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 <u>sometimes</u>,

you just <u>might</u> 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_i and x_2 are uncorrelated with ϵ
- x_i and x_2 they are correlated with each other such that $Cov(x_1, x_2) \neq 0$

So whats the problem?

- you don't actually observe x_2.

Uh oh.

The problem:

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

Regress y on just x_1:

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

where

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

The problem:

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

$cov(x_1, \nu) \neq 0 \Rightarrow \hat{\beta}_{1,OLS}$ 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:

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

The First Stage

$Cov(z, x_1) \neq 0$

- 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 β_1.

The exclusion restriction

$Cov(z, \nu) = 0$

- z has to affect y **only** through x_1.
- $\Rightarrow Cov(z, \epsilon) = 0$ (because we've already assumed that x_2 is uncorrelated with ϵ).

The IV estimator

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

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

Woot Woot! We have an unbiased estimator!

Chasing Unicorns

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

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 $\hat{\beta}_1$ in a situation that was otherwise hopeless.

Chasing Unicorns

It is also a rare, (some may argue imaginary) beast, that usually turns out to be a horse with an overly 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 are correlated with $Cor(x_1, x_2) = 0.75$
- z is correlated with x_1 such that $Cor(x_1, z) = 0.25$
- z is not correlated with x_2 (so $Cor(x_2, z) = 0$).

A simulation

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

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. 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(3221)
ivdat<- as.data.frame(mvrnorm(10000, mu = c(0,0,0),
                     Sigma = sigmaMat))

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

```
##           x_1          x_2            z
## x_1 1.0000000  0.753135403  0.237314050
## x_2 0.7531354  1.000000000 -0.008862925
## z   0.2373140 -0.008862925  1.000000000
```

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
0.9147512	0.8895539	0.6043023	-0.6100699	-9.253636
1.7837077	0.7578459	1.6537624	-0.9436084	1.736550
-0.6895438	-0.8811583	-0.0728647	1.2663896	11.994118
0.5523528	1.0688671	0.2623387	0.1740300	-15.679784
-2.3233713	-3.1155154	0.6425111	-0.5827891	38.493805
-0.2101972	-0.3597693	0.2199799	-0.6097235	4.483690

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 β?

$\Rightarrow$ Top Hat

A simulation:

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

Table 2

	Dependent variable:	
	Y	
	(1)	(2)
x_1	10.011^{***} (0.015)	-5.233^{***} (0.134)
x_2	-20.009^{***} (0.015)	
Constant	0.016 (0.010)	0.079 (0.134)
Note:		*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:

- $Cov(z, V_1) \neq 0$ (the first stage).
- $Cov(z, \nu) = 0$ (the exclusion restriction).

Our simulated data includes z, a variable with these properties

```r
cor(ivdat$z, ivdat$x_1)
```

```
[1] 0.237314
```

```r
#note: we can test this correlation because I am working with simulated data and observe x_2.
#In the wild x_2 would be unobservable and you would have to argue that this condition holds.

ivdat$nu<-B2*ivdat$x_2+ivdat$error
cor(ivdat$z, ivdat$nu)
```

```
[1] 0.008973809
```

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:

- I get an unbiased estimate of β_1!
- Careful: R^2 values get real funky (negative!?!)– don't use.

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

Table 3

	Dependent variable:		
	Y		
	OLS		*felm*
	(1)	(2)	(3)
x_1	10.011***	−5.233***	
	(0.015)	(0.134)	
x_2	−20.009***		
	(0.015)		
‘x_1(fit)‘			10.766***
			(0.878)
Constant	0.016	0.079	−0.036
	(0.010)	(0.134)	(0.209)
R^2	0.995	0.133	−1.111
Adjusted R^2	0.995	0.133	−1.112
Note:			*p<0.1; **p<0.05; ***p<0.01

2SLS:

How does β_{IV} uses the instrumental variable to retrieve an unbiased estimate?

To build intuition, let's look at the two-stage least squares (2SLS) estimator β_{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:

```r
sim2slsfs<-felm(x_1~z,ivdat)
summary(sim2slsfs)
```

```
##
## Call:
##    felm(formula = x_1 ~ z, data = ivdat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6955 -0.6508 -0.0001  0.6500  3.7093
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.006807   0.009723    0.70    0.484
## z           0.239464   0.009803   24.43   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9723 on 9998 degrees of freedom
## Multiple R-squared(full model): 0.05632   Adjusted R-squared: 0.05622
## Multiple R-squared(proj model): 0.05632   Adjusted R-squared: 0.05622
## F-statistic(full model):596.7 on 1 and 9998 DF, p-value: < 2.2e-16
## F-statistic(proj model): 596.7 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:

```
hatgamma0<-sim2slsfs$coefficients[1]
hatgamma1<-sim2slsfs$coefficients[2]
ivdat$hatx_1<-hatgamma0+hatgamma1*ivdat$z

sim2slsss<-felm(Y~hatx_1,ivdat)
```

The second stage:

```
stargazer(simiv1, simiv2, simiv3,sim2slsss, header=FALSE,  type='latex', omit.stat = "all", no.space=TRUE)
```

Table 4

	Dependent variable:			
	Y			
	OLS		*felm*	
	(1)	(2)	(3)	(4)
x_1	10.011***	−5.233***		
	(0.015)	(0.134)		
x_2	−20.009***			
	(0.015)			
‘x_1(fit)‘			10.766***	
			(0.878)	
hatx_1				10.766***
				(0.595)
Constant	0.016	0.079	−0.036	−0.036
	(0.010)	(0.134)	(0.209)	(0.141)
Note:				*p<0.1; **p<0.05; ***p<0.01

- Math Magic! $\hat{\beta}_{1,2SLS}$ consistently estimates β_1 and $\hat{\beta}_{1,2SLS} = \hat{\beta}_{1,IV}$!
- 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 is 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):

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

```
sim2slsrf<-felm(Y~z,ivdat)
stargazer(sim2slsfs, sim2slsss, sim2slsrf,  type='latex', header=FALSE, omit.stat = "all")
```

Table 5

	Dependent variable:		
	x_1	Y	
	(1)	(2)	(3)
z	0.239*** (0.010)		2.578*** (0.142)
hatx_1		10.766*** (0.595)	
Constant	0.007 (0.010)	−0.036 (0.141)	0.037 (0.141)
Note:			*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:

$$\hat{\beta}_1 = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{2.578}{0.239} = 10.786$$

- 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:

$$\hat{\beta}_1 = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{2.578}{0.239} = 10.786$$

- 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, Poisson 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

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

A weak instrument is an instrument with a weak first stage: $Cov(z, x_1)$, 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 $cov(z, \nu) = 0$ (the exclusion restriction) to hold.

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

If $cov(z, x_1)$ 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 $cov(z, x_1) = 0.03$
- a small violation of the exclusion restriction, $cov(z, x_2) = 0.01$

```
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

```r
#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

	Dependent variable:	
	x_1	Y
	OLS	*felm*
	(1)	(2)
z	0.025^{**}	
	(0.010)	
‘x_1(fit)‘		-0.415
		(5.685)
Constant	0.007	0.137
	(0.010)	(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 $cov(z, x_1) = 0.02473$ and $cov(z, \nu) = -0.25756$ 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.

$\Rightarrow$ be very cautious about results when there is a weak first stage.

What constitutes a “weak” instrument?

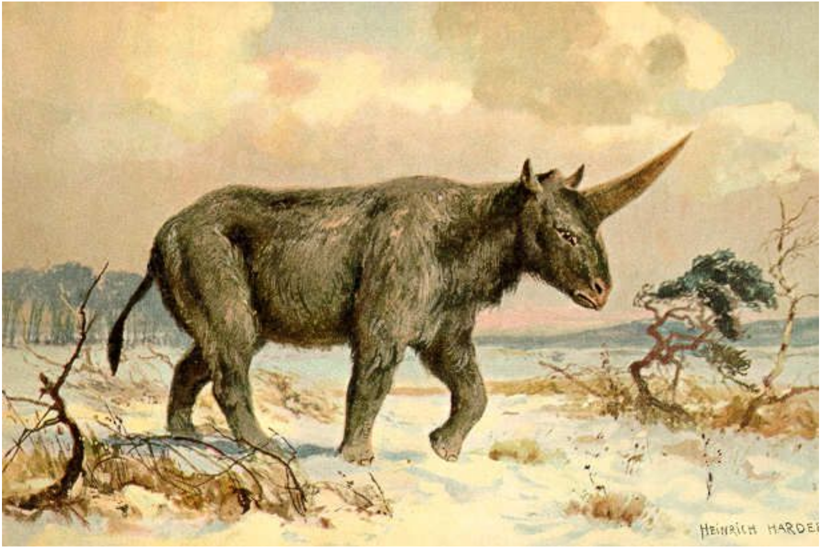
The standard benchmark is a first stage F-test that is less than 10.

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:

$$y_i = \beta_0 + \beta_1 d_i + \epsilon_i.$$

- y_i represents a medical outcome (continuous or discreet).
- d_i is generally a dummy variable (1 if treated and 0 if not).
- The error term, ϵ_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:

$$
\begin{aligned}
y &= dy_1 + (1-d)y_0 \\
y_0 &= \beta_0 + \epsilon \\
y_1 &= y_0 + \beta_1.
\end{aligned}
$$

Our goal:

- estimate the effect that the treatment (say a pill) has on our outcome (say blood pressure)
- our hope is that β_1 is large and negative.

Medical trials (RCT)

Options: Non-Experimental estimates:

- Selling the drug to the general population.
- Collect some data
- Regress blood pressure on whether or not you take the pill
- Problem:
  - people who take the pill are the ones who have high blood pressure to begin with!
  - We will likely get a positive estimate of β_1 (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 the pill and told to take it
- Control group are given a placebo (or nothing at all).

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 $y_i = \beta_0 + \beta_1 d_i + \epsilon$ 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 take the pill
- some (selected) people in the control group would obtain the pill from another source (even though they were not supposed to)

Non-compliance can bias in the estimate of β_1.

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 ν_i so $cov(z_i, \nu_i) = 0$ (the exclusion restriction).
- z_i is correlated with d_i, because you are going to be more likely to take the pill 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 β_1, the effect of taking the pill on blood pressure.