Causal Inference

Mixtape Session

Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
- Heterogeneity
  - Local average treatment effects
  - Covariates
  - Presentation suggestions
- Leniency design application
  - Introduction to leniency designs
  - Marginal Treatment Effects
- Other common applications
  - Lottery designs
  - Fuzzy RDD

Instrumental variables

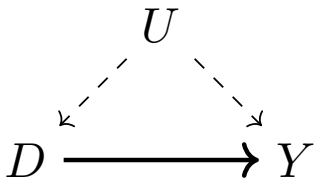
- If treatment is tied to an unobservable, then conditioning strategies, even RDD, are invalid
- Instrumental variables offers some hope at recovering the causal effect of D on Y
- The best instruments come from deep knowledge of institutional details (Angrist and Krueger 1991)
- Certain types of natural experiments can be the source of such opportunities and may be useful

When is IV used?

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

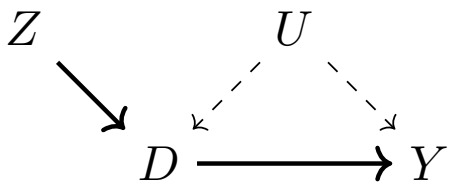
1. Omitted variable bias
2. Classical measurement error
3. Simultaneity (eg supply and demand)
4. Reverse causality
5. Randomized control trials with noncompliance
6. Fuzzy RDD

Selection on unobservables



Then D is endogenous due to backdoor path $D \leftarrow U \rightarrow Y$ and causal effect $D \rightarrow Y$ is not identified using the backdoor criterion.

Instruments



Notice how the path from $Z \to D \leftarrow U \to Y$ is blocked by a collider.

Phillip Wright

- Philip Wright was a renaissance man - published in JASA, QJE, AER, you name it, while on a very intense teaching load.
- Also published poetry, and even personally published Carl Sandburg's first book of poetry!
- Spent a long time at Tufts
- He was very concerned about the negative effects of tariffs and wrote a book about commodity markets

Elasticity of demand is unidentified

- James Stock notes that his publications had a theme regarding identification
- He knew, for instance, that he couldn't simple look at correlations between price and quantity if he wanted the elasticity of demand due to simultaneous shifts in supply and demand
- The pairs of quantity and price weren't demand, or supply - they were demand and supply equilibrium values and therefore didn't reflect the demand or the supply curve, both of which are counterfactuals
- Those points are nothing more than a bunch of numbers – no more, no less – that have no practical use, scientific or otherwise

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.

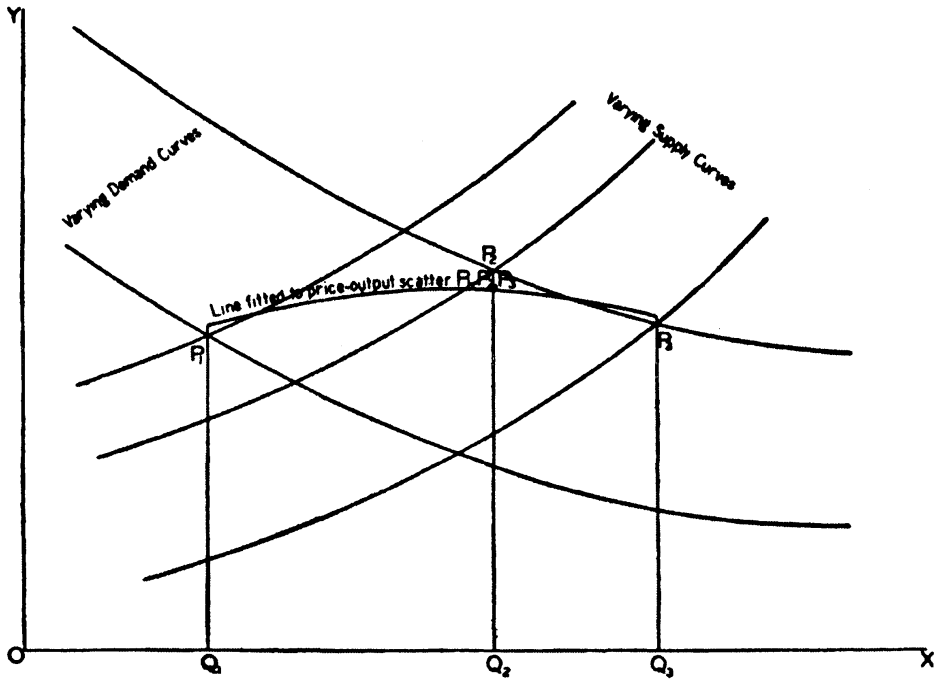
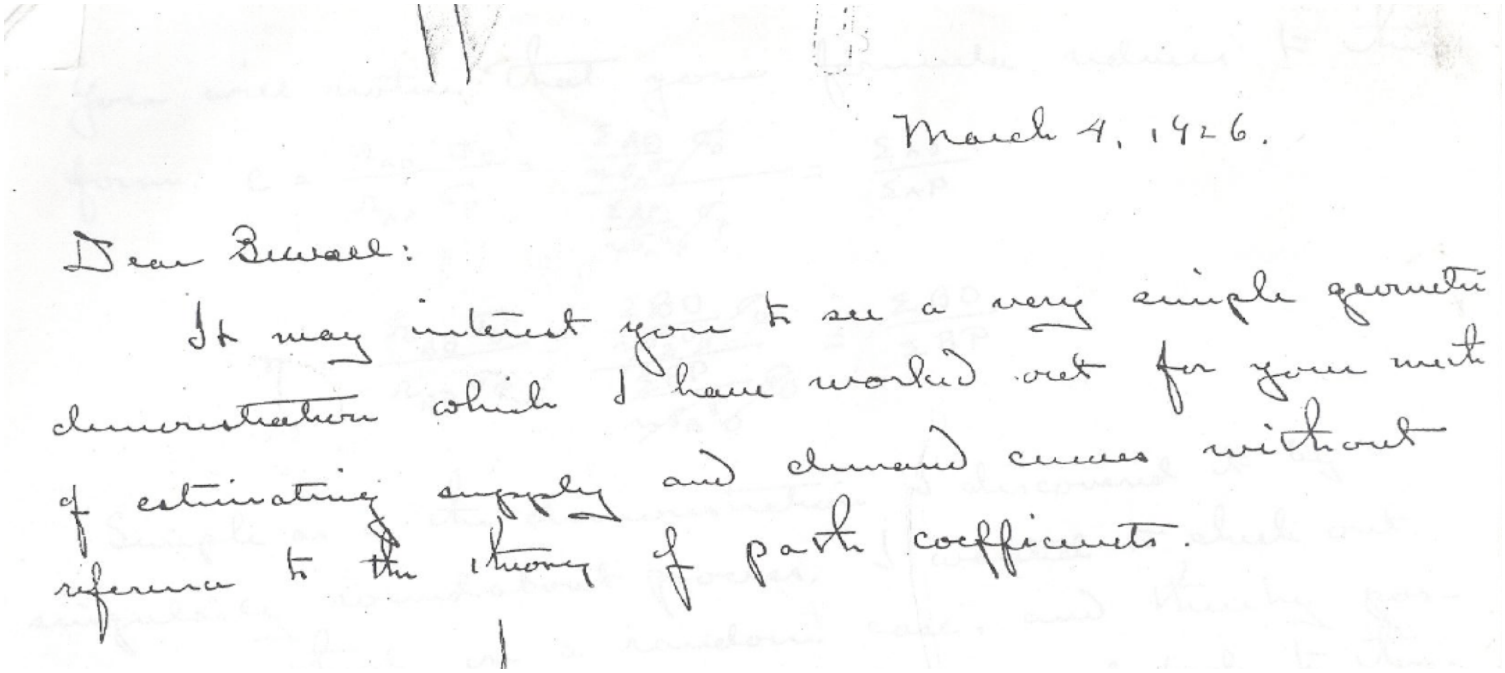


Figure: Wright's graphical demonstration of the identification problem

Sewell Wright

- Sewell was his son, who did *not* go into the family business
- Rather, he decided to become a genius and invent genetics
- Developed path diagrams (which Pearl revived 50 years later for causal inference)
- Father and son engage in letter correspondence as Philip tried to solve the “identification problem”



March 4, 1926.

Dear Sewall:

It may interest you to see a very simple geometric demonstration which I have worked out for your method of estimating supply and demand curves without reference to the theory of path coefficients.

Figure: Wright's letter to Sewell, his son

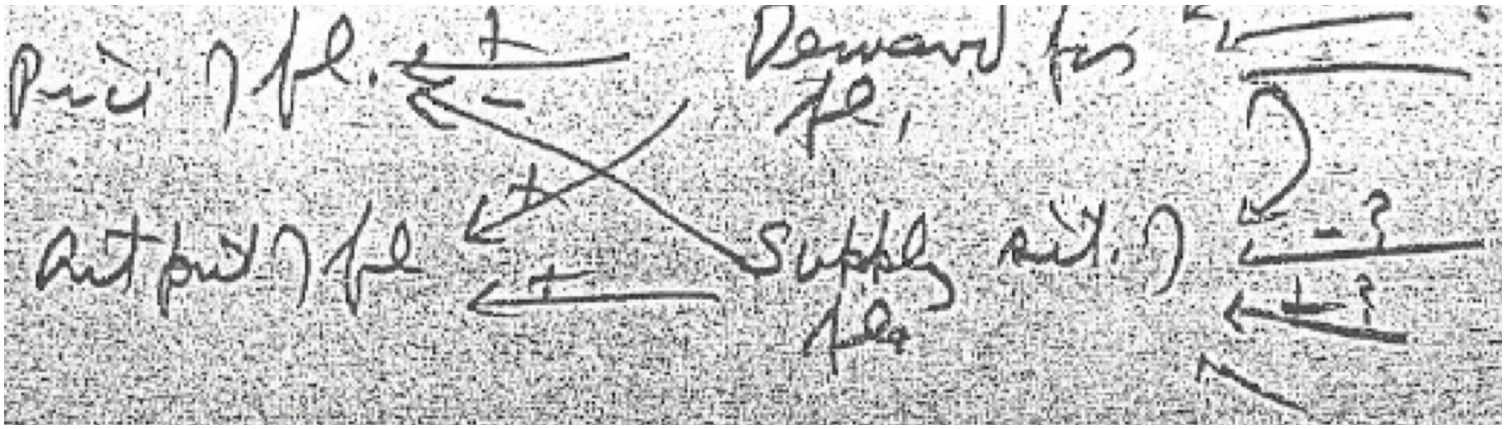


Figure: Recognize these?

QJE Rejects

- QJE misses a chance to make history and rejects his paper proving an IV estimator
- Sticks his proof in Appendix B of 1928 book,
  The Tariff on Animal and Vegetable Oils
- His work on IV is ignored, and is then rediscovered 15 years later (e.g., Olav Reiersøl).
- James Stock and others have helped correct the record

Sidebar: stylometric analysis

- Long standing question was who *wrote* Appendix B? Answer according to Stock and Trebbi (2003) using stylometric methods is that Philip *wrote* it.
- But who invented it? It was collaborative, but Sewell acknowledged he didn't know how to handle endogeneity and simultaneity (that was Philip)

Two things to keep in mind

- Traditionally econometric models tended to assume treatments effects, δ, were constant across units
- But a key contribution of Angrist and Imbens' work in the mid-1990s was exploring what happens when we relax that assumption
- The more contemporary approach does not take the position that treatment effects are the same for everyone
- This turned out to have a dramatic effect on interpretation

Constant treatment effects

- Constant treatment effects (i.e., δ 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 $\delta_i = \delta_{-i} = \beta$ for all i units

Heterogenous treatment effects

- Heterogeneous treatment effects (i.e., δ_i varies across individual units)
  - → Heterogeneous treatment effects means that the $ATE \neq ATT \neq ATU$ because δ_i differs across the population
  - → This is equivalent to assuming the coefficient, δ_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 girls; 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

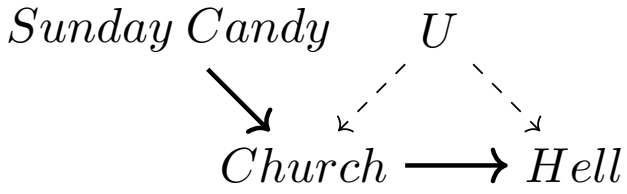
What are we missing?

“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”

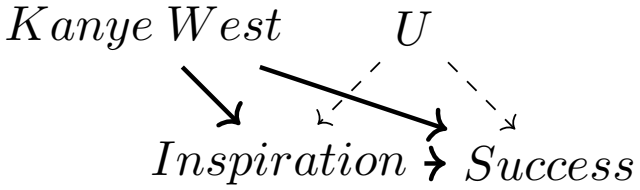
Sunday Candy DAG



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
  - Covariates
  - Presentation suggestions
- Leniency design application
  - Introduction to leniency designs
  - Marginal Treatment Effects
- Other common applications
  - Lottery designs
  - Fuzzy RDD

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. 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, $\widehat{S}$:

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

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

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

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
- 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

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

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

- 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

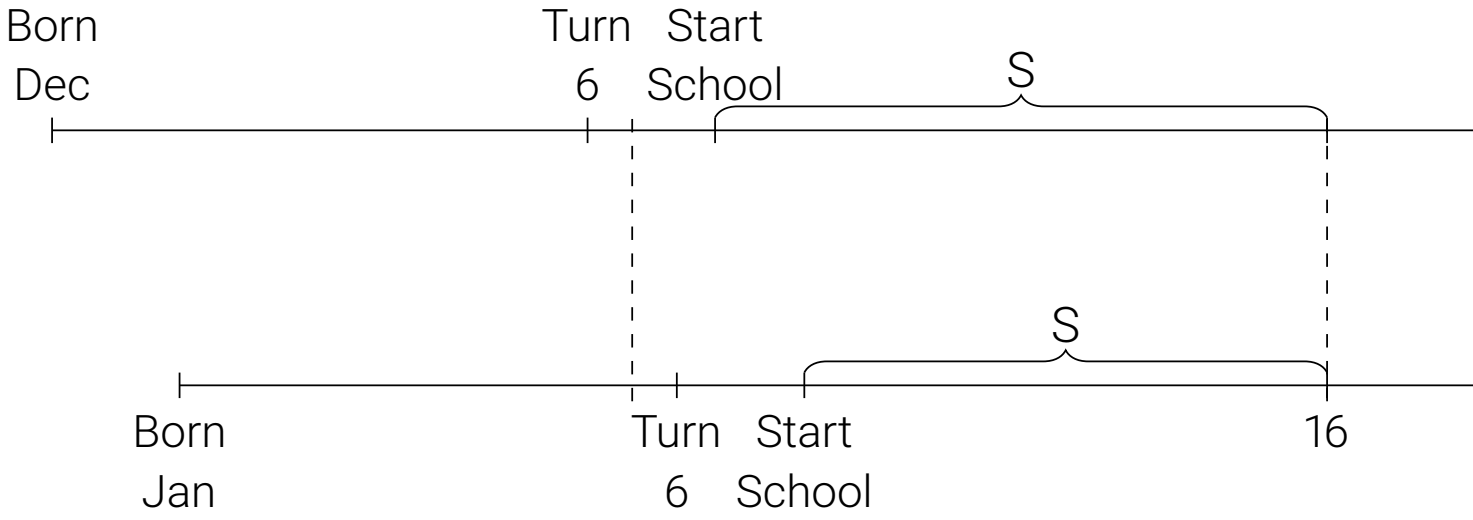
Somewhat inspiring to hear how Angrist reframed the weak instrument problem which his paper with Krueger brought into crisp focus

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):



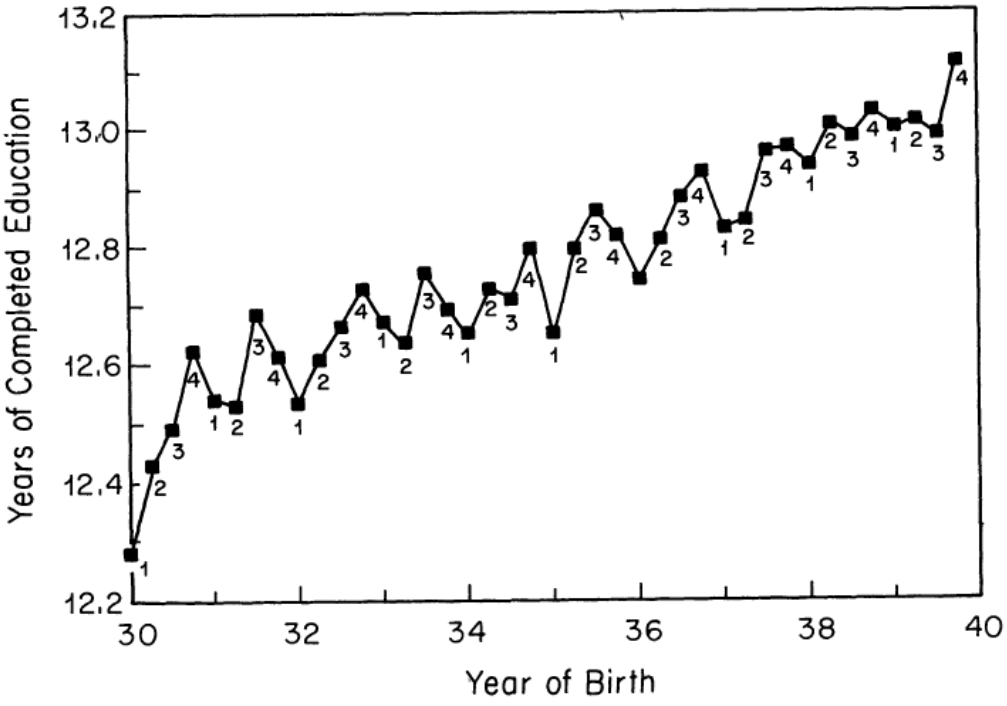
If you’re born in the fourth quarter, you hit 16 with more schooling than those born in the first quarter

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

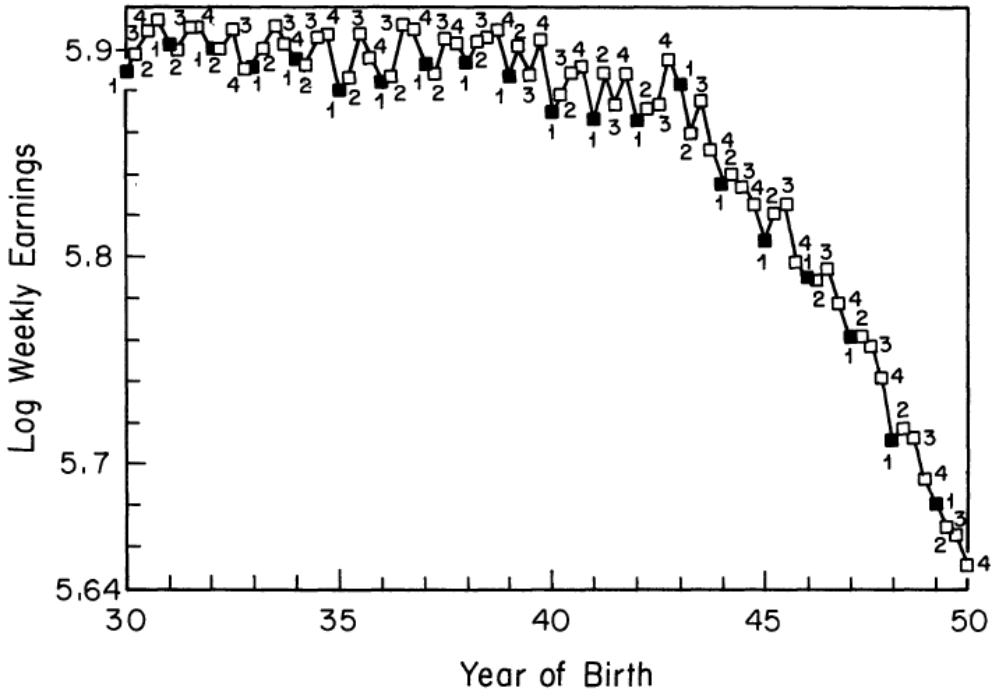
First Stage

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[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]

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.[18] 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
  - $\rightarrow$ What's the intuition here? The effect of quarter of birth may vary by birth year or by state
  - $\rightarrow$ By interacting their instrument with variables, they are "saturating" their 2SLS regression model (more on that later)
- It reduced the standard errors, but that comes at a cost of potentially having a weak instruments problem

More instruments

Table VII 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[a]

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

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

Weak Instruments

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

Matrices and instruments

- The causal model of interest is:

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

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

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

Weak instruments and bias towards OLS

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

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

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

Weak instruments and 2SLS bias towards OLS

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

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

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

Weak instruments and bias towards OLS

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

Weak Instruments - Adding More Instruments

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

Weak instrument problem

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

Bound, Jaeger and Baker (1995)

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

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

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

Adding instruments in Angrist and Krueger

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

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

Adding instruments in Angrist and Krueger

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

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

IV advice: Weak instruments

- Excellent review by Keane and Neal (2021) “A Practical Guide to Weak Instruments” 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
  - → 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 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

Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
- Heterogeneity
  - Local average treatment effects
  - Covariates
  - Presentation suggestions
- Leniency design application
  - Introduction to leniency designs
  - Marginal Treatment Effects
- Other common applications
  - Lottery designs
  - Fuzzy RDD

Internal and external validity

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

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

Constant vs heterogenous treatment effects

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

Potential treatment concept

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

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

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

Identification

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

SUTVA

SUTVA with respect to IV

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

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

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

Independence assumption

Independence assumption

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

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

Independence

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

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

Independence

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

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

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

Exclusion Restriction

Exclusion Restriction

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

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

Exclusion restriction

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

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

- Rewrite switching equation:

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

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

Know your treatment and instrument assignment mechanism

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

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

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

Strong first stage

Nonzero Average Causal Effect of Z on D

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

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

Monotonicity

Monotonicity

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

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

Local average treatment effect

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

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

Estimand

Instrumental variables (IV) estimand:

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

Local Average Treatment Effect

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

LATE and subpopulations

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

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

Never-Takers

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

Defier

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

Complier

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

Always-taker

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

Monotonicity Ensures that there are no defiers

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

LATE is not the ATE

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

Sensitivity to assumptions: exclusion restriction

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

Sensitivity to assumptions: Monotonicity violations

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

IV with covariates

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

Saturated regression models

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

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

Identification with covariates and 2SLS

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

2SLS estimand with covariates

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

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

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

Covariates in 2SLS models

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

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

Saturate and weight

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

Covariates going forward

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

Practical advice

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

IV advice: Pictures

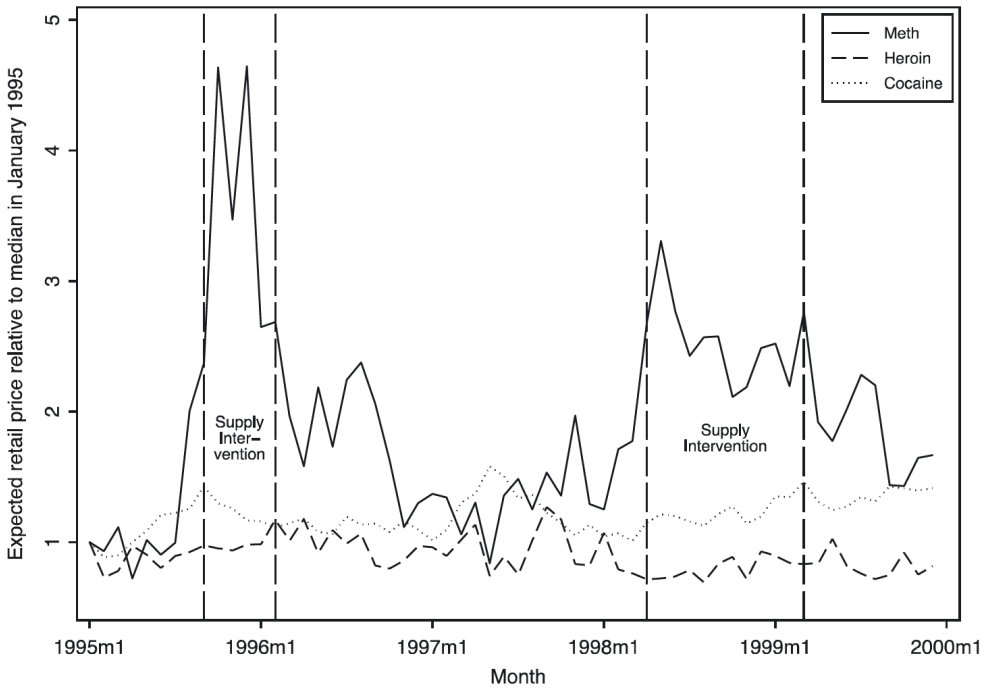
Present your main results in beautiful pictures

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

IV advice: Pictures

FIGURE 3

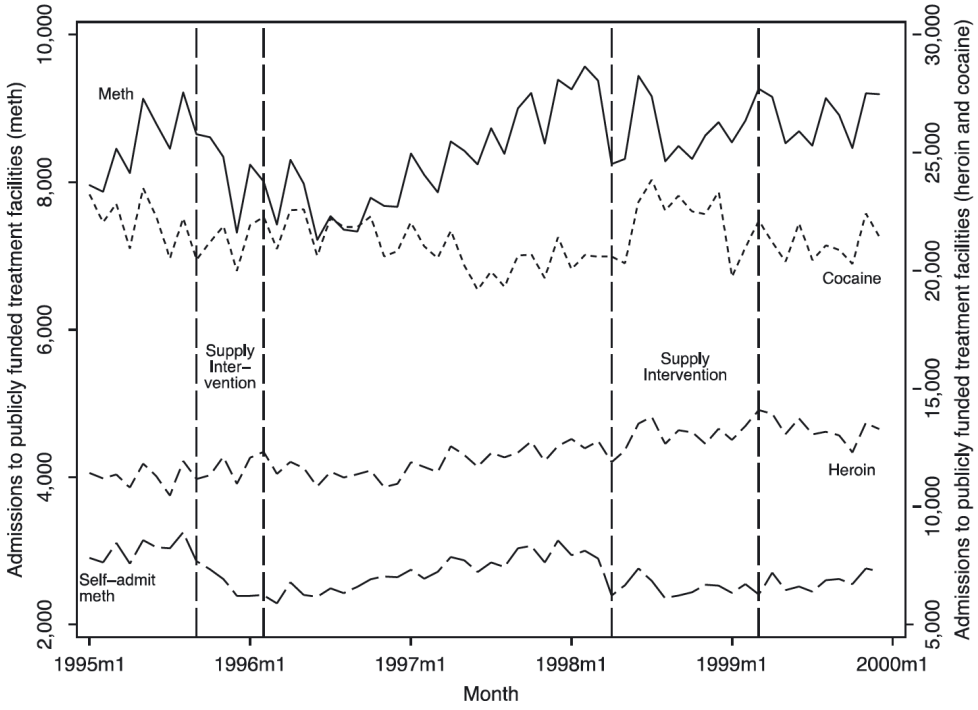
Ratio of Median Monthly Expected Retail Prices of Meth, Heroin, and Cocaine Relative to Their Respective Values in January 1995, STRIDE, 1995–1999



IV advice: Pictures

FIGURE 5

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

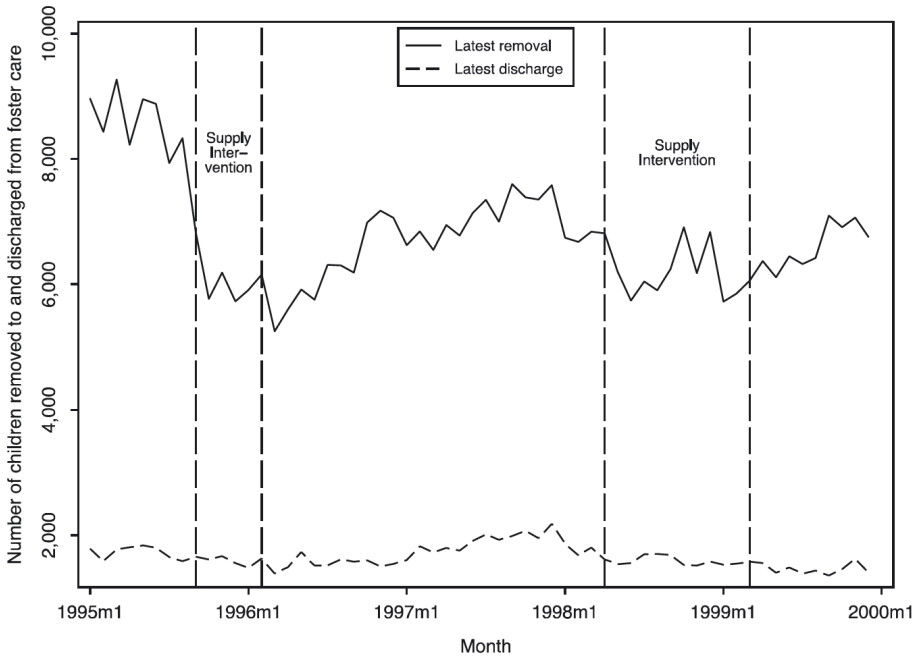


Notes: Authors' calculations from TEDS. Arizona, the District of Columbia, Kentucky, Mississippi, West Virginia, and Wyoming are excluded because of poor data quality. Patients can report the use of more than one drug.

IV advice: Pictures

FIGURE 4

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



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

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
Mean Dependent Variable	6.262	6.262
Std. Dev. Dependent Variable	0.444	0.444

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 δ 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 heterogenous 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
  - Covariates
  - Presentation suggestions
- Leniency design application
  - Introduction to leniency designs
  - Marginal Treatment Effects
- Other common applications
  - Lottery designs
  - Fuzzy RDD

Various types of IV designs

- Because of its broad usefulness, there are several IV-within-IV type models
- RCTs with noncompliance, fuzzy regression discontinuity, and Bartik instruments are such examples
- Not enough time to cover them all, so I'm going to focus on another one: leniency design or "judge fixed effects"
- I'll use this example to illustrate both new estimators as well as new things to learn

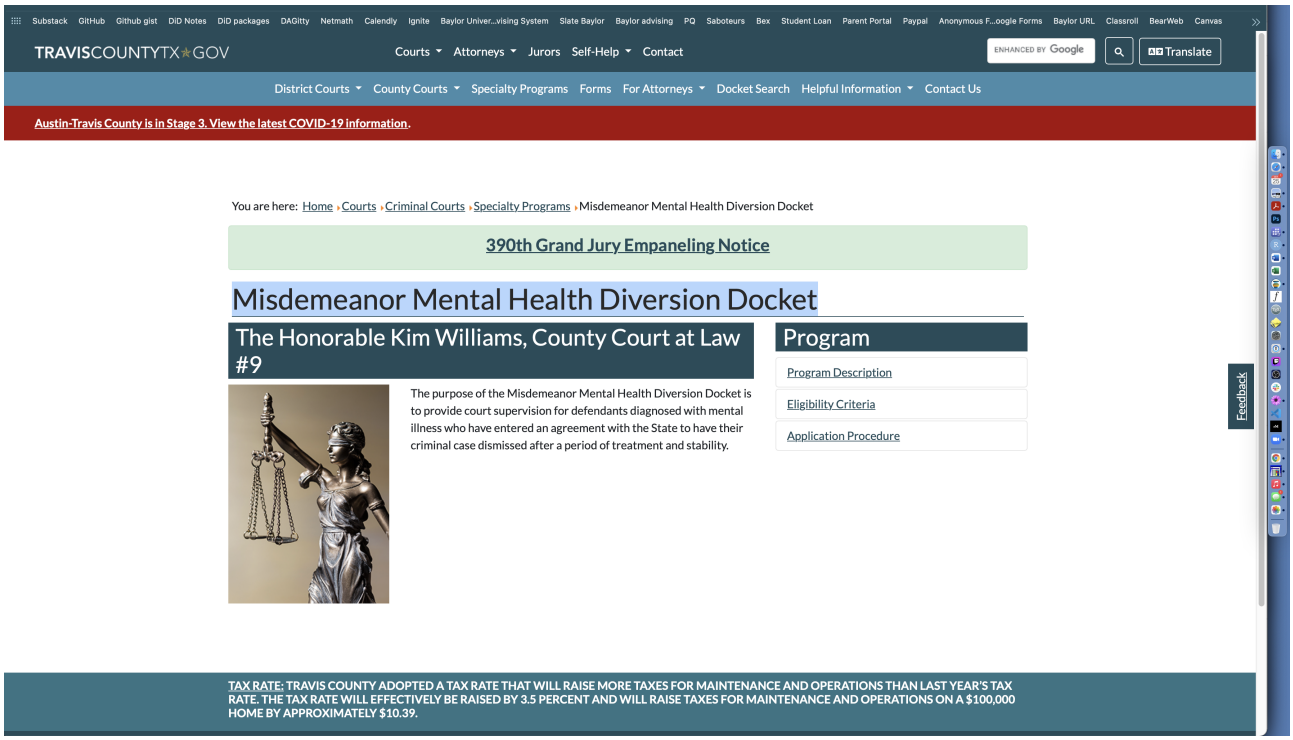
Mental health court and recidivism

- Mental health courts are in around 600 counties across the US
- Diversion courts that divert mentally ill inmates to a specialized court whose aim is to dismiss charges in exchange for treatment adherence
- High concentration of mentally ill in corrections, so important potential policy
- First evidence using quasi-random assignment within a jurisdiction
- Conclusion: mental health courts increase repeat offending

Jails and prisons are the mental health hospitals of last resort

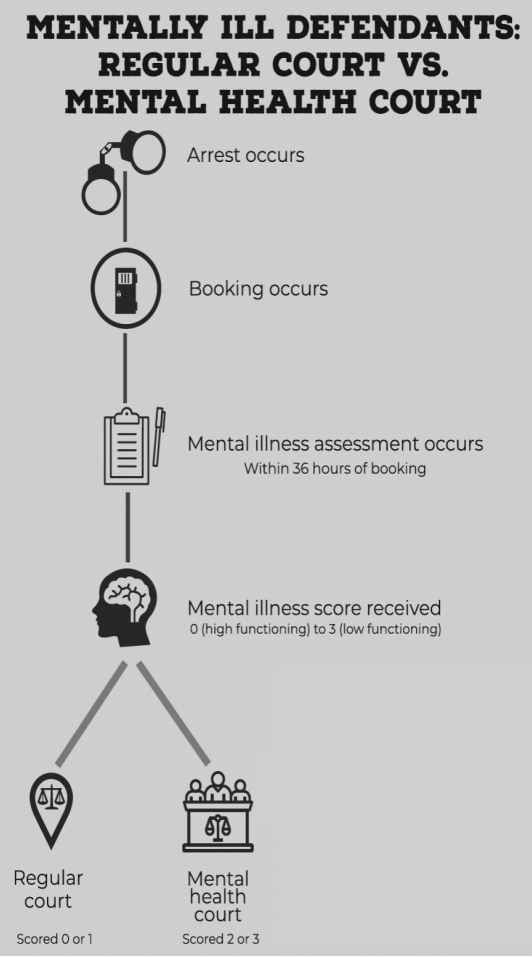
- Inmates are 64% or up to 12 times more likely to have a mental illness than the general community (Prins, 2014)
  - → In most states, there is at least one jail or prison that houses more mentally ill individuals than the largest psychiatric hospital in the area (Torrey et al. 2014)
  - → ∼20 percent of inmates in our data require treatment for their mental illness
- On any given day, 7 percent of inmates with mental illness are experiencing severe symptoms such as psychosis, delusions or suicidal thoughts (Corrections Officers Receive Specialized Mental Health Training, 2020)
  - → One study found a 77% prevalence rate of mental illness among inmates who attempted suicide (Goss et al. 2002)

Misdemeanor Mental Health Diversion Docket



Travis county correctional complex

From booking to mental health court



- Each inmate is randomly assigned a therapist who interviews them for 15 minutes within 36 hrs of booking
- Therapists assign a score (0-3) measuring the severity of mental and behavioral health symptoms
- Inmates with no (0) or mild (1) functioning related symptoms skip MHC and go to typical courts
- Inmates with moderate (2) or severe (3) symptoms go to MHC

What is a leniency design?

- Assume that in order to be assigned to counsel, an inmate must first be seen by a clinician who after interviewing them scores the severity of their symptoms (high/low)
- If symptoms were a blood test, then it wouldn't matter who saw the inmate – they'd all give the same score in counterfactual even if randomized
- But symptoms are based on observation, interpretation and professional judgment, and they're seeing them in high volume daily for only 15 minutes usually
- Leniency uses the clinicians average tendency to give inmates high scores as an instrument for what score they did give an inmate

IV Assumptions: (1) Independence

Independence – director of inmate mental health has explained that each day, they alphabetize clinicians and then one to one assign to inmates as they arrive (similar to Kremer and Miguel's deworming randomization)

Kind of interesting – I often wonder if since independence only requires that instruments be assigned independent of potential treatment status if that therefore means randomization is *necessary*, as opposed to just sufficient

IV Assumptions: (2) SUTVA

SUTVA – an inmate's assignment to mental health court is based on their own therapist, not someone else's

IV Assumptions: (3) Exclusion

Exclusion – randomized therapist during assessment can only impact repeat offending via assignment to mental health court

We hypothesized that if the scores are used to deliver services to inmates while in jail, then this could violate exclusion

One possibility is if high scores are associated with other treatments while in jail, such as housing, but for repeat offending upon release, it's hard to fathom what this would be doing.

IV Assumptions: (4) Non-zero first stage

Non-zero first stage – being assigned a therapist with a high average score significantly raises the probability the inmate gets a high score too

IV Assumptions: (5) Monotonicity

Monotonicity – if clinician A strictly gives more high scores than clinician B, then any time clinician B would have given a high score, clinician A would have as well (they do not change places in strictness)

Calculating the residualized leave-one-out mean

1. Regress observed *MHC* onto a vector of time controls (day of year time fixed effects)
2. Calculate the residual, $\widetilde{D}_{dkt}$, from this regression;['
3. Use the residualized propensity to recommend mental health court rate to calculate the therapist recommendation instrument $\widetilde{Z}_{cl}$ as a leave-one-out mean rate of mental health court associated with each randomly assigned therapist l and inmate c

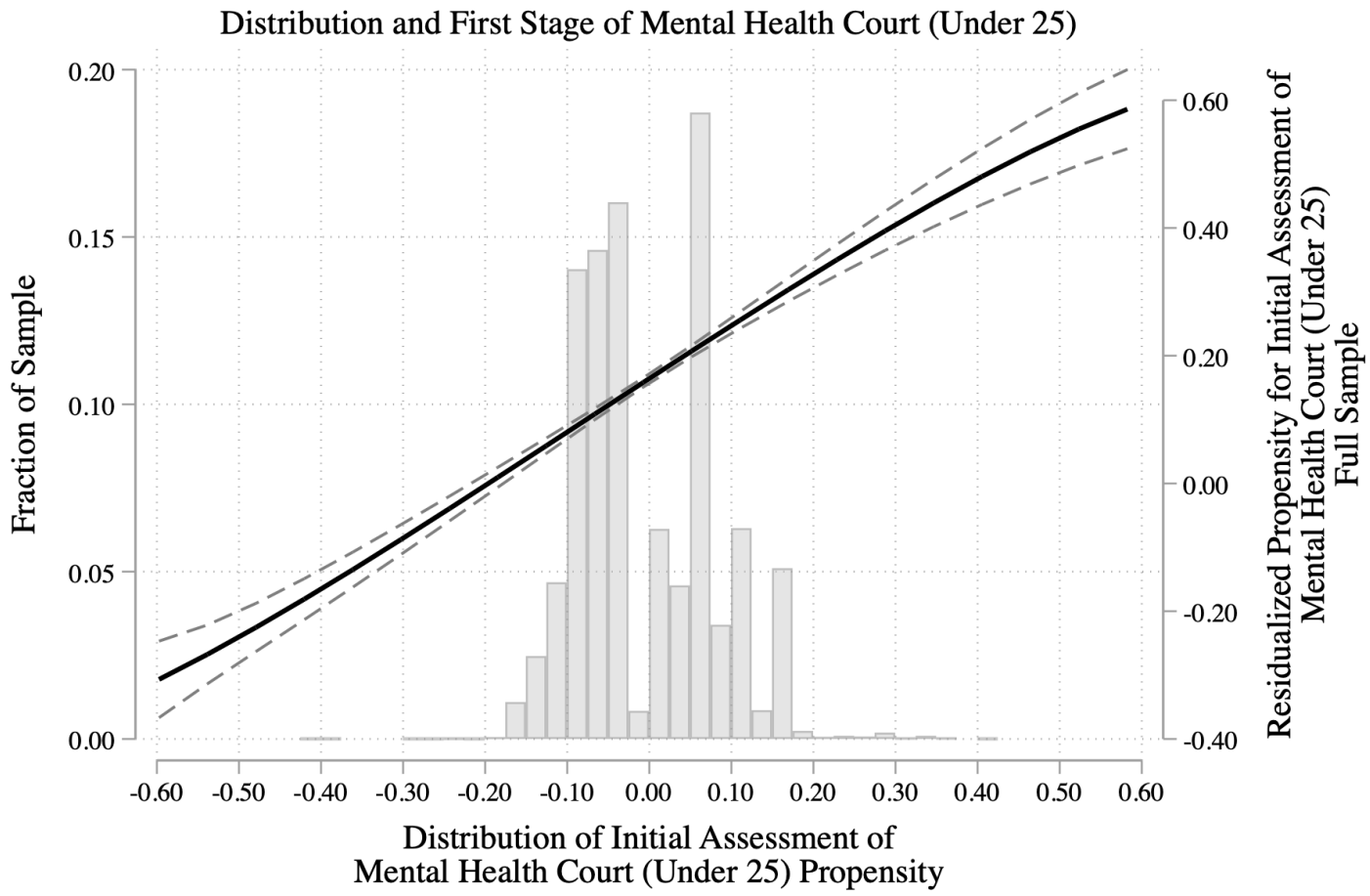
$$\begin{aligned} \widetilde{Z}_{cl} &= \left(\frac{1}{n_l - n_c}\right)\left(\sum_{k=0}^{n_l} \widetilde{D}_{dkt} - \sum_{k \in \{c\}} \widetilde{D}_{dkt}\right) \\ &= \frac{1}{n_l - 1} \sum_{k \neq c}^{n_l - 1} \widetilde{D}_{dkt} \end{aligned} \tag{1}$$

2SLS estimating equations

$$
\begin{aligned}
MHC_{dct} &= \beta \widetilde{Z}_{cl} + \psi X_{dct} + \tau_t + \varpi_{dct} && (2) \\
Y_{dct} &= \delta \widehat{MHC}_{dct} + \gamma X_{dct} + \tau_t + \varepsilon_{dct} && (3)
\end{aligned}
$$

where Y is the outcome of interest (e.g., repeat offending), MHC is an indicator equalling 1 if the inmate was assigned to the mental health court; X are pre-court controls; τ_t are time fixed effects; $\widetilde{Z}$ is the residualized "leave-one-out-mean" average assignment to mental health court and errors are at the end of each equation.

Visualizing residualized leave-one-out mean



First stage

Table: First Stage Regressions for Initial Assessment of Mental Health Court (Under 25)

	(1)	(2)
Z: Clinician's Leave-Out Mean Mental Health Score	0.627***	0.588***
	(0.094)	(0.092)
Kleibergen-Paap F	44.2100	41.0831
Time Fixed Effects	Yes	Yes
Baseline Controls	No	Yes
Observations	9,532	9,532

We report the first stage results of a linear probability model with outcome of interest being the initial assessment of an inmate's mental health being moderate to severe relative to none or mild. The propensity to assign a high score is estimated using data from other cases assigned to the clinician following the procedure described in the text. Column (1) shows the results by controlling only for day-of-week-month fixed effects, whereas Column (2) also includes 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 reported (which is equivalent to the effective F-statistic

Balance

Table: Balance of Instrument and Inmate Characteristics for Mental Health Court (Under 25)

	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	-0.087	-0.012	0.100	(0.000)	(0.000)
Inmate Characteristics					
Asian	0.013	0.011	0.011	(0.775)	(0.794)
Black	0.251	0.274	0.252	(0.329)	(0.914)
Race other	0.002	0.001	0.000	(0.221)	(0.180)
Hispanic	0.367	0.355	0.351	(0.555)	(0.518)
Male	0.708	0.723	0.684	(0.569)	(0.353)
Age at booking	21.053	21.069	21.000	(0.808)	(0.532)
Prior offense w/in 365 days	0.361	0.435	0.378	(0.302)	(0.628)
Number of offenses per booking	1.566	1.598	1.601	(0.555)	(0.275)
First time in jail	0.102	0.086	0.027	(0.690)	(0.024)
Prior treatment	0.094	0.128	0.060	(0.490)	(0.473)
Prior medications	0.090	0.108	0.049	(0.688)	(0.311)
Prior hospitalization	0.042	0.071	0.038	(0.297)	(0.993)
Homeless	0.021	0.043	0.015	(0.202)	(0.651)
Jobless	0.075	0.095	0.031	(0.593)	(0.229)

Data is from a large county correctional complex.

Time fixed effects include day-of-week-month fixed effects.

Clinician and inmate two-way clustered standard errors shown in parentheses.

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

Strict monotonicity test

- Frandsen, Lefgren and Leslie (2020) test joint null of exlcusion and monotonicity
- If you can argue one hold, then rejections mean the other doesn't hold
- We think exclusion plausibly holds given the research design, so rejecting the null speaks to strict monotonicity violations

Strict monotonicity test

- This test requires that the ATE among individuals who violate monotonicity be identical to the ATE among some subset of individuals who satisfy it.
- Their proposed test is based on two observations:
  1. the average outcomes, conditional on therapist assignment, should fit a continuous function of therapist propensities, and
  2. the slope of that continuous function should be bounded in magnitude by the width of the outcome variable's support
- Get ready to have your mind blown – spoiler alert, strict monotonicity doesn't hold even remotely in our data

Strict monotonicity

Table: Frandsen, Lefgren, Leslie (2020) Test of Joint Null of Exclusion and Monotonicity

Outcome	FLL P-Value
Recid after current booking	0.000
Recid within 1 year	0.000
Count of future recidivism	0.000
LOS	0.000
Days to recidivism	0.000
Next offense felony	0.000
Next booking mental health score improves	0.007

This table presents results from the test proposed in Frandsen, Lefgren, and Leslie (2020) for the joint null hypothesis that the monotonicity and exclusion restrictions hold. A failure to reject the null implies that we cannot reject the hypothesis that the monotonicity and exclusion restrictions jointly hold. This test was implemented in Stata via the package testjfe (Frandsen, 2020).

Average monotonicity

- We think maybe strict monotonicity doesn't hold because of the unbelievable discretion these therapists have, plus their inexperience – leniency designs give and the take away
- Historically, before Frandsen, Lefgren and Leslie (2020), though, people would test for monotonicity using a more ad hoc test
- You'd look at the first stage in subsets of the data (by covariates) and check if the sign was the same
- Personally, it seems almost impossible to fail this test

Average monotonicity

Table: Average Montonicity for Initial Assessment of Mental Health Court (Under 25)

	Male	Female	Black	White	Hispanic
	(1)	(2)	(3)	(4)	(5)
Z: Clinician's Leave-Out Mean Mental Health Score	0.523***	0.769***	0.579***	0.585***	0.442***
	(0.096)	(0.145)	(0.119)	(0.109)	(0.090)
Observations	6,717	2,815	2,468	6,946	3,411
Time Fixed Effects	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes

This table reports the first stage results by subsamples as listed in the column headers, which serves as informal evidence of average monotonicity if the estimate is significant across all subsamples. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

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 clinicians and with weak instruments, this creates finite sample bias for 2SLS due to *many instruments*
- 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

Alternative to 2SLS: LIML

- 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

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

Alternative to 2SLS: JIVE

- But JIVE can be extremely biased with numerous *covariates*.
- 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 inmate.
- This will cause JIVE to be biased when the number of covariates is large as is the case in our context – we have 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.

Main results

Table: Effects of Initial Assessment of Mental Health Court (Under 25) on Health Outcomes

	OLS		2SLS		JIVE	
	(1)	(2)	(3)	(4)	(5)	(6)
Recid after current booking	0.149***	0.105***	0.424**	0.403**	0.469***	0.444***
	(0.019)	(0.012)	(0.193)	(0.166)	(0.070)	(0.073)
			[0.065, 0.784]	[0.095, 0.711]		
Recid within 1 year	0.100***	0.110***	0.404***	0.465***	0.390***	0.481***
	(0.017)	(0.017)	(0.126)	(0.147)	(0.079)	(0.088)
			[0.144, 0.664]	[0.163, 0.738]		
Count of future recidivism	1.358***	1.186***	2.152**	2.261***	2.835***	2.999***
	(0.268)	(0.223)	(0.878)	(0.821)	(0.335)	(0.359)
			[0.516, 3.787]	[0.573, 3.789]		
LOS	6.796***	5.805***	5.263*	4.910*	7.665***	7.671***
	(0.967)	(0.986)	(2.947)	(2.880)	(2.459)	(2.526)
			[-0.800, 10.748]	[-1.010, 10.267]		
Days to recidivism	-41.232***	-36.629***	66.584	47.245	83.932*	84.951
	(9.828)	(7.895)	(82.581)	(61.076)	(49.133)	(56.420)
			[-135.476, 220.150]	[-101.952, 148.700]		
Next offense felony	0.017	0.003	0.106*	0.116**	0.123***	0.130**
	(0.011)	(0.011)	(0.060)	(0.055)	(0.047)	(0.052)
			[-0.018, 0.218]	[-0.007, 0.207]		
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Baseline Controls	No	Yes	No	Yes	No	Yes

This table reports the ordinary least squares, two-stage least squares, and jacknived instrumental variables (Angrist et al 1999) estimates of the impact of a clinician's initial assessment of a most severe mental health rating on inmates' subsequent mental health. The outcome variables of interest are given in each row along with the corresponding estimates of the impacts of an initial assessment of a most severe mental health rating. Two-stage least squares specifications instrument for severe mental health rating using a clinician leniency measure that is estimated using data from other cases assigned to a clinician as described in the text. We include day-of-week-month fixed effects for all specifications and baseline controls for Columns (2), (4), and (6). The clinician and inmate robust two-way clustered standard errors are shown in parentheses. For the 2SLS estimates, confidence intervals based on inversion of the Anderson-Rubin test are shown in brackets. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

Alternative to 2SLS: UJIVE

- To resolve this, we accompany our 2SLS and JIVE estimates with models that are more robust for large number of covariates as well as large number of instruments.
- Our 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.
- You can implement this using Kolesar's matlab code (and we do in another paper), but it's giving us problems here so I can't report the results

Alternatives: LASSO

- We also present 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.
- These machine models models are designed to minimize the problems of including a large number of *instruments* (columns 3-4) as well as a large number of *controls* (columns 5-6).
- We use the Stata command `lassopack` for its implementation

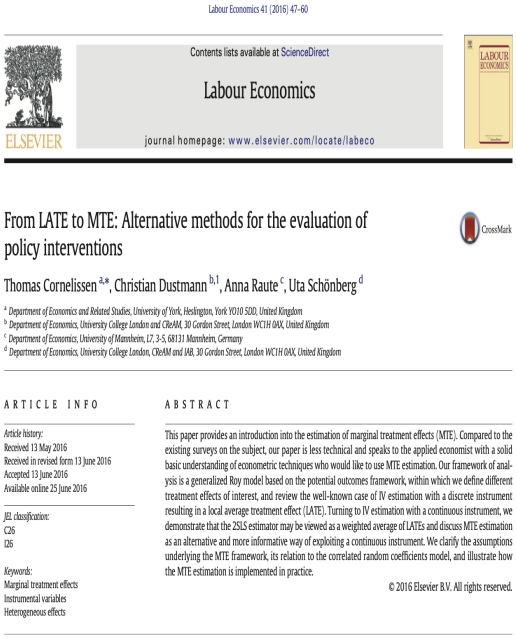
Main results

Table: IVLASSO Results for Initial Assessment of Mental Health Court (Under 25) and Suicidality Outcomes

	LASSO		Post-LASSO	
	(1)	(2)	(3)	(4)
Recid after current booking	0.246	0.553***	0.168	0.327
	(0.252)	(0.211)	(0.234)	(0.322)
Recid within 1 year	0.223*	0.203	0.180	0.160
	(0.115)	(0.132)	(0.112)	(0.134)
Count of future recidivism	2.294*	4.047***	1.824	2.888
	(1.379)	(1.156)	(1.288)	(1.810)
LOS	2.922	1.257	2.810	-1.407
	(2.282)	(2.220)	(2.396)	(4.171)
Days to recidivism	164.253**	113.031	158.323*	113.181
	(76.664)	(104.685)	(82.467)	(87.133)
Next offense felony	0.087	0.172***	0.076	0.171***
	(0.062)	(0.032)	(0.069)	(0.040)
Time fixed effects	Yes	Yes	Yes	Yes
Baseline controls	No	Yes	No	Yes

This table reports the Instrumental Variables LASSO (IVLASSO) estimates of the impact of a clinician's initial assessment of a most severe mental health rating on inmates' subsequent mental health and criminality. The outcome variables of interest are given in each row along with the corresponding estimates of the impacts of an initial assessment of a most severe mental health rating. We include day-of-week-month fixed effects and baseline controls for all specifications; however, the IVLASSO procedure penalizes the controls as well as the instruments and can penalize them to zero as discussed in the text. The IVLASSO procedure is run using both methods: lasso-orthogonalization and post-lasso-orthogonalization. The clinician and inmate robust two-way clustered standard errors are shown in parentheses for IVLASSO. * $p<0.10$, ** $p<0.05$, *** $p<0.01$

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
- Additive separability of treatment effects (i.e., unobserved plus observed heterogeneity)
- Only required if not full support, and we don't have full support

Marginal treatment effects

- In our context, we explore 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 court assignment is based on an individual latent index threshold in which the net benefits of mental health court 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 mental health court occurs when $D_i^* > 0$ otherwise they go to traditional adjudication

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 go to mental health court, 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 mental health court

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 mental health court
- 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

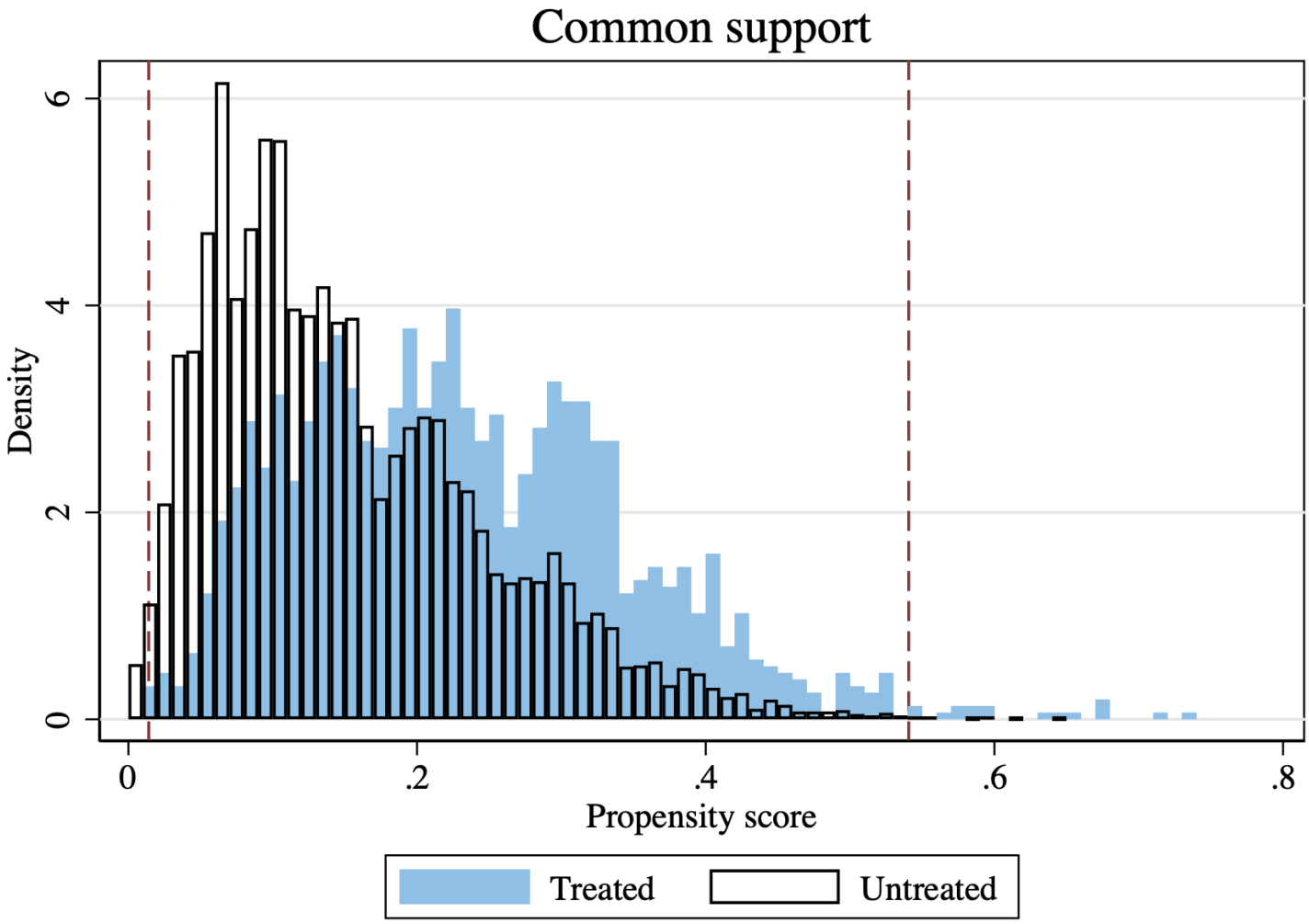
1. Estimate the propensity score using logit (here issues of common support arise as we want it across all cells of Z)
2. Model recidivism as a function of covariates and propensity score with second degree polynomials
3. Calculate the derivative of $\widehat{recidivism}$ with respect to the propensity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

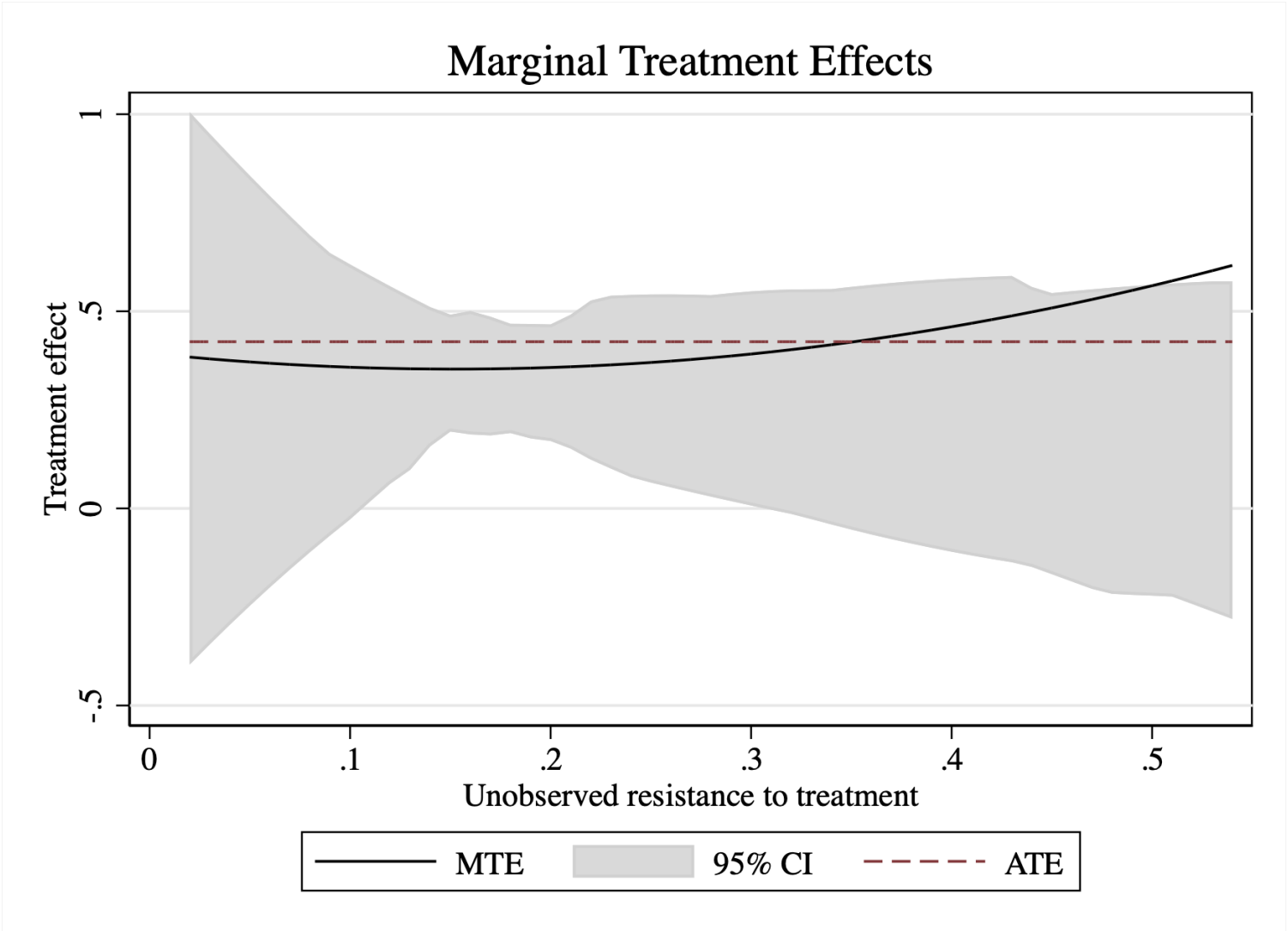
Calculating aggregates

- 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 mental health court (i.e., ATT>ATE>ATU)
- Upward slope means the reverse (i.e., ATU>ATE>ATT)

Common support



MTE and aggregate parameters for recidivism



Roadmap

- Instrumental variables
  - Background
  - Intuition
- Estimators
  - Two Step
  - Weak instruments
- Heterogeneity
  - Local average treatment effects
  - Covariates
  - Presentation suggestions
- Leniency design application
  - Introduction to leniency designs
  - Marginal Treatment Effects
- Other common applications
  - Lottery designs
  - Fuzzy RDD

IV in Randomized Trials

- In many randomized trials, participation is nonetheless voluntary among those randomly assigned to treatment
- Consequently, noncompliance is not uncommon and without correcting for it, creates selection biases
- IV designs may even be helpful when evaluating a randomized trial, even though treatment was randomly assigned
- The solution is to instrument for treatment with whether you “won the lottery” and estimate LATE

Lottery designs

- The instrument is your randomized lottery
- Examples might be randomized lottery for attending charter schools to study effect of charter schools on educational outcomes, or a randomized voucher to encourage the collection of health information
- Recall Thornton (2008) instrumented for getting HIV results to estimate causal effect of learning one was HIV+ on condom purchases
- We'll discuss two papers from 2012 and 2014 evaluating a lottery-based expansion of Medicaid health insurance on Oregon on numerous health and financial outcomes

Overarching question

- What are the effects of expanding access to public health insurance for low income adults?
  - → Magnitudes, and even the signs, associated with that question were uncertain
- Limited existing evidence
  - → Institute of Medicine review of evidence was suggestive, but a lot of uncertainty
  - → Observational studies are confounded by selection into health insurance
  - → Quasi-experimental work often focuses on elderly and children
  - → Only one randomized experiment in a developed country: the RAND health insurance experiment
    - 1970s experiment on a general population
    - Randomized cost-sharing, not coverage itself

The Oregon Health Insurance Experiment

Setting: Oregon Health Plan Standard

- Oregon's Medicaid expansion program for poor adults
- Eligibility
  - → Poor ($<$100% federal poverty line) adults 19-64
  - → Not eligible for other programs
  - → Uninsured $>$ 6 months
  - → Legal residents
- Comprehensive coverage (no dental or vision)
- Minimum cost-sharing
- Similar to other states in payments, management
- Closed to new enrollment in 2004

The Oregon Medicaid Experiment

Oregon held a lottery

- Waiver to operate lottery
- 5-week sign-up period, heavy advertising (January to February 2008)
- Low barriers to sign up, no eligibility pre-screening
- Limited information on list
- Randomly drew 30,000 out of 85,000 on list (March-October 2008)
- Those selected given chance to apply
  - → Treatment at household level
  - → Had to return application within 45 days
  - → 60% applied; 50% of those deemed eligible → 10,000 enrollees

Oregon Health Insurance Experiment

- Evaluate effects of Medicaid using lottery as randomized controlled trial (RCT)
  - → Intent-to-treat: Reduced form comparison of outcomes between treatment group (lottery selected individuals) and controls (not selected)
  - → LATE: IV using lottery as instrument for insurance coverage
    - First stage: about a 25 percentage point increase in insurance coverage
  - → Archived analysis plan
  - → Massive data collect effort – primary and secondary
- Similar to ACA expansion but limits to generalizability
  - → Partial equilibrium vs. General equilibrium
  - → Mandate and external validity
  - → Oregon vs. other states
  - → Short vs. Long-run

Examine Broad Range of Outcomes

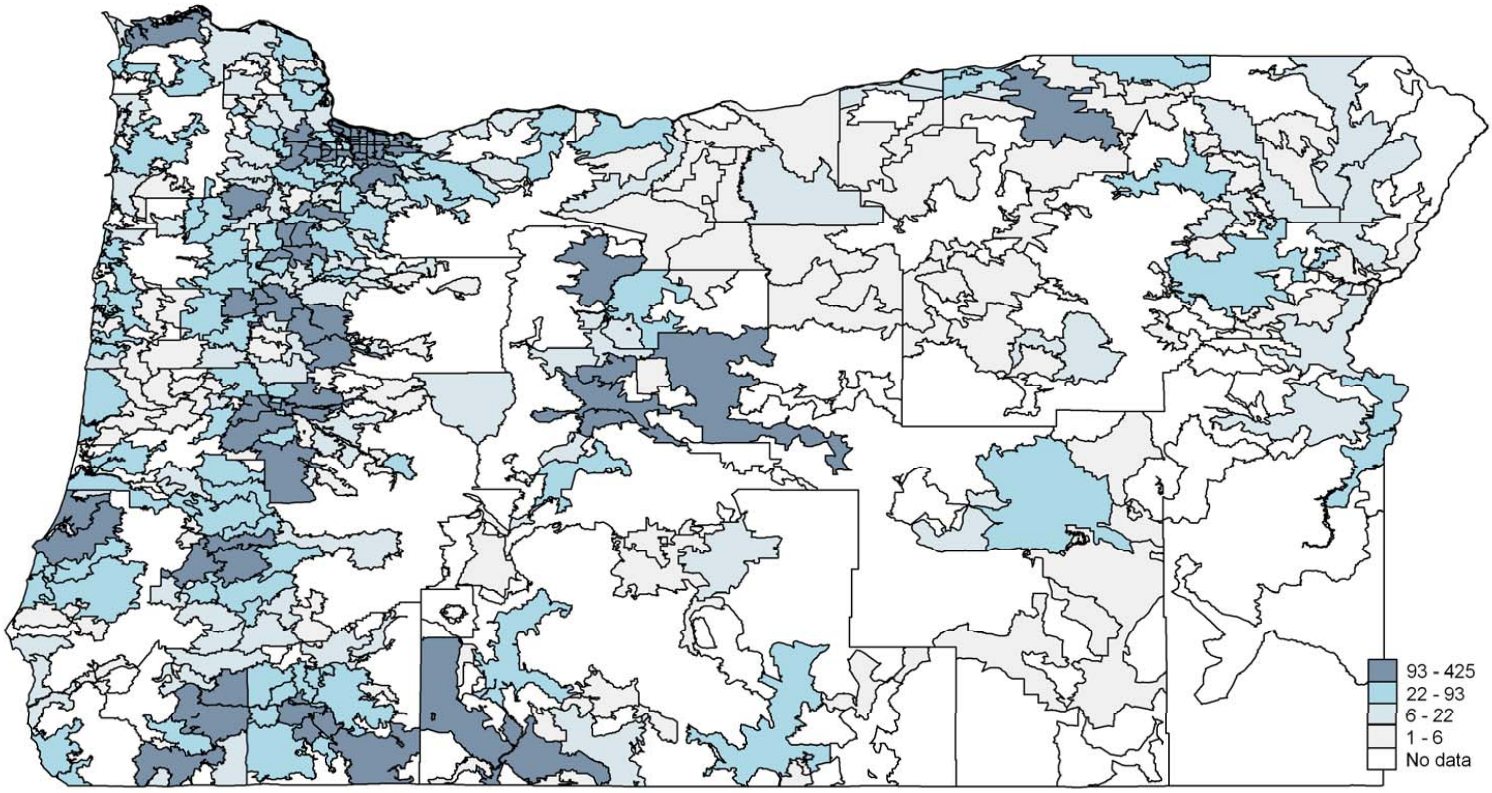
- Costs: Health care utilization
  - → Insurance increases resources (income) and lowers price, increasing utilization
  - → But improved efficiency (and improved health), decreasing utilization ("offset")
  - → Additional uncertainty when comparing Medicaid to no insurance
- Benefits I: Financial risk exposure
  - → Insurance supposed to smooth consumption
  - → But for very low income, is most care *de jure* or *de facto* free?
- Benefits II: Health
  - → Expected to improve (via increased quantity / quality of care)
  - → But could discourage health investments ("*ex ante* moral hazard")

Data

- Pre-randomization demographic information
  - → From lottery sign-up
- State administrative records on Medicaid enrollment
  - → Primary measure of first stage (i.e., insurance coverage)
- Outcomes
  - → Administrative data (∼16 months post-notification): Hospital discharge data, mortality, credit reports
  - → Mail surveys (∼15 months): some questions ask 6-month look-back; some ask current
  - → In-person survey and measurements (∼25 months): Detailed questionnaires, blood samples, blood pressure, body mass index

Lottery List

Distribution Across Zip Codes



Empirical Framework

- They present reduced form estimates of the causal effect of lottery selection

$$Y_{ihj} = \beta_0 + \beta_1 LOTTERY_h + X_{ih}\beta_2 + V_{ih}\beta_3 + \varepsilon_{ihj}$$

- Validity of experimental design: randomization; balance on treatment and control. This is what readers expect

Empirical framework

- They also present IV results because they want to isolate the causal effect of insurance coverage

$$
\begin{aligned}
INSURANCE_{ihj} &= \delta_0 + \delta_1 LOTTERY_{ih} + X_{ih}\delta_2 + V_{ih}\delta_3 + \mu_{ihj} \\
y_{ihj} &= \pi_0 + \pi_1 \widehat{INSURANCE}_{ih} + X_{ih}\pi_2 + V_{ih}\pi_3 + v_
\end{aligned}
$$

- Effect of lottery on coverage: about 25 percentage points
- We have independence guaranteed; now we need exclusion: the primary pathway of the lottery must be via being on Medicaid
  - → Could affect participation in other programs, but actually small
  - → "Warm glow" of winning – especially early
- Analysis plan, multiple inference adjustment

Effect of lottery on coverage (first stage)

	Full sample		Credit subsample		Survey respondents	
	Control mean	Estimated FS	Control mean	Estimated FS	Control mean	Estimated FS
Ever on Medicaid	0.141	0.256	0.135	0.255	0.135	0.290
		(0.004)		(0.004)		(0.007)
Ever on OHP Standard	0.027	0.264	0.028	0.264	0.026	0.302
		(0.003)		(0.004)		(0.005)
# of Months on Medicaid	1.408	3.355	1.352	3.366	1.509	3.943
		(0.045)		(0.055)		-0.09
On Medicaid, end of study period	0.106	0.148	0.101	0.151	0.105	0.189
		(0.003)		(0.004)		(0.006)
Currently have any insurance (self report)					0.325	0.179
						(0.008)
Currenty have private ins. (self report)					0.128	-0.008
						(0.005)
Currently on Medicaid (self report)					0.117	0.197
						(0.006)
Currently on Medicaid					0.093	0.177
						(0.006)

Amy Finkelstein, et al. (2012). “The Oregon Health Insurance Experiment: Evidence from the First Year”, *Quarterly Journal of Economics*, vol. 127, issue 3, August.

Effects of Medicaid

Use primary and secondary data to gauge 1-year effects

- Mail surveys: 70,000 surveys at baseline, 12 months
- Administrative data
  - → Medicaid enrollment records
  - → Statewide Hospital discharge data, 2007-2010
  - → Credit report data, 2007-2010
  - → Mortality data, 2007-2010

Mail survey data

- **Fielding protocol**
  - → ∼70,000 people, surveyed at baseline and 12 months later
  - → Basic protocol: three-stage male survey protocol, English/Spanish
  - → Intensive protocol on a 30% subsample included additional tracking, mailings, phone attempts (done to adjust for non-response bias)
- **Response rate**
  - → Effective response rate = 50%
  - → Non-response bias aways possible, but response rate and pre-randomization measures in administrative data were balanced between treatment and control

Administrative data

- **Medicaid records**
  - → Pre-randomization demographics from list
  - → Enrollment records to assess “first stage” (how many of the selected got insurance coverage)
- **Hospital discharge data**
  - → Probabilistically matched to list, de-identified at Oregon Health Plan
  - → Includes dates and source of admissions, diagnoses, procedures, length of stay, hospital identifier
  - → Includes years before and after randomization
- **Other data**
  - → Mortality data from Oregon death records
  - → Credit report data, probabilistically matched, de-identified

Sample

- 89,824 unique individuals on the waiting list
- Sample exclusions (based on pre-randomization data only)
  - → Ineligible for OHP Standard (out of state address, age, etc.)
  - → Individuals with institutional addresses on list
- Final sample: 79,922 individuals out of 66,385 households
  - → 29,834 treated individuals (surveyed 29,589)
  - → 40,088 control individuals (surveyed 28,816)

Sample characteristics

Variable	Mean	Variable	Mean
Panel A: Full sample			
% Female	0.56	Average Age	41
Panel B: Survey responders only			
Demographics:		*Health Status: Ever diagnosed with:*	
% White	0.82	Diabetes	0.18
% Black	0.04	Asthma	0.28
% Spanish/Hispanic/Latino	0.12	High Blood Pressure	0.40
% High school or less	0.67	Emphysema or Chronic Bronchitis	0.13
% don't currently work	0.55	Depression	0.56
Determinants of eligibility:			
Average hh income (2008)	13,050	% with any insurance	0.33
% below Federal poverty line	0.68	% with private insurance	0.13

Outcomes

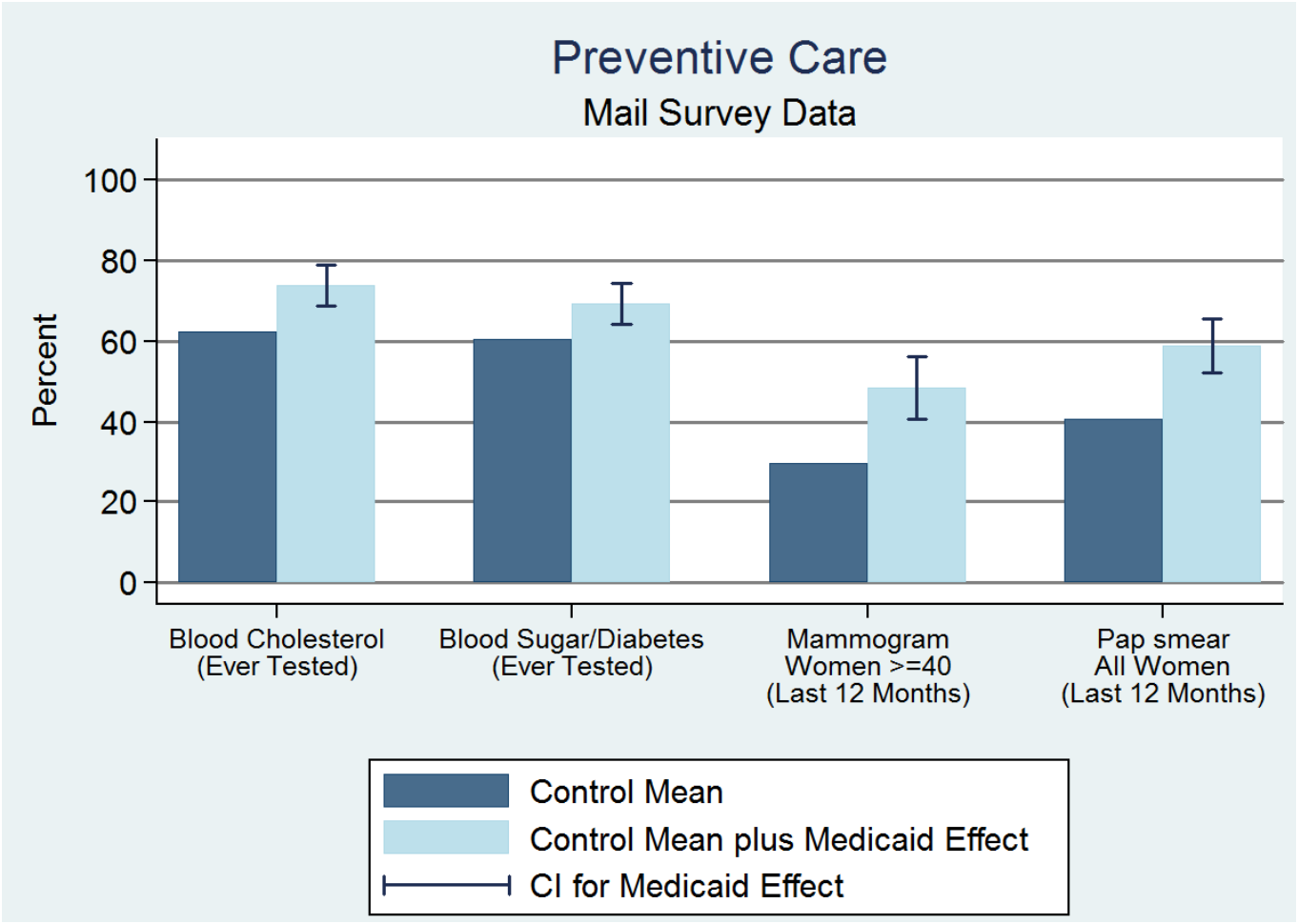
- **Access and use of care**
  - → Is access to care improved? Do the insured use more care? Is there a shift in the types of care being used?
  - → Mail surveys and hospital discharge data
- **Financial strain**
  - → How much does insurance protect against financial strain?
  - → What are the out-of-pocket implications?
  - → Mail surveys and credit reports
- **Health**
  - → What are the short-term impacts on self-reported physical and mental health?
  - → Mail surveys and vital statistics (mortality)

Effect of lottery on coverage

Gaining insurance resulted in better access to care and higher satisfaction with care (conditional on actually getting care)

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Have a usual place of care	49.9%	+9.9%	+33.9%	.0001
Have a personal doctor	49.0%	+8.1%	+28.0%	.0001
Got all needed health care	68.4%	+6.9%	+23.9%	.0001
Got all needed prescriptions	76.5%	+5.6%	+19.5%	.0001
Satisfied with quality of care	70.8%	+4.3%	+14.2%	.001

SOURCE: Survey data


Preventive Care
Mail Survey Data
Percent
100
80
60
40
20
0
Blood Cholesterol
(Ever Tested)
Blood Sugar/Diabetes
(Ever Tested)
Mammogram
Women >=40
(Last 12 Months)
Pap smear
All Women
(Last 12 Months)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

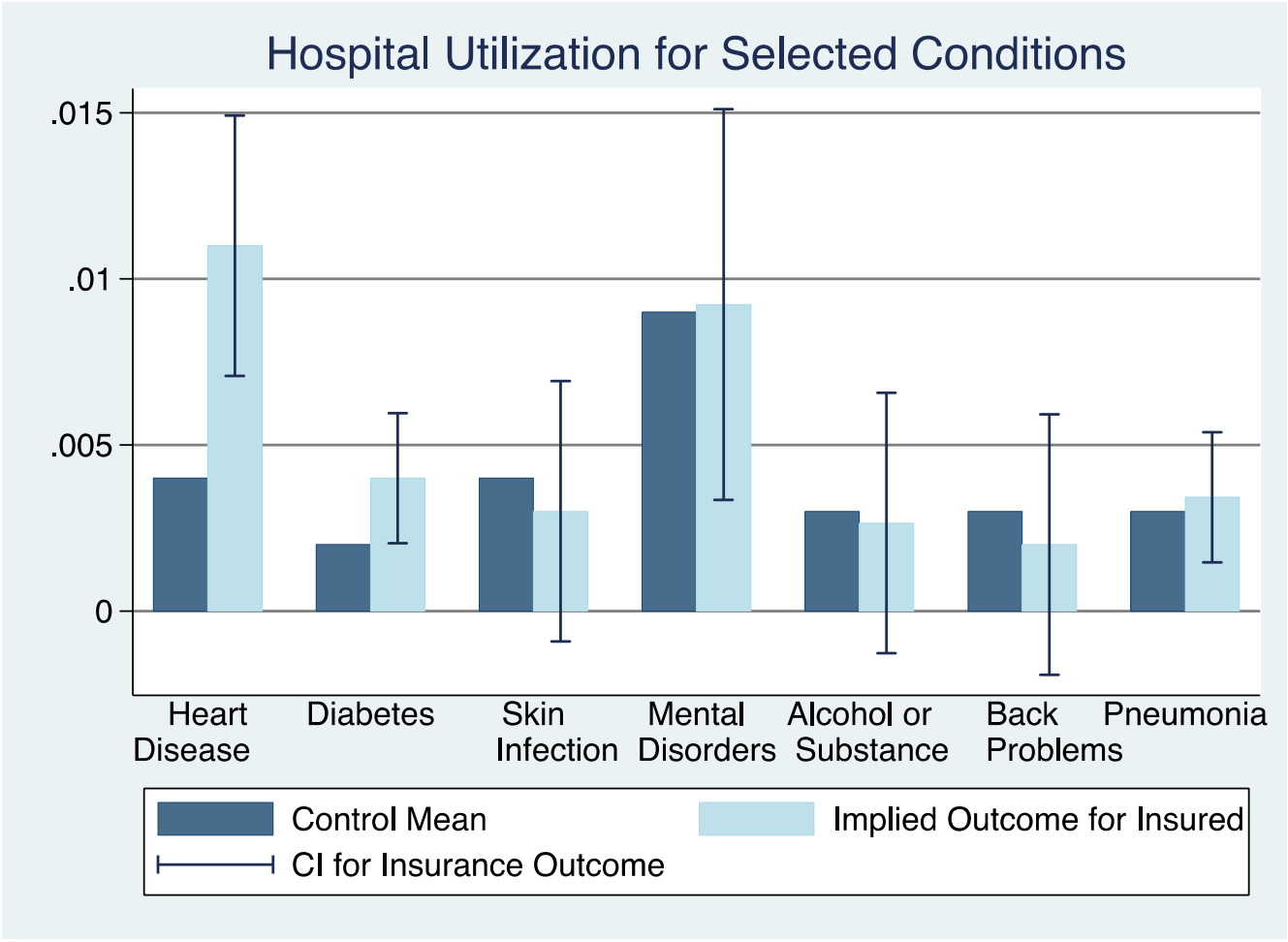
Effect of lottery on coverage

Gaining insurance resulted in increased probability of hospital admissions, primarily driven by non-emergency department admissions

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Any hospital admission	6.7%	+.50%	+2.1%	.004
--Admits through ED	4.8%	+.2%	+.7%	.265
--Admits NOT through ED	2.9%	+.4%	+1.6%	.002

SOURCE: Hospital Discharge Data

Overall, this represents a 30% higher probability of admission, although admissions are still rare events


Hospital Utilization for Selected Conditions
.015
.01
.005
0
Heart Disease
Diabetes
Skin Infection
Mental Disorders
Alcohol or Substance
Back Problems
Pneumonia
Control Mean
Implied Outcome for Insured
CI for Insurance Outcome

Summary: Access and use of care

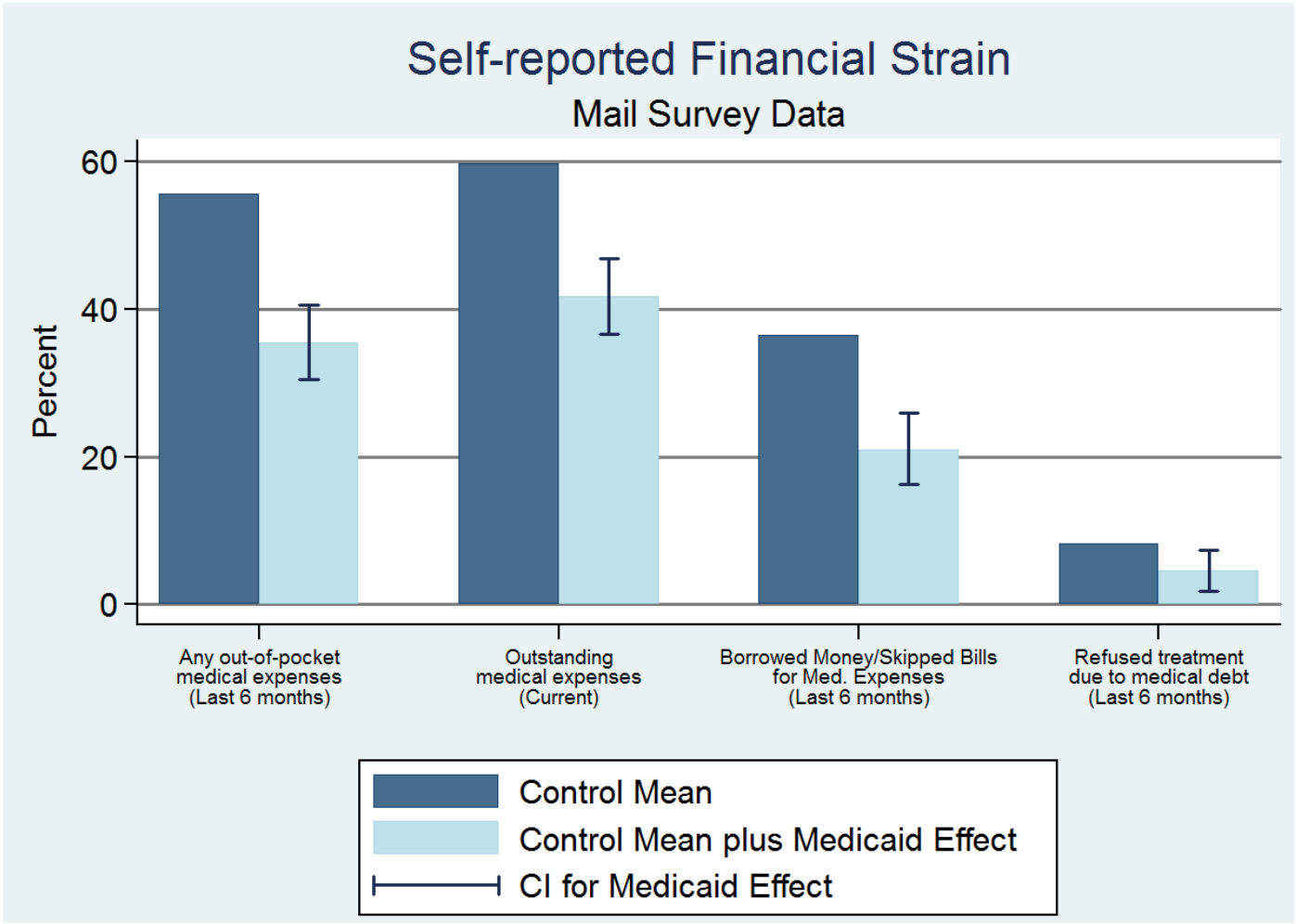
- Overall, utilization and costs went up relative to controls
  - → 30% increase in probability of an inpatient admission
  - → 35% increase in probability of an outpatient visit
  - → 15% increase in probability of taking prescription medications
  - → Total $777 increase in average spending (a 25% increase)
- With this increased spending, those who gained insurance were
  - → 35% more likely to get all needed care
  - → 25% more likely to get all needed medications
  - → Far more likely to follow preventive care guidelines, such as mammograms (60%) and PAP tests (45%)

Results: Financial Strain

Gaining insurance resulted in a reduced probability of having medical collections in credit reports, and in lower amounts owed

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Had a bankruptcy	1.4%	+0.2%	+0.9%	.358
Had a collection	50.0%	-1.2%	-4.8%	.013
--Medical collections	28.1%	-1.6%	-6.4%	.0001
--Non-medical collections	39.2%	-0.5	-1.8%	.455
$ owed medical collections	$1,999	-$99	-$390	.025

Source: Credit report data


Self-reported Financial Strain
Mail Survey Data
Percent
60
40
20
0
Any out-of-pocket medical expenses (Last 6 months)
Outstanding medical expenses (Current)
Borrowed Money/Skipped Bills for Med. Expenses (Last 6 months)
Refused treatment due to medical debt (Last 6 months)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Summary: Financial Strain

- Overall, reductions in collections on credit reports were evident
  - → 25% decrease in probability of a medical collection
  - → Those with a collection owed significantly less
- Household financial strain related to medical costs was mitigated
  - → Substantial reduction across all financial strain measures
  - → Captures “informal channels” people use to make it work
- Implications for both patients and providers
  - → Only 2% of bills sent to collections are ever paid

Results: Self-reported health

Self-reported measures showed significant improvements one year after randomization

	CONTROL	RF Model (ITT)	IV Model (LATE)	P-Value
Health good, v good, excellent	54.8%	+3.9%	+13.3%	.0001
Health stable or improving	71.4%	+3.3%	+11.3%	.0001
Depression screen NEGATIVE	67.1%	+2.3%	+7.8%	.003
CDC Healthy Days (physical)	21.86	+.381	+1.31	.018
CDC Healthy Days (mental)	18.73	+.603	+2.08	.003

Source: Survey data

Summary: Self-reported health

- Overall, big improvements in self-reported physical and mental health
  - → 25% increase in probability of good, very good or excellent health
  - → 10% decrease in probability of screening for depression
- Physical health measures open to several interpretations
  - → Improvements consistent with findings of increased utilization, better access, and improved quality
  - → BUT in their baseline surveys, results appeared shortly after coverage (∼2/3rds magnitude of full result)
  - → May suggest increase in *perception* of well-being rather than physical health
- Biomarker data can shed light on this issue

Discussion

- At 1 year, found increases in utilization, reductions in financial strain, and improvements in self-reported health
  - → Medicaid expansion had benefits and costs – didn't "pay for itself"
  - → Confirmed biases inherent in observational studies – would have estimated bigger increases in use and smaller improvements in outcomes
- Policy-makers may have different views on value of different aspects of improved well-being
  - → "I have an incredible amount of fear because I don't know if the cancer has spread or not."
  - → "A lot of times I wanted to rob a bank so I could pay for the medicine I was just so scared . . . People with cancer either have a good chance or no chance. In my case it's hard to recover from lung cancer but it's possible. Insurance took so long to kick in that I didn't think I would get it. Now there is a big bright light shining on me." (Anecdotes)
- Important to have broad evidence on multifaceted effects of Medicaid expansions

Baicker, Katherine, et al. (2014). “The Oregon Experiment – Effects of Medicaid on Clinical Outcomes”, *The New England Journal of Medicine*.

In-person data collection

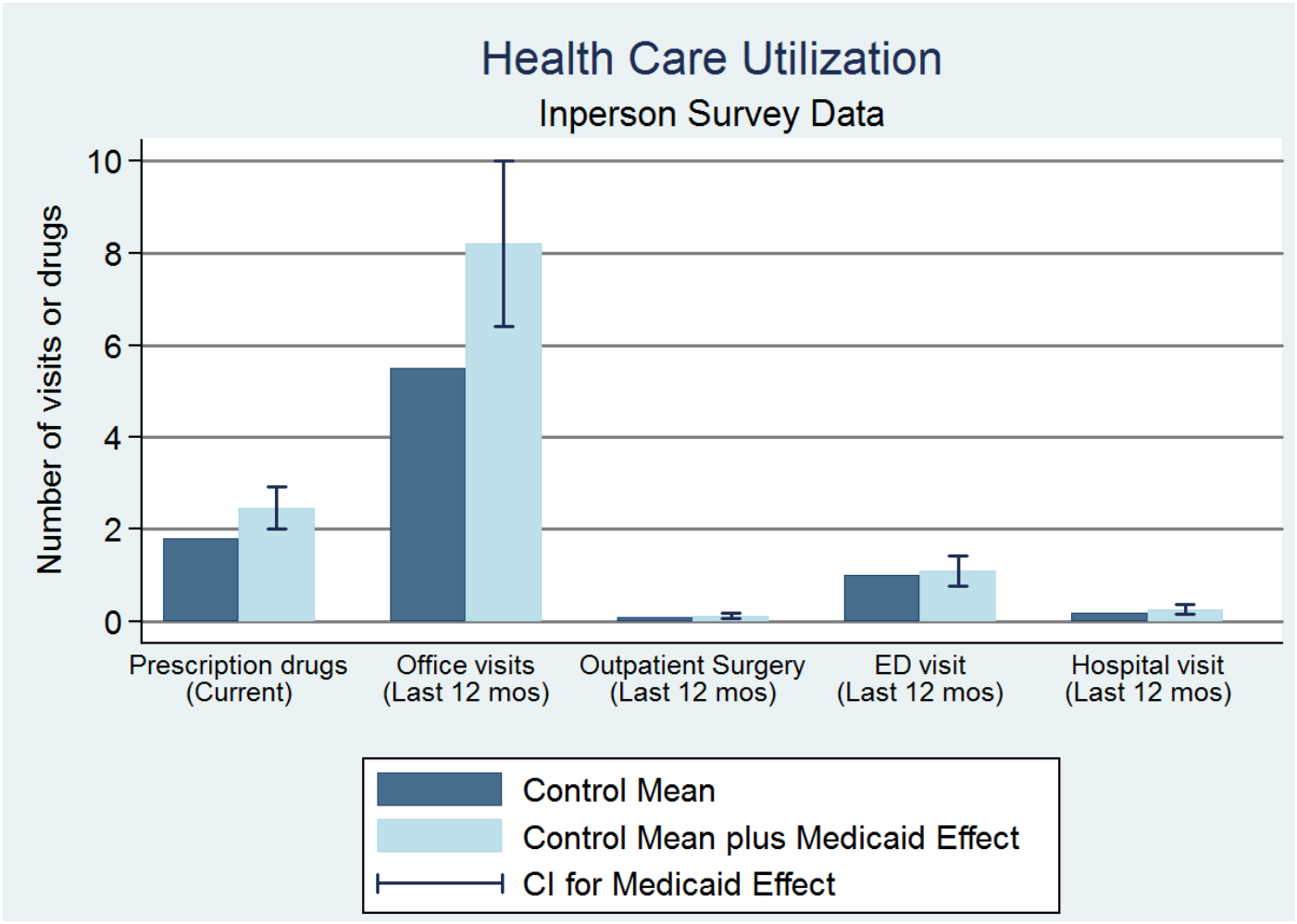
- Questionnaire and health examination including
  - → Survey questions
  - → Anthropometric and blood pressure measurement
  - → Dried blood spot collection
  - → Catalog of all medications
- Fielded between September 2009 and December 2010
  - → Average response ∼25 months after lottery began
- Limited to Portland area: 20,745 person sample
- 12,229 interviews for effective response rate of 73%

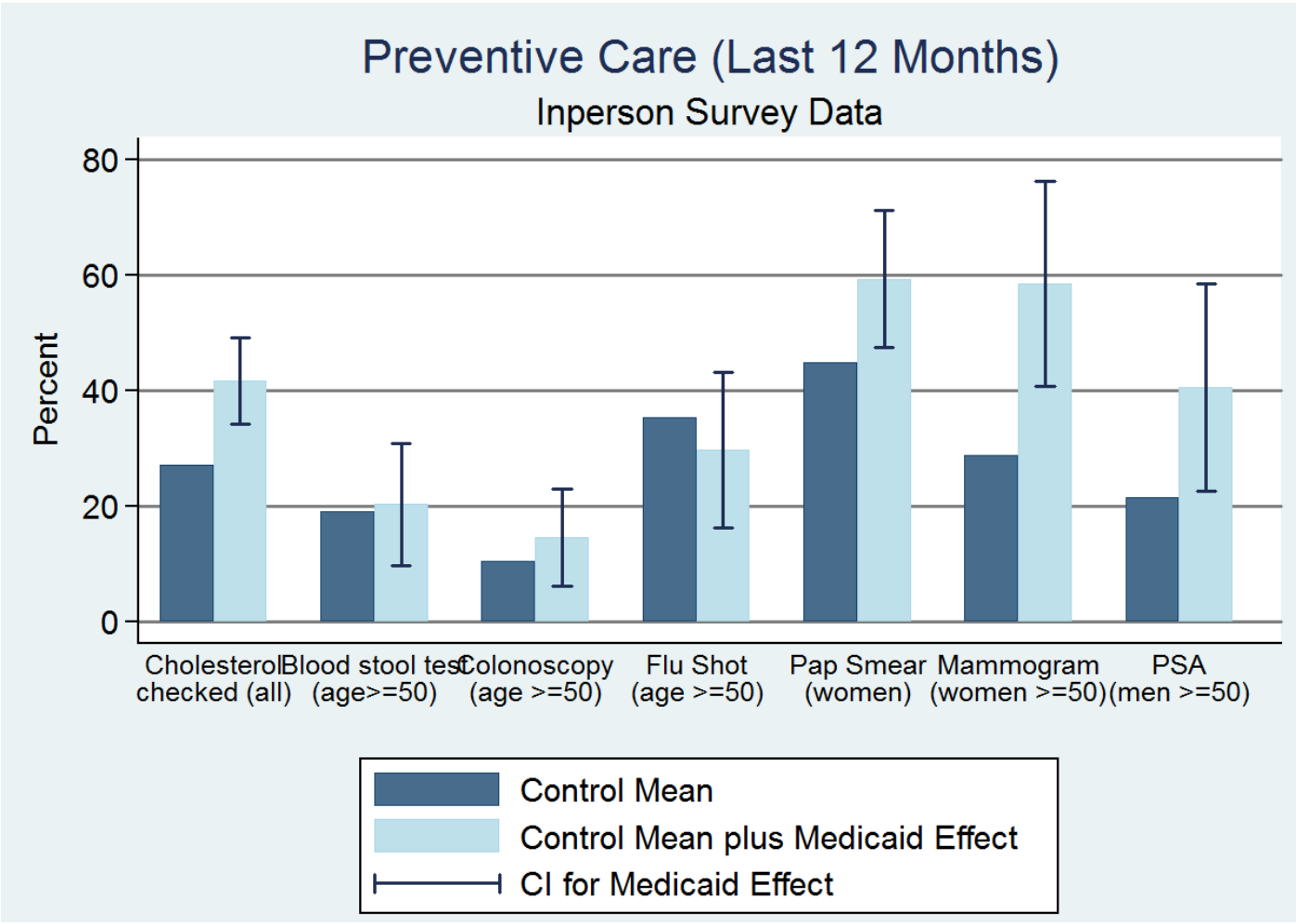
Analytic approach

- Intent to treat effect of *lottery selection*
  - → Comparing all selected with all not selected
  - → Random treatment assignment
  - → No differential selection for outcome measurement
- Local average treatment effect on *Medicaid coverage*
  - → Using lottery selection as an instrument for coverage
  - → ∼24 percentage point increase in Medicaid enrollment
  - → No change in private insurance (no crowd-out)
  - → No effect of lottery except via Medicaid coverage
- Statistical inference is the same for both

Results

1. *Health care use*
2. Financial strain
3. Clinical health outcomes


Health Care Utilization
Inperson Survey Data
Number of visits or drugs
10
8
6
4
2
0
Prescription drugs (Current)
Office visits (Last 12 mos)
Outpatient Surgery (Last 12 mos)
ED visit (Last 12 mos)
Hospital visit (Last 12 mos)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

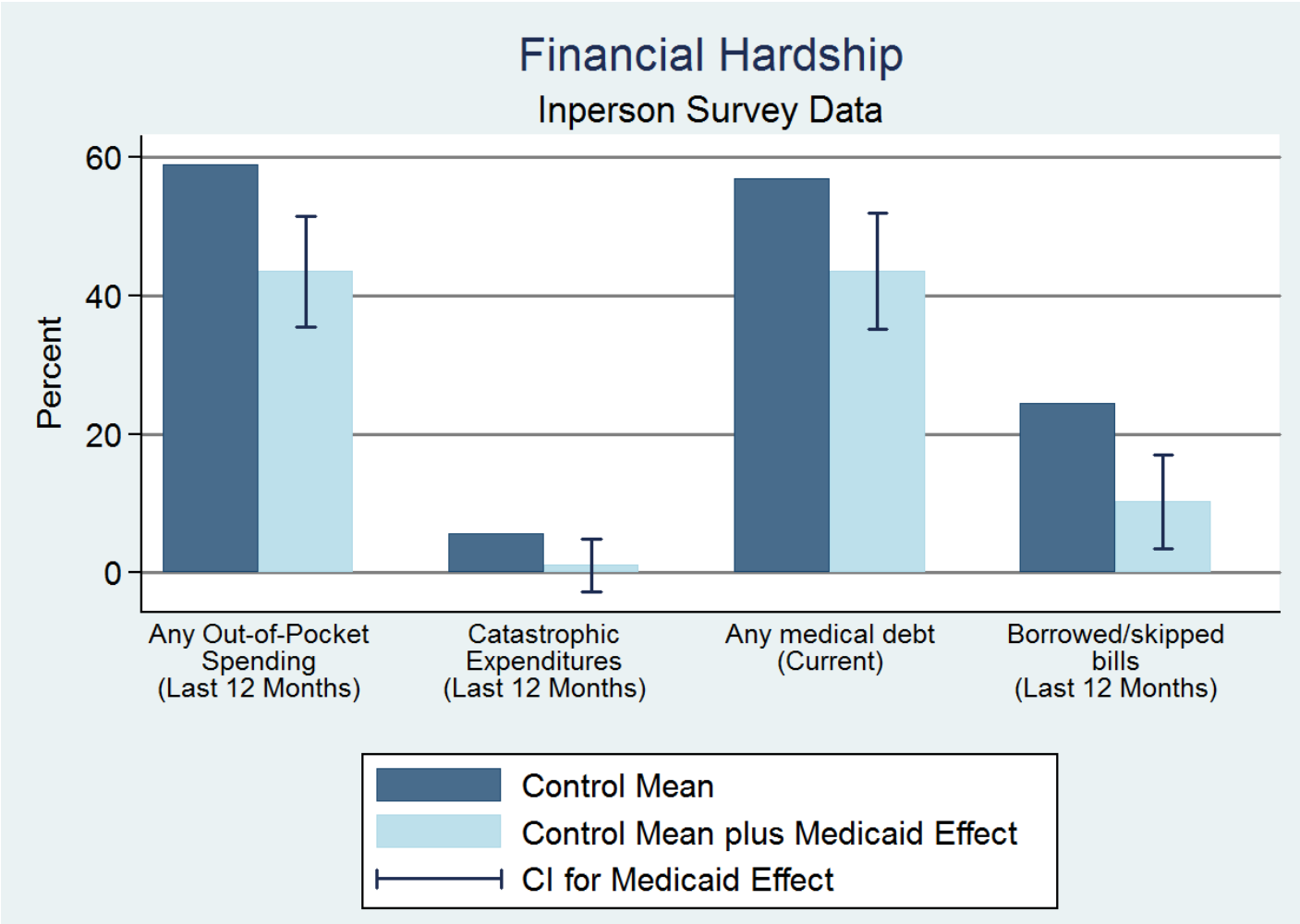

Preventive Care (Last 12 Months)
Inperson Survey Data
Percent
80
60
40
20
0
Cholesterol checked (all)
Blood stool test (age>=50)
Colonoscopy (age >=50)
Flu Shot (age >=50)
Pap Smear (women)
Mammogram (women >=50)
PSA (men >=50)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Health care use results

- Increases in use in various settings
  - → Increases in probability and number of outpatient visits
  - → Increases in probability and number of prescription drugs
  - → No discernible change in hospital or ED use (imprecise)
- Increases in preventive care across range of services
- Increases in perceived access and quality
- Implied 35% increase in spending for insured

Results

1. Health care use
2. *Financial strain*
3. Clinical health outcomes


Financial Hardship
Inperson Survey Data
Percent
60
40
20
0
Any Out-of-Pocket Spending (Last 12 Months)
Catastrophic Expenditures (Last 12 Months)
Any medical debt (Current)
Borrowed/skipped bills (Last 12 Months)
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Financial Hardship Results

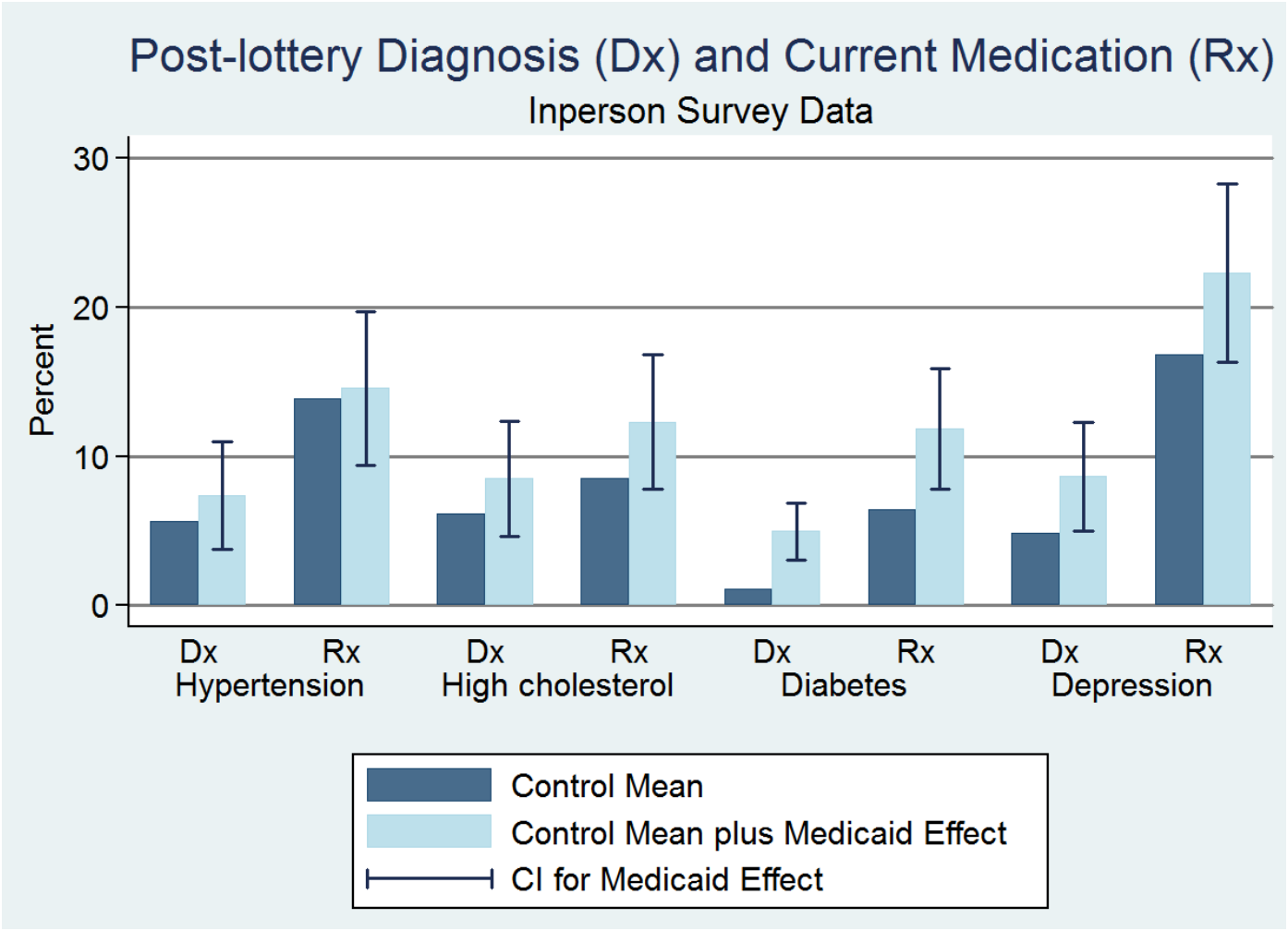
- Reduction in strain, out-of-pocket (OOP), money owed
  - → Substantial reduction across measures
  - → Elimination of catastrophic OOP health spending
- Implications for distribution of burden/benefits
  - → Some borne by patients, some by providers
  - → Non-financial burden of medical expenses and debt

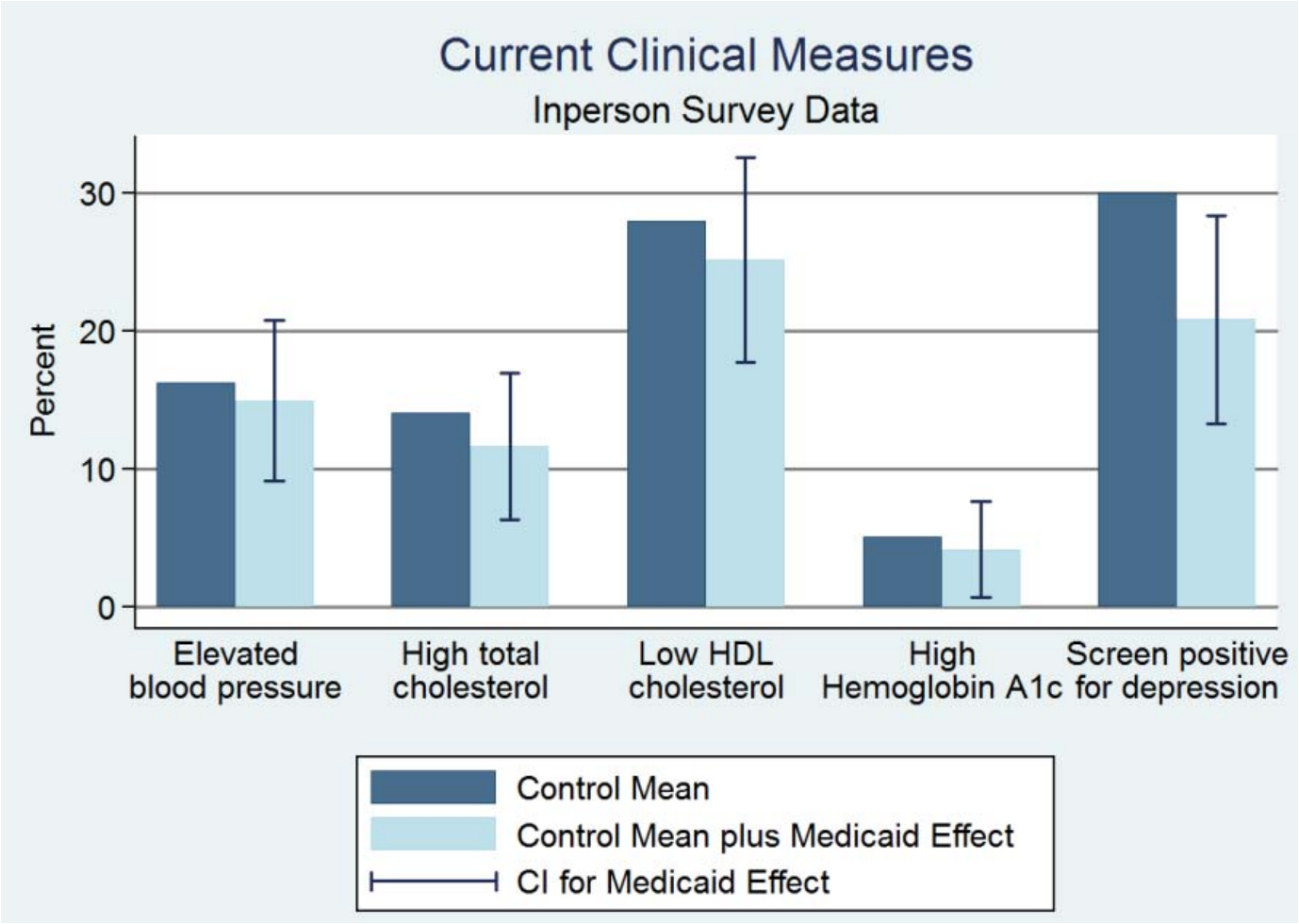
Results

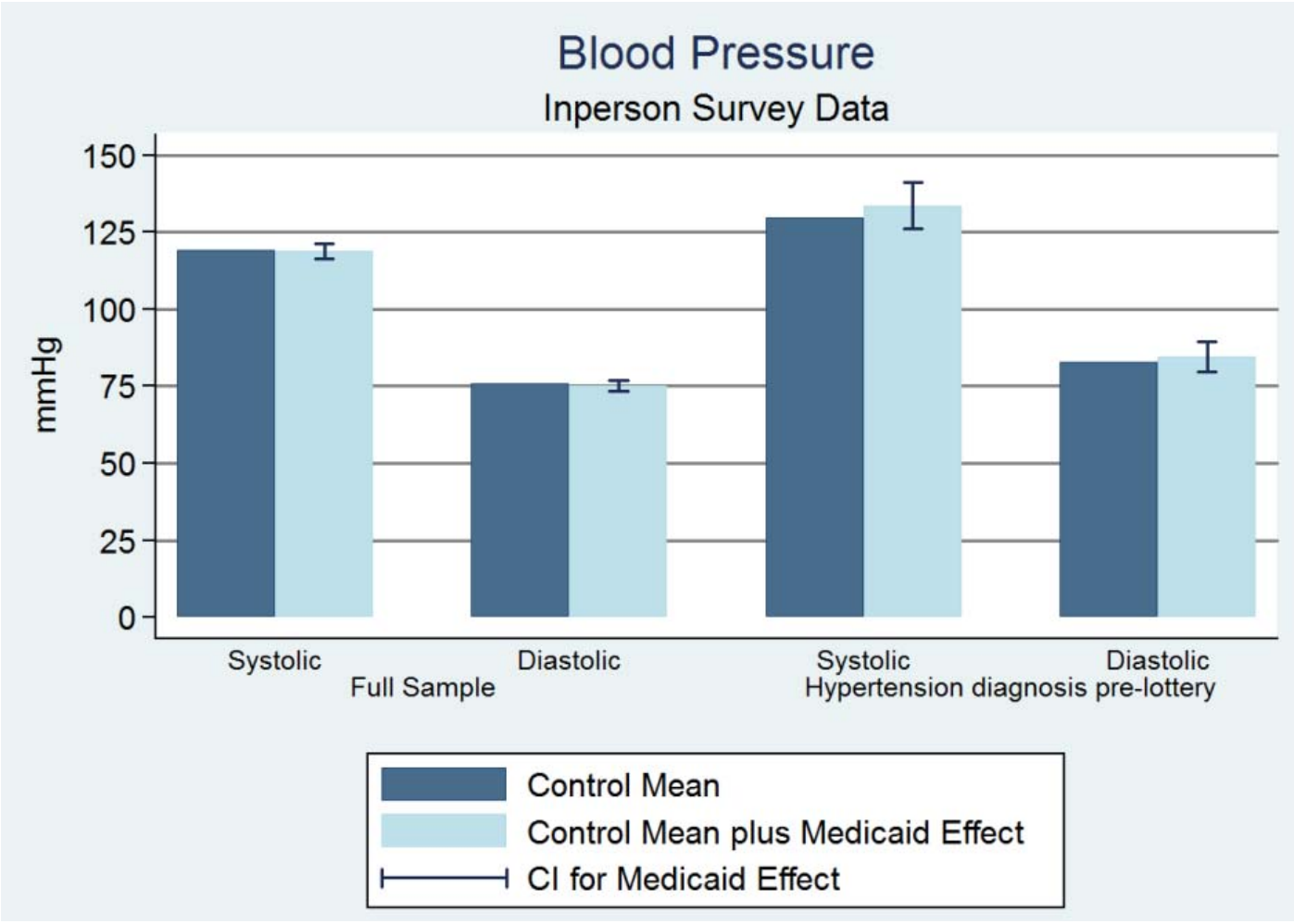
1. Health care use
2. Financial strain
3. *Clinical health outcomes*

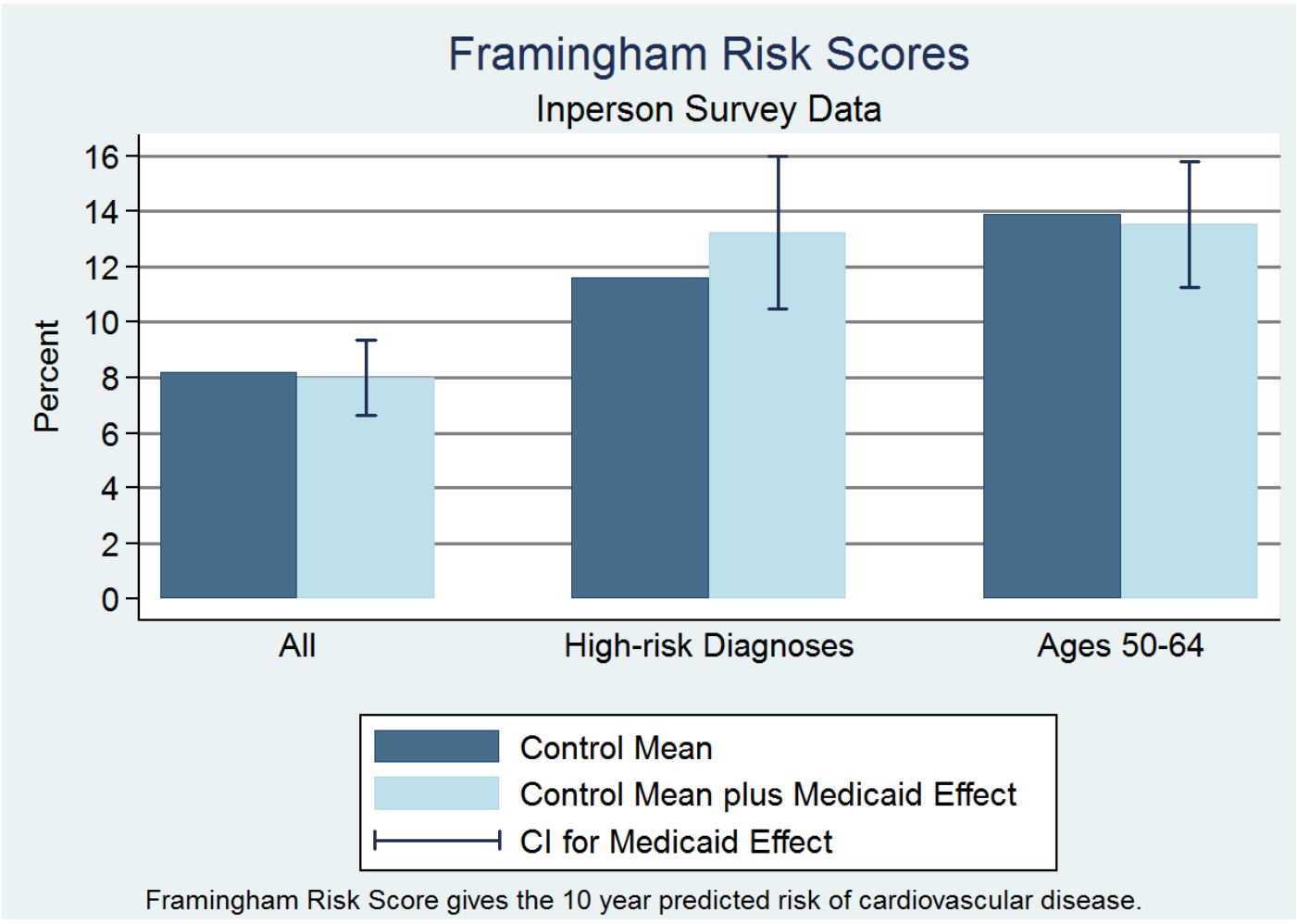
Focusing on specific conditions

- Measured:
  - → Blood pressure
  - → Cholesterol levels
  - → Glycated hemoglobin
  - → Depression
- Reasons for selecting these:
  - → Reasonably prevalent conditions
  - → Clinically effective medications exist
  - → Markers of longer term risk of cardiovascular disease
  - → Can be measured by trained interviewers and lab tests
- A limited window into health status


Post-lottery Diagnosis (Dx) and Current Medication (Rx)
Inperson Survey Data
Percent
30
20
10
0
Dx
Rx
Hypertension
High cholesterol
Diabetes
Depression
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect


Current Clinical Measures
Inperson Survey Data
Percent
30
20
10
0
Elevated blood pressure
High total cholesterol
Low HDL cholesterol
High Hemoglobin A1c
Screen positive for depression
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

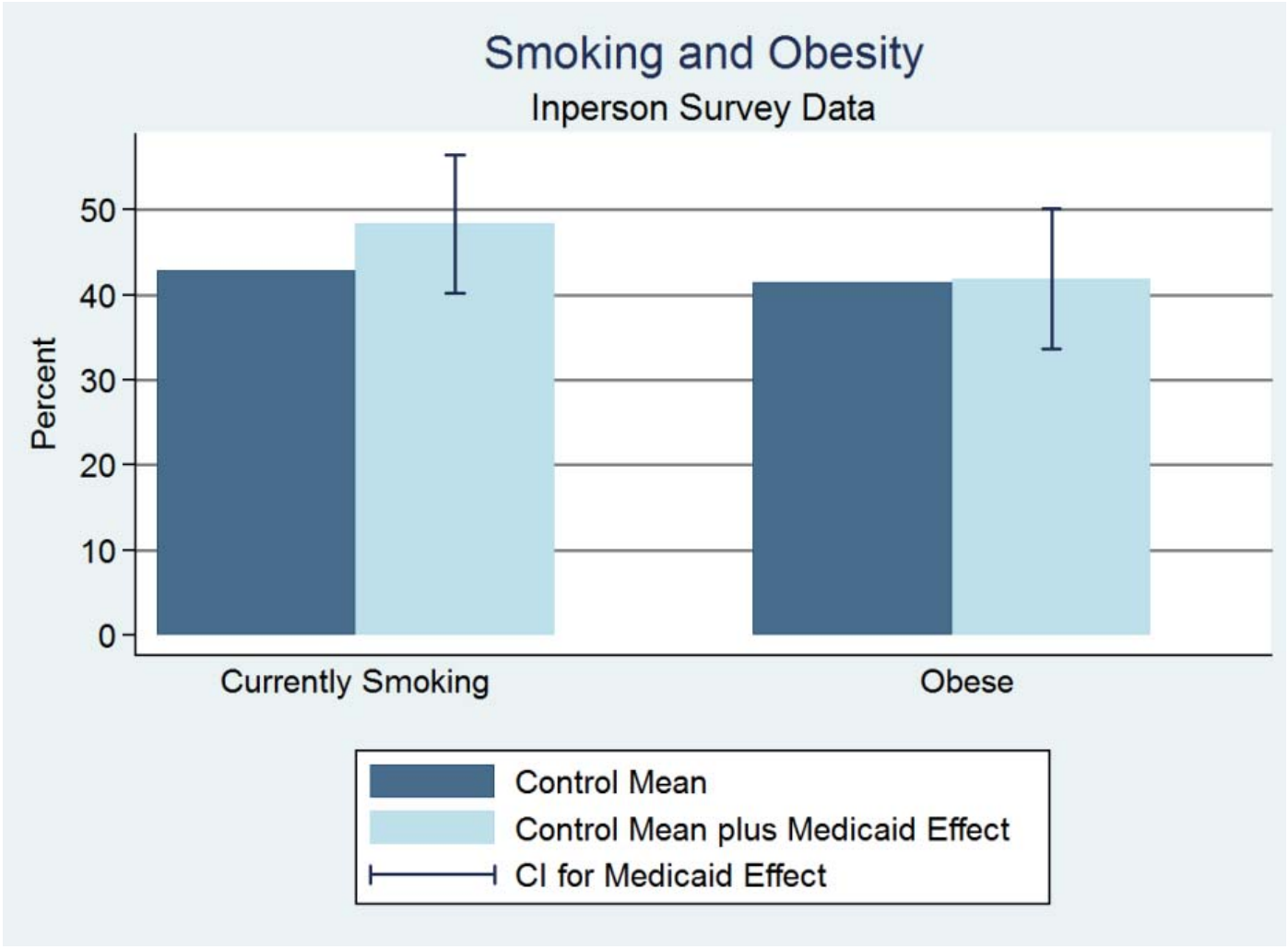

Blood Pressure
Inperson Survey Data
150
125
100
75
50
25
0
mmHg
Systolic
Diastolic
Full Sample
Systolic
Diastolic
Hypertension diagnosis pre-lottery
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect



Framingham Risk Score gives the 10 year predicted risk of cardiovascular disease.

Results on specific conditions

- Large reductions in depression
  - → Increases in diagnoses and medication
  - → In-person estimate of -9 percentage points in being depressed
- Glycated hemoglobin
  - → Increases in diagnosis and medication
  - → No significant effect on HbA1c; wide confidence intervals
- Blood pressure and cholesterol
  - → No significant effects on diagnosis or medication
  - → No significant effects on outcomes
- Framingham risk score
  - → No significant effect (in general or sub-poplulations)


Smoking and Obesity
Inperson Survey Data
Percent
50
40
30
20
10
0
Currently Smoking
Obese
Control Mean
Control Mean plus Medicaid Effect
CI for Medicaid Effect

Summary

- One to two years after expanded access to Medicaid:
  - → Increases in health care use and associated costs
  - → Increases in compliance with recommended preventive care
  - → Improvements in quality and access
  - → Reductions in financial strain
  - → Improvements in self-reported health
  - → Improvements in depression
  - → No significant change in specific physical measures
- Sense of the relative magnitude of the effects
  - → Use and access, financial benefits, general health, depression
  - → Physical measures of specific chronic conditions

Extrapolation to Obamacare (ACA) Expansion

- Context quite relevant for health care reform:
  - → States can choose to cover a similar population in planned 2014 Medicaid expansions (up to 138% of federal poverty line)
- But important caveats to bear in mind
  - → Oregon and Portland vs. US generally
  - → Voluntary enrollment vs. mandate
  - → Partial vs. general equilibrium effects
  - → Short-run (1-2 years) vs. medium or long run
- We will revisit this again later in the difference-in-differences section when discussing Miller, et al. (2019)

Updating Priors based on Study's Findings

- "Medicaid is worthless or worse than no insurance""
  - → Studies found increases in utilization and perceived access and quality
  - → Reductions in financial strain, improvement in self-reported health
  - → Improvement in depression
  - → Can reject large declines in several physical measures
- "Health insurance expansion saves money"
  - → In short run, studies showed increases in utilization and cost and no change in ED use
  - → Increases in preventive care, improvements in self-reported health, improvements in depression

Conclusion

- Effects of expanding Medicaid likely to be manifold
  - → Hard to establish with observational data and often misleading
- Expanding Medicaid generates both costs and benefits
  - → Increased spending
  - → Measurably improves *some* aspects of health but not others
    - Important caveats about generalizability
    - Weighing them depends on policy priorities
- Further research on alternative policies needed
  - → Many steps in pathway between insurance and outcome
  - → Role for innovation in insurance coverage
  - → Complements to health care (e.g., social determinants)

Fuzzy RDD, IV and ITT

- Fuzzy RDD is an IV estimator, and requires those assumptions
- You may be more comfortable with presenting the intent-to-treat (ITT) parameter which is just the reduced form regression of Y on Z, therefore
- Many papers will not present an IV-style parameter, but rather a blizzard of ITT parameters, out of a “fear” that the exclusion restrictions may not hold

Probability of treatment jumps at discontinuity

Probabilistic treatment assignment (i.e. “fuzzy RDD”)

The probability of receiving treatment changes discontinuously at the cutoff, c_0, but need not go from 0 to 1

$$lim_{X_i \to c_0} Pr(D_i = 1 | X_i = c_0) \neq lim_{c_0 \leftarrow X_i} Pr(D_i = 1 | X_i = c_0)$$

Examples: Incentives to participate in some program may change discontinuously at the cutoff but are not powerful enough to move everyone from non participation to participation.

Deterministic (sharp) vs. probabilistic (fuzzy)

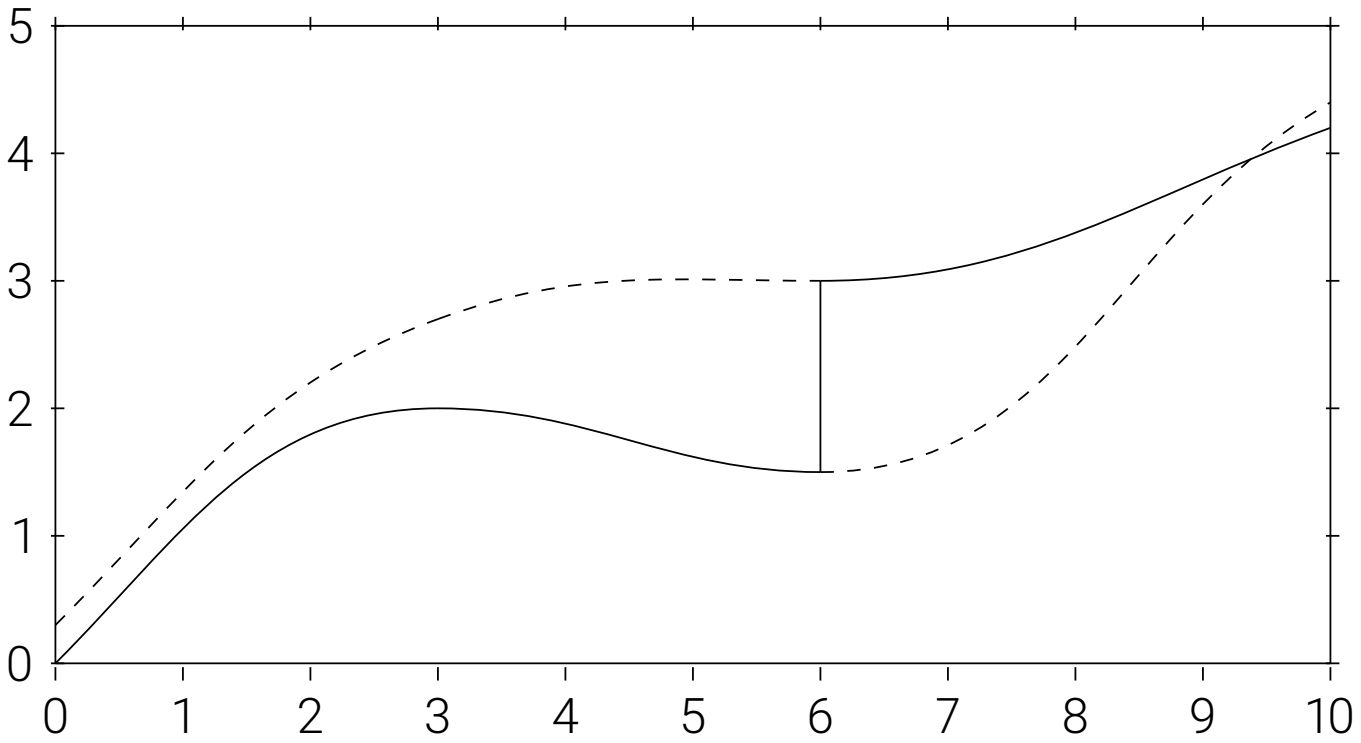
- In the sharp RDD, D_i was *determined* by $X_i \geq c_0$
- In the fuzzy RDD, the *conditional probability* of treatment *jumps* at c_0.
- The relationship between the conditional probability of treatment and X_i can be written as:

$$P[D_i = 1|X_i] = g_0(X_i) + [g_1(X_i) - g_0(X_i)]Z_i$$

where $Z_i = 1$ if $(X_i \geq c_0)$ and 0 otherwise.

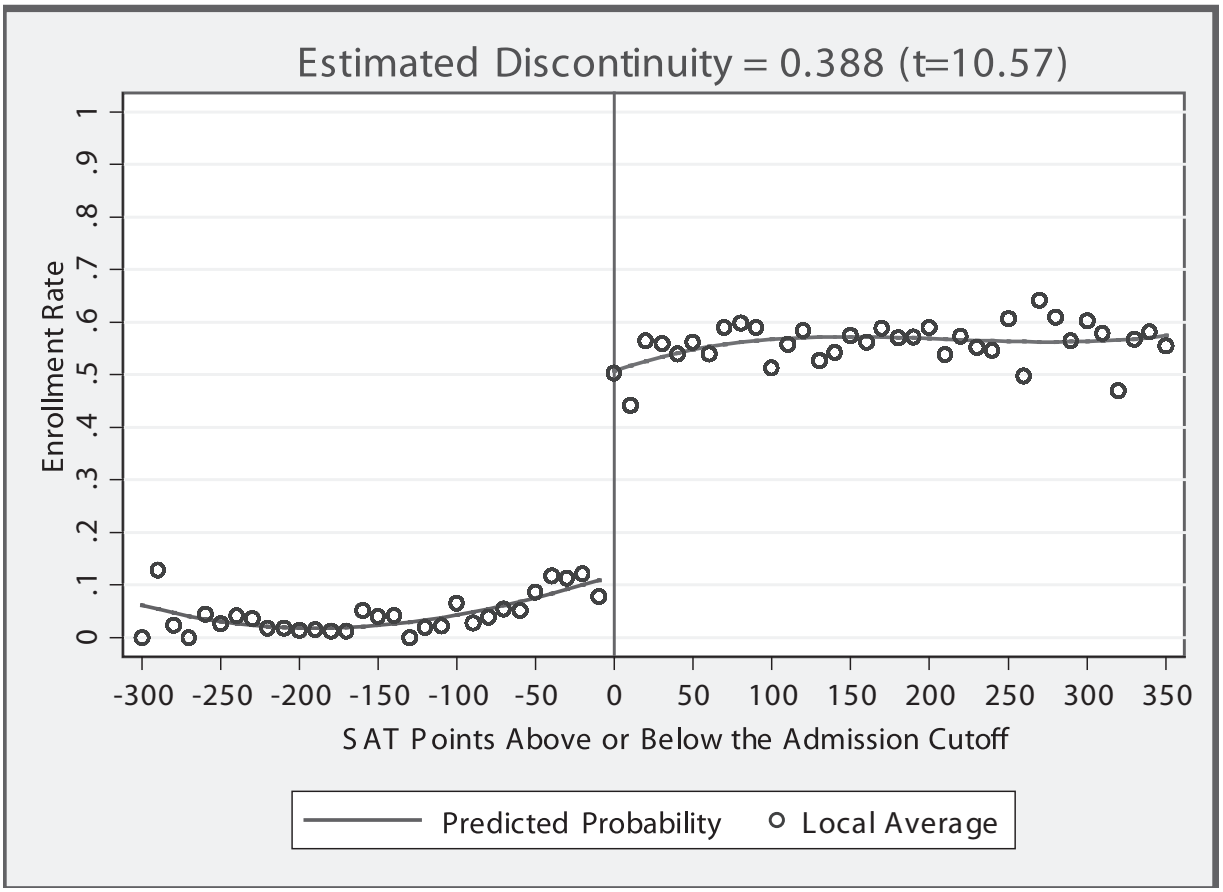
Visualization of identification strategy (i.e. smoothness)

- $E[Y^0|X]$ and $E[Y^1|X]$ for $D = 0, 1$ are the dashed/solid continuous functions
- $E[Y|X]$ is the solid which jumps at $X = 6$



Hoekstra flagship school

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Instrumental variables

- As said, fuzzy designs are numerically equivalent and conceptually similar to IV
  - → "Reduced form" Numerator: "jump" in the regression of the outcome on the running variable, X.
  - → "First stage" Denominator: "jump" in the regression of the treatment indicator on the running variable X.
- Same IV assumptions, caveats about compliers vs. defiers, and statistical tests that we will discuss in next lecture with instrumental variables apply here – e.g., check for weak instruments using F test on instrument in first stage, etc.

Wald estimator

Wald estimator of treatment effect under Fuzzy RDD

Average causal effect of the treatment is the Wald IV parameter

$$\delta_{\text{Fuzzy RDD}} = \frac{lim_{X \to c_0} E[Y|X = c_0] - lim_{c_0 \leftarrow X} E[Y|X = c_0]}{lim_{X \to c_0} E[D|X = c_0] - lim_{c_0 \leftarrow X} E[D|X = c_0]}$$

RDD's Relationship to IV

- Center X it's equal to zero at c_0 and define $Z = \mathbf{1}(X \geq 0)$
- The coefficient on Z in a regression like

```
. reg Y Z X X2 X3
```

  is the reduced form discontinuity, and

```
. reg D Z X X2 X3
```

  is the first stage discontinuity
- Ratio of discontinuities is estimate of $\delta_{\text{Fuzzy RDD}}$
- Simple way to implement is IV

```
. ivregress 2sls Y (D=Z) X X2 X3
```

First stage relationship between X and D

- One can use both Z_i as well as the interaction terms as instruments for D_i.
- If one uses only Z_i as IV, then the it is a "just identified" model which usually has good finite sample properties.
- In the just identified case, the first stage would be:

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \cdots + \gamma_p X_i^p + \pi Z_i + \varepsilon_{1i}$$

  where π is the causal effect of Z on the conditional probability of treatment.
- The fuzzy RD reduced form is:

$$Y_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \cdots + \kappa_p X_i^p + \rho\pi Z_i + \varepsilon_{2i}$$

Fuzzy RDD with varying Treatment Effects - Second Stage

- As in the sharp RDD case one can allow the smooth function to be different on both sides of the discontinuity.
- The second stage model with interaction terms would be the same as before:

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p \\ &\quad + \rho D_i + \beta_1^* D_i\tilde{x}_i + \beta_2^* D_i\tilde{x}_i^2 + \cdots + \beta_p^* D_i\tilde{x}_i^p + \eta_i \end{aligned}$$

- Where $\tilde{x}$ are now not only normalized with respect to c_0 but are also fitted values obtained from the first stage regression.

Fuzzy RDD with Varying Treatment Effects - First Stages

- Again one can use both Z_i as well as the interaction terms as instruments for D_i
- Only using Z the estimated first stages would be:

$$\begin{aligned} D_i &= \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p \\ &\quad + \pi Z_i + \gamma_1^*\tilde{X}_i Z_i + \gamma_2^*\tilde{X}_i^2 Z_i + \cdots + \gamma_p^* Z_i + \varepsilon_{1i} \end{aligned}$$

- We would also construct analogous first stages for $\tilde{X}_i D_i$, $\tilde{X}_i^2 D_i$, $\ldots$, $\tilde{X}_i^p D_i$.

Limitations of the LATE

- Fuzzy RDD has assumptions of all standard IV framework (exclusion, independence, nonzero first stage, and monotonicity)
- As with other binary IVs, the fuzzy RDD is estimating LATE: the local average treatment effect for the group of *compliers*
- In RDD, the compliers are those whose treatment status changed as we moved the value of x_i from just to the left of c_0 to just to the right of c_0
- Means we can use Medicare age cutoff to estimate the effect of public insurance on mortality (LATE) and still not know the effect of public insurance on mortality (ATE)