Lecture 24: Instrumental Variables

Pierre Biscaye

Fall 2022

Motivating Instrumental Variables

- Omitted Variables are a hard problem to solve. We
  - characterized the bias from omitted variabes
  - discussed the use of proxy variables
  - considered measurement error as an omitted variable
  - proposed assumptions on the nature of omitted variables which led to program evaluation methods
- today we will take a different approach: *instrumental variables* (IV) or *two stage least squares*
  - IV can generate unbiased estimates of β_1 even in the presence of omitted variables
  - IV will not need a panel, or a threshold, or randomization.
  - IV will just need a very special variable z.

Motivating IV: Section attendance and final exam scores

$$Final_i = b_0 + b_1 Attend_i + u_i \tag{1}$$

- What omitted variables are we concerned about?

Motivating IV: Section attendance and final exam scores

$$Final_i = b_0 + b_1 Attend_i + u_i \tag{2}$$
$$Final_i = \beta_0 + \beta_1 Attend_i + \beta_2 HrsStudied_i + v_i \tag{3}$$
$$Attend_i = \delta_0 + \delta_1 HrsStudied_i + e_1 \tag{4}$$
$$E[\hat{b_1} | Attend_i] = \beta_1 + \beta_2 \delta \tag{5}$$

- How to solve this problem?

IV flips the proxy intuition

$$Final_i = b_0 + b_1 Attend_i + u_i \tag{6}$$

$$\text{Our Problem} : cov(u_i, Attend_i) \neq 0 \tag{7}$$

- Instead of finding something correlated with $HrsStudied_i$ and controling for it we find something correlated with $Attend_i$ which is *not* correlated with study hours
- our *instrumental variable z* fits two conditions
  1. $cov(u, z) = 0$
  2. $cov(Attend, z) \neq 0$

IV in words

1. $cov(u, z) = 0$
2. $cov(Attend, z) \neq 0$

- Our problem: Section Attendance is correlated with an omitted variable in the error term (Study Hours)
- Our solution: We find a variable which is correlated with Section Attendance but uncorrelated with Study hours (or other omitted variables)
- With an instrumental variable, we say "x is not as good as random. But I can find something that is as good as random (z) which impacts x. I can use z to learn about the effects of x on y."

What kinds of variables might make a good z

- Wooldridge suggests distance to campus... concerns?

What kinds of variables might make a good z

- Other ideas:
  - PG&E preventative power cuts
  - local public health orders
  - if some people were more affected than others (e.g. PG&E power cuts happened only on a Wednesday but not a Friday, or only to some communities)
- key idea: we want something that is close to random, and should only impact section attendance
- PG&E cuts power when there are high winds
- Presumably, having PG&E cut power on the day of your section is uncorrelated with your usual studying behavior, but would cause you to miss a section

Using Instrumental Variables

- We have y, x, z, where $cov(z, u) = 0$

$$y = \beta_0 + \beta_1 x + u \tag{8}$$

$$cov(z, y) = \beta_1 cov(z, x) + cov(z, u) \tag{9}$$

$$\beta_1 = \frac{cov(z, y)}{cov(z, x)} \tag{10}$$

Estimating Instrumental Variables

$$\beta_1 = \frac{cov(z, y)}{cov(z, x)} \tag{11}$$

$$\widehat{\beta_1^{IV}} = \frac{\sum_i (y_i - \bar{y})(z_i - \bar{z})}{\sum_i (x_i - \bar{x})(z_i - \bar{z})} \tag{12}$$

Regression Interpretation

$$Attend_i = \pi_0 + \pi_1 Wind_i + v_i \tag{13}$$

- for some $\pi_1 \neq 0$ and $E[v|Wind] = 0$

$$Final_i = \beta_0 + \beta_1 Attend_i + \beta_2 HrsStudied_i + u_i \tag{14}$$
$$Final_i = \beta_0 + \beta_1(\pi_0 + \pi_1 Wind_i + v_i) + \beta_2 HrsStudied_i + u_i \tag{15}$$
$$Final_i = \beta_0 + \beta_1\pi_0 + \beta_1(\pi_1 Wind_i) + \beta_2 HrsStudied_i + u_i + \beta_1 v_i \tag{16}$$

Regression Interpretation

$$Final_i = \beta_0 + \beta_1\pi_0 + \beta_1(\pi_1 Wind_i) + \beta_2 HrsStudied_i + u_i + \beta_1 v_i \quad (17)$$

- We regress

$$Final_i = b_0 + b_1 Wind_i + e_i \quad (18)$$

$$E[\hat{b_1}|Wind] = \frac{cov(Wind, Final)}{var(Wind)} \quad (19)$$

$$E[\hat{b_1}|Wind] = \frac{1}{var(Wind)}[\beta_1\pi_1 var(Wind)$$

$$+\beta_2 cov(HrsStudied, Wind) + cov(u, Wind) + \beta_1 cov(v, Wind)] \quad (20)$$

$$E[\hat{b_1}|Wind] = \beta_1\pi_1 \quad (21)$$

Reduced Form and ITT

$$E[\hat{b_1}|Wind] = \beta_1 \pi_1 \tag{22}$$

- *Reduced Form* regression gives something like β_1 but not quite
- it tells us the effect of Wind on Final exam scores... which is the effect of attendance on exam scores weighted by the effect of Wind on section attendance
- This is the same as the ITT in Randomization with imperfect compliance

Estimating $\widehat{\beta_1^{IV}}$

$$E[\hat{b_1}|Wind] = \beta_1 \pi_1 \tag{23}$$

- π_1 is also estimable

$$Attend_i = \pi_0 + \pi_1 Wind_i + v_i \tag{24}$$

$$E[\hat{\pi}_1|Wind] = \frac{cov(Attend, Wind)}{var(Wind)} \tag{25}$$

$$\frac{E[\hat{b_1}|Wind]}{E[\hat{\pi}_1|Wind]} = \frac{\frac{cov(Final, Wind)}{var(Wind)}}{\frac{cov(Attend, Wind)}{var(Wind)}} = \frac{cov(Final, Wind)}{cov(Attend, Wind)} = E[\widehat{\beta_1^{IV}}] \tag{26}$$

IV and the ToT

$$\frac{E[\hat{b_1}|Wind]}{E[\hat{\pi_1}|Wind]} = \frac{cov(Final, Wind)}{cov(Attend, Wind)} = E[\widehat{\beta_1^{IV}}] \qquad (27)$$

Earlier: ToT

$$ToT = \frac{Y^{\bar{Prog}} - Y^{N\bar{o}Prog}}{Ed^{\bar{Prog}} - Ed^{N\bar{o}Prog}} \approx \frac{cov(Y, Prog)}{cov(Ed, Prog)} = E[\widehat{\beta_1^{IV}}] \qquad (28)$$

- We divide the relationship between y and z by the relationship between x and z
- This is the same as the ToT estimator: We assume that all of the effect of z was through changing x
  - And so we weight the relationship between z and y by the effect of z on x.
  - RCTs with imperfect compliance are the ideal case for IV

2 critical assumptions

1. $cov(u, z) = 0$
2. $cov(x, z) \neq 0$

- $cov(u, z) = 0$ is analogous to MLR 4
- But, instead of our variable of interest being unrelated to u, we just need a variable related to our variable of interest that is unrelated to u
- it means we need the *only* channel through which z effects y to be x - in other words $z \Rightarrow x \Rightarrow y$
- *Exclusion Restriction*

$cov(u, Wind) = 0$?

- Weather-based instruments are common: weather is related to many things we are interested in (farmer incomes, class attendance, customers at brick-and-mortar stores)
- Suppose there were more high wind days on Wednesdays. So, people enrolled in the Wednesday section have more cancellations due to preventative power outages.
- Would this have an effect on *Final* other than through section attendance?

Maybe?

- If $cov(wind, smoke) > 0$: Health effects on cognition?
- Selection into Wednesday vs. Friday sections?
- next time: using controls to address some of this.

Assumption 2: $cov(z, x) \neq 0$

- Unlike Assumption 1, Assumption 2 is testable

$$Attend_i = \pi_0 + \pi_1 Wind_i + u_i \tag{29}$$
$$H_0 : \pi_1 = 0 \tag{30}$$

- If we reject H_0 we have evidence in favor of Assumption 2
- if not, and $cov(z, y) \neq 0$ then Assumption 1 is unlikely to hold
  - If z influences y, it seems unlikely to be through x
- We will typically want a higher threshold for proof on this test (then 5%).

$var(\widehat{\beta_1^{IV}})$

- if we have homoskedastic errors ($E[u^2|z] = \sigma^2$)

$$var(\widehat{\beta_1^{IV}}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{31}$$

- $\sigma^2 = var(u)$
- $\sigma_x^2 = var(x)$
- $\rho_{x,z}^2 = (corr(x,z))^2$
- similar to before: except now we also know the variance will be large when $corr(x,z)$ is small

estimating $var(\widehat{\beta_1^{IV}})$

$$var(\widehat{\beta_1^{IV}}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{32}$$

$$\widehat{var(\widehat{\beta_1^{IV}})} = \frac{1}{n-2}\frac{\sum_i \hat{u_i}^2}{SST_x R_{x,z}^2} \tag{33}$$

- Note that this is the same as the OLS variance - except that the denominator is reduced by $R_{x,z}^2$
- it will always be larger than the OLS variance
- To Jupyter

Imperfect Instruments

- What if Instruments are imperfect?

$$y = \beta_0 + \beta_1 x + u \tag{34}$$

$$cov(z, y) = \beta_1 cov(z, x) + cov(z, u) \tag{35}$$

$$E[\widehat{\beta_1^{IV}}] = \frac{cov(z, y)}{cov(z, x)} = \beta_1 + \frac{cov(z, u)}{cov(z, x)} = \beta_1 + \frac{corr(z, u)}{corr(z, x)} \frac{\sigma_u}{\sigma_x} \tag{36}$$

- any bias is *magnified* by a low correlation between z and x

Quarter of Birth

- In a classic paper, Angrist and Krueger (1991) want to estimate $log(y_i) = \beta_0 + \beta_1 Ed_i + u_i$
- they are concerned, however, that $E[u_i|Ed_i] \neq 0$.
- they propose an instrument: quarter of birth
  - In the US, students are allowed to drop out of high school at age 16
  - students born late in the year turn 16 in 10th grade
  - but, students born earlier in the year turn 16 after 10th grade, or in 11th grade
  - quarter of birth might influence how many years of schooling you get but not otherwise be related to earnings

Weak Instruments

- It turns out quarter of birth is significantly correlated with schooling
  - but very weakly
- This means that even very small other relationships between quarter of birth and earnings may bias $\widehat{\beta_1^{IV}}$

$$\beta_1 + \frac{corr(z,u)}{corr(z,x)}\frac{\sigma_u}{\sigma_x} \tag{37}$$