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 <u>sometimes</u>,

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

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

[1] 0.237314

```
#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, $\beta_{IV} = \beta_{2SLS}$.