Econometrics

Lecture 10

Nathaniel Higgins

JHU

New topic

- Two lectures left (including today)
- One of the last topics I want to cover with you is instrumental variables
- I want to cover IV because it is such an important topic in applied econometrics (not the course, *applied econometrics*, but econometrics *as it is applied* professionally)
- IV is a technique that gives coefficients causal interpretation
- IV is a "solution" to the problem of endogeneity

Instrumental variables

"The method of instrumental variables is a signature technique in the econometrics toolkit." – Angrist and Krueger (2001)

Classic endogeneity problem

- An example of an endogenous regressor

$$earn = \beta_0 + \beta_1 educ + u$$

- OK, what's the problem?

Classic endogeneity problem

- An example of an endogenous regressor

$$earn = \beta_0 + \beta_1 educ + u$$

- OK, what's the problem?
- An omitted variable (which is thus part of *u*) that is correlated with *educ*

Classic endogeneity problem

- An example of an endogenous regressor

$$earn = \beta_0 + \beta_1 educ + u$$

- OK, what's the problem?
- An omitted variable (which is thus part of *u*) that is correlated with *educ*
- *ability*

Classic endogeneity problem

- An example of an endogenous regressor

$$earn = \beta_0 + \beta_1 educ + u$$

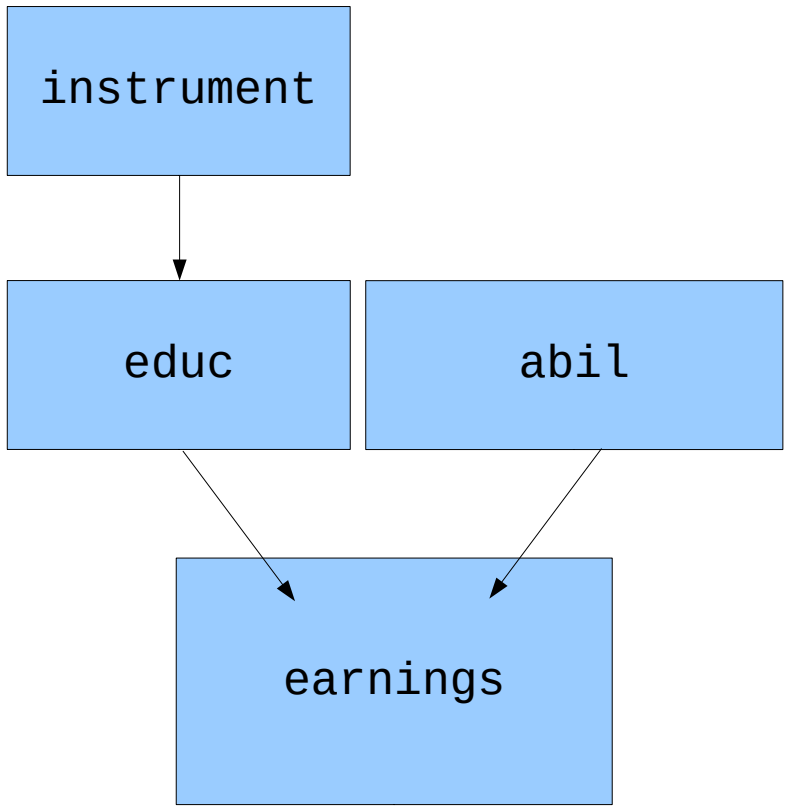
- OK, what's the problem?
- An omitted variable (which is thus part of *u*) that is correlated with *educ*
- *ability*
- A big problem in education research (and labor economics in general) is unobserved differences in innate ability
- It turns out that ability is an important determinant, but it is very difficult to observe/measure

Classic endogeneity problem

- OK, so what?
- If ability is correlated with both *educ* and *earn*, then *educ* is endogenous
- If *educ* is endogenous then the coefficient on *educ* is biased
- We would like to interpret the coefficient on *educ* as the effect of *educ* on *earn*
- In order to do this, we need *an instrument*
- An instrument is a variable that we can use to get rid of the endogeneity problem

What is an instrument?

- An instrument is a thing that is correlated with *educ*, but not correlated with *ability*
- Another way to say this (a more precise way): an instrument is a thing that is correlated with *educ*, but uncorrelated with *earn*, *except* through its effect on *educ*



Instruments — conceptual model

- The instrument moves *educ* and *educ* moves *earn*
- An instrument affects the dependent variable *only* through the variable we care about
- Simple, right?
- Not really . . .

Instruments — mathematical model

- Let's look at the model again

$$earn = \beta_0 + \beta_1 educ + u \quad (1)$$

$$educ = \alpha_0 + \alpha_1 instrument + v \quad (2)$$

- An instrument is (same thing said 3 different ways):
  - a variable that affects *earn*, but only *through educ*
  - a variable that moves *educ*, but has NO direct effect on *earn*
  - a variable that belongs in equation (2) but NOT in equation (1)
- What the heck kind of variable belongs in (2) but not in (1)? Can you think of one? It's not easy.

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*
  3. The effects of *educ* and *ability* are both assigned to *educ*

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*
  3. The effects of *educ* and *ability* are both assigned to *educ*
- Why an instrument solves the problem:

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*
  3. The effects of *educ* and *ability* are both assigned to *educ*
- Why an instrument solves the problem:
  1. When the instrument moves, *ability* . . .

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*
  3. The effects of *educ* and *ability* are both assigned to *educ*
- Why an instrument solves the problem:
  1. When the instrument moves, *ability* . . . *does not move*

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*
  3. The effects of *educ* and *ability* are both assigned to *educ*
- Why an instrument solves the problem:
  1. When the instrument moves, *ability* . . . *does not move*
  2. When the instrument moves, *educ* moves

Reminder: why endogeneity is bad

- Why endogeneity is a problem in the original regression:
  1. When *educ* moves, *ability* also moves (they *co-move*)
  2. When *educ* and *ability* move, they *both* affect *earn*
  3. The effects of *educ* and *ability* are both assigned to *educ*
- Why an instrument solves the problem:
  1. When the instrument moves, *ability* . . . *does not move*
  2. When the instrument moves, *educ* moves
  3. When *educ* moves, *earn* moves, and we identify the effect

If you've found a good instrument: First stage

- Think of the relationship between a good instrument and the endogenous variable as a "first stage" regression
- We run the first stage regression

$$educ = \alpha_0 + \alpha_1 instrument + v$$

  we get $\widehat{\alpha_0}$ and $\widehat{\alpha_1}$
- We use these coefficients to predict education

$$\widehat{educ} = \widehat{\alpha_0} + \widehat{\alpha_1} instrument$$

- Let's talk about $\widehat{educ}$

If you've found a good instrument: First stage

- $\widehat{educ}$ is eduction predicted by the instrument
- What do we know about the variation in education that is predicted by the instrument?

If you've found a good instrument: First stage

- $\widehat{educ}$ is eduction predicted by the instrument
- What do we know about the variation in education that is predicted by the instrument?
- The variation in education that is predicted by the instrument is *un*correlated with ability

If you've found a good instrument: First stage

- $\widehat{educ}$ is eduction predicted by the instrument
- What do we know about the variation in education that is predicted by the instrument?
- The variation in education that is predicted by the instrument is *un*correlated with ability
- We could call this the "clean" variation in *educ* — maybe we should rename the predictions to celebrate. We'll call the variable *cleanEduc*: $\widehat{educ} = cleanEduc$.

Second stage

- This new variable *cleanEduc* is then substituted for *educ* in the equation we care about (the "second stage")

$$earn = \beta_0 + \beta_1 cleanEduc + u$$

- We estimate this equation with least squares and obtain $\widehat{\beta_0}$ and $\widehat{\beta_1}$
- What is the interpretation of β_1 (and thus of our estimate $\widehat{\beta_1}$)?

Second stage

- This new variable *cleanEduc* is then substituted for *educ* in the equation we care about (the "second stage")

$$earn = \beta_0 + \beta_1 cleanEduc + u$$

- We estimate this equation with least squares and obtain $\widehat{\beta_0}$ and $\widehat{\beta_1}$
- What is the interpretation of β_1 (and thus of our estimate $\widehat{\beta_1}$)?
- It's exactly what we *wanted it to be* in the original regression
- β_1 is the effect of education on earnings
- The variation that we isolated by using an instrument is *still* variation in education

Examples

- It is very difficult to get a sense of how instrumental variables work without thinking through examples
- Instrumental variables is not a technique (like probit/logit/OLS/Maximum likelihood) that you simply impose on data
- Instrumental variables is a technique that involves creativity, intuition, and institutional knowledge
- Let's go through an example

Example of the earnings equation

- Stop thinking of instruments for just a second
- Suppose you were faced with the problem of estimating the earnings equation we've been discussing today

$$earn = \beta_0 + \beta_1 educ + u$$

- How (besides IV) could you estimate the effect of education on earnings?
- HINT: think of what you could do if you were all-powerful. What is the ideal experiment to determine the effect of education on earnings?

Ideal experiment

- Begin with the population of U.S. kindergarteners

Ideal experiment

- Begin with the population of U.S. kindergarteners
- Randomly assign every kindergartener to 1 of 17 different groups

Ideal experiment

- Begin with the population of U.S. kindergarteners
- Randomly assign every kindergartener to 1 of 17 different groups
  1. Group 1 gets kicked-out of kindergarten, and allowed in school nevermore

Ideal experiment

- Begin with the population of U.S. kindergarteners
- Randomly assign every kindergartener to 1 of 17 different groups
  1. Group 1 gets kicked-out of kindergarten, and allowed in school nevermore
  2. Group 2 gets to stay in kindergarten, but leave school after kindergarten

Ideal experiment

- Begin with the population of U.S. kindergarteners
- Randomly assign every kindergartener to 1 of 17 different groups
  1. Group 1 gets kicked-out of kindergarten, and allowed in school nevermore
  2. Group 2 gets to stay in kindergarten, but leave school after kindergarten
  3. Group 3 stays in kindergarten, then goes to first grade, but has to leave school after first grade

Ideal experiment

- Begin with the population of U.S. kindergarteners
- Randomly assign every kindergartener to 1 of 17 different groups
  1. Group 1 gets kicked-out of kindergarten, and allowed in school nevermore
  2. Group 2 gets to stay in kindergarten, but leave school after kindergarten
  3. Group 3 stays in kindergarten, then goes to first grade, but has to leave school after first grade
  4. . . . and so on. . .

Ideal experiment

- Begin with the population of U.S. kindergarteners
- Randomly assign every kindergartener to 1 of 17 different groups
  1. Group 1 gets kicked-out of kindergarten, and allowed in school nevermore
  2. Group 2 gets to stay in kindergarten, but leave school after kindergarten
  3. Group 3 stays in kindergarten, then goes to first grade, but has to leave school after first grade
  4. . . . and so on. . .

  17. Every member of group 17 must to to school for the next 16 years

Ideal experiment

- That'd work great!
- We would know that the variation in education (the number of years of school attended) was not correlated with ability
- Why?

Ideal experiment

- That'd work great!
- We would know that the variation in education (the number of years of school attended) was not correlated with ability
- Why?
- Because we randomly assigned students to a level of education
- By doing this random assignment, we break any natural correlation between education and ability
- Of course, this is impossible with a capital "I"
- The goal of choosing an instrument is to find a variable that acts as much like a random assignment tool as possible

Example of the earnings equation

- Angrist, Joshua D., and Alan B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4): 979-1014.
- They use quarter of birth (Q_{ij}) as an instrument for education
- Why might this work?

Example of the earnings equation

- Angrist, Joshua D., and Alan B. Krueger (1991). "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106(4): 979-1014.
- They use quarter of birth (Q_{ij}) as an instrument for education
- Why might this work?
- States have compulsory schooling laws
- Presumably, the quarter of birth has nothing to do with ability
- Think about how compulsory schooling laws work (next slide)

Example of the earnings equation

- Students typically start school in the year that they turn 5
- Therefore, starting date is a function of birth date
- Students who are born in January turn 5 in January, meaning that they are $\approx$ 5 3/4 when they start school (call these students Q1 students)
- Students who are born in December turn 5 in December, meaning that they are $\approx$ 4 3/4 when they start school (call these students Q4 students)
- Students are then allowed to drop out of school when they turn a certain age
- Suppose that certain age was 16 for all schools
- When a Q1 student turns 16 they have received 1 year less of education than a Q4 student

Angrist and Krueger 1991

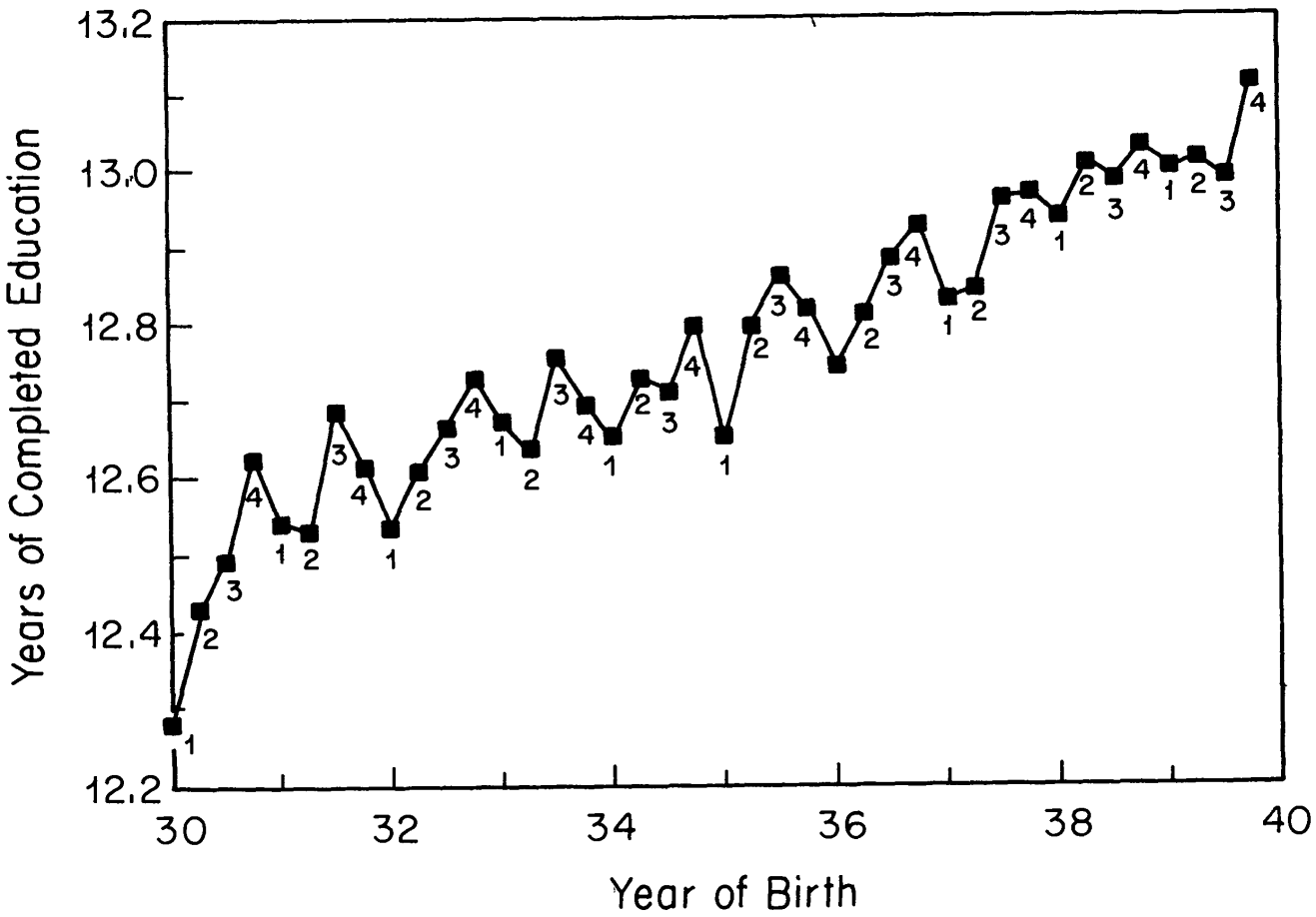


FIGURE I
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Angrist and Krueger 1991

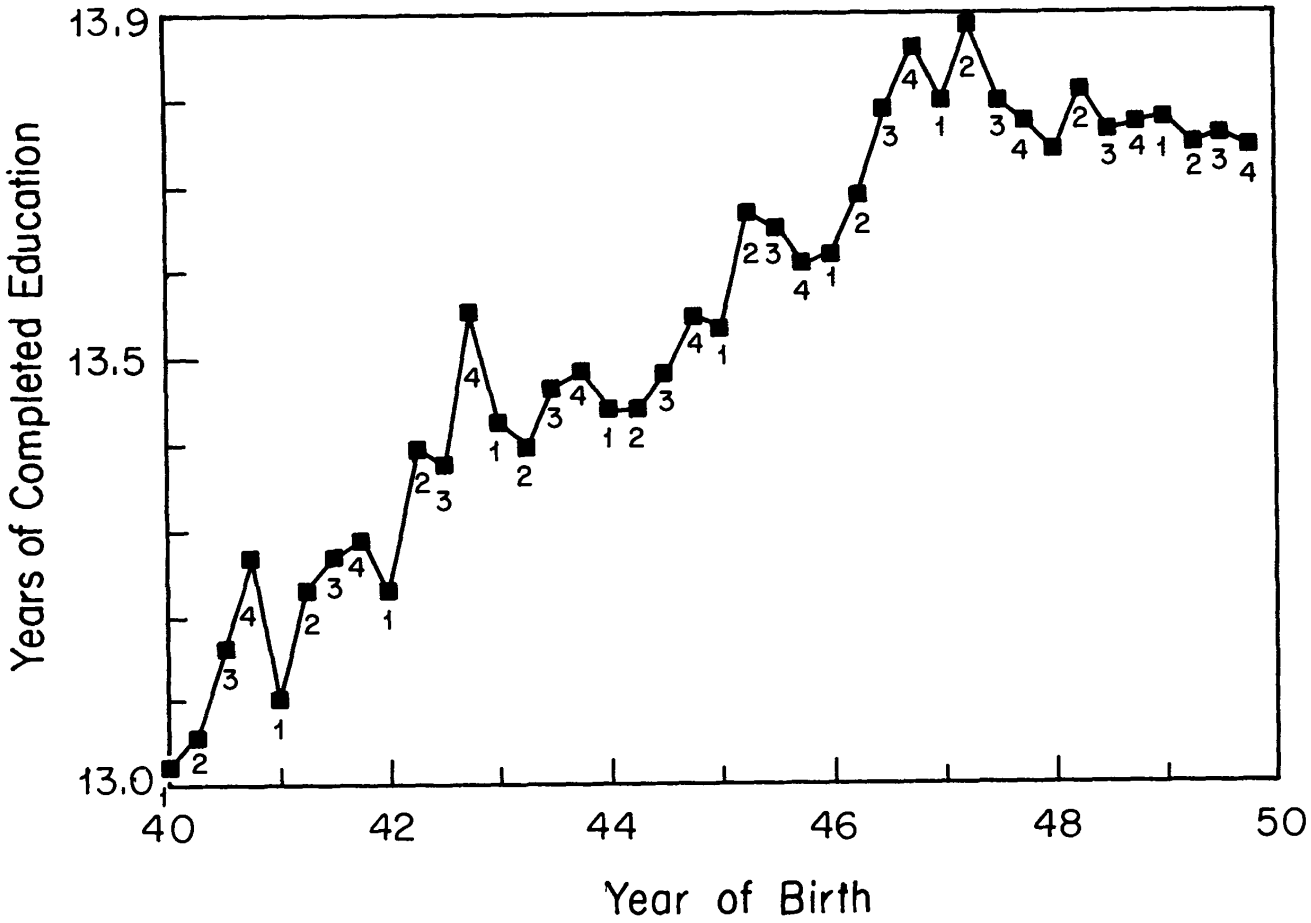


FIGURE II
Years of Education and Season of Birth
1980 Census
Note. Quarter of birth is listed below each observation.

Angrist and Krueger 1991

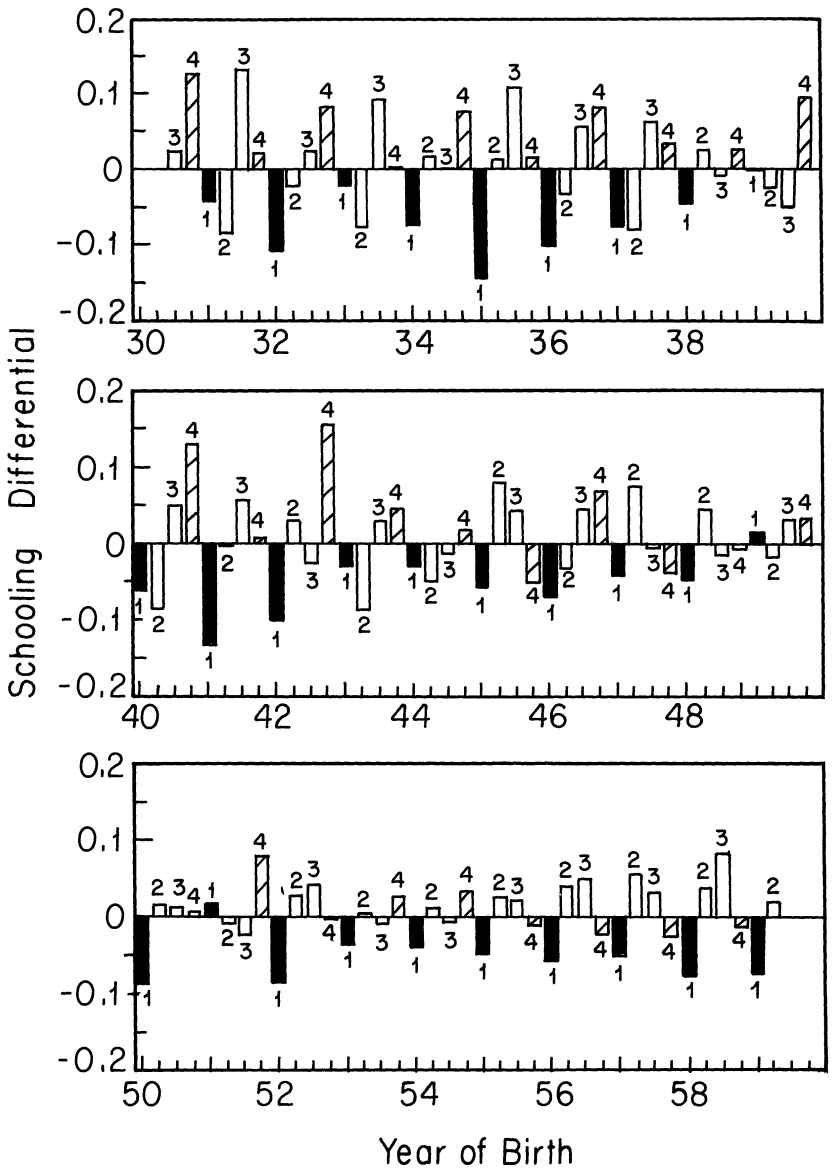


FIGURE IV
Season of Birth and Years of Schooling

Angrist and Krueger 1991

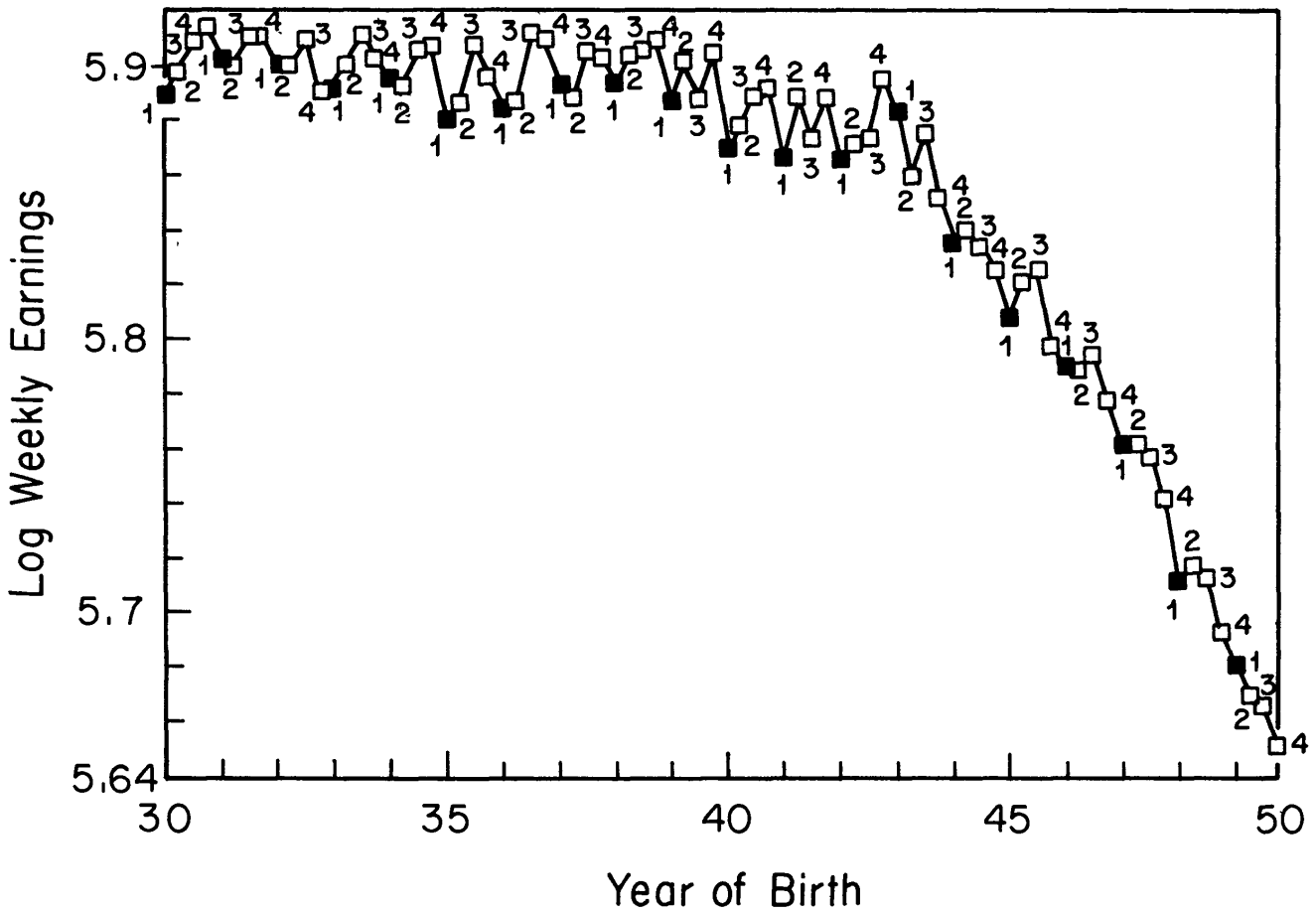


FIGURE V
Mean Log Weekly Wage, by Quarter of Birth
All Men Born 1930–1949; 1980 Census

Example of the earnings equation

- If quarter of birth is uncorrelated with ability (and we suppose that it is), then we have an instrument
- This instrument predicts education
- What might make this instrument *invalid*? (i.e. what might undermine the notion that quarter of birth is uncorrelated with ability?)

Example of the earnings equation

- If quarter of birth is uncorrelated with ability (and we suppose that it is), then we have an instrument
- This instrument predicts education
- What might make this instrument *invalid*? (i.e. what might undermine the notion that quarter of birth is uncorrelated with ability?)
- One example: if students move as a result of compulsory schooling laws
  - if low-ability students move to avoid schooling laws, the laws lose their impact
  - the relationship between quarter of birth and education breaks down (but this isn't very plausible)

Example of the earnings equation

- If quarter of birth is uncorrelated with ability (and we suppose that it is), then we have an instrument
- This instrument predicts education
- What might make this instrument *invalid*? (i.e. what might undermine the notion that quarter of birth is uncorrelated with ability?)
- One example: if students move as a result of compulsory schooling laws
  - if low-ability students move to avoid schooling laws, the laws lose their impact
  - the relationship between quarter of birth and education breaks down (but this isn't very plausible)
- IV Takeaway #1: the *validity* of an instrument is often a matter that can be argued — judgment is typically the ultimate arbiter (rather than a statistical test)

Validity

- It takes careful argument to defend the validity of an instrument
- Why no statistical test?

Validity

- It takes careful argument to defend the validity of an instrument
- Why no statistical test? Because we typically don't have the data! (we don't have *ability*)

Validity

- It takes careful argument to defend the validity of an instrument
- Why no statistical test? Because we typically don't have the data! (we don't have *ability*)
- The argument for validity is based on some combination of theory, common sense, and data
- In this case, I would say that the instrument is *valid*. I have no reason to suspect that schooling laws and ability are systematically related.
- Just because an instrument is valid doesn't mean things are all sunshine and lollipops
- Instruments *always* come with limitations
- Weakness of Q: Think of the variation in education predicted by quarter of birth. To whom is it relevant?

Example of the earnings equation

- To whom is it relevant?

Example of the earnings equation

- To whom is it relevant? Those who would like to drop out but are prevented from dropping out by compulsory schooling laws
- Another way to say this: Q is *irrelevant* to anyone who doesn't want to drop out before the minimum age in their particular state
- So would Q predict the level of education for any of you? Of course not.

Example of the earnings equation

- To whom is it relevant? Those who would like to drop out but are prevented from dropping out by compulsory schooling laws
- Another way to say this: Q is *irrelevant* to anyone who doesn't want to drop out before the minimum age in their particular state
- So would Q predict the level of education for any of you? Of course not.
- This means that the variation we are using (the "clean" variation) in education is not the variation we might be interested in if we are trying to predict the relationship between higher education and earnings
- BUT the relevant population here might actually be an important population — there is reason to care what these folks do

Validity tradeoffs

- IV Takeaway #2: Since instruments work by isolating variation, you need to be aware of the variation that you throw away, and how throwing it away limits the conclusions you can draw from your results

Example of the earnings equation

- Two potential weaknesses of the IV approach should be highlighted here
  1. Instruments can be "weak"
  2. Instruments can provide "local" estimates
- "Weak instruments" are instruments that are not "strong" predictors of the endogenous variable in the first stage regression (i.e. we throw out *a lot* of variation in a variable in order to clean up the endogeneity)
  - In general, we'd like to keep as much variation as possible while getting rid of the nasty variation (the stuff correlated with ability, in this case)

Example of the earnings equation

- Two potential weaknesses of the IV approach should be highlighted here
  1. Instruments can be "weak"
  2. Instruments can provide "local" estimates
- "Weak instruments" are instruments that are not "strong" predictors of the endogenous variable in the first stage regression (i.e. we throw out *a lot* of variation in a variable in order to clean up the endogeneity)
  - In general, we'd like to keep as much variation as possible while getting rid of the nasty variation (the stuff correlated with ability, in this case)
- Instruments provide "local" estimates if the estimates seem particularly relevant only for a specialized population (A&K is a vivid example of this)

Angrist and Krueger 1991

- It's no fun to go through an example unless you get to see the results
- Here are the actual equations estimated by A&K (second stage listed first)

$$\ln W_i = X_i\beta + \sum_c Y_{ic}\xi_c + \rho E_i + u$$

$$E_i = X_i\pi + \sum_c Y_{ic}\delta_c + \sum_c \sum_j Y_{ic}Q_{ij}\theta_{jc} + \epsilon_i$$

TABLE IV

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

...ent variable	(1) OLS	(2) TSLS	(3) OLS	(4) TSLS	(5) OLS	(6) TSLS	(7) OLS	(8) TSLS
...ation	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)
...ck)	—	—	—	—	0.2980	−0.3055	−0.2980	−0.2271
					(0.0043)	(0.0353)	(0.0043)	(0.0776)
...nter city)	—	—	—	—	0.1343	0.1362	0.1343	0.1163
					(0.0026)	(0.0092)	(0.0026)	(0.0198)
...married)	—	—	—	—	0.2928	0.2941	0.2928	0.2804
					(0.0037)	(0.0072)	(0.0037)	(0.0141)
...n dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
...sidence dummies	No	No	No	No	Yes	Yes	Yes	Yes
	—	—	0.1446	0.1409	—	—	0.1162	0.1170
			(0.0676)	(0.0704)			(0.0652)	(0.0662)
	—	—	−0.0015	−0.0014	—	—	−0.0013	−0.0012
			(0.0007)	(0.0008)			(0.0007)	(0.0007)
	—	36.0 [29]	—	25.6 [27]	—	34.2 [29]	—	28.8 [27]

...rors 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 ...e 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 ...age-squared are measured in quarters of years. Each equation also includes an intercept.

Angrist and Krueger 1991

- The previous slide is just a placeholder
- See the individual tables reproduced and posted on the website (unless you like tilting your head to the side and squinting)

Angrist and Krueger 1991

- What do the results suggest?
- That an extra (forced) year of school increases wages by somewhere between 7% - 8% (not bad!)
- We also learn that the omitted variable we were concerned about (ability) does not seem to have a large influence (the OLS and 2SLS estimates are very similar)
- Why does this make sense?

Angrist and Krueger 1991

- What do the results suggest?
- That an extra (forced) year of school increases wages by somewhere between 7% - 8% (not bad!)
- We also learn that the omitted variable we were concerned about (ability) does not seem to have a large influence (the OLS and 2SLS estimates are very similar)
- Why does this make sense?
- The variation in ability *among those student who want to drop out* does not explain a lot of the difference in earnings

Angrist and Krueger 1991

- Punchline 1: the instrumental variables approach is successful in isolating variation that we can be confident identified the effect of education on earnings
- Punchline 2: we need to be careful about how we interpret (and extrapolate) the findings — they appear most relevant for a particular population

Exercise

1. Install and load the package AER in R
2. Load the CollegeDistance dset
3. Examine the data and fit a wage model:
   wage $\sim$ urban + gender + ethnicity + unemp + education
4. Use `distance` as an instrument for education, doing 2sls "by hand" (without special commands, i.e. just using the regular `lm` regression command
5. Compare the coefficients