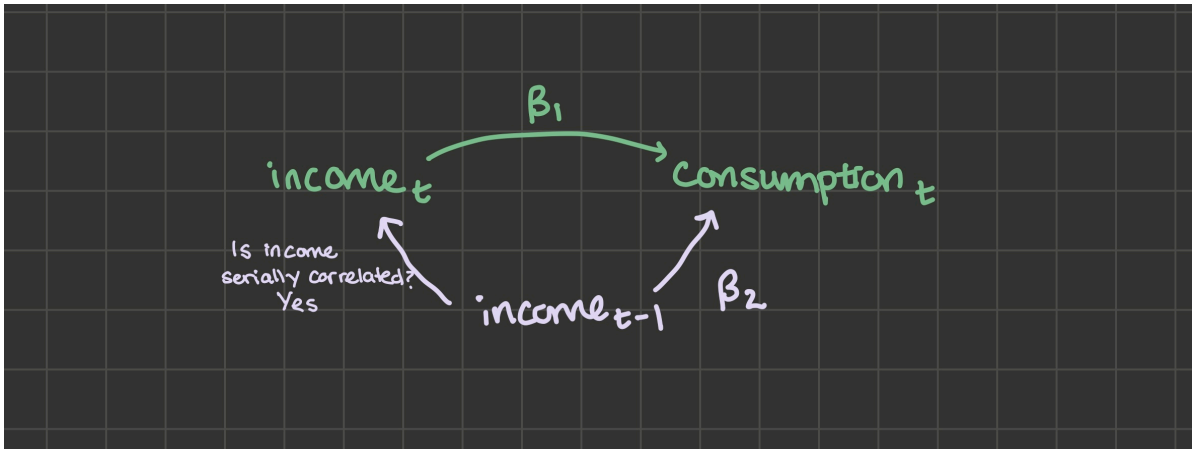
$$consumption_t = \beta_0 + \beta_1 income_t + \beta_2 income_{t-1} + \varepsilon_t$$

But we omit the lag of income and instead fit the static model:

$$consumption_t = \beta_0 + \beta_1 income_t + u_t$$

Will $\hat{\beta}_1$ suffer from omitted variable bias?

Notice that u_t becomes $\beta_2 income_{t-1} + \varepsilon_t$. So $\hat{\beta}_1$ suffers from OVB if we can draw the two lines out from u_t, that is, $income_{t-1}$:



$income_{t-1}$ and $consumption_t$ covary: the relationship between them is β_2. $income_t$ and $income_{t-1}$ also very likely covary: when you get a job, you hold it for multiple time periods and your income in that job is, if not constant, it's serially correlated.

So the answer to the question of "will $\widehat{\beta}_1$ suffer from omitted variable bias if a lag is omitted" is yes.

Should you then always include as many lags as possible? No, because multicollinearity becomes an issue and you'll get large standard errors. I'll provide 2 potential solutions:

1) First difference the data

2) Include a lag of the dependent variable as an explanatory variable

7.4 First Differences

Take the data generating process from above:

$$consumption_t = \beta_0 + \beta_1 income_t + \beta_2 income_{t-1} + u_t \tag{7.1}$$

Step the equation up one period, changing t to $t+1$ and $t-1$ to t:

$$consumption_{t+1} = \beta_0 + \beta_1 income_{t+1} + \beta_2 income_t + u_{t+1} \tag{7.2}$$

This lets us see the **long-run** effect of a 1 unit increase in $income_t$. If $income_t$ gets a 1-unit boost, then Equation 7.1 says $consumption_t$ will get a β_1 boost. And Equation 7.2 says $consumption_{t+1}$ will get a β_2 boost. Since the process only includes the current period's income and last period's income, the bump to $income_t$ won't affect consumption other than in the $consumption_t$ equation and in the $consumption_{t+1}$ equation. So the "long-run" (total) effect of the 1-unit bump to $income_t$ on consumption is $\beta_1 + \beta_2$.

The first-differences model allows you to estimate that long-run effect along with β_1 and β_2 without multicollinearity. Take the $consumption_t$ formula and add and subtract $\beta_2 income_t$:

$$\begin{aligned} consumption_t &= \beta_0 + \beta_1 income_t + \beta_2 income_t + \beta_2 income_{t-1} - \beta_2 income_t + u_t \\ &= \beta_0 + (\beta_1 + \beta_2) income_t - \beta_2 (income_t - income_{t-1}) + u_t \end{aligned}$$

So instead of running `lm(consumption ~ income + lag(income))`, you estimate `lm(consumption ~ income + I(income - lag(income)))`. The variable `income - lag(income)` is called the **first difference** of income: it's the amount by which income changes each period. It will be much less correlated with $income_t$ than $income_{t-1}$ was, so

the multicollinearity issue is solved. And you can back out β_1 and β_2 from the original model. Suppose you estimate that:

$$consumption_t = 3 + 5income_t - 2(income_t - income_{t-1}) + u_t$$

Then the parameters of the original model are 3, 3, and 2.

7.5 Include y_{t-1}

Instead of estimating this model (`y ~ x + lag(x)`):

$$consumption_t = \beta_0 + \beta_1 income_t + \beta_2 income_{t-1} + u_t$$

Estimate this one (`y ~ x + lag(y)`):

$$consumption_t = \beta_0 + \beta_1 income_t + \beta_2 consumption_{t-1} + u_t$$

When you include a lag of the dependent variable as an explanatory variable, you're implicitly including all lags of the explanatory variable. To see this, you can iteratively substitute:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + u_t \tag{7.3}$$

Step the equation back one period:

$$y_{t-1} = \beta_0 + \beta_1 x_{t-1} + \beta_2 y_{t-2} + u_{t-1} \tag{7.4}$$

And substitute Equation 7.4 for the y_{t-1} in Equation 7.3:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 \beta_0 + \beta_2 \beta_1 x_{t-1} + \beta_2^2 y_{t-2} + \beta_2 u_{t-1} + u_t$$

So by including y_{t-1}, we've implicitly included x_{t-1}.

We'll do this process again to substitute for y_{t-2}:

$$y_{t-2} = \beta_0 + \beta_1 x_{t-2} + \beta_2 y_{t-3} + u_{t-2}$$

$$y_t = \beta_0 + \beta_1 x_t + \beta_2(\beta_0 + \beta_1 x_{t-1} + \beta_2(\beta_0 + \beta_1 x_{t-2} + \beta_2 y_{t-3} + u_{t-2}) + u_{t-1}) + u_t$$

And you can see that by including y_{t-1}, we've implicitly included both x_{t-1} and x_{t-2}. We could keep going with this process, and we'd find that by including y_{t-1}, we've implicity included all lags of x_t. There's no multicollinearity issue because we only need to run this regression: `lm(y ~ x + lag(y))`.

Including a lag of the dependent variable has the benefit of letting every lag of x effect y without introducing multicollinearity, *but* it has some costs, which I'll discuss in turn:

1) Including y_{t-1} as an explanatory variable makes the restrictive assumption that lags of x_t effect y_t by way of a specific geometric series.
2) Including y_{t-1} as an explanatory variable makes estimates biased, but they may still be consistent as long as the unobservable term is not serially correlated.

7.5.1 Geometric Series Assumption

To see that including y_{t-1} as an explanatory variable makes the restrictive assumption that lags of x_t effect y_t by way of a specific geometric series, we'll consider the long-run (total) effect of a one-unit bump to x_t on the variable y:

$$y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + u_t$$

The one-unit bump to x_t creates a β_1 bump to y_t. What about y_{t+1}, y_{t+2}, etc? Stepping up the equation one period:

$$y_{t+1} = \beta_0 + \beta_1 x_{t+1} + \beta_2 y_t + u_{t+1}$$

The β_1 sized bump to y_t will create a $\beta_2\beta_1$ sized bump to y_{t+1}. So the total effect so far is $\beta_1 + \beta_2\beta_1$. Stepping up the equation to see the effect on y_{t+2}:

$$y_{t+2} = \beta_0 + \beta_1 x_{t+2} + \beta_2 y_{t+1} + u_{t+2}$$

The $\beta_2\beta_1$ jump to y_{t+1} creates a $\beta_2^2\beta_1$ jump to y_{t+2}. The total effect so far: $\beta_1 + \beta_2\beta_1 + \beta_2^2\beta_1$. If you continue stepping the equation up, you'll see the pattern continues: the effect of a one-unit bump to x_t on y is:

$\beta_1 + \beta_2\beta_1 + \beta_2^2\beta_1 + \beta_2^3\beta_1 + \beta_2^4\beta_1 + ...$

This is an infinite geometric series, which converges as long as $|\beta_2| < 1$. You can find the sum of this series using this trick:

I say that the geometric series assumption is restrictive because including a lag of the dependent variable may not be a good fit if you have reason to believe x_t does not impact y_t in this specific infinite geometric sequence sort of way.

7.5.2 With lagged dependent variables, estimates are biased

Take an even simpler data generating process:

$$y_t = \beta_1 y_{t-1} + u_t$$

With $|\beta_1| < 1$ (a condition we need for stationarity, the subject of the next chapter) and $u_t \sim N(0, \sigma^2)$. We know from chapter 2 that:

$$\hat{\beta}_1 = \beta_1 + \sum_t w_t u_t$$

Where $w_t = \frac{y_{t-1} - \bar{y}}{\sum_t (y_{t-1} - \bar{y})^2}$

The unbiasedness proof proceeds as follows: condition on all the explanatory variables across all observations X:

$$E[\hat{\beta}_1] = \beta_1 + E\left[\sum_t w_t u_t\right]$$

$$\sum_t E[w_t u_t | X] = \sum_t w_t E[u_t | X]$$

But in this case, y_t is included in X, and $E[u_t|y_t]$ is not a constant. Since u_t effects y_t directly, the expectation of u_t depends on y_t. In other words, a lagged dependent variable in X creates endogeneity and therefore OLS estimates are biased.

To develop more intuition about the bias from this dynamic effect, we'll continue: for further simplification, recognize that $\bar{y}$ will be near 0 because $E[y] = 0$:

$$E[y_t] = \beta_1 E[y_{t-1}] + u_t$$

$$E[y] = \beta_1 E[y] + E[u]$$

$$E[y](1 - \beta_1) = 0$$

$$E[y] = 0$$

So

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_t (y_{t-1} - \bar{y}) u_t}{\sum_t (y_{t-1} - \bar{y})^2}$$

Simplifies to:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_t y_{t-1} u_t}{\sum_t y_{t-1}^2}$$

$$E[\hat{\beta}_1] = \beta_1 + E\left[\left(\sum_t y_{t-1} u_t\right)\left(\sum_t y_{t-1}^2\right)^{-1}\right]$$

Evaluating the expectation on the right hand side further is beyond us, but I'll argue that the term is negative because $Cov(\sum_t y_{t-1} u_t, (\sum_t y_{t-1}^2)^{-1}) < 0$: when y_{t-1} and u_t are large, we can expect large y_t (and small y_t^{-1}). And when y_{t-1} and u_t are small, we can expect small y_t (and large y_t^{-1}). So when a lagged dependent variable is included, estimates will be biased *downward*. We'll simulate this result in class.

7.5.3 With lagged dependent variables, estimates may still be consistent

We learned that when a lagged dependent variable is included like this:

$$y_t = \beta_1 y_{t-1} + u_t$$

estimates for β_1 will be biased downward. But they still might be consistent: the key assumption for the consistency of OLS here is that **the unobserved term is not serially correlated**.

Consistency proof:

$$plim(\hat{\beta}_1) = \beta_1 + plim\left(\frac{\sum_t (y_{t-1} - \bar{y}) u_t}{\sum_t (y_{t-1} - \bar{y})^2}\right)$$

Multiply the numerator and denomenator of the fraction by $\frac{1}{n-1}$:

$$plim(\hat{\beta}_1) = \beta_1 + plim\left(\frac{\frac{1}{n-1}\sum_t (y_{t-1} - \bar{y}) u_t}{\frac{1}{n}\sum_t (y_{t-1} - \bar{y})^2}\right)$$

$$plim(\hat{\beta}_1) = \beta_1 + \frac{Cov(y_{t-1}, u_t)}{Var(y_{t-1})}$$

And since u_t is the fresh disturbance, it should not covary with y_{t-1}, which makes the estimator consistent:

$$plim(\widehat{\beta}_1) = \beta_1$$

7.5.4 Caveat: u_t autocorrelated

We just learned that when a lagged dependent variable is used as an explanatory variable, estimates will be biased, but consistent. But there's a caveat here: this will only be true as long as there's no autocorrelation in u_t.

If u_t is autocorrelated, estimates will be biased and inconsistent. Here's why:

Suppose u_t is autocorrelated in this way:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

For some nonzero ρ and iid disturbance ε.

If we're trying to fit the model:

$$y_t = \beta_0 + \beta_1 y_{t-1} + u_t$$

Then estimates will be biased just as before. But here, they'll also be inconsistent:

$$\begin{aligned} plim(\widehat{\beta}_1) &= \beta_1 + \frac{Cov(y_{t-1}, u_t)}{Var(y_{t-1})} \\ &= \beta_1 + \frac{Cov(y_{t-1}, \rho u_{t-1} + \varepsilon)}{Var(y_{t-1})} \\ &= \beta_1 + \frac{\rho Cov(y_{t-1}, u_{t-1})}{Var(y_{t-1})} \\ &= \beta_1 + \frac{\rho}{Var(y_{t-1})} \\ &\neq \beta_1 \end{aligned}$$

7.6 Consequences of autocorrelation in u_t

1) **Conventional standard errors are wrong.** Recall that no autocorrelation in u_t was a major assumption we used when we derived OLS standard errors in chapter 2. In the presence of autocorrelation in u_t, conventional standard errors will be incorrect. There's an adjustment you can make to them called Newey-West standard errors, which we won't go into detail about.

2) $\hat{\beta}_1$ **is unbiased along as we have exogeneity (no lagged dependent variables, no omitted variable bias, etc).** No autocorrelation in u_t was not something we needed to assume to get $\hat{\beta}_1$ to be unbiased.

3) **OLS is no longer BLUE because there exists a more efficient estimator FGLS (feasible generalized least squares).** We'll explore FGLS in the next section.

   Note: These 3 consequences should remind you of the consequences of heteroskedasticity!

7.6.1 FGLS

Feasible Generalized Least Squares (FGLS) is a more efficient estimator than OLS when u_t is autocorrelated, but you need to make an assumption about the type of autocorrelation u_t has. This should remind you about how WLS is a more efficient estimator than OLS under heteroskedasticity, but to use WLS, we had to make an assumption about $Var(u|X)$.

Suppose the model you're trying to fit is this:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

And u_t is autocorrelated. If you're willing to assume u_t is autocorrelated in this way, for example:

$$u_t = \rho u_{t-1} + \varepsilon_t$$

for some unknown value ρ and iid ε_t, then you can take your model and subtract ρy_{t-1} from both sides:

$$y_t - \rho y_{t-1} = \beta_0 + \beta_1 x_t + u_t - \rho y_{t-1}$$

Substitute on the right hand side $y_{t-1} = \beta_0 + \beta_1 x_{t-1} | u_{t-1}$:

$$y_t - \rho y_{t-1} = \beta_0 + \beta_1 x_t + u_t - \rho(\beta_0 + \beta_1 x_{t-1} + u_{t-1})$$

Moving ρy_{t-1} to the right hand side, collecting terms, and substituting ε_t for $u_t - \rho u_{t-1}$:

$$y_t = (\beta_0 - \rho\beta_0) + \rho y_{t-1} + \beta_1 x_t - \rho\beta_1 x_{t-1} + \varepsilon_t$$

So by running the regression `y ~ lag(y) + x + lag(x)`, you can in theory identify ρ, β_0, and β_1. And notice that the error term is only ε, which was assumed iid and therefore not autocorrelated.

7.6.2 Testing for autocorrelation in u_t

When we studied heteroskedasticity, we learned about 2 tests: the Goldfeld-Quandt test and the White test.

There is also a test you could use to detect autocorrelation in u_t:

The intuition is that if u_t is autocorrelated, it's likely that e_t will also be autocorrelated. So after fitting your original model, you could take the residuals and you could fit this model:

$$e_t = \beta_0 + \beta_1 e_{t-1} + \varepsilon_t$$

And do a hypothesis test where the null is $\beta_1 = 0$. That will tell you if the lag of e seems to have an effect on e ("first-order autocorrelation").

But there's a problem: if there's a lagged dependent variable in your original model, or another source of endogeneity, then biased estimates mean that e_t may not be a good enough estimate of u_t. The solution is to add all explanatory variables from the original model x_t to the test to correct for the endogeneity:

$$e_t = \beta_0 + \beta_1 e_{t-1} + \beta_2 x_t + \varepsilon_t$$

This is called the **Breusch-Godfrey test** for autocorrelation. The test statistic is nR^2, where n is the number of observations in the second regression and R^2 is the R-squared from the second regression. Under the null of no autocorrelation, the test statistic is distributed χ^2 with 1 degree of freedom if you're testing for first-order autocorrelation.

7.7 Exercises

Classwork 11: Dynamics (analytical)

Koans 17-18: lags and first differences

Classwork 12: Dynamics (R)

7.8 References

Dougherty (2016) Chapter 11: Models Using Time Series Data

Dougherty (2016) Chapter 12: Autocorrelation

8 Stationarity

8.1 Overview

What to expect in this chapter:

- In 8.2 we'll learn two functions from the tidyverse `reduce(.x, .f)` and `accumulate(.x, .f)`. We'll use these functions to generate autocorrelated data for time series simulations.
- Section 8.3 explains how to generate a random walk and the 3 conditions for a time series to be stationary.
- In 8.4 we'll explore some examples about how running regressions using nonstationary processes can result in spurious (nonsense) regressions.

Definition. **Random Walk**: a time series process y_t where

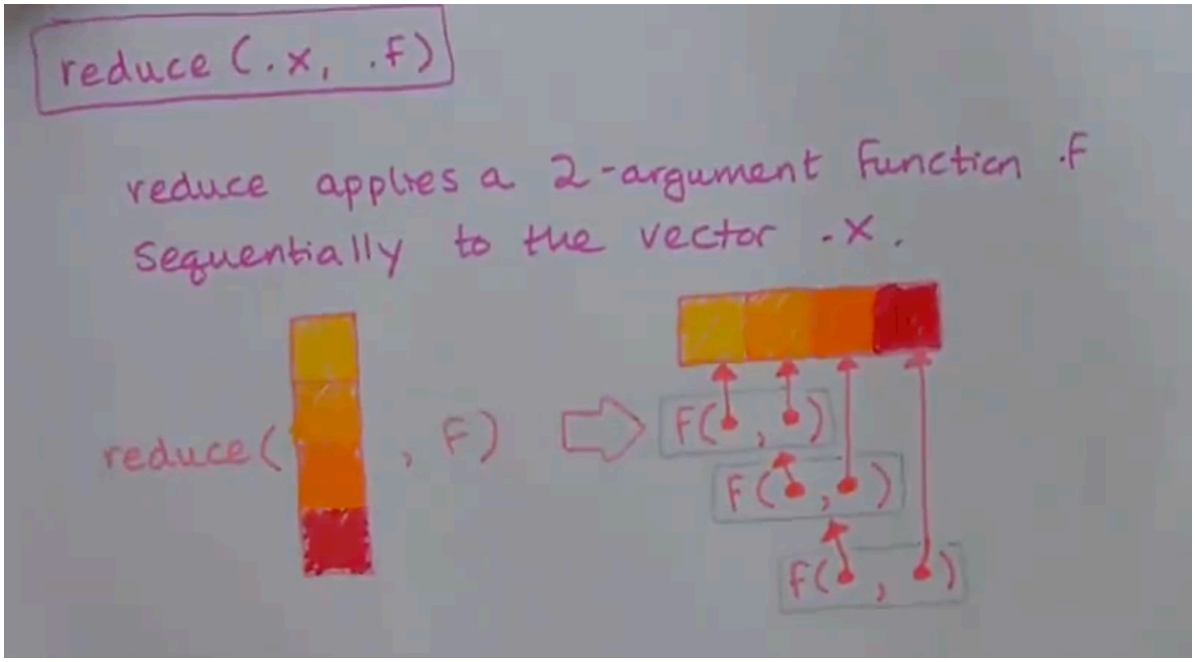
$$y_t = y_{t-1} + \varepsilon_t$$

Where ε_t is iid mean zero.

Definition. **Stationarity**: If a time series process meets all three of these conditions, you can say it is stationary. If it violates any, you can say it is **nonstationary**. 1) The expected value of the process is independent of time: $E[y_t] = E[y_{t-k}]$ for all k. 2) The variance of the process is independent of time: $Var(y_t) = Var(y_{t-k})$ for all k. 3) The series may be autocorrelated, but the nature of the autocorrelation can't be changing over time: $Cov(y_t, y_{t-k}) = Cov(y_s, y_{s-k})$ for all t, s, and k.

8.2 `reduce(.x, .f)` and `accumulate(.x, .f)`

The last two tidyverse functions we'll learn in this class are `reduce()` and `accumulate()`. They're from the same family of functions as `map()`: notice they take the same arguments, a vector `.x` to iterate over and a function `.f` to apply. The way that they apply the `.f` is a little different though.



8.2.1 .f can be named, anonymous, or a formula

Just like with `map()`, the `.f` in `reduce` can be a named function:

```
reduce(.x, intersect)
```

Or a (2-argument) formula:

```
reduce(.x, ~ intersect(.x, .y))
```

Or a (2-argument) anonymous function:

```
reduce(.x, function(x, y) {intersect(x, y)})
```

8.2.2 sum is a reduced +

8.2.3 `accumulate(.x, .f)`

8.3 Stationarity

How does `reduce()` and `accumulate()` help us with time series econometrics? We can use `accumulate()` to generate an autocorrelated series to do monte carlo simulations.

For example, if you wanted to generate data from a process like this:

$$y_t = y_{t-1} + u_t$$

Where u_t is iid $N(0, \sigma^2)$, you can use `accumulate()` to do that. By the way, this process is defined as a **random walk**, and it's how we'd model series driven by speculation like stock prices or housing prices. In markets where speculation is a major driver, the best guess you can make about the price of a stock tomorrow is its price today (if you had a better guess, you could make lots of money, but the point is, no one can consistently). That's what a random walk is: notice $E[y_{t+1}|y_t] = E[y_t + u_{t+1}|y_t] = y_t$ since we assume u_t has mean 0.

To generate data from a random walk, maybe you'd try this (but you'd get an error):

```
library(tidyverse)
```

```
tibble(
  u = rnorm(n = 100),
  y = lag(y) + u
)
#> Error in lag(y) : object 'y' not found
```

The error message says "object 'y' not found" because it can't evaluate `lag(y)` until y exists. What can you do instead?

Take `u = c(1, -1, 0, 1, 1)` and let $y_1 = u_1 = 1$.

What is y_2 if $y_t = y_{t-1} + u_t$?

$$y_2 = y_1 + u_2 = u_1 + u_2 = 1 - 1 = 0$$

How about y_3, y_4, and y_5?

$$y_3 = y_2 + u_3 = u_1 + u_2 + u_3 = 1 - 1 + 0 = 0$$

$$y_4 = y_3 + u_4 = u_1 + u_2 + u_3 + u_4 = 1 - 1 + 0 + 1 = 1$$

$$y_4 = y_4 + u_5 = u_1 + u_2 + u_3 + u_4 + u_5 = 1 - 1 + 0 + 1 + 1 = 2$$

So we should get $y = c(1, 0, 0, 1, 2)$, but the more important thing to notice here is that for a random walk, $y_t = \sum_t u_t$: a random walk is an accumulated sum!

The correct way to generate a random walk in the tidyverse is to accumulate a sum of u's:

```
  tibble(
    u = rnorm(n = 10),
    y = accumulate(u, `+`)
  )
```

```
# A tibble: 10 x 2
        u       y
    <dbl>   <dbl>
 1  0.308  0.308
 2 -0.774 -0.466
 3  0.505  0.0385
 4  1.88   1.92
 5  0.244  2.17
 6  2.19   4.36
 7  1.32   5.68
 8  0.611  6.29
 9  0.994  7.28
10  0.268  7.55
```

8.3.1 first difference a random walk to recover u

Notice what happens when we take the first difference of a random walk:

```
  tibble(
    u = rnorm(n = 10),
    y = accumulate(u, `+`),
    y_diff = y - lag(y)
  )
```

```
# A tibble: 10 x 3
         u      y  y_diff
     <dbl>  <dbl>   <dbl>
 1 -1.09   -1.09  NA
 2  0.331  -0.756  0.331
 3  1.58    0.825  1.58
 4  1.00    1.83   1.00
 5 -0.650   1.18  -0.650
 6  0.777   1.96   0.777
 7  0.861   2.82   0.861
 8  0.0804  2.90   0.0804
```

```
 9  0.318   3.22   0.318
10  1.69    4.91   1.69
```

Notice that `y_diff` is identical to `u` (except `u[1]` can't be identified)! Why?

$y_t = y_{t-1} + u_t$

Subtract y_{t-1} from both sides and you get that the first difference is equal to u:

$y_t - y_{t-1} = u_t$

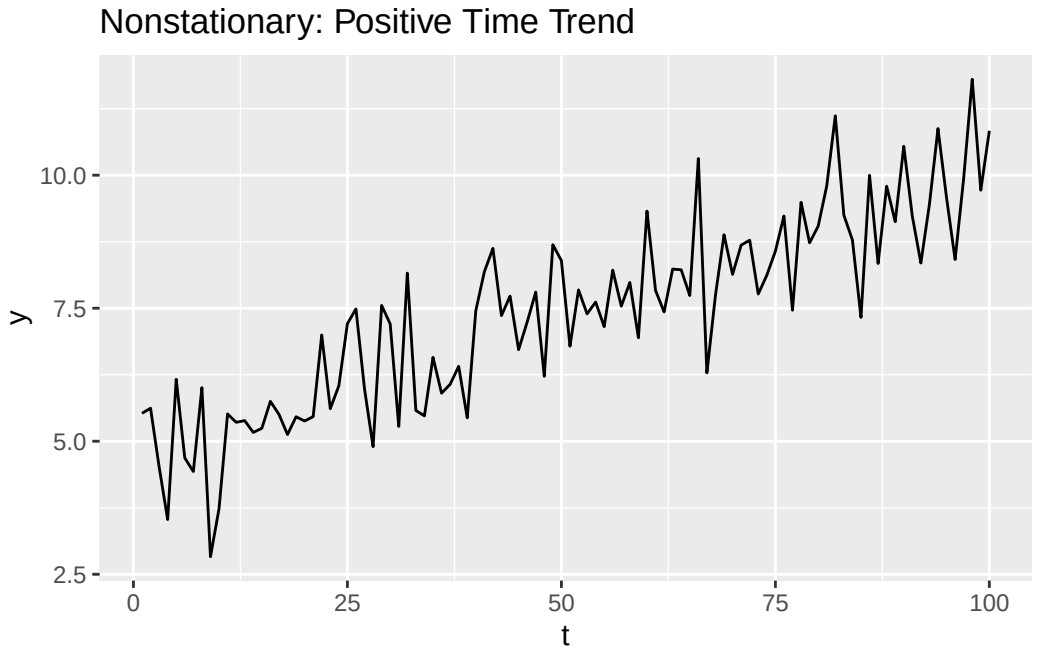
8.3.2 3 conditions for stationarity

There are 3 conditions for a process to be stationary:

1) **The expected value of the process is independent of time: $E[y_t] = E[y_{t-k}]$ for all k.**

If the time series process has a time trend, it will violate this condition. Example where $y_t = 5 + .05t + u_t$:

```
tibble(
    t = 1:100,
    y = 5 + .05 * t + rnorm(n = 100)
) %>%
    ggplot(aes(x = t, y = y)) +
    geom_line() +
    labs(title = "Nonstationary: Positive Time Trend")
```

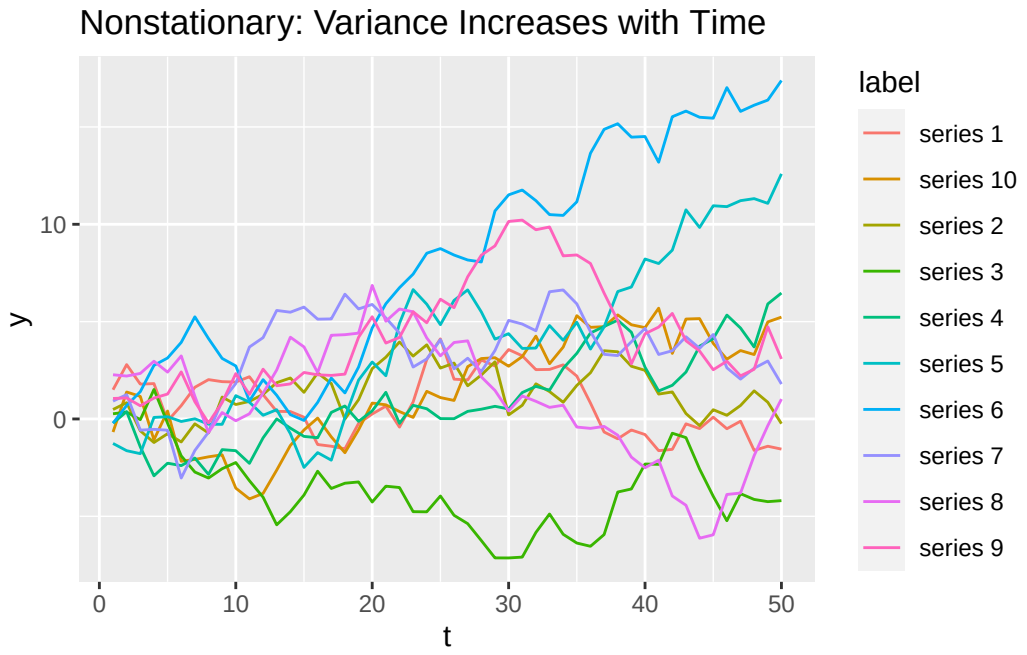


2. **The variance of the process is independent of time:** $Var(y_t) = Var(y_{t-k})$ **for all** k**.**

This is the condition that makes random walks nonstationary. To see this, let's see 10 random walks in one plot:

```
tibble(
    # t is 1:50 repeated 10 times, one for each random walk.

    t = rep(1:50, times = 10),

    # y is 10 random walks. I wanted to repeat the process 10 times so
    # I put it into a map() call. The thing I wanted to repeat 10 times
    # was an accumulated sum of random normals (a random walk).
    # map() outputs a list of length 10 where each element is a random
    # walk of length 50. I used unlist() to drop the structure and make
    # y a vector of length 500.

    y = map(1:10, function(...) accumulate(rnorm(n = 50), `+`)) %>% unlist(),

    # To differentiate the 10 random walks, I need to label them. rep()
    # with an "each" argument will repeat "series 1" 50 times, then
    # "series 2" 50 times, etc.
```

```
    label = rep(paste0("series ", 1:10), each = 50)
) %>%
    ggplot(aes(x = t)) +
    geom_line(aes(y = y, color = label)) +
    labs(title = "Nonstationary: Variance Increases with Time")
```



Notice that all the random walks start near 0 at t = 1, but then start (randomly walking) out and may end up very negative or very positive by t = 50. You'll prove the the variance of a random walk increases with time more rigorously in classwork 14.

3. **The series may be autocorrelated, but the nature of the autocorrelation can't be changing over time:** $Cov(y_t, y_{t-k}) = Cov(y_s, y_{s-k})$ **for all** t**,** s**, and** k**.**

If a series violates any of these 3 conditions for stationarity, it is called a **nonstationary process** and if you put it into a regression, you can often get spurious (nonsense) results. In general when it's possible, economists transform series that they think are nonstationary into stationary series before running regressions with them.

8.4 Spurious regressions

Take a look at this website for some examples of spurious correlations:

It's absolutely true that US spending on science correlates strongly with suicides by hanging, strangulation, and suffocation. But the relationship is obviously not causal. These two processes may both just have an upward time trend (they're nonstationary).

Scrolling down, the number of films Nicolas Cage appeared in correlates strongly with the number of people who drowned by falling into a pool. But we probably don't think there's any kind of causal relationship there. And neither series seem to have time trends. But both trends seem to be autocorrelated, and they may even be random walks, which would make them nonstationary.

8.4.1 Time Trends

Let x_t and y_t be two variables that are totally unrelated, except that they both have time trends.

If you fit the model:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

You'll likely be able to reject the null hypothesis that $\beta_1 = 0$. Why? Omitted variable bias, where t is the omitted variable.

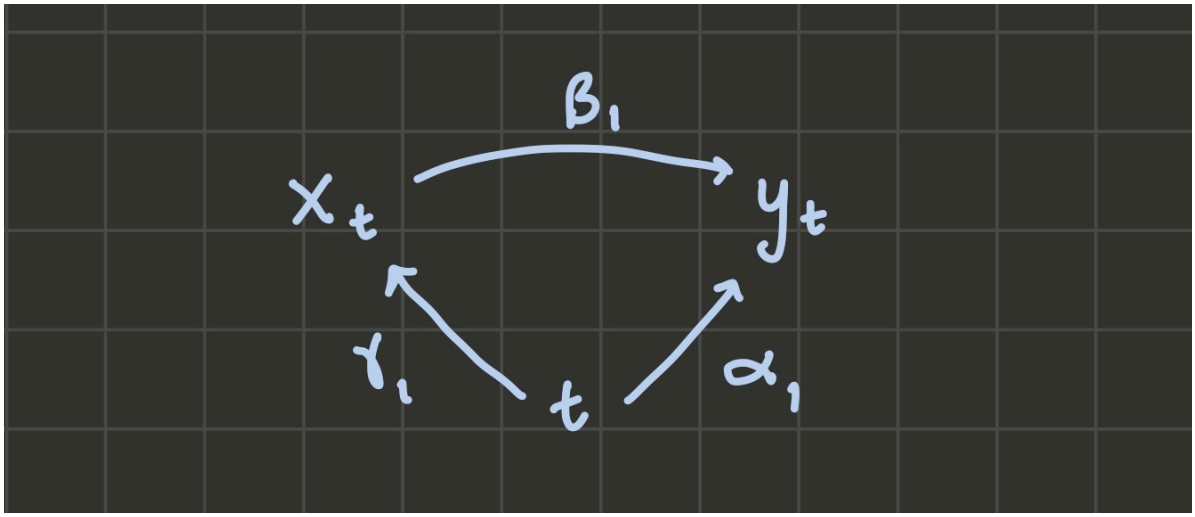
If the true DGP processes for x and y are:

$$y_t = \alpha_0 + \alpha_1 t + w_t$$

$$x_t = \gamma_0 + \gamma_1 t + v_t$$

Where w_t and v_t are both iid $N(0, \sigma^2)$, and you fit

$$y_t = \beta_0 + \beta_1 x_t + u_t$$



Omitting t means t gets absorbed into u_t, and since t covaries with x_t through γ_1 and y_t through α_1, $\hat{\beta}_1$ will be biased and inconsistent, and the sign of the bias will be the same sign as $\alpha_1\gamma_1$.

That is, if both x and y have positive time trends, it will look like and increase in x causes an increase in y. The same is true if both x and y have negative time trends. If x has a positive time trend and y has a negative time trend (or the other way around), $\hat{\beta}_1$ will be negative: it will look like increases in x cause decreases in y. But of course all these results would be spurious (nonsense) because there's no real relationship between x and y, it's just that both have time trends.

8.4.2 Random Walks

If x_t and y_t are two unrelated random walks:

$$x_t = x_{t-1} + v_t$$

$$y_t = y_{t-1} + w_t$$

Where v_t and w_t are iid mean 0.

As you'll show in classwork 14, $Var(x_t) = tVar(v_t)$ and $Var(y_t) = tVar(w_t)$. Random walks are nonstationary because their variances are not independent of time. And when we fit the model:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

$y_{t-1} + w_t$ gets absorbed into u_t. Since y_t is a random walk, y_{t-1} is also a random walk, and u_t will become highly autocorrelated. Recall the consequences of u_t being autocorrelated:

1) OLS estimates remain unbiased, but
2) Conventional standard errors will be incorrect, and
3) OLS isn't BLUE because FGLS is more efficient.

The first two consequences are the ones to focus on here: $\hat{\beta}_1$ is unbiased, so its distribution will be centered on 0, but because of consequence 2, conventional standard errors will be wrong (and in this case, they will be way too small). As a result, you'll incorrectly reject the null that $\hat{\beta}_1 = 0$ about 3/4 of the time. This is a very famous result from Granger and Newbold (1974) which you'll replicate in classwork 14.

How do you transform a random walk into a stationary series? You can take the first difference.

8.5 Exercises

Classwork 13: Time Trends

Koans 19-20: reduce and accumulate

Classwork 14: random walks

8.6 References

Dougherty (2016) Chapter 13: Introduction to Nonstationary Time Series

Part IV

Extensions

9 Instrumental Variables

9.1 Introduction

Recall back to chapter 3: Causal Inference. In that chapter we learned that if our explanatory variables are **exogenous**, then OLS provides unbiased estimates of the causal effect of X on Y.

But if X is **endogenous** (the opposite of exogenous), OLS will be biased. We learned that our options in that case include running an RCT: randomize X and observe differences in Y. But the ideal RCT is often unethical, it might take years or even decades, and it might be way too expensive:

- Randomizing education level and observing earnings down the road
- Randomizing whether someone can migrate to a different city and observing earnings after some time
- Randomizing how many friends of the opposite sex a high schooler can have and observing their GPA after a year or two

In this chapter, we'll learn about a possible alternative to the ideal experiment called **instrumental variables**. You'll learn in this chapter that if a valid instrument Z exists for the endogenous variable X, then we can make an unbiased and consistent estimate of the causal relationship between X and Y using a method called two-stage least squares (2SLS).

9.2 Instrument Validity

There are three conditions for an instrument Z to be valid:

1) z_i is exogenous: $E[u_i|Z] = 0$
2) z_i is relevant: z_i has a causal effect on x_i
3) z_i is excludable: z_i effects y_i only through how it effects x_i

A valid instrument z_i isolates some of the good, exogenous variation in x_i, and lets us ignore the bad, endogenous variation in x_i that leads to omitted variable bias.

For example, the bad, endogenous variation in *education* is the fact that there are high (low) ability people who get more (less) education and go on to get high (low) paying jobs. The good, exogenous variation in education would be the extra education a person gets by pure chance: maybe they happen to live near a university so they are more likely to go to college. Or perhaps the person got more education than a similar person because, while they both dropped out of school when they turned 16, the first person was born in a month that meant they had completed 10.7 years of education where the second person had only completed 10.2 years.

Whether someone lives near a university could be used as an instrument for education to understand the causal effect of education on earnings. So could birth month among 16 year old high school dropouts.

A valid instrument lets us isolate the good, exogenous variation in *education* that comes from the people who, by pure chance, got a little more or a little less education.

9.3 Two Stage Least Squares (2SLS)

To find an IV estimate, we use a method called two-stage least squares (2SLS). It has two stages. The first one checks the relevance of the instrument Z and finds the exogenous variation in X. The second one finds the effect of the exogenous variation in X on the outcome variable Y.

9.3.1 First Stage

In the first stage, we'll use OLS to estimate this model:

$$x_i = \gamma_0 + \gamma_1 z_i + v_i$$

Note that if $\hat{\gamma_1}$ is not statistically different from 0, then Z doesn't effect X strongly enough: it fails the **relevance** assumption and Z is not a valid instrument for X.

But if $\hat{\gamma_1}$ is statistically different from 0, we'll get the **fitted values** from this first stage regression, which we can think of as the good, exogenous variation in X.

$$x_i = x_i^{exog} + x_i^{endog}$$

Where $x_i^{exog} = \hat{x_i} = \hat{\gamma_0} + \hat{\gamma_1} z_i$

And $x_i^{endog} = v_i$.

9.3.2 Second Stage

Then in the second stage, we'll take the fitted values $\hat{x_i}$ from the first stage and we'll use OLS to fit this model:

$$y_i = \beta_0 + \beta_1 \hat{x_i} + w_i$$

And $\hat{\beta_1}$ is our estimate of the causal effect of X on Y.

9.4 Consistency of the IV Estimator

Recall that the OLS estimate of $\hat{\beta_1}$ is the sample covariance of x and y divided by the sample variance of x:

$$\hat{\beta_1}^{OLS} = \frac{sCov(x_i, y_i)}{sVar(x_i)}$$

So it should follow that, from the second stage, $\hat{\beta_1}^{IV} = \frac{sCov(\hat{x_i}, y_i)}{sVar(\hat{x_i})}$.

And taking probability limits:

$$plim(\hat{\beta_1}^{IV}) = \frac{Cov(\hat{x_i}, y_i)}{Var(\hat{x_i})} \tag{9.1}$$

Classwork #15.1: Using Equation 9.1 and the fact that $\hat{x_i} = \gamma_0 + \gamma_1 z_i$, show that $plim(\hat{\beta_1}^{IV}) = \frac{Cov(z_i, y_i)}{\gamma_1 Var(z_i)}$.

Classwork #15.2: Next show that: $plim(\hat{\beta_1}^{IV}) = \beta_1 + \frac{Cov(z_i, u_i)}{Cov(z_i, x_i)}$. It may be helpful to use the formula derived in the previous problem, the fact that $y_i = \beta_0 + \beta_1 x_i + u_i$ (as long as the exclusion restriction holds, y_i is not directly a function of z_i), and this formula for γ_1 from the first stage: $\gamma_1 = \frac{Cov(x_i, z_i)}{Var(z_i)}$.

Classwork #15.3: Using the formula derived above: $plim(\hat{\beta_1}^{IV}) = \beta_1 + \frac{Cov(z_i, u_i)}{Cov(z_i, x_i)}$, argue why the conditions for an instrument Z to be valid are the same as the conditions for $\hat{\beta_1}^{IV}$ to be consistent. We've used the exclusion restriction condition in the previous problem, but you should talk about the other two here.

9.5 IV Examples

9.5.1 Effect of Private School on Earnings using a Lottery Instrument

> Note: If you're also doing the extra credit workbook on DDC, you may be wondering at this point why, in the first stage, we're using a linear probability model when we could be using the much superior/internally consistent logit or probit. The answer, in short: econometricians have termed this approach the "forbidden regression" because it seems like a good idea, but it's not actually correct and can yield biased results. "The reason it doesn't work has to do with the fact that in a nonlinear regression, each variable's effect depends on the values of the other variables. So while the instrument Z and the second-stage error term u are unrelated, the fitted values $\hat{X}$ no longer get to borrow that nice unrelatedness from Z." – Huntington-Klein (2021)

9.5.2 Effect of Friends of the Opposite Sex on High School GPA

Hill (2015) *The Girl Next Door: The Effect of Opposite Gender Friends on High School Achievement.*

9.5.3 Effect of Military Service on Earnings

9.5.4 Effect of Meth on Foster Care Admissions

More details about this study here.

9.6 Bonus: IV estimates the LATE, not the ATE

In this bonus video, I work out a numerical example about the puzzle presented in the previous video: how IV estimates the causal effect for the complier subpopulation only: the *local* average treatment effect (LATE) instead of the overall average treatment effect (ATE).

This part is labeled "bonus" because I won't test you on this, it's just fun to explore :).

9.7 Exercises

Classwork 15: IV part 1

Classwork 16: IV part 2

9.8 References

Cunningham (2021) *Causal Inference: The Mixtape*

Hill (2015) *The Girl Next Door: The Effect of Opposite Gender Friends on High School Achievement.*

Huntington-Klein (2021) *The Effect: An Introduction to Research Design and Causality*

10 IV for Simultaneous Equations

10.1 Introduction

In your Principles of Micro or Macro classes, you probably learned about the "Marshallian Cross" model of the determination of market prices. The demand curve slopes down and the supply curve slopes up, and the equilibrium price and quantity exchanged are at the intersection of those two curves. So if you had data on prices and quantities exchanged and you ran a regression like this one: `lm(q ~ p)`, would you be estimating the supply curve or the demand curve? The answer is neither, because of *simultaneous equation bias*.

10.2 Biases We've Studied Thus Far

Simultaneous equation bias is something new, bus it's still based on a failure of exogeneity. Recall, these are the biases we've learned about thus far:

1) Omitted variable bias. Selection bias is a type of OVB where the omitted variable is someone's propensity to select in to treatment. OVB may bias an estimate up or down depending on the relationships between x and o, as well as y and o.

2) Measurement error. When X is measured with error, we'll estimate an effect of X on Y that is *attenuated* (closer to 0 than the true value).

3) Lagged dependent variable models: `lm(y ~ lag(y))`. When u_t is autocorrelated, we also get inconsistency for this model.

The root issue for all these cases is that exogeneity was violated: $E[u_i|X] \neq 0$. The same is true for the new source of bias for this chapter:

4) Simultaneity Bias: when data is generated by the interaction of two or more equations simultaneously.

To see how exogeneity would be violated in the market described in the video above, suppose we're trying to estimate the demand curve $q_i^d = \alpha_0 + \alpha_1 p_i + u_i$ where u_i refers to demand shifters like advertising. Exogeneity means:

$E[u_i|P] = 0$

That is, if you were told that in one location, the price was high, and in the other location, the price was low, *you would have no idea* which location may have had the advertising campaign. This is not a valid assumption, because advertising and price tend to be closely linked through many different channels: this is the simultaneity.

10.3 IV to the Rescue

In the previous chapter, we learned how an exogeneous instrument Z can help us estimate the causal effect of X on Y even when the data generating process suffers from OVB.

In the same way, an exogenous instrument Z can help us estimate the slope of the supply (or demand) curve even when the DGP is from two simultaneous equations.

10.4 Example 1: Market for Coffee

In this example, we'll use exogenous supply shocks w_i: **weather shocks in Brazil** as an instrument for the **price of coffee** to estimate the slope of the demand curve α_1. Basically we can see equilibrium P and Q for many supply curves but only one demand curve, and use that to find the slope of the demand curve.

$$q_i^d = \alpha_0 + \alpha_1 p_t + u_t$$

Where u_t are demand shifters like advertising campaigns or changes in the price of substitutes or complements.

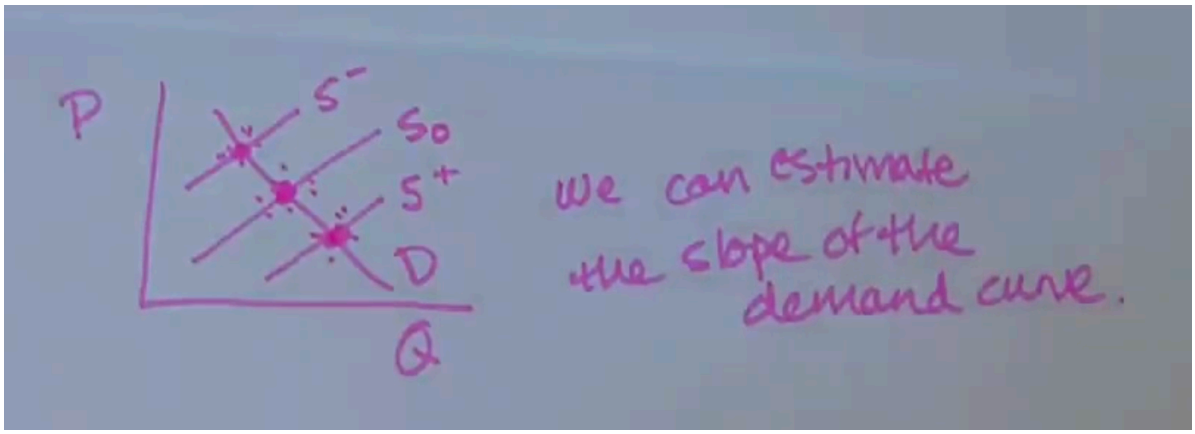
To check to make sure that **weather shocks in Brazil** is a valid instrument:

1) Is it relevant? Do weather shocks in Brazil effect the price of coffee? Yes.
2) Is it exogenous ($E[u_t|W] = 0$)? If you were told that in one time period, the weather was bad in Brazil, and in another, the weather was good. Does that give you any clue about which time period tea might have a higher or lower price? It shouldn't, so I'd say yes, W is exogenous.
3) Is it excludable? Or do weather shocks in Brazil effect the quantity demanded through any other channel than through changes in the price for coffee? W is excludable.

We'd then estimate the first stage, and get the fitted values: `fitted.values(lm(p ~ w))`. This is the exogenous variation in price that comes just from weather shocks in Brazil.

Then we run the second stage: `q ~ fitted.values(lm(p ~ w))` to see how exogenous supply shocks seem to effect the quantity exchanged. The data here are from the *supply* curve being

shifted up and down based on weather in Brazil, but the demand curve stays stationary. So in the second stage, we can trace out α_1 the slope of the demand curve.



10.5 Example 2: Market for Airline Tickets

10.6 References

Dougherty (2016) *Chapter 9: Simultaneous Equations Estimation*

11 Differences-in-differences

11.1 Introduction

In this final chapter, we'll learn about one more causal inference research design: the *diff-in-diff* estimator.

To quickly review, if you want to find how much some variable X effects another variable Y, running an RCT (randomized controlled trial) is the gold standard. The problem is that an RCT is often unethical and expensive, especially for many of the questions we're interested in researching as economists. For example, learning about the returns to education through an RCT would mean collecting a group of children and randomizing how much education they get to have, and then observing their earnings throughout their life.

An alternative research design is to use instrumental variables: if a valid instrument exists, a natural experiment happened. What's difficult about IV is that oftentimes, no valid instrument exists.

So in this final chapter, we'll discuss one more alternative: **diff-in-diff**. The special thing about this estimator is that we don't need strict exogeneity in order to get an unbiased estimator; we only need **panel data** and the **parallel trends assumption**.

11.2 Panel Data

Recall chapter 7 when we began studying time series. We learned that our focus for the first 6 chapters was on *cross-sectional* data, where we sampled many subjects at one time. For example, when you take a large group of people and ask them about their earnings, years of education, their sex, their race, and whether they are married, that builds a cross-sectional dataset.

Time series, on the other hand, observes one subject across time. So if you have data about the US (like gas prices in the US and the presidential approval rating) over a period of time, that's a time series.

The third and final type of data we'll discuss in this class is *panel data*, where you observe many subjects across time. We've actually been working with a panel dataset since the beginning

of the class: it's `gapminder`. The gapminder dataset has 142 subjects (countries) and makes observations about those countries in 12 different years.

11.3 Dr. John Snow

Dr. John Snow invented the diff-in-diff estimator in 1855 to prove that cholera spreads not through air, but through water. That was 85 years before RCT's were invented.

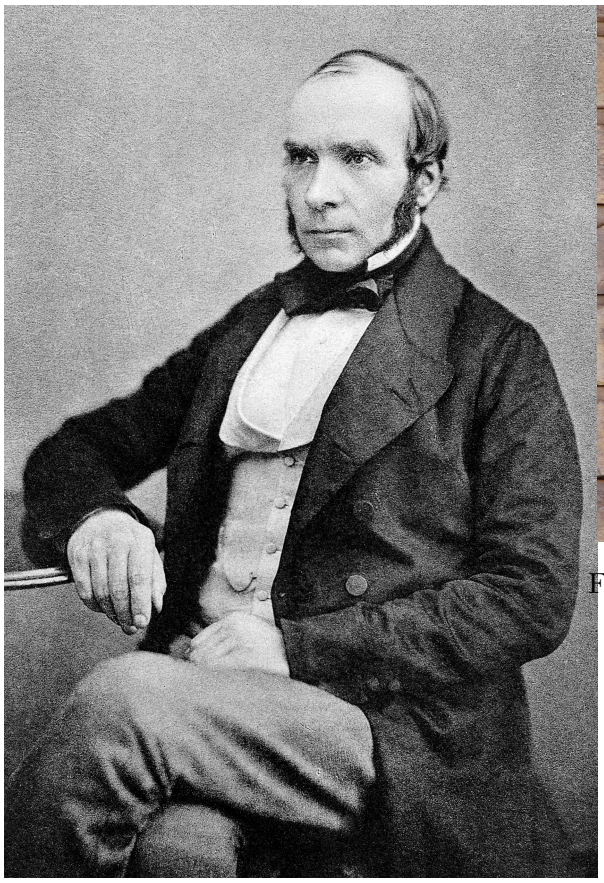


Figure 11.1: Originally from en.wikipedia; Public Domain, https://commons.wikimedia.org/w/index.php?curid=403227

Figure 11.2: By Tony Hisgett from Birmingham, UK - Dr John Snow, CC BY 2.0, https://commons.wikimedia.org/w/index.php?curid

11.4 References

Cunningham (2021) *Causal Inference: the Mixtape*

Part V

Summary and Resources

So there you have it: you're really starting to get the hang of the tidyverse, you probably feel a lot more confident with statistics, and if you pick up a research paper in applied economics, you'll be able to understand most of what they're talking about.

Where do you go from here? That depends on what you're interested in.

If you want to learn more about econometrics:

- I wrote a short workbook on Dynamic Discrete Choice in R
- The standard textbooks in grad school are these: Fumio Hayashi's Econometrics and William H. Greene's Econometric Analysis

If you want to learn more about the tidyverse:

- Take a look at the ggplot2 book: understanding this tool deeply is marketable and fun.
- But mostly, find a data project you're interested in and do it! Practice makes perfect. Just make sure you stay aligned with the principles of the tidyverse that you learned in this course.

If you want to become a better programmer in general: this is a great goal to have, even if you don't plan to become a software engineer. Computers are a major part of our lives, so learning to use them more effectively really pays off. Some people mistakenly seem to believe that becoming a better programmer means learning more programming languages, but this is not true. To be a good programmer, you just need to know how to do a good job solving the problems you're interested in. That means you should understand the nuances of the languages you love and their strengths and weaknesses. But mostly, it means figuring out what problems you're interested in and going out and doing them.

Classwork

CW1

Deriving OLS Estimators (analytical)

CW2

lm and qplot (R)

Download the R Script here

CW3

dplyr murder mystery (R)

Download the R Script here

CW4

hypothesis testing (analytical)

CW5

causal inference (analytical)

CW6

causal inference (R)

Download the R Script here

CW7

consistency (analytical)

CW8

heteroskedasticity (analytical)

Practice Midterm

CW9

heteroskedasticity (R)

Download the R Script here

CW10

simulation (R)

Download the R Script here

CW11

dynamics (analytical)

CW12

dynamics (R)

Download the R Script here

CW13

time trends (analytical)

CW14

random walks (half analytical, half R)

CW15

IV (analtyical)

CW16

IV (R)

Download the R Script here

CW17

Practice Final (analytical)

Koans

Here's a preview of the 20 koans developed for this class. To download them and the corresponding tests, follow the instructions in the preface to this workbook.

vectors, tibbles, and pipes

dplyr

ggplot2

lm() and statistical distributions

functions

map()

lags and first differences

reduce and accumulate

References

Bryan (n.d.)

Hadley Wickham and Grolemund (2017)

H. Wickham (2014)

Speegle and Clair (2021)

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

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

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:

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

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? $Var(X) \equiv E[(X - E[X])^2]$

Population Covariance

$Cov(X, Y) \equiv E[(X - E[X])(Y - E[Y])]$

Correlation

Correlation is a function of covariance:

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

If two random variables have 0 covariance, they will have 0 correlation.

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: $Cov(X, Y + Z) = Cov(X, Y) + Cov(X, Z)$

plim rules

Let c be a constant. Let x_n and y_n be sequences of random variables where $plim(x_n) = x$ and $plim(y_n) = y$. That is, for large x, the probability density function of x_n collapses to a spike on the value x and the same for y_n and y. Then:

1) The probability limit of a constant is the constant: $plim(c) = c$
2) $plim(x_n + y_n) = x + y$
3) $plim(x_n y_n) = xy$
4) $plim(\frac{x_n}{y_n}) = \frac{x}{y}$
5) $plim(g(x_n, y_n)) = g(x, y)$ for any function g.

Expectations

Let A and B be random variables, and let c be a constant.

1) $E[A + B] = E[A] + E[B]$
2) In general, $E[AB] \neq E[A]E[B]$
3) Constants can pass outside of an expectation: $E[cA] = cE[A]$

And continuing from 3), since $E[A]$ is a constant, $E[B\ E[A]] = E[A]E[B]$.

Conditional Expectations

If the conditional expectation of something is a constant, then the unconditional expectation is that same constant:

If $E[A|B] = c$, then $E[A] = c$.

Why? The **law of iterated expectations:**

$$\begin{aligned} E[A] &= E\left[E[A|B]\right] \\ &= E[c] \\ &= c \end{aligned}$$

Log rules

1. $log_e(e) = 1$
2. $log(ab) = log(a) + log(b)$
3. $log(\frac{a}{b}) = log(a) - log(b)$
4. $log(a^b) = b\ log(a)$

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.

Angrist, Joshua, and Jorn-Steffen Pischke. 2010. "A Note on Bias in Conventional Standard Errors Under Heteroskedasticity." https://econ.lse.ac.uk/staff/spischke/mhe/josh/Notes%20on%20conv%20std%20error.pdf.

Bryan, Jennifer. n.d. *Gapminder: Data from Gapminder*.

Cunningham, Scott. 2021. *The Mixtape*. New Haven: Yale University Press. https://doi.org/doi:10.12987/9780300255881.

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

Hill, Andrew J. 2015. "The Girl Next Door: The Effect of Opposite Gender Friends on High School Achievement." *American Economic Journal: Applied Economics* 7 (3): 147–77. https://doi.org/10.1257/app.20140030.

Huntington-Klein, Nick. 2021. *The Effect: An Introduction to Research Design and Causality*. 1st ed. Chapman; Hall/CRC. https://doi.org/10.1201/9781003226055.

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.

Rubin, Ed. 2022. "Economics 421: Introduction to Econometrics." Github. https://github.com/edrubin/EC421W22.

Speegle, Darrin, and Bryan Clair. 2021. "Data for the Dplyr Murder Mystery." https://rdrr.io/github/speegled/dplyrmurdermystery/.

Wickham, H. 2014. *Advanced r*. Chapman & Hall/CRC the r Series. Taylor & Francis. https://adv-r.hadley.nz/.

Wickham, Hadley, and Garrett Grolemund. 2017. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st ed. O'Reilly Media, Inc. https://r4ds.had.co.nz/.