STATISTICAL HORIZONS

Econometrics: A Causal Approach

Nick Huntington-Klein, Ph.D.

Upcoming Seminar:

January 20-22, 2022, Remote Seminar

Day 1: Econometrics, a Causal Approach

Estimating an Effect

Welcome!

- Welcome to Econometrics, a Causal Approach
- We are interested in questions about how *one thing causes another*
- This will require us to have good *research design* and also apply the proper *estimation techniques*
- The estimation technique makes sure we're doing a good job estimating
- The research design tells us that the thing we're doing a good job estimating actually answers our causal question

The Layout

- Day 1: Working with regression
- Day 2: Identification and research design
- Day 3: Common "back-door" research designs
- Day 4: Common "front-door" research designs

Approach

- We'll hit conceptual understanding first and foremost
- With important technical details coming up as necessary
- And plenty of implementation details for common research designs
- And coding practice as we go through, primarily in R, but alternative Stata and Python materials are also available
- For more technical depth, try my book The Effect or a book like Wooldridge's Econometric Analysis of Cross-Section and Panel Data or Bailey's Real Econometrics

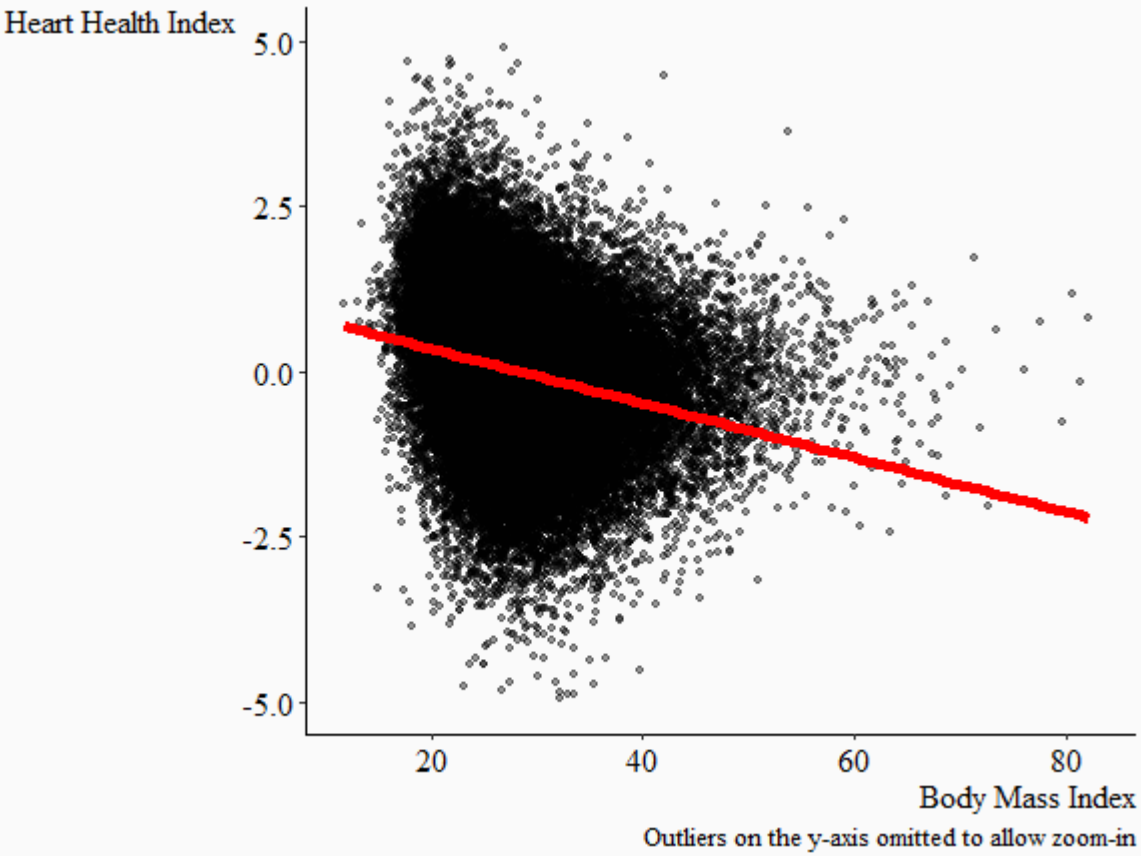
Starting Place

- I'll assume you have some familiarity with your coding language
- And also some basic statistical background
- If you already work with regression a lot, Day 1 may be a bit slow for you (although I will be presenting some aspects in a way that's likely new to you), but Day 2 onwards will have plenty of new stuff

With that in mind, let's get started!

- Today we'll be focusing on sharpening our understanding of regression analysis
- Regression is the most common tool used by econometricians, especially when working with causal effects
- Regression is a tool for estimating the *relationship between variables*.
- It does this by producing a "line of best fit"
- (*matching* is another common approach to getting causal effects, but less popular in econometrics. We'll largely be skipping it in this class)

A Basic Regression



Initial Questions

1. Why would we want this?
2. What is it actually doing?
3. How can we interpret it?

Why Would We Want This?

- We need a way of *describing the relationship between two variables*
- This will be key especially for causal analysis, since our goal is to say something like "an increase in X will (do something) to Y"
- There are plenty of ways we could describe a relationship, though. Why use a best-fit line?
- Let's explore the idea of relationships between variables generally, and see what other options we have

Relationships

- Two variables X and Y are *independent* if learning the value of one tells you nothing about the other
- For example, knowing the outcome of a roulette spin tells you nothing about what the next spin will be
- They are *dependent* if learning the value of one *changes the distribution* of the other
- For example, among all Americans, about 50.8\% are legally female and about 49.2\% are legally male
- But if we *learn that someone's name is Susan*, that distribution will change considerably in favor of female!

Conditional Values

- Another way of saying we *learn the value of one variable* is to say we *condition* on the value of that variable
- "Conditional on someone's name being Susan, the distribution of legal sex is 96% female and 4% male"
- P(Gender | Name = "Susan")
- Standard stuff about probabilities and probability distributions applies here
- It just applies only to *a portion of the data* (the Susans!)

Conditional Means

- In many fields, we are very interested in calculating *conditional means*
- i.e. *what is the mean of the conditional distribution?*
- And we want to know this not only for *one* value of the variable we're conditioning on, but *all* the values
- The *population* conditional mean is the *conditional expectation*, or $E(Y|X)$

Conditional Means/Expectations

- Talking about how two variables are related is just another way of talking about conditional values (usually, conditional means)
- If X and Y are positively related, that means "at higher values of X, the conditional expectation $E(Y|X)$ is higher"
- If they're negatively related, "at higher values of X, $E(Y|X)$ is lower"
- If their relationship is U-shaped (or similar), "at higher values of X, $E(Y|X)$ changes but not always in the same direction"

"Explaining"

- We can also say that $Y|X$ is "the part of Y that is *explained by* X"
- If $E(CoffeeCupsPerDay|\ Occupation = Professor) = 1.79$, and I drink 3 cups per day, then 1.79 of my cups are "explained by" the fact that I'm a professor, and $3 - 1.79 = 1.21$ of my cups are "not explained by" being a professor
- This can extend to multiple variables!
- If $E(CoffeeCupsPerDay|\ Occupation) = Professor$, $Gender = Male) = 2.13$, then 2.13 of my cups are "explained by" occupation and age, and $3 - 2.13 = .83$ of my cups are "not explained by" those two things
- This is a purely statistical explanation
- This doesn't necessarily mean that 2.13 of my cups are *because* I'm a professor and a man, but rather that 2.13 of my cups are *what would be expected* given what you know about me

Demonstrating Conditional Means

- When X only takes a few values, we can just show the distrbution of Y conditional on each of the values of X, and compare them to each other. Any distribution-showing method works - density plots, bar graphs...
- Scatterplots show all the data and imply the conditional mean
- For continuous X data, you can calculate the conditional mean $Y|X$ over different *ranges* of X, either splitting X into bins, or doing "local means"
- Regression

Example

- We'll be using data from Emily Oster's study of the relationship between taking Vitamin E and health outcomes
- (and whether that relationship changes as a result of Vitamin E being briefly recommended, then not!)
- We'll start with versions that can be used for *discrete* X values, like "smoking vs. non-smoking"
- Remember, the proportion of a binary variable *is* its mean