

Metrics-Math Bootcamp Day 3

Cameron Taylor

August 5, 2021

Stanford GSB

Overview

1. Instrumental Variables
2. Panel Data (FE, Diff-in-Diff)
3. Wrap-Up, IV and Panel in R, HW

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: many unobservable factors correlated with X and y
- 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 look at correlation between this induced variation in X and Y
- Can use multiple instruments. We will focus on single endogenous variable, single instrument
 - 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). Can read more in Mostly Harmless if interested.

IV Estimation: 2SLS I

- Many packages and functions in R (ivreg, felm) have iv estimators
 - Specify endogeneous variables X , instrument Z , other exogenous variables W
- The most standard estimation method for IV is called **2SLS** which uses

$$X_i = \alpha Z_i + W_i \gamma_1 + \nu_i$$

$$Y_i = \beta X_i + W_i \gamma_2 + \epsilon_i$$

- 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 step is IV estimator
- Watch out: if you run two separate regressions, your standard errors will be wrong! (So use a package)

IV Estimation: 2SLS II

- Consider a case with a single regressor X and a single instrument Z , no other covariates
- 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!
 - Doesn't seem to make 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 (not on all the variables!) 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 help understand results (complicated with heterogeneous treatment effects)
- 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 labor market 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
- There is now a huge literature and lots of work “lottery” ideas and methods, particularly for school choice and returns to different school types and 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

- Two main methods that I will cover
- First is fixed effects
- Second, I will cover diff-in-diff briefly (application of fixed effects)
- The main benefit of panel data and these methods is to utilize the individual and 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

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

Recent Examples of Fixed Effects: Two-Way Fixed Effects

- Two-way fixed effect models are very popular in applied work now
- Example: We want to know how a workers firm affects their wage (AKM / Card, Heining, Kline 2013)?

Write wage as

$$w_{it} = \alpha_i + \gamma_{j(i)} + \epsilon_i \quad (1)$$

where i is a worker index, $j(i)$ is the firm that worker i works at

- Challenge: how to separately identify α_i and $\gamma_{j(i)}$
 - We need workers that move between firms!

- 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:

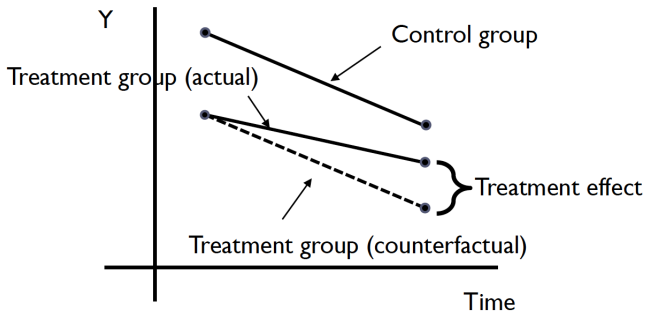
$$y_{it} = \alpha_i + \delta_t + \beta D_{it} + \epsilon_{it}$$

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

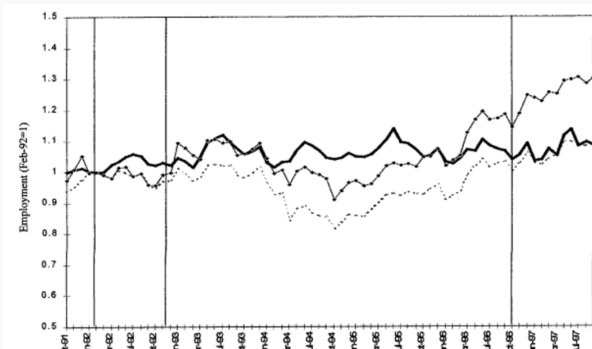


Diff-in-Diff Properties and Inference

- Placing in regression framework with fixed effects implies we can use the inference methods from before
- The key property is 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



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

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

- First: Pick one of these datasets (or find another dataset of interest on Kaggle)
 - US County COVID-19
 - Hotel Bookings Dataset
 - FIFA World Cup
 - S&P 500 Dataset
- Second: Download the data and load into R
- Third: Propose a model of the data or propose a hypothesis and test that hypothesis/estimate the model
- Fourth: Write up findings in 3-4 slides