Lecture 25: More on Instrumental Variables and Wrapping Up

Pierre Biscaye

Fall 2022

Reminder: Problem Statement

- We want to estimate $y_i = \beta_0 + \beta_1 x_i + u_i$
- But $E[u_i|x_i] \neq 0$
- Suppose we have z_i where
  1. $E[u_i|z_i] = 0$
  2. $E[x_i|z_i] \neq 0$
- Then

$$\beta_1^{IV} = \frac{cov(y,z)}{cov(x,z)} \tag{1}$$

$$\widehat{\beta_1^{IV}} = \frac{\sum_i (y_i - \bar{y})(z_i - \bar{z})}{\sum_i (x_i - \bar{x})(z_i - \bar{z})} \tag{2}$$

Regression Interpretation

$$Final_i = \beta_0 + \beta_1 Attend_i + \beta_2 HrsStudied_i + u_i \quad (3)$$

$$Attend_i = \pi_0 + \pi_1 Wind_i + v_i \quad (4)$$

$$Final_i = \beta_0 + \beta_1 \pi_0 + \beta_1 (\pi_1 Wind_i) + \beta_2 HrsStudied_i + u_i + \beta_1 v_i \quad (5)$$

Regression Interpretation(2)

- We regress

$$Final_i = b_0 + b_1 Wind_i + e_i \quad (6)$$

$$E[\hat{b_1}|Wind] = \frac{cov(Wind, Final)}{var(Wind)} \quad (7)$$

$$Final_i = \beta_0 + \beta_1\pi_0 + \beta_1(\pi_1 Wind_i) + \beta_2 HrsStudied_i + u_i + \beta_1 v_i \quad (8)$$

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

$$+\beta_2 cov(HrsStudied, Wind) + cov(u, Wind) + \beta_1 cov(v, Wind)] \quad (9)$$

$$E[\hat{b_1}|Wind] = \beta_1\pi_1 \quad (10)$$

Reduced Form and ITT

$$E[\hat{b_1}|Wind] = \beta_1 \pi_1 \quad (11)$$

- *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 $\widehat{\beta_1^{IV}}$

$$E[\hat{b_1}|Wind] = \beta_1 \pi_1 \tag{12}$$

- π_1 is also estimable

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

$$E[\hat{\pi}_1|Wind] = \frac{cov(Attend, Wind)}{var(Wind)} \tag{14}$$

$$\frac{E[\hat{b_1}|Wind]}{E[\hat{\pi}_1|Wind]} = \frac{\frac{cov(Final, Wind)}{var(Wind)}}{\frac{cov(Attend, Wind)}{var(Wind)}} = \frac{cov(Final, Wind)}{cov(Attend, Wind)} = E[\widehat{\beta_1^{IV}}] \tag{15}$$

IV and the ToT

$$\frac{E[\hat{b}_1|Wind]}{E[\hat{\pi}_1|Wind]} = \frac{cov(Final, Wind)}{cov(Attend, Wind)} = E[\widehat{\beta_1^{IV}}] \tag{16}$$

Earlier: ToT

$$ToT = \frac{Y^{\bar{P}rog} - Y^{N\bar{o}Prog}}{Ed^{\bar{P}rog} - Ed^{N\bar{o}Prog}} \approx \frac{cov(Y, Prog)}{cov(Ed, Prog)} = E[\widehat{\beta_1^{IV}}] \tag{17}$$

- 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?
- (in errata at end) can use controls to address some of this.

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

- Unlike Assumption 1, Assumption 2 is testable

$$Attend_i = \pi_0 + \pi_1 Wind_i + u_i \quad (18)$$

$$H_0 : \pi_1 = 0 \quad (19)$$

- 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 ($E[u^2|z] = \sigma^2$)

$$var(\widehat{\beta_1^{IV}}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{20}$$

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

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

$$var(\widehat{\beta_1^{IV}}) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{21}$$

$$\widehat{var(\widehat{\beta_1^{IV}})} = \frac{1}{n-2}\frac{\sum_i \hat{u}_i{}^2}{SST_x R_{x,z}^2} \tag{22}$$

- 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, if time allows

Another useful interpretation of IV

$$x_i = \pi_0 + \pi_1 z_i + v_i \tag{23}$$
$$\pi_1 \neq 0 \tag{24}$$
$$E[u_i|z_i] = 0 \tag{25}$$

- This is called the *first stage* regression. Part of the variation in x_i is explained by z_i
- That part is *exogenous*.
- The *endogenous* variation in x are in the v errors

2 stage Least Squares

$$\hat{x}_i = \hat{\pi}_0 + \hat{\pi}_1 z_i = x_i - \hat{v}_i \tag{26}$$
$$y_i = \beta_0 + \beta_1 x_i + u_i \tag{27}$$
$$y_i = \beta_0 + \beta_1(\hat{\pi}_0 + \hat{\pi}_1 z_i) + \beta_1 \hat{v}_i + u_i \tag{28}$$
$$E[u_i + \beta_1 \hat{v}_i | \hat{\pi}_0 + \hat{\pi}_1 z_i] = 0 \tag{29}$$

- So if we regress y on $\hat{x}$ we get an unbiased estimate for β_1.
- it turns out, this is the precise same estimate as $\widehat{\beta_1^{IV}}$
- (not today, but demonstration in notebook in datahub)

2 stage least squares summary

- Intuitively, when we run 2 stage least squares, we are predicting x using z
- and then only using the variation in z to understand the variation between x and y
  - in our section attendance example, section attendance was correlated with lots of things in u
  - but the weather also influenced attendance at particular sections, and maybe was uncorrelated with those things
  - so we predicted section attendance using the weather, and used those predictions to understand the relationship between section attendance and final exam scores
- Critically, it is important to run the actual *ivreg* and not separately run the two stages
  - Otherwise, the errors will be incorrectly estimated

Imperfect Instruments

- What if Instruments are imperfect?

$$y = \beta_0 + \beta_1 x + u \tag{30}$$

$$cov(z,y) = \beta_1 cov(z,x) + cov(z,u) \tag{31}$$

$$E[\widehat{\beta_1^{IV}}] = \frac{cov(z,y)}{cov(z,x)} = \beta_1 + \frac{cov(z,u)}{cov(z,x)} = \beta_1 + \frac{corr(z,u)}{corr(z,x)}\frac{\sigma_u}{\sigma_x} \tag{32}$$

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

$$\beta_1 + \frac{corr(z,u)}{corr(z,x)} \frac{\sigma_u}{\sigma_x} \tag{33}$$

Review outline: First Third P1

1. Random Variables, Expectations, Conditional Expectations
2. Simple and Multiple Linear Regression
   - Derivation
   - Predicted values, residuals
   - variance of $\hat{\beta_1}$ and heteroskedasticity
   - functional forms and interpretation of marginal effects
   - R^2
   - *Ceteris Paribus* interpretations and bad controls
   - MLR1-MLR4

Review outline: First Third P2

3. Confidence Intervals and Hypothesis tests
4. *p*-values
5. Distribution of $\hat{\beta_1}$ and MLR5-6
6. Hypothesis tests in regression framework

Review Outline: Second Third P1

1. F-tests and Chow tests
2. Adjusted R^2 and selecting between functional forms
3. Changing Units in indep and dependent variables
4. Qualitative Data
   - Categorical Variables as dependent variables
   - Categorical Variables as independent variables
   - Policy Analysis with qualitative data
5. Confidence intervals for predicted values

Review Outline: Second Third P2

5. Problems with MLR 4
   - Proxy Variables
   - Measurement Error
6. Potential Outcomes Framework

Review Outline: Third Third P1

1. Randomized Controlled Trials
   - spurious imbalance
   - imperfect compliance: ITT and ToT estimators
   - compliers vs. always-takers vs. never takers
2. Regression Discontinuity Design
   - identification strategy and assumption
   - sharp vs. fuzzy RDD
   - compliers and threats to identification
3. Difference-in-Differences
   - Identification assumption
   - graphical analysis
   - threats and strategies for identification

Review Outline: Third Third P2

4. Panel Data
   - First Differences
   - Fixed Effects
   - identification assumptions and strategy
   - unbalanced panels
5. Instrumental Variables (Not on exam this year)
   - Identification strategy and assumptions
   - 2sls
   - Weak Instruments
   - Measurement error

More on IV that we didn't get to this year

One more application: Measurement error in x

$$x_1 = x_1^* + e_1 \tag{34}$$

$$y = \beta_0 + \beta_1 x_1^* + u \tag{35}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_1 e_1 + u \tag{36}$$

$$E[e_1 | x_1^*] = 0 \quad E[e_1 | x_1] \neq 0 \tag{37}$$

- If we regress $y = \beta_0 + \beta_1 x_1 + \beta_1 e_1 + u$ we know that e_1 is an omitted variable and $\hat{\beta_1}$ is biased

We can use IV to overcome this bias

- Suppose we have a second measurement of x_1^*, z_1.

$$z_1 = x_1^* + a_1 \tag{38}$$
$$x_1 = \pi_0 + \pi_1 z_1 + b_1 \tag{39}$$
$$y = \beta_0 + \beta_1 x_1 + \beta_1 e_1 + u \tag{40}$$
$$y = \beta_0 + \beta_1(\hat{\pi}_0 + \hat{\pi}_1 z_1) + \beta_1(\hat{b_1} + e_1) + u \tag{41}$$
$$y = \beta_0 + \beta_1(\hat{\pi}_0 + \hat{\pi}_1(x_1^* + a_1)) + \beta_1(\hat{b_1} + e_1) + u \tag{42}$$

- $E[e_1|x_1^*] = 0$, $E[\hat{b_1}|z_1] = 0$ so the only question is if $E[e_1|a_1] = 0$
- e_1 and a_1 are the two measurement errors. If these are unrelated, we can estimate $\hat{\beta_1}$ where $E[\hat{\beta_1}] = \beta_1$

To Jupyter

Extending to Multiple Regression

$$log(wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 exper_i + u_i \tag{43}$$

- Suppose we are concerned that Education is endogenous ($corr(Educ_i, u_i) \neq 0$)
- but are confident that Experience is exogenous ($corr(exper_i, u_i) = 0$)
- what can we do?

What we can't do

- We can't:
  - run OLS and expect either $E[\hat{\beta_1}] = \beta_1$ or $E[\hat{\beta_2}] = \beta_2$
  - consider $log(wage_i) = b_0 + b_1 Educ_i + v_i$ and use $exper_i$ as an instrument for $Educ_i$
  - why?

What we can't do

- We can't:
  - run OLS and expect either $E[\hat{\beta_1}] = \beta_1$ or $E[\hat{\beta_2}] = \beta_2$
  - consider $log(wage_i) = b_0 + b_1 Educ_i + v_i$ and use $exper_i$ as an instrument for $Educ_i$
  - why?
  - $E[v_i | Exper_i] \neq 0$: *Exper* has an independent effect on wages

Instead, we need an instrument that *only* impacts wages through Education

$$E[u_i|z_{1i}] = 0 \tag{44}$$

$$E[Educ_i|z_{1i}, exper_i] \neq E[Educ_i|exper_i] \tag{45}$$

$$Educ_i = \pi_0 + \pi_1 z_{1i} + \pi_2 Exper_i + e_i \tag{46}$$

$$\widehat{Educ_i} = \hat{\pi}_0 + \hat{\pi}_1 z_{1i} + \hat{\pi}_2 Exper_i \tag{47}$$

$$log(wage_i) = \beta_0 + \beta_1 \widehat{Educ_i} + \beta_2 Exper_i + u_i \tag{48}$$

- always need to control for any exogenous variables in both stages
- once again, literally running two stages will get you the wrong errors
- but using ivreg calculates both correctly and the intuition is the same.

More endogenous and exogenous variables

- All of this generalizes to cases with more endogenous variables and more exogenous (control) variables
- In general, the rule is we need at least as many instrumental variables as we have endogenous variables
- in practice, a good instrument is hard to find - becomes very challenging if we have more than one endogenous variable.

IV summary

- IV solution to MLR4 is to divide our endogenous variable x into two parts: the part that is endogenous and the part that is exogenous
- we do that by finding a variable z which is correlated with x but exogenous, and predicting x using z
- We then use the part of x explained by z to learn about the relationship between x and y.