Metrics-Math Bootcamp Day 3

Cameron Taylor

August 14, 2019

Stanford GSB

Overview

1. Instrumental Variables
2. Panel Data (FE, RE, Diff-in-Diff)
3. Briefly: Both of these things in R

IV

Why IV?

- Recall one of the key questions and issues of interest: we are interested in the causal effect of X on Y
- We can use OLS to uncover this IF $E[\epsilon|X] = 0$
- But in many cases of interest, unlikely that this is true
- One way common way to address this: using an *instrument*
- The basic idea of an instrument is to find some quasi-random variation that shifts around X - then use this variation and its random variation with X to estimate the effect of X on Y
- Will work in single endogenous regressor and single instrument world for simplicity
    - When # instruments = # endogenous variables, call it "just identified"

IV Assumptions

- All IV's Z have two key assumptions
- First: **exclusion restriction** - $E[\epsilon|Z] = 0$ i.e. the Z does not belong in the full model
- Second: **relevance** - $\text{Cov}(X, Z) \neq 0$ i.e. the instrument actually shifts the X variable of interest (must be true after controlling for other variables too!)
- These assumptions form the basis of the estimation routine
- There are other assumptions that people generally "require" for IVs in more advanced settings (i.e. with heterogeneous causal effects)

IV Estimation: 2SLS I

- Many packages and functions in R (ivreg, felm) will have instrumental variables as an option
  - Only requires you to specify endogneous variables X, instrument Z and all other exogenous variables
- The most standard estimation method for IV is called **2SLS**
- First fit X as a function of Z and all other exogenous covariates
- Second fit Y on the fitted $\hat{X}$ from the first stage and all the other exogenous covariates
- Then regression coefficient on $\hat{X}$ in 2nd state is IV estimator
- But some potential tricky components: if you run two separate regressions, your standard errors will be wrong! (So use a package)
- Intuition: Only use quasi-random variation in X from Z to estimate effect

IV Estimation: 2SLS II

- Consider a case with a single regressor X and a single instrument Z
- First stage fitted values are

$$\hat{X} = Z'(Z'Z)^{-1}Z'X = P_Z X$$

  where P_Z is idempotent and so $P_Z^2 = P_Z$ and $P_Z' = P_Z$
- Then second stage is

$$\hat{\beta}_{IV} = ((P_Z X)'(P_Z X))^{-1}(P_Z X)'Y = (X'P_Z X)^{-1}X'P_Z Y = (Z'X)^{-1}Z'Y$$

- Then relatively easy to derive properties of estimator here

IV Properties

- As with the OLS estimator, the IV estimator has many properties to allow for inference
- In particular, it will have a limiting normal distribution due to $Z'\epsilon$ term which has mean 0
- The asymptotic variance-covariance matrix can again be derived, and we can use a homoscedastic and non-homoscedastic estimator as with OLS
- Most packages will automatically do inference for you
- Algebra and math looks similar to OLS, so won't go over here but can find in Mostly Harmless or other textbooks
- Importantly: IV is consistent *but* biased!
    - Though in practice I don't think this makes a huge difference for applied purposes

Assessing an IV: Relevance

- Recall two key assumptions: exclusion restriction and relevance
- We can actually test relevance since X and Z observable
- The rule of thumb is to look at the first stage (reg X on Z) and see if the F-stat on the instrument is more than 10
- If a single instrument, then F-stat is t-stat squared (so t-stat around 3.5 or so)
- If fails this rule-of-thumb test, then worry about *weak instruments* which have a large metrics literature

Assessing an IV: Exclusion Restriction

- The exclusion restriction cannot be tested
- But some ways to understand it (I'll briefly discuss 3 here)
- First, just thinking about logic of instrument and experimental ideal is a good way to think carefully about instruments
  - Many instruments are found through institutional details (company randomly assigns cases to workers, etc.)
- Next, can compare OLS and IV - if have theory that tells you direction of bias, this can confirm results
- Finally, some cases where IV should not predict something in the observables if it is truly quasi-random and not excluded - so can regress other observables on Z or check for observables of other characteristics based on values Z

IV Examples I: Card (1990)

- A question labor economists have occupied lots of their time with: what is the return to education?
- Obvious from data that education positively correlated with wages
- Problem: If people have innate unobservable abilities, might select into education based on abilities
- Card (1990) uses a quasi-random cost shifter to assess returns to education
- Instrument: Distance to a school
- Idea: This is a real cost shifter for attending school, but is plausibly not correlated with other variables in the wage model

IV Examples II: Angrist and Evans (1996)

- Question: How do fertility decisions affect labor supply?
- Problem: People with unobserved preferences for fertility or working will choose to have different number of kids
- Angrist and Evans (1996) use a quasi-random event that induces changes in fertility decisions to estimate effect
- Instrument: looking at families with 2 children, are both same-sex or different sex
- X here is whether the family has a third child
- Idea: Sex of child is randomly assigned, and people prefer to have diversity in sex of children

IV Examples III: Angrist (1990)

- Question: What is impact of military service on wages?
- Problem: People select into military based on factors that also affect wages
- Instrument: Use Vietnam draft lottery number as instrument for serving
- Idea: Lottery numbers randomly assigned but also affect chance of eventually serving in the military
- This "lottery" idea has been used in many subsequent empirical analyses, particularly for school choice and returns to different schools

Panel Data

What is Panel Data?

- Panel data is also called "longitudinal data"
- It involves individuals that are observed over multiple time periods
- Repeated cross-sections are NOT panel data (but can still be useful depending on the questions)
- Panels can be balanced (all people observed same amount of time) or unbalanced

Panel Data Methods

- Three main methods that I will cover
- First is fixed effects
- Second is random effects
- Finally, I will cover diff-in-diff very briefly
- The idea behind these methods in general is to utilize the time component to let us control for time-invariant unobserved features and to capture effects through time trends

Fixed Effects Model

- Panel data allows us to write a model relating y and x as

$$y_{it} = \beta x_{it} + \epsilon_{it}$$

- The value is that if x_{it} varies over time, then we can add a **fixed effect** α_i that controls for *time-invariant unobservables*

$$y_{it} = \alpha_i + \beta x_{it} + \nu_{it}$$

  where now $\epsilon_{it} = \alpha_i + \nu_{it}$
- The value is that now we only need ν_{it} to satisfy exogeneity requirements (deviations from individual means)
- This makes x_{it} variation potentially much more plausibly "exogenous"
  - But still need to argue exogeneity for ν as in previous models

Fixed Effects Costs

- Are there costs to fixed effects? Yes!
- First, it gets rid of all variation "between" individuals in estimation and only uses "within" variation
- This can be quite a high cost - if variation in x is not very large over time, will lose lots of precision
- Second, if hoping to identify parameters γ on other variables z_{it} that have no variation over time for each individual, then cannot separately identify γ and α_i
- There are ways to get around this second problem by parameterizing the fixed effects and running a second regression, but I have not seen this very commonly

Fixed Effect Estimation

- Many packages in R will automatically do this - I use the felm package to estimate fixed effect models in R (will go over this more)
- The most common way to estimate that I know of is to use the Least Square Dummy Variable (LSDV) method
- Just put an individual specific dummy for every observation, parameter on this dummy is then α_i
- Other ways: since dummies get rid of means, can take out means; also can do first-differencing

Fixed Effect Inference: Clustering

- An important issue when using panel data: errors are likely correlated within groups or individuals
- If we do not account for this in estimation, our standard errors will be way too small
- To deal with this, use **clustering**
- The algebra is a bit messy so I won't go through it, but most softwares (R, Stata) will allow you to do it pretty easily
- It is very rare to NOT cluster standard errors when using panel data
- A conservative rule of thumb: cluster at most aggregate level of fixed effect (more aggregate cluster is more conservative)
    - But can also reason based on your situation and using institutional details/experimental analogy for where clustering should happen

Random Effects Intro

- Random effects differs from fixed effects by relaxing some assumptions and then applying the logic of the generalized linear model
- Thus basically treat like OLS before without α_i fixed effects in model, but not have specific error structure (like heteroscedasticity) that we can use to improve efficiency of estimates
- I will go through it quickly - important to be aware of it, but FE is much more common

Random Effects Model

- Consider the model
$$y_{it} = \beta x_{it} + \alpha_i + \nu_{it}$$
- Now assume that x_{it} is independent of both α_i and ν_{it} (the first was not required in fixed effects)
- Assume otherwise i.i.d. draws and α and ν independent
- Then what this does is implies specific variance structures on the error term $\eta_{it} = \alpha_i + \nu_{it}$
- The particular error model is
$$V(\eta_{it}) = \sigma_a^2 + \sigma_n^2$$
where $V(\alpha) = \sigma_a^2$ and $V(\nu) = \sigma_n^2$ and
$$\text{Cov}(\eta_{it}, \eta_{is}) = \sigma_a^2$$

Random Effects Estimation

- As with GLS and dealing with heteroscedasticity before, need to estimate σ_a and σ_n
- Use FGLS method - first stage to get some objects that allow for producing $\hat{\sigma}$ and then plug those in as weights in running a second stage GLS
- We didn't go over FGLS but I recommend looking at Greene for the details

Random Effects Inference: Hausman Test

- Inference works as with heteroscedasiticty before in OLS - asymptotic distributions, etc.
- A useful tool is the **Hausman test** which tests between fixed effects and random effects
- Recall the major difference between the two models: α_i uncorrelated with x_{it}
- The null hypothesis of this test is that these are uncorrelated (i.e. random effects) - if reject, reject in favor of fixed effect model
- It is pretty standard to perform these tests in R and Stata

Diff-in-Diff Intro

- A very common way to use panel data for causal inference is **Differences-in-Differences**
- Basic idea: assuming that different units have parallel trends, and one unit receives treatment, we can take the difference in the difference in outcomes to estimate effect
- Usually done at a more aggregate level: cities, states, etc. instead of individuals

Diff-in-Diff Model

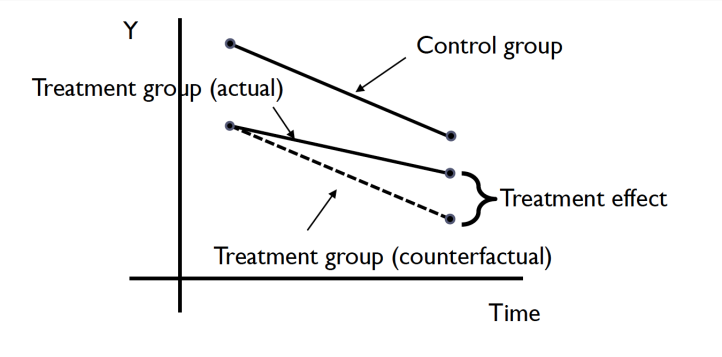
- Model:

$$y_{it} = \alpha_i + \delta_t + \beta D_{it} + \epsilon_{it}$$

  where D is an indicator for the treatment
- Thus β is a constant effect of the treatment - main parameter of interest
- Trends for outcome are captured by additive $\alpha_i + \delta_t$ for any unit i at time t
- If D is independent of ϵ then β is the causal effect

Diff-in-Diff Assumptions

- Two main identifying assumptions
- First, the units have parallel trends - if not, then whole idea fails
- Second, only captures the effect from the treatment of interest

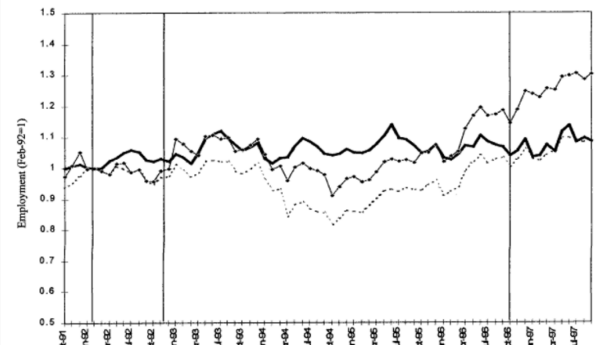


Diff-in-Diff Properties and Inference

- Placing in regression framework with fixed effects implies we can use the inference methods from before
- The key property it the identifying assumptions of parallel trends and intensity of treatment - gives plausible causal inference
- In general, I think this has become a bread-and-butter method for causal inference in economics

Diff-in-Diff Example

- A very famous diff-in-diff study: Card and Krueger on minimum wage and employment
- Compared restaurants in PA and NJ; NJ had a min wage increase
- Find no effect of min wage on employment in diff-in-diff



IV and Panel Data in R

IV in R

- I use the function ivreg() in R
- Works similar to lm()

```
> library(AER)
> set.seed(1)
> z <- rnorm(1000)
> u <- rnorm(1000)
> x <- z+u+rnorm(1000)
> y <- x+u+rnorm(1000)
> summary(iv <- ivreg(y~x | z))$coefficients
               Estimate Std. Error     t value      Pr(>|t|)
(Intercept) 0.0008183143 0.04618252  0.01771914  9.858665e-01
x           1.0284086843 0.04229140 24.31720634 5.984343e-103
> summary(reg <- lm(y~x))$coefficients
              Estimate Std. Error   t value  Pr(>|t|)
(Intercept) 0.004803324 0.04235475  0.113407 0.9097307
x           1.344653733 0.02303787 58.367094 0.0000000
```

Figure 3: IV

Panel Data in R

- I use the package lfe and the function felm() for fixed effect estimation
- Other people like the package and function plm because I think it has more panel data functionality (ex: random effects)

```
> library(lfe)
> set.seed(1)
> x <- rnorm(1000)
> i1 <- c(rep(1,500), rep(0,500))
> t <- c(1:500, 1:500)
> y <- 10*i1 + 0.01*t+2*x+rnorm(1000)
> summary(fe <- felm(y~x+t | i1 | 0 | i1))$coefficients
    Estimate Cluster s.e.  t value Pr(>|t|)
x 2.00593785 0.0325209506 61.68140        0
t 0.01010919 0.0001093155 92.47717        0
```

Figure 4: Fixed Effects

Wrap Up

- We learned about IV as a strategy to get at causal questions
- I recommend reading mostly harmless chapter 4 closely for more details
- Also looked at benefits of panel data, and how to use
- I also recommend mostly harmless chapter 5 for more on diff-in-diff and causal approaches to panel data

HW

- The "least optional" of all HWs so far
- You have three tasks
- First: Find a dataset that interests you; load it into your programming language and clean it. Make some basic summary statistics to understand it.
- Second: Propose a model of the data in some way and estimate it. What do we learn from the estimates?
- Last: Do a brief write-up and presentation of your results. Each of you will present very briefly (3 mins) to the class (presentation skills very important!)
- Help: I like using Kaggle or data.world to find random fun datasets. In general, newspaper articles and blogs have some interesting data. Also can just searching KEYWORD and then "data"