Lecture Notes: Causal Inference Fall 2020

Claire Duquennois

7/28/2020

Contents

1 Instrumental Variables **1**
- 1.1 Chasing Unicorns: IV Assumptions . . . 2
- 1.2 IV: A simulation . . . 3
- 1.3 2SLS . . . 6
- 1.4 Hints and warnings . . . 8
- 1.5 Dealing with multiples . . . 10
- 1.6 Control variables . . . 10
- 1.7 Unicorns and Work-horses . . . 14
- 1.8 Example: IV in practice . . . 17

1 Instrumental Variables

In the section above we saw that fixed effects can allow us to control for quite a number of different unobservables. However, as the example on crime and unemployment illustrated, there can still be concerns that certain types of unobservable variables could be biasing our estimates. What we would really like is to have 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 $cor(y_i, x_{ov}) \neq 0$. Sadly, 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.

Suppose I am interested in the relationship between y and x_1 but the true data generating process looks like this:

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

where x_i and x_2 are uncorrelated with ϵ but they are correlated with each other such that $Cov(x_1, x_2) \neq 0$, and, drum roll, you don't actually observe x_2. Uh oh. Houston, we have a problem! Even though we don't actually see x_2 we really, really want to estimate β_1!

The naive approach: regress y on just x_1-but you of course know better.[1] That would look like this:

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

where

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

[1] You'd better!

This means that our OLS estimator is biased since

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

x_2 is in our error term since $\nu = \beta_2 x_2 + \epsilon$ - and since x_1 and x_2 are correlated, we've got a problem: $cov(x_1, \nu) \neq 0$. x_1 has become *endogenous* and our OLS estimate of $\hat{\beta}_{1,OLS}$ is biased.

The good news is, you're project is not dead yet. The right IV could get you up and running in no time. An **instrumental variable** (IV) is a variable that drives/is correlated with the "good" or "*exogenous*" variation in x_1, but is unrelated to the "bad" or "*endogenous*" or "*related-to-x_2*" variation in x_1.

1.1 Chasing Unicorns: IV Assumptions

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 condition tells us that z and x_1 are correlated- if this weren't true, the IV would be useless and we would be back at square one: in trouble. Remember, 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.[2]

The second condition, commonly called the "exclusion restriction" says that z has to affect y **only** through x_1. (This also implies that $Cov(z, \epsilon) = 0$, because we've already assumed that x_2 is uncorrelated with ϵ).

With the IV estimator,

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

Since the exclusion restriction gives us that $cov(z, \nu) = 0$ by assumption, $\hat{\beta}_{1,IV} = \beta_1$ and we have an unbiased estimate of β_1.

It turns out that in real life, coming up with z's that satisfy the first condition is trivial-and the good news is that we can easily test the validity of this assumption. Coming up with z's that satisfy the exclusion restriction is extremely difficult. Because we don't observe ϵ, we can never test this assumption. This should make you highly skeptical of anybody doing instrumental variables regressions and wary of trying them yourself. IV is often more art than science and the quality of an IV project will hinge directly on your ability to convince people that the exclusion restriction is satisfied.

[2] If we are interested in how unemployment affects crime, I would not recommend using wind speed over the Pacific ocean as an instrument.

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.

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

1.2 IV: A simulation

To see how using an IV works in practice, let's generate some simulated data, with properties we fully understand as we did in section 1:

The data generating process is as follows. My outcome variable, Y depends on two variables, X_1 and X_2 such that

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

where x_1 and x_2 are correlated with $Cor(x_1, x_2) = 0.75$. In addition, I will also generate a variable, z, that is correlated with x_1 such that $Cor(x_1, z) = 0.25$. but not with x_2.

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

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

```
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
```

```
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
```

I can generate an unbiased estimate such that $E[\hat{\beta}_1] = \beta_1$ by estimating the correctly specified model. If I do not observe x_2, my estimate using the naive approach is biased.

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

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:29 AM

Table 1:

	Dependent variable:	
	Y	
	(1)	(2)
x_1	10.035***	−4.829***
	(0.015)	(0.131)
x_2	−20.021***	
	(0.015)	
Constant	−0.023**	0.195
	(0.010)	(0.131)
Observations	10,000	10,000
R^2	0.995	0.119
Adjusted R^2	0.995	0.119
Residual Std. Error	1.002 (df = 9997)	13.137 (df = 9998)
F Statistic	971,547.600*** (df = 2; 9997)	1,355.943*** (df = 1; 9998)
Note:		*p<0.1; **p<0.05; ***p<0.01

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

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

[1] 0.246359

```
#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

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

[1] -0.04227849

I can instrument my endogenous variable,x_1, with my instrumental variable z using the `felm` function as follows.

```
simiv3<-felm(Y~1|0|(x_1~z),ivdat)
stargazer(simiv1, simiv2, simiv3,  type='latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:29 AM

Table 2:

	Dependent variable:		
	Y		
	OLS		*felm*
	(1)	(2)	(3)
x_1	10.035***	−4.829***	
	(0.015)	(0.131)	
x_2	−20.021***		
	(0.015)		
‘x_1(fit)‘			9.828***
			(0.796)
Constant	−0.023**	0.195	0.141
	(0.010)	(0.131)	(0.197)
Observations	10,000	10,000	10,000
R^2	0.995	0.119	−0.981
Adjusted R^2	0.995	0.119	−0.981
Residual Std. Error	1.002 (df = 9997)	13.137 (df = 9998)	19.704 (df = 9998)
F Statistic	971,547.600*** (df = 2; 9997)	1,355.943*** (df = 1; 9998)	
Note:			*p<0.1; **p<0.05; ***p<0.01

Notice that using the instrumental variable allows me to retrieve an unbiased estimate of β_1, which is pretty

neat.[3]

1.3 2SLS

To build intuition about how the IV estimator, β_{IV} uses the instrumental variable to retrieve an unbiased estimate, I introduce another estimator, 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, not surprisingly, proceeds in two (least squares regression) stages. First, we run the "first stage," a regression of our endogenous variable on our instrument:[4]

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

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

```
##
## 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
```

```
hatgamma0<-sim2slsfs$coefficients[1]
hatgamma1<-sim2slsfs$coefficients[2]
```

We then use the estimated $\hat{\gamma}$ coefficients to generate predicted values, $\hat{x}_1$:

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

```
ivdat$hatx_1<-hatgamma0+hatgamma1*ivdat$z
```

Notice that since z is not correlated with ϵ, our new variable $\hat{x}_1$ is by construction also not correlated with ϵ. We have basically "partialed out" the "bad variation" in x_1 that was *endogenous*, leaving ourselves with only the *exogenous* "good variation" in $\hat{x}_1$.

We can now run the "second stage" where we regress

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

[3] R^2 values get real funky with IV regressions. They can be negative and should not be used for F-tests

[4] if you have other exogenous regressors, you will need to include them in both the first stage and the second stage regressions.

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

stargazer(simiv1, simiv2, simiv3,sim2slsss,  type='latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:29 AM

Table 3:

	Dependent variable:			
	Y			
	OLS		*felm*	
	(1)	(2)	(3)	(4)
x_1	10.035^{***}	-4.829^{***}		
	(0.015)	(0.131)		
x_2	-20.021^{***}			
	(0.015)			
‘x_1(fit)‘			9.828^{***}	
			(0.796)	
hatx_1				9.828^{***}
				(0.557)
Constant	-0.023^{**}	0.195	0.141	0.141
	(0.010)	(0.131)	(0.197)	(0.138)
Observations	10,000	10,000	10,000	10,000
R^2	0.995	0.119	−0.981	0.030
Adjusted R^2	0.995	0.119	−0.981	0.030
Residual Std. Error	1.002 (df = 9997)	13.137 (df = 9998)	19.704 (df = 9998)	13.787 (df =
F Statistic	$971{,}547.600^{***}$ (df = 2; 9997)	$1{,}355.943^{***}$ (df = 1; 9998)		
Note:				*p<0.1; **p<0.05; ***p<

Magical! $\hat{\beta}_{1,2SLS}$ consistently estimates β_1 and $\hat{\beta}_{1,2SLS} = \hat{\beta}_{1,IV}$![5]

1.3.1 The Reduced Form and more cool IV intuition

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

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

In our simulated data we get

```
sim2slsrf<-felm(Y~z,ivdat)

stargazer(sim2slsfs, sim2slsss, sim2slsrf,  type='latex')
```

[5]Note: The standard errors reported from the second stage of 2SLS will not be correct. This is 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.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:29 AM

Table 4:

	Dependent variable:		
	x_1	Y	
	(1)	(2)	(3)
z	0.249***		2.443***
	(0.010)		(0.138)
hatx_1		9.828***	
		(0.557)	
Constant	0.003	0.141	0.167
	(0.010)	(0.138)	(0.138)
Observations	10,000	10,000	10,000
R^2	0.061	0.030	0.030
Adjusted R^2	0.061	0.030	0.030
Residual Std. Error (df = 9998)	0.971	13.787	13.787
Note:			*p<0.1; **p<0.05; ***p<0.01

It turns out that we can recover our estimate of $\hat{\beta}_1$ by taking the $\hat{\pi}_1$from the reduced form and dividing it by $\hat{\gamma}_1$ from the first stage:

$$\hat{\beta}_1 = \frac{\hat{\pi}_1}{\hat{\gamma}_1} = \frac{2.443}{0.249} = 9.807$$

Again, we get the right coefficient on x_1. Why does this work? The reduced form is (essentially) the effect of z on y. What we are doing is 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).[6]

1.4 Hints and warnings

1.4.1 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.

1.4.2 Weak Instruments:

A weak instrument is an instrument with a weak first stage, meaning that the correlation between z, the instrument, and the endogenous variable x_1, $Cov(z, x_1)$, is small. There are several reasons why weak instruments are a problem.

First, a weak instrument will amplify any endogeneity that exists in your model. Recall that

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

[6]Note: this won't work if you have multiple endogenous variables and multiple instruments or additional exogenous variables.

For our IV estimator to return an unbiased estimate of β_1, we need the exclusion restriction, that $cov(z, \nu) = 0$ 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. This wouldn't severely bias our estimates unless we had a weak first stage. 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.

We can see this in a quick simulated example. Below, I generate a simulated dataset with a week first stage $cov(z, x_1) = 0.03$ and a small violation of the exclusion restriction, such that $cov(z, x_2) = 0.01$, and proceed to estimate $\hat{\beta}_{1,IV}$.

```
library(MASS)
library(ggplot2)

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")
ivdatwk$error<-rnorm(10000, mean=0, sd=1)

ivdatwk$nu=(-20)*ivdatwk$x_2+ivdatwk$error
cov(ivdatwk)
```

```
##                x_1          x_2            z        error          nu
## x_1     0.98285285   0.74457303  0.024729361 -0.004001840 -14.8954624
## x_2     0.74457303   1.00432318  0.012771206 -0.019418031 -20.1058816
## z       0.02472936   0.01277121  0.988920198 -0.002137254  -0.2575614
## error  -0.00400184  -0.01941803 -0.002137254  1.003682461   1.3920431
## nu    -14.89546238 -20.10588160 -0.257561369  1.392043072 403.5096751
```

```
#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')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:30 AM

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

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

Since it is rare that an instrument would be perfectly independent of all confounding factors, and it is impossible to test the exclusion restriction, this is a major problem. For this reason, you should be very

Table 5:

	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)
Observations	10,000	10,000
R^2	0.001	0.020
Adjusted R^2	0.001	0.020
Residual Std. Error (df = 9998)	0.991	14.137
F Statistic	6.295** (df = 1; 9998)	
Note:		*p<0.1; **p<0.05; ***p<0.01

cautious about engaging in a project that has a weak first stage.

The standard benchmark for a "weak" instrument is a first stage F-test that is less than 10, though this number should not be taken as iron law. It is common to see papers where the first stage F's are numbers like 10.1. This is usually a sign that someone has been running a lot of regressions.[7]

1.5 Dealing with multiples

Thus far we have kept things simple, working with one endogenous regressor and one instrument. More complicated models can features multiple endogenous regressors and multiple instruments and control variables.

1.6 Control variables

Model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 C_1 + \nu$$

First stage:

$$x_1 = \gamma_0 + \gamma_1 z_1 + \gamma_2 C_1 + u$$

Estimation:

```
library(MASS)
library(ggplot2)

sigmaMat<-matrix(c(1,0.75,0.25,0.2,0.75,1,0,0,0.25,0,1,0,0.2,0,0,1 ), nrow=4)
sigmaMat
```

[7]There is a large econometric literature on the properties of weak instruments. There are also additional problems that come up when running regressions with many weak instruments that we will not discuss here but consider yourself warned.

```
##      [,1] [,2] [,3] [,4]
## [1,] 1.00 0.75 0.25  0.2
## [2,] 0.75 1.00 0.00  0.0
## [3,] 0.25 0.00 1.00  0.0
## [4,] 0.20 0.00 0.00  1.0
```

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

names(ivc)<-c("x_1","x_2","z", "c")
ivc$error<-rnorm(10000, mean=0, sd=1)

ivc$nu=(-20)*ivc$x_2+ivc$error

#The data generating process
B1<-10
B2<-5
B3<-(-20)

ivc$Y<-ivc$x_1*B1+ivc$x_2*B3+B2*ivc$c+ivc$error
```

```
simivc<-felm(Y~c|0|(x_1~z+c),ivc)
```

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

```
stargazer(simivc,  type='latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:30 AM

Table 6:

	Dependent variable:
	Y
c	4.624^{***}
	(0.271)
‘x_1(fit)‘	11.115^{***}
	(0.860)
Constant	-0.024
	(0.210)
Observations	10,000
R^2	-0.772
Adjusted R^2	-0.773
Residual Std. Error	20.982 (df = 9997)
Note:	*p<0.1; **p<0.05; ***p<0.01

1.6.1 Multiple instruments

A word of caution: be cautious about using multiple weak instruments. Even if jointly they give you a strong first stage they can still generate substantial bias in $\hat{\beta}_{IV}$.

Model:

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

First stage:

$$x_1 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + u$$

Estimation:

```
library(MASS)
library(ggplot2)

sigmaMat<-matrix(c(1,0.75,0.25,0.5,0.75,1,0,0,0.25,0,1,0.3,0.5,0,0.3,1 ), nrow=4)
sigmaMat
```

```
##      [,1] [,2] [,3] [,4]
## [1,] 1.00 0.75 0.25  0.5
## [2,] 0.75 1.00 0.00  0.0
## [3,] 0.25 0.00 1.00  0.3
## [4,] 0.50 0.00 0.30  1.0
```

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

names(ivmi)<-c("x_1","x_2","z_1", "z_2")
ivmi$error<-rnorm(10000, mean=0, sd=1)

ivmi$nu=(-20)*ivmi$x_2+ivmi$error

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

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

```
simivmifs<-felm(x_1~z_1+z_2,ivmi)
simivmi<-felm(Y~1|0|(x_1~z_1+z_2),ivmi)
stargazer(simivmifs, simivmi,  type='latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:30 AM

1.6.2 Multiple endogenous variables and multiple instruments

It is possible to estimate models that include several endogenous variables. For things to go well though, you will want to have at least as many instruments as there are endogenous variables in your model (otherwise the model is *under identified* and you will not be able to estimate all of your coefficients).

Model:

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \nu$$

Table 7:

	Dependent variable:	
	x_1	Y
	(1)	(2)
z_1	0.101***	
	(0.009)	
z_2	0.476***	
	(0.009)	
‘x_1(fit)‘		10.120***
		(0.388)
Constant	−0.006	−0.099
	(0.009)	(0.200)
Observations	10,000	10,000
R^2	0.265	−0.998
Adjusted R^2	0.265	−0.998
Residual Std. Error	0.858 (df = 9997)	20.009 (df = 9998)
Note:		*p<0.1; **p<0.05; ***p<0.01

First stage:

$$x_1 = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + u x_3 = \lambda_0 + \lambda_1 z_1 + \lambda_2 z_2 + u$$

Estimation:

```
library(MASS)
library(ggplot2)

sigmaMat<-matrix(c(1,0.75,0.25,0.1,0.2,
                   0.75,1,0,0,0.4,
                   0.25,0,1,0.3,0.15,
                   0.1,0,0.3,1,0.35,
                   0.2,0.4,0.15,0.35,1), nrow=5)
sigmaMat
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,] 1.00 0.75 0.25 0.10 0.20
## [2,] 0.75 1.00 0.00 0.00 0.40
## [3,] 0.25 0.00 1.00 0.30 0.15
## [4,] 0.10 0.00 0.30 1.00 0.35
## [5,] 0.20 0.40 0.15 0.35 1.00
```

```
set.seed(5500)
ivme<- as.data.frame(mvrnorm(10000, mu = c(0,0,0,0,0),
                     Sigma = sigmaMat))

names(ivme)<-c("x_1","x_2","z_1", "z_2","x_3")
ivme$error<-rnorm(10000, mean=0, sd=1)

ivme$nu=(-20)*ivme$x_2+ivme$error
```

```
#The data generating process
B1<-10
B2<-(-20)
B3<-(-30)

ivme$Y<-ivme$x_1*B1+ivme$x_2*B2+ivme$x_3*B3+ivme$error
```

```
simivmefs1<-felm(x_1~z_1+z_2,ivme)
simivmefs2<-felm(x_3~z_1+z_2,ivme)
#Underidentified
simivmeunder1<-felm(Y~1|0|(x_1|x_3~z_2),ivme)
```

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

```
simivmeunder2<-felm(Y~1|0|(x_1|x_3~z_1),ivme)
```

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

```
simivme<-felm(Y~1|0|(x_1|x_3~z_1+z_2),ivme)
stargazer(simivmefs1,simivmefs2,simivmeunder1,simivmeunder2, simivme,  type='latex')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Wed, Oct 28, 2020 - 10:21:30 AM

Table 8:

	Dependent variable:				
	x_1	x_3		Y	
	(1)	(2)	(3)	(4)	(5)
z_1	0.239***	0.048***			
	(0.010)	(0.010)			
z_2	0.030***	0.352***			
	(0.010)	(0.010)			
‘x_1(fit)‘				−8.304***	10.162*
				(1.492)	(0.939
‘x_3(fit)‘			−27.630***		−30.399
			(0.420)		(0.635
Constant	−0.013	0.002	−0.109	−0.037	0.045
	(0.010)	(0.009)	(0.153)	(0.368)	(0.201
Observations	10,000	10,000	10,000	10,000	10,000
R^2	0.061	0.134	0.840	0.079	0.726
Adjusted R^2	0.061	0.134	0.840	0.079	0.726
Residual Std. Error	0.974 (df = 9997)	0.933 (df = 9997)	15.302 (df = 9998)	36.732 (df = 9998)	20.021 (df =
Note:					*p<0.1; **p<0.05; ***p

As you can see, the under identified models cannot estimate coefficients for all of your endogenous variables.

To get estimates for all of the coefficients in your model, you need to include both instruments.

1.7 Unicorns and Work-horses

IV estimations show up in two different types of situations.

There are IV projects. These are projects in which the validity of the instrumental variable is central to the identification strategy in the paper. These projects can be very interesting because they are often looking at an important but highly endogenous variable and then arguing that they have a valid instrument to make causal statements about said endogenous variable. The validity of the causal claims, however, depends **heavily** on the validity of the instrument. And valid instruments are hard to find.

IV estimations also make cameo appearances in many other types of projects, namely in randomized control trials (RCT) and in regression discontinuity (RD) projects. In these scenarios, researchers use the random assignment of treatment as an instrument to estimate treatment effects on certain groups of subjects.

1.7.1 IV intuition and medical trials

This section is a good segway into our section on randomized control trials. It is also a good way to see how "work horse" IV's are used as well as build intuition on what exactly an IV is doing.

For a variety of reasons, medical trials are a fantastic example of an application of instrumental variables. First, they are socially important (perhaps the most important application of IV to date). Furthermore, they are very clean in terms of experimental design, so they make a great teaching example for conveying the intuition behind what the IV estimator is doing.

The model for a medical trial is the same simple regression model that we are accustomed to:

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

In this case, Y_i represents a medical outcome, which could either be a continuous variable such as blood pressure, joint pain or a discreet variable such as survival. The variable D_i is generally a dummy variable that is 1 if you receive the treatment and 0 if you do not. The error term, ϵ_i represents all other factors that affect the health outcome. Note that this regression model corresponds to the potential outcome model with constant treatment effects

$$\begin{aligned} Y_i(D_i) &= D_i Y(1) + (1 - D_i)Y(0) \\ Y_i(0) &= \alpha + \epsilon_i \\ Y_i(1) &= Y_i(0) + \tau. \end{aligned}$$

Let Y_i be knee pain and D_i represent a theraputic stretching excercise that is designed to decrease knee pain, such that $D_i = 1$ if individual i does the stretching excercises and $D_i = 0$ if individual i does not. Our goal is to estimate the effect that the stretching excercises have on lowering knee pain-our hope is that τ is large and negative.

One way to estimate the effect is to start selling promoting the excercises to the general population and then collect some data and run a regression of knee pain on whether or not you do the stretching excercises. However, this estimate will clearly suffer from a selection issue- people who do the stretching excercises are the ones who have knee pain to begin with! We will likely get a positive estimate of τ from this procedure, even if the true τ is large and negative. This may be true even if we condition on observable covariates. Therefore in order to accurately estimate τ, we design a medical trial in which we randomly assign some patients to the treatment group and others to the control group. The patients in the treatment group are instructed on the stretching excercises and told to do them regularly, while the patients in the control group are not.

Back in the old days, people estimated the effect of a therapy by simply subtracting the mean of Y_i for the control group from the mean of Y_i for the treatment group (which is the equivalent to regressing Y_i on a

variable that is 1 if you are assigned to the treatment groups and 0 if you are in the control group). This is what is known as an **intention to treat** (or ITT) analysis, because you are taking the difference between the group that you intended to treat and the group that you do not intend to treat. But there was the problem of **non-compliance**- some people in the treatment group would fail to do the stretching excercises and others in the control group would get instruction on how to do the stretching excercises from another source, even though they were not supposed to. This non-compliance can cause bias in the estimate of τ and it was not immediately clear how to fix this bias until it became obvious that what we were looking at was actually a simple IV problem.

In this case, the instrument, Z_i is the intention to treat, ie $Z_i = 1$ if you are assigned to the treatment group (we intend to treat you), and $Z_i = 0$ if you are assigned to the control group (we do not intend to treat you). It is easy to see that z_i satisfies the two properties of a good instrument. First, Z_i is randomly assigned so by construction will be uncorrelated with ν_i so $cov(Z_i, \nu_i) = 0$. Second, 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 $cov(Z_i, D_i) \neq 0$. Therefore, Z_i is a valid instrument for D_i and the IV estimator gives us a consistent estimate of τ, the effect of taking the stretching excercises on knee pain.

How does this fix the non-compliance problem? To facilitate understanding, assume the non-compliance problem only exists for the people in the treatment group. That is to say, assume that only half the people in the treatment group do the stretching excercises (ie half of the treatment group fails to comply and does not do the stretching excercises while the other half do do the stretching excercises as they were told to). What will the IV look like?

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

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

Since zero people in the control group did the stretching excercises while half in the treatment group did stretching excercises, it should be intuitive that our estimate for $E[\hat{\gamma}_0] = 0$ and $E[\hat{\gamma}_1] = 0.5$.

Now recall that the IV estimate is the reduced form scaled by the first stage. In this case, 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). So the reduced form is

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

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? Well, we know that the reduced form estimates the causal effect of the instrument on Y_i, so in our case the reduced form is estimating the effect that being assigned to the treatment group has on knee pain. If there were a perfect correlation between being assigned to the treatment group and doind the stretching excercises (ie everyone complies with their treatment assignment), then the reduced form estimate would be the effect of the stretching excercises because the first stage would give us $\hat{\gamma}_1 = 1$ and the IV estimate would be $\hat{\tau}_{IV} = \frac{\hat{\pi}_1}{1} = \hat{\pi}_1$.

In our case however, the correlation is not perfect so the reduced form is estimating the effect on your knee pain of increasing the probability that you do the stretching excercises by 50 percentage points. This means that the reduced form is not estimating the full effect of the stretching excercises, but rather half of the effect of the stretching excercises.

For example: suppose there are 10 people in the treatment group. 5 do the stretching excercises and 5 do not. Of the 10 people in the control group, no one does the stretching excercises. The (expected) mean knee pain for the treatment group will be $\frac{5\alpha+5(\alpha+\tau)}{10} = \alpha + \frac{\tau}{2}$, while the (expected) mean knee pain for the control group will be just α. So the reduced form coefficient $\hat{\pi}_1$ will be the difference of means between the treatment and control groups, or $\frac{\tau}{2}$. This, of course, is half the effect of the stretching excercises.

Therefore, $E[\tau_{IV}] = \frac{\pi_1}{\gamma_1} = \frac{0.5\tau}{0.5} = \tau$ which is exactly what we want. We can see that the IV estimate gives us a consistent estimate precisely because it is scaling the reduced form by the first stage. In this example what

this means in practice is that we are re scaling the reduced form to account for the fact that being in the treatment group only increases your probability of doind the stretching excercises by 50 percentage points not by a full 100 percentage points. So the reduced form only represents half of the effect of the stretching excercises and it must be re-scaled by (divided by) 0.5 in order to estimate the full effect of the stretching excercises.

Another important clarifying point. Note how IV is different from simply taking the mean of Y_i for those in the treatment group who did the stretching excercises and subtracting the mean of Y_i for those in the control group who did not do the stretching excercises. The estimator I just described, ie the naive estimator, is affected by the same selection issues as a simple OLS regression of Y_i on D_i. Specifically, it may be the case that the people in the treatment group who choose not to do the stretching excercises do so because their knee pain was not very severe to begin with. Thus the group of people that actually did the stretching excercises are the ones that all had severe knee pain to begin with, and we will tend to estimate that the stretching excercises does not have much of an effect.

The IV estimator does not suffer from this selection problem because it does not release the people in the treatment group who choose not to do the stretching excercises. To understand this, imagine for the moment that there are two types of people in our sample: high knee pain types and low knee pain types. Assume that they occur with equal frequency, so that when we randomly assign our sample to the treatment and control groups, half of the treatment group is high knee pain, half of the treatment group is low knee pain, half of the control group is high knee pain, and half of the control group is low knee pain. The half of the treatment group that does the stretching excercises all have high knee pain, so when we apply the naive estimator and compare their average knee pain to the average knee pain of the control group, we underestimate the effect of the stretching excercises because we are comparing a group of high knee pain people (who did the stretching excercises) to a group that is a 50/50 mix of high knee pain and low knee pain people (who did not do the stretching excercises). In contrast, what IV does is compare the mean of the treatment group (which is half high knee pain people and half low knee pain people) to the mean of the control group (which is half high knee pain people and half low knee pain people) in the reduced form. It then re scales this difference in means by the first stage to account for the fact that not all of the treated group did the stretching excercises. So unlike the naive estimator, which deceptively compares a high knee pain group to a half-high/half-low knee pain group, IV compares two comparable groups, and that is why it gives us a consistent estimate of the effect of the stretching excercises.

1.8 Example: IV in practice

We have actually already encountered an IV estimate in these notes. Recall the section on the Arseneaux, Gerber and Green (2006) paper. In this paper, they used data from a large-scale voter "Get out the Vote" mobilization effort that randomly calls households and encourages them to vote. They generated several different estimates that controlled for observable characteristics and compared these estimates to "experimental" estimates in order to gauge bias that was generated by unobservables. These "experimental" estimates were generated using an instrumental variable. They are interested in estimating how getting contacted by the "Get out the Vote" mobilization affect the likelihood of actually voting:

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

where X_j is a vector of exogenous control variables that control for the sampling group the observation is drawn from (which is based on the state and whether they are voting in a competitive race).

Who gets contacted however is not random, as not everyone will pick up the phone. They instrument $Contacted_i$, the endogenous explanatory variable, with whether that household was randomly assigned to receive a call from the campaign. Thus the first stage is

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

Notice that the way this works is very similar to the hypothetical medical trial example described in the previous section.

The code and results are replicated below.

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 3.6.3
```

```
library(here)
```

```
## Warning: package 'here' was built under R version 3.6.3
```

```
## here() starts at C:/Users/Claire/Dropbox/MQE_Causal_Inf/MQE_Causal_Inf
```

```
library(lfe)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:MASS':
##
##     select
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
agg_data<-read_dta("../../data/data_M3_IV/IA_MI_merge040504.dta")
nrow(agg_data)
```

```
[1] 2474927
```

```
#scalling 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
```

```
stargazer(regols1,regiv1,type='latex', se = list( regols1$rse, regiv1$rse), header=FALSE)
```

It is clear that the OLS estimates are substantially biased due to selection of who picks up the phone.

Table 9:

	Dependent variable:	
	vote02	
	(1)	(2)
contact	6.207***	
	(0.306)	
state	6.671***	7.388***
	(0.347)	(0.350)
comp_mi	4.836***	4.911***
	(0.098)	(0.098)
comp_ia	6.353***	6.083***
	(0.177)	(0.178)
‘contact(fit)‘		0.360
		(0.498)
Constant	46.128***	46.081***
	(0.126)	(0.126)
Observations	1,905,320	1,905,320
R^2	0.012	0.012
Adjusted R^2	0.012	0.012
Residual Std. Error (df = 1905315)	49.486	49.491
Note:	*p<0.1; **p<0.05; ***p<0.01	