

# MQE: Economic Inference from Data:

## Module 3: Instrumental Variables

Claire Duquennois

6/9/2020

# You can't always get what you want

Even with fixed effects, certain types of unobservables can still bias our estimates.

For OVB to not be a problem, we want a treatment variable  $x_i$  where we know that there does not exist some omitted variable  $x_{ov}$  such that

►  $cor(x_i, x_{ov}) \neq 0$  and  $cor(y_i, x_{ov}) \neq 0$ .

This is a tall order...

You can't always get what you want

But if you try sometimes,

You can't always get what you want

But if you try sometimes,  
you just might find,

You can't always get what you want

But if you try sometimes,  
you just might find,  
you get what you need: a good instrumental variable.

## An instrument for what?

I am interested in the relationship between  $y$  and  $x_1$ .

The true data generating process looks like this:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

- ▶  $x_1$  and  $x_2$  are uncorrelated with  $\epsilon$
- ▶  $x_1$  and  $x_2$  are correlated with each other such that  $\text{Cov}(x_1, x_2) \neq 0$

So whats the problem?

- ▶ you don't actually observe  $x_2$ .

Uh oh.

## The problem:

The naive approach (but you of course know better than to do this. . .)

Regress  $y$  on just  $x_1$ :

$$y_i = \beta_0 + \beta_1 x_1 + \nu$$

where

$$\nu = \beta_2 x_2 + \epsilon.$$

## Cov Math Rules:

$$\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$$

$$\text{Cov}(X, X) = \text{Var}(X)$$

$$\text{Cov}(cX, Y) = c\text{Cov}(X, Y)$$



## The problem:

$$\begin{aligned}\hat{\beta}_{1,OLS} &= \frac{\text{cov}(x_1, y)}{\text{var}(x_1)} \\&= \frac{\text{cov}(x_1, \beta_0 + \beta_1 x_1 + \nu)}{\text{var}(x_1)} \\&= \frac{\text{cov}(x_1, \beta_0) + \text{cov}(x_1, \beta_1 x_1) + \text{cov}(x_1, \nu)}{\text{var}(x_1)} \\&= \frac{\beta_1 \text{var}(x_1) + \text{cov}(x_1, \nu)}{\text{var}(x_1)} \\&= \beta_1 + \frac{\text{cov}(x_1, \nu)}{\text{var}(x_1)}\end{aligned}$$

$\text{cov}(x_1, \nu) \neq 0$  since

$\text{cov}(x_1, \nu) = \text{cov}(x_1, \beta_2 x_2 + \epsilon) = \beta_2 \text{cov}(x_1, x_2) + \text{cov}(x_1, \epsilon)$  and  
 $\text{cov}(x_1, x_2) \neq 0$

$\Rightarrow \hat{\beta}_{1,OLS}$  is biased.

# All is not lost!

An **instrumental variable** (IV) is a variable that

- ▶ is correlated with the “good” or “*exogenous*” variation in  $x_1$
- ▶ is unrelated to the “bad” or “*endogenous*” or “*related-to- $x_2$* ” variation in  $x_1$ .

## Formally

An IV is a variable,  $z$  that satisfies two important properties:

- ▶  $\text{Cov}(z, x_1) \neq 0$  (the first stage).
- ▶  $\text{Cov}(z, \nu) = 0$  (since  $\text{Cov}(z, x_2) = 0$ ) (the exclusion restriction)

.

# The First Stage

$$\text{Cov}(z, x_1) \neq 0$$

- ▶  $z$  and  $x_1$  are correlated
- ▶ the IV is useless without a first stage.

We are trying to get a  $\hat{\beta}_1$  such that  $E[\hat{\beta}_1] = \beta_1$ . If our instrument is totally unrelated to  $x_1$ , we won't have any hope of using it to get at  $\beta_1$ .

## The exclusion restriction

$$\text{Cov}(z, \nu) = 0$$

- ▶  $z$  has to affect  $y$  **only** through  $x_1$ .
- ▶ Since  $\text{Cov}(z, \nu) = \beta_2 \text{Cov}(z, x_2) + \text{Cov}(z, \epsilon)$   
 $\Rightarrow \text{Cov}(z, x_2) = 0$  and  $\text{Cov}(z, \epsilon) = 0$   
 $\Rightarrow$  all of the influence of  $z$  on  $y$  must be channelled through  $x_1$ .

## The IV estimator

$$\begin{aligned}\hat{\beta}_{1,IV} &= \frac{\text{cov}(z, y)}{\text{cov}(z, x)} \\&= \frac{\text{cov}(z, \beta_0 + \beta_1 x_1 + \nu)}{\text{cov}(z, x_1)} \\&= \beta_1 \frac{\text{cov}(z, x_1)}{\text{cov}(z, x_1)} + \frac{\text{cov}(z, \nu)}{\text{cov}(z, x_1)} \\&= \beta_1 + \frac{\text{cov}(z, \nu)}{\text{cov}(z, x_1)}.\end{aligned}$$

With the exclusion restriction:  $\text{cov}(z, \nu) = 0 \Rightarrow E[\hat{\beta}_{1,IV}] = \beta_1$

Woot Woot! We have an unbiased estimator!

Do the following variables satisfy the properties of a good instrument?

1. You are interested how the length of a prison sentence affects recidivism 5 years later. You instrument with how strict the randomly assigned judge to the case is.
2. You are interested in how college attendance influences wages. You instrument with family home proximity to a college.
3. You are interested in how stress affects blood pressure. You instrument with COVID rates in the local community.

# Some instruments:

An (old) survey of instrumental variables: Angrist and Krueger (2001)

*Table 1*

**Examples of Studies That Use Instrumental Variables to Analyze Data From Natural and Randomized Experiments**

| <i>Outcome Variable</i>       | <i>Endogenous Variable</i>             | <i>Source of Instrumental Variable(s)</i>                    | <i>Reference</i>                      |
|-------------------------------|--|--|---------------------------------------|
| <i>1. Natural Experiments</i> |  |  |                                       |
| Labor supply                  | Disability insurance replacement rates | Region and time variation in benefit rules                   | Gruber (2000)                         |
| Labor supply                  | Fertility                              | Sibling-Sex composition                                      | Angrist and Evans (1998)              |
| Education, Labor supply       | Out-of-wedlock fertility               | Occurrence of twin births                                    | Bronars and Grogger (1994)            |
| Wages                         | Unemployment insurance tax rate        | State laws   | Anderson and Meyer (2000)             |
| Earnings                      | Years of schooling                     | Region and time variation in school construction             | Duflo (2001)                          |
| Earnings                      | Years of schooling                     | Proximity to college   | Card (1995)                           |
| Earnings                      | Years of schooling                     | Quarter of birth   | Angrist and Krueger (1991)            |
| Earnings                      | Veteran status                         | Cohort dummies   | Imbens and van der Klaauw (1995)      |
| Earnings                      | Veteran status                         | Draft lottery number   | Angrist (1990)                        |
| Achievement test scores       | Class size                             | Discontinuities in class size due to maximum class-size rule | Angrist and Lavy (1999)               |
| College enrollment            | Financial aid                          | Discontinuities in financial aid formula                     | van der Klaauw (1996)                 |
| Health                        | Heart attack surgery                   | Proximity to cardiac care centers                            | McClellan, McNeil and Newhouse (1994) |
| Crime                         | Police                                 | Electoral cycles   | Levitt (1997)                         |
| Employment and Earnings       | Length of prison sentence              | Randomly assigned federal judges                             | Kling (1999)                          |
| Birth weight                  | Maternal smoking                       | State cigarette taxes  | Evans and Ringel (1999)               |



# Chasing Unicorns

- ▶  $z$ 's that satisfy the first condition are easy to find, and we can test that  $\text{Cov}(z, x_1) \neq 0$
- ▶  $z$ 's that satisfy the exclusion restriction are rare and we cannot test that  $\text{Cov}(z, \nu) = 0$  since we don't observe  $\epsilon$ .

# Chasing Unicorns

A good IV is not unlike a unicorn. It is quite powerful/magical as it will allow you to recover a consistent estimate of  $\beta_1$  in a situation that was otherwise hopeless.



# Chasing Unicorns

It is also a rare, (some may argue imaginary) beast, that often turns out to be a horse with an optimistic rider (author).



- ▶ be skeptical of instrumental variables regressions
- ▶ be wary of trying them yourself
- ▶ be prepared to convince people the exclusion restriction is satisfied

## A simulation

I generate some simulated data, with properties I fully understand:

The DGP:  $Y$  depends on two variables,  $X_1$  and  $X_2$  such that

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

- ▶  $x_1$  and  $x_2$  covary with  $\text{Cov}(x_1, x_2) = 0.75$
- ▶  $z$  covaries with  $x_1$  such that  $\text{Cov}(x_1, z) = 0.25$
- ▶  $z$  does not covary with  $x_2$  (so  $\text{Cov}(x_2, z) = 0$ ).

# A simulation

```
library(MASS)
library(ggplot2)
library(stargazer)
```

```
##
## Please cite as:

## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
sigmaMat<-matrix(c(1,0.75,0.25,0.75,1,0,0.25,0,1), nrow=3)
sigmaMat
```

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.75 0.25
## [2,] 0.75 1.00 0.00
## [3,] 0.25 0.00 1.00
```

```
set.seed(5000)
ivdat<- as.data.frame(mvrnorm(10000, mu = c(0,0,0),
                                Sigma = sigmaMat))
```

```
names(ivdat)<-c("x_1", "x_2", "z")
cov(ivdat)
```

```
##           x_1           x_2           z
## x_1 1.0036443 0.745103511 0.246358965
## x_2 0.7451035 0.981180725 0.002127948
## z    0.2463590 0.002127948 0.991118191
```

# A simulation

```
ivdat$error<-rnorm(10000, mean=0, sd=1)

#The data generating process
B1<-10
B2<-(-20)

ivdat$Y<-ivdat$x_1*B1+ivdat$x_2*B2+ivdat$error

knitr::kable(head(ivdat))
```

| x_1        | x_2        | z          | error      | Y         |
|------------|------------|------------|------------|-----------|
| -0.1480778 | -0.5218521 | 1.2549131  | -0.0470256 | 8.909239  |
| -0.8240697 | -1.1709528 | -0.7069437 | 0.3014639  | 15.479822 |
| -2.2167131 | -1.7736810 | -0.7409791 | 0.5561088  | 13.862597 |
| 1.2223555  | 0.2629376  | -0.0761498 | 0.2886668  | 7.253469  |
| -0.1527584 | -0.5661063 | -0.6523414 | 0.8400475  | 10.634589 |
| -0.4315813 | -0.7802528 | 1.3235641  | -0.3380069 | 10.951237 |

## A simulation:

```
simiv1<-lm(Y~x_1+x_2, data=ivdat)  
simiv2<-lm(Y~x_1, data=ivdat)
```

How will our estimate of  $\hat{\beta}_1$  in model 2 compare to the true  $\beta$ ?

⇒ Top Hat

## A simulation:

```
cov(ivdat$Y, ivdat$x_2)
```

```
## [1] -12.1676
```

```
cov(ivdat$x_1, ivdat$x_2)
```

```
## [1] 0.7451035
```

How will our estimate of  $\hat{\beta}_1$  in model 2 compare to the true  $\beta$ ?

- ▶  $\text{Cov}(Y, x_2) < 0$
- ▶  $\text{Cov}(x_1, x_2) > 0$
- ▶  $\hat{\beta}_1$  is downward biased



## A simulation:

```
stargazer(simiv1, simiv2, header=FALSE, type='latex', omit.stat = "all", single.row = TRUE)
```

Table 2

| <i>Dependent variable:</i>                  |                    |                   |
|---|--------------------|-------------------|
| Y   |                    |                   |
|   | (1)                | (2)               |
| x_1   | 10.035*** (0.015)  | -4.829*** (0.131) |
| x_2   | -20.021*** (0.015) |                   |
| Constant                                    | -0.023** (0.010)   | 0.195 (0.131)     |
| <i>Note:</i> * p<0.1; ** p<0.05; *** p<0.01 |                    |                   |

- ▶ With the correctly specified model  $E[\hat{\beta}_1] = \beta_1$ .
- ▶ If I do not observe  $x_2$ , the naive approach is biased.

## A simulation:

Suppose there exists a variable  $z$  that satisfies the two conditions outlined above:

- ▶  $\text{Cov}(z, x_1) \neq 0$  (the first stage).
- ▶  $\text{Cov}(z, \nu) = 0$  (the exclusion restriction).

Our simulated data includes  $z$ , a variable with these properties

## A simulation:

I instrument my endogenous variable,  $x_1$ , with my instrument  $z$ :

```
library(lfe)
```

```
## Loading required package: Matrix
```

```
simiv3<-felm(Y~1|0|(x_1~z),ivdat)
```

# A simulation:



```
stargazer(simiv1, simiv2, simiv3, header=FALSE,
           type='latex', omit.stat = c("n", "f", "ser"))
```

Table 3

|                         | <i>Dependent variable:</i> |                      |                     |
|-------------------------|----------------------------|----------------------|---------------------|
|                         | Y                          |                      |                     |
|                         | <i>OLS</i>                 |                      | <i>feIm</i>         |
|                         | (1)                        | (2)                  | (3)                 |
| x_1                     | 10.035***<br>(0.015)       | -4.829***<br>(0.131) |                     |
| x_2                     | -20.021***<br>(0.015)      |                      |                     |
| 'x_1(fit)'              |                            |                      | 9.828***<br>(0.796) |
| Constant                | -0.023**<br>(0.010)        | 0.195<br>(0.131)     | 0.141<br>(0.197)    |
| R <sup>2</sup>          | 0.995                      | 0.119                | -0.981              |
| Adjusted R <sup>2</sup> | 0.995                      | 0.119                | -0.981              |

Note: \* p<0.1; \*\* p<0.05; \*\*\* p<0.01

► I get  
an un-  
biased  
esti-  
mate  
of  $\beta_1$ !

► Careful:  
 $R^2$   
values  
get  
real  
funky  
(negative!?!)-  
don't  
use

## 2SLS:

How does  $\beta_{IV}$  use the instrumental variable to retrieve an unbiased estimate?

To build intuition, let's look at the two-stage least squares (2SLS) estimator  $\beta_{2SLS}$ .

When we are working with only one instrument and one endogenous regressor,  $\beta_{IV} = \beta_{2SLS}$ .

## 2SLS:

2SLS proceeds in two (least squares regression) stages:

- ▶ the “first stage,” a regression of our endogenous variable on our instrument

$$x_1 = \gamma_0 + \gamma_1 z + u.$$

- ▶ using the estimated  $\hat{\gamma}$  coefficients we generate predicted values,  $\hat{x}_1$ :

$$\hat{x}_1 = \hat{\gamma}_0 + \hat{\gamma}_1 z$$

- ▶ the “second stage” where we regress our outcome on the predicted values of the endogenous variable

$$y = \beta_0 + \beta_1 \hat{x}_1 + \epsilon$$

# The first stage:

```
sim2splsfs<-felm(x_1~z,ivdat)
summary(sim2splsfs)
```

```
##
## Call:
##   felm(formula = x_1 ~ z, data = ivdat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1596 -0.6478  0.0059  0.6504  3.4744
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.002705   0.009708   0.279   0.781
## z            0.248567   0.009752  25.488 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9708 on 9998 degrees of freedom
## Multiple R-squared(full model): 0.06101   Adjusted R-squared: 0.06092
## Multiple R-squared(proj model): 0.06101   Adjusted R-squared: 0.06092
## F-statistic(full model):649.7 on 1 and 9998 DF, p-value: < 2.2e-16
## F-statistic(proj model): 649.7 on 1 and 9998 DF, p-value: < 2.2e-16
```

## The second stage:

```
hatgamma0<-sim2slsfs$coefficients[1]  
hatgamma1<-sim2slsfs$coefficients[2]  
ivdat$hatx_1<-hatgamma0+hatgamma1*ivdat$z  
  
sim2slsss<-felm(Y~hatx_1,ivdat)
```



## The second stage:

- Math Magic!

$\hat{\beta}_{1,2SLS}$

consis-

tently

esti-

mates

$\beta_1$  and

$\hat{\beta}_{1,2SLS} =$

$\hat{\beta}_{1,IV}!$

```
stargazer(simiv1, simiv3, sim2slsss, header=FALSE, type='latex',  
omit.stat = "all", no.space=TRUE)
```

Table 4

|   | <i>Dependent variable:</i> |                     |                     |
|---|----------------------------|---------------------|---------------------|
|   | Y                          |                     |                     |
|   | <i>OLS</i>                 | <i>felm</i>         |                     |
|   | (1)                        | (2)                 | (3)                 |
| x_1   | 10.035***<br>(0.015)       |                     |                     |
| x_2   | -20.021***<br>(0.015)      |                     |                     |
| 'x_1(fit)'                                  |                            | 9.828***<br>(0.796) |                     |
| hatx_1                                      |                            |                     | 9.828***<br>(0.557) |
| Constant                                    | -0.023**<br>(0.010)        | 0.141<br>(0.197)    | 0.141<br>(0.138)    |
| <i>Note:</i> * p<0.1; ** p<0.05; *** p<0.01 |                            |                     |                     |

- Note: The standard errors reported from the second stage of 2SLS will not be correct (because they are based on  $\hat{x}_1$  rather than  $x_1$ ). (There are ways to correct this but the math and coding are a bit complicated.)

# The Reduced Form (and more cool IV intuition)

The **reduced form** regresses the outcome directly on the exogenous instrument (and any other exogenous variables if you have them):

$$y_i = \pi_0 + \pi_1 Z_i + \eta$$

```
sim2slsrf<-felm(Y~z,ivdat)
stargazer(simiv3, sim2slsfs, sim2slsrf, type='latex', header=FALSE, omit.stat = "all")
```

Table 5

|  | <i>Dependent variable:</i> |                     |                     |
|--|----------------------------|---------------------|---------------------|
|  | Y                          | x_1                 | Y                   |
|  | (1)                        | (2)                 | (3)                 |
| 'x_1(fit)'                               | 9.828***<br>(0.796)        |                     |                     |
| z  |                            | 0.249***<br>(0.010) | 2.443***<br>(0.138) |
| Constant                                 | 0.141<br>(0.197)           | 0.003<br>(0.010)    | 0.167<br>(0.138)    |
| <i>Note:</i> *p<0.1; **p<0.05; ***p<0.01 |                            |                     |                     |

## The Reduced Form (and more cool IV intuition)

We can recover  $\hat{\beta}_1$  by taking the  $\hat{\pi}_1$  from the reduced form and dividing it by  $\hat{\gamma}_1$  from the first stage:

$$\hat{\beta}_1 = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{2.442}{0.249} = 9.807$$

- ▶ Math Magic!
- ▶ Why does this work?

## The Reduced Form (and more cool IV intuition)

We can recover  $\hat{\beta}_1$  by taking the  $\hat{\pi}_1$  from the reduced form and dividing it by  $\hat{\gamma}_1$  from the first stage:

$$\hat{\beta}_1 = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{2.442}{0.249} = 9.807$$

- ▶ Math Magic!
- ▶ Why does this work? We are taking the effect of  $z$  on  $y$  and scaling it by the effect of  $z$  on  $x_1$  (since  $z$  affects  $y$  via  $x_1$ ).

We saw the good. Now for the bad and ugly.

- ▶ The  
Forbidden  
Regression
- ▶ Weak  
Instruments



# The Bad: The Forbidden Regression

Be weary of the **forbidden regression**!

People sometimes try to run a logit, probit, or some other non-linear regression as the first stage of a 2SLS procedure. This is a bad idea. Don't do it.

## The Ugly: Weak Instruments

Recall that

$$\hat{\beta}_{IV} = \beta + \frac{\text{cov}(z, \nu)}{\text{cov}(z, x_1)}.$$

A weak instrument is an instrument with a weak first stage:  
 $\text{Cov}(z, x_1)$ , is small.

Why is this a problem?

## The Ugly: Weak Instruments

**A weak instrument will amplify any endogeneity that exists in your model.**

For  $E[\hat{\beta}_{IV}] = \beta_1$ , we need  $\text{cov}(z, \nu) = 0$  (the exclusion restriction) to hold.

Suppose this assumption is violated in a small way, meaning that  $\text{cov}(z, \nu) \neq 0$  but that it was a very small value.

If  $\text{cov}(z, x_1)$  is also small, the violation of the exclusion restriction will get amplified leading to potentially severe bias in our estimator.



# The Ugly: Simulation

I generate a simulated dataset with:

- ▶ a week first stage  $\text{cov}(z, x_1) = 0.03$
- ▶ a small violation of the exclusion restriction,  $\text{cov}(z, x_2) = 0.01$

```
sigmaMat<-matrix(c(1,0.75,0.03,0.75,1,0.01,0.03,0.01,1), nrow=3)
sigmaMat
```

```
##      [,1] [,2] [,3]
## [1,] 1.00 0.75 0.03
## [2,] 0.75 1.00 0.01
## [3,] 0.03 0.01 1.00
```

```
set.seed(5000)
ivdatwk<- as.data.frame(mvrnorm(10000, mu = c(0,0,0),
                                Sigma = sigmaMat))
```

```
names(ivdatwk)<-c("x_1", "x_2", "z")
cov(ivdatwk)
```

```
##           x_1           x_2           z
## x_1 0.98285285 0.74457303 0.02472936
## x_2 0.74457303 1.00432318 0.01277121
## z    0.02472936 0.01277121 0.98892020
ivdatwk$error<-rnorm(10000, mean=0, sd=1)
ivdatwk$nu=(-20)*ivdatwk$x_2+ivdatwk$error
```

# The Ugly: Simulation

```
#The data generating process
```

```
B1<-10
```

```
B2<-(-20)
```

```
ivdatwk$Y<-ivdatwk$x_1*B1+ivdatwk$x_2*B2+ivdatwk$error
```

```
simivweakfs<-lm(x_1~z,ivdatwk)
```

```
simivweak<-felm(Y~1|0|(x_1~z),ivdatwk)
```

```
stargazer(simivweakfs,simivweak, type='latex', omit.stat = c("n", "adj.rsq", "rsq", "ser"),  
          header=FALSE)
```

Table 6

|                     | <i>Dependent variable:</i>     |                   |
|---------------------|--------------------------------|-------------------|
|                     | <i>x_1</i>                     | <i>Y</i>          |
|                     | <i>OLS</i>                     | <i>felm</i>       |
|                     | (1)                            | (2)               |
| <i>z</i>            | 0.025**<br>(0.010)             |                   |
| ' <i>x_1</i> (fit)' |                                | -0.415<br>(5.685) |
| Constant            | 0.007<br>(0.010)               | 0.137<br>(0.146)  |
| F Statistic         | 6.295** (df = 1; 9998)         |                   |
| Note:               | * p<0.1; ** p<0.05; *** p<0.01 |                   |

## The Ugly: Simulation

```
cov(ivdatwk$z, ivdatwk$x_1)
```

```
[1] 0.02472936
```

```
cov(ivdatwk$z, ivdatwk$nu)
```

```
[1] -0.2575614
```

We can see that  $\text{cov}(z, x_1) = 0.02473$  and  $\text{cov}(z, \nu) = -0.25756$  so

$$\hat{\beta}_{1, IV} = 10 + \frac{-0.25756}{0.02473} = -0.415 \neq \beta_1 = 10.$$



# The Ugly: Weak Instruments

This is a major problem because:

- ▶ It is rare that an instrument would be perfectly independent of all confounding factors, and
- ▶ it is impossible to test the exclusion restriction.

⇒ be very cautious about results when there is a weak first stage.

What constitutes a “weak” instrument?

The standard benchmark is a first stage F-test that is less than 10 (ie you want the F-stat to be large).

# Dealing with Multiples

See the lecture notes for example of how to deal with more complicated specifications:

- ▶ Control variables
- ▶ Multiple Instruments
- ▶ Multiple endogenous variables and multiple instruments

## Unicorns and Work-horses



The real Siberian unicorn, *Elasmotherium sibiricum*, 29,000 BC.

# Unicorns and Work-horses

IV estimations show up in two different types of situations:

- ▶ IV projects:
  - ▶ the validity of the instrumental variable is central to the identification strategy
  - ▶ can be very interesting because they are often looking at an important but highly endogenous variable
  - ▶ the validity of the causal claims, depends heavily on the validity of the instrument.
- ▶ Cameo appearances in other projects:
  - ▶ in randomized control trials (RCT)
  - ▶ in regression discontinuity (RD) projects
  - ▶ the random assignment of treatment is used as an instrument to estimate treatment effects

## “Work Horse” IV intuition and medical trials (RCT)

Medical trials are a fantastic example of an application of instrumental variables:

- ▶ socially important (perhaps the most important application of IV to date)
- ▶ very clean experimental design

And this is a good segway to the RCT module.



# Medical trials (RCT)

The model for a medical trial:

$$Y_i = \alpha + \tau D_i + \epsilon_i.$$

- ▶  $Y_i$  represents a medical outcome (continuous or discrete).
- ▶  $D_i$  is generally a dummy variable (1 if treated and 0 if not).
- ▶ The error term,  $\epsilon_i$  represents all other factors that affect the health outcome

## Medical trials (RCT)

This regression model corresponds to the potential outcome model with constant treatment effects:

$$Y_i(D_i) = D_i Y(1) + (1 - D_i) Y(0)$$

$$Y_i(0) = \alpha + \epsilon_i$$

$$Y_i(1) = Y_i(0) + \tau.$$

Our goal:

- ▶ estimate the effect that the treatment (say a physical therapy regimen of stretching exercises) has on our outcome (say knee pain)
- ▶ our hope is that  $\tau$  is negative and large in magnitude.

# Medical trials (RCT)

Options: Non-Experimental estimates:

- ▶ Promote the stretching excercises to the general population.
- ▶ Collect some data
- ▶ Regress knee pain on whether or not you did the stretching excercises
- ▶ **What is the problem with this approach?**

# Medical trials (RCT)

Options: Non-Experimental estimates:

- ▶ Promote the stretching excercises to the general population.
- ▶ Collect some data
- ▶ Regress knee pain on whether or not you did the stretching excercises
- ▶ **What is the problem with this approach?**
  - ▶ people who do the stretching excercises are the ones who have high knee pain to begin with!
  - ▶ We will likely get a positive estimate of  $\tau$  (even with controls).

# Medical trials (RCT)

Options: A Medical Trial:

- ▶ randomly assign some patients to the treatment group and others to the control group.
- ▶ Treatment group is given instruction on the stretching exercises and told to do them daily
- ▶ Control group are not.

# Medical trials (RCT)

Back in the old days: the **intention to treat** (or ITT) estimate:

- ▶ Calculate  $\bar{Y}_{i,treat} - \bar{Y}_{i,control}$
- ▶ This is the equivalent to estimating  $Y_i = \alpha + \tau D_i + \epsilon_i$  where  $D_i = 1$  if in the treatment group and 0 if in the control group
- ▶ **intention to treat** because you compare the group that you intended to treat and the group that you do not intend to treat.

**What is the problem with this?**

# Medical trials (RCT)

## **Non-compliance:**

- ▶ some (selected) people in the treatment group would fail to do the stretching exercises
- ▶ some (selected) people in the control group would obtain learn about the stretching exercises from another source (even though they were not supposed to)

Non-compliance can bias the estimate of  $\tau$ .

**How can this bias be corrected?**

# The “Work Horse” IV:

This is actually a simple IV problem:

- ▶ The instrument,  $Z_i$  is the intention to treat:
  - ▶  $Z_i = 1$  if you are assigned to the treatment group (we intend to treat you)
  - ▶  $Z_i = 0$  if you are assigned to the control group (we do not intend to treat you)

**Does  $Z_i$  satisfies the two properties of a good instrument?**



## The “Work Horse” IV:

**Does  $Z_i$  satisfies the two properties of a good instrument?**

- ▶  $Z_i$  is randomly assigned so by construction will be uncorrelated with  $\epsilon_i$  so  $cov(Z_i, \epsilon_i) = 0$  (the exclusion restriction).
- ▶  $Z_i$  is correlated with  $D_i$ , because you are going to be more likely to do the stretches if you are in the treatment group so  $cov(Z_i, D_i) \neq 0$  (the first stage).

$\Rightarrow Z_i$  is a valid instrument for  $D_i$  and the IV estimator gives us a consistent estimate of  $\tau$ , the effect of the stretching excercises on knee pain.

# The “Work Horse” IV: Intuition

## **How does this fix the non-compliance problem?**

Example:

- ▶ Assume the non-compliance problem only exists for the people in the treatment group.
- ▶ Only half the people in the treatment group do the stretching exercises (ie half of the treatment group fails to comply and does not do the stretching exercises while the other half do the exercises as they were told to)

**What will the IV look like?**

## The “Work Horse” IV: Intuition

The first stage will regress whether you did the exercises on whether you were in the treatment group  $D_i$  on  $Z_i$ :

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

Zero people in the control group did the stretches while half in the treatment group did the stretches:

$$\Rightarrow E[\hat{\gamma}_0] = 0 \text{ and } E[\hat{\gamma}_1] = 0.5.$$

## The “Work Horse” IV: Intuition

The IV estimate is the reduced form scaled by the first stage:

The reduced form is a regression of  $Y_i$  (your knee pain) on  $Z_i$  (whether you were assigned to the treatment or control group):

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

Therefore our IV estimate,

$$\hat{\tau}_{IV} = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{\hat{\pi}_1}{0.5}.$$

**How is this fixing the non-complier problem?**

# The “Work Horse” IV: Intuition

## **How is this fixing the non-complier problem?**

The reduced form is estimating the effect that being assigned to the treatment group has on knee pain.

But this is not what we are interested in.

We want to know the effect of doing the stretching exercises.

# The “Work Horse” IV: Intuition

## How is this fixing the non-complier problem?

If there were perfect compliance:

- ▶ the reduced form estimate would be the effect of doing the stretching exercises
  - ▶ the first stage would give us  $\hat{\gamma}_1 = 1$
  - ▶ the IV estimate would be  $\hat{\tau}_{IV} = \frac{\hat{\pi}_1}{1} = \hat{\pi}_1$ .

If compliance isn't perfect:

- ▶ the reduced form estimates the effect of increasing the probability you do the stretching exercises
  - ▶ This means that the reduced form is not estimating the full effect of the stretching exercises

## The “Work Horse” IV: Intuition: Example

- ▶ 10 people in the treatment group. 5 do the exercises and 5 do not.
  - ▶ The (expected) mean knee pain for the treatment group:

$$E[Y_i | TreatGroup_i] = \frac{5\alpha + 5(\alpha + \tau)}{10} = \alpha + \frac{\tau}{2}$$

- ▶ 10 people in the control group, no one does the stretching exercises.
  - ▶ The (expected) mean knee pain for the control group:

$$E[Y_i | ContGroup_i] = \alpha$$

## The “Work Horse” IV: Intuition: Example

- ▶ 10 people in the treatment group. 5 take do the stretches and 5 do not.
- ▶ The (expected) mean knee pain for the treatment group:

$$E[Y_i | TreatGroup_i] = \frac{5\alpha + 5(\alpha + \tau)}{10} = \alpha + \frac{\tau}{2}$$

- ▶ 10 people in the control group, no one does the stretches.
- ▶ The (expected) mean knee pain for the control group:

$$E[Y_i | ContGroup_i] = \alpha$$

So the reduced form coefficient(=the difference of means):

$$\hat{\pi}_1 = \alpha + \frac{\tau}{2} - \alpha = \frac{\tau}{2}$$

is half the effect of taking the excercises (since only half the treated group did them).



## The “Work Horse” IV: Intuition: Example

Thus,

$$E[\tau_{IV}] = \frac{\pi_1}{\gamma_1} = \frac{0.5\tau}{0.5} = \tau.$$

The IV estimate gives us a consistent estimate because it is scaling the reduced form by the first stage.

We re-scale the reduced form because:

- ▶ being in the treatment group only increases your probability of doing the exercises by 50 percentage points not by a full 100 percentage points.
- ▶ the reduced form only represents half of the effect of doing the stretching exercises

## The “Work Horse” IV: Intuition:

Note:

- ▶ the IV is different from simply taking the mean of  $Y_i$  for those in the treatment group who did the stretches and subtracting the mean of  $Y_i$  for those in the control group who did not.
- ▶ This would be affected by the same selection issues as a simple OLS regression of  $Y_i$  on  $D_i$ .

## The “Work Horse” IV: In practice

Recall Arseneaux, Gerber and Green (2006):

Evaluate a “Get out the Vote” mobilization:

- ▶ Who gets called ( $Call_i$ ) is random
- ▶ Who answers the call ( $Contact_i$ ) is not

In the paper they generate:

- ▶ estimates that controlled for observable characteristics
- ▶ “experimental” estimates in order to gauge bias from unobservables

The “experimental” estimates use an instrumental variable.

## The “Work Horse” IV: In practice

They are interested in estimating how getting contacted by the “Get out the Vote” mobilization affect the likelihood of actually voting:

$$Votes_i = \beta_0 + \beta_1 Contacted_i + \beta_j X_j + \epsilon_i.$$

But who gets contacted is not random (not everyone will pick up the phone!)

They instrument  $Contacted_i$ , with being randomly assigned to receive a call from the campaign. Thus the first stage is

$$Contacted_i = \gamma_0 + \gamma_1 Called_i + \gamma_j X_j + u_i$$

# The “Work Horse” IV: In practice

Replicating results of columns 1 of p.49 and p.50:

```
library(haven)

agg_data<-read_dta("../IA_MI_merge040504.dta")

#scaling the vote02 variable to remove excess 0's from tables
agg_data$vote02<-100*as.numeric(agg_data$vote02)

regols1<-felm(vote02~contact+state+comp_mi+comp_ia,agg_data)
regiv1<-felm(vote02~state+comp_mi+comp_ia|0|(contact~treat_real+state+comp_mi+comp_ia),agg_data)

## Warning in chol.default(mat, pivot = TRUE, tol = tol): the matrix is either
## rank-deficient or indefinite
```

# The “Work Horse” IV: In practice

Replicating results of columns 1 of p.49 and p.50:

```
stargazer(regols1, regiv1, type='latex', se = list( regols1$rse, regiv1$rse), header=FALSE,
          omit.stat = "all")
```

Table 7

|   | <i>Dependent variable:</i> |                      |
|---|----------------------------|----------------------|
|   | vote02                     |                      |
|   | (1)                        | (2)                  |
| contact                                     | 6.207***<br>(0.306)        |                      |
| state                                       | 6.671***<br>(0.347)        | 7.388***<br>(0.350)  |
| comp_mi                                     | 4.836***<br>(0.098)        | 4.911***<br>(0.098)  |
| comp_ia                                     | 6.353***<br>(0.177)        | 6.083***<br>(0.178)  |
| 'contact(fit)'                              |                            | 0.360<br>(0.498)     |
| Constant                                    | 46.128***<br>(0.126)       | 46.081***<br>(0.126) |
| <i>Note:</i> * p<0.1; ** p<0.05; *** p<0.01 |                            |                      |