Applied Econometrics: A Crash Course

Prof. Jetson Leder-Luis

May, 2024

You can download this PDF here to follow along:

`https://tinyurl.com/mentoring-metrics`

Please ask questions! This is a conversation, not just a lecture.

1 Motivation

- You need to be able to do **great** econometrics to write influential papers
- Without good metrics, your story is not going to be credible, your interpretation is likely to be wrong, and you may have a bad time sending your research to journals or presenting it at talks
- Economics is different from public health because we take causal inference seriously

  By "causal inference," I mean the question "does X cause Y"
- Causal inference is necessary for policy work.

  If X causes Y, then changing X will change Y

  If X and Y are just correlated, changing X will not necessarily change Y

  Policy questions center on "how should we change X" and without a causal understanding, you get ineffective policy
- The reason that economics has an outsized impact on policy in many fields – education, health, labor – is that we can answer causal questions

1.1 Intellectual History

- Much of today's lecture is sourced from Josh Angrist's work
- I TA'd Angrist's 14.387 applied metrics seminar, and we wrote a paper together in grad school
- I **highly recommend** you buy and read Mostly Harmless Econometrics if you have not yet done so

2 Selection Bias

2.1 Motivating Example

- Important to distinguish between correlation and causation
- Let's talk about this recent study

> JAMA Netw Open. 2022 Aug 1;5(8):e2228510. doi: 10.1001/jamanetworkopen.2022.28510.

Association of Leisure Time Physical Activity Types and Risks of All-Cause, Cardiovascular, and Cancer Mortality Among Older Adults

Eleanor L Watts [1], Charles E Matthews [1], Joshua R Freeman [1], Jessica S Gorzelitz [1], Hyokyoung G Hong [1], Linda M Liao [1], Kathleen M McClain [1], Pedro F Saint-Maurice [1], Eric J Shiroma [2], Steven C Moore [1]

Affiliations + expand

PMID: 36001316 PMCID: PMC9403775 DOI: 10.1001/jamanetworkopen.2022.28510

Is this analysis **causal**? By which I mean, does this tell us that increasing exercise will reduce mortality?

To discuss with the class

2.2 Selection Bias and Regression

- **Selection bias**, which also gets called **omitted variable bias**, occurs when those who are different in the X variable are also different along other variables
- Consider the following regression:

  $$Sunburn_i = a + b * IceCream_i$$

  This regression is run with data at a person-day level. Sunburn is an indicator for whether they got a sunburn that day, and ice cream is an indicator for whether they ate ice cream that day.

  Suppose b is a positive, statistically significant coefficient. Can we conclude that ice cream causes sunburns?

  Of course not! Ice cream and sunburns are **correlated**, but ice cream doesn't cause sunburns. Eating ice cream tells us something about the weather – it was hotter out – and we know the sun causes sunburns.

- Consider the following regression:

  $$Death60_i = a + b * SleepDep_i$$

  This regression is run at an individual level across all people in the US. $Death60$ is an indicator variable if the individual died by age 60, and SleepDep is an indicator for whether the individual regularly gets less than 6 hours of sleep.

  Suppose b is a positive, statistically significant coefficient. Can we conclude from this regression that sleep deprivation causes early death?

  The answer is **no**, for the exact same reason as the above .

  Sleep deprivation is not a random condition. Maybe individuals who don't sleep are more likely to be doing dangerous drugs, or maybe they work more dangerous jobs, or maybe they aren't sleeping because they have chronic illnesses

- Regressions always have a correlational interpretation, but they do not have a causal interpretation unless X is randomly or quasi-randomly assigned

3 Experiments Solve Selection Bias

- Experiments solve selection bias
- Why? Because if treatment is randomly assigned, then nothing is correlated with it
- And therefore, there can be nothing else that is causing Y when X changes except X itself
- Sometimes, you can run experiments, (see e.g. the work on development aid)
- Experiments are rare in aging & health because you cannot, e.g., randomize who smokes
- There is some great experimental work on health insurance, but mostly you can't plan for this to be your approach

4 Applied Econometrics

4.1 Major Methods

- Most of applied econometrics is about solving selection bias
- We want **quasi-experimental variation** which means that treatment is **as good as randomly assigned**
- Generally, there are only a few methods used by economists, and we will cover them

  1. Difference-in-difference
  2. Instrumental variables
  3. Regression discontinuity

  There are a few advanced methods we will also get to at the end

4.2 Apply the Methods

- You should probably not try to invent something new with your job market paper
- Trust me, I did a new version of synthetic controls for my job market paper. (Jetson will tell a story about his flyout at Wharton)
- Your goal is to look for data and quasi-experimental variation that will allow you to answer causal questions in a credible way

5 Difference in Difference

5.1 Setup

- Difference in difference occurs when you have treated and control units, before vs after the treatment
- In the basic model, you just compare Before vs After, Treated vs Control

  E.g. Card and Krueger 1994 uses this to show that the rise in minimum wage had no effect on employment. It's just before vs after, New Jersey vs Pennsylvania
- The **identifying assumption** is called the parallel trends assumption – that the treated and untreated units were on parallel paths, (the levels don't have to be the same, but the trends do), in the pre-period, and the assumption that this would have continued without the treatment
- Nobody wants to see 2-period diff-in-diff before vs after, because you can't assess **pretrends**. Instead, the field has moved toward staggered difference-in-difference – many units treated at different times, and some untreated units.
- In staggered diff-in-diff, the not-yet-treated units also act as controls for the treated units at the time of treatment

5.2 How to Code

- We set up our data as a panel – repeated observations of the treated and untreated units. You generally want a complete panel – one observation per unit per time period.
- Then, we run the following regressions:

  First, we run the **Static TWFE**

  $$Y_{it} = \alpha + \beta D_{it} + \alpha_i + \alpha_t + \epsilon_{it}$$

  Here, Y_{it} is the outcome variable, D_{it} is an indicator for being in the post-treatment period, α_i are dummies for the unit, and α_t are dummies for the period

  We call this **two-way fixed effects** because we have both unit and time fixed effects. A **fixed effect** is just an indicator variable for belonging to that group, and it takes out the mean of that group
- This regression gives a single estimate, β, which is the treatment effect
- We also run the **Dynamic difference-in-difference**, which does the same thing but year by year

- When you do this year by year, for a few years before vs after the event, then you can check for pre-trends! This is critical and you must do it.
- We estimate:

$$Y_{it} = \sum_{j=-K, j \neq -i}^{L} \beta_j T_{ij} + \alpha_i + \alpha_t + \epsilon_{it}$$

  Here, the α_i and α_t are still the unit and time fixed effects. Instead of using D_{it}, and indicator for being in the post period, we instead have indicators of being in years $-K$ through $+L$, excluding year -1, where T_{ij} and indicators for each **relative year j**
- The numbers we care about are the β_j, which we can plot as an **event study figure**, which shows us the effect for each year before and after each treatment. The year -1 is normalized to 0.
- We include all units, even untreated units. Untreated units are set to have all years equal to relative time -1. Why? First, since -1 is our omitted level, we aren't really estimating anything for them. Then why include them? Because they help estimate the year fixed effects for all units, i.e. to help understand any time trends that are not due to the treatment.

5.3 Caveats

- Binning vs trimming – in general, you can't use an infinitely long panel. Instead, we want to do, e.g. plus or minus 24 months on either side, or plus or minus 5 years if you have a yearly panel.

  You have two choices in how to handle this:
  - In **trimming**, you drop all observations outside of some window – e.g. relative years +/- 5. Because all untreated units are year -1, you keep them.
  - In **binning**, instead you code all of the years outside the window as the endpoints; for example, year -6, -7, -8 all get tagged as year -5.
  - In general, there is debate about how to do this, but trimming is safe
- Modern TWFE: a number of studies have shown that regular diff-in-diff has some problems with how things get weighted. In general, it's not a good idea to let the already-treated be the control units.
  - The math here gets super complicated
  - In practice, you want to re-do your diff-in-diff with a package like Sun and Abraham or Callaway and Sant'Anna that handles these problems

5.4 Diff-in-Diff Example – Ambulance Taxis

Wherein students review the methods section of Ambulance Taxis: The Impact of Litigation and Regulation on Health Care Fraud, Eliason, League, Leder-Luis, McDevitt and Roberts, 2021 NBER WP

Topics to cover

- Context of the study
- Diff-in-diff setup
- Event study figures
- Robustness using modern TWFE

6 Instrumental Variables

6.1 Setup

- Remember, the main problem from a regression of Y on some treatment X is that treatment is not randomly assigned, so the regression is not causal
- With instrumental variables, we look for some 3rd variable, Z, that shift the values of X
- We conduct what's called **two-stage least squares**.

  Some people think about this as regressing X on Z, then regressing Y on the fitted values.
- It can be shown that this is equivalent to the following two-step regression:

  The **first stage** regression

  $$X_i = a_1 + b_1 Z_i + e_i$$

  The **reduced form** regression

  $$Y_i = a_2 + b_2 Z_i + e_j$$

  Then what we care about is

  $$\text{Treatment Effect} = \frac{b_2}{b_1} = \frac{\text{Reduced Form Coefficient}}{\text{First Stage Coefficient}}$$

  We look at the effect of Z on Y, blown up by the effect of Z on X
- This requires a set of assumptions and tests
  - **Independence**: your instrument must be as good as randomly assigned, i.e. it is not correlated with the potential outcomes (in the state of the world where treatment hadn't happened)
  - **Relevance**: Z affects X. We look at the **F statistic of the first stage** to check that this is true. If Z doesn't affect X, we have what's called a **weak instruments** problem. Rule of thumb: $F > 10$
  - **Exclusion restriction**: Z only affects Y through X.
  - **Monotonicity**: the effect of Z on X always points in the same direction. I.e. nobody is less likely to be treated because the instrument made some people more likely to be treated.
- What does this actually identify?

  IV identifies the **LATE of the Compliers**. LATE means local average treatment effect. That means it is the effect of the treatment of X on Y, on average, of the population of people who are induced into treatment because of the instrument.

6.2 How to Code

Do not do least squares manually. Use a package. Doing the 2SLS by hand often gets the standard errors wrong. Use ivregress or another equivalent package.

6.3 Caveats

- You need to check for weak instruments. If the first stage is too weak, then you are dividing by a small number, which makes your estimates too big.
- Unless you really know what you're doing, both the first stage and the reduced form should be linear regressions. Nonlinearity has some other issues – don't make it up, follow good papers/see Woolridge, 2015, JHR, "Control Function Methods in Applied Econometrics"
- Often we include controls in these regressions. Any controls in the first stage need to be in both the first stage and the reduced form. It is acceptable if your instrument is only random *conditional* on controls.
- The exclusion restriction is not truly testable, so be prepared to argue. You need to think deeply about the exclusion restriction, and make arguments.
- Sometimes people want different tests, like balance tests to show the instrument is plausibly random

6.4 Extensions

- You can characterize the set of compliers – the people who take up treatment. See: Mostly Harmless Economics, "Abadie Kappa". In short, it's about the first stage among different subgroups
- Mountjoy (2022, AER) shows how to use multiple instruments to show effects along compliers from different treatment margins. Advanced, beyond the scope of today. Gruber et al (example below) implement this

6.5 Example – Dying or Lying?

Wherein students look at an IV example from: Dying or Lying? For-Profit Hospice and End of Life Care, Gruber, Howard, Leder-Luis and Caputi, NBER WP 2023

Topics to cover

- Context of the study
- IV study
- IV tables
- Interpreting LATE and complires

7 Regression Discontinuity

7.1 Setup

- Regression discontinuity is used when the world presents a **threshold** in some variable, above which treatment applies

  We call that variable the **running variable**

- The identifying assumption is that the threshold is somewhat random, so people/units just above and just below the threshold are similar
- We compare the treated units above to the untreated units below to understand the causal effect of the treatment
- There are two version of RD: sharp and fuzzy
- **Sharp RD** applies when individuals just above the threshold are always treated, and those below are never treated
- **Fuzzy RD** applies when going above the threshold drastically increases your probability of treatment.

7.2 How to Code

Generally, there are good packages for this.

RDRobust is available in Stata, Python and R

Note that if you have fuzzy RD, this is equivalent to an instrumental variables problem! Being above the threshold is an **instrument** for being treated, and we can run this as an IV if you prefer. It gives the same results.

7.3 Caveats

- **Bandwidth selection** is an ongoing question. How far above/below the threshold do you want to use data from? There are some papers on optimal bandwidth. Packages given an optimal bandwidth as well.

  You can either choose a reasonable bandwidth, use the optimal bandwidth, or show your results stay the same no matter the bandwidth

- You **must** make the plot of the treatment variable vs the running variable to show that there is a jump at the discontinuity
- **Running variable manipulation** is a concern – if there is any reason that someone can choose whether a unit is above/below the threshold, then your treatment assignment is no longer random. This is an issue because this nonrandomness may be correlated with other factors

  To make sure this isn't a case, it's standard to do the **McCrary Test**, where you plot the histogram of the running variable, and show no bunching/discontinuities in the running variable left/right of the threshold. There are packages that will turn this into a p-value.

- **Donut RD**: If you have an RD where you are worried about small manipulation right around the threshold, you can re-run, dropping the observations right next to the threshold. That means you're comparing above to below, but ignoring the units most likely to be shifted by a little. This is referred to as a "donut" because you are taking out a hole in the middle of the running variable.

7.4 Example – The Effects of NICU

Estimating Marginal Returns to Medical Care: Evidence from At-Risk Newborns. Almond, Currie, Kowalski and Williams, QJE 2010

To Discuss:

- Context
- Identifying the running variable
- Look at plots
- Problems with the study

8 Advanced Methods

There are some advanced methods we did not cover today that are still within the “normal enough” scope, but you will have to get quite good at ecometrics

1. **Synthetic controls** – use when you have one treated unit, and multiple possible control units. Synthetic controls automatically chooses weights for your control units, to produce a single “synthetic” control unit that is a good match for your treated unit. Then, you compare treated vs control, before vs after. There are packages that do this; in Stata, use `synth`

   This is new-ish and not regularly used in health economics. See: Leder-Luis 2023 (ReStat)for my “staggered synthetic controls” but again, I don’t suggest you go that path

2. **Dynamic RD** – like regression discontinuity, but with a time-varying component. Suppose you have an RD every year, where units come in and out of treatment – then, you need to handle the effects of treatment year by year, with dynamic and long-term effects

   Best Citation: Cellini, Ferreira and Rothstein, 2010 (QJE)

   Newer: Barron, Hyman, Vasquez, Forthcoming Restat on school quality and adult crime.

   Less used in health econ – possibly a good space to explore

3. **Machine Learning- based tools**. There is a huge literature here. Machine learning can be used for everything from variable construction to causal analysis. Some new-fangled things include **causal forest**. C.f. the work on Susan Athey, and the new Finkelstein, Einav, Mahoney and Ji paper on home health fraud (NBER Summer Institute, 2024).

9 Final Thoughts

This is not that good of a reference guide. This is a crash course. You need to dive into this stuff, get serious, read textbooks, read papers, and attend seminars to understand the pitfalls in your (and others') studies

Imitation is your friend. Your goal is to do defensible econometrics on a new topic that has relevance to the real world. There is not a huge return to creating new methods.

There are other ways that people do economics not covered here, including theory, applied theory, and structural modeling. Those are beyond my expertise, but they are useful in some contexts – particularly where there is no quasi-experimental variation to exploit.