Dealing with Endogeneity

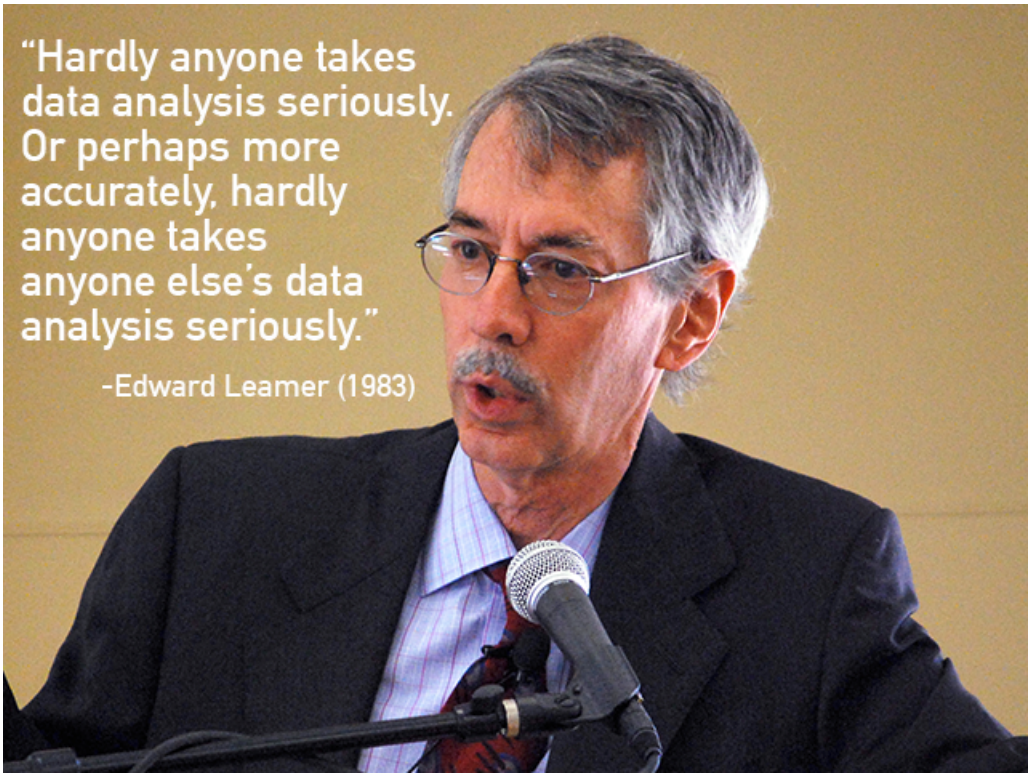
- Omitted variable bias is a major obstacle for most data-driven research.
  - We usually care about causality.
  - We want to change things in the world, make decisions to improve an outcome.
- Social researchers, in particular, have to deal with omitted variable bias.
  - If you study physics, it might be that every photon is the same as every other photon.
  - But humans have unobserved attributes that affect their behavior.

Dealing with Endogeneity (cont.)

- Surprisingly, our ability to deal with endogeneity bias in the social sciences is actually quite recent.
- Part of the problem was computational limits. For a long time, we didn't have the processing power to run the more advanced techniques we use today.
  - A naïve regression was usually all you needed to publish a paper. (This is still true in many fields.)
- Researchers knew about endogeneity, but it wasn't regarded as a huge problem.
  - A touch of willful blindness?

The Early 1980s Crisis

- As computing power increased, many fancy techniques were developed through the mid-1900s.
  - They often rested on heroic assumptions.
- The proliferation of naïve regressions and untenable assumptions led to something of a crisis in the early 1980s.



The Identification Revolution

- These days, a lot of social sciences take endogeneity bias much more seriously.
- Microeconomics is a leading example.
  - In the mid-1990s, economists embraced the need for "well-identified" econometrics whenever interpreting something causally.
  - "Well-identified" is a term from simultaneous equation modeling.
  - Here, "identify" means that we can consistently estimate a coefficient in a causal model.
- An **identification strategy** is our plan for consistently measuring a causal effect in the face of endogeneity.

What Are Identification Strategies? (Part One)

1. A true experiment

- Key feature: We randomize the treatment variable x.
  - Otherwise, it's not an experiment.
  - E.g., we flip a coin.
  - A randomized RV is independent of every other RV we measure.
- This implies $\text{cov}(x,u) = 0$.

What Are Identification Strategies? (Part Two)

- True experiments are regarded as the gold standard for causal inference.
- Unfortunately, we can't always run a true experiment.
    - It could be expensive.
    - It could be infeasible.
        - E.g., Can you randomly change governments to see what the effect of democracy is?
    - It could be unethical.
        - E.g., Can you make some people smoke and some people not smoke to test the effects of smoking?

What Are Identification Strategies? (Part Three)

There are some purists who will say that experiments are the only way to determine a causal effect.

- More researchers would disagree with this.
- Experiments are the gold standard, but we have other identification strategies that are important, especially when experiments are impossible:

- Difference in difference
- Instrumental variables
- Regression discontinuity

Each of these is a huge topic. I'm just going to tell you enough to know that they exist.

Difference in Difference

- A good example of a difference-in-difference design:
  - Card, David, and Alan B. Krueger. *Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania*. 1993.
- On April 1, 1992, New Jersey's minimum wage increased from $4.24 to $5.05.
  - Card and Krueger wanted to understand the effect on employment.
- The problem was that the number of jobs could have been increasing/decreasing anyway.
  - We can't directly compare two time periods.

Difference in Difference: Data

	Stores by state		
Variable	PA (i)	NJ (ii)	Difference, NJ − PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	−2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	−0.14 (1.07)
3. Change in mean FTE employment	−2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Difference in Difference (cont.)

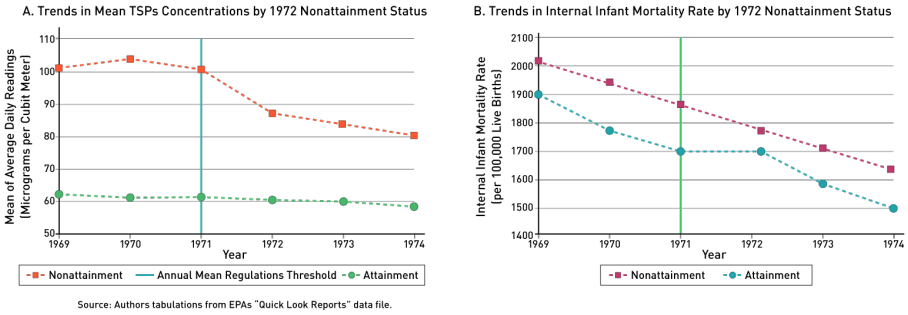
- Another idea would be to compare employment in New Jersey with neighboring Pennsylvania, which is similar in some ways.
    - But the states have different characteristics.
- Instead, one can look at how much employment changed in New Jersey, and compare it to how much it changed in Pennsylvania.
    - As long that the characteristics of each state don’t change in time, the difference in the differences can be attributed to the policy chance.
    - Relative to stores in Pennsylvania, stores in new Jersey increased employment by 13%.
- This is a difference-in-difference design. It's a tool that allows us to remove the effects of time-constant confounders.

Instrumental Variables

- Instrumental variables allow us to study situations in which we can’t randomly assign treatment, but we can identify a source of variation that we believe is exogenous and affects the treatment—that's called an instrumental variable.
- This means that part of the treatment is exogenous.

Instrumental Variables: Example

Chay and Greenstone, 2003 "Air Quality, Infant Mortality, and the Clean Air Act of 1970"



Source: Authors tabulations from EPAs "Quick Look Reports" data file.

- This study leveraged the Clean Air Act of 1970, which created a set of national standards.
- The change in law affected some states, but not others, depending on previous regulations there.
- This is a good instrumental variable—whether a given state is affected by the Clean Air Act seems functionally random, so we can interpret the effects as causal.
- "We estimate that a one percent decline in TSPs results in a 0.5 percent decline in the infant mortality rate."
- Instrumental variables are a workhorse technique, especially in labor economics, but they are finding uses in may other areas.

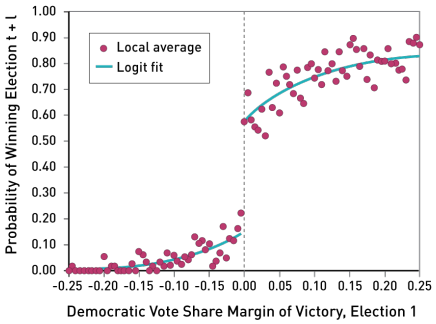
Regression Discontinuity

Idea: Exploit a decision rule that determines what individuals receive a treatment based on an attribute.

E.g., Lee, David S. "Randomized experiments from non-random selection in US House elections." *Journal of Econometrics* 142.2 (2008): 675-697

Lee is interested in whether a party that holds a seat in Congress has an advantage in the next election.

- Being an incumbent is endogenous—there's a reason the incumbent won before, maybe because voters like them.
- But if we look at the previous election, we might think that a party that wins 49% of the vote, is pretty similar to one that wins 51% of the vote. One wins and one loses.
- So we can compare candidates that are on either side of the discontinuity.



Conclusion

- All of these identification strategies have strengths and weaknesses.
- One of the best places to learn more is *Mostly Harmless Econometrics* by Angrist and Pishke.