Webinar Series in Applied Quantitative Analysis - Updated

Date	Topic	
	Session One	
February 29	Potential Outcomes and Omitted Variable Bias I (Theory)	✓
March 7	Potential Outcomes and Omitted Variable Bias II (Application)	
	Session Two	
March 21	Difference-in-differences I (Theory)	✓
March 28	Difference-in-differences II (Application)	
	Session Three	
April 25	Power analysis, clustering and sample size calculations I (Theory)	✓
May 2	Power analysis, clustering and sample size calculations II (Application)	
	Session Four	
May 23	Propensity score matching (Theory)	✓
May 30	Propensity score matching (Application)	
	Session Five	
June 20	Fixed-effects I (Theory)	✓
June 27	Fixed-effects II (Application)	
	Session Six	
July 25	Instrumental variables I (Theory)	
August 1	Instrumental Variables II (Application)	
	Session Seven	
August 22	Lagged dependent variables and the Arellano-Bond Estimator I (Theory)	
August 29	Lagged dependent variables and the Arellano-Bond Estimator II (Application)	

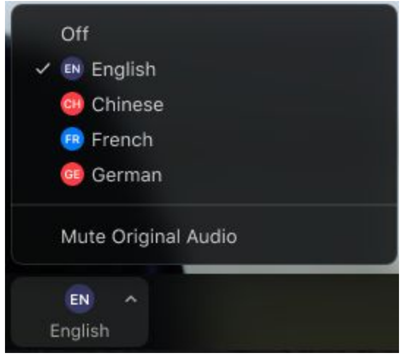
AIR®
Advancing Evidence.
Improving Lives.

https://www.airp3-africa.org/applied-quantitative-analysis-webinar-series

AERC CREA

Écoute de l'interprétation d'une langue Windows | macOS

1. Dans les contrôles de votre réunion/webinaire, cliquez sur Interprétation .
2. Cliquez sur la langue que vous souhaitez entendre. (Nous aurons le français) Pas besoin de choisir l'anglais, c'est la langue de la salle Zoom principale



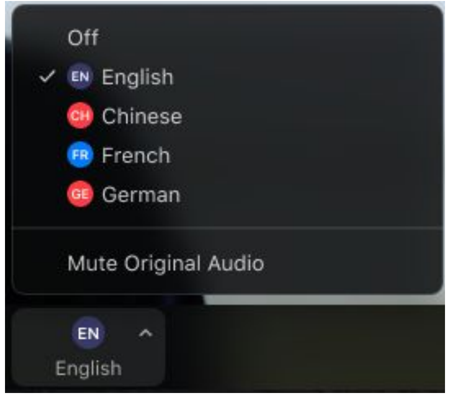
3. (Facultatif) Pour entendre uniquement la langue interprétée, cliquez sur Couper le son original.

Remarques:

- Vous devez rejoindre l’audio de la réunion via l’audio/VoIP de votre ordinateur. Vous ne pouvez pas écouter l’interprétation linguistique si vous utilisez les fonctions audio de connexion ou d’appel téléphonique.
- En tant que participant rejoignant une chaîne linguistique, vous pouvez retransmettre sur le canal audio principal canal si vous réactivez votre audio et parlez.

Listing to language interpretation
Windows | macOS

1. In your meeting/webinar controls, click Interpretation .
2. Click the language that you would like to hear. (We will have French) No need to choose English, that is the language in the main Zoom room



3. (Optional) To hear the interpreted language only, click Mute Original Audio.

Notes:

- You must join the meeting audio through your computer audio/VoIP. You cannot listen to language interpretation if you use the dial-in or call me phone audio features.
- As a participant joining a language channel, you can broadcast back into the main audio channel if you unmute your audio and speak.

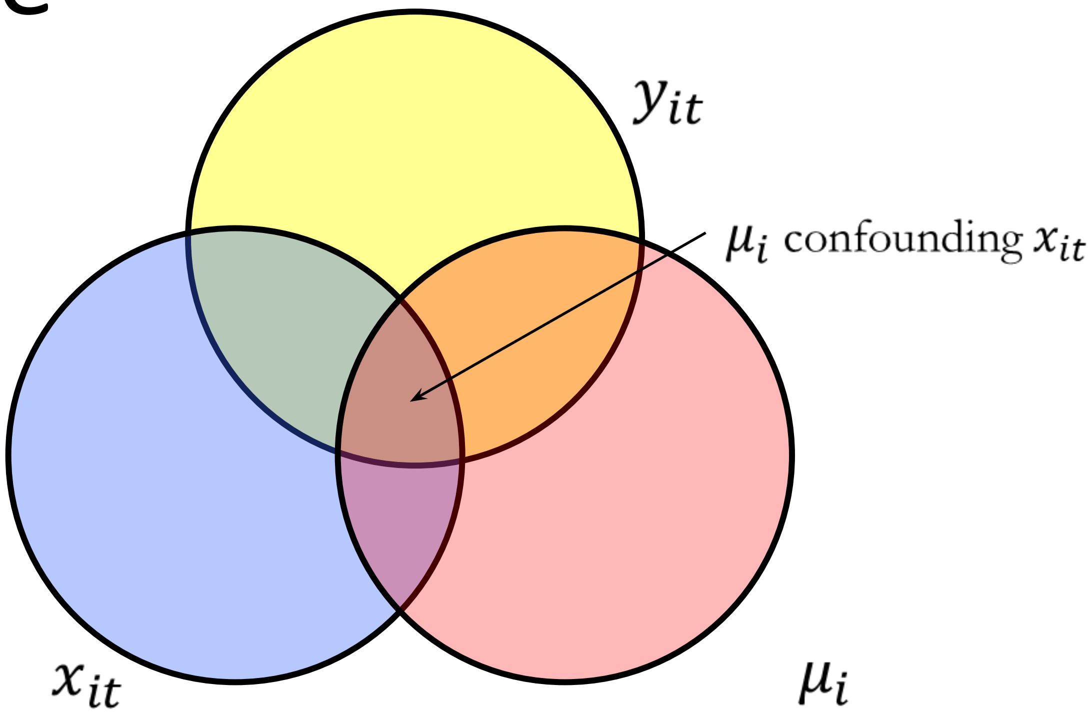
Instrumental Variables Theory

Professor Jeremy Moulton
Department of Public Policy
University of North Carolina at Chapel Hill
July 25, 2024

Advancing Evidence.
Improving Lives.

The Set Up From Last Time

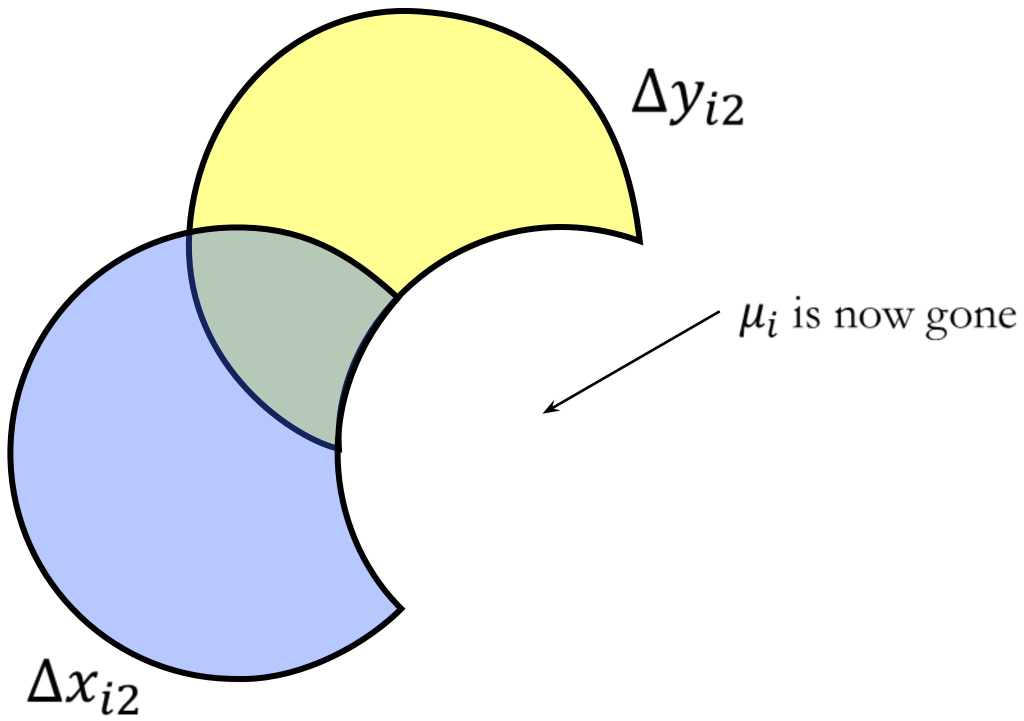
- $y_{it} = \alpha + \beta x_{it} + \varepsilon_{it}$
  - $\varepsilon_{it} = \nu_{it} + \mu_i$



- x_{it} = independent variable of interest
- μ_i = time-invariant individual specific error
  - Example: innate ability or health endowment

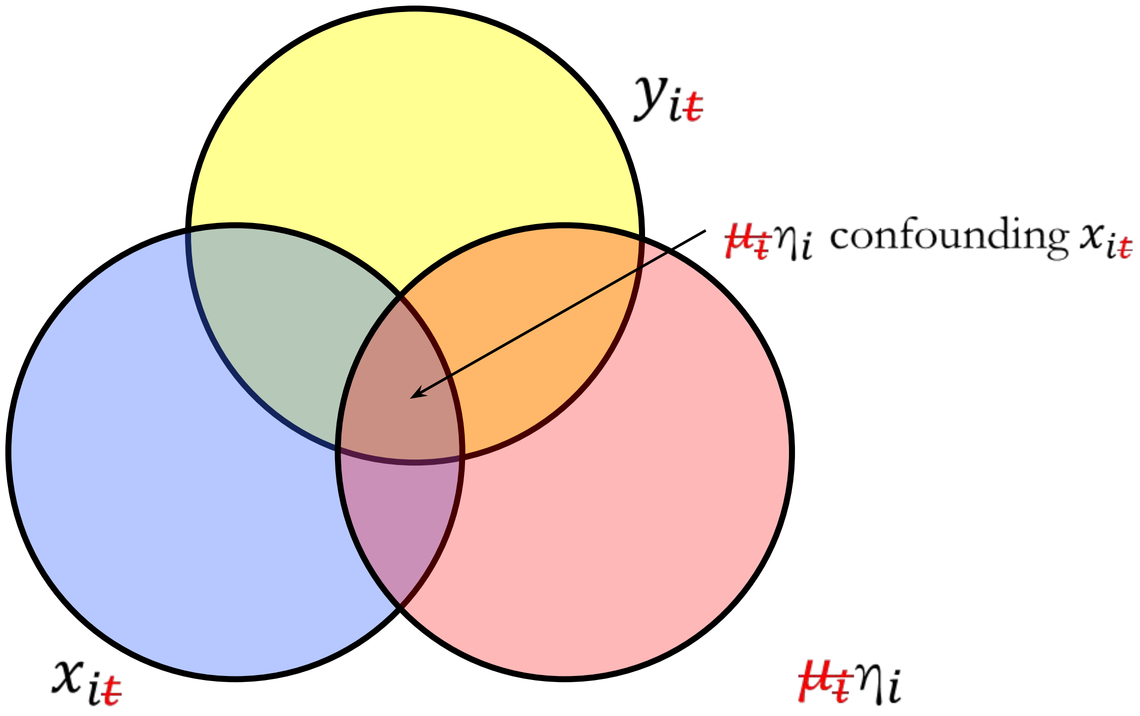
The Old Solution (Fixed Effects/First Differences)

$$\Delta y_{i2} = \beta \Delta x_{i2} + \Delta \nu_{i2}$$



Change in the Set Up

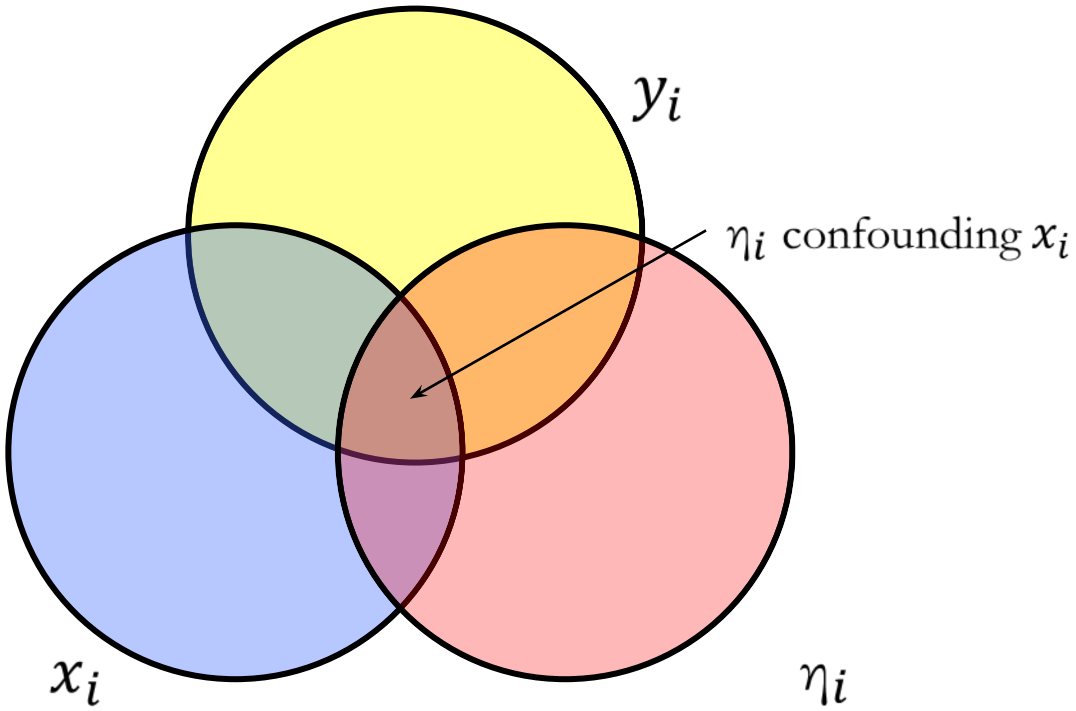
- $y_{i\cancel{t}} = \alpha + \beta x_{i\cancel{t}} + \varepsilon_{i\cancel{t}}$
  - $\varepsilon_{i\cancel{t}} = \nu_{i\cancel{t}} + \cancel{\mu_{\cancel{t}}} + \eta_i$



- $x_{i\cancel{t}}$ = independent variable of interest
- ~~$\mu_{\cancel{t}}$ = time-invariant individual specific~~ η_i = error that is correlated with x_i
  - Example: innate ability ~~or health endowment~~

The New Set Up

- $y_i = \alpha + \beta x_i + \varepsilon_i$
  - $\varepsilon_i = \nu_i + \eta_i$



- x_i = independent variable of interest
- η_i = error that is correlated with x_i
  - Example: innate ability

- **Main point: We no longer need panel data to "fix" the problem in our error term, but you can use one.**

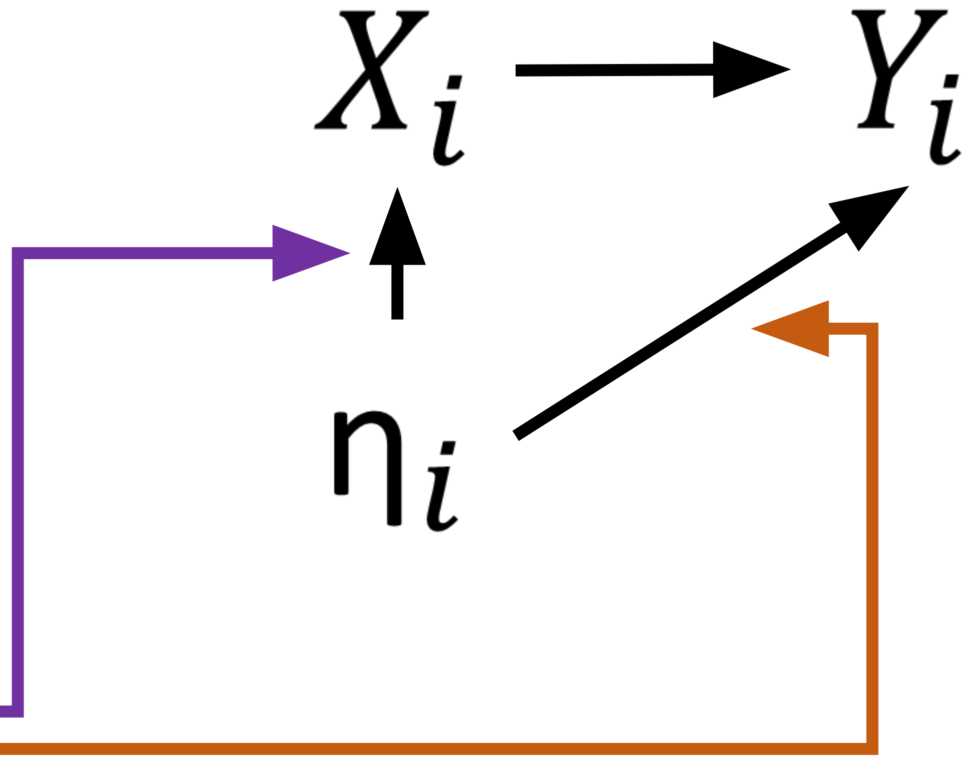
The New Solution (Instrumental Variables)

Second Stage

- $y_i = \alpha + \beta\hat{x}_i + \varepsilon_i$
  - $\varepsilon_i = \nu_i + \eta_i$

First Stage

- $x_i = \delta + \gamma IV_i + \epsilon_i$
  - IV_i = Instrumental Variable
  - $\hat{x}_i$ = Predicted X from First Stage
- **Requires:**
  - $\mathbf{cov(X, \varepsilon) \neq 0}$



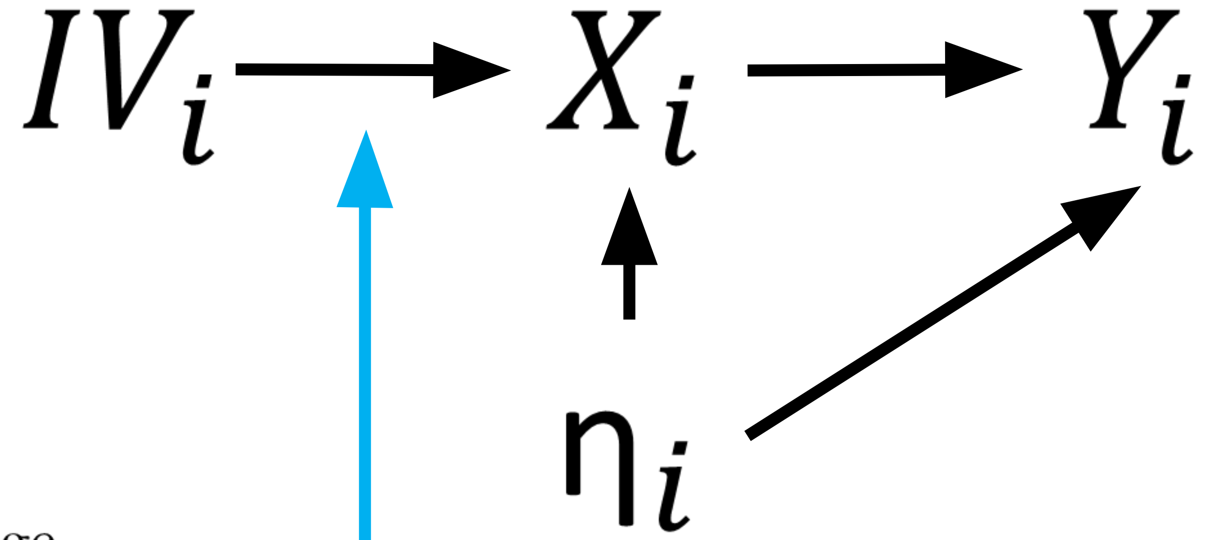
The New Solution (Instrumental Variables)

Second Stage

- $y_i = \alpha + \beta\hat{x}_i + \varepsilon_i$
  - $\varepsilon_i = \nu_i + \eta_i$

First Stage

- $x_i = \delta + \gamma IV_i + \epsilon_i$
  - IV_i = Instrumental Variable
  - $\hat{x}_i$ = Predicted X from First Stage
- **Requires:**
  - **cov(X, $\boldsymbol{\varepsilon}$) $\neq$ 0**
  - **cov(IV, X) $\neq$ 0**



The New Solution (Instrumental Variables)

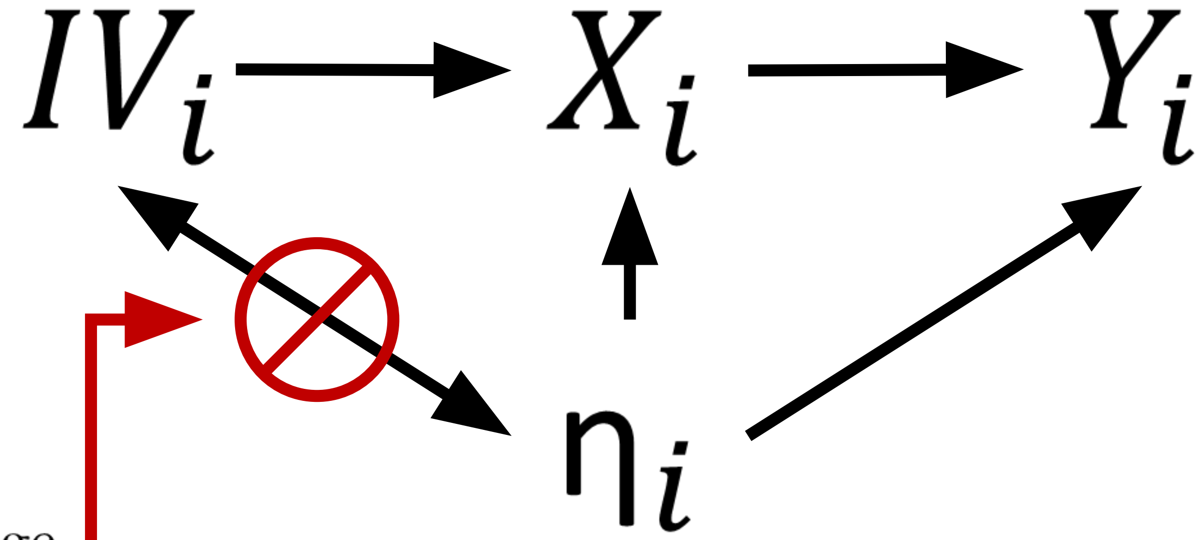
Second Stage

- $y_i = \alpha + \beta \hat{x}_i + \varepsilon_i$
  - $\varepsilon_i = \nu_i + \eta_i$

First Stage

- $x_i = \delta + \gamma IV_i + \epsilon_i$
  - IV_i = Instrumental Variable
  - $\hat{x}_i$ = Predicted X from First Stage

- **Requires:**
  - $\mathbf{cov(X, \varepsilon) \neq 0}$
  - $\mathbf{cov(IV, \varepsilon) = 0}$
  - $\mathbf{cov(IV, X) \neq 0}$



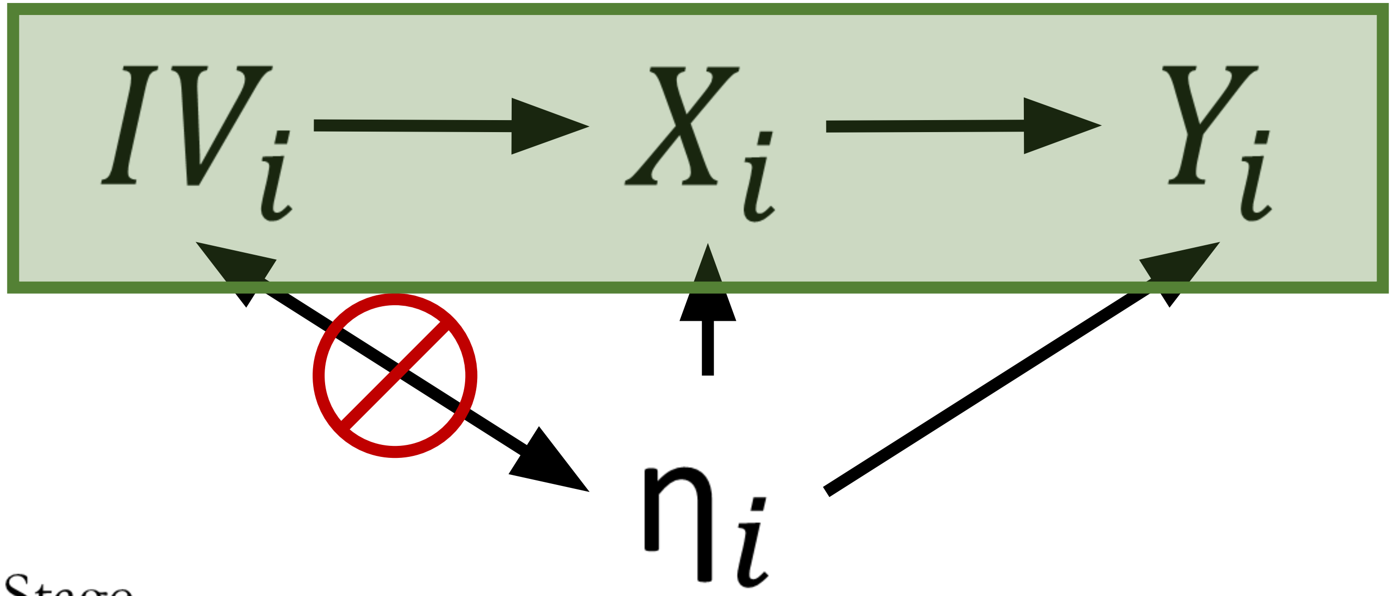
The New Solution (Instrumental Variables)

Second Stage

- $y_i = \alpha + \beta \hat{x}_i + \varepsilon_i$
  - $\varepsilon_i = \nu_i + \eta_i$

First Stage

- $x_i = \delta + \gamma IV_i + \epsilon_i$
  - IV_i = Instrumental Variable
  - $\hat{x}_i$ = Predicted X from First Stage
- **Requires:**
  - **cov(X, $\boldsymbol{\varepsilon}$) ≠ 0**
  - **cov(IV, $\boldsymbol{\varepsilon}$) = 0**
  - **cov(IV, X) ≠ 0**



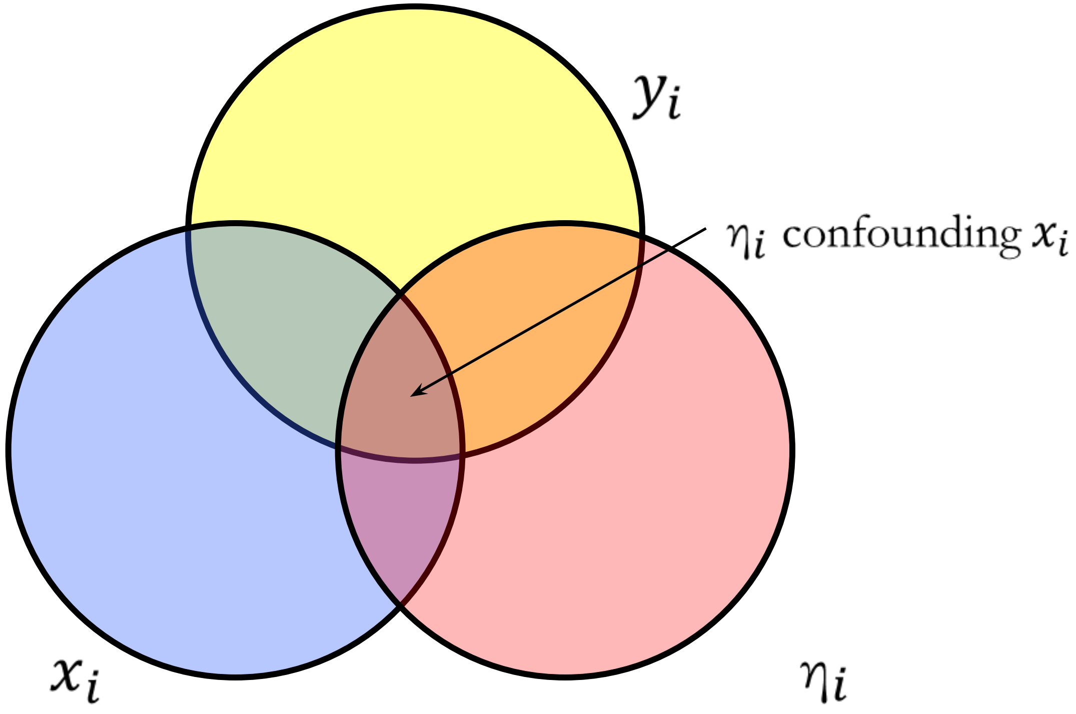
The New Solution (Instrumental Variables)

Second Stage

- $y_i = \alpha + \beta\hat{x}_i + \varepsilon_i$
  - $\varepsilon_i = v_i + \eta_i$

First Stage

- $x_i = \delta + \gamma IV_i + \epsilon_i$
  - IV_i = Instrumental Variable
  - $\hat{x}_i$ = Predicted X from First Stage



The New Solution (Instrumental Variables)

First Stage

$x_i = \delta + \gamma IV_i + \epsilon_i$

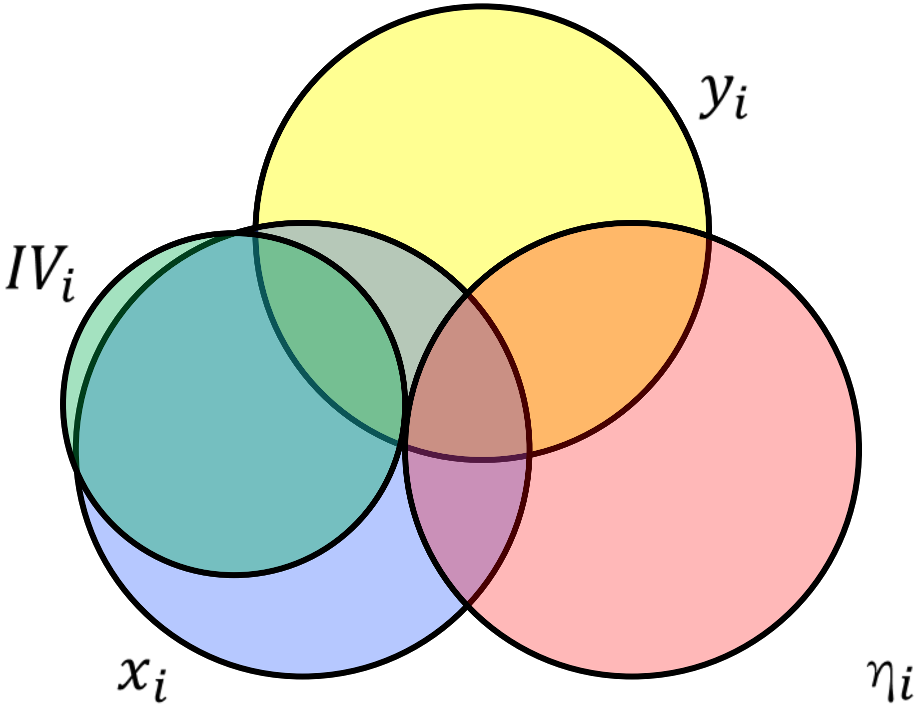
IV_i = Instrumental Variable

Requires:

$\mathbf{cov(X, \varepsilon) \neq 0}$

$\mathbf{cov(IV, \varepsilon) = 0}$

$\mathbf{cov(IV, X) \neq 0}$



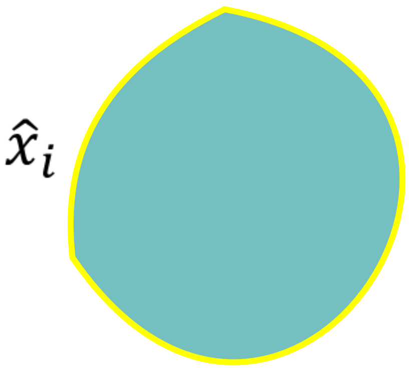
The New Solution (Instrumental Variables)

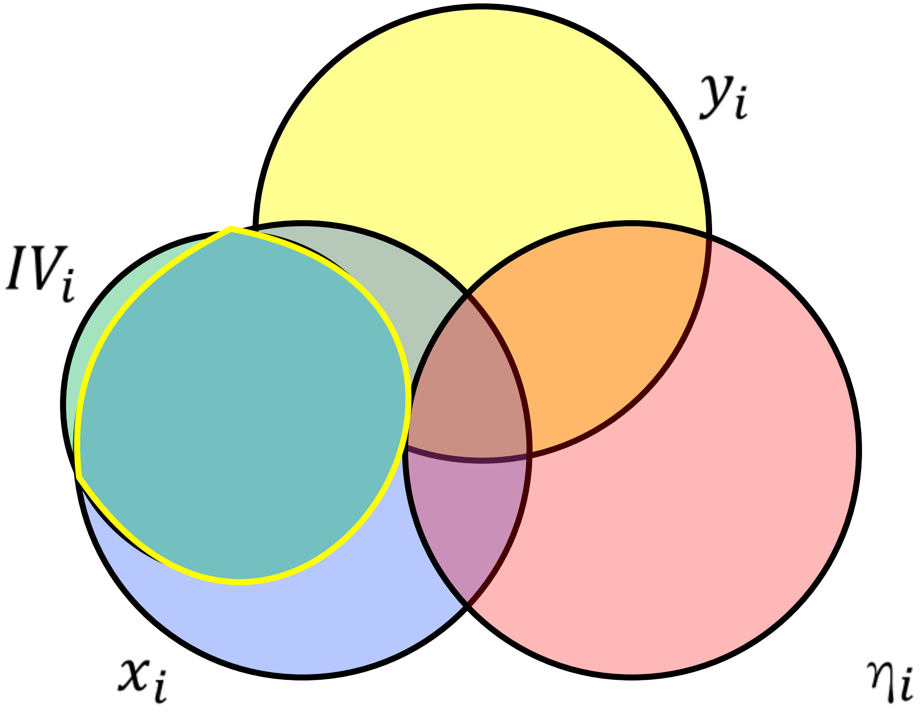
First Stage

$x_i = \delta + \gamma IV_i + \epsilon_i$

IV_i = Instrumental Variable

$\hat{x}_i$ = Predicted X from First Stage





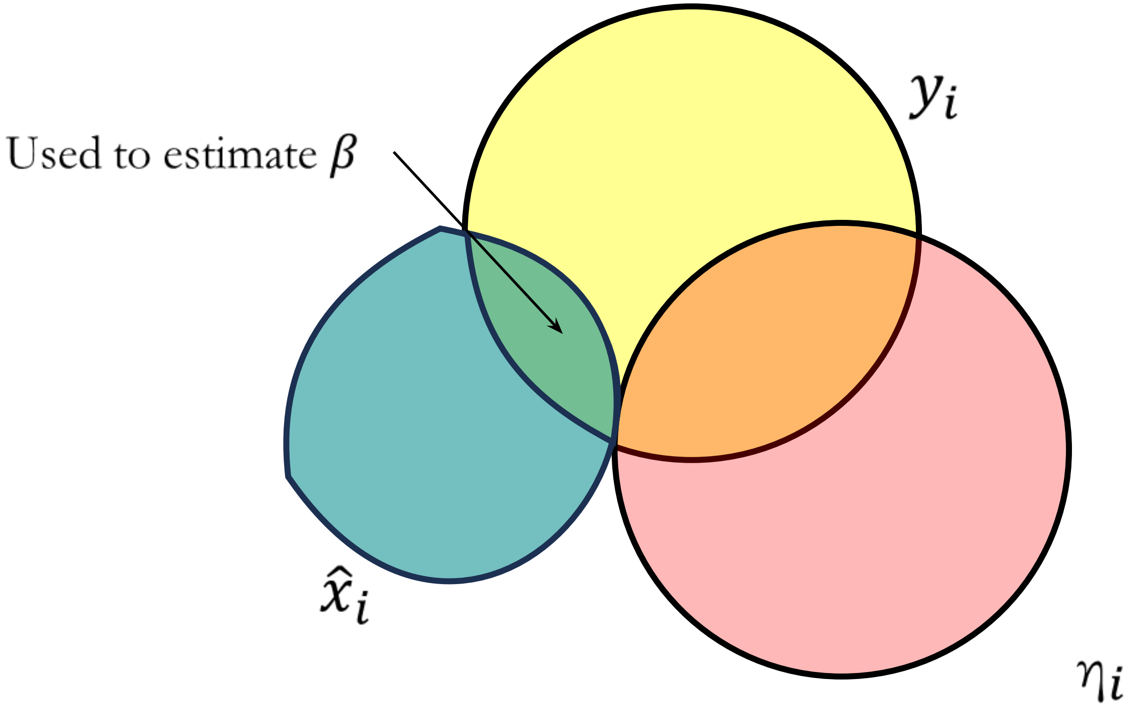
The New Solution (Instrumental Variables)

Second Stage

$$y_i = \alpha + \beta \hat{x}_i + \varepsilon_i$$

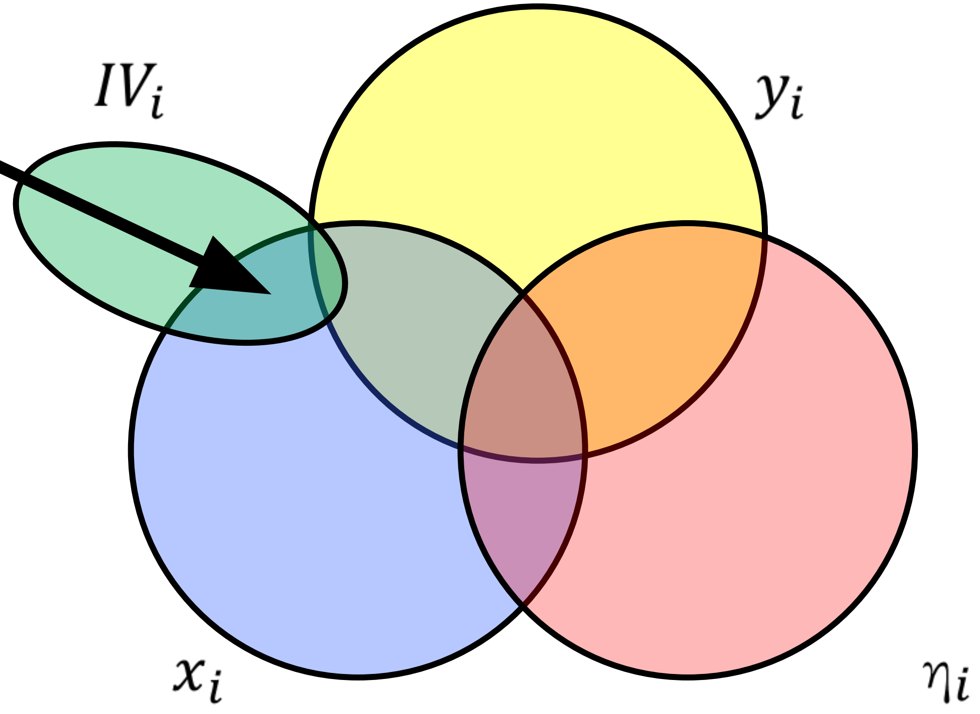
$$\varepsilon_i = \nu_i + \eta_i$$

Note that $\text{cov}(\hat{x}_i, \varepsilon) = 0$, so we solved the problem.



Can There Be Any Problems?

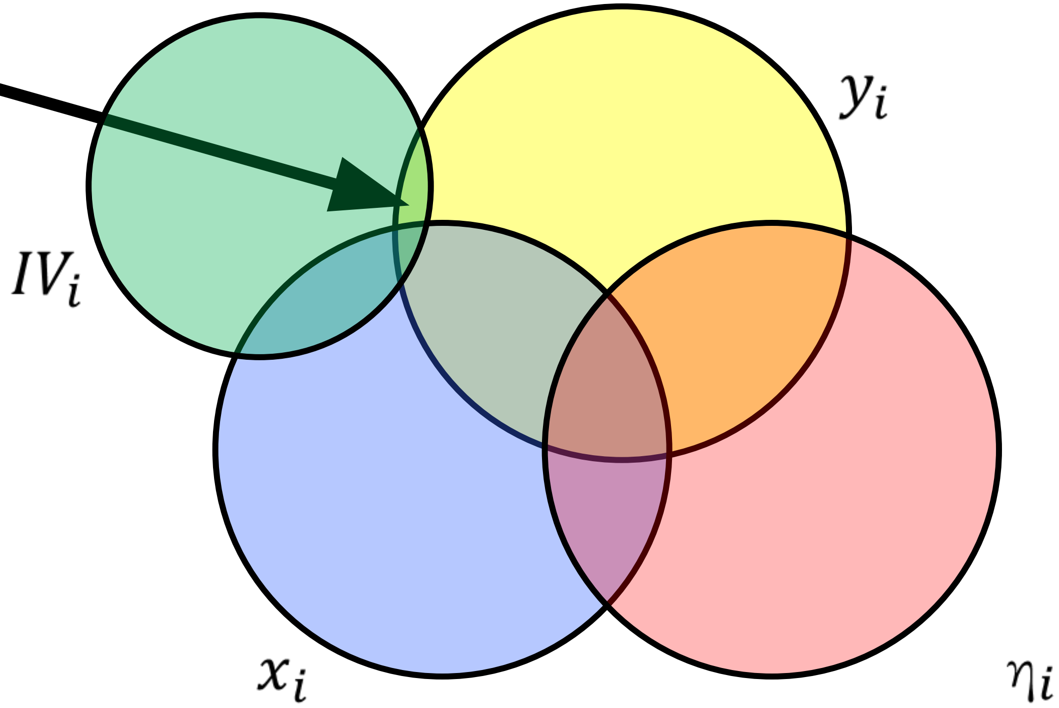
Small cov(IV, x)
Larger standard errors



Can There Be Any Problems?

$\mathbf{cov(IV, \mathcal{E}) \neq 0}$

IV variable belongs in second stage
Biased coefficient

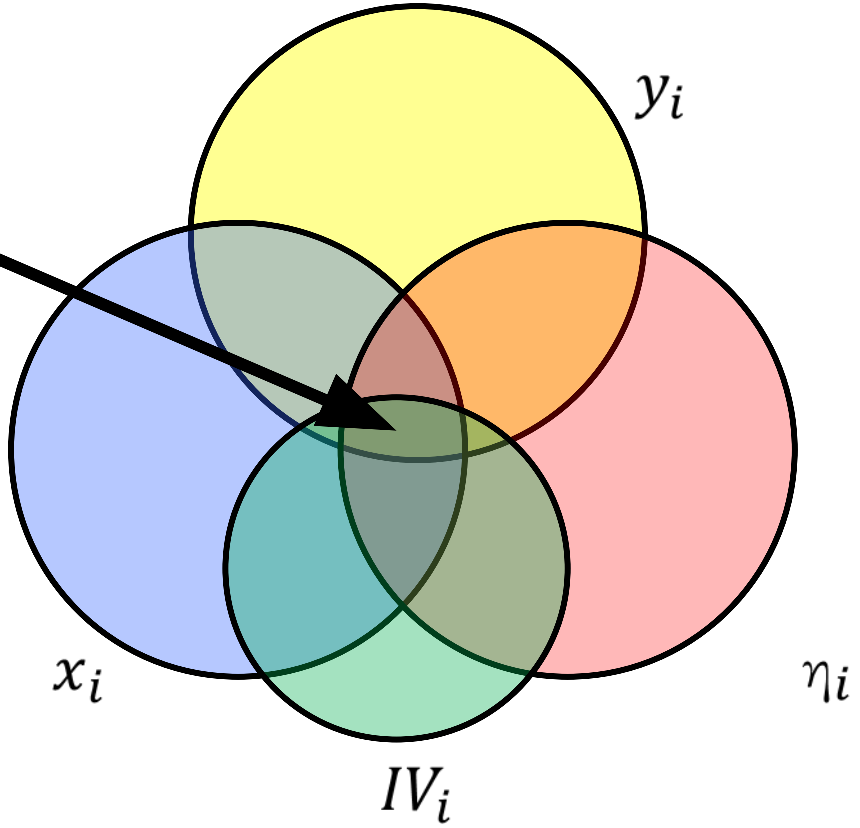


Can There Be Any Problems?

cov(IV, $\mathcal{E}$) ≠ 0

Note that η is part of $\boldsymbol{\varepsilon}$

Biased coefficient



Second Stage

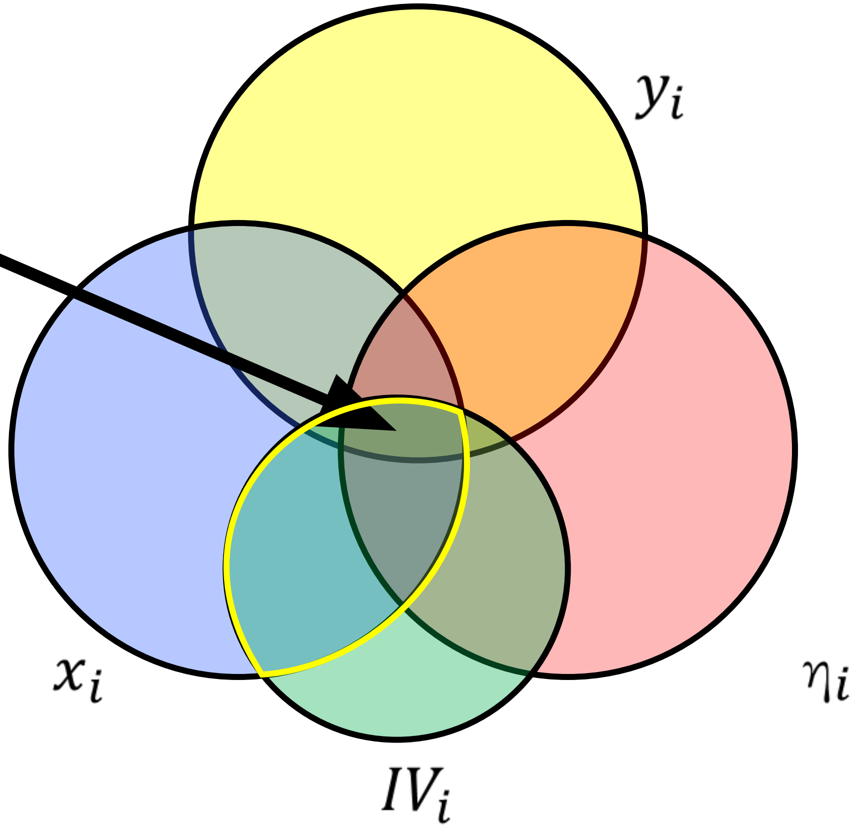
$$y_i = \alpha + \beta\hat{x}_i + \varepsilon_i$$
$$\varepsilon_i = \nu_i + \eta_i$$

Can There Be Any Problems?

cov(IV, $\mathcal{E}$) ≠ 0

Note that η is part of $\boldsymbol{\varepsilon}$

Biased coefficient

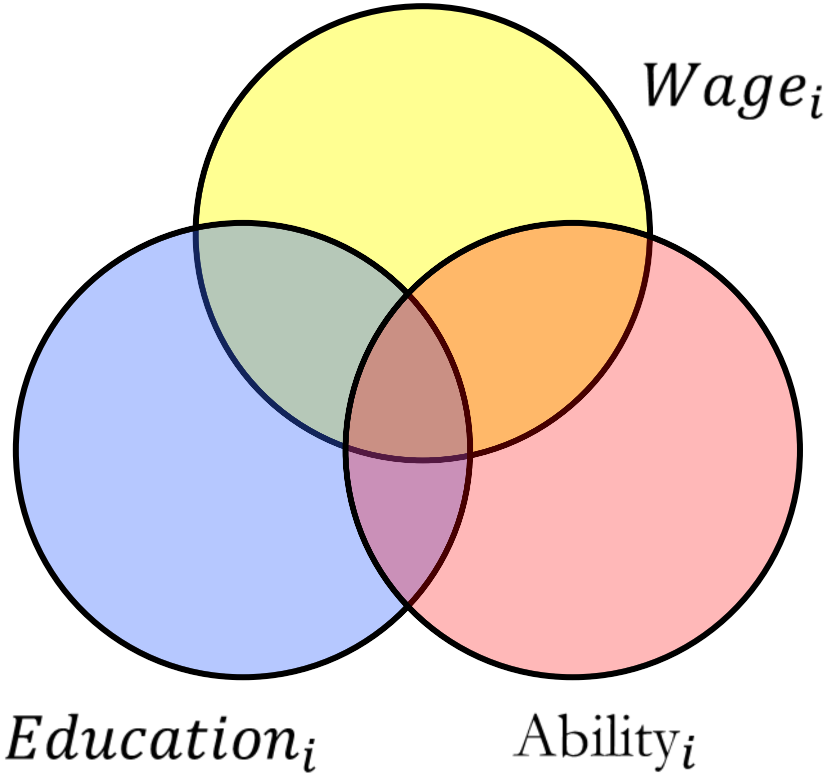


Second Stage

$$y_i = \alpha + \beta\hat{x}_i + \varepsilon_i$$
$$\varepsilon_i = \nu_i + \eta_i$$

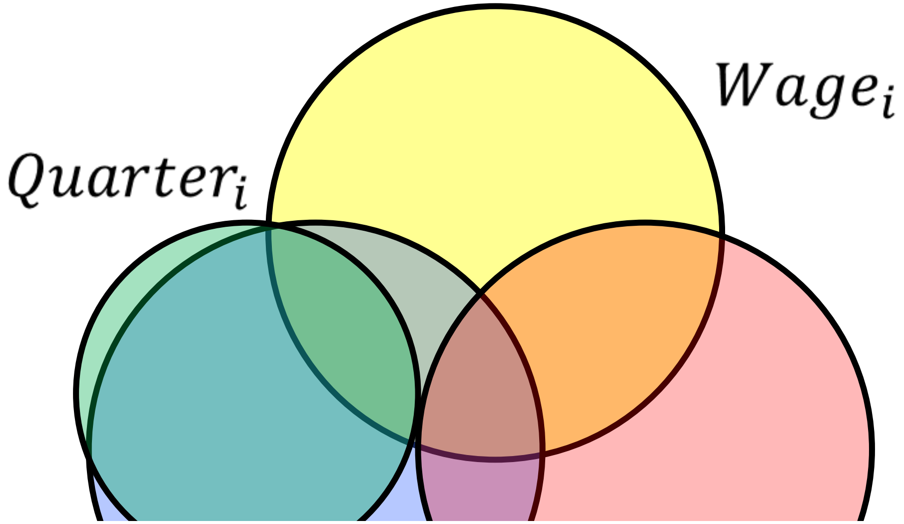
Angrist & Krueger (1991)

- Question: Returns to Education
- Method: IV to counter OVB
  - Quarter of Birth
- Data: U.S. Census
- Results: IV and OLS similar (no OVB)



IV = Quarter of Birth

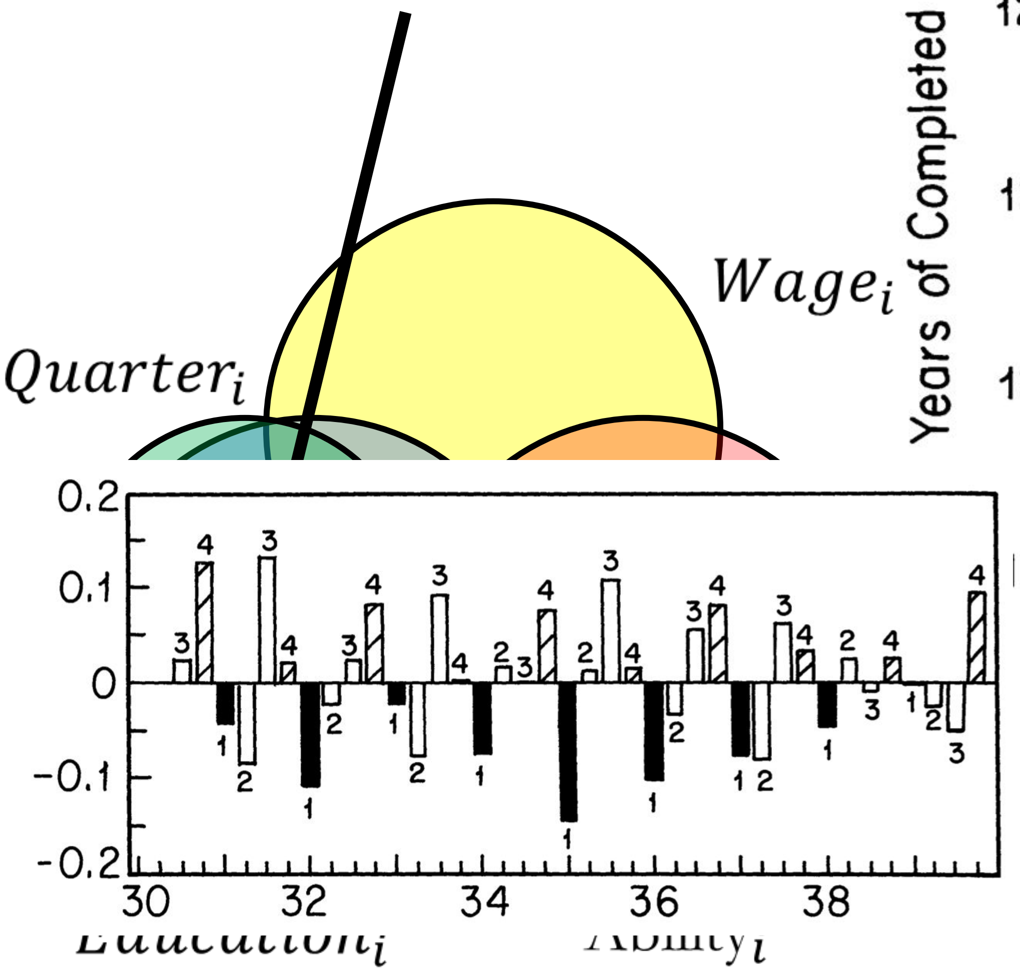
- Most states have age cutoffs to enter school
  - January 1st during this time
- States also had cutoffs to exit school
  - At this time most were age 16
- For example, if you were born at the beginning of the year (QTR 1) you would be the oldest kid in school
  - You can drop out earlier than anyone else in your grade and those born later in the year are forced to get more education.



State	1960	1970	1980	Notes
18 Indiana	16	16	16	
19 Iowa	16	16	16	
20 Kansas	16	16	16	
21 Kentucky	16	16	16	
22 Louisiana	16	16	16	
23 Maine	15	17	17	
24 Maryland	16	16	16	
25 Massachusetts	16	16	16	
26 Michigan	16	16	16	
27 Minnesota	16	16	16	
28 Mississippi	—	—	—	Age 14 starting 1983
29 Missouri	16	16	16	
30 Montana	16	16	16	
31 Nebraska	16	16	16	
32 Nevada	17	17	17	
33 New Hampshire	16	16	16	
34 New Jersey	16	16	16	
35 New Mexico	16	17	17	
36 New York	16	16	16	May be changed by city districts
37 North Carolina	16	16	16	
38 North Dakota	17	16	16	
39 Ohio	18	18	18	
40 Oklahoma	18	18	18	
41 Oregon	18	18	18	
42 Pennsylvania	17	17	17	
44 Rhode Island	16	16	16	
45 South Carolina	—	16	16	Reinstated in 1967

First Stage

$\mathbf{cov(IV}, \mathcal{X}\mathbf{)} \neq \mathbf{0}$



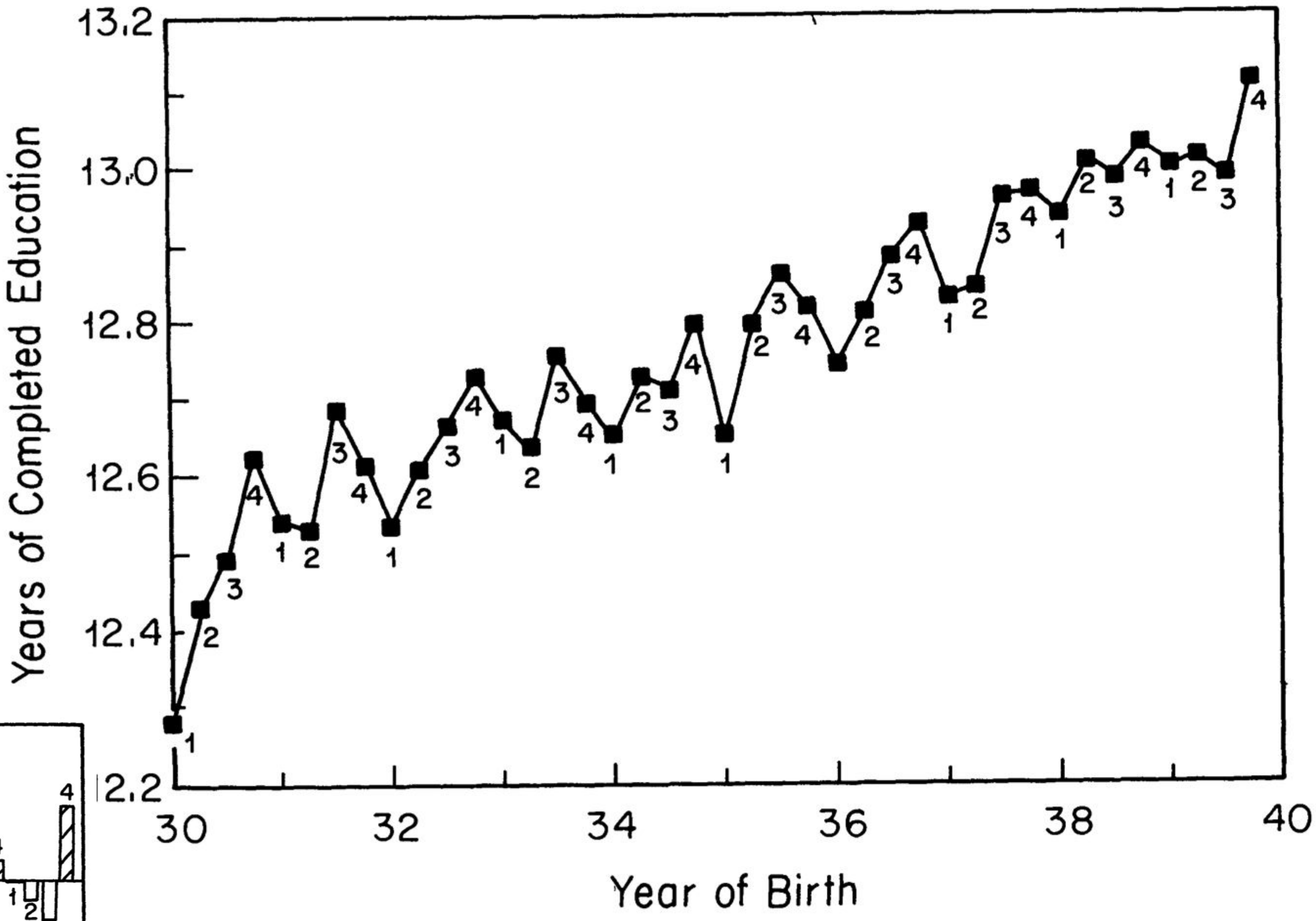


FIGURE I

Years of Education and Season of Birth

1980 Census

Note. Quarter of birth is listed below each observation.

First Stage

cov(IV, $\mathcal{X}$) ≠ 0

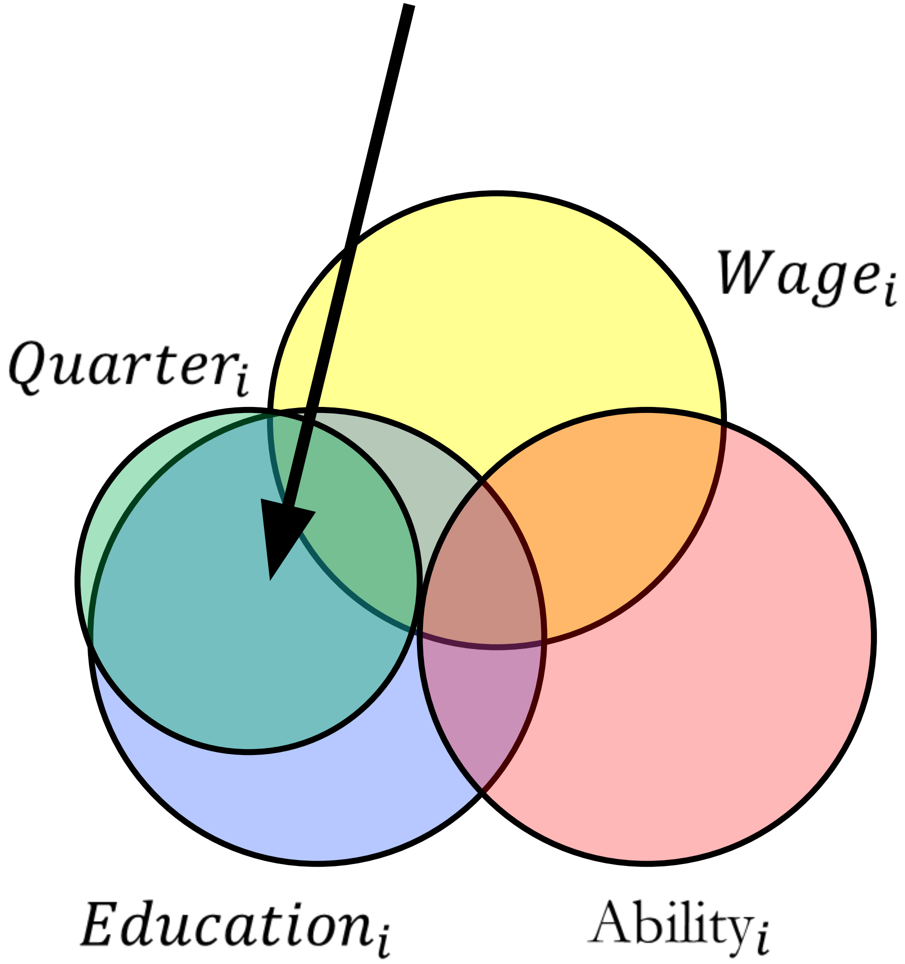
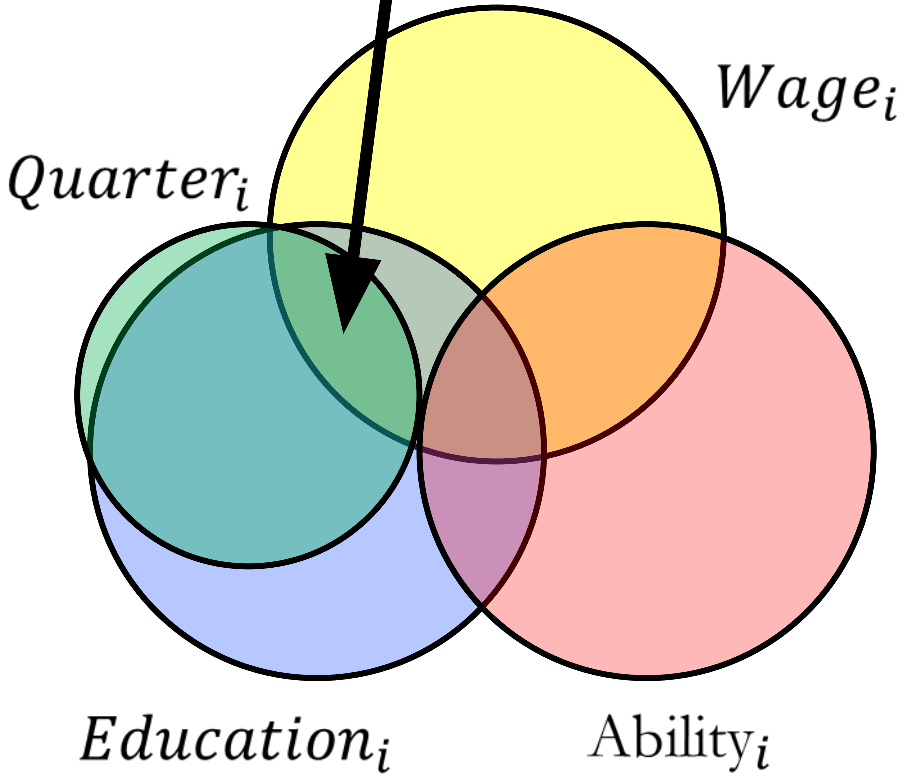


TABLE I
THE EFFECT OF QUARTER OF BIRTH ON VARIOUS EDUCATIONAL OUTCOME VARIABLES

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	−0.086	−0.015	24.9
			(0.017)	(0.017)	(0.016)	[0.0001]
	1940–1949	13.56	−0.085	−0.035	−0.017	18.6
			(0.012)	(0.012)	(0.011)	[0.0001]
High school graduate	1930–1939	0.77	−0.019	−0.020	−0.004	46.4
			(0.002)	(0.002)	(0.002)	[0.0001]
	1940–1949	0.86	−0.015	−0.012	−0.002	54.4
			(0.001)	(0.001)	(0.001)	[0.0001]
Years of educ. for high school graduates	1930–1939	13.99	−0.004	0.051	0.012	5.9
			(0.014)	(0.014)	(0.014)	[0.0006]
	1940–1949	14.28	0.005	0.043	−0.003	7.8
			(0.011)	(0.011)	(0.010)	[0.0017]
College graduate	1930–1939	0.24	−0.005	0.003	0.002	5.0
			(0.002)	(0.002)	(0.002)	[0.0021]
	1940–1949	0.30	−0.003	0.004	0.000	5.0
			(0.002)	(0.002)	(0.002)	[0.0018]
Completed master's degree	1930–1939	0.09	−0.001	0.002	−0.001	1.7
			(0.001)	(0.001)	(0.001)	[0.1599]
	1940–1949	0.11	0.000	0.004	0.001	3.9
			(0.001)	(0.001)	(0.001)	[0.0091]
Completed doctoral degree	1930–1939	0.03	0.002	0.003	0.000	2.9
			(0.001)	(0.001)	(0.001)	[0.0332]
	1940–1949	0.04	−0.002	0.001	−0.001	4.3
			(0.001)	(0.001)	(0.001)	[0.0050]

Reduced Form of 2^{nd} Stage

$\mathbf{cov(IV,}\ y\mathbf{)} \neq \mathbf{0}$



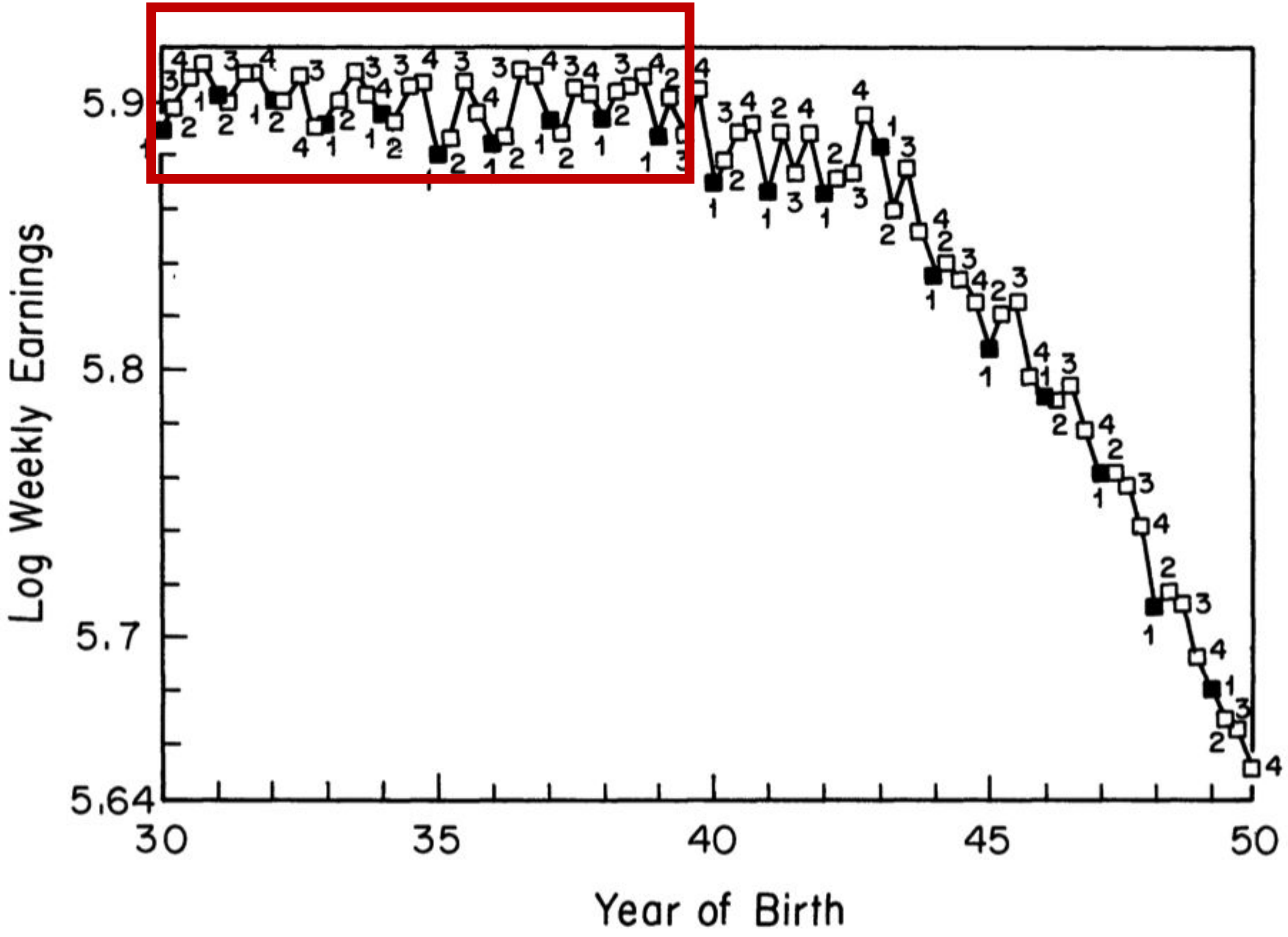


FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

Wald Estimator (Simplified IV)

- Convert Instrument (IV) into dichotomous (0/1) variable
- Calculate
  - Difference in X for IV = 1, IV = 0
  - Difference in Y for IV =1, IV = 0
- Wald $= \frac{E(Y(IV=1) - E(Y(IV=0)}{E(X(IV=1) - E(X(IV=0)}$
- $\frac{5.1484 - 5.1484}{11.3996 - 11.5252}$
- $= \frac{-0.00898}{-0.1256} = 0.0715$
- A 1 year increase in education is associate with a 7.15% increase in wages.

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)

IV Regression

- E is for Education
- X is for Controls
- Y is for Year FE
- YQ is for Year by Quarter (these are the IV)
- ln W is for Logged Weekly Wage

$$(1) \qquad E_i = X_i\pi + \Sigma_c Y_{ic}\delta_c + \Sigma_c \Sigma_j Y_{ic} Q_{ij} \theta_{jc} + \epsilon_i$$

$$(2) \qquad \ln W_i = X_i\beta + \Sigma_c Y_{ic}\xi_c + \rho E_i + \mu_i,$$

IV Results

TABLE IV
OLS AND TSLS ESTIMATES OF THE RETURN TO EDUCATION FOR MEN BORN 1920–1929: 1970 CENSUS[a]

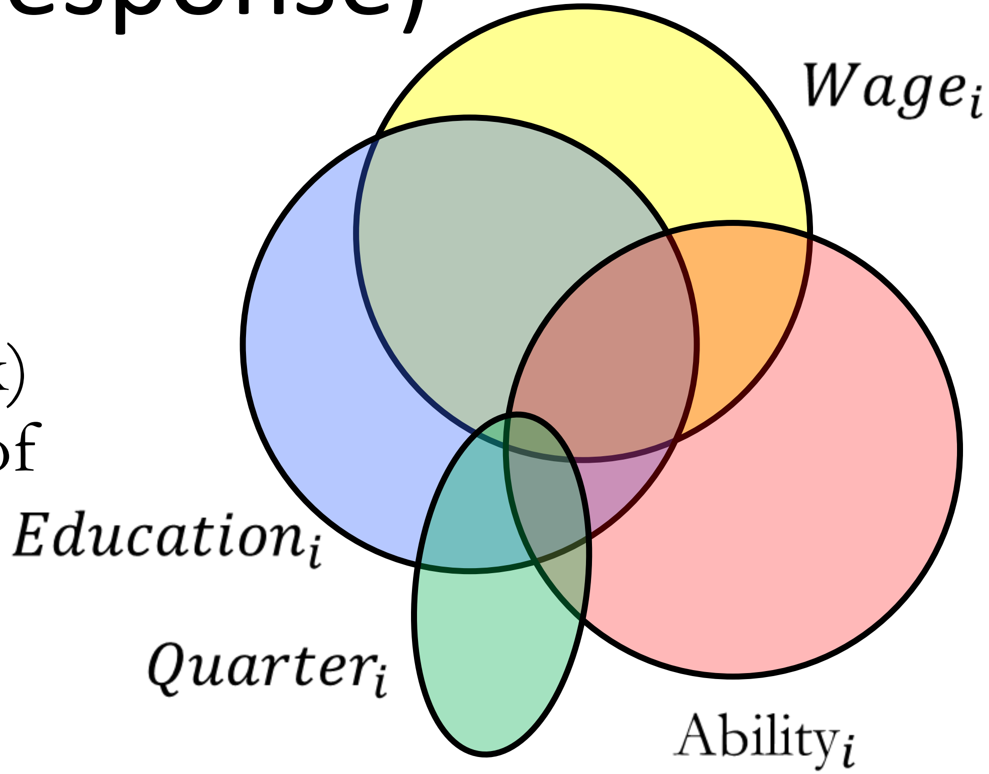
Independent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
Years of education	0.0802	0.0769	0.0802	0.1310	0.0701	0.0669	0.0701	0.1007
	(0.0004)	(0.0150)	(0.0004)	(0.0334)	(0.0004)	(0.0151)	(0.0004)	(0.0334)
Race (1 = black)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
SMSA (1 = center city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
Married (1 = married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
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
Age	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
Age-squared	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
χ^2 [dof]	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

a. Standard errors are in parentheses. Sample size is 247,199. Instruments are a full set of quarter-of-birth times year-of-birth interactions. The sample consists of males born in the United States. The sample is drawn from the State, County, and Neighborhoods 1 percent samples of the 1970 Census (15 percent form). The dependent variable is the log of weekly earnings. Age and age-squared are measured in quarters of years. Each equation also includes an intercept.

Conclusions

- Do not see that OLS is biased upward (if anything downward)
- Do not find effects for Annual Salary or Weeks Worked
- NOTE that all IV provides a Local Average Treatment Effect (LATE)
  - In this case all variation in X (education) is coming from kids on the margin of dropping out of high school

Bound, Jaeger, and Baker (the response)



- Question: Are there issues with instrumental variables analysis?
- Weak Instruments combined with any (even weak) relationship between Instrument and Y (outside of X) can cause bias
- In finite samples IV is biased toward OLS
  - Magnitude of bias is related to R^2 between IV and X

Back to the Math (OLS)

- *Assume just a single X and no controls for simplicity*
- $Y = \alpha + \beta X + \varepsilon$
- $Cov(Y,X) = Cov(\alpha + \beta X + \varepsilon, X)$
- $Cov(Y,X) = \beta Cov(X,X) + Cov(\varepsilon, X)$
  - Note: $Cov(\alpha, X) = 0$ because α does not vary
- $\sigma_{YX} = \beta\sigma^2_X + \sigma_{\varepsilon X}$
- $\sigma_{YX}/\sigma^2_X = \beta + \sigma_{\varepsilon X}/\sigma^2_X$
- $\beta_{OLS} = \beta + \sigma_{\varepsilon X}/\sigma^2_X$
- Bias if $\sigma_{\varepsilon X}$ does not = 0 (this could be due to Omitted Variable Bias)

$$\text{plim } \hat{\beta}_{\text{ols}} = \beta + \frac{\sigma_{x,\varepsilon}}{\sigma_x^2}$$

Convergence in probability

The basic idea behind this type of convergence is that the probability of an “unusual” outcome becomes smaller and smaller as the sequence progresses.

The concept of convergence in probability is used very often in statistics. For example, an estimator is called consistent if it converges in probability to the quantity being estimated. Convergence in probability is also the type of convergence established by the weak law of large numbers.

Same Math, but do IV

- $Y = \alpha + \beta X + \varepsilon$
- $Cov(Y,IV) = Cov(\alpha + \beta X + \varepsilon, IV)$
- $Cov(Y,IV) = \beta Cov(X,IV) + Cov(\varepsilon,IV)$
- $\sigma_{YIV} = \beta\sigma_{XIV} + \sigma_{\varepsilon IV}$
- $\sigma_{YIV}/\sigma_{XIV} = \beta + \sigma_{\varepsilon IV}/\sigma_{XIV}$
- $\beta_{iv} = \beta + \sigma_{\varepsilon IV}/\sigma_{XIV}$
- In Instrumental Variables we assume that $\sigma_{\varepsilon IV} = 0$

$$\text{plim } \hat{\beta}_{iv} = \beta + \frac{\sigma_{\hat{x},\varepsilon}}{\sigma_x^2}$$

Note that $\hat{x}_i = Cov(X, IV)$

Conclusion

- The level of inconsistency is related directly to the fit of the First Stage ($R^2_{X,IV}$)
- In Angrist & Krueger's paper the R^2 is only 0.0001 to 0.0002
- In practice:
- The R^2 from your First Stage (only including your instruments) is a measure of the instruments' strength (bigger is better)
- Test strength directly using an F-test of all instruments from first stage (**F > 10** is industry standard, but bigger is better – as long as cov(IV, ε) = 0

$$\frac{\text{plim } \hat{\beta}_{\text{iv}} - \beta}{\text{plim } \hat{\beta}_{\text{ols}} - \beta} = \frac{\sigma_{\hat{x},\varepsilon} / \sigma_{x,\varepsilon}}{R^2_{x,\,\mathbf{IV}}}$$

Note that in Bound, Jaeger, Baker's paper they referred to the IV variable as Z. This is common in the literature.

What were the F and R^2 in Angrist & Krueger?

Table 1. *Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings (standard errors of coefficients in parentheses)*

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) OLS	(6) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)	.063 (.000)	.060 (.029)
F (excluded instruments)		13.486		4.747		1.613
Partial R^2 (excluded instruments, ×100)		.012		.043		.014
F (overidentification)		.932		.775		.725
Age Control Variables						
Age, Age^2	x	x			x	x
9 Year of birth dummies			x	x	x	x
Excluded Instruments						
Quarter of birth		x		x		x
Quarter of birth × year of birth				x		x
Number of excluded instruments		3		30		28

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. *F* (first stage) and partial R^2 are for the instruments in the first stage of IV estimation. *F* (overidentification) is that suggested by Basmann (1960).

Reexamine Angrist & Krueger

- Paper hinges on the law being the only reason for the differences in education
- Problem is that the relationship is weak, so any other reason can be problematic
    - Attendance Rates
    - Behavioral Difficulties
    - Mental Health
    - Reading, Writing, Math
    - Schizophrenia
    - Mental Retardation
    - Dyslexia
    - Multiple Sclerosis
    - Manic Depression
    - Some show IQ
    - Region of Birth
    - Parental Income

Let's Focus on Just Parental Income

- Solon (1992): children's education (X) increases 0.014 for each 1% increase in father's income
  - Difference in Family Income between QTR 1 and QTR 2 can explain 0.03 education years (this is 1/3 of the IV First Stage, so not really a weak relationship)
- Intergenerational correlation of long-run income > 0.4
  - Difference in Family Income between QTR 1 and QTR 2 can explain 0.95% lower wages for children
  - A&K: 1980 QTR 1 earn 1.1% less using IV
- Most of the "effect" can be explained by parental income

Some Thoughts

- IV is asymptotically unbiased, but that does not mean unbiased
  - You need a large sample for IV to work correctly
- IV is an art rather than a science
  - You need to convince the reader that your instrument does not belong in the Second Stage (is not correlated with error – Exclusion Restriction)
  - You will need to appeal to theory, rather than directly test this

- Next time
  - More tests – particularly when you have > 1 instrument
  - More Stata – ivregress