

# Causal Inference I

MIXTAPE SESSION

---



# Roadmap

Instrumental variables

- Background

- Intuition

Estimators

- Two Step

- Weak instruments

Heterogeneity

- Local average treatment effects

- Presentation suggestions

Leniency design

# When is IV used?

Instrumental variables methods are typically used to address the following kinds of problems encountered in naive regressions

1. Omitted variable bias
2. Classical measurement error
3. Simultaneity (eg supply and demand)
4. Reverse causality
5. Randomized control trials with noncompliance
6. Fuzzy RDD

# Instrumental variables

- If treatment is tied to an unobservable, then RDD and DiD may be invalid (they do not solve all problems)
- The best instruments come from deep knowledge of institutional details (Angrist and Krueger 1991)
- Certain types of natural experiments can be the source of such opportunities and may be useful, but not all natural experiments will satisfy IV even if they are themselves interpretable as causal

# Constant treatment effects

- Constant treatment effects (i.e.,  $\delta$  is constant across all individual units)
  - Constant treatment effects is the traditional econometric pedagogy when first learning instrumental variables, and doesn't need the potential outcomes model or notation to get the point across
  - Constant treatment effects is identical to assuming that  $ATE=ATT=ATU$  because constant treatment effects assumes  $\delta_i = \delta_{-i} = \delta$  for all  $i$  units

# Heterogenous treatment effects

- Heterogeneous treatment effects (i.e.,  $\delta_i$  varies across individual units)
  - Heterogeneous treatment effects means that the  $ATE \neq ATT \neq ATU$  because  $\delta_i$  differs across the population
  - This is equivalent to assuming the coefficient,  $\delta_i$ , is a random variable that varies across the population
  - Heterogenous treatment effects is based on work by Angrist, Imbens and Rubin (1996) and Imbens and Angrist (1994) which introduced the "local average treatment effect" (LATE) concept

# Data requirements

- 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*.
- That is, the researcher knows of a variable ( $Z$ ) that actually *is* randomly assigned and that affects the endogenous variable but not the outcome (except via the endogenous variable)
- Such a variable is called an "instrument".

# Picking a good instrument

- The best instruments you think of first, then you seek the data second (but often students go in the reverse order which is basically guaranteed to be a crappy instrument)
- If you want to use IV, then ask:  
*What moves around the covariate of interest that might be plausibly random?*
- Is there any element in the treatment that could be construed as random?
- If you were to find that random piece, then you have found an instrument
- Once you have identified such a variable, begin to think about what data sets might have information on an outcome of interest, the treatment, and the instrument you have put your finger on.

# Does family size reduce labor supply or is it selection?

Angrist and Evans (1998), "Children and their parents' labor supply"  
*American Economic Review*,

- They want to know the effect of family size on labor supply, but need exogenous changes in family size
- So what if I told you if the first two children born were of the same gender, then you're less likely to work. What?!

## Angrist and Evans cont.

- Many parents have a preference for having at least one child of each gender
  - Consider a couple whose first two kids were both boys; they will often have a third, hoping to have a girl
  - Consider a couple whose first two kids were girls; they will often have a third, hoping for a boy
  - Consider a couple with one boy and one girl; they will often not have a third kid
- The gender of your kids is arguably randomly assigned (maybe not exactly, but close enough)

## 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 *strangeness of the instrument*, which is the inference drawn by an intelligent layperson with no particular knowledge of the phenomena or background in statistics.
- You need more information, in other words, otherwise the layperson can't understand what same gender of your children has to do with working

## When a good IV strategy finally makes sense

- But then the researchers point out that women whose first two children are of the same gender are more likely to have additional children than women whose first two children are of different genders
- The layperson then asks himself, “Hm. I wonder if the labor market differences are due *solely* to the differences in the number of kids the woman has...”

# Sunday Candy is a good instrument

- Let's listen to a few lines from "Ultralight Beam" by Kanye West.  
Chance the Rapper sings on it and says  
*"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 making a song have to do with hell? What does meeting Kanye West have to do with success? Let's consider each in order

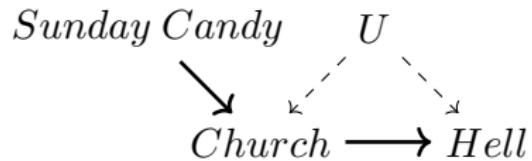
# What are we missing?

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

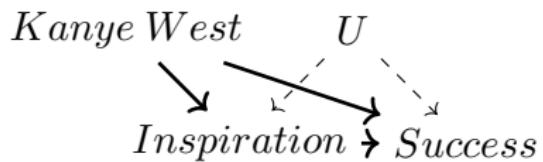
# Sunday Candy DAG



# 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



## Foreshadowing the questions you need to be asking

1. Is our instrument highly correlated with the treatment? With the outcome? Can you test that?
2. Are there random elements within the treatment? Why do you think that?
3. Is the instrument exogenous? Why do you think that?
4. Could the instrument affect outcomes directly? Why do you think that?
5. Could the instrument be associated with anything that causes the outcome even if it doesn't directly? Why do you think that?

# Roadmap

Instrumental variables

Background

Intuition

Estimators

Two Step

Weak instruments

Heterogeneity

Local average treatment effects

Presentation suggestions

Leniency design

## Two step vs Minimum Distance

- The two-stage least squares (2SLS) estimator was developed by Theil (1953) and Basman (1957) independently
- Kolesář 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  $\nu$  is the error term

- Suppose there exists a variable,  $Z_i$ , that is correlated with  $S_i$  .
- We can estimate  $\delta$  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  $\delta$ , 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  $Z$  is orthogonal to the factors in  $\nu$ , such as unobserved ability,  $A$ , as well as the structural disturbance term,  $\nu$
- Combine  $A$  and  $\nu$  into a composite error term  $\eta$  for simplicity
- Assuming the first stage exists and that the exclusion restriction holds, then we can estimate  $\delta$  with  $\hat{\delta}_{Wald}$ :

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

## Two Sample IV

- Wald can be implemented in exotic ways, even across datasets
  1. Dataset 1 needs information on the outcome and the instrument –  $\text{Cov}(Y, Z)$
  2. Dataset 2 needs information on the treatment and the instrument –  $\text{Cov}(D, Z)$
- 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. Sometimes called the structural model:

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

- 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,  $\hat{S}$ :

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

## Reduced form

- Some people like a simpler approach because they don't want to defend IV's assumptions
- Reduced form a regression of  $Y$  onto the instrument:

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

- This would be like regressing hell onto Sunday Candy, as opposed to regressing hell onto church with Sunday Candy instrumenting for church

## 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

$$Y_i = \alpha + \delta S_i + \eta_i \text{ (causal model)}$$

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

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

## Two-stage least squares language

$$Y_i = \psi + \pi Z_i + \varepsilon_i \text{ (reduced form)}$$

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

We can calculate the ratio of “reduced form” ( $\pi$ ) to “first stage” coefficient ( $\rho$ ) using the Wald IV estimator:

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

## Two-stage least squares

Carry over from previous slide

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

Rewrite  $\hat{\rho}$  as

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

## Two-stage least squares

Multiply Wald IV by  $\frac{\hat{\rho}}{\bar{\rho}}$  (also note the subscript – we are moving now into 2SLS)

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

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

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

## Two-stage least squares

Recall

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

So after estimation, we get

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

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

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

## Proof.

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

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

$$Cov(\widehat{S}, Y) = \widehat{\delta}Cov(Y, Z)$$



## Intuition of 2SLS

- Intuition is that 2SLS replaces  $S$  with the fitted values  $\hat{S}$  from the first stage regression of  $S$  onto  $Z$  and all other covariates
- I prefer the intuition of 2SLS to the intuition of the ratio of reduced form to first stage, though your mileage may vary
- 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 (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

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

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

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

# Finite sample problems with 2SLS

Plug in covariance and write out the following:

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

## Finite sample problems with 2SLS

Substitute the causal model definition of  $Y$  to get:

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

Where did the first term go? Why did the second term become  $\delta$ ? 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.,  $\hat{S}$ ) 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

- 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>

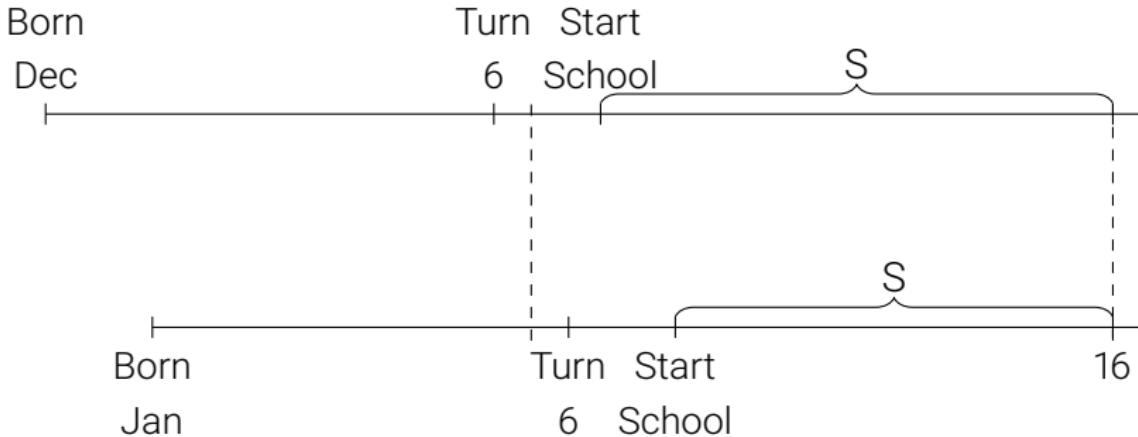
Somewhat inspiring to hear how Angrist reframed the weak instrument problem which his paper with Krueger brought into crisp focus

## 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 Chance quote - what the heck would birth quarter have to do with earnings such that it was an excludable instrument?

# Compulsory schooling

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



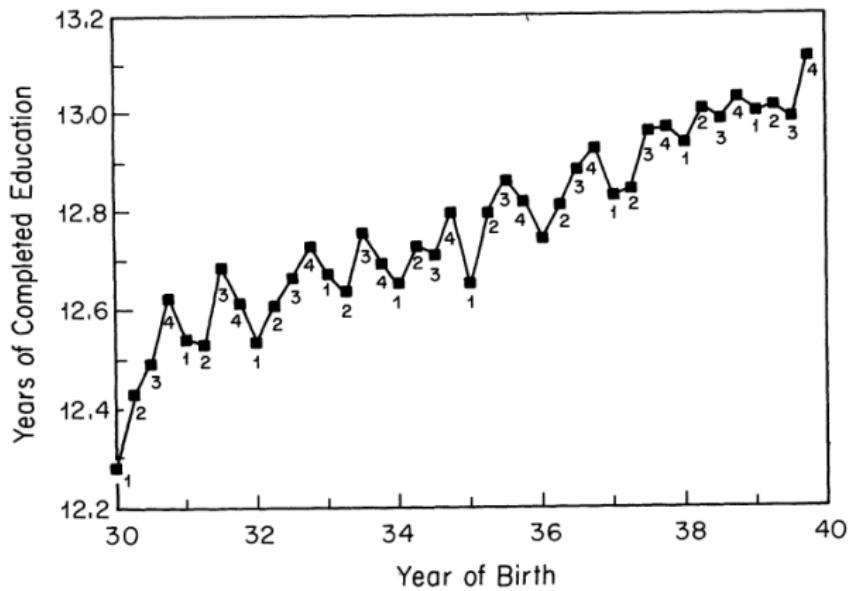
If you're born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

# Visuals

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

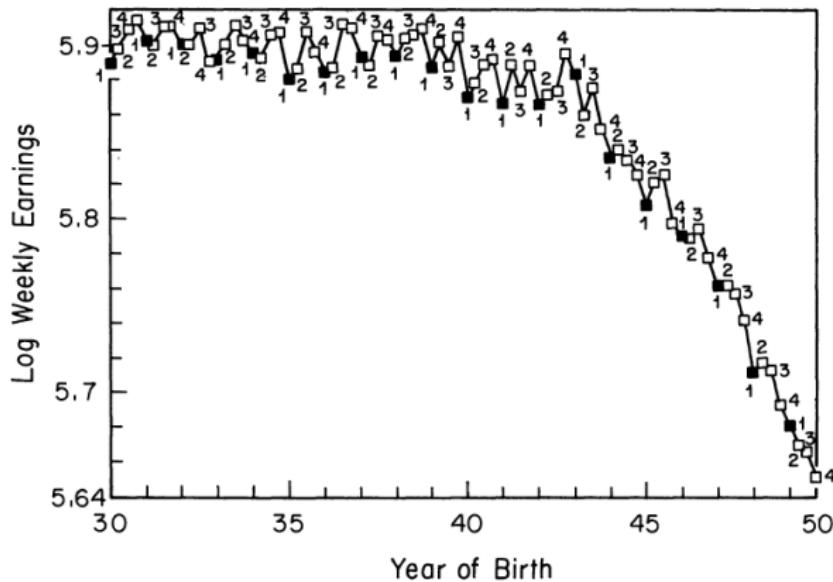
## First Stage

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



# Reduced Form

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



# Two Stage Least Squares model

- The causal model is

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

- The first stage regression is:

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

- The reduced form regression is:

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

- The covariate adjusted IV estimator is the sample analog of the ratio,

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

## Two Stage Least Squares

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

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

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

# First stage regression results

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

Outcome variable	Birth cohort	Mean	Quarter-of-birth effect <sup>a</sup>			F-test <sup>b</sup> [P-value]
			I	II	III	
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 (0.0003)	0.0891 (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	—	—
$\chi^2$ [dof]	—	25.4 [29]

## More instruments

To incorporate the cross-state seasonal variation in education, we computed TSLS estimates that use as instruments for education a set of three quarter-of-birth dummies interacted with fifty state-of-birth dummies, in addition to three quarter-of-birth dummies interacted with nine year-of-birth dummies.<sup>18</sup> The estimates also include fifty state-of-birth dummies in the wage equation, so the variability in education used to identify the return to education in the TSLS estimates is solely due to differences by season of birth. Unlike the previous TSLS estimates, the seasonal differences are now allowed to vary by state as well as by birth year.

## Problem enters with many quarter of birth interactions

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

## More instruments

Table VII presents the TSLS and OLS estimates of the new specification for the sample of 40–49 year-old men in the 1980 Census. This is the same sample used in the estimates in Table V. Freeing up the instruments by state of birth and including 50 state-of-birth dummies in the wage equation results in approximately a 40 percent reduction in the standard errors of the TSLS estimates. Furthermore, in the specifications in each of the columns in Table VII, the estimated return to education in the TSLS model is slightly greater than the corresponding TSLS estimate in Table V, whereas in each of the OLS models the return is slightly smaller in Table VII than in Table V. As a consequence, the difference between the TSLS and OLS estimates is of greater significance. For example, the TSLS estimate in column (6) of Table VII is 0.083 with a standard error of 0.010, and the OLS estimate is 0.063 with a standard error of 0.0003: the TSLS estimate is nearly 30 percent greater than the OLS estimate.

# More instruments

TABLE VII  
OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1930–1939: 1980 CENSUS<sup>a</sup>

Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0673 (0.0003)	0.0928 (0.0093)	0.0673 (0.0003)	0.0907 (0.0107)	0.0628 (0.0003)	0.0831 (0.0095)	0.0628 (0.0003)	0.0811 (0.0109)
Race (1 = black)	—	—	—	—	-0.2547 (0.0043)	-0.2333 (0.0109)	-0.2547 (0.0043)	-0.2354 (0.0122)
SMSA (1 = center city)	—	—	—	—	0.1705 (0.0029)	0.1511 (0.0095)	0.1705 (0.0029)	0.1531 (0.0107)
Married (1 = married)	—	—	—	—	0.2487 (0.0032)	0.2435 (0.0040)	0.2487 (0.0032)	0.2441 (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 (0.0617)	-0.0880 (0.0624)	—	—	-0.0778 (0.0603)	-0.0876 (0.0609)
Age-squared	—	—	0.0008 (0.0007)	0.0009 (0.0007)	—	—	0.0008 (0.0007)	0.0009 (0.0007)
$\chi^2$ [dof]	—	163 [179]	—	161 [177]	—	164 [179]	—	162 [177]

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

## Weak Instruments

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

# Matrices and instruments

- The causal model of interest is:

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

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

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

## Weak instruments and bias towards OLS

- If  $\nu_i$  and  $\eta_i$  are correlated, estimating the first equation by OLS would lead to biased results, wherein the OLS bias is:

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

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

## Weak instruments and 2SLS bias towards OLS

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

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

- Consider the intuition all that work bought us now: if the first stage is weak (i.e.,  $F \rightarrow 0$ ), 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 ( $F \rightarrow \infty$ ) 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 → 3 instruments
2. Quarter of birth dummies + (quarter of birth) × (year of birth) + (quarter of birth) × (state of birth) → 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 (.000)	.142 (.033)	.063 (.000)	.081 (.016)
F (excluded instruments)		13.486		4.747
Partial R <sup>2</sup> (excluded instruments, ×100)		.012		.043
F (overidentification)		.932		.775
<i>Age Control Variables</i>				
Age, Age <sup>2</sup>	x	x		
9 Year of birth dummies			x	x
<i>Excluded Instruments</i>				
Quarter of birth		x		x
Quarter of birth × year of birth			x	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 (.000)	.083 (.009)
<i>F</i> (excluded instruments)	2.428	
Partial <i>R</i> <sup>2</sup> (excluded instruments, ×100)	.133	
<i>F</i> (overidentification)	.919	
<i>Age Control Variables</i>		
Age, Age <sup>2</sup>		
9 Year of birth dummies	x	x
<i>Excluded Instruments</i>		
Quarter of birth	x	
Quarter of birth × year of birth	x	
Quarter of birth × 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 has poor properties here
  - Under powered
  - Artificially low standard errors when endogeneity is severe
  - This causes  $t$ -tests to be misleading

## IV advice: Weak instruments

*"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  $\beta_{2SLS}$  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  $F$ -test to gauge instrument strength
- Andrews, et al. (2019) suggest the Olea and Pflueger (2013) effective first-stage  $F$  statistic
- Single instrument just-identified case reduces to the conventional robust  $F$  and the Kleibergen and Paap (2006) Wald

# Roadmap

Instrumental variables

- Background

- Intuition

Estimators

- Two Step

- Weak instruments

Heterogeneity

- Local average treatment effects

- Presentation suggestions

Leniency design

## 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

$$Y(D, Z) = Y(D, Z') \text{ for all } Z, Z', \text{ and for all } D$$

- 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  $\nu$  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):

$$Y_i^1 = Y_i(1, 1) = Y_i(1, 0)$$

$$Y_i^0 = Y_i(0, 1) = Y_i(0, 0)$$

- Rewrite switching equation:

$$Y_i = Y_i(0, Z_i) + [Y_i(1, Z_i) - Y_i(0, Z_i)]D_i$$

$$Y_i = Y_i^0 + [Y_i^1 - Y_i^0]D_i$$

$$Y_i = Y_i^0 + \delta_i D_i$$

- 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  $i = 1, \dots, N$

- Recall that  $\pi_{1i}$  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  $Y$ :

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

# 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)

## Practical advice

- Some people are good at econometrics
- Some people know how a rug can really tie a room together
- Both of these talents are helpful when it comes to data visualization of causal inference

## Practical advice

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

## IV advice: Pictures

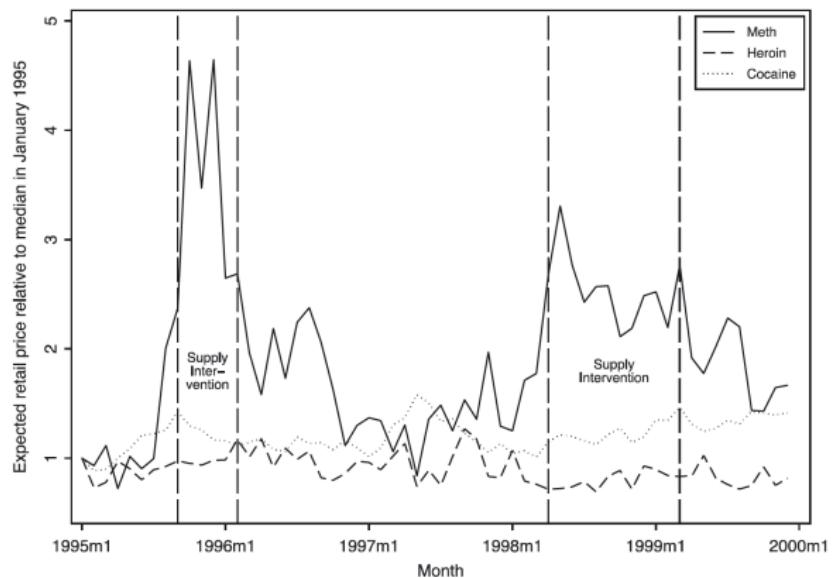
Present your main results as Wald quantites in beautiful pictures even if you're estimating with 2SLS

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

# IV advice: Pictures

**FIGURE 3**

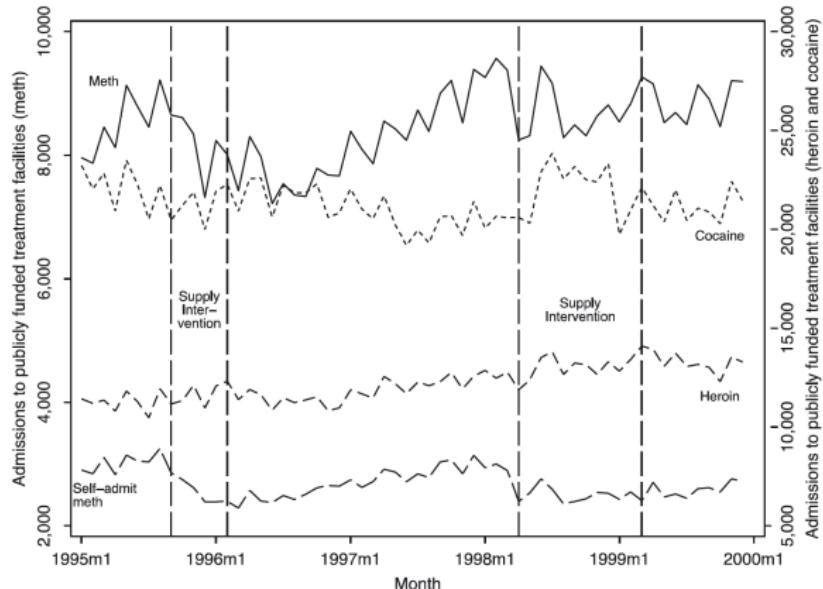
Ratio of Median Monthly Expected Retail Prices of Meth, Heroin, and Cocaine Relative to Their Respective Values in January 1995, STRIDE, 1995–1999



# IV advice: Pictures

**FIGURE 5**

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

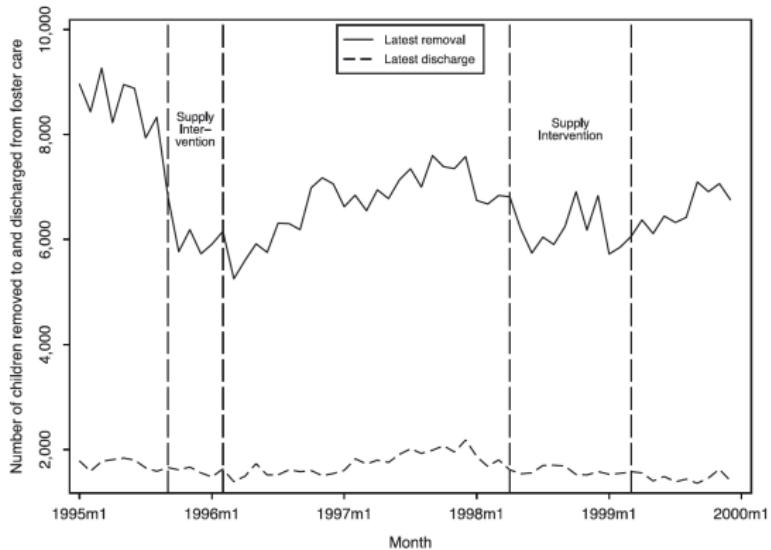


*Notes:* Authors' calculations from TEDS. Arizona, the District of Columbia, Kentucky, Mississippi, West Virginia, and Wyoming are excluded because of poor data quality. Patients can report the use of more than one drug.

# IV advice: Pictures

**FIGURE 4**

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



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

# 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

<b>Dependent variable</b>	<b>Log wage</b>	
	OLS	2SLS
educ	0.071*** (0.003)	0.124** (0.050)
exper	0.034*** (0.002)	0.056*** (0.020)
black	-0.166*** (0.018)	-0.116** (0.051)
south	-0.132*** (0.015)	-0.113*** (0.023)
married	-0.036*** (0.003)	-0.032*** (0.005)
smsa	0.176*** (0.015)	0.148*** (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

## Looking forward

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

# Summarizing

- The potential outcomes framework gives a more subtle interpretation of what IV is measuring
  - In the constant coefficients world, IV measures  $\delta$  which is “the” causal effect of  $D_i$  on  $Y_i$ , and assumed to be the same for all  $i$  units
  - In the random coefficients world, IV measures instead an average of heterogeneous causal effects across a particular population –  $E[\delta_i]$  for some group of  $i$  units
  - IV, therefore, measures the *local average treatment effect* or LATE parameter, which is the average of causal effects across the subpopulation of *compliers*, or those units whose covariate of interest,  $D_i$ , is influenced by the instrument.

## Summarizing

- Under heterogeneous treatment effects, Angrist and Evans (1996) identify the causal effect of the gender composition of the first two kids on labor supply
- This is not the same thing as identifying the causal effect of children on labor supply; the former is a LATE whereas the latter might be better described as an ATE
- *Ex post* this is probably obvious, but like many obvious things, it wasn't obvious until it was worked out. This was a real breakthrough (see Angrist, Imbens and Rubin 1996; Imbens and Angrist 1994)

# Roadmap

Instrumental variables

- Background

- Intuition

Estimators

- Two Step

- Weak instruments

Heterogeneity

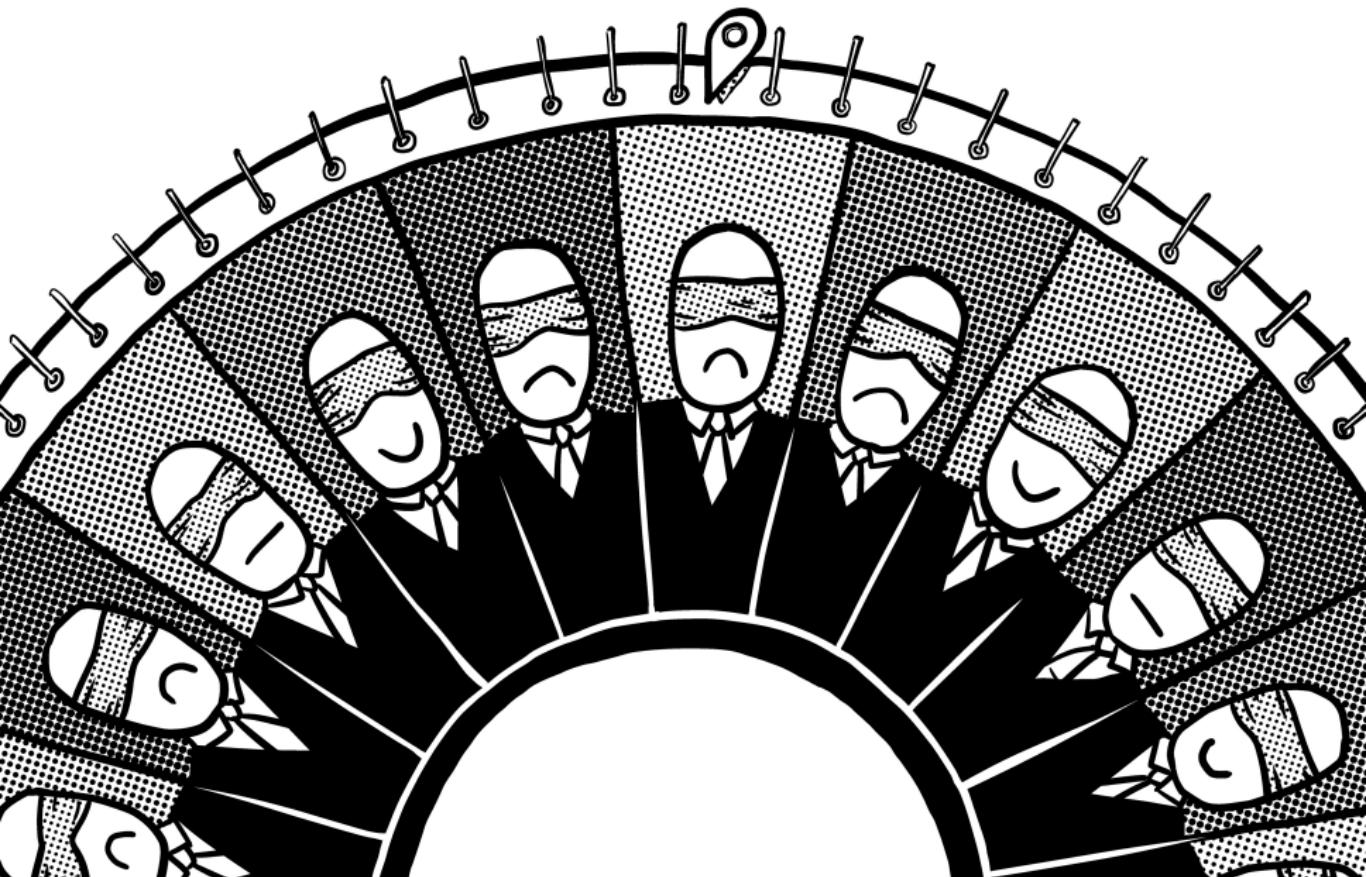
- Local average treatment effects

- Presentation suggestions

Leniency design

# Leniency designs

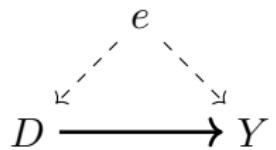
- Imagine the following:
  1. A person moves through a pipeline and hits a critical point where treatment occurs as a result of some decision-maker
  2. There are many different decision-makers and you're assigned randomly to one of them
  3. Each decision-maker differs in terms of their *leniency* in 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

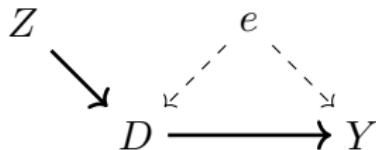
- Aizer and Doyle (2015) were interested in the causal effect of juvenile imprisonment on future crime and human capital accumulation
- Extremely important policy question given the US has the world's highest incarceration rate and prison population of any country in the world by a significant margin (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  $\varepsilon_i$  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

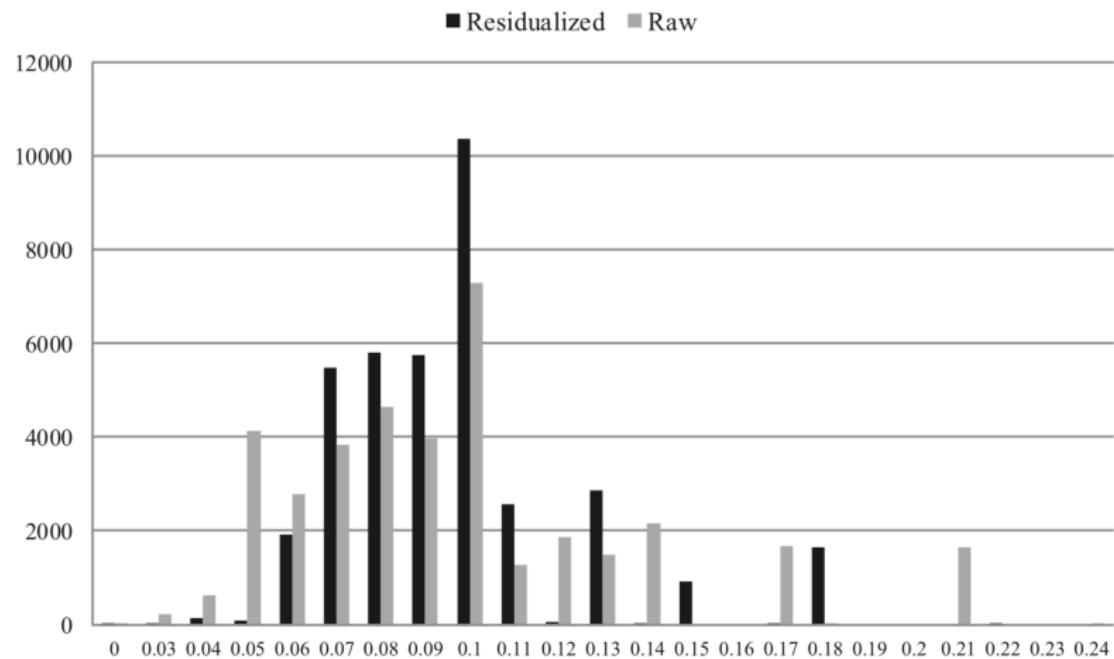


FIGURE I  
Distribution of Z: Judge Incarceration Rate

# Balance test

TABLE II  
INSTRUMENT VERSUS JUVENILE CHARACTERISTICS

	Z distribution			Middle vs.	Top vs.
	Bottom tercile	Middle tercile	Top tercile	bottom p-value	bottom p-value
Z: first judge's leave-out mean incarceration rate in first cases	0.062	0.094	0.147	(.000)	(.000)
<b>Juvenile characteristics</b>					
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 (0.102)	1.082 (0.095)	1.060 (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			
Juvenile incarceration	OLS	OLS	Inverse propensity score weighting	OLS	OLS	2SLS	2SLS
	-0.389 (0.0066)	-0.292 (0.0065)	-0.391 (0.0055)	-0.088 (0.0043)	-0.073 (0.0041)	-0.108 (0.044)	-0.125 (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 (0.0082)	0.350 (0.0064)	0.219 (0.013)	0.200 (0.0072)	0.155 (0.0073)	0.260 (0.073)	0.234 (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 (0.0031)	0.021 (0.0030)	0.035 (0.030)	0.138 (0.0046)	0.061 (0.0050)	0.149 (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 (0.0040)	0.047 (0.0038)	0.142 (0.044)	0.183 (0.011)	0.078 (0.0068)	0.097 (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)	(2)	(3)	(4)	(5)	(6)
	Ever present in CPS school at least 1 year after Initial hearing	Transferred to another CPS high school in years after hearing	Ultimate transfer: adult correctional facility	OLS	2SLS	OLS
Juvenile incarceration	-0.025 (0.0063)	-0.215 (0.069)	0.055 (0.010)	-0.115 (0.243)	0.127 (0.006)	0.243 (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

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

## Concluding remarks

- Sad, but important, paper - the marginal kid shouldn't have been incarcerated
- More generally, leniency designs are very powerful and very common if you know how to look for them
- Bottleneck, influential decision-makers, discretion - these are the three elements of the design