Instrumental Variables Estimation

Tyler Ransom

Univ of Oklahoma

Mar 14, 2019

Today's plan

1. Review reading topics on when $E(u|\mathbf{x}) \neq 0$
   1.1 What are Instrumental Variables?
   1.2 How to estimate causal effects with IV
2. In-class activity: basic IV estimation in R

What are Instrumental Variables?

Instrumental Variables

- Suppose we have cross-sectional data, and one x might be endogenous
- We have basically two choices to resolve this problem:
    1. Collect good controls, hope that the variable becomes exogenous
    2. Find one or more **instrumental variables** for the endogenous x variable
- An IV (call it z) is a variable correlated with x, but not with u
    - IV's typically come out of so-called **natural experiments**
    - e.g. exogenous change in laws; school choice lotteries; military conscription

Example: Class size and student performance

- Consider a model in the population:

$$score = \beta_0 + \beta_1 classize + u$$

  where we think *classize* is endogenous:

$$Cov(classize, u) \neq 0$$

- Why would *classize* be endogenous?
  - More motivated parents choose to live in better-funded school districts
  - Teachers prefer to teach in better districts, so can have more classrooms

Example: Class size and student performance

- Could try to put in proxies for family background, SES, etc.

- But probably won't be able to capture everything in u that affects *score*

- A solution: collect data on a variable z that satisfies

  1. z is **exogenous** to the equation:

     $$Cov(z, u) = 0$$

  2. z is **relevant** for explaining x:

     $$Cov(z, x) \neq 0$$

Testability of IV assumptions

- We **cannot** test condition (1)
    - Must appeal to theory or qualitative evidence
- We can test condition (2)
    - Can easily compute $Corr\left(z,x\right)$ and use a t-test

Deriving the formula for the IV estimator

- Take our population model and take $Cov(z, \cdot)$ to both sides:

$$\begin{aligned} y &= \beta_0 + \beta_1 x + u \\ Cov(z,y) &= \underbrace{Cov(z,\beta_0)}_{=0} + Cov(z,\beta_1 x) + \underbrace{Cov(z,u)}_{=0 \text{ (cond. 1)}} \\ &= \beta_1 Cov(z,x) \end{aligned}$$

then solving for β_1 we get:

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$$

Deriving the formula for the IV estimator

- Translating the previous formula from population to sample gives:

$$\hat{\beta}_{1,IV} = \frac{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{n^{-1}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

$$= \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$$

Properties of IV

- $\hat{\beta}_{1,IV}$ is consistent, but biased! Bias $\uparrow$ as $|Corr(z,x)| \downarrow$

- $Var(\hat{\beta}_{1,IV}) > Var(\hat{\beta}_{1,OLS})$

$$Var(\hat{\beta}_{1,IV}) \approx \frac{\sigma_u^2}{n\sigma_x^2\rho_{x,z}^2}$$

$$Var(\hat{\beta}_{1,OLS}) \approx \frac{\sigma_u^2}{n\sigma_x^2}$$

where $\rho_{x,z}^2$ is correlation between x and z

How much larger is $Var(\hat{\beta}_{1,IV})$ than $Var(\hat{\beta}_{1,OLS})$?

- Rule of thumb:

$$se(\hat{\beta}_{1,IV}) \approx \frac{se(\hat{\beta}_{1,OLS})}{|r_{xz}|}$$

  where r_{xz} is the sample correlation between x and z

- This is the cost of doing IV when we could be doing OLS

- A type of bias-variance tradeoff

- Often $|r_{xz}|$ is small, so IV standard error is “large”; can offset with large N

How to estimate causal effects with IV

How to do IV in R, generally

Suppose our variables are y, x, and z

```
library(AER)

est.ols <- lm(y ~ x, data=df)
est.iv  <- ivreg(y ~ x | z, data=df)
```

To check if x and z are correlated:

```
est.iv1  <- lm(x ~ z , data=df)
```

Example: Kids and Labor supply

- Just because a var. is randomized does not make it exogenous to a model
- Economic agents can change their behavior!
- Angrist and Evans (1998) look at Mom's hours worked with number of kids:

$$hours = \beta_0 + \beta_1 kids + u$$

  for those who have at least 2 children

- IV: dummy for if first two kids are of same sex (call it *samesex*)
- Thought process: marginal cost of 2nd kid lower if of same sex as 1st

Example: Kids and Labor supply

- Use their data ($N = 666,384$)

```
df1 %>% select(KIDCOUNT,HOURSMOM,SAMESEX)
    %>% as.data.frame
    %>% stargazer(type='text')
```

```
========================================
Statistic     N      Mean   St. Dev. Min Max
----------------------------------------
KIDCOUNT   666,384 2.454    0.758     2   12
HOURSMOM   666,384 23.618   18.913    0   99
SAMESEX    666,384 0.504    0.500     0    1
----------------------------------------
```

Descriptive stats look reasonable

Example: Kids and Labor supply

OLS estimates:

```
 est.ols <- lm(HOURSMOM ~ KIDCOUNT, data=df1)

 tidy(est.ols)
# A tibble: 2 x 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)    32.0     0.0777      412.      0.
2 KIDCOUNT       -3.41    0.0303     -113.      0.
```

More kids $\implies$ fewer hours worked

Example: Kids and Labor supply

Check that *samesex* is correlated with *kidcount*:

```
 est.iv1 <- lm(KIDCOUNT ~ SAMESEX, data=df1)
 tidy(est.iv1)
# A tibble: 2 x 5
  term          estimate std.error statistic p.value
  <chr>            <dbl>     <dbl>     <dbl>   <dbl>
1 (Intercept)     2.41     0.00132    1832.        0.
2 SAMESEXTRUE     0.0806   0.00186      43.4       0.
```

So having same gender kids $\implies$ couple will have more kids

Example: Kids and Labor supply

Now compute the IV estimates:

```
 est.iv <- ivreg(HOURSMOM ~ KIDCOUNT | SAMESEX, data=df1)
 tidy(est.iv)
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)    30.7      1.40       22.0  4.37e-107
2 KIDCOUNT       -2.90     0.570      -5.09 3.51e-  7
```

Notice: t-stat went from **-113 (OLS) to -5 (IV)**

Another way of viewing regression output

```
stargazer(est.ols,est.iv1,est.iv, type="text")

=========================================================================
                                           Dependent variable:
                                   --------------------------------------
                                     HOURSMOM     KIDCOUNT     HOURSMOM
                                       OLS          OLS      instrumental
                                                               variable
                                       (1)          (2)          (3)
-------------------------------------------------------------------------
KIDCOUNT                             -3.415***                 -2.902***
                                      (0.030)                   (0.570)

SAMESEX                                           0.081***
                                                  (0.002)

Constant                             31.998***    2.414***     30.739***
                                      (0.078)     (0.001)       (1.398)

-------------------------------------------------------------------------
Observations                          666,384     666,384       666,384
R2                                     0.019       0.003         0.018
Adjusted R2                            0.019       0.003         0.018
Residual Std. Error (df = 666382)     18.735       0.757         18.739
F Statistic (df = 1; 666382)       12,733.180*** 1,886.461***
=========================================================================
Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Comparing OLS and IV SEs

From the previous example, can compute $Corr(kidcount, samesex)$

```
cor(df1$KIDCOUNT,df1$SAMESEX)
[1] 0.05313105

# Actual ratio of IV se to OLS se:

.570/.030
[1] 19

# Ratio from rule-of-thumb:
1/0.05313105
[1] 18.82139
```

One more time about the assumptions

- In the previous example, no way to test if *samesex* is exogenous
- We must assume it is in order to trust IV to be consistent
- In some cases, we can use other info to determine if IV is exogenous
- You'll explore this in PS4 (due next time!)
- Typically, need a richer data set to be able to do this

Multiple instruments, conditional exogeneity

- Nothing stops us from using multiple instruments
- In fact, the more instruments the better (so long as they're all exogenous!)
- Sometimes an instrument is only exogenous *conditional on* other x's
- In this case, z must be *partially correlated* with the endogenous x
- We'll talk more next time about each of these cases

References

Angrist, Joshua D and William N Evans. 1998. "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size." *American Economic Review* 88 (3):450–477. URL `http://www.jstor.org/stable/116844`.