Causal Inference I

Mixtape Session

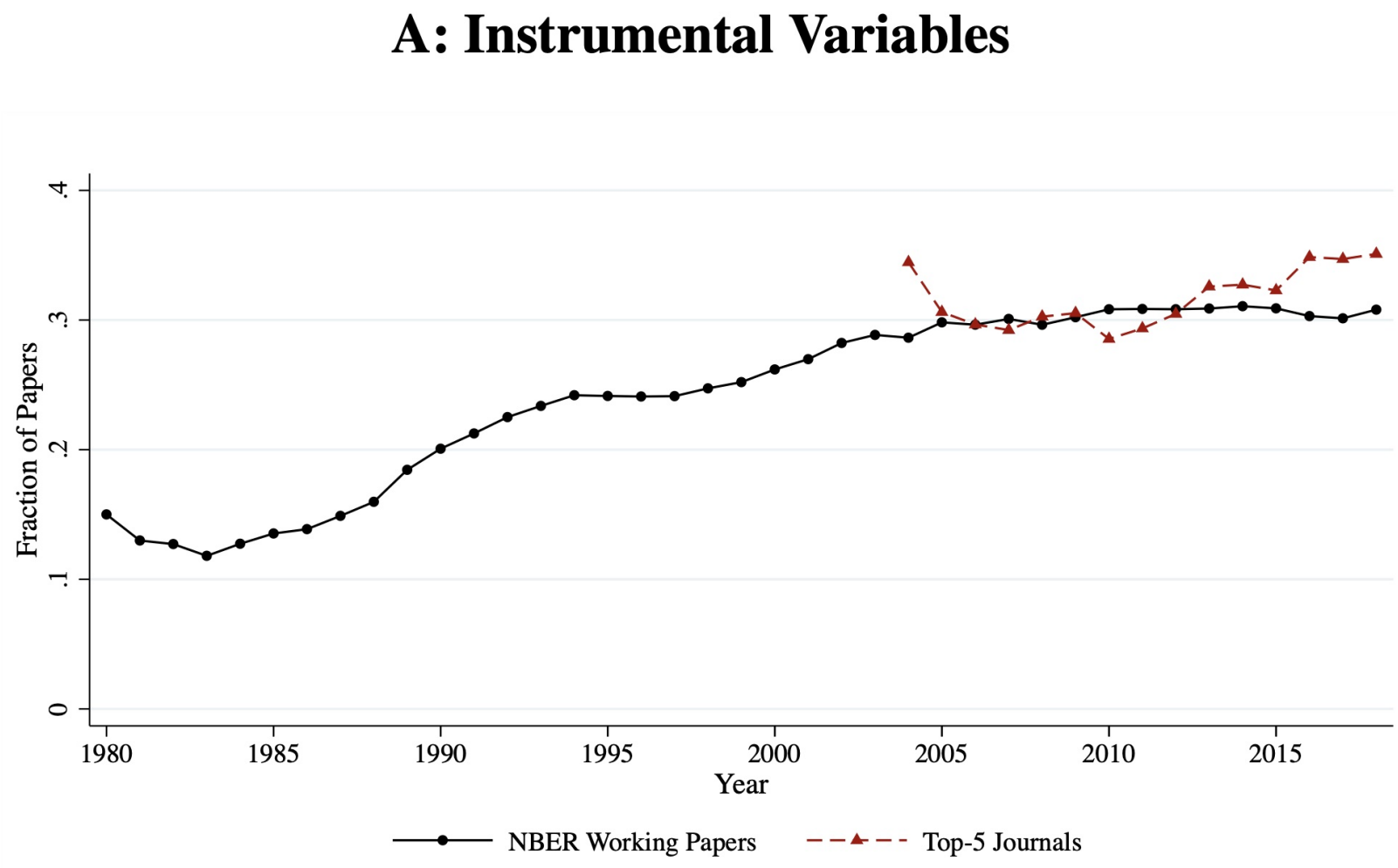
Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
  - Local average treatment effects
- Application
  - Data visualization and necessary evidence
  - Leniency design
  - Marginal Treatment Effects
  - Covariates
  - Price elasticity of demand
- Conclusion

Instrumental variables

- When you cannot observe all the confounders (or don't know them), or the concept of CIA doesn't seem plausible, then selection on observable methods do not solve the problem
- Alternative methods were developed but they themselves also have specific instances where they are suitable and when they aren't
- One popular method is the instrumental variables method which is the most popular of all designs (see next slide)

IV Popularity



What is instrumental variables?

- Often associated with “natural experiments” because an instrument is when we discovered our treatment variable randomly changed when a third seemingly inconsequential event occurred
- Natural experiments are rarely “good instruments” because they often are associated with too much destruction to be useful, but the metaphor can be helpful because we often use instruments to recreate the conditions of a randomized experiment outside the lab
- Do we find instruments or do they find us? They tend to become apparent to us when we understand what they are as otherwise they’re chameleons blending into the background

IV from my research

- Estimated the effect of growing online sex work on street prostitution arrests using commercial broadband adoption as an instrument for online sex work (Cunningham and Kendall 2010)
- Estimated the effect of parental meth abuse on foster care and child abuse/neglect using changes in meth prices caused by shortages of ephedrine and pseudoephedrine as an instrument for meth abuse (Cunningham and Finlay 2010)
- Estimated the price elasticity of demand for meth using the same instrument (Cunningham and Finlay 2014)
- Estimated the effect of being classified mentally ill in jail on suicide attempts (Seward, Cunningham, Vigliotti and Clay 2022)

Uses of instrumental variables

- When you have unobserved confounders, IV may work
- When you have reverse causality, IV could help
- When you have variables being determined simultaneously (like in markets with supply and demand), then IV can step in
- When you run an experiment (RCT, AB tests, etc.) but not everyone obeys their treatment assignment, IV will help you
- When your treatment variable is measured with error, IV can help

My Teaching Style

1. Intuition first – what is an instrument and what do you do with it?
2. Estimation second – what are most common methods employed?
3. Interpretation – constant treatment effects versus heterogeneous treatment effects and the LATE parameter
4. Applications – price elasticity of demand, leniency

Constant versus heterogenous treatment effects

Angrist and Imbens won the Nobel Prize (with Card) in October 2021 for their work on instrumental variables which focused on interpretation when treatment effects differed across a population

- **Constant treatment effects**: you and I both benefit the same from a college degree
- **Heterogenous treatment effects**: college degree *causes* my wages to rise by 5% but caused yours to rise by 18%

We'll start with the first but then move into the second.

Instruments aren't labeled

- Your data isn't going to come with a codebook saying "instrumental variable". So how do you find it?
- Well, sometimes the researcher just *knows*
- You know instruments when you see them because you know what their general structure is, and then notice them in the events and surroundings of the thing you're studying (Angrist and Krueger 2001).

Picking a good instrument

- Typically instruments are random events that are captured with a covariate that are highly predictive of your treatment variable but are in no other way related to the outcome
- If you want to use IV, then ask yourself this questions:

  Is there any part of your treatment variable that is blown by the wind (i.e., random)? If so, is there a variable you know of that is associated with that randomness?

- In other words, is there any element in the treatment that could be construed as driven by a non-confounder (i.e., random)?

Strangeness Example

- What if I told you if the first two children born were of the same sex, then the mother is less likely to work.
- What does sex composition of first two born have to do with a mother's willingness or ability to work?

Strangeness Example

- What if I told you if the first two children born were of the same sex, then the mother is less likely to work.
- What does sex composition of first two born have to do with a mother's willingness or ability to work?
- Unclear to an intelligent layperson why it should matter

Strangeness Example

- What if I told you if the first two children born were of the same sex, then the mother is less likely to work.
- What does sex composition of first two born have to do with a mother's willingness or ability to work?
- Unclear to an intelligent layperson why it should matter which is why it is a good instrument potentially

Angrist and Evans cont.

- Many parents have both a “stopping rule” and a preference for having at least one child of each sex
  - → If a couple whose first two kids were both boys, they will often have a third, hoping to have a girl
  - → If a couple whose first two kids were both girls, they will often have a third, hoping to have a boy
  - → But if it was boy/girl or girl/boy, they will often **stop**
- Sex of your kids is essentially as good as random (around 0.51 historically chance of a boy)
- Josh Angrist and Bill Evans used sex ratio of first two kids as an instrument for family size to estimate effect of family size on whether a woman worked

Good instruments must be a bit strange

- On its face, it's puzzling that the first two kids' gender predicts labor market participation
- Instrumental variables strategies formalize this *strangeness*,
- Strangeness principle is the inference drawn by an intelligent layperson with no particular knowledge of the phenomena or background in statistics.
- Without understanding the research question (family size $\rightarrow$ maternal work), the instrument's correlation with the outcome (sex ratio of first two kids predicting maternal work) makes no sense

Sunday Candy is a good instrument

- Let's listen to a few lines from "Ultralight Beam" by Kanye West (skip to 2:30, but then it's around 3:15)
- Chance the Rapper sings the following two lines

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

- What does a song have to do with hell, or meeting Kanye to do with success?
- These are instruments, but for what and are they good ones?

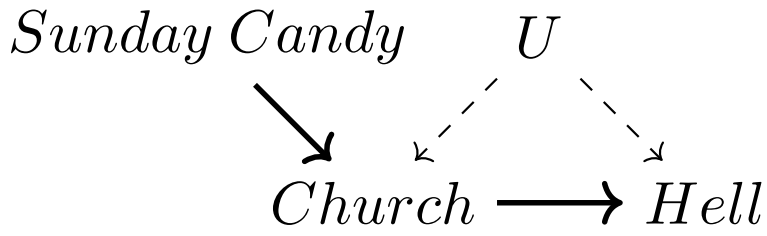
What are we missing?

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

- There must be more to this story, right?
- So what if it’s something like this

  “I made Sunday Candy
  a pastor invited me to church on Sunday,
  I’m never going to hell”

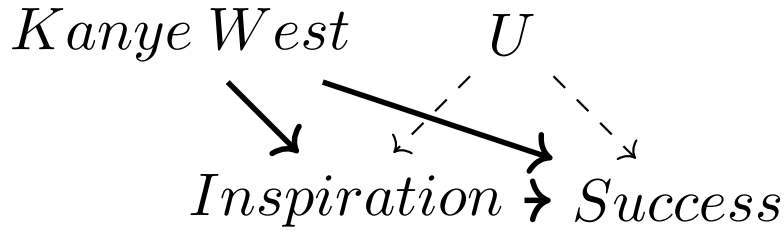
Sunday Candy DAG



Kanye West is a bad instrument

- Chance long idolized and was inspired by Kanye West – both Chicago, both very creative hip hop artists
- Kanye West is not a good instrument for Chance's inspiration, though, because Kanye West can singlehandedly make a person's career
- Kanye is not strange enough

Kanye West DAG



Some questions you need to be asking

1. Is our instrument highly correlated with the treatment? With the outcome? Can we see evidence for this?
2. Are there random reasons why our treatment changes? Why do you think that?
3. Is the instrument independent of confounders? Why do you think that?
4. Could the instrument affect outcomes directly? Why do you think that?

Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
  - Local average treatment effects
- Application
  - Data visualization and necessary evidence
  - Leniency design
  - Marginal Treatment Effects
  - Covariates
  - Price elasticity of demand
- Conclusion

Two step vs Minimum Distance

- The two-stage least squares (2SLS) estimator was developed by Theil (1953) and Basman (1957) independently
- Kolesaŕ has a helpful distinction: two step (Wald, 2 Sample IV, JIVE, UJIVE, 2SLS) vs minimum distance estimators (LIML)
- Too much to review as IV 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

Wald estimator

$$Y = \alpha + \delta S + \gamma A + \nu$$

where Y is log earnings, S is years of schooling, A is unobserved ability, and ν is the error term

- Suppose there exists a variable, Z_i, that is correlated with S_i .
- We can estimate δ with this variable, Z:

Deriving Wald

$$
\begin{aligned}
Cov(Y,Z) &= Cov(\alpha + \delta S + \gamma A + \nu, Z) \\
&= E[(\alpha + \delta S + \gamma A + \nu)Z] - E[\alpha + \delta S + \gamma A + \nu]E[Z] \\
&= \{\alpha E(Z) - \alpha E(Z)\} + \delta\{E(SZ) - E(S)E(Z)\} \\
&\quad +\gamma\{E(AZ) - E(A)E(Z)\} + E(\nu Z) - E(\nu)E(Z) \\
Cov(Y,Z) &= \delta Cov(S,Z) + \gamma Cov(A,Z) + Cov(\nu,Z)
\end{aligned}
$$

Divide both sides by $Cov(S,Z)$ and the first term becomes δ, the LHS becomes the ratio of the reduced form to the first stage, plus two other scaled terms.

Consistency

- What conditions must hold for a valid IV design?
  - → $Cov(S, Z) \neq 0$ – "first stage" exists. S and Z are correlated
  - → $Cov(A, Z) = Cov(\nu, Z) = 0$ – "exclusion restriction". This means Z that orthogonal to the factors in ν, such as unobserved ability, A, as well as the structural disturbance term, ν
- Combine A and ν into a composite error term η for simplicity
- Assuming the first stage exists and that the exclusion restriction holds, then we can estimate δ with δ_{Wald}:

$$\begin{aligned} \text{plim } \widehat{\delta}_{Wald} &= \delta + \gamma \frac{Cov(\eta, Z)}{Cov(S, Z)} \\ &= \delta \end{aligned}$$

Two Sample IV

- Wald can be implemented in exotic ways, even across datasets
  1. Dataset 1 needs information on the outcome and the instrument – Cov(Y,Z)
  2. Dataset 2 needs information on the treatment and the instrument – Cov(D,Z)
- This is known as “Two sample IV” because there are two *samples* involved, rather than the traditional one sample.
- Once we define what IV is measuring carefully, you will see why this works.

Two-stage least squares concepts

- Causal model. Your main research question:

$$Y_i = \alpha + \delta S_i + \eta_i$$

- First-stage regression. Gets the name because of two-stage least squares:

$$S_i = \gamma + \rho Z_i + \zeta_i$$

- Second-stage regression. Notice the fitted values, $\widehat{S}$:

$$Y_i = \beta + \delta \widehat{S}_i + \nu_i$$

- Reduced form regression: Y regressed onto the instrument:

$$Y_i = \psi + \pi Z_i + \varepsilon_i$$

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)

Two-stage least squares language

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

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

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

Two-stage least squares

Carry over from previous slide

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

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 $\frac{\widehat{\rho}}{\widehat{\rho}}$ (also note the subscript – we are moving now into 2SLS)

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

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

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

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)

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

Proof.

We will show that $\widehat{\delta} Cov(Y, Z) = Cov(\widehat{S}, Y)$. I will leave it to you to show that $Var(\widehat{\delta} Z) = Var(\widehat{S})$

$$
\begin{aligned}
Cov(\widehat{S}, Y) &= E[\widehat{S}Y] - E[\widehat{S}]E[Y] \\
&= E(Y[\widehat{\rho} + \widehat{\delta} Z]) - E(Y)E(\widehat{\rho} + \widehat{\delta} Z) \\
&= \widehat{\rho} E(Y) + \widehat{\delta} E(YZ) - \widehat{\rho} E(Y) - \widehat{\delta} E(Y)E(Z) \\
&= \widehat{\delta}[E(YZ) - E(Y)E(Z)] \\
Cov(\widehat{S}, Y) &= \widehat{\delta} Cov(Y, Z)
\end{aligned}
$$

□

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
- Our previous slides showed that 2SLS and the ratio of reduced form and first stage were equivalent though
- By using the fitted values of the endogenous regressor from the first stage regression, our regression now uses *only* the quasi-random variation in the treatment due to the instrumental variable itself (only the random parts of schooling remain)

Finite sample problems with 2SLS

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

$$\begin{aligned} Y_i &= \alpha + \delta S_i + \eta_i \\ S_i &= \gamma + \rho Z_i + \zeta_i \end{aligned}$$

where $Cov(Z, \eta_i) = 0$ and $\rho \neq 0$.

Finite sample problems with 2SLS

Plug in covariance and write out the following:

$$
\begin{aligned}
\widehat{\delta_{2sls}} &= \frac{Cov(Z,Y)}{Cov(Z,S)} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})(Y_i-\overline{Y})}{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})(S_i-\overline{S})} \\
&= \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})Y_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})S_i}
\end{aligned}
$$

Finite sample problems with 2SLS

Substitute the causal model definition of Y to get:

$$\begin{aligned}
\widehat{\delta_{2sls}} &= \frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})\{\alpha + \delta S_i + \eta_i\}}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})S_i} \\
&= \delta + \frac{\frac{1}{n}(Z_i - \overline{Z})\eta_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i - \overline{Z})S_i} \\
&= \delta + \text{"small if } n \text{ is large"}
\end{aligned}$$

Where did the first term go? Why did the second term become δ? Why might the second term not be zero even under exclusion?

Intuition of 2SLS

- Two stage least squares is nice because in addition to being an estimator, there's also great intuition contained in it which you can use as a device for thinking about IV more generally.
- The intuition is that 2SLS estimator replaces S with the fitted values of S (i.e., $\widehat{S}$) from the first stage regression of S onto Z and all other covariates.
- 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

Intuition of IV in 2SLS

- . . . but think about it – that variation was there before, but was just a subset of all the variation in the regressor
- Go back to what we said in the beginning - we need the endogenous variable to have pieces that are random, and IV finds them.
- Instrumental variables therefore reduces the variation in the data, but that variation which is left is *exogenous*

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

- 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

Weak instruments

“In instrumental variables regression, the instruments are called weak if their correlation with the endogenous regressors, conditional on any controls, is close to zero.” – Andrews, Stock and Sun (2018)

Weak instruments

- Whereas exclusion restriction is not testable, the non-zero first stage
- Weak instruments 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
- This brought into sharp focus with Angrist and Krueger (1991) quarter of birth study and some papers that followed

My March 2022 Interview with Angrist

Before we dive into the paper, though, let's listen to Angrist discuss the history

https://youtu.be/ApNtXe-JDfA?t=2348

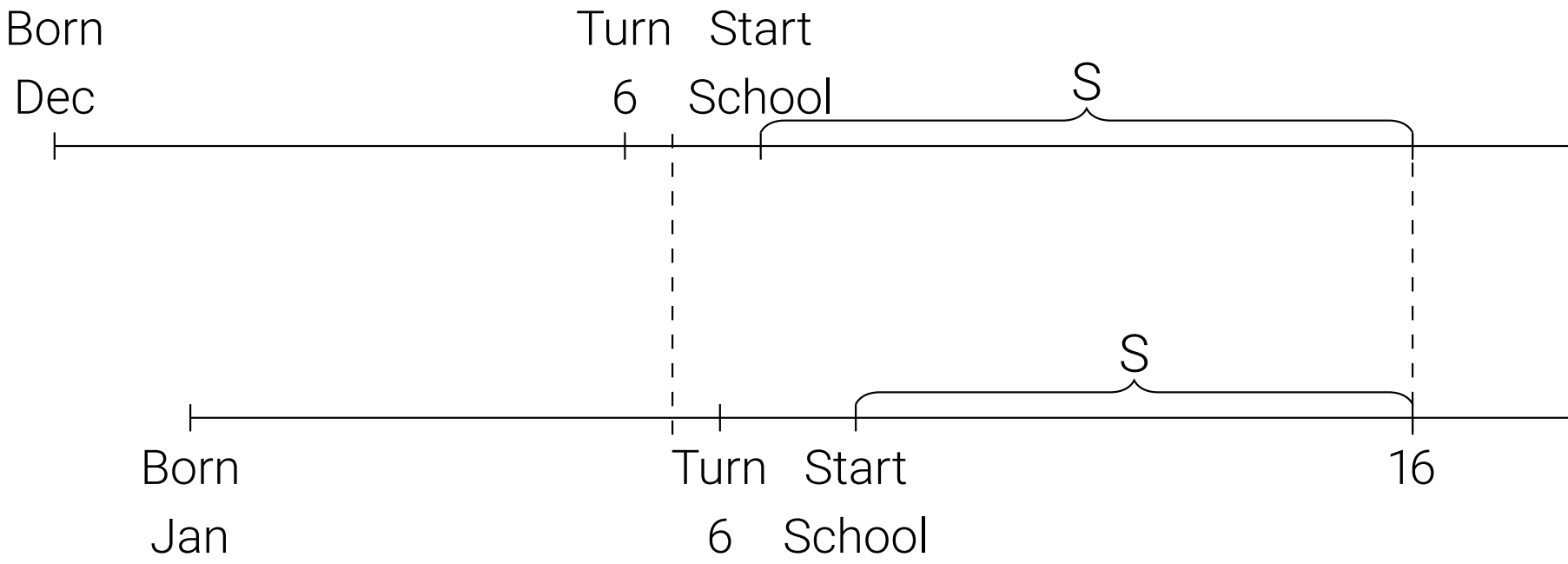
Somewhat inspiring to hear how Angrist reframed the weak instrument problem which his paper with Krueger brought into crisp focus

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 "strangeness principle" - why would birth quarter cause earnings in the reduced form?

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):



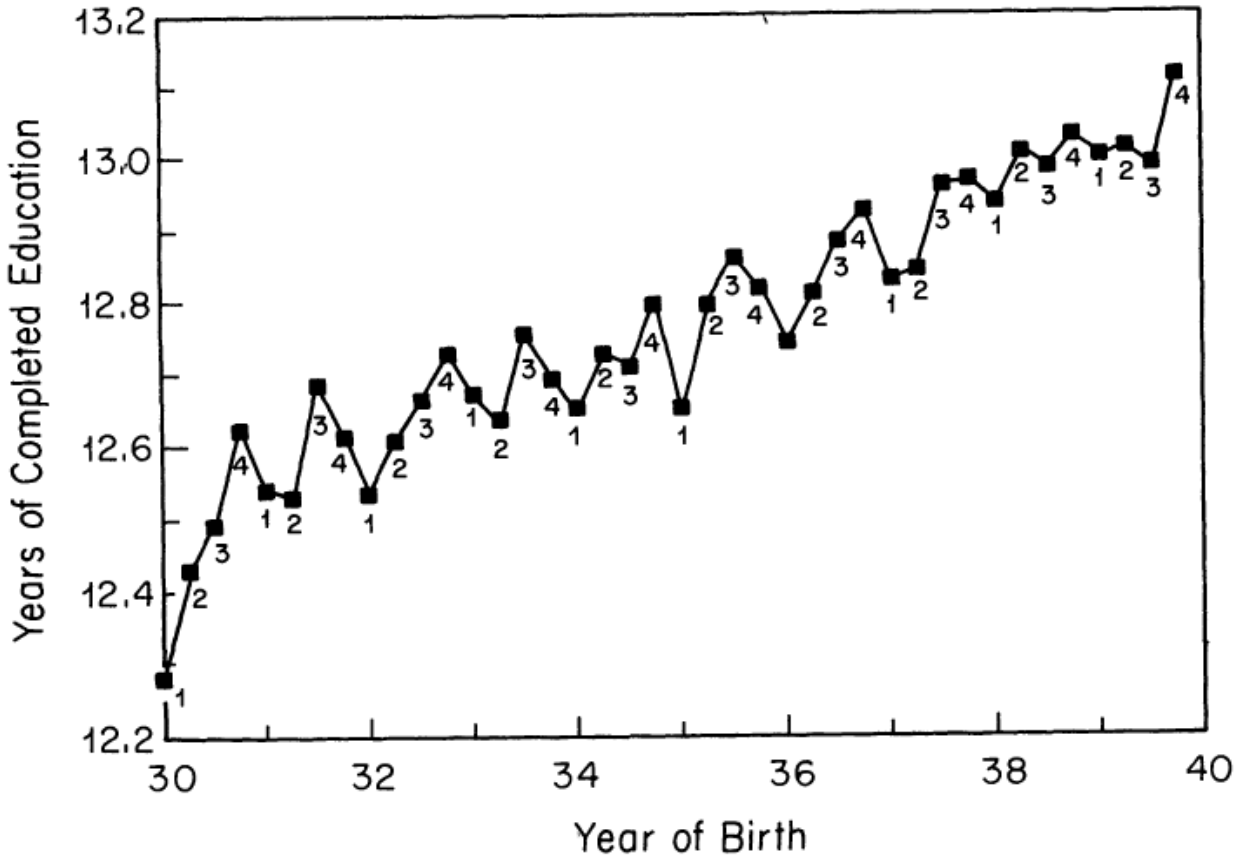
If you’re born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

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

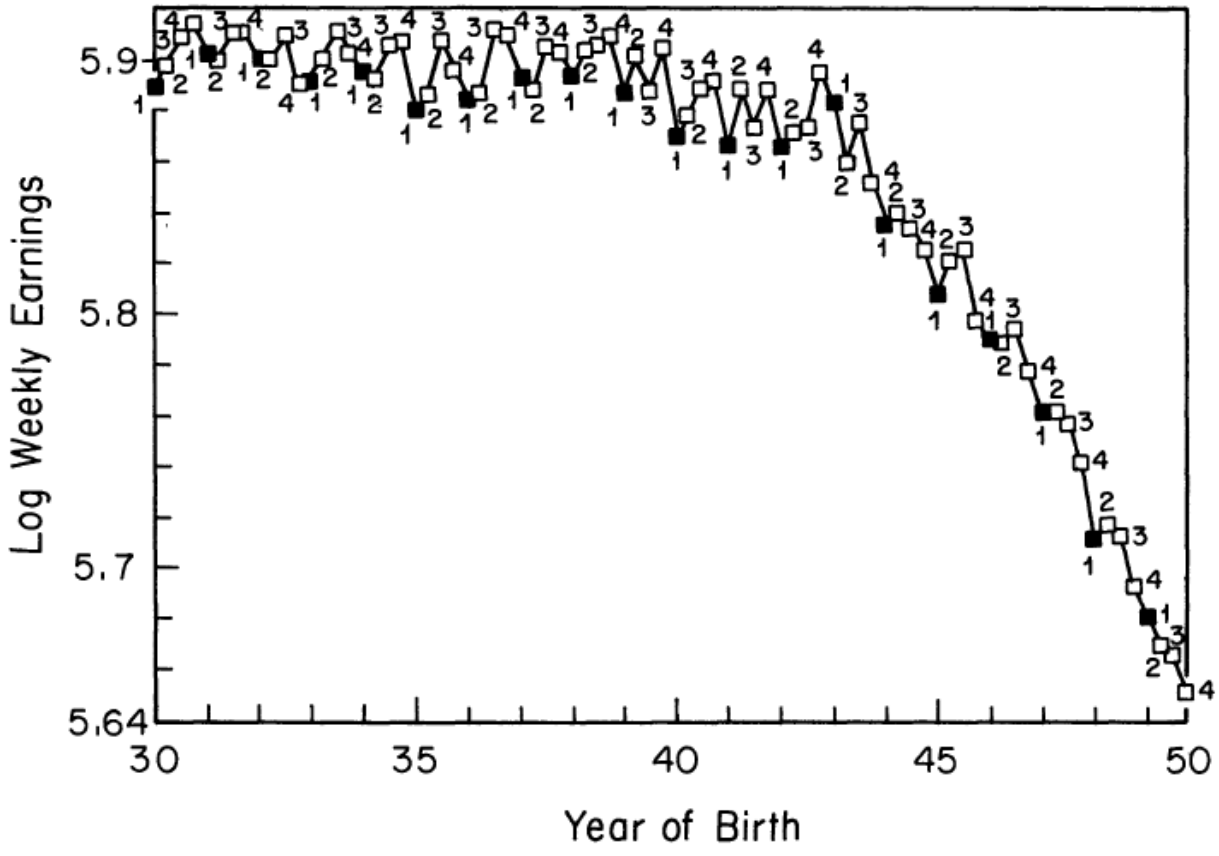
First Stage

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 sample analog of the Wald estimator that adjusts for covariates:

$\frac{\pi_{21}}{\pi_{11}}$

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 initial 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 (including all controls X), but the fitted values are from the new first stage

$$Y_i = X\pi + \delta\widehat{S}_i + \epsilon$$

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]

180 instruments

- To improve precision in their two stage least squares model, they include more instruments (causes 40 percent reduction in standard errors in 2SLS)
- More instruments can increase variation in the predicted schooling variable, lowering standard errors and tightening confidence intervals
- Three QoB dummies interacted with 50 state-of-birth dummies plus 3 QoB dummies interacted with 9 year-of-birth dummies (180 instruments)
- Includes 50 state-of-birth dummies so variability in education in 2SLS is solely due to differences in seasons of birth and this is allowed to vary by state and birth year for the first time

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, plus introduces modern notion of seeking “plausibly exogenous instruments”
- But in the early 1990s, a number of papers showed that IV can be *severely* biased with weak instruments and many instruments for one endogenous variable
- 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 from causal model and η_i from first stage model are correlated, then OLS estimated $\widehat{\beta}_{OLS}$ in causal model is biased
- To show the bias of OLS, take the population mean difference in β minus estimated β_{OLS}:

$$E[\widehat{\beta}_{OLS} - \beta] = \frac{Cov(\nu, X)}{Var(X)} = \frac{\sigma_{\nu\eta}}{\sigma_{\eta}^2}$$

- Our hope is that with 2SLS, we can drive this bias to zero in the finite sample and have a reasonably unbiased estimate of β

Weak instruments and 2SLS bias towards OLS

- Strong instruments shrink the bias term, $\frac{\sigma_{\nu\eta}}{\sigma_\eta^2}$, to an inconsequential scaled value (but cannot go to zero)
- 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" as well as Andrews, Stock and Sun (2018)
- 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
  - $\rightarrow$ Under powered
  - $\rightarrow$ Artificially low standard errors when endogeneity is severe
  - $\rightarrow$ This causes t-tests to be misleading

IV advice: Weak instruments

"In the leading case with a single endogenous regressor, we recommend that researchers judge instrument strength based on the effective F-statistic of Montiel Olea and Pflueger (2013). If there is a single instrument, we recommend reporting identification robust Anderson-Rubin confidence intervals. These are effective regardless of the strength of the instruments, and so should be reported regardless of the value of the first stage F. Finally, if there are multiple instruments, the literature has not yet converged on a single procedure, but we recommend choosing from among the several available robust procedures that are efficient when the instruments are strong." – Andrews, Stock and Sun (2018)

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."

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-tsest 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

Constant vs heterogenous treatment effects

- IV was modeled using realized outcomes, which clouded causal inference
- But also tended to assume constant treatment effects
- When you introduced heterogenous treatment effects, IV became more complex

Some background

- October 2021's Nobel Prize in economics went to Card, Angrist and Imbens (the last two for work 1990s work on IV)
- Angrist writes a dissertation using randomized instruments (Vietnam draft), goes to Harvard, overlaps with Imbens for a year, they are mentored by Gary Chamberlain, work with Don Rubin, write their famous LATE paper
- Chamberlain recommends modifying Rubin's potential outcomes framework (instead of their original latent index modeling) and that seems to make the work more generally attractive (outside economics)
- Let's spend twenty minutes listening to them

Angrist, Imbens and Harvard

Josh Angrist on the negative results at the time (10 min)

https://youtu.be/ApNtXe-JDfA?t=1885

Guido Imbens on the reception of their work (10 min)

https://youtu.be/cm8V65AS5iU?t=799

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

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)

Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
  - Local average treatment effects
- Application
  - Data visualization and necessary evidence
  - Leniency design
  - Marginal Treatment Effects
  - Covariates
  - Price elasticity of demand
- Conclusion

Practical advice

- Before we move into applications, let's talk about pictures
- It's very easy for causal inference to become a black box, but the more it's a black box, the less people will believe your analysis
- 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)
- Pictures are crucial, but it's particular kinds of pictures you need to show for IV that I want to emphasize (not just any data visual)

Show Wald Quantities

Present your main results as Wald quantites in beautiful pictures of simple correlations even if you're estimating with 2SLS

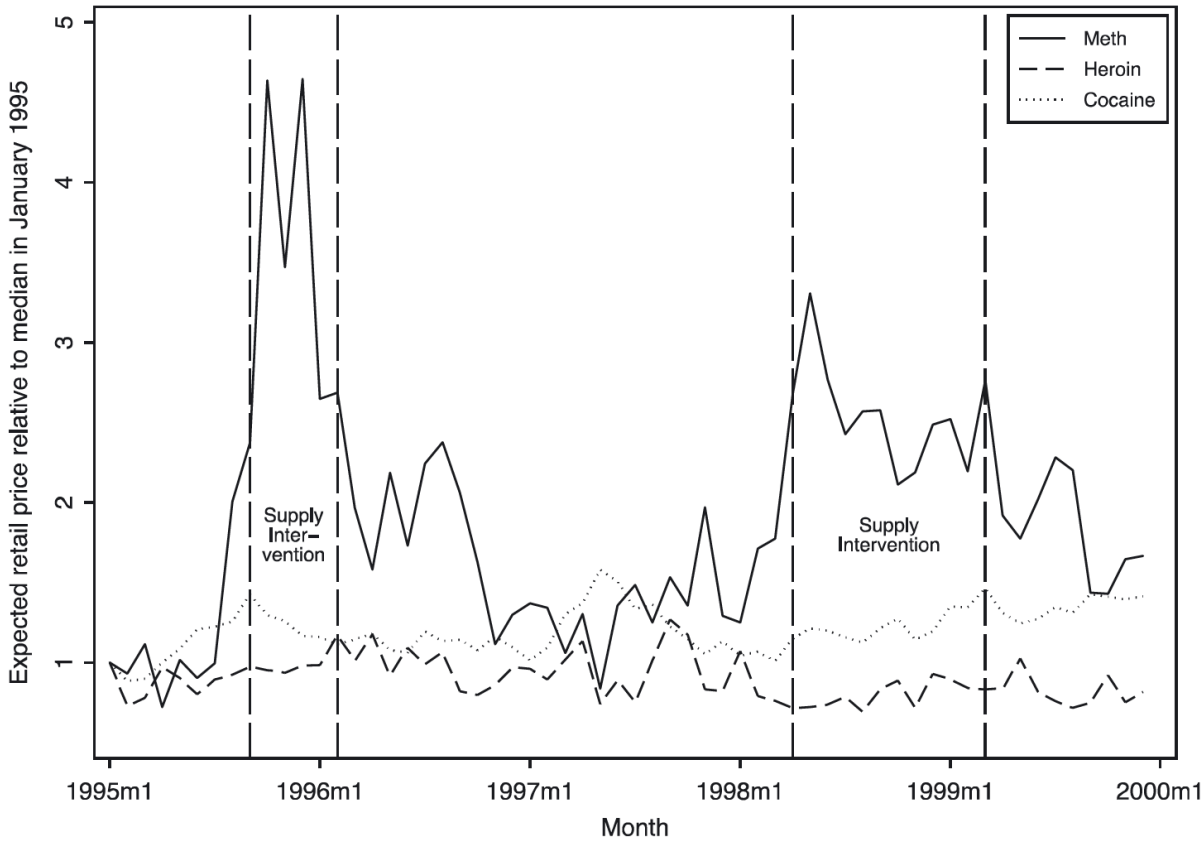
- Show pictures of the first stage. If you can't see the correlation in the first stage, you have a weak instrument problem
- Show instead pictures of the reduced form. If you can't see the correlation in the reduced form, it's likely not there.

This can be challenging as not every IV design will lend itself to easy pictures though which is why it helps if you can familiarize yourself with a range of pictures for inspiration

IV advice: Picturing my instrument

FIGURE 3

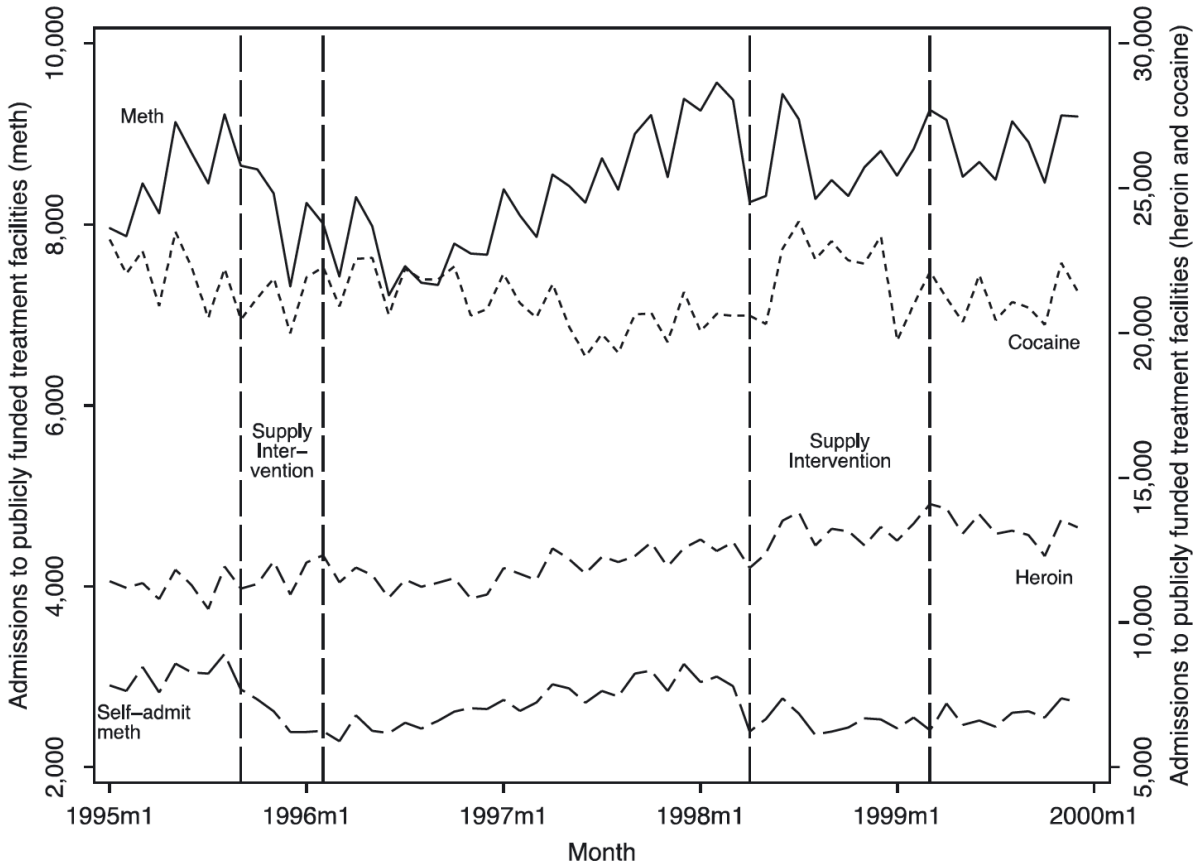
Ratio of Median Monthly Expected Retail Prices of Meth, Heroin, and Cocaine Relative to Their Respective Values in January 1995, STRIDE, 1995–1999



IV advice: Picturing the first stage

FIGURE 5

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

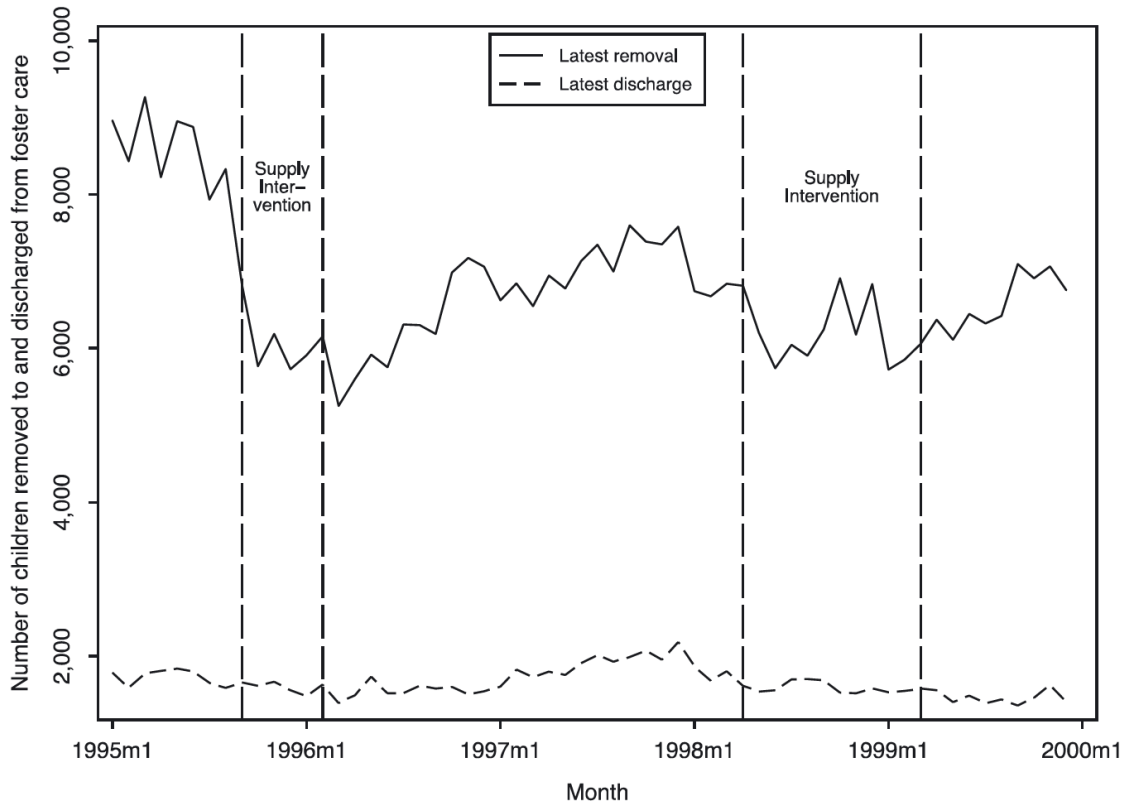


Notes: Authors' calculations from TEDS. Arizona, the District of Columbia, Kentucky, Mississippi, West Virginia, and Wyoming are excluded because of poor data quality. Patients can report the use of more than one drug.

IV advice: Picturing the reduced form

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.

Tables

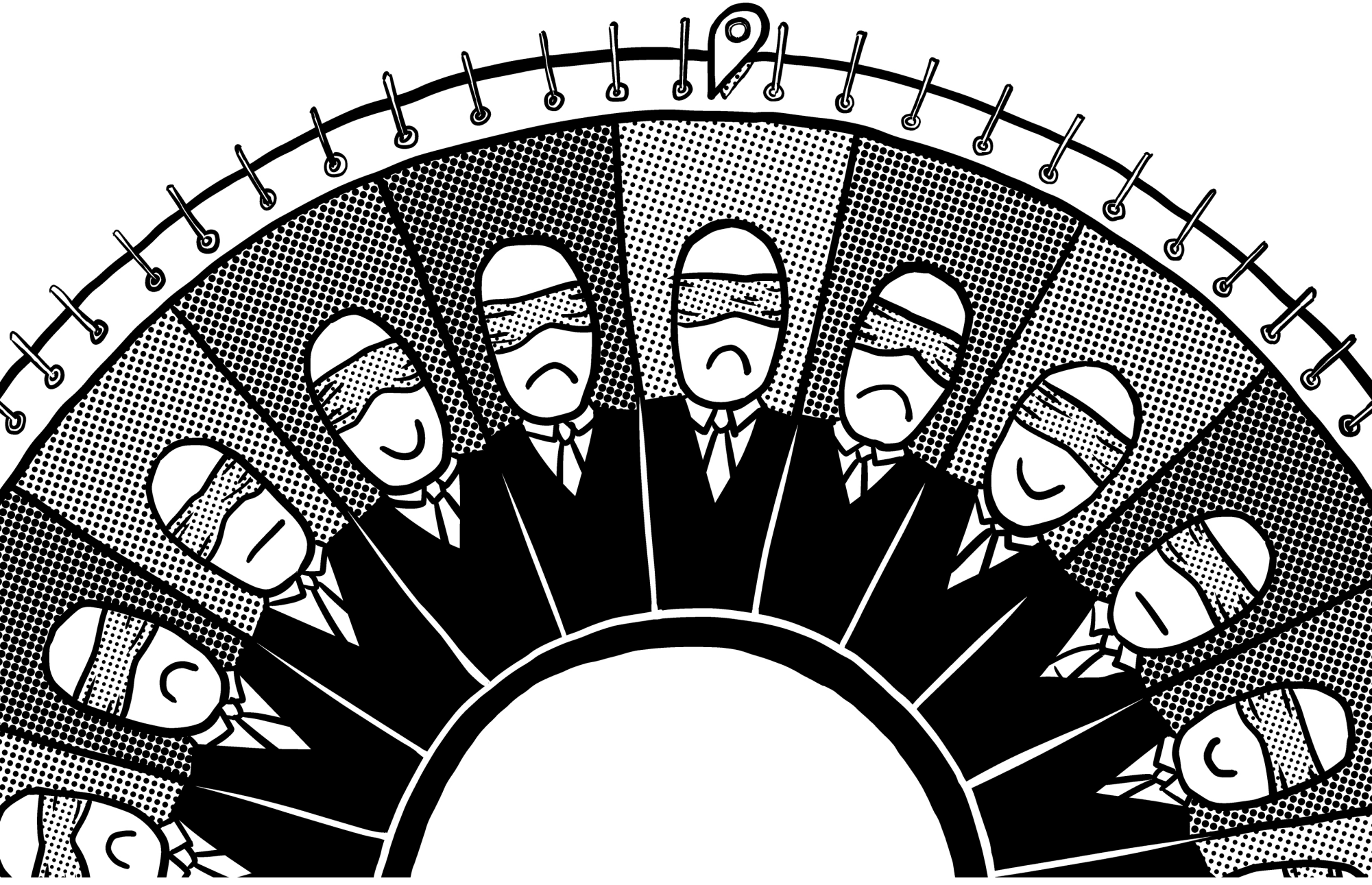
1. Naive OLS model (though with heterogeneity this may not be informative of same parameter with IV)
2. Reduced Form
3. First stage
4. Weak instrument tests
5. IV model

Table: OLS and 2SLS regressions of Log Earnings on Schooling

Dependent variable	**Log wage**	
	OLS	2SLS
educ	0.071***	0.124**
	(0.003)	(0.050)
exper	0.034***	0.056***
	(0.002)	(0.020)
black	-0.166***	-0.116**
	(0.018)	(0.051)
south	-0.132***	-0.113***
	(0.015)	(0.023)
married	-0.036***	-0.032***
	(0.003)	(0.005)
smsa	0.176***	0.148***
	(0.015)	(0.031)
First Stage Instrument		
College in the county		0.327***
Robust standard error		0.082
F statistic for IV in first stage		15.767
N	3,003	3,003

Leniency designs

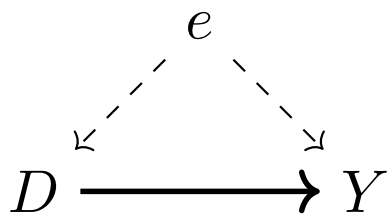
- Imagine the following:
  1. A person moves through a pipeline and hits a critical point where treatment occurs as a result of some decision-maker
  2. There are many different decision-makers and you're assigned randomly to one of them
  3. Each decision-maker differs in terms of their *leniency* in assigning the treatment
- Very popular in criminal justice bc of how often judges are randomly assigned to defendants (Kling 2006; Mueller-Smith 2015; Dobbie, et al. 2018) or even children to foster care case workers (Doyle 2007; Doyle 2008)



Juvenile incarceration

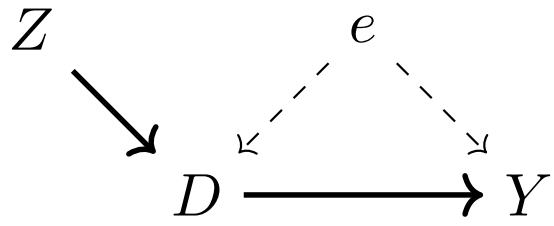
- Aizer and Doyle (2015) were interested in the causal effect of juvenile imprisonment on future crime and human capital accumulation
- Extremely important policy question given the US has the world's highest incarceration rate and prison population of any country in the world by a significant margin (500 prisoners per 100,000, over 2 million adults imprisoned, 4.8 million under supervision)
- High rates of incarceration extend to juveniles: in 2010, the stock of juvenile detainees stood at 70,792, a rate of 2.3 per 1,000 aged 10-19.
- Including supervision, US has a juvenile corrections rate 5x higher than the next highest country, South Africa

Confounding



- We are interested in the causal effect of juvenile incarceration (D) on life outcomes, like adult crime and high school completion
- But youth *choose* to commit crimes, and that choice may be due to unobserved criminogenic factors like poverty or underlying criminal propensities which are themselves causing those future outcomes

Leniency as an instrument



- Aizer and Doyle (2015) propose an instrument - the propensity to convict by the judge the youth is randomly assigned
- If judge assignment is random, and the various assumptions hold, then the IV strategy identifies the local average treatment effect of juvenile incarceration on life outcomes

The Main Idea

- “Plausibly exogenous” variation in juvenile detention stemming from the random assignment of cases to judges who vary in their sentencing
- Consider two juveniles randomly assigned to two different judges with different incarceration tendencies (Scott and Bob)
- Random assignment ensures that differences in incarceration between Scott and Bob are due to the judge, not themselves, because remember, they’re identical

Data

- 35,000 juveniles administrative records over 10 years who came before a juvenile court in Chicago (Juvenile Court of Cook County Delinquency Database)
- Data were linked to public school data for Chicago (Chicago Public Schools) and adult incarceration data for Illinois (Illinois Dept. of Corrections Adult Admissions and Exits)
- They wanted to know the effect of juvenile incarceration on high school completion (2nd data needed) and adult crime (3rd data needed) using randomized judge assignment (1st data needed)
- They need personal identifying information in each data set to make this link (i.e., name, DOB, address)

Preview of findings

- Juvenile incarceration decreased high school graduation by 13 percentage points (vs. 39pp in OLS)
- Increased adult incarceration by 23 percentage points (vs. 41pp in OLS)
- Marginal cases are high risk of adult incarceration and low risk of high school completion as a result of juvenile custody
- Unlikely to ever return to school after incarcerated, but when they do return, they are more likely to be classified as special ed students, and more likely to be classified for special ed services due to behavioral/emotional disorders (as opposed to cognitive disability)

“Plausibly” exogenous

- Very common in these studies for the assignment to some decision-maker to be *arbitrary* but not clearly random (i.e., not random no. generator)
- In this case, juveniles charged with a crime are assigned to a calendar corresponding to their neighborhood and calendars have 1-2 judges who preside over them
- 1/5 of hearings are presided over by judges who cover the calendar when the main judge can’t, known as swing judges
- Judge assignment is a function of the sequence with which cases happen to enter into the system and judge availability that is set in advance
- No scope for which judge you see first; conversations with court administrators confirm its random

Structural equation

$$Y_i = \beta_0 + \beta_1 JI_i + \beta_2 X_i + \varepsilon_i$$

where X_i is controls and ε_i is an error term. In this, juvenile incarceration is likely correlated with the error term.

This is the "long" causal model. But note, from the prior DAG, we cannot control for e because it is unobserved. But it is confounding the estimation of juvenile incarceration's effect on outcomes.

Incarceration Propensity as an Instrument

- The instrument is based on the randomized judge equalling the propensity to incarcerate from the randomly assigned judge
- "Leave-one-out mean"

$$Z_{j(i)} = \left(\frac{1}{n_{j(i)} - 1}\right)\left(\sum_{k \neq i}^{n_{j(i)}-1} \widetilde{JI}_k\right)$$

- The $n_{j(i)}$ terms is the total number of cases seen by judge k, and $\widetilde{JI}_k$ is equal to 1 if the juvenile was incarcerated during their first case
- Thus the instrument is the judge's incarceration among first cases based on all their other cases
- It's basically a judge fixed effect given the likelihood two judges have precisely the same propensity is small

Information about the instrument

- There are 62 judges in the data, and the average number of initial cases per judge is 607
- Substantial variation in the data - raw measure ranges from 4% to 21%
- Residualized measure based on controls still has substantial variation from 6% to 18%
- Variation comes from two sources: variation among the regular (nonswing) judges (80% of cases) and variation from the swing judges (20% of cases)

Distribution of IV

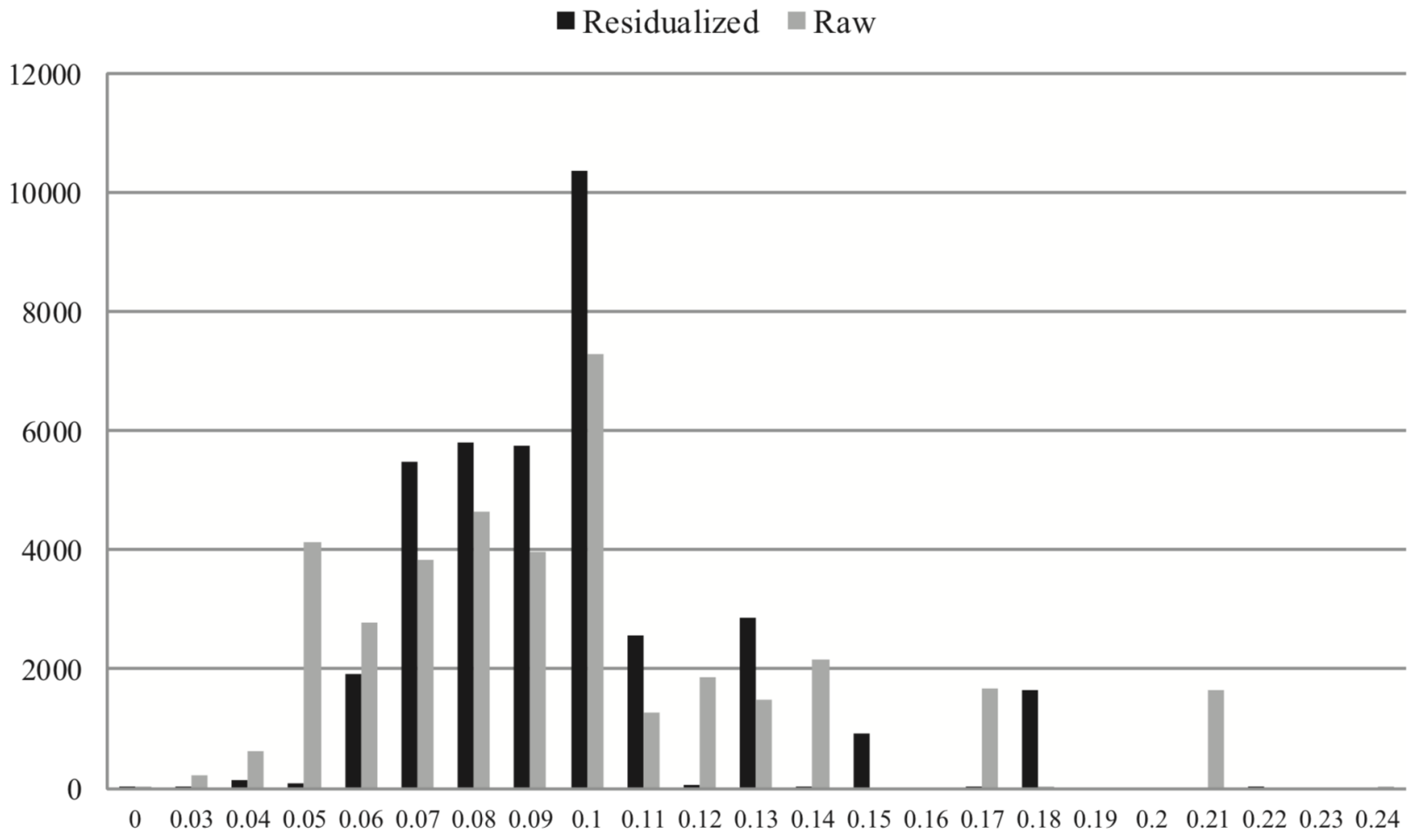


FIGURE I

Distribution of Z: Judge Incarceration Rate

Balance test

TABLE II

INSTRUMENT VERSUS JUVENILE CHARACTERISTICS

	Z distribution				
	Bottom tercile	Middle tercile	Top tercile	Middle vs. bottom p-value	Top vs. bottom p-value
Z: first judge's leave-out mean incarceration rate in first cases	0.062	0.094	0.147	(.000)	(.000)
Juvenile characteristics					
Male	0.827	0.830	0.833	(.561)	(.311)
African American	0.724	0.737	0.742	(.096)	(.249)
Hispanic	0.189	0.176	0.172	(.061)	(.272)
White	0.078	0.079	0.078	(.833)	(.957)
Other race/ethnicity	0.009	0.008	0.007	(.352)	(.345)
Special education	0.241	0.237	0.252	(.549)	(.130)
U.S. census tract poverty rate	0.264	0.265	0.265	(.572)	(.696)
Age at offense	14.8	14.8	14.8	(.437)	(.434)
P(Juvenile incarceration | X)	0.219	0.221	0.220	(.251)	(.516)
Observations	37,692				

First stage

TABLE III

FIRST STAGE

	(1)	(2)	(3)
Dependent variable: juvenile incarcerations		OLS	
First judge's leave-out mean incarceration rate among first cases	1.103	1.082	1.060
	(0.102)	(0.095)	(0.097)
Demographic controls	No	Yes	Yes
Court controls	No	No	Yes
Observations	37,692		
Mean of dependent variable	0.227		

High school completion

TABLE IV

JUVENILE INCARCERATION AND HIGH SCHOOL GRADUATION

	Dependent variable: graduated high school						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample			Juvenile court sample			
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	−0.389	−0.292	−0.391	−0.088	−0.073	−0.108	−0.125
	(0.0066)	(0.0065)	(0.0055)	(0.0043)	(0.0041)	(0.044)	(0.043)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440,797	440,797	420,033	37,692			
Mean of dependent variable	0.428	0.428	0.433	0.099			

Adult crime

TABLE V

JUVENILE INCARCERATION AND ADULT CRIME

	Dependent variable: entered adult prison by age 25						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Full CPS sample			Juvenile court sample			
	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
Juvenile incarceration	0.407	0.350	0.219	0.200	0.155	0.260	0.234
	(0.0082)	(0.0064)	(0.013)	(0.0072)	(0.0073)	(0.073)	(0.076)
Demographic controls	No	Yes	Yes	No	Yes	No	Yes
Court controls	N/A	N/A	N/A	No	Yes	No	Yes
Observations	440797	440797	420033	37692			
Mean of dependent variable	0.067	0.067	0.057	0.327			

Crime type

TABLE VI

JUVENILE INCARCERATION AND ADULT CRIME TYPE

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent variable: entered adult prison by age 25 for crime type					
	Homicide			Violent		
	OLS	OLS	2SLS	OLS	OLS	2SLS
Juvenile incarceration	0.051	0.021	0.035	0.138	0.061	0.149
	(0.0031)	(0.0030)	(0.030)	(0.0046)	(0.0050)	(0.041)
Sample	Full CPS	Juvenile court	Juvenile court	Full CPS	Juvenile court	Juvenile court
Mean of dep. var.: JI = 0	0.008	0.043	0.043	0.024	0.121	0.121
Observations	440,797	37,692	37,692	440,797	37,692	37,692
	Property			Drug		
Juvenile incarceration	0.079	0.047	0.142	0.183	0.078	0.097
	(0.0040)	(0.0038)	(0.044)	(0.011)	(0.0068)	(0.052)
Sample	Full CPS	Juvenile Court	Juvenile Court	Full CPS	Juvenile Court	Juvenile Court
Mean of dep. var.	0.013	0.060	0.060	0.034	0.176	0.176
Observations	440,797	37,692	37,692	440,797	37,692	37,692

High school transfers

TABLE VIII

INTERMEDIATE SCHOOLING OUTCOMES: HIGH SCHOOL TRANSFERS

Dependent variable:	(1) Ever present in CPS school at least 1 year after Initial hearing		(3) Transferred to another CPS high school in years after hearing		(5) Ultimate transfer: adult correctional facility	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
Juvenile incarceration	−0.025	−0.215	0.055	−0.115	0.127	0.243
	(0.0063)	(0.069)	(0.010)	(0.243)	(0.006)	(0.060)
Mean of dependent variable	0.666		0.242		0.175	
Observations	37,692		18,195		37,692	

Developing emotional problems

TABLE IX

INTERMEDIATE SCHOOLING OUTCOMES: SPECIAL EDUCATION STATUS

	(1)	(2)	(3)	(4)	(5)	(6)
	Special education type observed in years after initial hearing					
Dependent variable:	Any Special Education		Emotional/ behavioral disorder		Learning disability	
	OLS	2SLS	OLS	2SLS	OLS	2SLS
Juvenile incarceration	−0.024	−0.003	0.027	0.133	−0.040	−0.097
	(0.004)	(0.037)	(0.003)	(0.043)	(0.004)	(0.039)
Mean of dependent variable	0.193		0.082		0.085	
Observations	29,794					

Examples from our paper

- Just accepted at JHR, “Adverse Impacts of Mental Health Needs Assessment on Jail Outcomes: Evidence from Transition Age Youth and Adults” with Cunningham, Seward, Clay and Vigliotti (randomized)
- We investigated the effect of mental health needs scores on suicidality in jail and recidivism and explored possible mechanisms like length of stay
- We also used a leniency design coming from randomized clinicians who scored on score of 0 to 3 the severity of their daily functioning
- Rather than review the paper, I want to highlight the main exhibits of a leniency paper today

Write down your structural equation

First equation states your research question, not your instrumental variable strategy. Let's consider a model in which an inmate i at booking b remains in jail longer, attempts suicide or recidivates based on interventions given in response to their mental health needs score, $Score_{ibt}$.

$$Y_{ibt} = \beta_0 + \beta_1 Score_{ibt} + \beta_2 X_{ibt} + \tau_t + \epsilon_{ibt}, \tag{1}$$

where Y_{ibt} is the outcome of interest for individual i in booking b in month-year t: (1) length of stay, (2) suicide attempt, or (3) recidivism within a year of release. The matrix X_{ibt} contains baseline inmate and booking characteristics, τ_t is a vector of month-year dummies and ε_{ibt} is an error term.

Present your instrument equation

To isolate the effect of clinician scoring on the mental health needs score of inmate i at booking b in time t, we first estimate a linear model that includes a vector of month-of-year fixed effects. These fixed effects account for any county-wide trends that might affect mental health needs scores:

$$Score_{ibt} = \gamma_1 + \gamma_2 \tau_t + \varepsilon_{ibt} \tag{2}$$

where ε is an idiosyncratic error.

Present your instrument equation

We then residualize the inmate's mental health needs score using the fitted coefficients from previous equation

$$Score^{*}_{ibt} = Score_{ibt} - \widehat{\gamma_1} - \widehat{\gamma_2}\tau_t \tag{3}$$

This residualized score isolates the part of the inmate's mental health needs score that is due to the clinician's scoring, independent of the month-of-year fixed effects.

Residualized leave-one-out mean IV

We construct a measure of residualized clinician c propensity to assign a high score Z_{btc} as:

$$Z_{btc} = \left(\frac{1}{n_{tc} - n_{itc}}\right)\left(\sum_{k=0}^{n_{it}}(Score^*_{ikt}) - \sum_{b=0}^{n_{itc}}(Score^*_{ibt})\right), \tag{4}$$

where n_{ct} is the number of cases seen by clinician c in month-year t and n_{it} is the number of bookings of inmate i seen by clinician c in month-year t.

$Score^*_{ikt}$ is the residualized mental health needs score given by the clinician at time t, and $Score^*_{ibt}$ is the residualized score by inmates i. In other words, we remove the residualized mental health needs score assignment of all of an inmate's bookings seen by clinician c in each month.

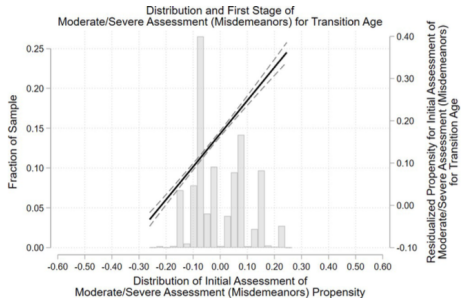
Present the first stage equation

We explore this positive association between the residualized leave-one-out-mean clinician scores and an inmate's own score by estimating the following linear probability model:

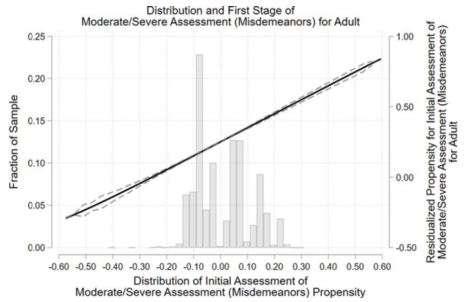
$$Score_{ibt} = \alpha + \pi Z_{btc} + X_{ibt} + \tau_t + \varepsilon_{ibt} \tag{5}$$

where $Score_{ibt}$ is the binary treatment variable ("classification score") indicating whether an inmate received a mental health needs score of "Moderate" or "Severe," and Z_{btc} is a vector of the residualized leave-one-out-mean clinician score, X_{ibt} is an array of pre-booking inmate characteristics, including race, sex, age at booking, whether they had a prior offense in the last year, the number of offenses per booking, τ_t are month-of-year fixed effects, and ε_{ibt} is the inmate specific error term. Standard errors are two-way clustered by clinician and inmate.

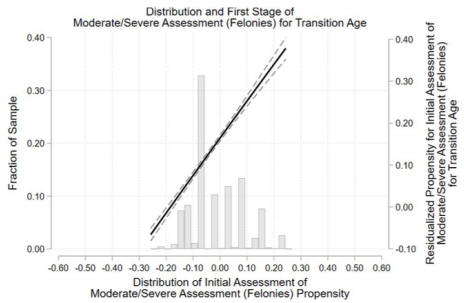
First Stage Figures



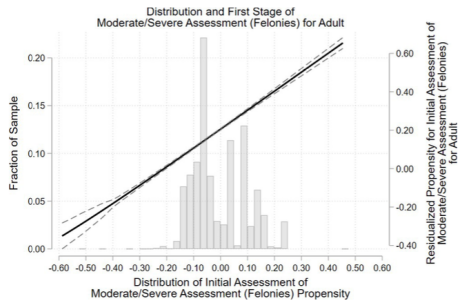
(a) First Stage Youth (Misdemeanors)



(b) First Stage Adults (Misdemeanors)



(c) First Stage Youth (Felonies)



(d) First Stage Adults (Felonies)

Figure 1: Distribution of Clinician Leniency and First Stage

First Stage Tests

Table 4: First Stage Regressions for Initial Assessment of Moderate/Severe (Misdemeanors)

	Transition Age		Adults	
	(1)	(2)	(3)	(4)
Z: Clinician's Leave-Out	0.783***	0.758***	0.981***	0.861***
Mean Mental Health Score	(0.033)	(0.028)	(0.030)	(0.030)
Kleibergen-Paap F	575	715	1,048	808
Time Fixed Effects	Yes	Yes	Yes	Yes
Baseline Controls	No	Yes	No	Yes
Observations	11,147	11,147	30,642	30,642

Notes: We report the first stage results of a linear probability model stratified by age group. The binary outcome of interest is the initial assessment of an inmate's mental health needs being either none/low or moderate/severe. The propensity to assign the most severe score is estimated using data from other cases assigned to the clinician following the procedure described in the text. Columns (1) and (2) limit the sample to TAY whereas columns (3) and (4) limit the sample to adults. Columns (1) and (3) show the results by controlling only for month-year fixed effects, whereas Columns (2) and (4) also include the inmate baseline controls as shown in Table 1. Each column gives the corresponding clinician and inmate robust two-way clustered standard errors in parentheses. Robust (Kleibergen-Paap) first stage F is reported. Note, this is equivalent to the effective F-statistic of Montiel Olea and Pflueger (2013) in this case of a single instrument. * p$<$0.10, ** p$<$0.05, *** p$<$0.01

First Stage Tests

Table 6: Test of Randomization for Moderate/Severe (Misdemeanors)

	Transition Age		Adults	
	(1) Moderate/ Severe (Mis-demeanors)	(2) Z: Moderate/ Severe (Mis-demeanors)	(3) Moderate/ Severe (Mis-demeanors)	(4) Z: Moderate/ Severe (Mis-demeanors)
Asian	0.036	0.004	-0.035	-0.001
	(0.041)	(0.017)	(0.037)	(0.007)
Black	0.012	-0.002	0.009	-0.003**
	(0.014)	(0.002)	(0.010)	(0.001)
Race other	0.080	-0.043**	-0.015	-0.020
	(0.110)	(0.019)	(0.097)	(0.018)
Hispanic	-0.043***	-0.004	-0.075***	-0.004
	(0.011)	(0.004)	(0.011)	(0.003)
Male	-0.056***	-0.012**	-0.066***	-0.011**
	(0.014)	(0.006)	(0.011)	(0.005)
Age at booking	0.001	0.000	0.003***	0.001***
	(0.001)	(0.000)	(0.000)	(0.000)
Number of priors	0.012***	0.000	0.011***	0.001***
	(0.002)	(0.000)	(0.001)	(0.000)
Time fixed effects	Yes	Yes	Yes	Yes
F-test	11	2	33	4
Observations	11,147	11,147	30,642	30,642

Notes: These linear probability models control for the baseline characteristics used in the instrumental variables analyses. The binary dependent variable in columns (1) and (3) is being assigned a moderate-to-severe mental illness score at initial assessment. The dependent variable in columns (2) and (4) is the propensity to assign a high or low score to inmates. Time fixed effects include month-year fixed effects. Clinician and inmate two-way clustered standard errors shown in parentheses. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Balance Tests Showing Independence

Table 8: Balance of Instrument and Inmate Characteristics for Moderate/Severe (Misdemeanors)

Transition Age

	Bottom Tercile	Middle Tercile	Top Tercile	Middle v. Bottom P-Value	Top v. Bottom P-Value
Z: Clinician's Leave-Out Mean Mental Health Score	-0.097	-0.015	0.119	(0.000)	(0.000)
Inmate Characteristics					
Asian	0.013	0.009	0.012	(0.257)	(0.772)
Black	0.253	0.271	0.257	(0.261)	(0.837)
Race other	0.001	0.001	0.000	(0.689)	(0.143)
Hispanic	0.355	0.349	0.353	(0.651)	(0.955)
Male	0.720	0.697	0.690	(0.438)	(0.285)
Age at booking	21.569	21.656	21.647	(0.182)	(0.193)
Number of priors	2.754	3.245	3.179	(0.279)	(0.115)

Adults

	Bottom Tercile	Middle Tercile	Top Tercile	Middle v. Bottom P-Value	Top v. Bottom P-Value
Z: Clinician's Leave-Out Mean Mental Health Score	-0.097	-0.009	0.126	(0.000)	(0.000)
Inmate Characteristics					
Asian	0.011	0.012	0.012	(0.151)	(0.417)
Black	0.251	0.262	0.247	(0.105)	(0.464)
Race other	0.001	0.001	0.000	(0.412)	(0.081)
Hispanic	0.246	0.248	0.240	(0.967)	(0.636)
Male	0.743	0.723	0.726	(0.379)	(0.401)
Age at booking	38.392	39.082	39.718	(0.053)	(0.005)
Number of priors	4.920	6.198	7.018	(0.012)	(0.007)

Notes: Data is from a large county correctional complex. Time fixed effects include month-year fixed effects. Clinician and inmate two-way clustered standard errors shown in parentheses.
* p<0.10, ** p<0.05, *** p<0.01

Average monotonicity tests

Table 10: Average Monotonicity Tests (Misdemeanors)

Transition Age

	Male	Female	Black	White	Hispanic
	(1)	(2)	(3)	(4)	(5)
Z: Clinician's Leave-Out Mean	0.738***	0.796***	0.824***	0.729***	0.632***
Mental Health Score	(0.045)	(0.083)	(0.062)	(0.040)	(0.045)
Observations	7,841	3,306	2,897	8,118	3,930
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

Adults

	Male	Female	Black	White	Hispanic
	(1)	(2)	(3)	(4)	(5)
Z: Clinician's Leave-Out Mean	0.846***	0.901***	0.941***	0.841***	0.734***
Mental Health Score	(0.035)	(0.057)	(0.049)	(0.045)	(0.059)
Observations	22,380	8,262	7,762	22,498	7,502
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

Notes: This table presents a test for satisfying average monotonicity as proposed in Frandsen, Lefgren, and Leslie (2020) where they show that average monotonicity can suffice in lieu of strict monotonicity if the average treatment propensities move in the same direction as their potential treatment decisions. In the context of our paper, we can relax strict monotonicity for any given clinician if the individual monotonically complies with enough other judges. Thus, the coefficients should all be significant and the same direction to support average monotonicity.

2SLS and one instrument

- When we have one endogenous variable and one instrument, we say the equation is “just identified” and if we have more instruments than endogenous variables we say it is “overidentified”
- 2SLS is best when just identified as weak instruments can be better controlled (and tests for it exist like reporting Anderson-Rubin CI as well as reporting appropriately robust F statistics like we saw)
- Single instrument is also helped with visualization and balance tests which I think shouldn’t be downplayed
- But it has problems too which we will now explore

Continuous instrument

- LATE theorem (Imbens and Angrist 1994) references a binary treatment and binary instrument technically making tests for weak instruments easier with broad agreement
- But with a continuous instrument, things change in the leniency design – for one, the LATE interpretation gets a little more challenging, and the monontonicity assumption more complicated

Weak instruments

- But it may even be that the just identified model is a bait and switch in the first place
- Peter Hull and others, for one, have noted that the residualized leave-one-out mean used as the instrument is a collapsing of judge fixed effects into a single scalar as a quasi-propensity score (recall it's dimension reducing)
- So really, the correct specification is judge fixed effects potentially, not the residualized leave-one-out-mean, and 2SLS has severe bias with multiple instruments

Criticism of 2SLS: Over identification and bias

- Even though the 2SLS model is just identified with residualized leave-one-out-mean, our instrument is actually multi-dimensional in the number of judges and with weak instruments, this creates finite sample bias for 2SLS due to *many instruments*; see Sun's `manyweakiv` at https://github.com/lsun20/manyweakiv
- To help pin this down, consider that the propensity score theorem which states the propensity score is a scalar based on dimension reduction in X (Rosenbaum and Rubin 1983)
- This isn't really a just identified model so we should explore alternative models that are more appropriate for our data and design: we consider three

Many instruments

- High-dimensional instruments can arise when there is inherently large number of potentially relevant instruments or when it’s unclear how these instruments should be specified (e.g. dummy variables, interaction effects).

Alternative to 2SLS: LIML

- Two branches of IV: minimum distance IV (very old; limited information maximum likelihood) and “two step” procedures (Wald, 2SLS, JIVE, etc.)
- If LATEs vary, then two step IV estimators like 2SLS estimate a convex combination of them
- Minimum distances estimators like LIML may be outside the convex hull of the LATEs and may not be interpretable as causal
- This calls into the question the use of LIML when there are large number of instruments, so not advised

Alternative to 2SLS: JIVE

- One popular alternative to the 2SLS model in these applications has been the jackknife IV estimator (JIVE) (Angrist, Imbens and Krueger 1999)
- JIVE removes bias using leave-out first stage fits for each observation
- That is the fitted value for observation i is $Z_i\widehat{\pi}_i$ where $\widehat{\pi}_i$ is an estimate that throws out observation i
- The other appeal is that it can handle large number of instruments, but ironically *not* large number of covariates

Many control variables

- The primary interest in an econometric analysis often lies in one or a few regressors, for which we want to estimate the causal effect on an outcome variable.
- However, to allow for a causal interpretation we need to control for confounding factors.
- Lasso-type techniques can be employed to appropriately select controls and thus improve the robustness of causal inference.

Alternative to 2SLS: JIVE

- Kolesar (2013) notes that in a finite sample, JIVE will be noisy and this estimation error will be correlated with the outcome since it depends on the treatment status of a particular person (unit of observation for me is an inmate).
- This will cause JIVE to be biased when the number of covariates is large as is the case in our context – in my recent paper, for instance, I had 14 covariates and 84 time fixed effects.
- We face therefore a tradeoff between a set of time fixed effects that ensure conditional randomization and the biases created by large numbers of covariates for our JIVE estimator.

Alternative to 2SLS: UJIVE

- Researchers therefore often accompany just identified 2SLS with another model that is robust for both large number of covariates and large number of instruments.
- First alternative is to estimate LATEs using the UJIVE estimator (Kolesar 2013)
- UJIVE estimates are robust to a large set of *covariates* and *instruments* by excluding inmate i from estimation, thus guaranteeing that the prediction error be uncorrelated with the outcome (Kolesar 2013)
- This means that UJIVE estimates are consistent for a convex combination of LATEs even when we have a large number of covariates.

Alternative to 2SLS: UJIVE

- You can implement this using Kolesar's matlab code https://github.com/kolesarm/ivreg/blob/master/ivreg.m, but it's giving us problems here so I can't report the results
- You can also use Stata's `manyiv` and we will review that together,
- R: Kolesar's github does not yet have this updated for R https://github.com/kolesarm/ivreg/tree/master, see the ivreg.pdf

Alternatives: LASSO

- But there are also two machine learning selection IV models:
  - → the post-double-selection model and
  - → the post-lasso-orthogonalization method described by Chernozhukov (2015) which we loosely term our LASSO and post-LASSO models, respectively.
- We use the Stata command `lassopack` for its implementation

Why pdslasso?

- `pdslasso` and `ivlasso` are routines for estimating structural parameters in linear models with many controls and/or many instruments.
- The routines use methods for estimating sparse high-dimensional models, specifically the LASSO and the square-root-LASSO (Belloni et al. 2011, 2014).
- Purpose of `pdslasso` is to improve causal inference when the aim is to assess the effect of one or a few (possibly endogenous) regressors on the outcome variable.
- `pdslasso` allows to select control variables and/or instruments.

LASSOPACK

- These estimators can be used to select controls (pdslasso) or instruments (ivlasso) from a large set of variables (possibly numbering more than the number of observations), in a setting where the researcher is interested in estimating the causal impact of one or more (possibly endogenous) causal variables of interest.
- Two approaches are implemented in pdslasso and ivlasso: (1) The ”post- double-selection” (PDS) methodology of Belloni et al. (2012, 2013, 2014, 2015, 2016). (2) The ”post-regularization” (CHS) methodology of Chernozhukov, Hansen and Spindler (2015). For instrumental variable estimation, ivlasso implements weak-identification-robust hypothesis tests and confidence sets using the Chernozhukov et al. (2013) sup-score test.

Results

Table 12: Effects of Initial Assessment of Moderate/Severe (Misdemeanors)

	Transition Age			Adults		
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	2SLS	IVLASSO	OLS	2SLS	IVLASSO
LOS	5.649***	6.082**	7.650***	6.510***	-0.128	1.339
	(0.473)	(2.928)	(2.256)	(0.323)	(2.068)	(1.900)
		[0.641, 11.524]			[-3.978, 3.722]	
Suicide attempt (SA)	0.019***	0.046***	0.040***	0.012***	0.021***	0.018**
	(0.006)	(0.011)	(0.013)	(0.002)	(0.006)	(0.007)
		[0.025, 0.067]			[0.011, 0.032]	
SA/(LOS + 1)	0.003	0.007***	0.007***	0.002***	0.004***	0.004***
	(0.002)	(0.002)	(0.002)	(0.000)	(0.001)	(0.001)
		[0.003, 0.012]			[0.001, 0.007]	
Recid within 1 year	0.053***	0.049	0.054	0.162***	0.245***	0.249***
	(0.008)	(0.056)	(0.066)	(0.02)	(0.065)	(0.071)
		[-0.055, 0.154]			[0.124, 0.366]	
Recid within 18 months	0.064***	-0.004	-0.019	0.149***	0.216***	0.195***
	(0.011)	(0.069)	(0.061)	(0.02)	(0.063)	(0.075)
		[-0.131, 0.123]			[0.099, 0.332]	
Recid within 2 years	0.074***	0.031	-0.039	0.140***	0.245***	0.157
	-0.01	-0.072	-0.044	-0.017	-0.069	-0.099
		[-0.102, 0.164]			[0.132, 0.385]	
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Controls	Yes	Yes	Yes	Yes	Yes	Yes

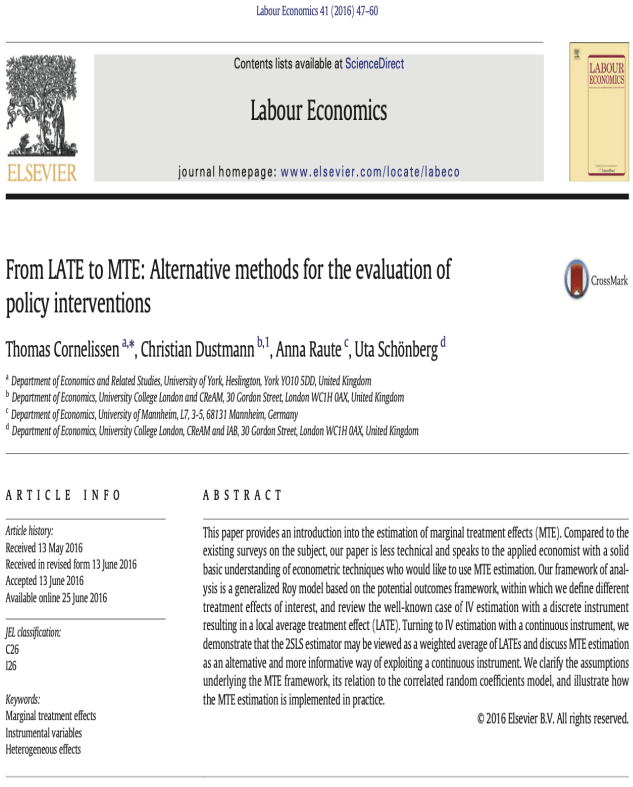
Notes: This table reports the ordinary least squares (OLS), two-stage least squares (2SLS), and the instrumental variables LASSO (IVLASSO) estimates of the impact of being assigned a moderate/severe mental health needs rating. The dependent variable is listed in each row. The recidivism within 1 year sample is truncated to 1 year prior to the last date our dataset. Similarly, the recidivism within 18 months and 2 years samples are truncated to 18 months and 2 years prior to the last date. The 2SLS and IVLASSO specifications instrument for assignation of a high mental health needs score using a clinician leniency measure that is estimated using data from other cases assigned to a clinician as described in the text. We include month-year fixed effects and baseline controls for all specifications. The clinician and inmate robust two-way clustered standard errors are shown in parentheses. For the 2SLS estimates, confidence intervals based on the inversion of the Anderson-Rubin test are shown in brackets.

* p<0.10, ** p<0.05, *** p<0.01

Marginal treatment effects and judge IV

- Increasingly common to see people go from estimating LATE effects with judges to MTE and then even to ATT, ATE and ATU
- Literature goes back to early 2000s by Ed Vycatil and Jim Heckman in a series of papers
- Works only with a continuous instrument, which you have with the residualized leave-one-out mean

Marginal Treatment Effects



Labour Economics 41 (2016) 47–60

Contents lists available at ScienceDirect

Labour Economics

journal homepage: www.elsevier.com/locate/labeco

From LATE to MTE: Alternative methods for the evaluation of policy interventions

Thomas Cornelissen [a,*], Christian Dustmann [b,1], Anna Raute [c], Uta Schönberg [d]

[a] Department of Economics and Related Studies, University of York, Heslington, York YO10 5DD, United Kingdom
[b] Department of Economics, University College London and CReAM, 30 Gordon Street, London WC1H 0AX, United Kingdom
[c] Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany
[d] Department of Economics, University College London, CReAM and IAB, 30 Gordon Street, London WC1H 0AX, United Kingdom

ARTICLE INFO

Article history:
Received 13 May 2016
Received in revised form 13 June 2016
Accepted 13 June 2016
Available online 25 June 2016

JEL classification:
C26
I26

Keywords:
Marginal treatment effects
Instrumental variables
Heterogeneous effects

ABSTRACT

This paper provides an introduction into the estimation of marginal treatment effects (MTE). Compared to the existing surveys on the subject, our paper is less technical and speaks to the applied economist with a solid basic understanding of econometric techniques who would like to use MTE estimation. Our framework of analysis is a generalized Roy model based on the potential outcomes framework, within which we define different treatment effects of interest, and review the well-known case of IV estimation with a discrete instrument resulting in a local average treatment effect (LATE). Turning to IV estimation with a continuous instrument, we demonstrate that the 2SLS estimator may be viewed as a weighted average of LATEs and discuss MTE estimation as an alternative and more informative way of exploiting a continuous instrument. We clarify the assumptions underlying the MTE framework, its relation to the correlated random coefficients model, and illustrate how the MTE estimation is implemented in practice.

- Cornelissen, et al. (2016) is a phenomenal review of this literature
- I've found this literature really fascinating
- If there is heterogeneity in unobservables
- To get a distribution of treatment effects
- Calculate aggregate parameters like the ATE

Marginal treatment effects

What assumptions do we need?

- Heckman and Vytacil (2005, 2006, 2007, etc.) show that all policy relevant parameters (e.g., ATE, ATT, ATU) can be reconstructed using weighted averages over the MTE
- Technically you can identify LATE with average monotonicity, but you need strict for MTE (new AER by Frandsen, Lefgren and Leslie forthcoming)
- Additive separability of treatment effects (i.e., unobserved based on covariates plus observed heterogeneity)
- Only required if not full support, and we don't have full support

Marginal treatment effects

- Rooted in the LATE revolution in the 1990s focused on identifying models with heterogenous treatment effects
- Imbens and Angrist are known for the LATE theorem and clarified interpretation of IV estimates
- Heckam and Vytlacil, Card and others proposed a control function approach as an alternative to linear IV which under stronger assumptions than IV would allow the estimation of the ATE (which IV could not do)

Marginal treatment effects

- Marginal treatment effect (MTE) appears in the late 80s in the context of a switching regression model where ”marginal gain” was the gain from treatment for people who were shifted into (or out of) treatment status by a marginal change in the cost of the treatment (which was the instrument)
- Heckman and Vytlacil in four articles from 1999 to 2007 define the MTE as the “gain from treatment for individuals shifted into treatment or out of treatment by a marginal change in the *propensity score* (i.e., the predicted probability of treatment which is a function of the instrument)
- They then develop nonparametric estimation methods and clarify the connection of the switching regime self selection model and of MTE with IV and LATE

Marginal treatment effects

- Since it's familiar to the audience, I'll focus on recidivism from the mental health paper I mentioned not suicide
- In our context, we could have explored the heterogeneity in treatment effects across inmates' underlying mental illness proxied by the propensity score
- Consider the following equations decomposing an inmate i's potential recidivism into the conditional means of potential outcomes, $\mu^j(X_i)$, based on inmate characteristics X_i as well as deviations from the mean U_i^j

$$
\begin{aligned}
Y_i^0 &= \mu^0(X_i) + U_i^0 \\
Y_i^1 &= \mu^1(X_i) + U_i^1
\end{aligned}
$$

Selection

Mental health assignment is based on an individual latent index threshold in which the net benefits of “severe mental illness” (obviously odd way of putting it) are exactly equal to

$$D_i^* = \mu^D(X_i, Z_i) - V_i$$

where X_i and Z_i are the inmate’s observed determinants of treatment choice, but Z_i is the instruments, and V_i the unobserved characteristics that makes treatment choice less like (“unobserved resistance to treatment”)

Assignment to the severe mental illness (low functioning) occurs when $D_i^* > 0$ otherwise they are classified as high functioning

Selection and expected gains

- Very common for people in this literature to use language about “selection based on gains” as in choosing the treatment bc they expect to gain from it
- In our context, we don’t think you can take that literally because their moderate to severe mental illness is doing the choosing
- We treat “choice” and “moderate to severe mental illness” are treated as synonyms, but that’s just in our context

Selection

- The directions on our variables can be a little hard to interpret because for us higher values of $\mu_D(X, Z)$ basically cause them to get a severe mental illness score, but they correspond to worse symptoms
- When the therapist believes the inmate's functioning is above her own reservation threshold for that inmate, V_i, the evaluator assigns a high score which assigns him to a high severe mental illness score

Selection

- Indifference condition is $\mu_D(X, Z) = V$, and thus when $D^* > 0$, $\mu_D(X, Z) > V$, and the therapist assigns him to a severe mental illness score
- We then apply the cdf of V to this inequality to get $F_V(\mu^D(X_i, Z_i)) \geq F_V(V_i)$ which bounds both sides b/w 0 and 1
- The left side is the propensity score (i.e., the conditional probability of treatment) which is $p_i(X_i, Z_i)$ and the right is the quantiles of the distribution of the unobserved resistance to treatment, U_i^D
- Rewrite the selection equation as $p_i(X_i, Z_i) \geq U_i^D$ which means an inmate i selects into treatment when their propensity score is greater than their unwillingness to participate due to their slightly higher functioning

Some sources of confusion I had

- So it may help if you replace the phrase “high propensity score / low resistance to treatment” with “moderate to severe mental illness” (high scores)
- If I have schizophrenia with extreme displays of psychosis, then I am **less** resistance to treatment because the treatment is a high score, and I will almost certainly get a high score (high propensity score)
- If I come in with depression but am functional, my score is lower and so I both have a lower propensity score *and* therefore have a *higher* resistance to treatment
- Note the subtle language: “propensity score” and “resistance to treatment” are reversed – people with less resistance have higher propensity scores because their illness is so severe

Switching equation

Write down the choice of either potential outcome according to which treatment was chosen using a switching equation

$$
\begin{aligned}
Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 \\
Y_i &= Y_i^0 + (Y_i^1 - Y_i^0) D_i
\end{aligned}
$$

which is a regression equation. If we substitute our earlier potential outcomes into this we get:

$$
Y_i = \mu^0(X_i) + D_i(\mu^1(X_i) - \mu^0(X_i) + U_i^1 + U_i^0) + U_i^0
$$

Steps

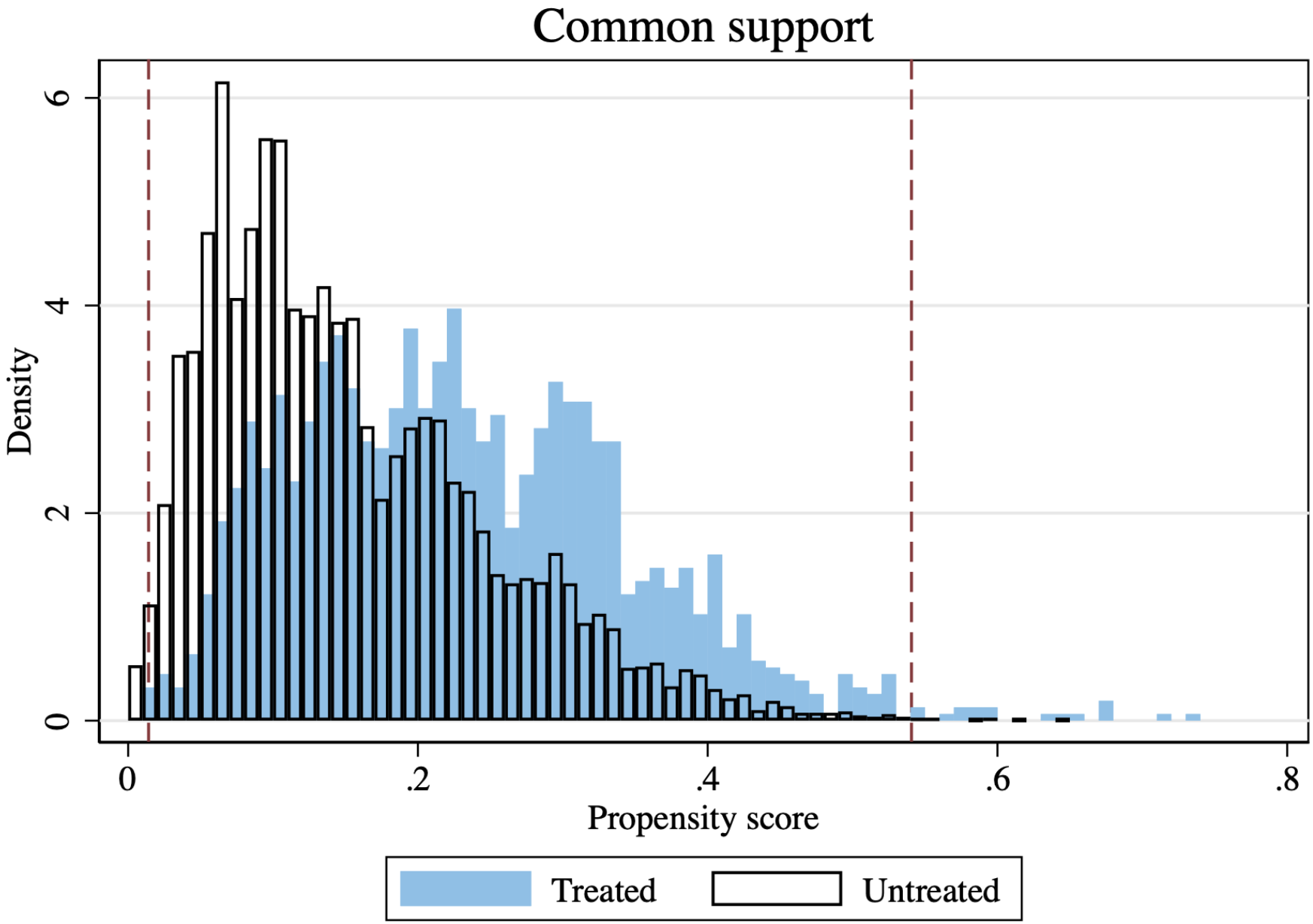
1. Estimate the propensity score using logit (here issues of common support arise as we want it across all cells of Z)
2. Model recidivism as a function of covariates and propensity score with second degree polynomials
3. Calculate the derivative of $\widehat{recidivism}$ with respect to the propensity score and plot it as a curve

Since we don't have full support, we rescaled the weights so that they integrate to one over the trimmed sample so that we can calculate different weighted averages of the MTEs

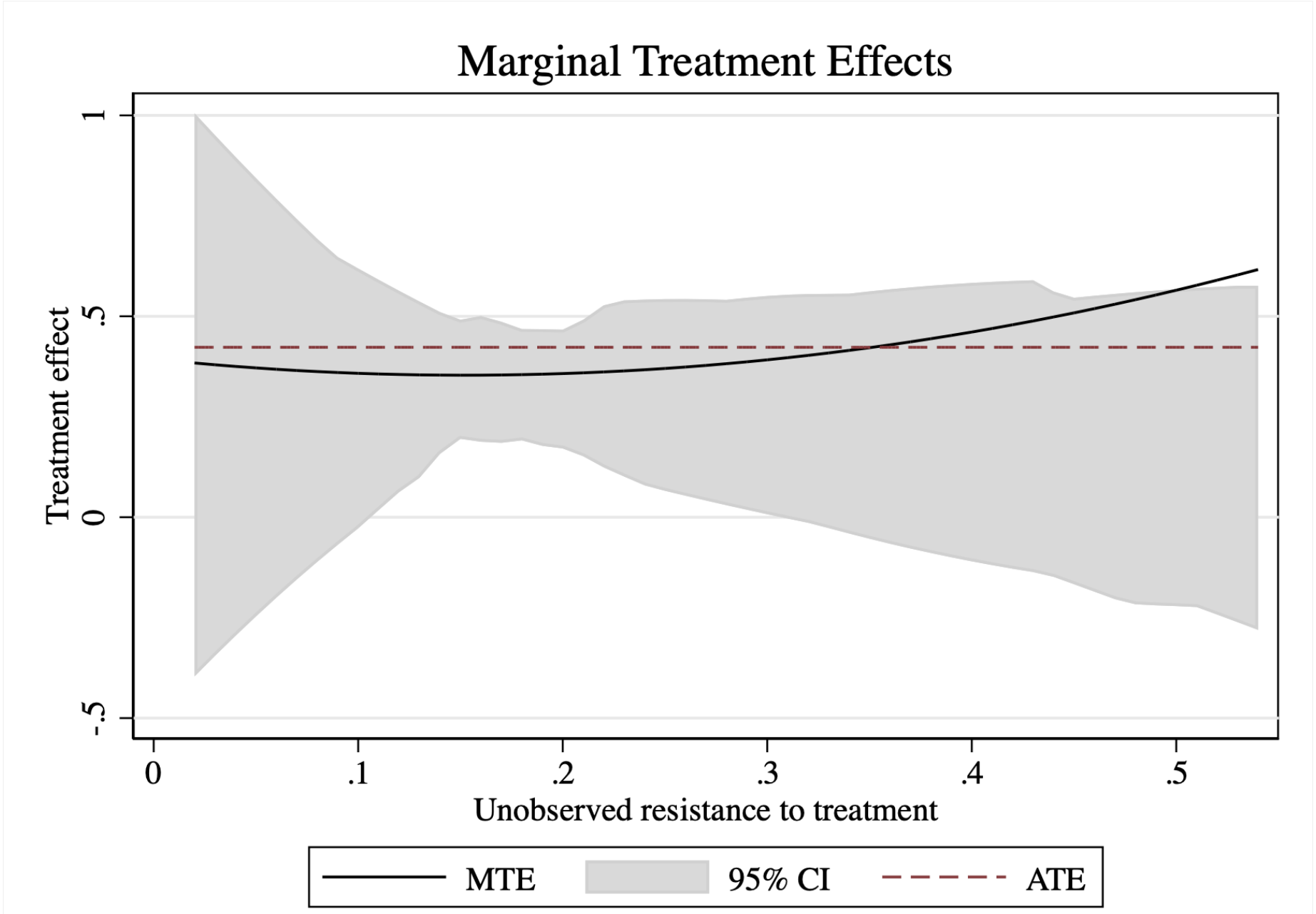
Calculating aggregates

- ATE is the equally weighted average over the entire MTE curve, ATT as a weighted average over the left group of “low resistance, high propensity score”, and ATU as the weighted average of the right group of (“high resistance, low propensity”)
- Downward slope means selection on unobserved returns to receiving a mental illness score (i.e., ATT>ATE>ATU)
- Upward slope means the reverse (i.e., ATU>ATE>ATT)

Common support



MTE and aggregate parameters for recidivism



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"

Concluding remarks

- Identifying causal effect without an instrument is likely not possible given the deep selection issues associated with crime as a child and adult
- Leniency designs are everywhere, even in tech, but you need to know how to look for them
- Bottleneck, influential decision-makers, discretion - these are the three elements of the design

Supply and Demand

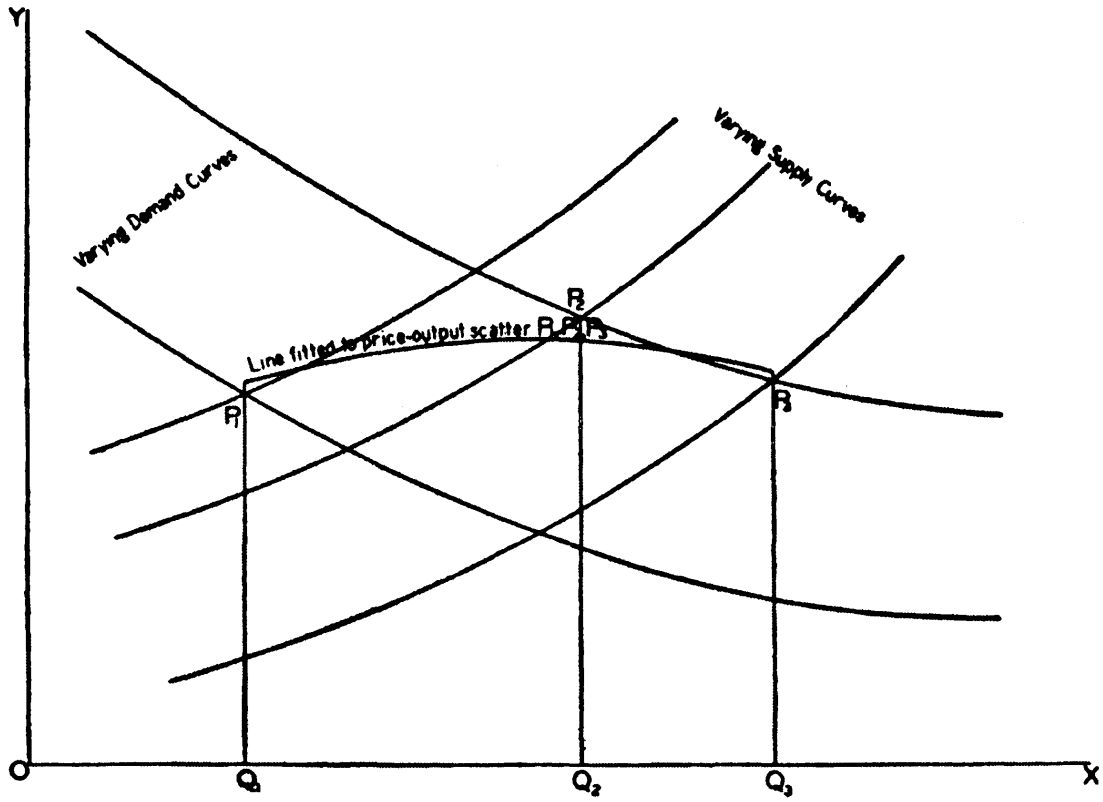
- Instrumental variables was developed in the 1920s largely to address problems created by supply and demand
- Demand curves are causal functions, but so are supply curves
- We do not observe all the prices and quantities to be able to calculate the slope or shape of the demand curve because we only observe the “realized prices and quantities” in equilibrium
- But if we did know the price elasticity of demand, we could set more optimal policies like tax policy or profit maximization

Supply and Demand

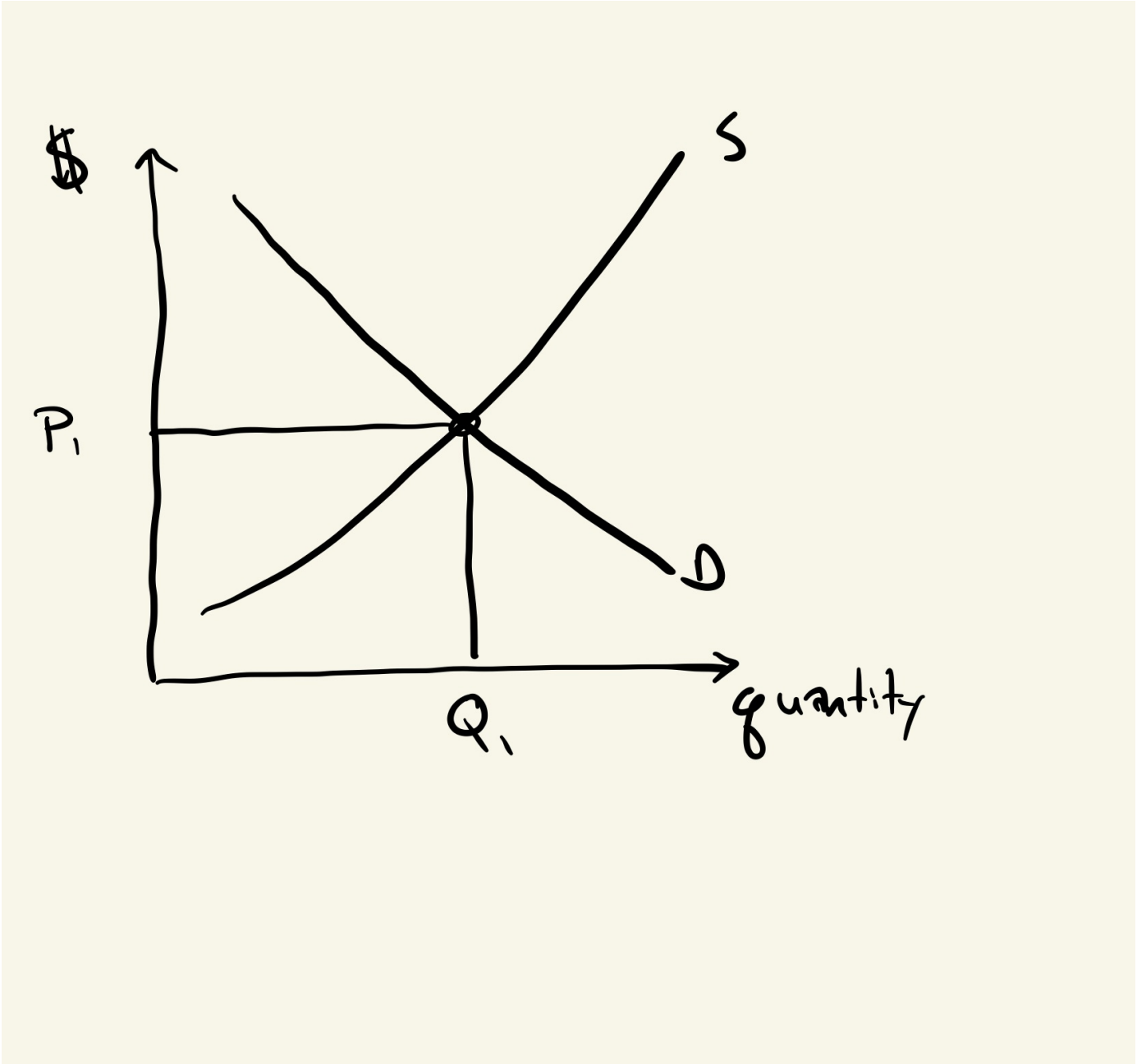
Exhibit 1

The Graphical Demonstration of the Identification Problem in Appendix B (p. 296)

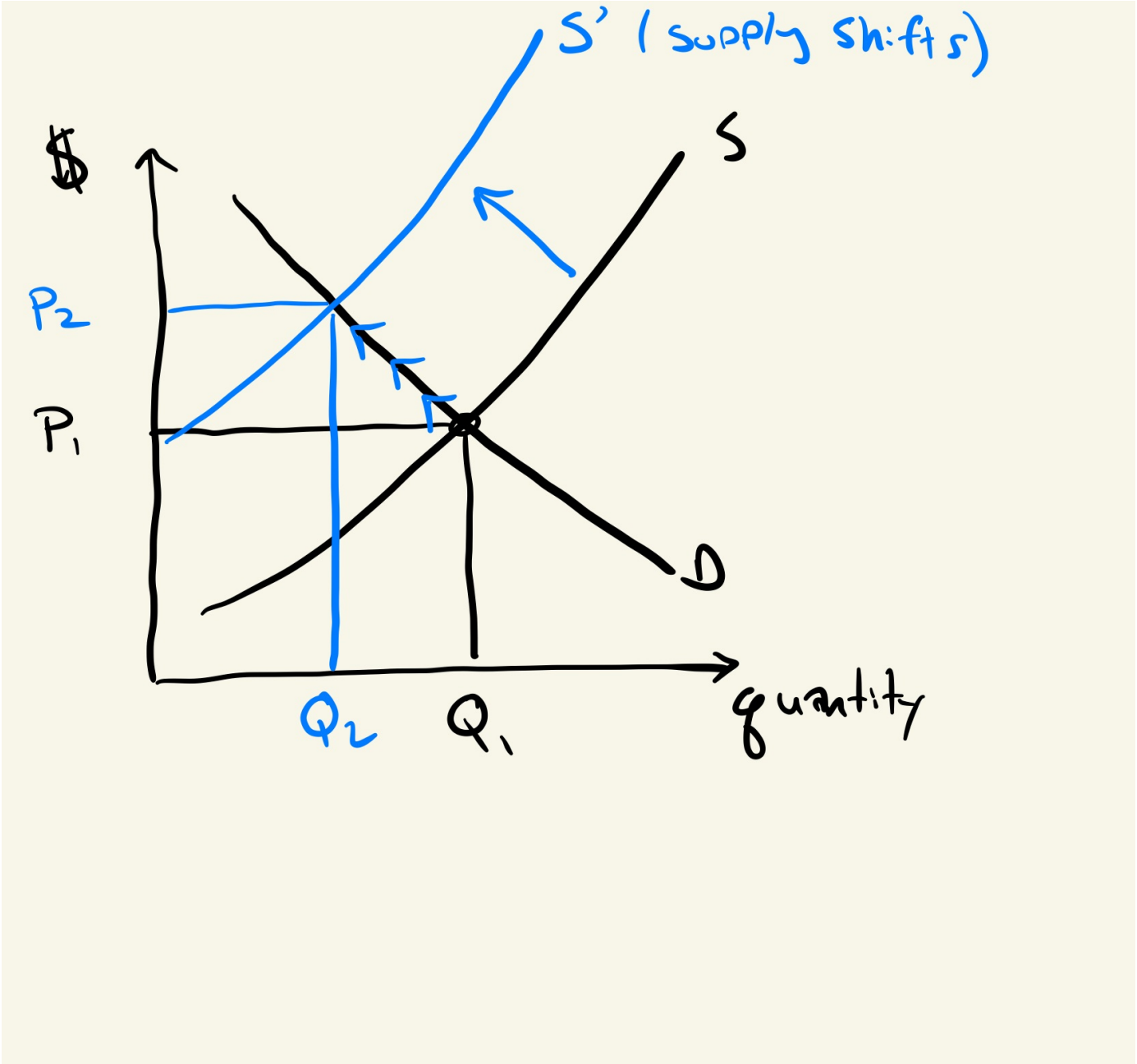
FIGURE 4. PRICE-OUTPUT DATA FAIL TO REVEAL EITHER SUPPLY OR DEMAND CURVE.



Initial price and quantity



Supply shift



Price elasticity of demand

$$\delta = \frac{Q_2 - Q_1}{P_2 - P_1}$$

Can be estimated with log-log regressions:

$$LnQ_{it} = \alpha + \delta LnP_{it} + \psi_{it}$$

But you need an instrument for price and it must be a supply shifter only

Supply shift

- **Supply shifters**: Firm input costs and technology are typical candidates
- Demand shifters: Other consumer good prices, consumer income, availability of substitutes, expectations about the future
- Good instruments must shift **only** supply – not demand

Elasticity of meth demand

HEALTH ECONOMICS
Health Econ. **25**: 1268–1290 (2016)
Published online 27 July 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.3213

IDENTIFYING DEMAND RESPONSES TO ILLEGAL DRUG SUPPLY INTERDICTIONS

SCOTT CUNNINGHAM[a] and KEITH FINLAY[b,*]

[a]*Department of Economics, Baylor University, Waco, TX, USA*

[b]*Center for Administrative Records Research and Applications, US Census Bureau, Room 6H216F, 4600 Silver Hill Road, Washington, DC, USA*

SUMMARY

Successful supply-side interdictions into illegal drug markets are predicated on the responsiveness of drug prices to enforcement and the price elasticity of demand for addictive drugs. We present causal estimates that targeted interventions aimed at methamphetamine input markets ('precursor control') can temporarily increase retail street prices, but methamphetamine consumption is weakly responsive to higher drug prices. After the supply interventions, purity-adjusted prices increased then quickly returned to pre-treatment levels within 6–12 months, demonstrating the short-term effects of precursor control. The price elasticity of methamphetamine demand is −0.13 to −0.21 for self-admitted drug treatment admissions and between −0.24 and −0.28 for hospital inpatient admissions. We find some evidence of a positive cross-price effect for cocaine, but we do not find robust evidence that increases in methamphetamine prices increased heroin, alcohol, or marijuana drug use. This study can inform policy discussions regarding other synthesized drugs, including illicit use of pharmaceuticals.

Received 20 December 2013; Revised 14 February 2015; Accepted 20 May 2015

JEL Classification: I12; I18; K42

KEY WORDS: illegal drugs; addiction; demand; substitution; war on drugs; methamphetamine

1. INTRODUCTION

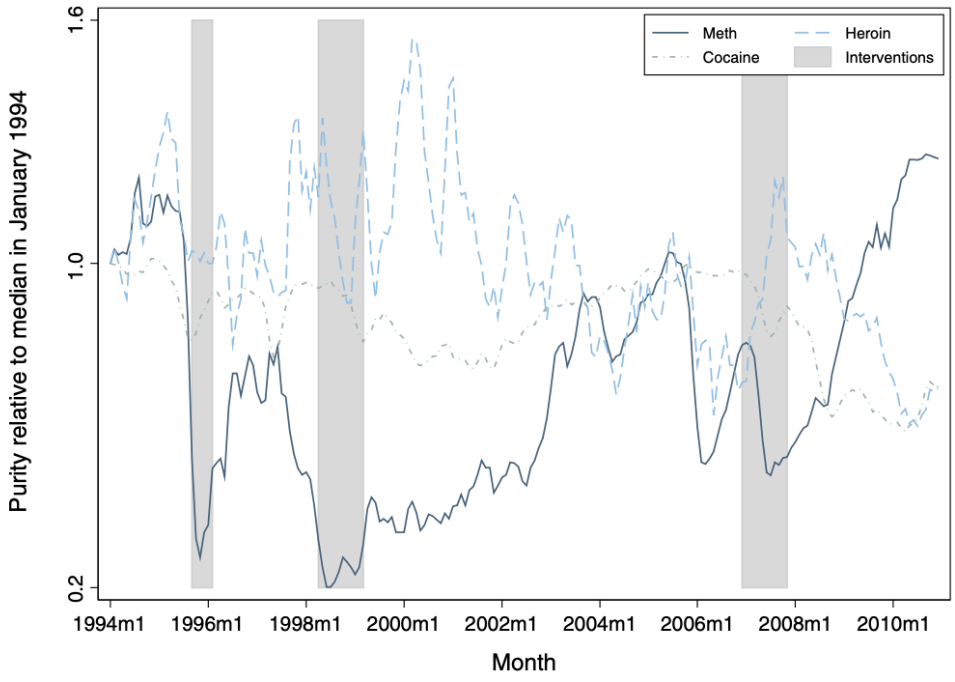
Policymakers trade off the social costs of addiction with the costs of enforcement when designing optimal drug policy. Costs for the enforcement of drug laws are as much as $40 billion annually in the USA (Miron and Waldock, 2010). The US incarceration rate per 100,000 residents grew from 100 in 1980 to 492 in 2011 as the share of prisoners convicted of drug offenses increased from 22% to 48% (Blumstein and Beck, 1999; Carson and Sabol, 2012). Although violence associated with drug trafficking is a major urban problem, the marginal efficacy of enforcement-oriented interventions is uncertain given evidence of diminishing returns to incarceration (Johnson and Raphael, 2012). Policies that attempt to reduce demand by increasing drug prices may also be ineffective if drug addicts have inelastic demand with respect to prices.

There are few causal estimates of illegal drug demand because of the difficulty of obtaining exogenous variation in prices and reliable indicators of use. The simultaneity of supply and demand for each drug confounds estimates of demand elasticities as a causal measure of demand response to price changes. For example, suppose that the government chooses an enforcement policy for reducing illegal drug consumption and then

Cooking meth

- d-methamphetamine is a chemical product synthesized from either ephedrine or pseudoephedrine
- 1995, 1997 and 2003 there were several federal regulations that restricted access to these precursors as an effort combat meth epidemic
- Undercover meth seizure data showed massive increases in real price of meth on the street in response

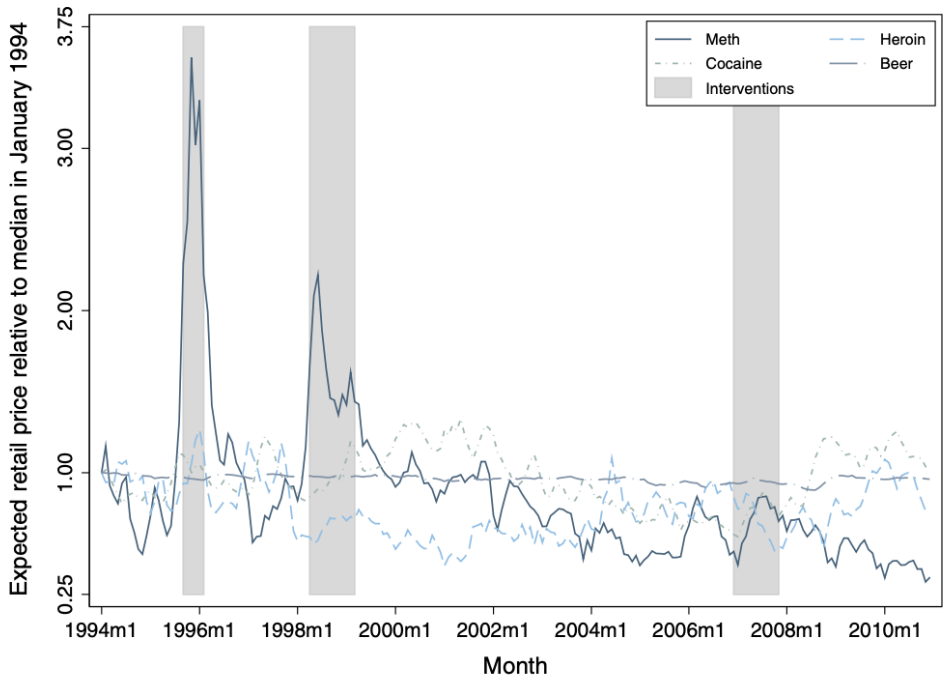
Meth purity plummetted



Notes: Authors' calculations from STRIDE. Month-of-year fixed effects have been partialled out from the raw series to improve presentation. The 1995 and 1997 interdictions represent the time windows after significant federal seizures when real prices deviated from trend. The 2006 interdiction represents the window after the effective date of the Combat Methamphetamine Epidemic Act when real prices deviated from trend.

Figure 1. Ratio of median purities of meth, heroin, and cocaine relative to their respective values in January 1994, System to Retrieve Information from Drug Evidence (STRIDE), 1994–2010

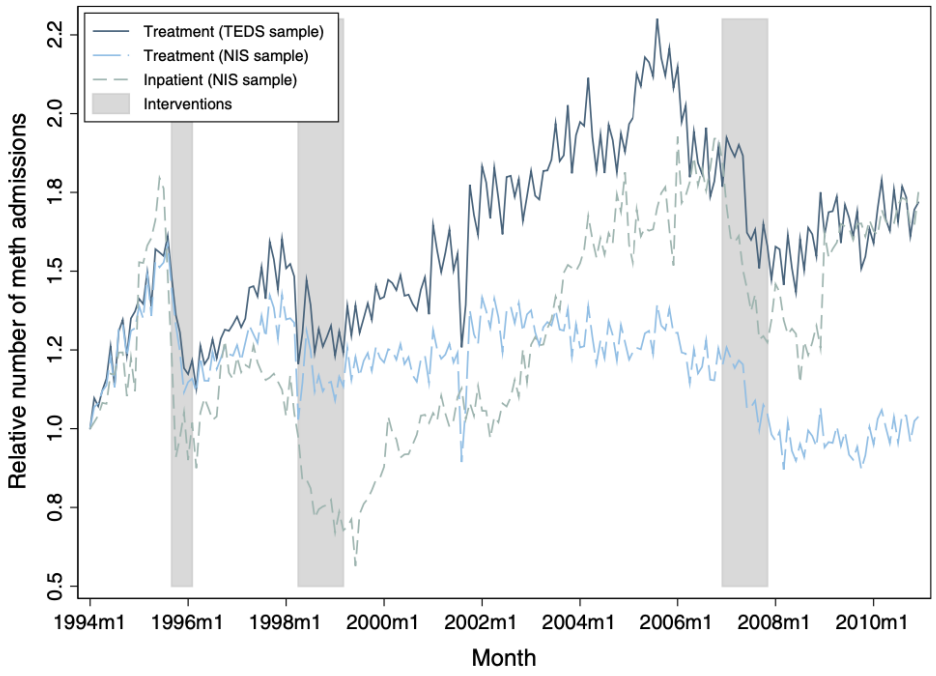
Meth prices skyrocketed



Notes: Authors' calculations from STRIDE and ACCRA. Month-of-year fixed effects have been partialled out from the raw series to improve presentation. Prices are inflated to 2013 dollars by the All Urban CPI series before calculating the ratio. The 1995 and 1997 interdictions represent the time windows after significant federal seizures when real prices deviated from trend. The 2006 interdiction represents the window after the effective date of the Combat Methamphetamine Epidemic Act when real prices deviated from trend.

Figure 2. Ratio of median monthly expected retail prices of meth, heroin, and cocaine, and retail price of beer relative to their respective values in January 1994, System to Retrieve Information from Drug Evidence (STRIDE) and ACCRA, 1994–2010

Meth admissions to treatment and hospitals fell



Notes: Month-of-year fixed effects have been partialled out from the raw series to improve presentation. The NIS sample includes only states that participated in the NIS during the entire sample period: Arizona, California, Colorado, Connecticut, Illinois, Iowa, Kansas, Maryland, Massachusetts, New Jersey, New York, Oregon, Pennsylvania, South Carolina, Washington, and Wisconsin. The 1995 and 1997 interdictions represent the time windows after significant federal seizures when real prices deviated from trend. The 2006 interdiction represents the window after the effective date of the Combat Methamphetamine Epidemic Act when real prices deviated from trend.

Figure 3. Hospital inpatient and self-admitted treatment proxies for meth use relative to January 1994, Treatment Episode Data Set (TEDS) and Nationwide Inpatient Sample (NIS), various subsamples, 1994–2010

OLS and 2SLS Estimation

Table IV. Regressions of log self-admitted methamphetamine treatment and log hospital inpatient admissions on log drug prices, TEDS and NIS samples, 1994–2010

Outcome	Drug treatment						Hospital inpatient					
Estimator	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS	OLS	2SLS
Covariates	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Log meth price (1 month lag)	−0.09***	−0.21***	−0.06***	−0.20***	−0.06***	−0.20***	−0.07	−0.24***	−0.09*	−0.25***	−0.09*	−0.26***
	(0.02)	(0.06)	(0.01)	(0.06)	(0.01)	(0.06)	(0.05)	(0.07)	(0.04)	(0.07)	(0.04)	(0.07)
Log unemployment rate	0.29**	0.24**	0.23**	0.17*	0.23**	0.16**	−0.25*	−0.32***	−0.07	−0.14	−0.09	−0.17
	(0.11)	(0.11)	(0.10)	(0.09)	(0.09)	(0.08)	(0.13)	(0.12)	(0.12)	(0.11)	(0.13)	(0.11)
Log cigarette tax	−0.02	−0.02	0.00	−0.01	0.03	0.01	0.02	−0.02	−0.12	−0.17**	−0.15**	−0.21***
	(0.07)	(0.07)	(0.08)	(0.08)	(0.07)	(0.07)	(0.15)	(0.15)	(0.08)	(0.08)	(0.07)	(0.07)
Log population 15–49	1.59**	1.44**	2.44**	2.03*	4.66***	3.77***	0.09	−0.07	3.20***	3.02***	4.53***	4.01***
	(0.75)	(0.71)	(1.20)	(1.05)	(1.42)	(1.24)	(1.49)	(1.44)	(0.96)	(0.80)	(1.20)	(0.96)
Linear national trend	×	×					×	×				
Linear state trends			×	×					×	×		
Quadratic state trends					×	×					×	×
First stage												
1995 intervention indicator		0.89***		0.90***		0.93***		0.99***		0.99***		1.05***
(1 month lag)		(0.13)		(0.12)		(0.13)		(0.21)		(0.22)		(0.21)
1997 intervention indicator		0.62***		0.61***		0.58***		0.65***		0.65***		0.64***
(1 month lag)		(0.05)		(0.05)		(0.06)		(0.07)		(0.07)		(0.07)
CMEA indicator		0.11**		0.12**		0.13**		0.12*		0.12*		0.08
(1 month lag)		(0.05)		(0.05)		(0.05)		(0.06)		(0.07)		(0.07)
First-stage F-statistic		90		65		45		41		37		32
First-stage p-value		0.00		0.00		0.00		0.00		0.00		0.00
Hansen χ^2-statistic		1.83		2.06		1.55		0.33		0.33		0.34
Hansen p-value		0.40		0.36		0.46		0.85		0.85		0.85
Specification												
R^2	0.92		0.94		0.95		0.93		0.95		0.95	
N (state-months)	8,532	8,532	8,532	8,532	8,532	8,532	2,830	2,830	2,830	2,830	2,830	2,830
N (states)	44	44	44	44	44	44	15	15	15	15	15	15
Mean of dep. var.	3.99	3.99	3.99	3.99	3.99	3.99	3.73	3.73	3.73	3.73	3.73	3.73
Std. dev. of dep. var.	1.71	1.71	1.71	1.71	1.71	1.71	1.74	1.74	1.74	1.74	1.74	1.74

Notes: All models include state and month-of-year fixed effects. Standard errors that account for arbitrary, within-state heteroskedasticity are shown in parentheses. Stars indicate statistical significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. TEDS, Treatment Episode Data Set; NIS, Nationwide Inpatient Sample.

Discussion

- Highly inelastic (-0.21 to -0.26) and robust across two measures of meth consumption
- Need data on prices and quantities, required FOIA requests to Drug Enforcement Agency and purchasing data on hospitalizations and treatment
- But crucially needed something that would’ve raised firm costs through inputs that was not related to consumer demand (theory guided)
- Instruments were deviations in the price from longterm trends

Discussion

- There are other ways to estimate demand, but this method is one that should immediately be exploited when supply shocks happen
- Focus needs to be on inputs for which there are not instantly the ability to substitute
- But can’t be so correlated with broader demand shocks that you are back shifting supply and demand (e.g., COVID)

Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
  - Local average treatment effects
- Application
  - Data visualization and necessary evidence
  - Leniency design
  - Marginal Treatment Effects
  - Covariates
  - Price elasticity of demand
- Conclusion

Conclusion

- “With a long enough lever, I can move the world” – Archimedes
- With a strong enough and strange enough instrument, you can identify the LATE even outside of the laboratory
- Need to know what to look for so keep that DAG in mind all the time – write it on the wall
- But like selection on observables, we need plausible assumptions, properly measured data and appropriate estimators