Two-Stage Least squares (2SLS)

Tyler Ransom

Univ of Oklahoma

Mar 26, 2019

Today's plan

1. Review reading topics on when $E(u|\mathbf{x}) \neq 0$
    1.1 Two Stage Least Squares (2SLS)
2. In-class activity: 2SLS estimation in R

Quick Review

Instrumental Variables (IV)

- 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
- Allows us to estimate a causal effect, even when A4 is violated

IV conditions

- The instrument z must satisfy

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

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

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

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

Two Stage Least Squares (2SLS)

Two Stage Least Squares (2SLS)

- When we have more z's than needed, IV = **two stage least squares (2SLS)**
- `ivreg()` in R works the same
- Any variables not included before the "|" are instruments
- Any variables excluded after the "|" are endogenous x's
- e.g. `est.iv <- ivreg(HOURSMOM ~ KIDCOUNT | SAMESEX)`
- `SAMESEX` is the instrument; `KIDCOUNT` is the endogenous x

Why it's called 2SLS

- There are two "stages" of estimation:
    1. Regress endogenous *x*('s) on instrument(s)
    2. Regress y on *x*'s and predicted values of the endogenous *x*('s)
- 1st stage is a "relevance" regression
- 2nd stage relies on the exogeneity assumption
- Don't compute 2SLS by hand (this will give you wrong SEs)

Weak Instruments

- So-called **weak instruments** cause a litany of problems
- biased estimates, large SEs are foremost
- How can you tell if your instrument(s) is (are) weak?
- If a single instrument, first-stage $|t| > 3.2 \approx \sqrt{10}$
- If multiple instruments, first-stage $F > 10$
- The above are rules of thumb from Stock and Yogo (2005)

Multiple instruments, multiple endogenous x's

- Things get slightly more complicated with multiple z's and endogenous x's
- To even proceed, we must satisfy the **order condition**:
- Must exclude at least as many z's from our equation as endogenous x's
- Example: Female labor supply (from last time)
- `KIDCOUNT` is endogenous x, `SAMESEX` is excluded z
- Could also come up with additional instruments for `KIDCOUNT`

Example: Distance to college and returns to education

- Influential paper: Card (1995)
- Use distance to college (as teenager) as instrument for education
- “I happened to grow up close to a college”
- “This makes me more likely to attend when college aged”
- Wooldridge data set `card`

Example: Distance to college and returns to education

```
df <- as_tibble(card)
df %>% select(nearc2,nearc4,fatheduc,motheduc)
   %>% as.data.frame
   %>% stargazer(type='text')

=======================================
Statistic   N    Mean  St. Dev. Min Max
---------------------------------------
nearc2    3,010 0.441   0.497    0   1
nearc4    3,010 0.682   0.466    0   1
fatheduc  2,320 10.003  3.721    0  18
motheduc  2,657 10.348  3.180    0  18
---------------------------------------
```

Example: Distance to college and returns to education

OLS estimation:

```
est.ols <- lm(lwage ~ educ + exper + expersq +
              black + smsa + south + smsa66 +
              reg662 + reg663 + reg664 +
              reg665 + reg666 + reg667 +
              reg668 + reg669, data=df)
```

Example: Distance to college and returns to education

better code (create factor variable):

```
df %<>% mutate(reg66cat = case_when(reg661==1 ~ "1",
                                    reg662==1 ~ "2",
                                    reg663==1 ~ "3",
                                    reg664==1 ~ "4",
                                    reg665==1 ~ "5",
                                    reg666==1 ~ "6",
                                    reg667==1 ~ "7",
                                    reg668==1 ~ "8",
                                    reg669==1 ~ "9"),
               reg66cat = factor(reg66cat))
est.ols <- lm(lwage ~ educ + exper + expersq +
               black + smsa + south + smsa66 +
               reg66cat, data=df)
```

Example: Distance to college and returns to education

OLS estimates:

```
stargazer(est.ols, keep=c("educ"), type="text")
================================================
                        Dependent variable:
                    ----------------------------
                               lwage
------------------------------------------------
educ                         0.075***
                             (0.003)

------------------------------------------------
Observations                   3,010
R2                             0.300
Adjusted R2                    0.296
Residual Std. Error      0.372 (df = 2994)
F Statistic          85.476*** (df = 15; 2994)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Example: Distance to college and returns to education

Now estimate the first stage regression with two instruments and do *F*-test:

```
est.iv1 <- lm(educ ~ nearc2 + nearc4 + exper + expersq +
              black + smsa + south + smsa66 + reg66cat, data=d

linearHypothesis(est.iv1, c("nearc2", "nearc4"), vcov=hccm)
```

Note the use of robust standard errors (`vcov=hccm`)

Example: Distance to college and returns to education

First-stage results reveal $F = 8.26 < 10$, but maybe it's "close enough" to 10

```
Linear hypothesis test
Hypothesis:
nearc2 = 0
nearc4 = 0

Model 1: restricted model
Model 2: educ ~ nearc2 + nearc4 + exper + expersq + black + smsa +
    south + smsa66 + reg66cat

Note: Coefficient covariance matrix supplied.

  Res.Df Df      F    Pr(>F)
1   2995
2   2993  2 8.2562 0.0002656 ***
```

Example: Distance to college and returns to education

Now do the IV estimation:

```
est.iv <- ivreg(lwage ~ educ + exper + expersq + black +
                  smsa + south + smsa66 + reg66cat |
                  nearc2 + nearc4 + exper + expersq + black +
                  smsa + south + smsa66 + reg66cat, data=df)
```

Example: Distance to college and returns to education

OLS and IV estimates:

```
stargazer(est.ols, est.iv, keep=c("educ"), type="text")

================================================================
                                Dependent variable:
                            ------------------------------------
                                        lwage
                                  OLS            instrumental
                                                   variable
                                  (1)                (2)
----------------------------------------------------------------
educ                            0.075***           0.157***
                                (0.003)            (0.053)

----------------------------------------------------------------
Observations                      3,010              3,010
R2                                0.300              0.170
Adjusted R2                       0.296              0.166
Residual Std. Error (df = 2994)   0.372              0.405
F Statistic                85.476*** (df = 15; 2994)
================================================================
Note:                              *p<0.1; **p<0.05; ***p<0.01
```

Do the results make sense?

- The IV estimate of the returns to education that is double the OLS estimate
- Think about why this might be
- Are *nearc2* and *nearc4* exogenous?
- Whenever you use IV estimation, you must think deeply about the assumptions
- Finding truly valid IV's is really difficult

References

Card, David. 1995. "Using Geographic Variation in College Proximity to Estimate the Return to Schooling." In *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, edited by Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky. Toronto: University of Toronto Press, 201–222.

Stock, James H. and Motohiro Yogo. 2005. "Testing for Weak Instruments in Linear IV Regression." In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, edited by D.W.K. Andrews and J.H. Stock. Cambridge: Cambridge University Press, 80–108.