Econ 210 - Introduction

Sidharth Sah[1]

September 8, 2023

[1] Thanks to Thomas Wiemann for useful material

Goal of Econometrics

- Econometrics can be broadly defined as the study of how to empirically answer economic questions
- Modern econometrics is largely geared towards how to answer *causal* questions about the world

Descriptive vs Causal Questions

- Descriptive questions are questions about how the world *is* - the realized state
  - What is the unemployment rate? How many people are watching House of the Dragon?
- Causal questions are questions about the world *would be* under different conditions - potential states
  - What *would* the unemployment rate be *if* there was a higher national minimum wage? How many people *would be* watching House of the Dragon *if* there was no Rings of Power?

Answering Causal Questions

- Answering causal questions necessitates three (generally) distinct steps:
  - Defining the causal parameter of interest
  - Determining how to identify the parameter from a hypothetically infinite amount of data
  - Determining how to infer the parameter from a finite sample of actual data

Defining the Causal Parameter

- Can represent any outcome of interest as:

$$Y_i(s, u)$$

where Y is the outcome, i is an individual unit of observation, $s \in S$ is a state of the world, and u is every other characteristic of the individual that enters into the outcome

-

$$Y_i(s', u) - Y_i(s, u)$$

gives the individual-level treatment effect of s' compared to s - the difference in outcome for i between two potential states of the world, holding all else equal

Defining the Causal Parameter

- Could have Y_i be employment status, s represent current minimum wage policies, and s' be a state of the world with a \$15 national minimum wage. The previous slide would then show the causal impact of a minimum wage policy change on individual i, holding everything else equal
- While we likely won't care that much about any individual i, we might care about the average over the population:

$$E[Y(s', U) - Y(s, U)]$$

which would be the average change in employment status from the current world to the \$15 minimum wage world

Fundamental Problem of Causal Inference

- We can only ever observe a given unit of observation under a single state at a given time - one value $Y_i(s, u)$ but not the other $\{Y_i(s', u) | s' \neq s\}$
- Because of this we never actually observe $Y_i(s', u) - Y_i(s, u)$ in the data. This is called the Fundamental Problem of Causal Inference
- Attempting to deal with this is the second step of answering a causal question - identifying the parameter of interest

Identification

- Identification proceeds by making certain assumptions about the data-generating process (the nature of the $Y_i(s, u)$'s) and demonstrating mathematically that we can solve for the parameter of interest from the type of data we have, if we had infinite data (infinite number of observations)
- At this stage of the process, we are only worried about what *type* of data we see, not *how much* data we see. For instance, we might only see certain characteristics and outcomes, which we take into consideration, but not how many people we see those characteristics and outcomes for

Inference

- Finally, we need to acknowledge that we never actually have an infinite amount of data. Because of this, we experience statistical uncertainty
- Inference refers to what conclusions we can reach from the finite samples of data we have and how certain we can be about those conclusions

Example - Setup

- Lets demonstrate these three steps with an example (just to see what the three steps are - no need to worry too much about the details yet, that's what the rest of the course is about!)
- Say we are interested in the effect of a particular job training program on wages

Example - Defining the Parameter

- First we have to define the parameter of interest. Say the program will take place over the next few months, concluding by 2023. Then we might care about

$$E[Y(s', U) - Y(s, U)]$$

  where Y is income in 2023, s' indicates having done the job training, s indicates not having done the training, and U is a random variable representing all other characteristics that help determine wages
- This is the average effect of job training on 2023 income, averaged over the whole population, holding all else equal

Example - Defining the Parameter

- Notice we've already made several choices that could be questioned. Why do we care about 2023 wages instead of longer term outcomes? We also considered average effect over everyone - should we focus on low-wage workers who might be the most inclined to take up such a program and who we might be most interested in trying to help?

Example - Identification

- We can break up our parameter of interest:

$$E[Y(s', U) - Y(s, U)] = E[Y(s', U)] - E[Y(s, U)]$$

- However, due to the Fundamental Problem, we don't see either of the addends above. We do see $E[Y(s', U)|S = s']$ and $E[Y(s, U)|S = s]$, the mean salary of people who did the job training among people who did the job training and the mean salary of people who didn't do the job training among people who didn't do the job training

Example - Identification

- In order to proceed we'll have to make an assumption about the nature of the data. Let's assume $S \perp U$, meaning training program uptake is independent of everything else. In the regular world this would be a terrible assumption - people would opt in and out of the program for reasons! If we did an experiment, putting people in or out of the program randomly, it might make sense

Example - Identification

- With that assumption in place, and further assuming for the same of example that $U \in (-\infty, \infty)$ we can say:

$$\begin{aligned} E[Y(s', U)|S = s'] &= \int_{-\infty}^{\infty} Y(s', u) f_{U|S}(u|s') du \\ &= \int_{-\infty}^{\infty} Y(s', u) f_U(u) du \quad (S \perp U) \\ &= E[Y(s', U)] \end{aligned}$$

- We can similarly show $E[Y(s, U)|S = s] = E[Y(s, U)]$

Example - Identification

- Thus we can show that our parameter of interest is completely a function of things we can observe in the data:

$$E[Y(s', U) - Y(s, U)] = E[Y(s', U)|S = s'] - E[Y(s, U)|S = s]$$

- All that remains is to estimate the two expectations on the right-hand side from data - inference

Example - Inference

- $E[Y(s,U)|S=s']$ and $E[Y(s,U)|S=s]$ are *population*-level objects - they are scalars defined by an entire probability distribution - we can't see them directly and we couldn't calculate them from infinite observations, even if we had those!
- Lets say, however, that we see the 2023 incomes of a random sample of n people

Example - Inference

- Then we can use estimators of those population-level parameters. In this case, those estimators might be simple averages of the values from our sample
- Define the notation that $W = 1\{S = s'\}$ - so W is an *indicator function* that the individual has done the training - this will be 1 for people who did it and 0 for people who didn't. So, we'll see (y_i, w_i) for the n people in our sample

Example - Inference

- Our estimators can then be

$$\hat{E}[Y(s', U)|S = s'] = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} y_i w_i$$

$$\hat{E}[Y(s, U)|S = s] = \frac{1}{\sum_{i=1}^{n} 1 - w_i} \sum_{i=1}^{n} y_i (1 - w_i)$$

- From those we can form an estimator of our parameter of interest:

$$\hat{T} = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} y_i w_i - \frac{1}{\sum_{i=1}^{n} 1 - w_i} \sum_{i=1}^{n} y_i (1 - w_i)$$

Example - Inference

- Notice that all of our estimators are functions of our random sample of n people, so they themselves are random variables! Later in the course we'll figure out the distributions of those kinds of random variables, so that we can form notions of confidence in our estimates - things like standard errors and confidence intervals

Example - Conclusion

- With that we've gone from a causal question to an answer, with some sense of how confident we are in our answer. Along the way, we had to employ assumptions about the data-generating process (hopefully based in economic theory), probability theory, and statistics. These are the tools we will further develop for the rest of the course