Intstrumental Variables

Sidak Yntiso sgy210@nyu.edu

April 13, 2020

Identifying assumptions

Second stage: $Y_i = \alpha_0 + \alpha_1 D_i + \epsilon_i$

First stage: $D_i = \beta_0 + \beta_1 Z_i + \mu_i$

- Exogenous instrument
  - $Cov(Z_i, \mu_i) = 0$ i.e., $D_0, D_1 \perp\!\!\!\perp Z$
- Exclusion restriction
  - $Cov(Z_i, \epsilon_i) = 0$ i.e., $Y_0, Y_1 \perp\!\!\!\perp Z$
- First stage
  - $\beta_1 \neq 0$ i.e, $0 < P(Z = 1) < 1$ and $P(D_1 = 1) \neq P(D_1 = 0)$
- Monotonicity ($D_1 \geq D_0$)

Overview

- Wald estimator
  - Constant treatment effects & binary instrument
  - Tests for first stage
  - Placebo regressions for exclusion restriction
- Preliminaries on 2SLS estimator
  - More in class
  - Heteregeneous treatment effects
  - Two papers

Paper 1: Bloom et al 1997

- What is the effect of participation in job training programs on earnings?
- Leverage random assignment of admission to training program
  - 21,000 person RCT commisioned by US Dept of Labor in 1986
  - 16 local areas across the country between 1987 and 1989
  - Sample consists of economically disadvantaged adults and out-of-school youths
- Outcomes: total earnings and educational attainment
- Problems with compliance (not a perfect experiment)

Load the Data

```
library(haven)
library(estimatr)
rm(list=ls())
setwd("C:\\Users\\Sidak Yntiso\\Dropbox\\CI\\Week 10\\Lab")
load("jtpa.RDA")

#imperfect compliance
mean(d$training[d$assignmt==1])
```

```
## [1] 0.6415976
```

```
mean(d$training[d$assignmt==0])
```

```
## [1] 0.01452785
```

```
#naive OLS maybe biased
summary(lm_robust(earnings~training,data=d))$coefficients
```

```
##              Estimate Std. Error   t value    Pr(>|t|)
## (Intercept) 14605.085   206.8771 70.597879 0.00000e+00
```

First stage effect

```r
#regression effect of Z on D
summary(lm_robust(training~assignmt,data=d))$coefficients
```

```
##                Estimate  Std. Error    t value     Pr(>|t
## (Intercept) 0.01452785 0.001962840   7.401441 1.443646e-
## assignmt    0.62706980 0.005879983 106.644835 0.000000e-
##                CI Upper    DF
## (Intercept) 0.01837536 11201
## assignmt    0.63859560 11201
```

```r
#$\frac{Cov(D,Z)}{Var(Z)}$
vmat <- cov(d[,c("earnings","training","assignmt")])
vmat[3,2]/vmat[3,3]
```

```
## [1] 0.6270698
```

Reduced form/Intent to Treat Effect

```
#regression effect of Z on Y
summary(lm_robust(earnings~assignmt,data=d))$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 15040.504   265.3927 56.67264 0.0000000000 1
## assignmt     1161.417   330.4793  3.51434 0.0004425883
##                DF
## (Intercept) 11201
## assignmt    11201
```

```
#$\frac{Cov(Y,Z)}{Var(Z)}$
vmat[1,3]/vmat[3,3]
```

```
## [1] 1161.417
```

Wald Estimator

- Effect of D on Y using only exogenous variation in D induced by Z:

$$\rho = \frac{\frac{Cov(Y,Z)}{Var(Z)}}{\frac{Cov(D,Z)}{Var(Z)}} = \frac{\text{Reduced form}}{\text{First stage}}$$

$$= \frac{Cov(Y,Z)}{Cov(D,Z)} = \frac{\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{N}(z_i - \bar{z})(D_i - \bar{D})}$$

Estimation

Focusing on the numerator...

$$
\begin{aligned}
\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y}) &= \sum_{i=1}^{N} z_i(y_i - \bar{y}) - (\sum_{i=1}^{N} \bar{z}(y_i - \bar{y})) \\
&= \sum_{i=1}^{N}(z_i y_i - z_i\bar{y}) - \bar{z}(\sum_{i=1}^{N}(y_i - \bar{y})) \\
&= \sum_{z_i=1}(z_i y_i - z_i\bar{y}) - \bar{z}(n\bar{y} - n\bar{y}) \\
&= \sum_{z_i=1}(z_i y_i - z_i\bar{y})
\end{aligned}
$$

The ratio

$$\rho = \frac{\sum_{z_i=1}(z_i y_i - z_i \bar{y})}{n_1} \Big/ \frac{\sum_{z_i=1}(z_i D_i - z_i \bar{D})}{n_1}$$
$$= \frac{\bar{y_1} - \bar{y}}{\bar{D_1} - \bar{D}}$$

Using the fact that $\bar{y} = \frac{n_1 \bar{y_1} + n_0 \bar{y_0}}{n}$

$$\rho = \frac{\bar{y_1} - \bar{y_0}}{\bar{D_1} - \bar{D_0}}$$

Converges in probability to...

$$= \frac{E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0]}{E[D_i|Z_i = 1] - E[D_i|Z_i = 0]}$$

Wald Estimate

```r
#$\frac{Cov(Y,Z)}{Var(Z)} / $\frac{Cov(D,Z)}{Var(Z)}$
(vmat[1,3]/vmat[3,3])/(vmat[2,3]/vmat[3,3])
```

```
## [1] 1852.133
```

Variance

- The asymptotic standard error of the Wald estimates is derived from the limiting distribution of $\sqrt{n}\frac{(\bar{y_1}-\bar{y_0})}{(\bar{D_1}-\bar{D_0})}$.
- The numerator has a nondegenerate limiting distribution, while $(\bar{D_1}-\bar{D_0})$ converges to a constant.
- The standard error is therefore equal to $1/(\bar{D_1}-\bar{D_0})$ times the standard error of the numerator

Standard Error of Wald Estimate

```
#variance of Y1
var1 = var(d$earnings[d$assignmt==1])/(length(d$earnings[d$

#variance of Y0
var0 = var(d$earnings[d$assignmt==0])/(length(d$earnings[d$

#difference in compliance
diffcom = mean(d$training[d$assignmt==1]) - mean(d$trainin

#variance of wald estimate
(var1+var0)^0.5/diffcom
```

```
## [1] 527.0215
```

Test for first stage

- In contrast to OLS, the IV estimator is not unbiased in small (finite) samples even when instrument is perfectly exogenous
- Because of sampling variability in first stage estimation of fitted values, some part of the correlation between errors in first and second stage seeps into 2SLS estimates (correlation disappears in large samples)
- Finite sample bias can be considerable (e.g., 20 - 30%), even when the sample size is over 100,000 if the instrument is weak

Empirical papers typically report first-stage F-statistics

```
library(lmtest,quietly = T)
fs1 <- lm_robust(training~ sex + age2225+age2629+age3035+
                   age3644+age4554+married +assignmt,data=d
fs2 <- lm_robust(training~ sex +age2225+age2629+age3035+
                   age3644+age4554+married,data=d)
waldtest(fs1, fs2)
```

```
## Wald test
##
## Model 1: training ~ sex + age2225 + age2629 + age3035 +
##     married + assignmt
## Model 2: training ~ sex + age2225 + age2629 + age3035 +
##     married
##   Res.Df Df Chisq Pr(>Chisq)
## 1  11194
## 2  11195 -1 11314  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Continuous IV example

- For our example with IV, we will start with AJR (2001) - Colonial Origins of Comparative Development
- Treatment is average protection from expropriation
- Exogenous covariates are dummies for British/French colonial presence
- Instrument is settler mortality
- Outcome is log(GDP) in 1995

Continuous IV example

- For our example with IV, we will start with AJR (2001) - Colonial Origins of Comparative Development
- Treatment is average protection from expropriation
- Exogenous covariates are dummies for British/French colonial presence
- Instrument is settler mortality
- Outcome is log(GDP) in 1995

```
require(foreign,quietly=TRUE)
dat <- read.dta("AJR 2001\\maketable5.dta")
dat <- subset(dat, baseco==1)
```

2SLS Estimator

- Fit first stage and obtain fitted values $E[D|Z]$
- Plug into second stage: $Y = \alpha_0 + \alpha_1 E[D|Z] + \epsilon_i$
- Standard errors incorrect (ignore estimation uncertainty in first stage).
- Canned packages estimate 2SLS in one step

Estimate IV via 2SLS

```
#first stage
first <- lm_robust(avexpr~logem4+f_brit+f_french,dat)

#IV
iv2sls<-iv_robust(logpgp95~avexpr+f_brit+f_french|logem4+f_
```

Examine First Stage

```
summary(first)
```

```
##
## Call:
## lm_robust(formula = avexpr ~ logem4 + f_brit + f_french,
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Low
## (Intercept)   8.7466     0.7639 11.4502 9.909e-17   7.21
## logem4       -0.5344     0.1612 -3.3148 1.559e-03  -0.85
## f_brit        0.6293     0.3740  1.6825 9.766e-02  -0.11
## f_french      0.0474     0.4044  0.1172 9.071e-01  -0.76
##
## Multiple R-squared:  0.3081 ,    Adjusted R-squared:  0.
## F-statistic: 7.762 on 3 and 60 DF,  p-value: 0.0001837
```

Examine Output

```
summary(iv2sls)
```

```
##
## Call:
## iv_robust(formula = logpgp95 ~ avexpr + f_brit + f_frenc
##     f_brit + f_french, data = dat)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Low
## (Intercept)   1.3724     1.6481  0.8327 4.083e-01  -1.92
## avexpr        1.0779     0.2553  4.2214 8.353e-05   0.56
## f_brit       -0.7777     0.3852 -2.0188 4.798e-02  -1.54
## f_french     -0.1170     0.3484 -0.3358 7.382e-01  -0.81
##
## Multiple R-squared:  0.04833 ,   Adjusted R-squared:  0.
## F-statistic: 8.342 on 3 and 60 DF,  p-value: 0.0001011
```

Final example

- We're going to be looking at Ananat (2011) in AEJ
- This study looks at the effect of racial segregation on economic outcomes.
- Outcome: Poverty rate & Inequality (Gini index)
- Treatment: Segregation (level of dismilarity)
  - What percentage of blacks (or nonblacks) would have to move to another census tract in order for the proportion black in equal tract to be constant
  - dism $= 1/2$ $|$(blacks in i /blacks total) - (non blacks in i/nonblacks total)$|$
- Instrument: "railroad division index"
  - herf $=$ 1 - ($\sum$ (Area of Neighborhood i)/ (Area Total))^2
- Main covariate of note: railroad length in a town

```
require(foreign)
d<-read.dta("Ananat 2011\\aej_maindata.dta")
```

Main effects for Black Subsample

```
#OLS
ols <- lm_robust(lngini_b ~ dism1990 +lenper,d)

#first stage for all areas
first.stage <- lm_robust(dism1990~herf+lenper,d)

#IV for gini and poverty
gini.iv <- iv_robust(lngini_b~dism1990+lenper|herf+lenper,d
pov.iv <- iv_robust(povrate_b~dism1990+lenper|herf+lenper,d
```

Base Results

```r
round(summary(ols)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##      0.449      0.095      4.704      0.000      0.260
```

```r
round(summary(first.stage)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##      0.357      0.114      3.139      0.002      0.132
```

```r
round(summary(gini.iv)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##      0.875      0.441      1.982      0.050      0.001
```

```r
round(summary(pov.iv)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##      0.258      0.112      2.302      0.023      0.036
```

Effects for whites

```
ols.v2 <- lm_robust(lngini_w~dism1990+lenper,d)

first.stage.v2 <- lm_robust(dism1990~herf+lenper,d)

gini.iv.v2 <- iv_robust(lngini_w~dism1990+lenper|herf+lenpe

pov.iv.v2 <- iv_robust(povrate_w~dism1990+lenper|herf+lenpe
```

Base Results for White Subsample

```
round(summary(ols.v2)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##     -0.075      0.039     -1.912      0.058     -0.152
```

```
round(summary(first.stage.v2)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##      0.357      0.114      3.139      0.002      0.132
```

```
round(summary(gini.iv.v2)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##     -0.334      0.129     -2.591      0.011     -0.590
```

```
round(summary(pov.iv.v2)$coefficients[2,],3)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)   CI Lower
##     -0.196      0.070     -2.811      0.006     -0.334
```