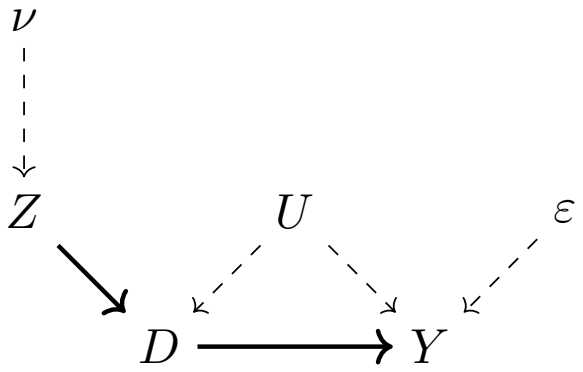
Roadmap

- Intuition
  - Causal Graphs
  - Random Lightning Bugs in a Jar
  - Strangeness Principle
- Two sub IV designs
  - Externalities on Airbnb Platform
  - Limiting a user's choice set
  - Non-compliance in an RCT
- Estimators
  - Two Step
  - Weak instruments
- Heterogenous treatment effects
  - Potential outcomes
  - Assumptions
  - Local average treatment effects
- Data visualization

So what is instrumental variables?

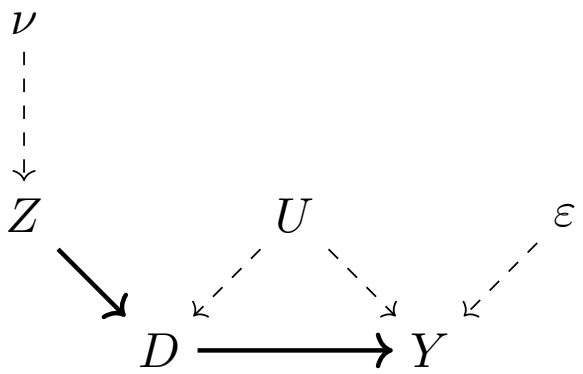
- As we saw, selection on observables has a high, not a low, bar because of the confidence we need in the DAG to justify CIA
- Alternative approaches are available and instrumental variables is perhaps the oldest and one of the most powerful
- Instrumental variables estimates an average causal effect using variation in a treatment caused by a third variable, the instrument
- Estimation has been straightforward for a while
- Challenge isn't estimation; it's the assumptions

Valid Instruments I: Causal Graphs



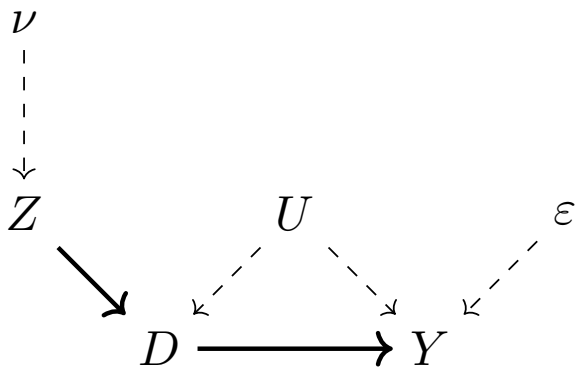
Notice how Z is determined by randomness ν (independence)

Valid Instruments I: Causal Graphs



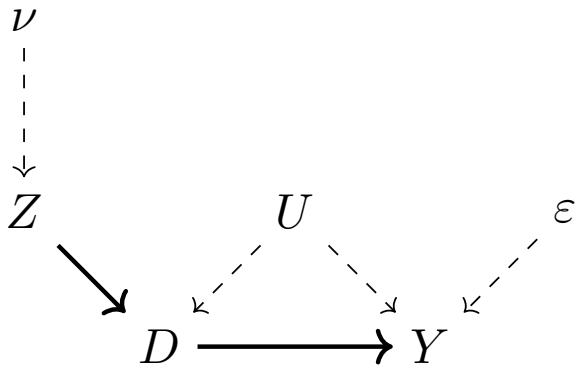
Notice how Z is independent of ε and U, as well as blocked by a collider, D (exclusion I)

Valid Instruments I: Causal Graphs



Notice how along the path $Z \to D \to Y$, D is a mechanism and Z only operates on the outcome via D (exclusion II)

Valid Instruments I: Causal Graphs



Notice $Z \to D$ (nonzero first stage)

Valid Instruments II: Random Lightning Bugs

- Imagine a jar filled with two kinds of bugs Z producing light Y
- Some of move in patterns; but some of move around randomly
- Instruments are special honey Z jostling only the random bugs
- Where can we find data with all three: D, Y, and Z?

Valid Instruments III: Strangeness Principle

- Over Thanksgiving dinner, I say: “Did you know if a woman’s first two kids are a boy, she is less likely to work?”
- You say “*Compared to who*?”
- I say “Compared to a mother whose first two kids had been boy and girl.”
- You are an intelligent lay person, but you are puzzled and think “why would gender composition of first two kids matter for work?”

Valid Instruments III: Strangeness Principle

- Be explicit: why should gender composition of first two kids predict labor market participation in women?
- You're right to be puzzled. It is legitimately puzzling. It isn't clear what is going on at all.
- You need more information, in other words, otherwise the layperson can't understand what same gender of your children has to do with working

Valid Instruments III: Strangeness Principle

- So then I point out that women whose first two children are of the same gender are more likely to have additional children than women whose first two children are of different genders
- You say, “Oh I bet that’s it: she exits the market because she had more kids, not just because of gender composition. Gender composition just made her have more kids.”

Valid Instruments III: Strangeness Principle

Strangeness principle

The confusion an intelligent layperson with familiarity in the outcome feels when they hear about a correlation between the instrument and the outcome before learning of the treatment

- Instrumental variables strategizes formalizes *strangeness* when using an instrument to predict an outcome
- If two variables don't seem to go together whatsoever, but one of them is highly correlated with the other, you might have stumbled upon a valid instrument
- May be why ironically valid instruments get criticized as "cute" also

Valid Instruments III: Strangeness Principle

- Let's listen to a few lines from "Ultralight Beam" by Kanye West.
- Chance the Rapper is featured and sings

  "I made Sunday Candy, I'm never going to hell
  I met Kanye West, I'm never going to fail."
  - Chance the Rapper

- Assume Sunday Candy and Kanye West are the instruments in their verses. Let's consider each in order

Valid Instruments III: Strangeness Principle

“I made Sunday Candy,
I’m never going to hell”,

I’m confused. What does that song have to do with the afterlife, even in Christianity? There must be more to this story, right?

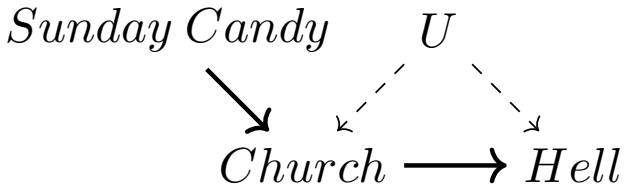
Valid Instruments III: Strangeness Principle

What if it's something like this

> *"I made Sunday Candy*
> *this pastor invited me to church on Sunday,*
> *I'm never going to hell"*

You may disagree, but at least it no longer feels strange once the whole story is revealed. Good instruments feel like that – confusion followed by understanding

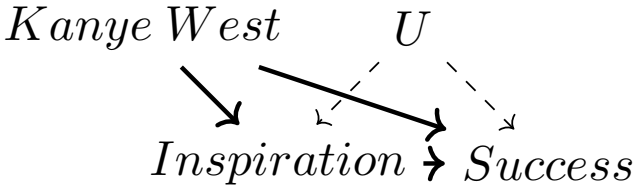
Valid Instruments III: Strangeness Principle



Valid Instruments III: Strangeness Principle

- Chance idolized Kanye West and said Kanye inspired him ever since high school
- Kanye is an amazing artist, but is he a good instrument for Chance's inspiration
- To be an instrument for Chance's inspiration, Kanye can only affect Chance's success through his inspiration
- I am not surprised knowing Kanye determines success – nothing about it seems strange
- Kanye is not a good instrument for Chance's inspiration because Kanye can directly affect success too

Valid Instruments III: Strangeness Principle



Roadmap

- Intuition
  - Causal Graphs
  - Random Lightning Bugs in a Jar
  - Strangeness Principle
- Two sub IV designs
  - Externalities on Airbnb Platform
  - Limiting a user's choice set
  - Non-compliance in an RCT
- Estimators
  - Two Step
  - Weak instruments
- Heterogenous treatment effects
  - Potential outcomes
  - Assumptions
  - Local average treatment effects
- Data visualization

Quality externalities on Airbnb platform

- When a guest has a negative experience with a host on the Airbnb platform, they may (or may not) be less likely to return to the platform as a whole
- If a host on Airbnb causes a guest to stop coming to Airbnb altogether, then it is a “quality externality” called “guest return propensity” (GRP)
- Joffe, et al. (2019) theorize that GRP will have a stronger effect on the beliefs of new users who have less info abt platform quality

Findings

- They construct GRP metrics for each accommodation on Airbnb and found significant variation across listings, even controlling for guest and trip characteristics
    1. GDP has a larger effect on inexperienced guests return probability than experienced ones
    2. But, experienced guests get impacted because they tend to be heavy travelers which causes small belief shocks to cascade
- Conclude “from the platforms POV, experienced guests are attractive targets for high-GRP listings, even if their beliefs are less affected by a listing’s GRP.”

Instrumenting for GRP

*"We **instrument for the GRP** of the booked listing with the **average GRP across the first page of search results**, using only variation generated by multiple guests searching for listings in the same market, for the same travel day, with comparable trip lead time."*

What is going on?

- You go to Airbnb, you type in “Hot Springs, AR” and a bunch of places show up by the lake
- Another person does it that day, and they see a different choice set often for the same search
- You tend to pick from the front page choice set – they’re going to instrument for what you picked with the average GRP of that first page
- Where you booked is the treatment; what you were presented from your search is the instrument
- Find booking a listing with a 1SD higher GRP causes you to take 0.32 additional future trips

Table 9: First stage and IV results:
Effect of search results on booking characteristics, and instrumented effect of booking characteristics on subsequent trips

	First stage		
	GRP Booked	Rating Booked	Trips After (IV)
Quality booked			0.318^{***}
			(0.047)
Rating booked			0.102
			(0.104)
Avg quality shown	0.474^{***}	0.002	
	(0.003)	(0.001)	
Avg Jan 1 Rating	0.003	0.326^{***}	
	(0.005)	(0.002)	
Avg Price	-0.00003^{***}	-0.00001^{***}	-0.00004
	(0.00000)	(0.00000)	(0.00002)
1 Past Trips	0.0003	0.007^{***}	0.345^{***}
	(0.001)	(0.001)	(0.006)
2 Past Trips	0.001	0.010^{***}	0.636^{***}
	(0.002)	(0.001)	(0.008)
3+ Past Trips	0.004^{**}	0.014^{***}	2.365^{***}
	(0.002)	(0.001)	(0.009)
F-Stat	.	.	16289.5
Observations	3,729,162	3,704,788	3,738,152
R^2	0.346	0.380	0.414
Adjusted R^2	0.062	0.109	0.160

*p<0.1; **p<0.05; ***p<0.01

Note: The first three columns show the effect of search results on the characteristics of the booking, the first stage for the IV regression. The second column is limited to trips at listings for whom the GRP measure is based on at least 20 trips. The last column shows the IV results where we use the average GRP and rating of search results to instrument for the GRP and rating booked. "Past Trips" refers to the number of trips

Sub IV Designs: Is this One?

- Oftentimes in causal inference, specific *types* of instrumental variables situations arise over and over
- When this happens, it tends to attract focused attention as we study these particular kinds of applications
- This paper seems to outline a type of IV that I could imagine is unique to many platforms based on two sided matching via search

Explanation

- Recall the problem that Blake, et al. (2015) were facing when evaluating the causal effect of paid search advertising on revenue – selection bias
- Anyone searching may be already inclined to have certain unobserved and heterogenous preferences for those items such that absent the treatment they still would have had different potential outcomes as others
- But let's consider this possible design from previously – what exactly is the instrument?

Randomized consideration sets as instruments

- On Talkspace, a platform matching therapists with clients, you an algorithm presents you with several therapists to choose from
- You can only choose from among them – this is your consideration set
- Insofar as the box has quasi random elements, then you may be able to construct instruments to answer questions that otherwise are flawed by selection bias
- For instance, a box consisting of different modalities (e.g., mindfulness) or therapist race would seem to raise the probability that you picked one of those

Randomized consideration sets as instruments

- You may be able to take advantage of these types of platform style instruments to address the selection bias
- Randomized consideration sets that limit search *options* could instrument for the option you did choose
- Measurement, though, will be key here

Sub-IV designs: Non-compliance with field experiments

- A second type of sub-IV design is using randomized nudges (e.g., vouchers) as instruments for treatment participation in field experiments
- Oftentimes in an experiment humans refuse to comply to treatment assignment
- Instrumenting for treatment assignment with the randomized vouchers will allow us to identify aggregated causal effects
- Let's dive deeper into precisely how these work though by exploring some explicit estimators

Roadmap

- Intuition
  - Causal Graphs
  - Random Lightning Bugs in a Jar
  - Strangeness Principle
- Two sub IV designs
  - Externalities on Airbnb Platform
  - Limiting a user's choice set
  - Non-compliance in an RCT
- Estimators
  - Two Step
  - Weak instruments
- Heterogenous treatment effects
  - Potential outcomes
  - Assumptions
  - Local average treatment effects
- Data visualization

Two step vs Minimum Distance

- Wald, Two Sample, the jackknife, Two Stage Least Squares are forms of two step procedures in which first and second stages are calculated separately
- Too much to review as IV *estimation* is a *huge* area, so I will focus on a few things, starting with two stage least squares (2SLS)
- 2SLS is basically the workhorse IV model, though it can have some issues because of its finite sample bias with weak instruments

Two-stage least squares language

Suppose you have a sample of data on Y, S, and Z. For each observation i we assume the data are generated according to

$$\begin{aligned} Y_i &= \alpha + \delta S_i + \eta_i \text{ (causal model)} \\ S_i &= \gamma + \rho Z_i + \zeta_i \text{ (first stage)} \end{aligned}$$

where $Cov(Z, \eta_i) = 0$ (strangeness, hereafter exclusion) and $\rho \neq 0$ (relevance, hereafter non-zero first stage)

2SLS and Wald

$$
\begin{aligned}
Y_i &= \psi + \pi Z_i + \varepsilon_i \text{ (reduced form)} \\
S_i &= \gamma + \rho Z_i + \zeta_i \text{ (first stage)}
\end{aligned}
$$

We can calculate the ratio of “reduced form” (π) to “first stage” coefficient (ρ) using the Wald IV estimator:

$$
\widehat{\delta}_{2sls} = \frac{Cov(Z,Y)}{Cov(Z,S)} = \frac{\frac{Cov(Z,Y)}{Var(Z)}}{\frac{Cov(Z,S)}{Var(Z)}} = \frac{\widehat{\pi}}{\widehat{\rho}}
$$

Two-stage least squares

Carry over from previous slide

$$\widehat{\delta}_{2sls} = \frac{Cov(Z,Y)}{Cov(Z,S)} = \frac{\frac{Cov(Z,Y)}{Var(Z)}}{\frac{Cov(Z,S)}{Var(Z)}} = \frac{\widehat{\pi}}{\widehat{\rho}}$$

Rewrite $\widehat{\rho}$ as

$$\begin{aligned} \widehat{\rho} &= \frac{Cov(Z,S)}{Var(Z)} \\ \widehat{\rho} Var(Z) &= Cov(Z,S) \end{aligned}$$

Two-stage least squares

Multiply Wald IV by $\frac{\widehat{\rho}}{\widehat{\rho}}$

$$\widehat{\delta}_{2sls} = \frac{Cov(Z,Y)}{Cov(Z,S)} = \frac{\widehat{\rho}Cov(Z,Y)}{\widehat{\rho}Cov(Z,S)}$$

Substitute $Cov(Z,S) = \widehat{\rho}Var(Z)$ and simplify as constants disappear in covariance and variance

$$\begin{aligned}\widehat{\delta}_{2sls} &= \frac{\widehat{\rho}Cov(Z,Y)}{\widehat{\rho}Cov(Z,S)} = \frac{\widehat{\rho}Cov(Z,Y)}{\widehat{\rho}^2Var(Z,S)} \\ &= \frac{Cov(\widehat{\rho}Z,Y)}{Var(\widehat{\rho}Z)}\end{aligned}$$

Two-stage least squares

Recall

$$S_i = \gamma + \rho Z_i + \zeta_i \text{ (first stage)}$$

So after estimation, we get

$$\widehat{S} = \widehat{\gamma} + \widehat{\rho} Z \text{ (fitted values)}$$

Substitute for $\widehat{S}$ for $\widehat{\rho} Z$ ($\widehat{\gamma}$ drops out)

$$\widehat{\delta}_{2sls} = \frac{Cov(\widehat{\rho} Z, Y)}{Var(\widehat{\rho} Z)} = \frac{Cov(\widehat{S}, Y)}{Var(\widehat{S})}$$

Intuition of 2SLS

- Intuition is that 2SLS replaces S with the fitted values $\widehat{S}$ from the first stage regression of S onto Z and all other covariates
- I prefer the intuition of 2SLS to the intuition of the ratio of reduced form to first stage, even though as we just showed the cross walk between the two exists – just how I was brought up!
- By using the fitted values of the endogenous regressor from the first stage regression, our regression now uses *only* the exogenous variation in the regressor due to the instrumental variable itself

Breakout II: College in the county

- David Card (2021 Nobel Laureate in economics) wrote an interesting paper estimating the causal effect of attending college on log earnings
- Attending college and earnings are likely confounded by unobserved factors like “ability” (labor economist catch-all) which both makes one more likely to attend college and have higher wages even without attending
- Instruments for college attendance using “college in the county” as an instrument

Breakout: Estimate with software

Probably not a bad idea to estimate both reduced form and first stage, just to check everything is sensible, but ultimately you want to use software because second stage standard errors are wrong

Several software options – use one of them for 2SLS as they calculate correct standard errors

- Estimate this in Stata using -ivregress 2sls-.
- Estimate this in R -ivreg()- which is in the AER package
- Lots of options, like -linearmodels-, in python

Breakout: College in the county

1. Draw a canonical IV DAG – is college in the county “strange”? Why/why not?
2. Estimate the effect using Google Collab example illustrating python code using -linearmodels-
3. Answer questions in part 6

https://colab.research.google.com/github/scunning1975/python-mixtape/blob/main/Instrumental_Variables.ipynb#scrollTo=0nTjwGx9y6eW

Weak instruments

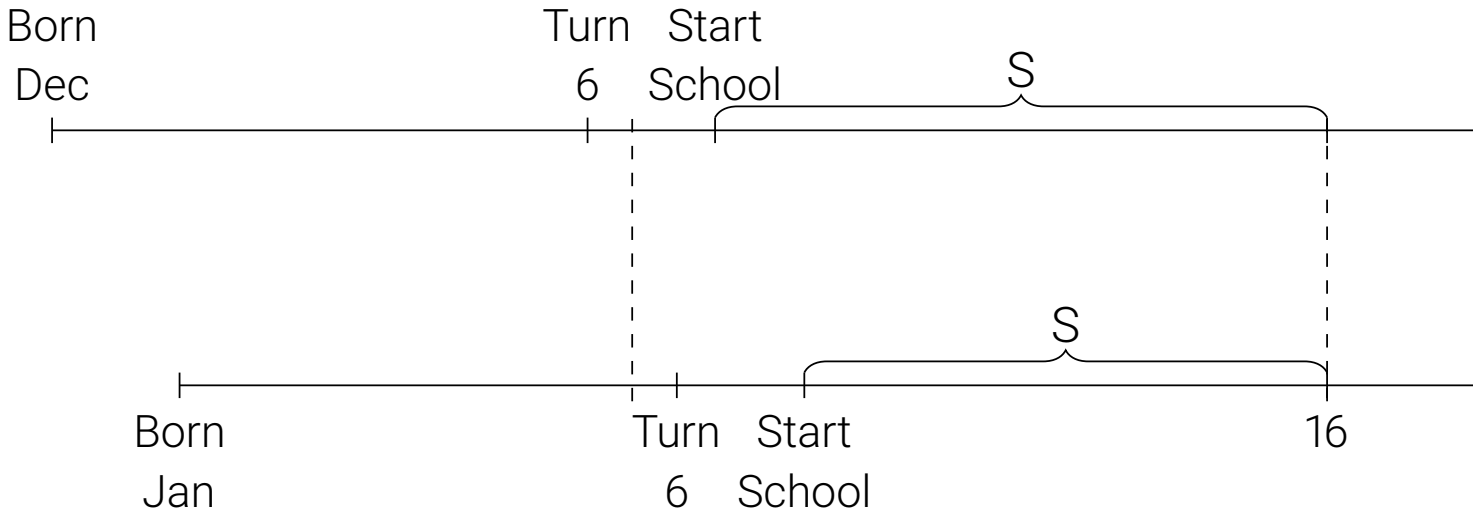
- A weak instrument is one that is not strongly correlated with the endogenous variable in the first stage
- This can happen if the two variables are independent or the sample is small
- If you have a weak instrument, then the bias of 2SLS is centered on the bias of OLS and the cure ends up being worse than the disease
- We knew this was a problem, but it was brought into sharp focus with Angrist and Krueger (1991) and some papers that followed

Angrist and Krueger (1991)

- In practice, it is often difficult to find convincing instruments – usually because potential instruments don't satisfy the exclusion restriction
- But in an early paper in the causal inference movement, Angrist and Krueger (1991) wrote a very interesting and influential study instrumental variable
- They were interested in schooling's effect on earnings and instrumented for it with *which quarter of the year you were born*
- Remember Chance quote - what the heck would birth quarter have to do with earnings such that it was an excludable instrument?

Compulsory schooling

- In the US, you could drop out of school once you turned 16
- “School districts typically require a student to have turned age six by January 1 of the year in which he or she enters school” (Angrist and Krueger 1991, p. 980)
- Children have different ages when they start school, though, and this creates different lengths of schooling at the time they turn 16 (potential drop out age):



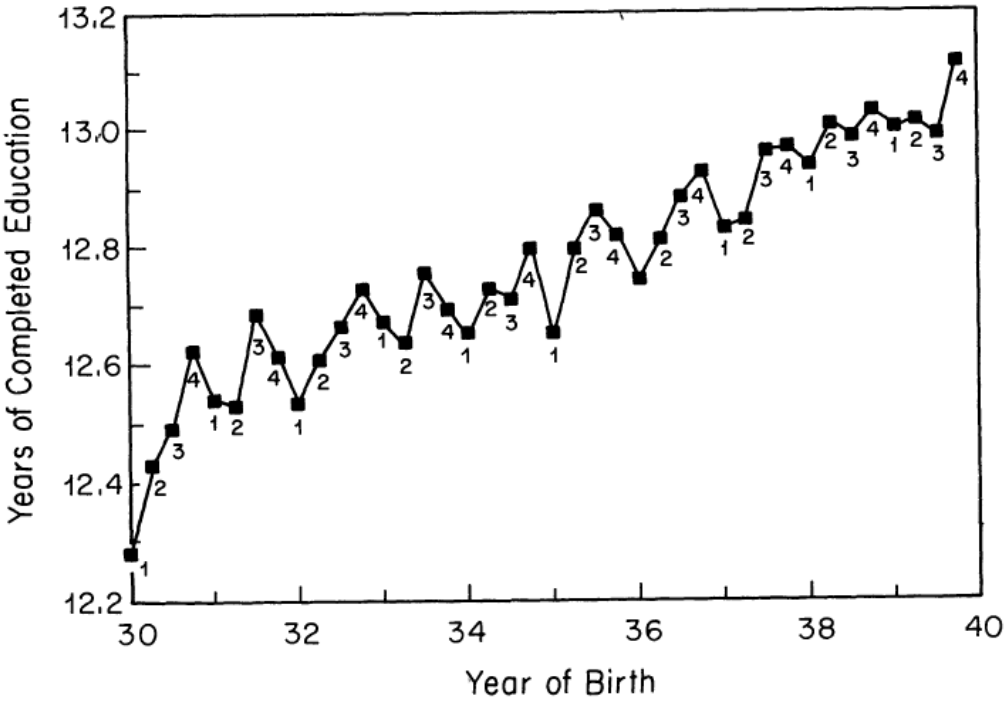
If you’re born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

Visuals

- You need good data visualization for IV partly because of the scrutiny around the design
- The two pieces you should be ready to build pictures for are the first stage and the reduced form
- Angrist and Krueger (1991) provide simple, classic and compelling pictures of both

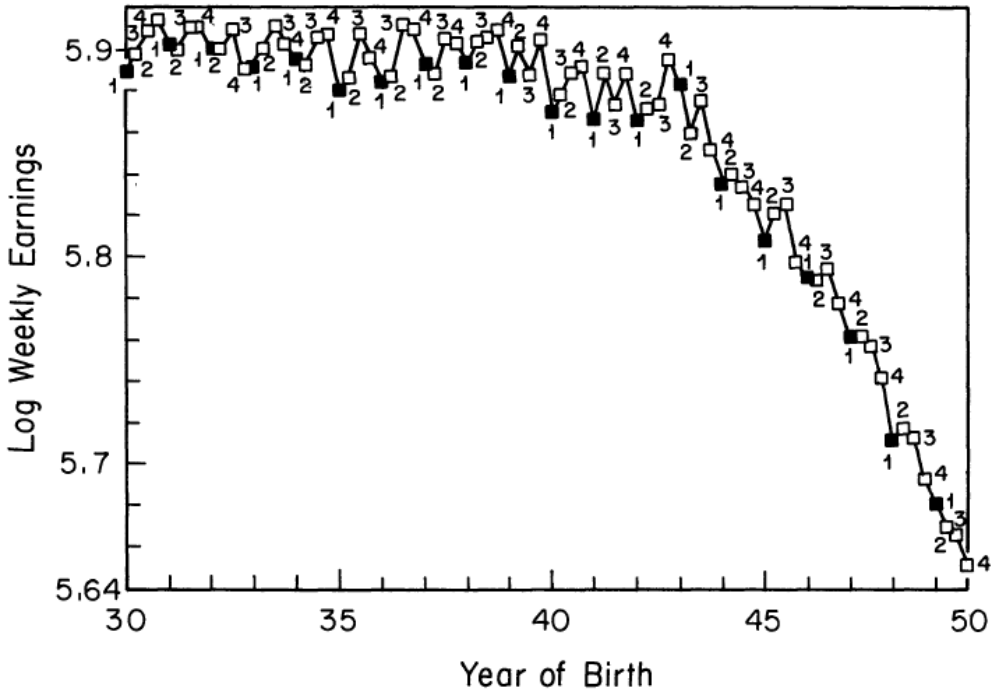
First Stage

Men born earlier in the year have lower schooling. This indicates that there is a first stage. Notice all the 3s and 4s at the top. But then notice how it attenuates over time . . .



Reduced Form

Do differences in schooling due to different quarter of birth translate into different earnings?



Two Stage Least Squares model

- The causal model is

$$Y_i = X\pi + \delta S_i + \varepsilon$$

- The first stage regression is:

$$S_i = X\pi_{10} + \pi_{11} Z_i + \eta_{1i}$$

- The reduced form regression is:

$$Y_i = X\pi_{20} + \pi_{21} Z_i + \eta_{2i}$$

- The covariate adjusted IV estimator is the sample analog of the ratio, $\frac{\pi_{21}}{\pi_{11}}$

Two Stage Least Squares

- Angrist and Krueger instrument for schooling using three quarter of birth dummies: a dummies for 1st, 2nd and 3rd qob
- Their estimated first-stage regression is:

$$S_i = X\pi_{10} + Z_{1i}\pi_{11} + Z_{2i}\pi_{12} + Z_{3i}\pi_{13} + \eta_1$$

- The second stage is the same as before, but the fitted values are from the new first stage

First stage regression results

Quarter of birth is a strong predictor of total years of education

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect[a] I	II	III	F-test[b] [P-value]
Total years of education	1930–1939	12.79	−0.124 (0.017)	−0.086 (0.017)	−0.015 (0.016)	24.9 [0.0001]
	1940–1949	13.56	−0.085 (0.012)	−0.035 (0.012)	−0.017 (0.011)	18.6 [0.0001]
High school graduate	1930–1939	0.77	−0.019 (0.002)	−0.020 (0.002)	−0.004 (0.002)	46.4 [0.0001]
	1940–1949	0.86	−0.015 (0.001)	−0.012 (0.001)	−0.002 (0.001)	54.4 [0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004 (0.014)	0.051 (0.014)	0.012 (0.014)	5.9 [0.0006]
	1940–1949	14.28	0.005 (0.011)	0.043 (0.011)	−0.003 (0.010)	7.8 [0.0017]
College graduate	1930–1939	0.24	−0.005 (0.002)	0.003 (0.002)	0.002 (0.002)	5.0 [0.0021]
	1940–1949	0.30	−0.003 (0.002)	0.004 (0.002)	0.000 (0.002)	5.0 [0.0018]

IV Estimates Birth Cohorts 20-29, 1980 Census

Independent variable	(1) OLS	(2) TSLS
Years of education	0.0711	0.0891
	(0.0003)	(0.0161)
Race (1 = black)	—	—
SMSA (1 = center city)	—	—
Married (1 = married)	—	—
9 Year-of-birth dummies	Yes	Yes
8 Region-of-residence dummies	No	No
Age	—	—
Age-squared	—	—
χ^2 [dof]	—	25.4 [29]

Problem enters with many quarter of birth interactions

- They want to increase the precision of their 2SLS estimates, so they load up their first stage with more instruments
- Specifications with 30 (quarter of birth $\times$ year) dummy variables and 150 (quarter of birth $\times$ state) instruments
  - $\rightarrow$ What's the intuition here? The effect of quarter of birth may vary by birth year or by state
  - $\rightarrow$ By interacting their instrument with variables, they are "saturating" their 2SLS regression model (more on that later)
- It reduced the standard errors, but that comes at a cost of potentially having a weak instruments problem

More instruments

TABLE VII

OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1930–1939: 1980 CENSUS[a]

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0673	0.0928	0.0673	0.0907	0.0628	0.0831	0.0628	0.0811
	(0.0003)	(0.0093)	(0.0003)	(0.0107)	(0.0003)	(0.0095)	(0.0003)	(0.0109)
Race (1 = black)	—	—	—	—	−0.2547	−0.2333	−0.2547	−0.2354
					(0.0043)	(0.0109)	(0.0043)	(0.0122)
SMSA (1 = center city)	—	—	—	—	0.1705	0.1511	0.1705	0.1531
					(0.0029)	(0.0095)	(0.0029)	(0.0107)
Married (1 = married)	—	—	—	—	0.2487	0.2435	0.2487	0.2441
					(0.0032)	(0.0040)	(0.0032)	(0.0042)
9 Year-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
8 Region-of-residence dummies	No	No	No	No	Yes	Yes	Yes	Yes
50 State-of-birth dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Age	—	—	−0.0757	−0.0880	—	—	−0.0778	−0.0876
			(0.0617)	(0.0624)			(0.0603)	(0.0609)
Age-squared	—	—	0.0008	0.0009	—	—	0.0008	0.0009
			(0.0007)	(0.0007)			(0.0007)	(0.0007)
χ^2 [dof]	—	163 [179]	—	161 [177]	—	164 [179]	—	162 [177]

a. Standard errors are in parentheses. Excluded instruments are 30 quarter-of-birth times year-of-birth dummies and 150 quarter-of-birth times state-of-birth interactions. Age and age-squared are measured in quarters of years. Each equation also includes an intercept term. The sample is the same as in Table VI. Sample size is 329,509.

Weak Instruments

- Important paper suggesting OLS and 2SLS were pretty similar, as well as the power of natural experiments ("plausibly exogenous")
- But in the early 1990s, a number of papers highlighted that IV can be *severely* biased – in particular, when instruments have only a weak correlation with the endogenous variable of interest and when many instruments are used to instrument for one endogenous variable (i.e., there are many overidentifying restrictions).
- In the worst case, if the instruments are so weak that there is no first stage, then the 2SLS sampling distribution is centered on the probability limit of OLS

Matrices and instruments

- The causal model of interest is:

$$Y = \beta X + \nu$$

- Matrix of instrumental variables is Z with the first stage equation:

$$X = Z'\pi + \eta$$

Weak instruments and bias towards OLS

- If ν_i and η_i are correlated, estimating the first equation by OLS would lead to biased results, wherein the OLS bias is:

$$E[\beta_{OLS} - \beta] = \frac{Cov(\nu, X)}{Var(X)}$$

- If ν_i and η_i are correlated the OLS bias is therefore: $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$

Weak instruments and 2SLS bias towards OLS

- We can derive the approximate bias of 2SLS as:

$$E[\widehat{\beta}_{2SLS} - \beta] \approx \frac{\sigma_{\nu\eta}}{\sigma_\eta^2} \frac{1}{F+1}$$

- Consider the intuition all that work bought us now: if the first stage is weak (i.e, $F \to 0$), then the bias of 2SLS approaches $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$

Weak instruments and bias towards OLS

- This is the same as the OLS bias as for $\pi = 0$ in the second equation on the earlier slide (i.e., there is no first stage relationship) $\sigma_x^2 = \sigma_\eta^2$ and therefore the OLS bias $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$ becomes $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$.
- But if the first stage is very strong ($F \to \infty$) then the 2SLS bias is approaching 0.
- Cool thing is – you can test this with an F test on the joint significance of Z in the first stage
- It's absolutely critical therefore that you choose instruments that are strongly correlated with the endogenous regressor, otherwise the cure is worse than the disease

Weak Instruments - Adding More Instruments

- Adding more weak instruments will increase the bias of 2SLS
  - → By adding further instruments without predictive power, the first stage F-statistic goes toward zero and the bias increases
  - → We will see this more closely when we cover the leniency design
- If the model is "just identified" – mean the same number of instrumental variables as there are endogenous covariates – weak instrument bias is less of a problem

Weak instrument problem

- After Angrist and Krueger study, there were new papers highlighting issues related to weak instruments and finite sample bias
- Key papers are Nelson and Startz (1990), Buse (1992), Bekker (1994) and especially Bound, Jaeger and Baker (1995)
- Bound, Jaeger and Baker (1995) highlighted this problem for the Angrist and Krueger study.

Bound, Jaeger and Baker (1995)

Remember, AK present findings from expanding their instruments to include many interactions (i.e., saturated model)

1. Quarter of birth dummies $\rightarrow$ 3 instruments
2. Quarter of birth dummies + (quarter of birth) $\times$ (year of birth) + (quarter of birth) $\times$ (state of birth) $\rightarrow$ 180 instruments

So if any of these are weak, then the approximate bias of 2SLS gets worse

Adding instruments in Angrist and Krueger

	(1) OLS	(2) IV	(3) OLS	(4) IV
Coefficient	.063	.142	.063	.081
	(.000)	(.033)	(.000)	(.016)
F (excluded instruments)		13.486		4.747
Partial R^2 (excluded instruments, $\times 100$)		.012		.043
F (overidentification)		.932		.775
	Age Control Variables			
Age, Age^2	x	x		
9 Year of birth dummies			x	x
	Excluded Instruments			
Quarter of birth		x		x
Quarter of birth $\times$ year of birth				x
Number of excluded instruments		3		30

Adding more weak instruments reduced the first stage F-statistic and increases the bias of 2SLS. Notice its also moved closer to OLS.

Adding instruments in Angrist and Krueger

	(1) OLS	(2) IV
Coefficient	.063	.083
	(.000)	(.009)
F (excluded instruments)		2.428
Partial R^2 (excluded instruments, $\times 100$)		.133
F (overidentification)		.919
Age Control Variables		
Age, Age^2		
9 Year of birth dummies	x	x
Excluded Instruments		
Quarter of birth		x
Quarter of birth $\times$ year of birth		x
Quarter of birth $\times$ state of birth		x
Number of excluded instruments		180

More instruments increase precision, but drive down F, therefore we know the problem has gotten worse

IV advice: Weak instruments

- Excellent review by Keane and Neal (2021) “A Practical Guide to Weak Instruments”
- Stock, Wright and Yogo (2002) found that F statistics on the excludability of the instrument from the first stage greater than 10 performed well in Monte Carlos with homoskedasticity, but 2SLS is has poor properties here
  - → Under powered
  - → Artificially low standard errors when endogeneity is severe
  - → This causes t-tests to be misleading

IV advice: Weak instruments

- Anderson-Rubin greatly alleviate this problem and should be used even with very strong instruments provided the first-stage F is well above 10 (Lee, et al. 2020 say 104.7)
- Higher thresholds are recommended, and even then robust tests are suggested unless F is in the thousands
- Keane and Neal (2021) write, "to avoid over-rejecting the null when β_{2SLS} is shifted in the direction of the OLS bias, one should rely on the Anderson-Rubin test rather than the t-test even when the first-stage F-statistic is in the thousands."
- Good news – in your world with massive industry data, first stage issues could be less problematic

Heteroskedastic DGP

- Assessing acceptable first stage F statistics means in practice considering the impact of heteroskedasticity
- With multiple instruments, it is inappropriate to use either a conventional or heteroskedasticity robust F-test to gauge instrument strength
- Andrews, et al. (2019) suggest the Olea and Pflueger (2013) effective first-stage F statistic
- Single instrument just-identified case reduces to the conventional robust F and the Kleibergen and Paap (2006) Wald

Roadmap

- Intuition
  - Causal Graphs
  - Random Lightning Bugs in a Jar
  - Strangeness Principle
- Two sub IV designs
  - Externalities on Airbnb Platform
  - Limiting a user's choice set
  - Non-compliance in an RCT
- Estimators
  - Two Step
  - Weak instruments
- Heterogenous treatment effects
  - Potential outcomes
  - Assumptions
  - Local average treatment effects
- Data visualization

Internal and external validity

Familiar terms, but listen closely, as they are mainly about heterogenous treatment effects in IV context

1. Internal validity: If all assumptions hold, IV will identify a very specific average causal effect found within your data
2. External validity: This average causal effect may or may not be policy relevant for reasons I'll get into

Constant vs heterogenous treatment effects

- Constant treatment effects made things very simple because if you identified an average causal effect using IV, you estimated the ATE
- But not the case with heterogenous treatment effects, which is the context in which I interpret Angrist and Imbens work
- What parameter did we even estimate using IV when there were heterogenous treatment effects? Let's look more closely using "potential treatment" notation

Potential treatment concept

"Potential treatment status" (D^j) is like potential outcomes the thought experiment; it's not the observed treatment status D until we switch between them with the instrument's assignment

- $D_i^1 = i$'s treatment status when $Z_i = 1$
- $D_i^0 = i$'s treatment status when $Z_i = 0$

We'll represent outcomes as a function of both treatment status and instrument status. In other words, $Y_i(D_i = 0, Z_i = 1)$ is represented as $Y_i(0, 1)$

Identification

1. Stable Unit Treatment Value Assumption (SUTVA)
2. Random Assignment
3. Exclusion Restriction
4. Nonzero First Stage
5. Monotonicity

SUTVA

SUTVA with respect to IV

In the IV context, SUTVA means the **potential treatments** for any unit do not (1) vary with the instruments assigned to other units, and for each unit, (2) there are no different forms of versions of each instrument level, which lead to different potential treatments

Once you make D_i^1, D_i^0 based on a scalar, you've invoked SUTVA because this means your potential outcome is not based on other's assignment and it means there's no hidden variation in the instrument

Example: The instrument is a randomly generated draft number. When your friend, i', gets drafted, you, i, somehow get drafted too even though you didn't get assigned with your draft number

Independence assumption

Independence assumption

$$\{Y_i(D_i^1, 1), Y_i(D_i^0, 0), D_i^1, D_i^0\} \perp\!\!\!\perp Z_i$$

- Instruments are assigned independent of potential treatment status and potential outcomes
- Independence is ensured by physical randomization, but perhaps other assignments could too (e.g., alphabetized assignment)
- Example: Random draft numbers generated by a random number generator

Independence

Implications of independence: First stage measures the causal effect of Z_i on D_i:

$$\begin{aligned} E[D_i|Z_i=1] - E[D_i|Z_i=0] &= E[D_i^1|Z_i=1] - E[D_i^0|Z_i=0] \\ &= E[D_i^1 - D_i^0] \end{aligned}$$

Independence

Implications of independence: Reduced form measures the causal effect of Z_i on Y_i

$$\begin{aligned} E[Y_i|Z_i=1] - E[Y_i|Z_i=0] &= E[Y_i(D_i^1, 1)|Z_i=1] \\ &\quad -E[Y_i(D_i^0, 0)|Z_i=0] \\ &= E[Y_i(D_i^1, 1)] - E[Y_i(D_i^0, 0)] \end{aligned}$$

But independence is not enough to for this to mean we've identified the causal effect of D on Z as Z could be operating directly not "only through" the treatment – for that we need exclusion

Exclusion Restriction

Exclusion Restriction

Y(D,Z) $=$ **Y(D,Z')** for all **Z**, **Z'**, and for all **D**

- Notice how in the notation, Z is changing to Z', but D is held fixed and as a result of it being held fixed, Y does not change?
- That's the "only through" part. Any effect of Z on Y must be via the effect of Z on D.
- Recall the DAG and the *missing arrows* from Z to ν and from Z to Y directly
- **Violation example**: Your draft number causes you to go to graduate school to avoid the draft, but graduate school changes your wages, therefore exclusion is violated even though instrument was random

Exclusion restriction

- Use the exclusion restriction to define potential outcomes indexed solely against treatment status (regardless of instrument assignment):

$$\begin{aligned} Y_i^1 &= Y_i(1,1) = Y_i(1,0) \\ Y_i^0 &= Y_i(0,1) = Y_i(0,0) \end{aligned}$$

- Rewrite switching equation:

$$\begin{aligned} Y_i &= Y_i(0,Z_i) + [Y_i(1,Z_i) - Y_i(0,Z_i)]D_i \\ Y_i &= Y_i^0 + [Y_i^1 - Y_i^0]D_i \\ Y_i &= Y_i^0 + \delta_i D_i \end{aligned}$$

- Notice here that D_i will only change if the instrument assignment causes it to change, and thus the average causal effect picked up will only be for those who reply to their instrument assignment

Know your treatment and instrument assignment mechanism

People tend to target exclusion arguments when they see them, because except under very special situations like homogenous treatment effects with overidentification, they're based on untestable assumptions

Angrist and Krueger (2001) note "In our view, good instruments often come from detailed knowledge of the economic mechanism and institutions determining the regressor of interest."

You simply can't avoid the importance of deep knowledge of treatment and instrument assignment, as those are literally in the identifying assumptions (e.g., independence, exclusion)

Strong first stage

Nonzero Average Causal Effect of Z on D

$$E[D_i^1 - D_i^0] \neq 0$$

- Recall the weak instrument literature from earlier (AR, F very large)
- D^1 means instrument is turned on, and D^0 means it is turned off. We need treatment to change when instrument changes.
- Z has to have some statistically significant effect on the average probability of treatment
- Example: Check whether a high draft number makes you more likely to get drafted and vice versa
- Finally – a testable assumption. We have data on Z and D

Monotonicity

Monotonicity

Either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all $i = 1, \ldots, N$

- Recall that π_{1i} is the reduced form causal effect of the instrumental variable on an individual i's treatment status.
- Monotonicity requires that the instrumental variable (weakly) operate in the same direction on all individual units.
- "changing the instrument's value does not induce two-way flows in and out treatment" – Michal Kolesar (2013)
- Anyone affected by the instrument is affected *in the same direction* (i.e., positively or negatively, but not both).
- **Example of a violation**: People with high draft number dodge the draft but would have volunteered had they gotten a low number

Local average treatment effect

If all 1-5 assumptions are satisfied, then IV estimates the **local average treatment effect (LATE)** of D on Y:

$$\delta_{IV,LATE} = \frac{\text{Effect of Z on Y}}{\text{Effect of Z on D}}$$

Estimand

Instrumental variables (IV) estimand:

$$\begin{aligned}\delta_{IV,LATE} &= \frac{E[Y_i(D_i^1, 1) - Y_i(D_i^0, 0)]}{E[D_i^1 - D_i^0]} \\ &= E[(Y_i^1 - Y_i^0)|D_i^1 - D_i^0 = 1]\end{aligned}$$

Local Average Treatment Effect

- The LATE parameters is the average causal effect of D on Y for those whose treatment status was changed by the instrument, Z
- For example, IV estimates the average effect of military service on earnings for the subpopulation who enrolled in military service because of the draft but would not have served otherwise.
- LATE does not tell us what the causal effect of military service was for patriots (volunteers) or those who were exempted from military service for medical reasons

LATE and subpopulations

IV estimates the average treatment effect for only one of these subpopulations:

1. Always takers: My family have always served, so I serve regardless of whether I am drafted
2. Never takers: I'm a contentious objector so under no circumstances will I serve, even if drafted
3. Defiers: When I was drafted, I dodged. But had I not been drafted, I would have served. I am a man of contradictions.
4. **Compliers**: I only enrolled in the military because I was drafted otherwise I wouldn't have served

Never-Takers

$D_i^1 - D_i^0 = 0$
$Y_i(0,1) - Y_i(0,0) = 0$
By Exclusion Restriction, causal effect of Z on Y is zero.

Defier

$D_i^1 - D_i^0 = -1$
$Y_i(0,1) - Y_i(1,0) = Y_i(0) - Y_i(1)$
By Monotonicity, no one in this group

Complier

$D_i^1 - D_i^0 = 1$
$Y_i(1,1) - Y_i(0,0) = Y_i(1) - Y_i(0)$
Average Treatment Effect among Compliers

Always-taker

$D_i^1 - D_i^0 = 0$
$Y_i(1,1) - Y_i(1,0) = 0$
By Exclusion Restriction, causal effect of Z on Y is zero.

Monotonicity Ensures that there are no defiers

- Why is it important to not have defiers?
  - → If there were defiers, effects on compliers could be (partly) canceled out by opposite effects on defiers
  - → One could then observe a reduced form which is close to zero even though treatment effects are positive for everyone (but the compliers are pushed in one direction by the instrument and the defiers in the other direction)
- Monotonicity assumes there are no defiers (there are weak and strong versions of it too)

LATE is not the ATE

- IV estimates the average causal effect for those units affected by the instrument (i.e., complier causal effects)
- Work in the mid-2000s found that with continuous instruments, it could be possible to extrapolate from the LATE to the aggregate parameter (marginal treatment effect literature)
- I'll wait to discuss that literature but know it's coming and important to learn

Sensitivity to assumptions: exclusion restriction

- Someone at risk of draft (low lottery number) changes education plans to retain draft deferments and avoid conscription.
- Increased bias to IV estimand through two channels:
  - → Average direct effect of Z on Y for compliers
  - → Average direct effect of Z on Y for noncompliers multiplied by odds of being a non-complier
- Severity depends on:
  - → Odds of noncompliance (smaller → less bias)
  - → "Strength" of instrument (stronger → less bias)
  - → Effect of the alternative channel on Y

Sensitivity to assumptions: Monotonicity violations

- Someone who would have volunteered for Army when not at risk of draft (high lottery number) chooses to avoid military service when at risk of being drafted (low lottery number)
- Bias to IV estimand (multiplication of 2 terms):
  - → Proportion defiers relative to compliers
  - → Difference in average causal effects of D on Y for compliers and defiers
- Severity depends on:
  - → Proportion of defiers (small → less bias)
  - → "Strength" of instrument (stronger → less bias)
  - → Variation in effect of D on Y (less → less bias)

IV with covariates

- What if you think you need to control for covariates? Can't you just control for it in your 2SLS specification? But how?
- Blandhol, et al. (2022) as well as Stoczynski (2021) bring up some issues with typical 2SLS specifications with covariates
- This is a decently sized literature going back at least to Abadie (2003), Frolich (2007), as well as to a degree Imbens and Angrist (1995)
- The punchline is that controlling for covariates can be somewhat hazardous when using 2SLS

Saturated regression models

- Remember Angrist and Krueger's QoB instrument specification where they interacted Z with region of birth and year of birth? This was almost entirely a saturated model (they didn't interact Z with age I don't think)
- Saturated models are the full set of interactions on all discrete covariates as well as each one independently

  "Saturated regression models are regression models with discrete explanatory variables, where the model includes a separate parameter for all possible values taken on by the explanatory variables." (Angrist and Pischke 2009, p. 48-49)

Identification with covariates and 2SLS

- We have to modify independence and exclusion (which isn't all that surprising), but we also have to introduce new types of first stage and common support assumptions
- Assume conditional independence since we're controlling for X, exclusion conditional on X, positive correlation with covariates and treatment
- Common support assumptions: there are units with $Z = 1$ across distribution of X and units in both treatment and control across X
- The last two parts of that requires that there is variation in the instrument as well as a distinct number of compliers and defiers at every value of covariates

2SLS estimand with covariates

If you assume this and monotonicity, then Sloczyn'ski (2021), Angrist and Imbens (1995) and Kolesar 2013) shows that a saturated 2SLS model identifies a convex combination of conditional LATEs with weights equal to the conditional variance of the first stage

$$\delta_{2SLS} = \frac{E[\sigma^2(X) \cdot \tau(X)]}{E[\sigma^2(X)]}$$

where σ^2 is $E\Big[(E[D|X,Z] - E[D|X])^2|X\Big]$ and $\tau(x)$ is the conditional LATE. Notice the variances weighting the conditional LATEs

Covariates in 2SLS models

- So the Angrist and Imbens (1995) approach to interacting the instrument with all possible dummies combining covariates in a saturated 2SLS model is not only sufficient to recover weighted combination of LATEs – it's also necessary
- But though Angrist and Imbens (1995) did it this way, it's very rare to see covariates controlled for in a nonparametric way like this because overidentification with 2SLS raises issues with weak instruments

  " Bound, Jaeger and Baker (1995) write, "[our results] indicate that the common practice of adding interaction terms as excluded instruments may exacerbate the [weak instruments] problem."
- Another possibility is to run first stages for every value of X combination (these get huge quickly) and weight them so as to avoid curse of dimensionality issues

Saturate and weight

- Only one that isn't is the saturate and weight method which requires interacting dummies for values of continuous X_k with all $X_{k'}$ which in a finite sample runs into curse of dimensionality
- Some cells won't have any variation in Z conditional on X
- They show it's necessary and sufficient for estimate to be weighted average over all individual LATEs, otherwise negative weights enter

Covariates going forward

- When all covariates are discrete, then the Angrist and Imbens (1995) saturated method recovers convex combination of conditional LATEs
- 2SLS will in general reflect treatment effects for compliers and always/never takers, and some of the treatment effects for the always/never-takers will necessarily be negatively weighted
- Sloczyn'ski (2021) introduces a new procedure called "reordered IV" but it doesn't guarantee that the resulting estimand will be similar to the unconditional LATE
- There are a variety of alternatives to 2SLS like Abadie (2003), which uses a propensity score (for Z) to construct "kappa weights"

Roadmap

- Intuition
  - Causal Graphs
  - Random Lightning Bugs in a Jar
  - Strangeness Principle
- Two sub IV designs
  - Externalities on Airbnb Platform
  - Limiting a user's choice set
  - Non-compliance in an RCT
- Estimators
  - Two Step
  - Weak instruments
- Heterogenous treatment effects
  - Potential outcomes
  - Assumptions
  - Local average treatment effects
- Data visualization

Practical advice

- Before I conclude, I wanted to just make a strong suggestion to you
- It's very easy for IV to become a black box, but no one is helped by that
- There's also recent evidence that IV papers show signs of publication bias with a large spike in p-values at 0.05 (unlike RCT and RDD)
- So in addition to all I said, I'd like to make some aesthetic suggestions

IV advice: Pictures

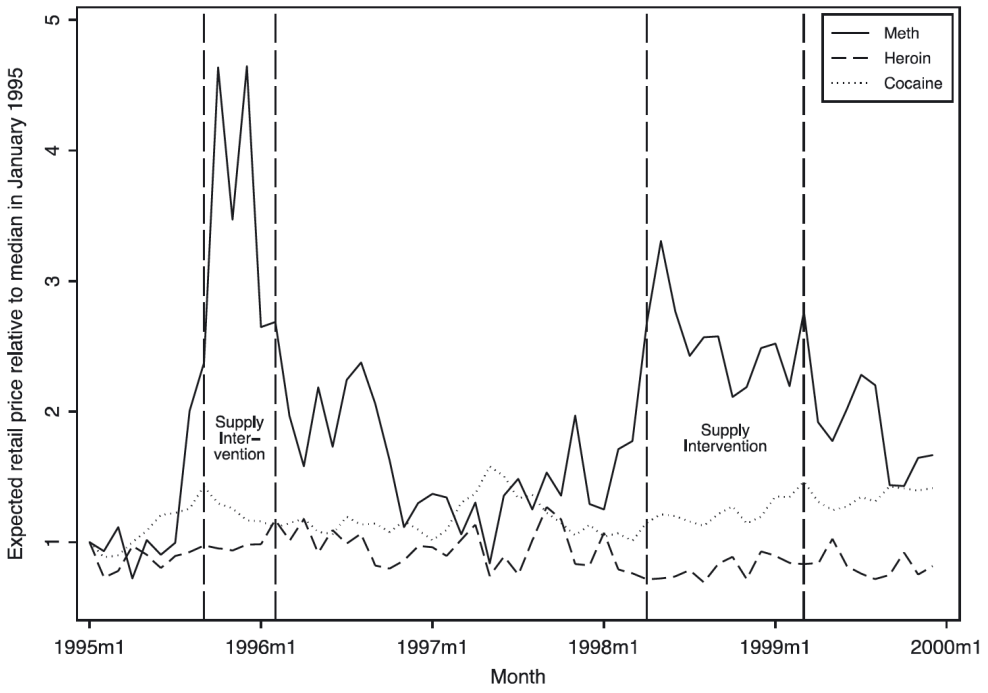
Present your main results in beautiful pictures

- Show pictures of the first stage. If you can’t see it there, then weak instruments are likely
- You can’t show a second stage with raw data, so instead show pictures of the reduced form. Same as above

IV advice: Pictures

FIGURE 3

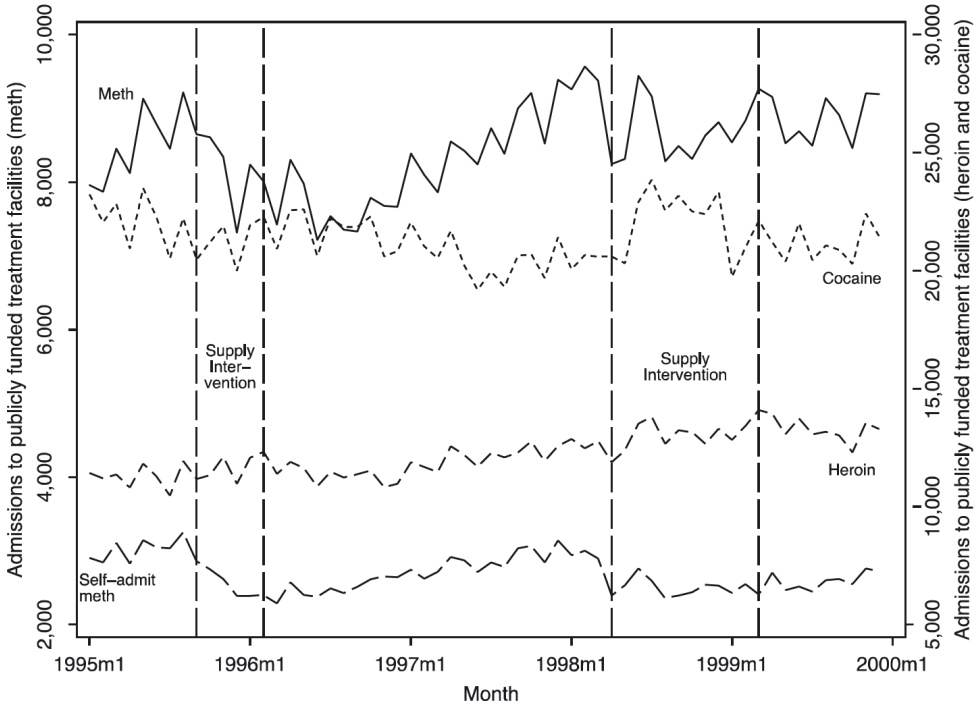
Ratio of Median Monthly Expected Retail Prices of Meth, Heroin, and Cocaine Relative to Their Respective Values in January 1995, STRIDE, 1995–1999



IV advice: Pictures

FIGURE 5

Total Admissions to Publicly Funded Treatment Facilities by Drug and Month, Selected States, Whites, TEDS, Seasonally Adjusted, 1995–1999

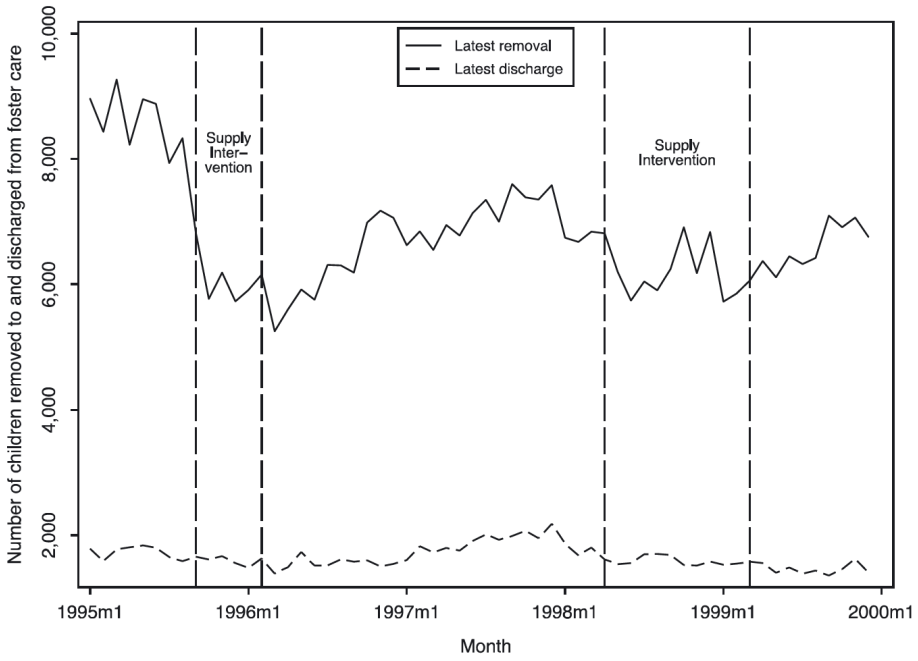


Notes: Authors' calculations from TEDS. Arizona, the District of Columbia, Kentucky, Mississippi, West Virginia, and Wyoming are excluded because of poor data quality. Patients can report the use of more than one drug.

IV advice: Pictures

FIGURE 4

Number of Children Removed to and Discharged from Foster Care in a Set of Five States by Month, AFCARS, Seasonally Adjusted, 1995–1999



Sources: Authors' calculations from AFCARS. This figure contains AFCARS data only from California, Illinois, Massachusetts, New Jersey, and Vermont. These states form a balanced panel through the entire sample period.

Looking forward

- IV is an old method and quite powerful; when conditions hold, it can recover the LATE
- Heterogeneity has made much of this challenging, be it jumping over monotonicity and exclusion, addressing weak first stages which now must be even stronger, and these issues around covariates and 2SLS
- I remain optimistic – literature on MTE shows amazingly we can recover aggregate parameters with instrument intensity, as well as information about compliers (which I couldn't cover due to time constraints)
- My hope is that by learning, also, about the leniency design, Amazon employees might see them in more places
- Thank you for having me!