# Lecture 14: Instrumental Variables (continued)

April 23, 2018

# IV Intuition

- I will formalize the idea and the math behind IVs with a simple example
- Suppose that we are interested in investigating the effect of studyng on grades:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- where
  - $Y_i$: is GPA
  - $X_i$: is study time (hours per day)
- We expect that our OLS estimator $\hat{\beta}_1$ will be severely biased here (why?)

# IV Model

- To get rid of OVB, we shall use an IV
- One possible IV: whether your roomate has a N64 (or playstation/whatever video game console kids use today)
- Important feature: Roomates are randomly assigned in college
  - At least at Berea College (Kentucky) where this example comes from (see Stinebrickner and Stinebrickner (2008))

# Instrument Valdity

- First thing for any instrument is to think of validity. Two conditions:
  - Instrument is $N64$, which is indicator variable that your roomate has N64

1. Relevance: $Cov(Z, X) \neq 0$; here $Cov(N64, study) \neq 0$
   - Seems likely to hold as everyone prefers playing Mario Kart to studying
   - Testable
2. Exogeneity: $Cov(Z, U) = 0$; here $Cov(N64, U) = 0$
   - Untestable

# Thinking About Exogeneity

- Exogeneity: $Cov(Z, U) = 0$
- "Storytime" should talk about how two things hold
  1. People do not select into $Z$ in some manner that is likely to be correlated with $Y$
     - Concern: People that pick roomates that are "fun" and have N64s are likely people who do not care too much about grades
  2. $Z$ only affects $Y$ through $X$
     - Concern: Having a roomate with a video game affects your GPA through other means than affecting your study hours
- Either story "invalidates" the instrument (but in different ways)

# Thinking About Exogeneity

1. Selection story: there is an omitted variable related to both $Z$ and $Y$
   - Example: omitted variable is effort because students that really put in a lot of effort make sure they do not get a roomate with N64
2. Other channel story:
   - Possible Story 1: Mario Kart helps me grasp physics, so I ace my physics exam
     - so $Z$ directly affects Y *independent* of $X$
   - Possible Story 2: Other people hang out in our room due to our N64, making it really loud and so I cannot study effectively
     - so $Z$ affects Y through *another* $X$

# IV Directly into Model

- Suppose that our IV assumptions hold
- Then we can just directly replace $X$ with our IV in our model:

$$Y_i = \beta_0 + \beta_1 Z_i + u_i$$

- where
  - $Y_i$: is GPA
  - $Z_i$: is whether roomate has N64
- **Interpretation of $\hat{\beta}_1$:** Having a roomate with a N64 **causes** a $\hat{\beta}_1$ decrease in GPA
  - Causal effect because if instrument is valid since $cov(Z, U) = 0$ (so $\mathbb{E}[U|Z] = 0$)
- But we want the effect of studying on GPA!

# Two Stage Least Squares

- We estimate IV models with Two Stage Least Squares (2SLS)
- This effectively runs regression $Y_i = \beta_0 + \beta_1 Z_i + u_i$, but *scales* $\beta_1$ by how much $Z$ affects $X$
  - If $Z$ affects $X$ a lot, $\beta_1$ in above regression is fine
  - If $Z$ affects $X$ only a litte, we need to scale $\beta_1$ up a lot to say what affect of $X$ on $Y$ is
- We therefore estimate in two steps:
  1. Regress $X$ on $Z$ to capture effect of $Z$ on $X$
  2. Regress residuals of step 1 to capture effect of $X$ on $Y$
     - Effectively put **scaled** $Z$ into $Y_i = \beta_0 + \beta_1 Z_i + u_i$ regression
- Under IV assumptions, this gives us **causal** effect of $X$ on $Y$

# 2SLS Implementation

- Step 1: regress

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- Step 2: take your predicted values from Step 1, $\hat{X}_i$, and regress

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$$

- Under IV assumptions, step 1 finds the "good part" of $X$ (that is not related to $U$), and then step 2 takes that "good part" of $X$ as a regression to find the causal relationship between $X$ and $Y$

# 2SLS Estimator Intuition

- Step 1: regress

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$

- Step 2: For intuition, you could regress:

$$Y_i = \gamma_0 + \gamma_1 Z_i + u_i$$

- Our IV estimator is:

$$\hat{\beta}_1^{2SLS} = \frac{Cov(Y,Z)}{Cov(X,Z)} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} \quad (since\ X_i = \pi_0 + \pi_1 Z_i)$$

- We need relevance assumption ($Cov(X,Z) \neq 0$) else $\hat{\beta}_1^{2SLS} = \infty$

# 2SLS Estimator

- Why is $\hat{\beta}_1^{2SLS} = \frac{Cov(Y,Z)}{Cov(X,Z)}$?
- True model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Take covariance of both sides with respect to $Z$:

$$Cov(Y_i, Z_i) = Cov(\beta_0 + \beta_1 X_i + u_i, Z_i)$$
$$Cov(Y_i, Z_i) = Cov(\beta_0, Z_i) + Cov(\beta_1 X_i, Z_i) + Cov(u_i, Z_i)$$
$$Cov(Y_i, Z_i) = 0 + \beta_1 Cov(X_i, Z_i) + 0 \quad \textit{since } Cov(u_i, Z_i) \textit{ by assumption}$$
$$\implies \beta_1 = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

# Unbiasedness

- **Question:** Under exogeneity assumption ($Cov(Z, U) = 0$), show that the IV estimator $\hat{\beta}_1^{2SLS} = \frac{Cov(Y,Z)}{Cov(X,Z)}$ is unbiased

# Workspace

# Math Sidenote:

- "Actual" IV Estimator (that R implements using data):

$$\hat{\beta}_1^{2SLS} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}$$

- You can do proof using above equation, OR start with following equation:

$$\hat{\beta}_1^{2SLS} = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

- Either proof is acceptabe in this course (see next slides for explanation why)
  - It is because they are the same thing (as $n \to \infty$)!

# Two Unbiasedness Proofs:

- "Proper" summation proof:

$$\mathbb{E}[\hat{\beta}_1^{2SLS}] = \mathbb{E}\left[\frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}\right]$$

$$= \mathbb{E}\left[\beta_1 + \frac{\sum_{i=1}^{N}(z_i - \bar{z})u_i}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}\right]$$

$$= \beta_1 + \frac{\sum_{i=1}^{N}(z_i - \bar{z})\mathbb{E}[u_i]}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})} = \beta_1 + 0$$

- (obviously, I skipped a few steps here)

# Two Unbiasedness Proofs:

- "Shortcut" summation proof:

$$\mathbb{E}[\hat{\beta}_1^{2SLS}] = \mathbb{E}\left[\frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})(z_i - \bar{z})}\right] = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

- Then we proceed as I did on the board for slide 12
- Why is this the same? Because we showed earlier in course that:

$$s_{ZY} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{n-1} \quad \textit{is an unbiased estimator of } \sigma_{ZY}$$

- Or, put another way: $\mathbb{E}[s_{ZY}] = \sigma_{ZY} \equiv Cov(Z, Y)$
  - We do not worry about $n-1$ denominator here, as it is in both numerator and denominator of $\hat{\beta}_1^{2SLS}$ so it cancels out

# Single Variable IV in R: Manually

- Data:
  - $Y_i$: is GPA
  - $X_i$: is study time (hours per day)
  - $Z_i$: is indicator for roomate having N64
- Running 2SLS by doing both steps manually:
- Step 1: Regress $X_i = \pi_0 + \pi_1 Z_i + v_i$ and get predicted study time, $\hat{X}_i$
  - Data\$xHat = predict(lm(X∼Z, data=Data))
- Step 2: Regress $y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$
  - m3 = lm(Y∼xHat, data=Data)
  - coeftest(m3, vcov = vcovHC(m3, type = "HC1"))
    - Note: Standard errors will be wrong!
  - `predict` gives the fitted values of $X$ from the first stage
  - Then `lm` can be used to get the second stage
  - Problem: the standard errors (even robust) will be wrong

# Doing Single Variable IV in R: All at once

- Need a new package called `AER` for this
  - install.packages("AER")
  - library(AER)
- `R` can run 2SLS all at once:
- m4 = ivreg(Y $\sim$ X | Z, data=Data)
- Use new standard error command to get "right" standard errors:
  - coeftest(m4, vcov=sandwich)
  - `ivreg` will do 2SLS directly (called from the `AER` package)
  - Command starts the same as usual `lm` formula
  - The "pipe", |, (NOT A COMMA) is how you tell `R` what are the instruments
  - Write `sandwich` proper IV standard errors

# An IV Example

- Economists long been interested in effect of competition on school performance
- Milton Friedman: Why not just give a (government funded) voucher to every student?
  - Will raise competition
  - Problem: might impact peers or parental voice (see Hirschman (1976))
- Identifying effect of competition will find the "benefit" of Friedman's proposal
- One measure of competition: number of school districts in a city

# An IV Example

- Regression we would like to run:

$$scores_{ic} = \beta_0 + \beta_1 \#districts_c + u_{ic}$$

- where
  - $i$ indexes students, $c$ indexes cities
  - $scores_{ic}$: test score of student $i$ in city $c$
  - $\#districts_c$: Number of school districts in city $c$ (measure of competition)

# An IV Example

- Our data is:
  - $i$ indexes students, $c$ indexes cities
  - $scores_{ic}$: test score of student $i$ in city $c$
  - $\#districts_c$: Number of school districts in city $c$ (measure of competition)
  - $\#streams_c$: Number of streams in city $c$ (will be our IV)
    - A stream is defined as a non-navigable body of water that cannot be "jumped" across
- Suggested solution: use # of streams as IV for # of districts

# Questions:

- **Question 1:** In our original regression without an IV ($scores_{ic} = \beta_0 + \beta_1 \#districts_c + u_{ic}$) give an example of an omitted variable that would cause upward bias and one that would cause downward bias
- **Question 2:** Describe how to run this regression (formally write out the equations)
- **Question 3:** Discuss the validity of the instrumental variable

# Workspace

# Workspace