Chapter 7: Instrumental Variables (IV)

Jonathan Roth

Mathematical Econometrics I
Brown University

Motivation

- We've seen how to estimate average treatment effects when treatment is as good as randomly assigned conditional on observable characteristics
- We've also seen how we can relax the assumption of conditional unconfoundedness if we have panel data and assume the selection bias is constant over time (parallel trends)
- But sometimes these assumptions won't be plausible
  - We might be worried about unobserved confounding variables
  - We might not have panel data, or might be worried that the confounders are time-varying (so parallel trends fails)
- What can we do then?

"Local Experiments"

- The gold standard for causal inference is to run an experiment
- But often an experiment is not possible
- Luckily, sometimes we have a (natural) experiment that affects whether some people take up the treatment we're interested in
- If that's the case, then (under certain conditions) we can use this experiment to learn about the effect of the treatment for the "compliers" who are induced to take-up treatment by the experiment

Example - the effects of Medicaid

- Important question for policy: what is the effect of Medicaid on health outcomes?
- The ideal situation for learning about the causal effects of Medicaid would be to randomize who gets it
  - Not possible for moral / budgetary reasons
- However, we have the Oregon Health Insurance Experiment (OHIE) which randomized *eligibility* for Medicaid for some people (i.e. those with earnings between 100 and 138% of the FPL)

Reminder - background about OHIE

In 2008, a group of uninsured low-income adults in Oregon was selected by lottery to be given the chance to apply for Medicaid. This lottery provides an opportunity to gauge the effects of expanding access to public health insurance on the health care use, financial strain, and health of low-income adults using a randomized controlled design. In the year after random assignment, the treatment group selected by the lottery was about 25 percentage points more likely to have insurance than the control group that was not selected. We find that in this first year, the treatment group had substantively and statistically significantly higher health care utilization (including primary and preventive care as well as hospitalizations), lower out-of-pocket medical expenditures and medical debt (including fewer bills sent to collection), and better self-reported physical and mental health than the control group. *JEL* Codes: H51, H75, I1.

Winning the lottery vs Medicaid

- Previously, we used the OHIE to estimate the causal effect of *winning the eligibility lottery*
- But what if we actually care about the causal effect of *enrolling in Medicaid*?
- Winning the lottery is not the same as enrolling in Medicaid:

	Winners	Losers
Ever on Medicaid	0.397	0.141

- Some people who win the lottery do not enroll (don't return form; no longer eligible)
- Some people who lose the lottery enroll anyway (become eligible by other criteria)

The effect of Medicaid enrollement on depression

	Winners	Losers
Ever on Medicaid	0.397	0.141
Depression symptoms	0.306	0.329

- What is the estimated effect of winning the lottery on Medicaid enrollment? $0.397 - 0.141 = 0.256$
- What is the estimated effect of winning the lottery on depression? $0.306 - 0.329 = -0.023$
- Say you want to estimate the effect of *Medicaid enrollment* on depression? What would you do?
- A natural estimate is to divide the effect on depression by the effect on enrollment: $-0.023/0.256 \approx -0.09 \rightarrow$ a 9 pp reduction per enrollee
- This is called an **instrumental variables (IV)** estimate. When does it work? And how exactly do we interpret it?

Local Average Treatment Effects (LATE)

- We will show that under certain assumptions, instrumental variables lets us identify a **local average treatment effect** (LATE).
- This is the average treatment effect for **compliers**, i.e. people who are induced to take-up Medicaid by winning the lottery
- This is a "local" effect in the sense that it doesn't tell us about the treatment effect for people whose Medicaid status is not affected by the experiment
- Next, we'll go over the assumptions we need to identify the LATE with instrumental variables

Four types of people

We can divide the population into four types of people:

- **Always takers:** people who will enroll in Medicaid regardless of the outcome of the lottery
- **Never takers:** people who will never enroll in Medicaid regardless of the lottery
- **Compliers:** people who enroll in Medicaid only if they win the lottery
- **Defiers:** people who enroll in Medicaid only if they lose the lottery
  - Defiers are weird, and often we will assume they don't exist (called *monotonicity*)

Four types of people – With Math

- Let D_i be an indicator for whether you are on Medicaid (treatment). Let Z_i be an indicator for whether you won the lottery (*instrument*)

- Introduce potential treatments, $D_i(1)$ and $D_i(0)$, analogous to the potential outcomes
  - $D_i(1)$ is treatment status if $Z_i = 1$ (win the lottery)
  - $D_i(0)$ is treatment status if $Z_i = 0$ (lost the lottery)

- **Always takers:** people who will enroll in Medicaid regardless of the outcome of the lottery. Always takers have $D_i(1) = D_i(0) = 1$

- **Never takers:** people who will never enroll in Medicaid regardless of the lottery. Never takers have $D_i(1) = D_i(0) = 0$

- **Compliers:** people who enroll in Medicaid only if they win the lottery. Compliers have $D_i(1) = 1$ and $D_i(0) = 0$

- **Defiers:** people who enroll in Medicaid only if they lose the lottery. Defiers have $D_i(1) = 0$ and $D_i(0) = 1$

Key assumptions

- **Independence:** The instrument (e.g. whether you win the lottery) is as good as randomly assigned, $Z_i \perp\!\!\!\perp (Y_i(0), Y_i(1), D_i(0), D_i(1))$
- **Relevance** Instrument affects the probability of treatment: $P(D_i = 1|Z_i = 1) \neq P(D_i = 1|Z_i = 0)$
  - This one is testable – we saw that it does!
- **Monotonicity:** No defiers: $D_i(1) \geq D_i(0)$ for all i
  - Seems reasonable that anyone who enrolls without winning would also enroll if they won

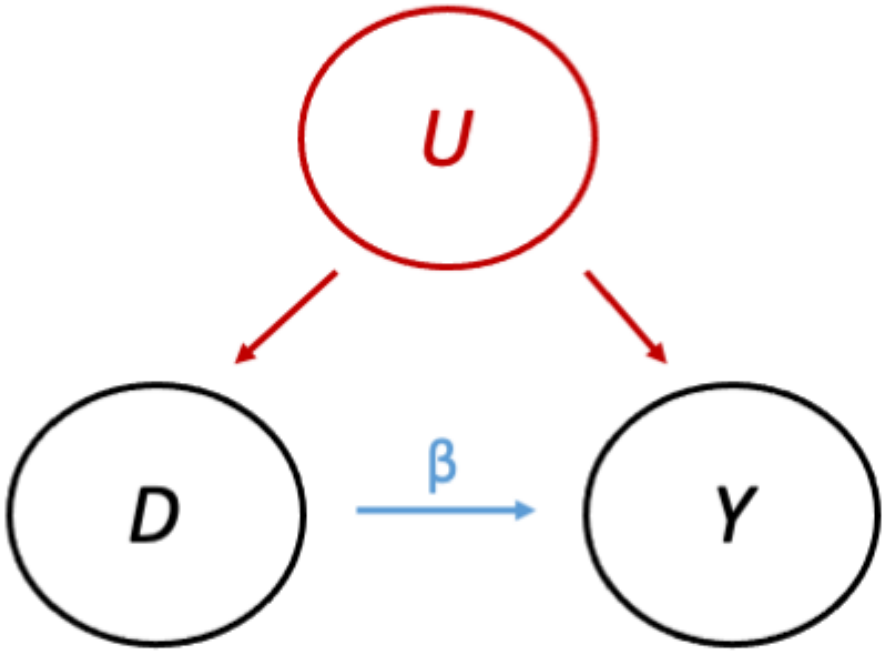
Final key assumption — exclusion restriction!

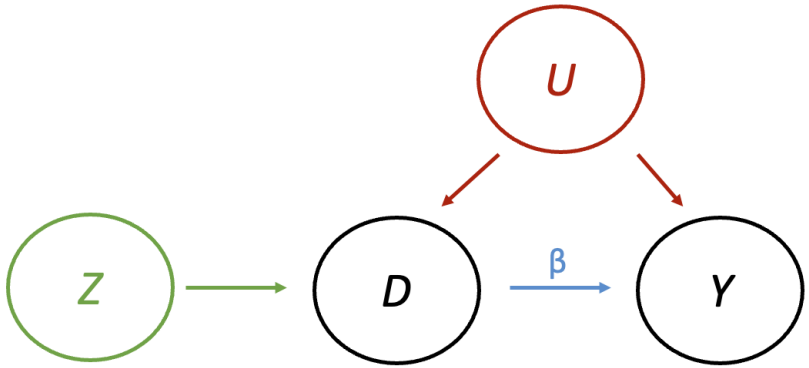
- The final key assumption, and often the most tenuous, is what's called the **exclusion restriction**
- Intuitively, this says that the instrument Z_i (whether you won the lottery) affects your outcomes *only through the treatment* (whether you enrolled in Medicaid)
- Mathematically, we have that $Y_i = D_i(Z_i)Y_i(1) + (1 - D_i(Z_i))Y_i(0)$
  - Equivalently, can write the POs as $Y_i(d, z)$, then state the exclusion restriction as $Y_i(d, z) = Y_i(d)$.
- This implies that the outcomes for the always-takers and never-takers doesn't depend on Z_i

Why might exclusion fail?

- Intuitively, exclusion fails if the instrument (e.g. winning the lottery) can affect your outcome even without changing your treatment sattus
- Suppose in the OHIE that some always-takers would have had to wait longer to get on Medicaid if they hadn't won the lottery. Would that violate exclusion?
- Yes, if wait-time for Medicaid affects your health.

Graphical Illustration





- Unconfoundedness fails if there is an unobservable U that affects both treatment D and Y (denoted by arrows)

The LATE Theorem

- If Independence, Relevance, Exclusion, and Monotonicity hold, then

$$\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]} = E[Y_i(1)-Y_i(0)|D_i(1)=1, D_i(0)=0]$$

  or in words:

$$\frac{\text{Effect of } Z \text{ on } Y}{\text{Effect of } Z \text{ on } D} = \text{Local average treatment effect for compliers}$$

- This result won Joshua Angrist and Guido Imbens the 2021 Nobel Prize in Economics!

Intuition for the LATE theorem

- Remember the theorem says that

$$\frac{E[Y_i|Z_i=1]-E[Y_i|Z_i=0]}{E[D_i|Z_i=1]-E[D_i|Z_i=0]} = E[Y_i(1)-Y_i(0)|D_i(1)=1, D_i(0)=0]$$

- By independence, the numerator is the causal effect of Z on Y
- However, the exclusion restriction says that the effect of Z on Y is zero for always-takers and never-takers (their treatment doesn't change!). Further, we've assumed that there are no defiers
- Thus, the numerator will be the average effect for compliers times the fraction of compliers in the population, i.e. $LATE \times P(Complier)$
- But the denominator is the effect of Z on D. This effect is zero for ATs and NTs, so the denominator is the share of compliers, $P(Complier)$

More formal derivation

- Remember that we can divide the population into three groups, always-takers (ATs), never-takers (NTs), and compliers (Cs) — we assumed no defiers!

- Let $\alpha_{AT} = P(AT), \alpha_{NT} = P(NT), \alpha_C = P(C)$ be the shares of ATs, NTs, and Cs

- Since Z is randomly assigned (independence), we have that $P(AT|Z = 1) = P(AT) = \alpha_{AT}$.
  Likewise, $P(AT|Z = 0) = P(AT) = \alpha_{AT}$

- By similar arguments, share of NTs and Cs is the same in the $Z = 1$ and $Z = 0$ groups.

More formal derivation

- Next step: show that the numerator is $\alpha_C \times LATE$
- By the Law of Iterated Expectations,

  $E[Y_i|Z_i = 1] =$
  $\alpha_C E[Y_i|Z_i = 1, C] + \alpha_{AT} E[Y_i|Z_i = 1, AT] + \alpha_{NT} E[Y_i|Z_i = 1, NT] =$
  $\alpha_C E[Y_i(1)|Z_i = 1, C] + \alpha_{AT} E[Y_i(1)|Z_i = 1, AT] + \alpha_{NT} E[Y_i(0)|Z_i = 1, NT] =$
  $\alpha_C E[Y_i(1)|C] + \alpha_{AT} E[Y_i(1)|AT] + \alpha_{NT} E[Y_i(0)|NT]$

  where the last line uses independence
- Similarly,

$$E[Y_i|Z_i = 0] = \alpha_C E[Y_i(0)|C] + \alpha_{AT} E[Y_i(1)|AT] + \alpha_{NT} E[Y_i(0)|NT]$$

- Thus, the numerator in our ratio is

$$E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \alpha_C \underbrace{(E[Y_i(1)|C] - E[Y_i(0)|C])}_{LATE}$$

More formal derivation

- We showed that the numerator is $\alpha_C \times LATE$
- To finish proof, we show that the *denominator* is α_C
- The denominator is

$$\begin{aligned} E[D_i|Z_i=1] - E[D_i|Z_i=0] &= Pr(\text{AT or C}|Z_i=1) - P(\text{AT}|Z_i=0) \\ &= (\alpha_C + \alpha_{AT}) - \alpha_{AT} = \alpha_C \end{aligned}$$

Hence,

$$\frac{numerator}{denominator} = \frac{\alpha_C \times LATE}{\alpha_C} = LATE$$

Estimating LATE

- We've shown that under the IV assumptions, the LATE is identified as a function of population means

$$\frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]} = \text{LATE}$$

- How can we estimate LATE?

- Plug in sample means!

$$\hat{\beta} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{D}_{Z=1} - \bar{D}_{Z=0}},$$

  where, e.g., $\bar{Y}_{Z=1}$ is the sample mean of Y_i for units with $Z_i = 1$

- $\hat{\beta}$ is called the IV estimator, or more precisely, the two-stage least squares (2SLS) estimator (for reasons that will become clear soon!)

Example - Medicaid

	Winners	Losers
Ever on Medicaid	0.397	0.141
Depression symptoms	0.306	0.329

- Let's calculate the 2SLS estimator in the Medicaid example.

$$\frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{D}_{Z=1} - \bar{D}_{Z=0}} = \frac{0.306 - 0.329}{0.397 - 0.141} = -0.09$$

Two-stage least squares

- Consider the two linear regressions:

$$Y_i = \gamma_0 + Z_i\gamma_1 + \varepsilon_i \quad (1)$$
$$D_i = \pi_0 + Z_i\pi_1 + u_i \quad (2)$$

- What are the OLS estimates $\hat{\gamma}_1$ and $\hat{\pi}_1$?

$$\hat{\gamma}_1 = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$$
$$\hat{\pi}_1 = \bar{D}_{Z=1} - \bar{D}_{Z=0}$$

- Thus, the 2SLS estimator is the ratio of these two OLS coefficients:

$$\hat{\beta} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\bar{D}_{Z=1} - \bar{D}_{Z=0}} = \frac{\hat{\gamma}_1}{\hat{\pi}_1}$$

- Equation (2) is typically called the *first-stage*, while equation (1) is called the *reduced form*

Example - Medicaid

	Winners	Losers
Ever on Medicaid	0.397	0.141
Depression symptoms	0.306	0.329

- What is the "first-stage" coefficient from

$$D_i = \pi_0 + Z_i\pi_1 + u_i?$$

- $\hat{\pi}_1 = 0.397 - 0.141 = 0.256$
- What is the "reduced-form" coefficient from

$$Y_i = \gamma_0 + Z_i\gamma_1 + \varepsilon_i?$$

- $\hat{\gamma}_1 = 0.306 - 0.329 = -0.023$
- So the 2SLS estimator is $\hat{\gamma}_1/\hat{\pi}_1 = -0.023/0.256 = -0.09$.

IV conditional on covariates

- Oftentimes, the assumption that the instrument is as-good-as-randomly assigned will only be plausible conditional on observable characteristics
- In fact, in the OHIE, the probability of winning the lottery depended on family size!
- Suppose that we replace the independence assumption with conditional independence, $Z_i \perp\!\!\!\perp (Y_i(1), Y_i(0), D_i(1), D_i(0)) | X_i$
- By similar arguments as before, we can identify the *conditional LATE* by taking the same ratio within groups with the same value of x

$$\frac{E[Y_i|Z_i = 1, X_i = x] - E[Y_i|Z_i = 0, X_i = x]}{E[D_i|Z_i = 1, X_i = x] - E[D_i|Z_i = 0, X_i = x]} = \underbrace{E[Y_i(1) - Y_i(0)|D_i(1) = 1, D_i(0) = 0, X_i = x]}_{\text{LATE(x)}}$$

Estimation for LATEs conditional on covariates

- In practice, when we have conditional independence for the instrument, people estimate a modified version of 2SLS that includes covariates.

- That is, they use $\hat{\beta}_{2SLS} = \hat{\gamma}_1/\hat{\pi}_1$, for the OLS estimates from

$$Y_i = \gamma_0 + Z_i\gamma_1 + \boldsymbol{X}_i'\boldsymbol{\gamma}_2 + \varepsilon_i \tag{3}$$
$$D_i = \pi_0 + Z_i\pi_1 + \boldsymbol{X}_i'\pi_2 + u_i \tag{4}$$

- When $\boldsymbol{X}_i$ is a set of dummy variables for different categories (e.g. family size), this gives a weighted average of the 2SLS estimates for each covariate value

- More generally, this will be approximately consistent for a weighted average of $LATE(x)$ when (4) is a good approximation to the CEF

Example - Medicaid

- In the OHIE, the lottery was actually only random condition on family size
- Finkelstein et al (2012) therefore estimate 2SLS with

$$Y_i = \gamma_0 + Z_i\gamma_1 + \boldsymbol{X}_i'\boldsymbol{\gamma}_2 + \varepsilon_i \tag{5}$$
$$D_i = \pi_0 + Z_i\pi_1 + \boldsymbol{X}_i'\pi_2 + u_i \tag{6}$$

where $\boldsymbol{X}_i$ includes fixed effects for family size and some other demographic variables.

TABLE IV

HOSPITAL UTILIZATION

	Control mean (1)	ITT (2)	LATE (3)	p-values (4)
Panel A: Extensive margin				
All hospital admissions	0.067	0.0054	0.021	[0.004]
	(0.250)	(0.0019)	(0.0074)	
Admissions through ER	0.048	0.0018	0.0070	[0.265]
	(0.214)	(0.0016)	(0.0062)	
Admissions not through ER	0.029	0.0041	0.016	[0.002]
	(0.167)	(0.0013)	(0.0051)	

TABLE IX

HEALTH

	Control mean (1)	ITT (2)	LATE (3)	*p*-values (4)
Panel A: Administrative data				
Alive	0.992	0.00032	0.0013	[0.638]
	(0.092)	(0.00068)	(0.0027)	
Panel B: Survey data				
Self-reported health good/very good/excellent (not fair or poor)	0.548	0.039	0.133	[<0.0001]
	(0.498)	(0.0076)	(0.026)	{<0.0001}
Self-reported health not poor (fair, good, very good, or excellent)	0.86	0.029	0.099	[<0.0001]
	(0.347)	(0.0051)	(0.018)	{<0.0001}
Health about the same or gotten better over last six months	0.714	0.033	0.113	[<0.0001]
	(0.452)	(0.0067)	(0.023)	{<0.0001}
# of days physical health good, past 30 days*	21.862	0.381	1.317	[0.019]
	(10.384)	(0.162)	(0.563)	{0.018}
# days poor physical or mental health did not impair usual activity, past 30 days*	20.329	0.459	1.585	[0.009]
	(10.939)	(0.175)	(0.606)	{0.015}
# of days mental health good, past 30 days*	18.738	0.603	2.082	[0.001]
	(11.445)	(0.184)	(0.64)	{0.003}
Did not screen positive for depression, last two weeks	0.671	0.023	0.078	[0.001]
	(0.470)	(0.0071)	(0.025)	{0.003}
Standardized treatment effect		0.059	0.203	[<0.0001]
		(0.011)	(0.039)	

Notes. Standard errors in parentheses; per comparison *p*-values in square brackets; family-wise *p*-values in curly brackets. Column (2) reports the coefficient and standard error on *LOTTERY* from estimating equation (1) by OLS. Column (3) reports the coefficient and standard error on *INSURANCE* from estimating equation (3) by IV; for the IV estimates in column (3), the endogenous variable *INSURANCE* is defined as "ever on Medicaid" during our study period and the first stage is given in the first row of Table III. Column (4) reports the per comparison *p*-value and the family-wise *p*-value across the different measures used to create the standardized treatment effect. Standardized treatment effect reports results based on equation (2). All regressions include household size fixed effects and standard errors are clustered on the household. The regressions in Panel A include lottery draw fixed effects, and the dependent variable "alive" is measured from the notification date through September 2009 ($N = 74{,}922$). The regressions in Panel B include survey wave fixed effects and the interaction of survey wave fixed effects with household size fixed effects, and are weighted using the survey weights ($N = 23{,}741$).

*These questions were worded to ask about number of days health "not good" or "impaired"; we switched the sign for consistency with the other measures. See Online Appendix Figure A4 for the exact survey wording.

"Quasi-experimental" Instruments

- In the Oregon Medicaid setting, the instrument Z_i was explicitly randomly assigned
- In other settings, researchers use an instrument that is not directly randomly but may be the result of idiosyncratic factors that are potentially "as good as random"
- This expands the set of cases where we can use IV, but means the required assumptions deserve extra scrutiny!

Example – Angrist and Krueger (1991)

- AK (1991) are interested in a classic question in labor economics: what are the labor market returns to an additional year of schooling?
- Why can't we just compare earnings for people who get more or less schooling?
  - Schooling is not randomly assigned! Choice of schooling may depend on many confounding factors, such as ability or family background, that would directly affect earnings as well.
- What do we need if we want to estimate the effects of schooling on earnings using IV?
- We need to find an instrument that is as-good-as-randomly assigned, and affects earnings only through years of schooling.

Compulsory Schooling Laws

- Most states have compulsory schooling laws that require people to stay in school until their 16th or 17th birthday
- AK argue that these laws will lead people planning to drop out to have less schooling if they are born earlier in the year
- That is, if the oldest kid in a class is born on January 1 and the youngest kid is born on December 31, then the oldest kid can legally drop out after getting 1 fewer year of education than the youngest kid
- AK argue that the part of the year in which you are born is effectively random, and therefore instrument for years of schooling with quarter of birth

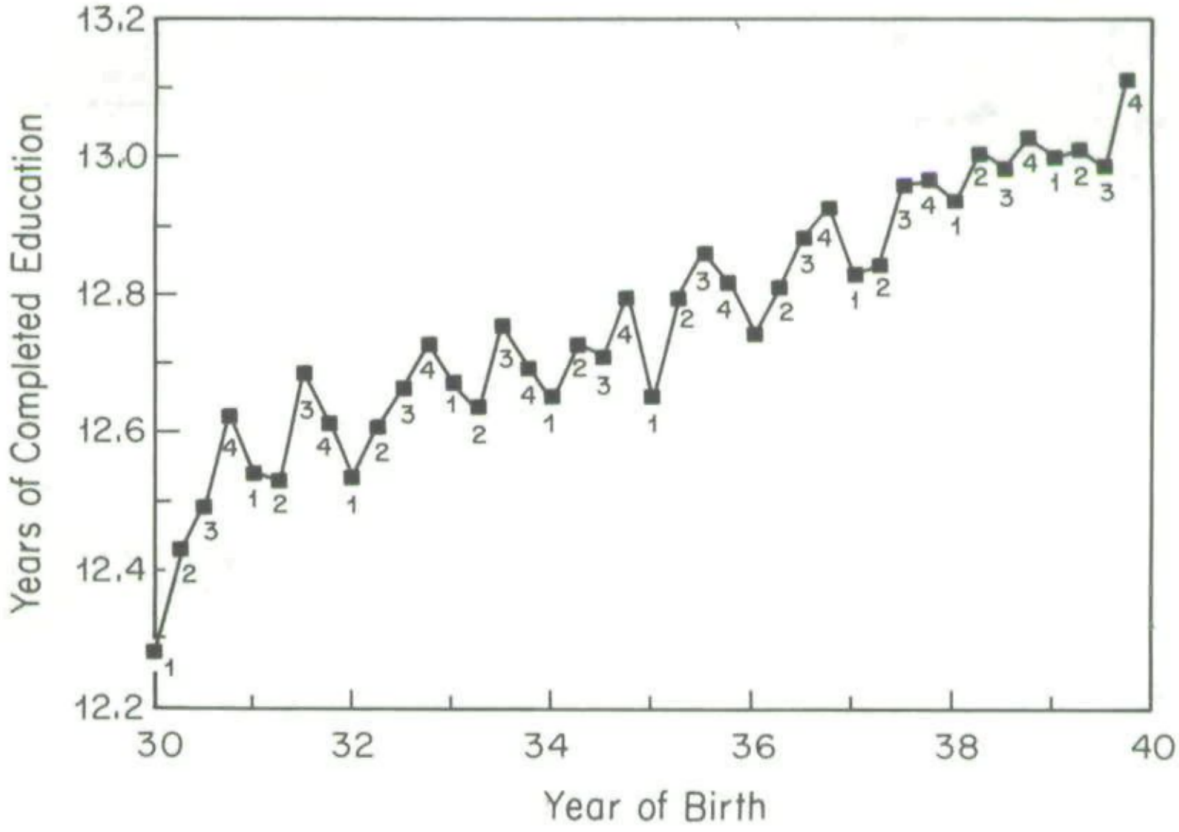


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

- Indeed, AK find that people born in the first quarter the of the year tend to have less schooling on average than people born in the remaining three quarters of the same year

PANEL A: WALD ESTIMATES FOR 1970 CENSUS—MEN BORN 192

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year
ln (wkly. wage)	5.1484	5.1574
Education	11.3996	11.5252

- What is the first stage estimate? $11.5252 - 11.3996 = 0.1256$
- What is the reduced form estimate? $5.1574 - 5.1484 = 0.0090$
- What is the 2SLS estimate? $0.0090/0.1256 = 0.0715$

TABLE III

PANEL A: WALD ESTIMATES FOR 1970 CENSUS—MEN BORN 1920–1929[a]

	(1) Born in 1st quarter of year	(2) Born in 2nd, 3rd, or 4th quarter of year	(3) Difference (std. error) (1) − (2)
ln (wkly. wage)	5.1484	5.1574	−0.00898
			(0.00301)
Education	11.3996	11.5252	−0.1256
			(0.0155)
Wald est. of return to education			0.0715
			(0.0219)
OLS return to education[b]			0.0801
			(0.0004)

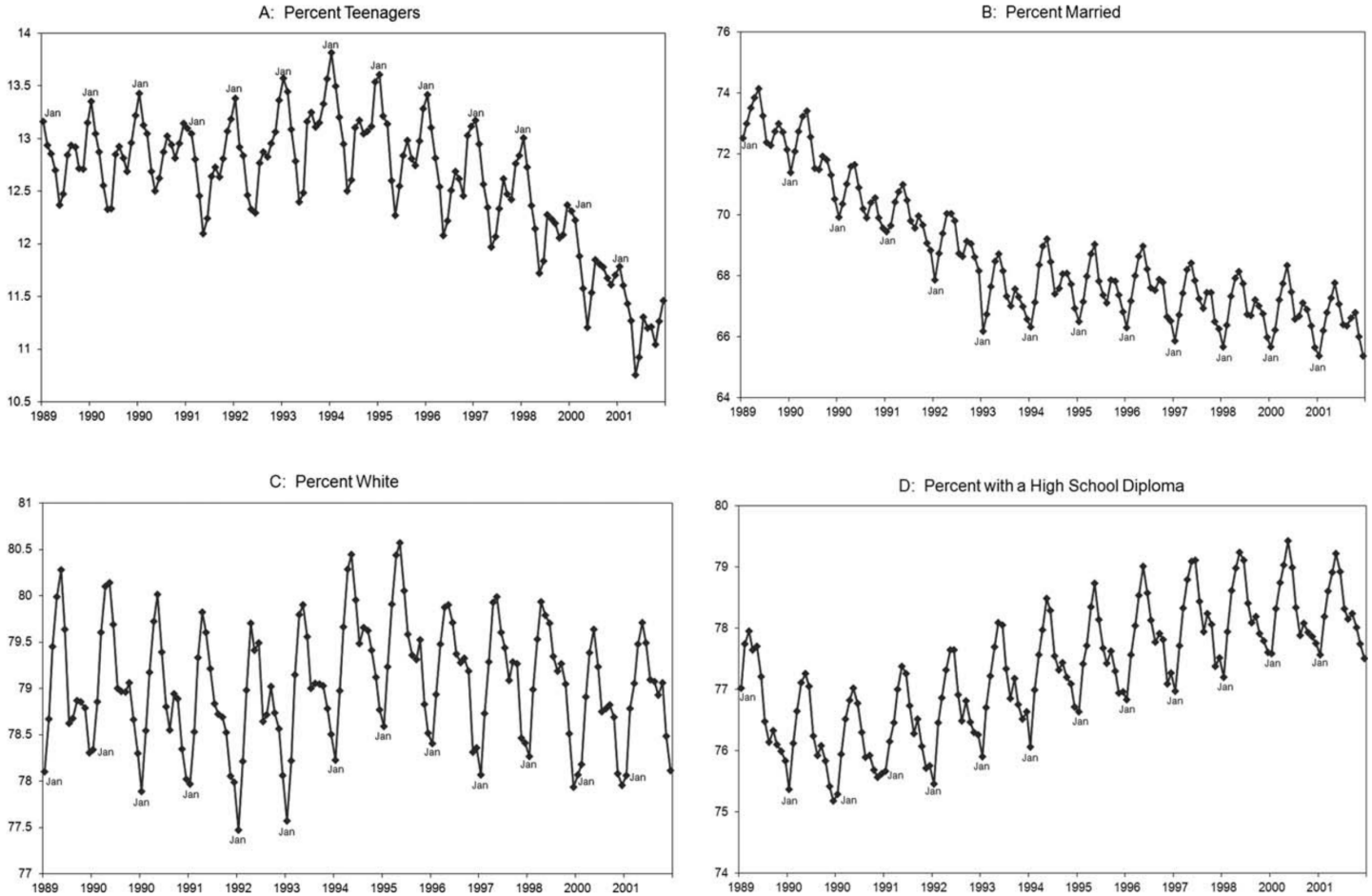
Evaluating the assumptions

- **Relevance:** Need quarter of birth to be correlated with years of education.
  ✓ This one we can check and indeed is the case (t-statistic of 12).

- **Independence:** Need quarter of birth to be independent of determinants of wages or responses to compulsory schooling laws ($Z_i \perp\!\!\!\perp (Y(.), D(.))$).

- Independence seems plausible if parents can't time the births of their children exactly. But...

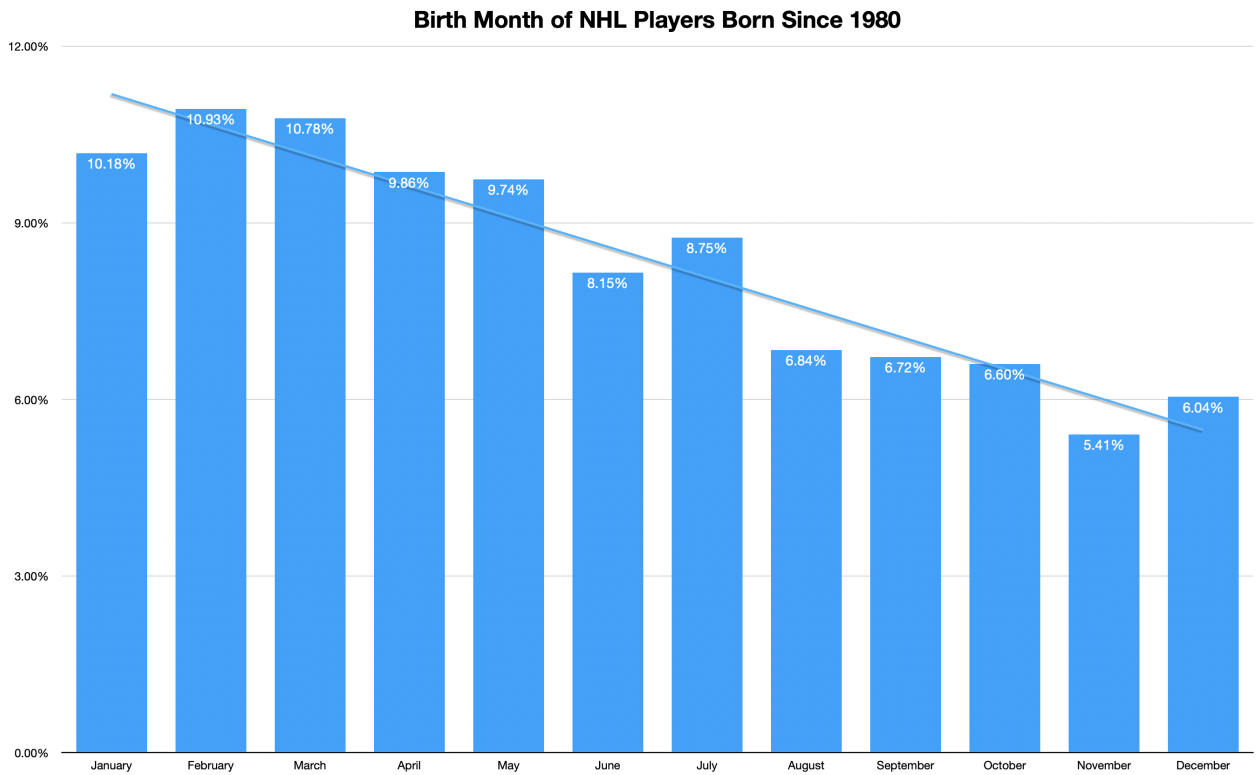
FIGURE 1.—MATERNAL CHARACTERISTICS BY MONTH, NATALITY FILES, 1989–2001

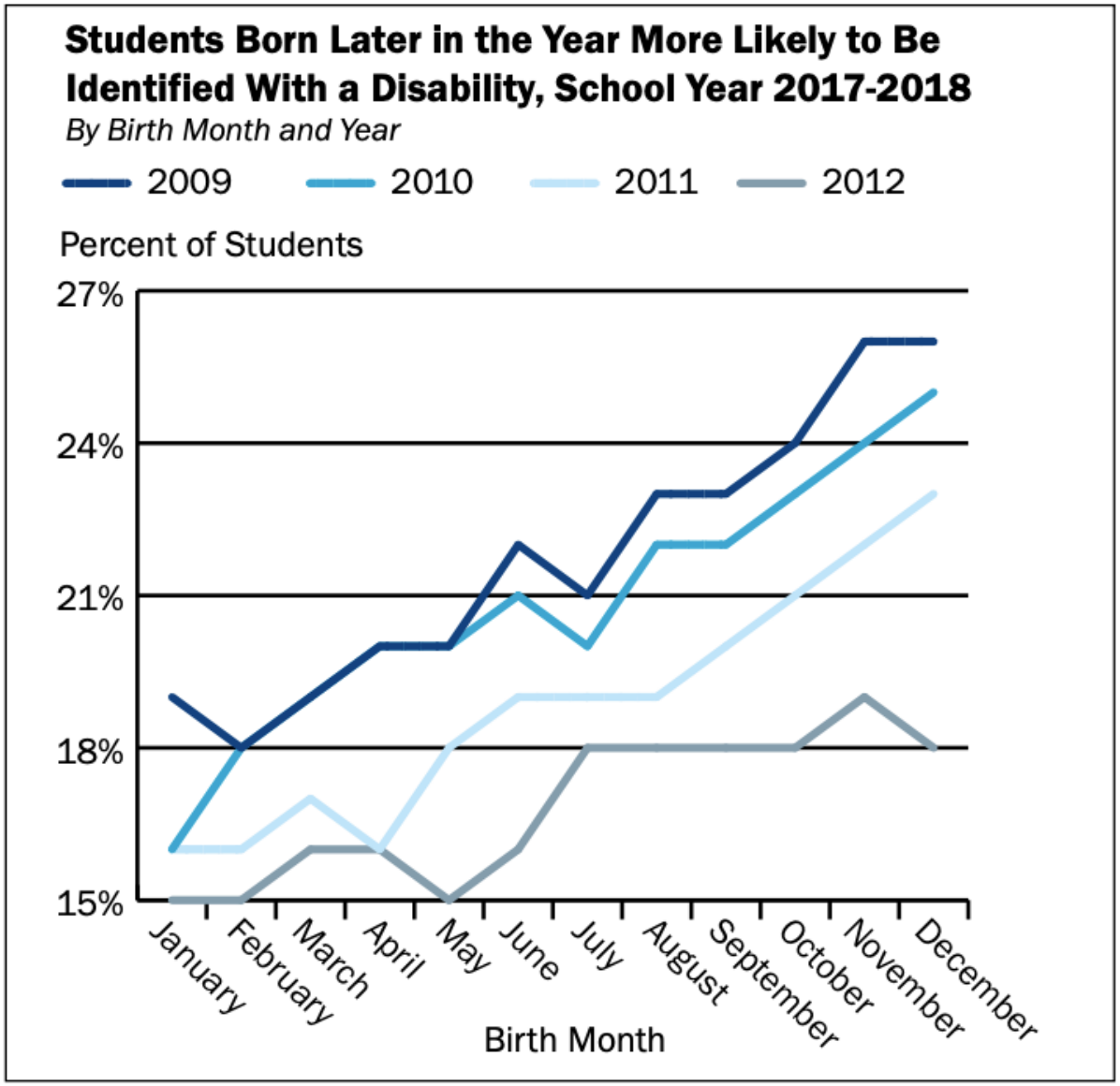


- Quarter of birth is correlated with some demographic characteristics
- Suggests some possible violations of independence

Evaluating the assumptions...

- **Exclusion:** Quarter of birth affects earnings only through number of years of education
- Implies that people whose years of schooling are unaffected by QOB would have same wages if born at different time of the year
- Seems generally plausible but... being older or younger in your grade may affect the quality of your education directly


Birth Month of NHL Players Born Since 1980
12.00%
9.00%
6.00%
3.00%
0.00%
10.18%
10.93%
10.78%
9.86%
9.74%
8.15%
8.75%
6.84%
6.72%
6.60%
5.41%
6.04%
January
February
March
April
May
June
July
August
September
October
November
December


Students Born Later in the Year More Likely to Be Identified With a Disability, School Year 2017-2018
By Birth Month and Year
2009
2010
2011
2012
Percent of Students
27%
24%
21%
18%
15%
January
February
March
April
May
June
July
August
September
October
November
December
Birth Month

Evaluating the assumptions

- **Monotonicity:** Everyone would get at least as many years of schooling if not born in the first quarter of the year
- Seems reasonable if all schools use a January 1 cutoff for school grades
- If some schools use a Sept 1 cutoff, it could be that being born in Q4 is even worse

TABLE II

PERCENTAGE OF AGE GROUP ENROLLED IN SCHOOL BY BIRTHDAY AND LEGAL DROPOUT AGE[a]

Date of birth	Type of state law[b]: School-leaving age: 16 (1)	Type of state law[b]: School-leaving age: 17 or 18 (2)	Column (1) − (2)
	Percent enrolled April 1, 1960		
1. Jan 1–Mar 31, 1944	87.6	91.0	−3.4
(age 16)	(0.6)	(0.9)	(1.1)
2. Apr 1–Dec 31, 1944	92.1	91.6	0.5
(age 15)	(0.3)	(0.5)	(0.6)
3. Within-state diff.	−4.5	−0.6	−4.0
(row 1 − row 2)	(0.7)	(1.0)	(1.2)

- Comparing states with different-aged CSLs suggested the effects of QOB on education operates through the CSL $\rightarrow$ assuages some of (but not all) concerns about independence

A treatment effect for whom?

- Suppose that we believe the IV assumptions hold (approximately).
- How do we interpret the treatment effect we estimate?
- It is a Local Average Treatment Effect (LATE) for compliers — i.e. people who would have gotten an extra year of school if they'd been born in the latter part of the year
- Why might the LATE not correspond with the ATE for the whole population?
  - Treatment effects could be different for people on the margin of dropping out

Multiple Instruments

- So far, we have considered binary instrumental variables (win vs lose lottery)
- Sometimes we may have an instrument that takes on multiple values (or multiple different instruments)
- Next we'll talk about how we can use similar ideas to exploit the variation from multi-valued instruments

Example – Angrist 1990

- Angrist (1990) is interested in the following Q: how does serving in the military (particularly the Vietnam War) affect your labor market earnings later in life?
- Why can't we we just compare veterans to non-veterans?
  - Veteran status is non-random! Veterans may differ in career goals, family background, physical ability, etc.
- If we want to use IV, what do we need? A plausibly random instrument that affects earnings only through whether you enroll in the military

Angrist 1990 – Draft Lotteries

- Angrist (1990) exploits the fact that during the Vietnam War there was a compulsory draft (for men)

- In the 1970s, the military conducted lotteries where each birthday was randomly assigned a priority number. People born on a birthday with a lower number were eligible to be drafted first

- Lottery numbers did not entirely determine whether one served in the military or not:
  - Some people with low numbers were given medical exemptions or left the country
  - Some people with high numbers volunteered to serve in the military anyway

- Angrist first considers a binary version of the instrument (eligible vs non-eligible bdays), then generalizes to multiple instruments (one for each bday)

Evaluating the IV assumptions

- **Independence:** Need lottery numbers to not be systematically related to potential earnings / potential enrollment
  - Randomization of the lottery makes this plausible
  - People with high/low numbers are similar on observable characteristics
- **Exclusion:** Need lottery number to affect earnings only through enrollment
  - Exclusion in this context is a bit tricky
  - Never-takers might have to flee the country if they have a low number but not a high one
  - Always-takers might have a different army experience if they enroll voluntarily
- **Relevance:** Lottery numbers must affect whether someone enrolls
  - Reasonable – and we'll see it's true!
- **Monotonicity:** Everyone who would enroll with a high number would also enroll with a low number
  - Seems reasonable

- Angrist first considers a binarized version of the instrument where he groups birthdays into eligible and non-eligible

- Here is his first stage (see $\hat{p}^e - \hat{p}^n$)

TABLE 2—VETERAN STATUS AND DRAFT ELIGIBILITY

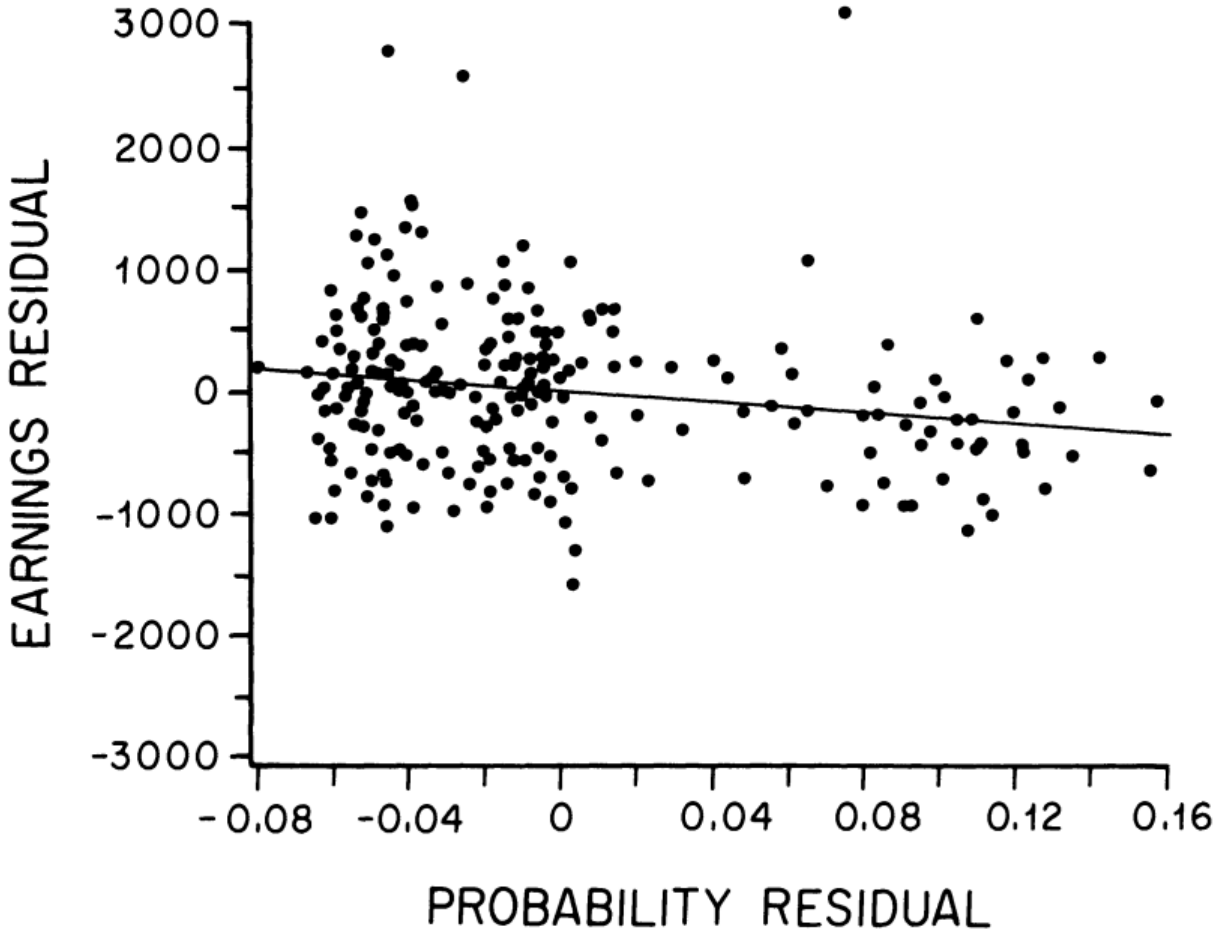
Whites						
Data Set	Cohort	Sample	P(Veteran)	$\hat{p}^e$	$\hat{p}^n$	$\hat{p}^e - \hat{p}^n$
SIPP (84)[a]	1950	351	0.2673	0.3527	0.1933	0.1594
			(0.0140)	(0.0325)	(0.0233)	(0.0400)
	1951	359	0.1973	0.2831	0.1468	0.1362
			(0.0127)	(0.0390)	(0.0180)	(0.0429)
	1952	336	0.1554	0.2310	0.1257	0.1053
			(0.0114)	(0.0473)	(0.0146)	(0.0495)
	1953	390	0.1298	0.1581	0.1153	0.0427
			(0.0106)	(0.0339)	(0.0152)	(0.0372)
DMDC/CWHS[b]	1950	16119	0.0633	0.0936	0.0279	0.0657
			(0.0019)	(0.0032)	(0.0019)	(0.0037)
	1951	16768	0.1176	0.2071	0.0708	0.1362
			(0.0025)	(0.0053)	(0.0024)	(0.0059)
	1952	17703	0.1515	0.2683	0.1102	0.1581
			(0.0027)	(0.0065)	(0.0027)	(0.0071)
	1953	17749	0.1343	0.1548	0.1268	0.0280
			(0.0026)	(0.0053)	(0.0029)	(0.0060)

TABLE 3—WALD ESTIMATES

Cohort	Year	Draft-Eligibility Effects in Current $: FICA Earnings (1)	Adjusted FICA Earnings (2)	Total W-2 Earnings (3)	$\hat{p}^e - \hat{p}^n$ (4)	Service Effect in 1978 $ (5)
1950	1981	−435.8	−487.8	−589.6	0.159	−2,195.8
		(210.5)	(237.6)	(299.4)	(0.040)	(1,069.5)
	1982	−320.2	−396.1	−305.5		−1,678.3
		(235.8)	(281.7)	(345.4)		(1,193.6)
	1983	−349.5	−450.1	−512.9		−1,795.6
		(261.6)	(302.0)	(441.2)		(1,204.8)
	1984	−484.3	−638.7	−1,143.3		−2,517.7
		(286.8)	(336.5)	(492.2)		(1,326.5)
1951	1981	−358.3	−428.7	−71.6	0.136	−2,261.3
		(203.6)	(224.5)	(423.4)	(0.043)	(1,184.2)
	1982	−117.3	−278.5	−72.7		−1,386.6
		(229.1)	(264.1)	(372.1)		(1,312.1)
	1983	−314.0	−452.2	−896.5		-2,181.8
		(253.2)	(289.2)	(426.3)		(1,395.3)
	1984	−398.4	−573.3	−809.1		−2,647.9
		(279.2)	(331.1)	(380.9)		(1,529.2)
1952	1981	−342.8	−392.6	−440.5	0.105	−2,502.3
		(206.8)	(228.6)	(265.0)	(0.050)	(1,556.7)
	1982	−235.1	−255.2	−514.7		−1,626.5
		(232.3)	(264.5)	(296.5)		(1,685.8)
	1983	−437.7	−500.0	−915.7		−3,103.5
		(257.5)	(294.7)	(395.2)		(1,829.2)
	1984	−436.0	−560.0	−767.2		−3,323.8
		(281.9)	(330.1)	(376.0)		(1,959.3)

Visualizing LATE in this example

See diagrams here



Notes: The figure plots the history of FICA taxable earnings for the four cohorts born 1950–53. For each cohort, separate lines are drawn for draft-eligible and draft-ineligible men. Plotted points show average real (1978) earnings of working men born in 1953, real earnings + $3000 for men born in 1950, real earnings + $2000 for men born in 1951, and real earnings + $1000 for men born in 1952.

FIGURE 1. SOCIAL SECURITY EARNINGS PROFILES BY DRAFT-ELIGIBILITY STATUS

Formalizing 2SLS with Multiple Instruments

- We can formalize this procedure with what's called two-stage least squares with multiple instruments.
- Suppose we have a vector of instruments $\boldsymbol{Z}_i$ (e.g. Z_1 is birthday 1, Z_2 is birthday 2)

2SLS with multiple instruments

- Step 1 (first stage): Estimate $E[D_i|\boldsymbol{Z}_i]$ by estimating the OLS regression specification:

$$D_i = \gamma_0 + \boldsymbol{Z}_i'\boldsymbol{\gamma}_1 + u_i$$

  E.g., estimate mean enrollment for each birthday

- Step 2: Construct the estimate of $E[D_i|Z_i]$ for each unit i:

$$\hat{D}_i = \hat{\gamma}_0 + \boldsymbol{Z}_i'\hat{\boldsymbol{\gamma}}_1$$

  E.g., $\hat{D}_i$ is average enrollment for people with i's birthday

- Step 3 (second stage): use OLS to regress Y_i on $\hat{D}_i$:

$$Y_i = \beta_0 + \hat{D}_i\beta_1 + \varepsilon_i$$

  E.g., regress earnings on average enrollment per birthday

- $\hat{\beta}_1$ is our estimate of the (weighted) LATE

TABLE 4—TWO-STAGE INSTRUMENTAL VARIABLES ESTIMATES

	Whites		
Cohort	FICA Taxable Earnings	Adjusted FICA Earnings	Total W-2 Compensation
Model 1			
1950	−1709.2	−2093.7	−1895.0
	(946.8)	(1108.8)	(1333.1)
1951	−1457.1	−1983.7	−2431.4
	(959.3)	(1036.1)	(1152.1)
1952	−1724.0	−1943.0	−2058.7
	(863.1)	(927.2)	(1001.9)
1953	1223.8	900.7	−488.6
	(3232.1)	(3505.3)	(3936.0)

Martin and Yurukoglu (2017)

- Martin and Yurukoglu are interested in the following Q: how does exposure to Fox News (conservative TV channel) affect voting patterns in the US?

- We could compare places where people watch more Fox News to places where people watch less Fox News. Would that give us a causal effect?
  - Probably not! There is likely a strong confounding variable of political preference (people who are more conservative watch more Fox News and vote Republican)

- We need an instrument that is as good as randomly assigned and affects voting patterns only through its impact on Fox News viewership. Any ideas?

- Martin and Yurukoglu propose to use the position of Fox News in the channel lineup as an IV
  - People watch more Fox News if it appears earlier in the guide
  - Argue that channel lineup was determined by idiosyncratic factors during rollout of cable news in the 1990s

Martin and Yurukoglu's 2SLS Specification

First stage:

$$D_i = \pi Z_i + \boldsymbol{\gamma}' \boldsymbol{X}_i + u_i$$

where D_i is average minutes of Fox News watched per week in Zip code i, $\boldsymbol{X}_i$ is a vector of control variables including state or county FEs, the position of MSNBC, and some demographic variables (income, age, race, etc)

Second stage:

$$Y_i = \beta \hat{D}_i + \boldsymbol{\gamma}' \boldsymbol{X}_i + \varepsilon_i$$

where Y_i is Republican vote share in the 2008 election and $\hat{D}_i$ is the prediction from the first-stage

Evaluating the assumptions

- **Independence:** Need that Fox News position is as good as random (conditional on observable characteristics). This is the most tenuous: seems plausible, but worry that Fox News might be put earlier in places with higher demand
- **Exclusion:** Need that Fox News position impacts voting only through Fox News viewership. Seems relatively plausible
- **Relevance:** Need that cable position impacts Fox News viewership. We'll see that it does
- **Monotonicity:** Need that being in a lower cable position only increases viewership. Seems plausible, although could be violated if being later increases proximity to other popular channels

First Stage

TABLE 2—FIRST-STAGE REGRESSIONS: NIELSEN DATA

	FNC minutes per week					
	(1)	(2)	(3)	(4)	(5)	(6)
FNC position	−0.146	−0.075	−0.174	−0.167	−0.097	−0.111
	(0.043)	(0.039)	(0.028)	(0.025)	(0.033)	(0.030)
MSNBC position	0.078	0.073	0.064	0.070	0.019	0.020
	(0.036)	(0.032)	(0.025)	(0.022)	(0.034)	(0.035)
Has MSNBC only	1.904	1.137	−3.954	−2.804	−1.220	−1.562
	(3.697)	(3.713)	(4.255)	(3.416)	(6.180)	(5.397)
Has FNC only	31.423	26.526	23.460	22.011	15.141	15.069
	(2.677)	(2.546)	(2.278)	(1.864)	(2.697)	(2.314)
Has both	24.859	23.118	18.338	16.168	15.159	14.486
	(2.919)	(2.687)	(2.361)	(1.991)	(3.216)	(2.842)
Satellite FNC minutes				0.197		0.173
				(0.013)		(0.015)
Fixed effects	Year	State-year	State-year	State-year	County-year	County-year
Cable controls	Yes	Yes	Yes	Yes	Yes	Yes
Demographics	None	None	Extended	Extended	Extended	Extended
Robust F-stat	11.39	3.72	39.02	44.7	8.86	13.43
Number of clusters	5,789	5,789	4,830	4,761	4,839	4,770
Observations	71,150	71,150	59,541	52,053	59,684	52,165
R^2	0.030	0.074	0.213	0.377	0.428	0.544

- Being lower in the line-up predicts lower Fox News viewership. A 1SD improvement in lineup position corresponds to roughly 2.5 minutes per week more viewership

Some evidence on independence

TABLE 5—FNC CABLE POSITION PLACEBO TESTS

	Predicted viewing		Predicted voting		1996 contributions		1996 vote
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
FNC position	0.100	0.033	0.027	−0.0004	0.0002	0.001	−0.006
	(0.034)	(0.026)	(0.022)	(0.017)	(0.0002)	(0.0003)	(0.012)
Fixed effects	State-year	County-year	State-year	County-year	State-year	County-year	State-year
Demographics	Extended	Extended	Extended	Extended	Extended	Extended	Extended
Number of clusters	4,830	4,839	4,814	4,827	4,830	4,839	4,830
Observations	59,551	59,694	17,400	17,451	59,551	59,694	59,551
R^2	0.380	0.827	0.339	0.729	0.176	0.436	0.571

- Fox News position is not significantly related to vote share in 1996 (before cable news), and to most (but not all) demographic characteristics

Validation using satellite viewers

TABLE 6—SATELLITE PLACEBO FIRST STAGE: NIELSEN DATA

	FNC minutes per week			
	(1)	(2)	(3)	(4)
FNC position × cable	−0.155	−0.264	−0.151	−0.219
	(0.043)	(0.035)	(0.048)	(0.051)
FNC position × sat	0.031	−0.050	0.037	0.045
	(0.049)	(0.041)	(0.063)	(0.067)
MSNBC position × cable	0.102	0.092	0.035	0.046
	(0.036)	(0.032)	(0.049)	(0.048)
MSNBC position × sat	−0.004	−0.029	−0.029	−0.033
	(0.040)	(0.033)	(0.072)	(0.074)
Fixed effects	State-year	State-year	County-year	County-year
Cable controls	Yes	Yes	Yes	Yes
Demographics	None	Extensive	None	Extensive
Chow test p-value	0	0	0.011	0.001
Number of clusters	5,786	4,830	5,786	4,830
Observations	127,072	107,829	127,072	107,829
R^2	0.032	0.077	0.232	0.278

- Fox News position predicts viewership only on cable and not on satellite TV (which has different lineup)

2SLS results

TABLE 4—SECOND STAGE REGRESSIONS: ZIP CODE VOTING DATA

	2008 McCain vote percentage			
	(1)	(2)	(3)	(4)
Predicted FNC minutes	0.152 (0.056, 0.277)	0.120 (0.005, 0.248)	0.157 (−0.126, 0.938)	0.098 (−0.121, 0.429)
Satellite FNC minutes		−0.021 (−0.047, 0.001)		−0.015 (−0.073, 0.022)
Fixed effects	State	State	County	County
Cable system controls	Yes	Yes	Yes	Yes
Demographics	Extended	Extended	Extended	Extended
Number of clusters	4,814	3,993	4,729	4,001
Observations	17,400	12,417	17,283	12,443
R^2	0.833	0.841	0.907	0.919

A 1SD improvement in channel position increases viewership by about 2.5 minutes/wk, which is estimated to move vote share by 0.3 percentage pts

Inference for 2SLS

- How do we get standard errors for two-stage least squares estimates?
- We just showed that two-stage least squares is equivalent to running OLS with the regressor $\hat{D}$ from the first-stage?
- You might be tempted to just run this second stage regression and use the OLS standard errors. But this will not give you the correct answer because it won't account for estimation error in $\hat{D}$
- Luckily, it is easy to get correct 2SLS standard errors from Stata or other software packages
  - `ivreg y (d=z) x, r` in Stata estimates 2SLS for treatment d with instrument z and controls x
- Where do these standard errors come from?

Normality of reduced form and first stage

- Remember that with a single instrument, $\hat{\beta}_{IV} = \hat{\gamma}_1/\hat{\pi}_1$, where $\hat{\gamma}_1, \hat{\pi}_1$ are OLS estimates of

$$Y_i = \gamma_0 + Z_i\gamma_1 + \varepsilon_i$$
$$D_i = \pi_0 + Z_i\pi_1 + u_i$$

- We've shown that OLS estimates are asymptotically normally distributed (and these convergences hold jointly), so we will have

$$\sqrt{N}\left(\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\pi}_1 \end{pmatrix} - \begin{pmatrix} \gamma_1 \\ \pi_1 \end{pmatrix}\right) \to_d \mathrm{N}(0, \Sigma)$$

- Can we learn from this the asymptotic distribution of $\hat{\gamma}_1/\hat{\pi}_1$?

Delta method

- Let $g(\hat{\gamma}_1, \hat{\pi}_1) = \hat{\gamma}_1 / \hat{\pi}_1$. For $(\hat{\gamma}_1, \hat{\pi}_1) \approx (\gamma_1, \pi_1)$ a first-order Taylor expansion tells us that

$$g(\hat{\gamma}_1, \hat{\pi}_1) \approx g(\gamma_1, \pi_1) + \nabla g(\gamma_1, \pi_1)(\hat{\gamma}_1 - \gamma_1, \hat{\pi}_1 - \pi_1)'$$

  where $\nabla g(\gamma_1, \pi_1)$ is the gradient of g evaluated at (γ_1, π_1).

- This implies that

$$\sqrt{N}\left(\underbrace{\frac{\hat{\gamma}_1}{\hat{\pi}_1}}_{g(\hat{\gamma}_1, \hat{\pi}_1)} - \underbrace{\frac{\gamma_1}{\pi_1}}_{g(\gamma_1, \pi_1)} \right) \approx \nabla g(\gamma_1, \pi_1) \sqrt{N} \left(\begin{pmatrix} \hat{\gamma}_1 \\ \hat{\pi}_1 \end{pmatrix} - \begin{pmatrix} \gamma_1 \\ \pi_1 \end{pmatrix} \right)$$

- By the continuous mapping theorem, this converges in distribution to $\mathrm{N}(0, \nabla g(\gamma_1, \pi_1) \Sigma \nabla g(\gamma_1, \pi_1)')$.

- We can estimate the variance with sample analogs (e.g. $\nabla g(\hat{\gamma}_1, \hat{\pi}_1)$)).

Weak identification

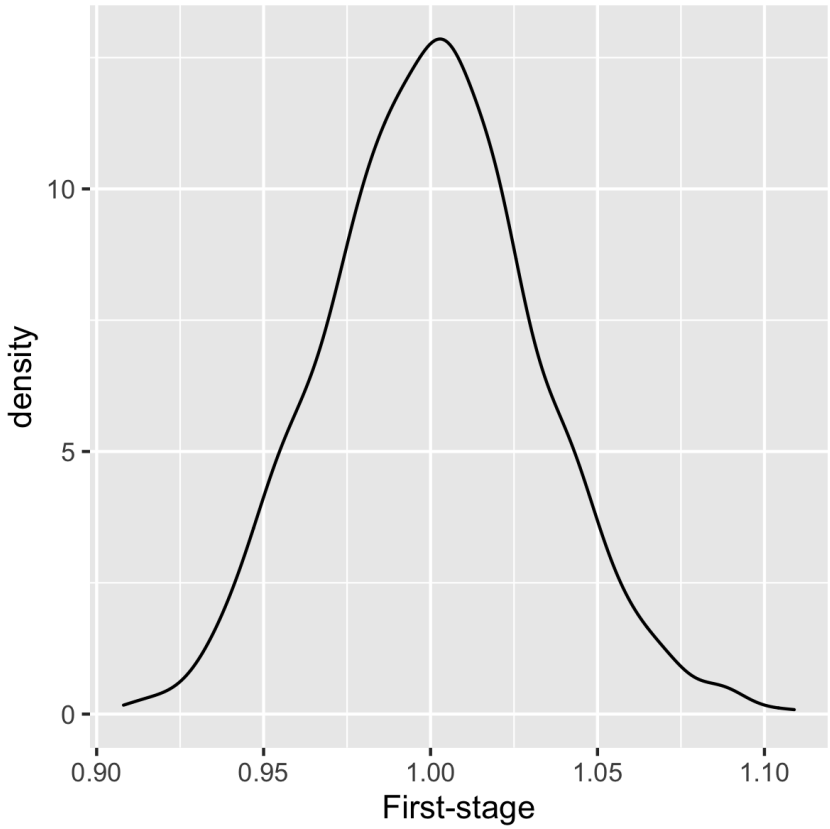
- Note that $\hat{\beta}_{IV} = \hat{\gamma}_1 / \hat{\pi}_1$ is only well-defined for $\hat{\pi}_1 \neq 0$.
- As $\hat{\pi}_1 \to 0$, $|\hat{\beta}_{IV}| \to \infty$, so the IV estimator will be poorly behaved if $\hat{\pi}_1 \approx 0$.
- This wasn't a problem in our asymptotics because we assumed $\pi_1 \neq 0$ (relevance), and as N gets large, $\hat{\pi}_1 \to_p \pi_1$, so asymptotically $\hat{\pi}_1$ is near-zero with probability approaching zero.
- But in practice, $\hat{\pi}_1$ may sometimes be close to zero (relative to its standard error).
- In this case, the normal distribution above may provide a poor approximation to the distribution of $\hat{\beta}_{IV}$. This is a problem known as *weak identification*.

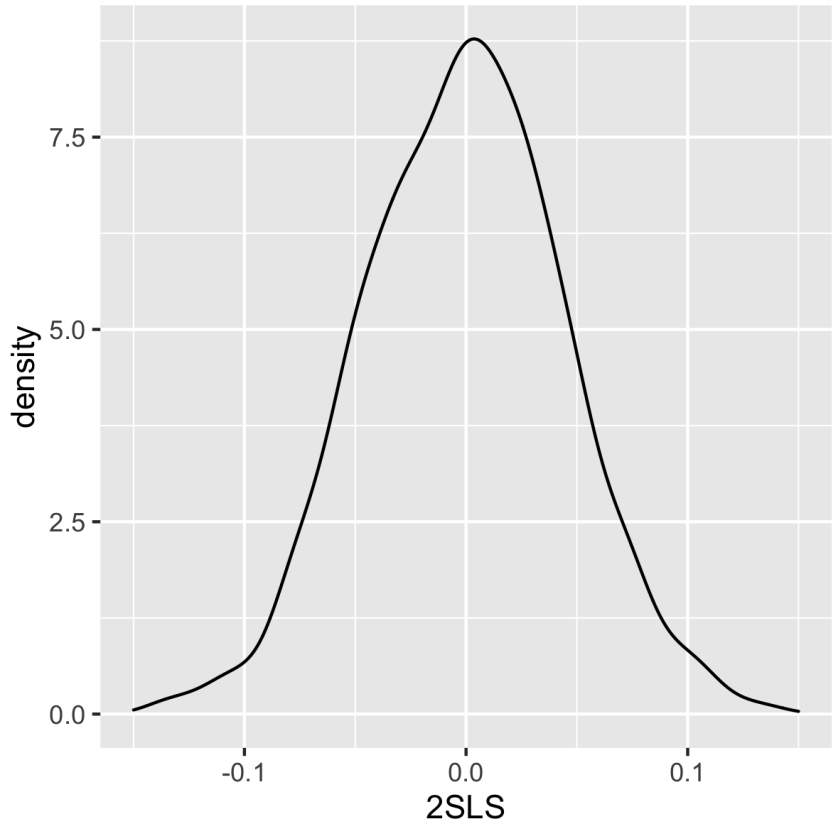
Weak Instruments - Monte Carlo

Monte Carlo Simulation: No True Treatment Effect

$Y_i(d) = \eta + \nu; \qquad D_i = \pi_1 Z_i + \eta; \qquad \eta, \nu, Z_i \sim N(0,1)$

Consider first a strong first-stage: $\pi_1 = 1 \rightarrow \frac{\pi_1}{SD(\pi_1)} \approx 22$



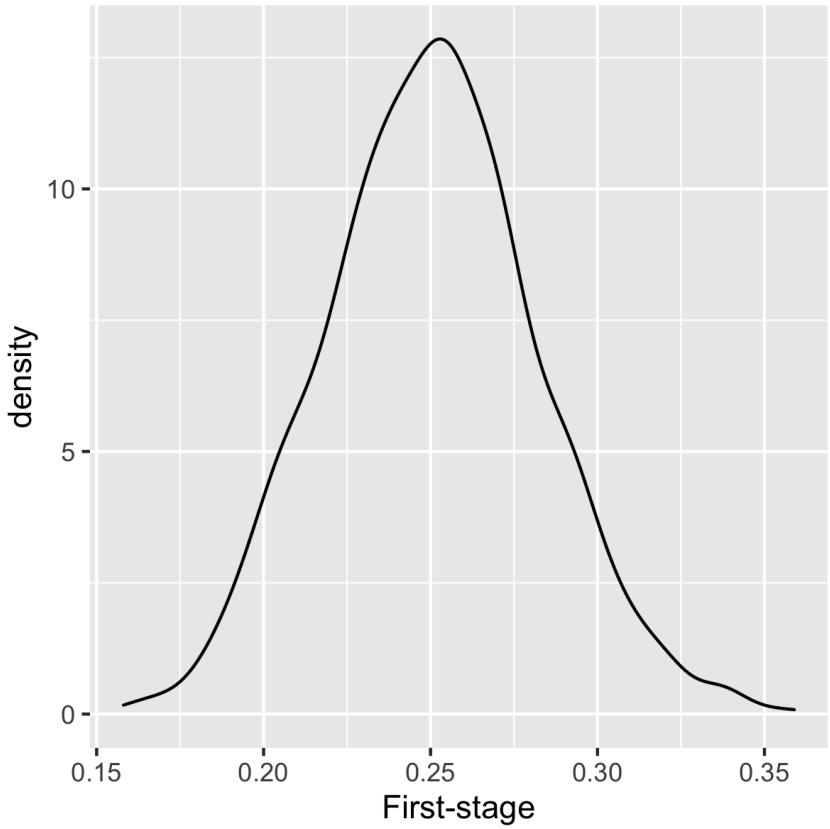


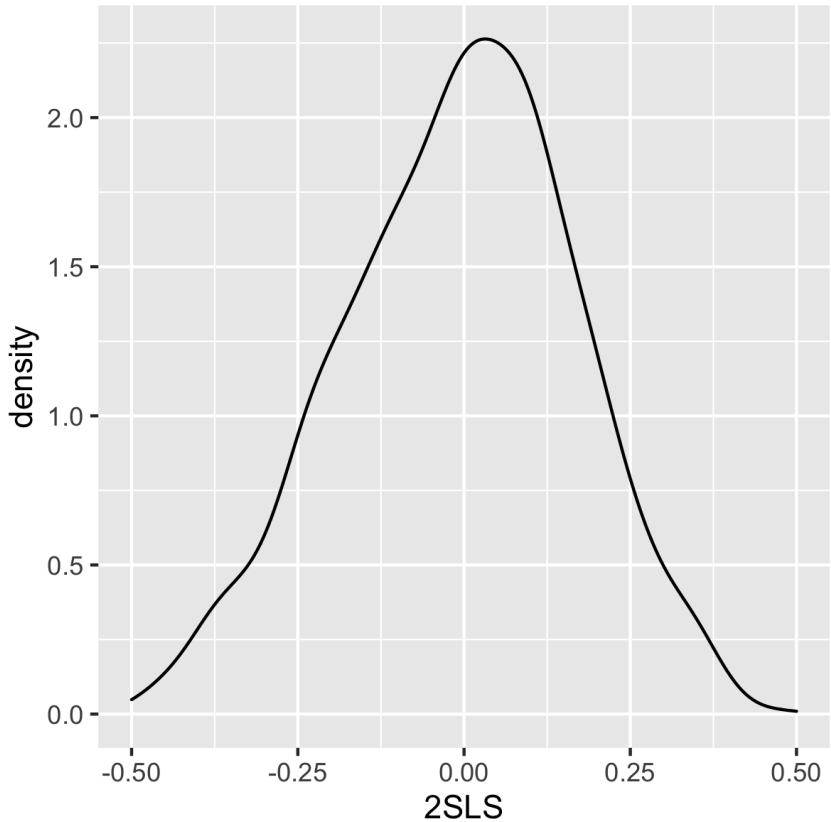
Weak Instruments - Monte Carlo

Monte Carlo Simulation: No True Treatment Effect

$Y_i(d) = \eta + \nu; \qquad D_i = \pi_1 Z_i + \eta; \qquad \eta, \nu, Z_i \sim N(0,1)$

Consider next a medium first-stage: $\pi_1 = 0.25 \rightarrow \frac{\pi_1}{SD(\pi_1)} \approx 7$



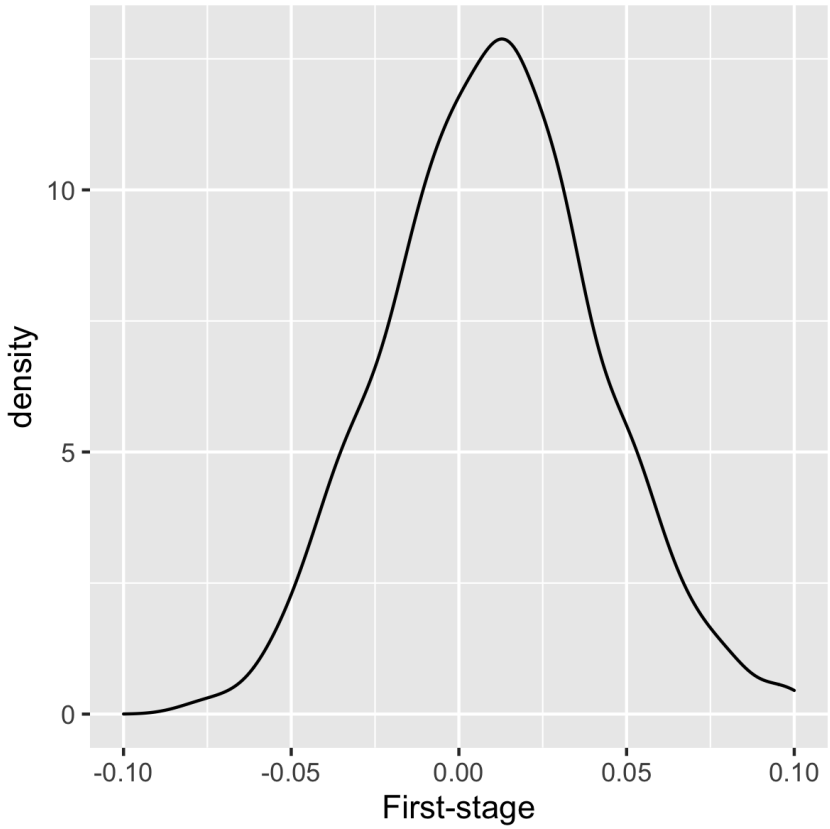


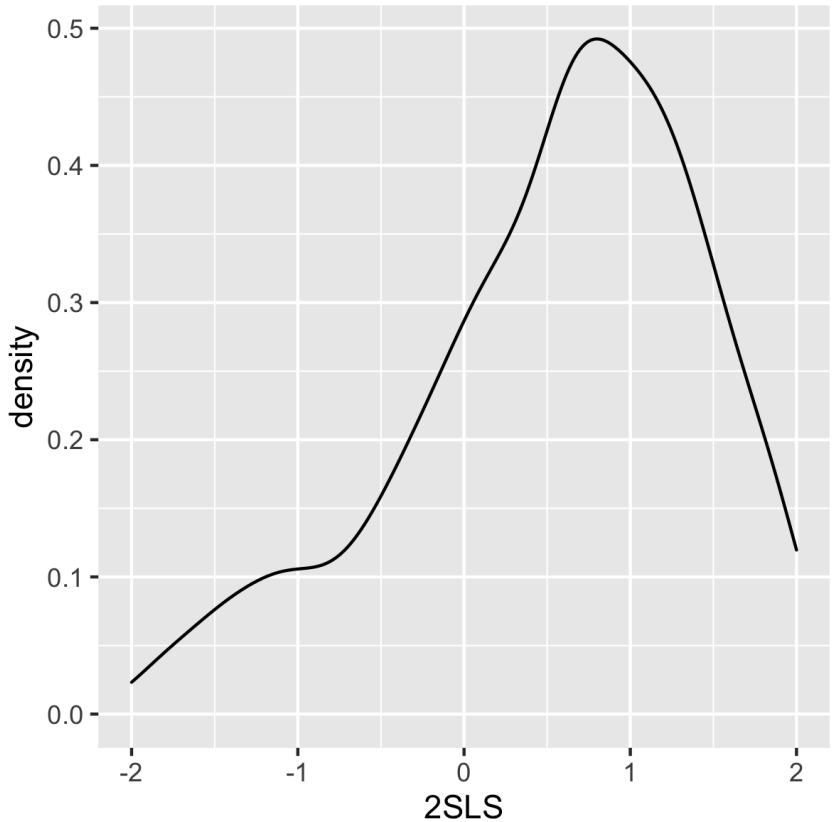
Weak Instruments - Monte Carlo

Monte Carlo Simulation: No True Treatment Effect

$Y_i(d) = \eta + \nu; \qquad D_i = \pi_1 Z_i + \eta; \qquad \eta, \nu, Z_i \sim N(0,1)$

Consider next a very weak first-stage: $\pi_1 = 0.01 \rightarrow \frac{\pi_1}{SD(\pi_1)} \approx 0.3$





Weak IV with multiple instruments

A conceptually related problem arises when we have many instruments

- Canonical example / cautionary tale: Angrist and Krueger (1991) interact the quarter-of-birth instruments with year and state of birth
- This decreased their SEs: better in-sample prediction of D_i
- Bound et al. (1995) famously show they get the same 2SLS estimates from randomly generated instruments

Intuition: when we have many instruments in the first-stage, we will "overfit" D_i.

- Imagine the extreme case, where every observation gets its own instrument (i.e. interact quarter-of-birth with individual)
- First-stage fit will be *perfect*: $\hat{D}_i = \hat{\boldsymbol{\pi}}_1' Z_i = D_i$. So 2SLS = OLS numerically
- Thus the weak instrument problem arises from "overfitting" in the first stage

Testing for weak IV

- A typical rule of thumb is to worry about weak instruments if the F-statistic on the instruments in the first stage is < 10
- That is, run the first-stage regression

$$D_i = \pi_0 + \boldsymbol{Z}_i'\boldsymbol{\pi}_1 + \boldsymbol{X}_i'\boldsymbol{\pi}_2 + u_i$$

  and construct the F-statistic for the null $H_0 : \boldsymbol{\pi}_1 = 0$.

- With one instrument, the F-statistic is just the t-statistic on the instrument squared, so $F > 10$ corresponds with $|t| > 3.2$.

```
. reg education D highnum_region_D post high_intensity highnum_region highnum_regi
> on_post highnum_region_high, r

Linear regression                               Number of obs     =     31,310
                                                F(7, 31302)       =     133.29
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0282
                                                Root MSE          =     3.8458

------------------------------------------------------------------------------
             |               Robust
   education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           D |  -.1623344   .1236149    -1.31   0.189    -.4046245    .0799557
highnum_regio~D|   .5457673   .1761518     3.10   0.002     .2005028    .8910318
        post |   .3354358   .0763198     4.40   0.000     .1858459    .4850256
high_intensity|  -1.309835   .0951979   -13.76   0.000    -1.496426   -1.123243
highnum_region|  -.3988454   .0838969    -4.75   0.000    -.5632866   -.2344043
highnum_regio~t|  -.2082078   .1109533    -1.88   0.061    -.4256806     .009265
highnum_regio~h|   .0932633   .1348633     0.69   0.489    -.1710742    .3576008
         _cons |   10.05763   .0590164   170.42   0.000     9.941953    10.1733
------------------------------------------------------------------------------

. test D highnum_region_D

 ( 1)  D = 0
 ( 2)  highnum_region_D = 0

       F(  2, 31302) =    5.53
            Prob > F =    0.0040
```

Here, the instruments are D and highnumregion_D (confusing labels - sorry!). First-stage F is 5.53

```
. ivreg2 log_wage (education=D highnum_region_D) post high_intensity highnum_regio
> n highnum_region_post highnum_region_high, r

IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics robust to heteroskedasticity

                                                      Number of obs =    31310
                                                      F(  6, 31303) =   276.17
                                                      Prob > F      =   0.0000
Total (centered) SS     =  13459.19497                Centered R2   =  -0.2431
Total (uncentered) SS   =  4570266.616                Uncentered R2 =   0.9963
Residual SS             =  16731.05867                Root MSE      =     .731

------------------------------------------------------------------------------
             |               Robust
   log_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .1856081   .0584726     3.17   0.002     .0710038    .3002124
        post |  -.2773642   .0197758   -14.03   0.000     -.316124   -.2386044
 high_intensity |   .0949207   .0819745     1.16   0.247    -.0657465    .2555878
 highnum_region |  -.0947804   .0319689    -2.96   0.003    -.1574383   -.0321225
highnum_regio~t |  -.0734455   .0165758    -4.43   0.000    -.1059335   -.0409576
highnum_regio~h |   .0379194   .0261998     1.45   0.148    -.0134312      .08927
          _cons |   10.45536   .5899054    17.72   0.000     9.299167    11.61155
------------------------------------------------------------------------------

Underidentification test (Kleibergen-Paap rk LM statistic):             11.057
                                                      Chi-sq(2) P-val =   0.0040
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                5.617
                         (Kleibergen-Paap rk Wald F statistic):          5.530
Stock-Yogo weak ID test critical values: 10% maximal IV size            19.93
                                         15% maximal IV size            11.59
                                         20% maximal IV size             8.75
                                         25% maximal IV size             7.25
Source: Stock-Yogo (2005).  Reproduced by permission.
NB: Critical values are for Cragg-Donald F statistic and i.i.d. errors.
------------------------------------------------------------------------------
```

We can get the same F-stat directly from the ivreg2 command

What do about weak instruments?

- There are some ways to reducing over-fitting in the first-stage by splitting the sample (split sample or jacknife IV)
- There are alo some better ways to get confidence intervals when you have weak instruments
  - Most common is what are called Anderson-Rubin confidence sets (we won't have time to cover)
- Can also try to increase the strength of the instrument by:
  - Getting a larger sample
  - Adding control variables that correlate with D (but not strongly with Z)
  - Thinking of a new instrument ;-)