Econometrics

Chapter 12: Instrumental Variables

Recall some possible big problems for OLS:
1. Omitted variable bias - in case we cannot find adequate controls
2. Measurement error in our X
3. Simultaneity

All lead to $E(u|X) \neq 0$ and we will have biased estimates of β

Set up

Say we care about the following relationship:

$Y_i = \beta_0 + \beta_1 X_i + u_i$

But X and u are correlated.

What we will do is use an additional variable Z, an "instrumental" variable to isolate the part of X that is *not* correlated with u.

Endogenous variables - variables correlated with the error term

Exogenous variables - variables uncorrelated with the error term

We need our variable Z to meet two conditions in order for it to be a valid instrument:

1. Instrument relevance: $corr(Z_i, X_i) \neq 0$
2. Instrument Exogeneity: $corr(Z_i, u_i) = 0$

If an instrument is relevant, then some of its variation is related to variation in X.
If it is also exogenous, then the variation of X captured by Z is also exogenous.
We use this to get an estimate of β_1.

Two Staged Least Squares Estimator

Step 1: Find the part of X that is uncorrelated with u - ie. the part that is correlated with Z.

How? OLS:

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

Since Z is exogenous, so is $\pi_0 + \pi_1 Z_i$ - and this is also an estimator of X_i
(hopefully a 'good' one, we will come back to this).

Get predictions of X_i, $\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$.

Step 2: Replace X_i with $\hat{X}_i$ in our main regression:

$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$

Now since $\hat{X}_i$ is exogenous, then our OLS assumptions hold and we can estimate β_1.

This estimator of β_1 is denoted: $\hat{\beta}_1^{TSLS}$. And it is consistent.

Another Way to See this

$$\begin{aligned} cov(Y_i, Z_i) &= cov(\beta_0 + \beta_1 X_i + u_i, Z_i) \\ &= cov(\beta_0, Z_i) + cov(\beta_1 X_i, Z_i) + cov(u_i, Z_i) \\ &= 0 + \beta_i cov(X_i, Z_i) \end{aligned}$$

Now solve for: $\beta_1 = \frac{cov(Y_i,Z_i)}{cov(X_i,Z_i)}$, and replace with sample covariances.

One more way to see this

$$X_i = \pi_0 + \pi_1 Z_i + v_i$$
$$Y_i = \gamma_0 + \gamma_1 Z_i + w_i$$

So a change of one unit of Z gives a change in X of π_1 units and a change in Y of γ_1 units. And this change in X is exogenous because Z is.

Now this means a exogenous change in X of π_1 units is associated with a change in Y of γ_1 units. So the effect on Y of a change of one unit of X is $\gamma_1/\pi_1 = \beta_1$.

Example: Studying and GPA

Taken from:
Stinebrickner, Ralph and Stinebrickner, Todd R. (2008) “The Causal Effect of Studying on Academic Performance,” *The B.E. Journal of Economic Analysis & Policy* Vol. 8: Iss. 1 (Frontiers), Article 14.

Data:
n = 210 Freshman at Berea College in Kentucky U.S. in 2001
Y = first semester GPA (4 pt scale)
X = Average study hours per day (taken from a survey)
Z = 1 if their roommate brought a video game console to the dorm, = 0 otherwise (roommates are randomly assigned).

Do you thing there is any OVB if we just regress Y on X?

Is Z likely a relevant instrument (related to X)?

Is Z likely exogenous (uncorrelated with u)?

$$\begin{aligned} X &= \pi_0 + \pi_1 Z + v_i \\ Y &= \gamma_0 + \gamma_1 Z + w_i \end{aligned}$$

Findings:

$\hat{\pi}_1 = -0.668$; $ga\hat{m}ma_1 = -0.241$
$\rightarrow \hat{\beta}^{IV} = \frac{\hat{\gamma}_1}{\hat{\pi}_1} = \frac{-0.241}{-0.668} = 0.360$

Do these make sense? What are the units?

TSLS is consistent

We noted above that we can write our estimator (when there is a single X and single Z) as:

$\hat{\beta}^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$

The sample covariances are consistent for the actual covariances, so:

$\hat{\beta}^{TSLS} = \frac{s_{YZ}}{s_{XZ}} \rightarrow \frac{cov(Y,Z)}{cov(X,Z)} = \beta_1$

The fact that the instrument in relevant makes sure the denominator is not zero.

And in large samples the TSLS estimator is normally distributed.

Note: usually don't actually do TSLS in two stages, but all at once. Moreover the SEs from the second stage are not right - need SEs specifically that take into account the 1st stage.

General IV Regression

Want to extend the above to include:
1. Additional control variables (ie. other exogenous variables in our equation of interest)
2. Possibly multiple instruments (ie. maybe have more than one valid IV)
3. Multiple endogenous variables? (be very careful, and not advisable)

Included exogenous variables (call them W) - the additional X variables in main equation that are not the problematic endogenous X
A similar issue arises whether these are exogenous ($E(u|W) = 0$) or if they are just 'controls' that ensure the IV is exogenous ($E(u|Z, W) = E(u|W)$). We will focus on the former here.

If # IVs = # Endogenous variables (example before had 1 and 1), then the model is exactly identified

If # IVs > # Endogenous variables, then the model is over identified

If # IVs < # Endogenous variables, then the model is under identified - then you cannot get an answer

The interest is in:

$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + ... + \beta_{1+r} W_{r+1} + u_i$

So the Ws are exogenous but X is correlated with the error term.

Say we have multiple IVs - say m Zs

1st Stage: $X_i = \pi_0 + \pi_1 Z_{1i} + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + ... + \pi_{m+r} W_{ri} + v_i$

Use this to get predicted values of X; $\hat{X}_i$

2nd Stage: $Y_i = \beta_0 + \beta_1 \hat{X}_i + \beta_2 W_{1i} + ... + \beta_{1+r} W_{r+1} + u_i$

So in first stage regress X on all Ws and Zs.

Can extend to multiple endogenous X's. Need at least as many Zs, and each gets its own 1st stage - each X regressed on all Ws and all Zs.
But again, this is something to be cautioned against - why are you trying to deal with 2 problems at once?

And as noted above, in the software it does not actually do two separate stages.

Instrument relevance and exogeneity are still the same in this case.

Instrument Validity

You need strong instruments - ie. they are good predictors of X.

If your instruments are weak, then the normal distribution is not a good approximation for your estimator even in large samples, and you will end up with biased (in direction of OLS) estimators with large SEs.

Rule of Thumb - you want your F-stat from excluding all instruments in 1st stage to be bigger than 10 (at least).

If you have multiple instruments - less is better. Meaning a few strong ones is better than a few strong ones and a few weak ones.

Are your instruments exogenous? Good question.

Say you have a single endogenous variable. Now say you have a single IV. Then there is no way to test if this is exogenous. You have to rely on theory.

However, if you now have 2 IVs, one of which you *know* is exogenous, you can test whether the second is also. This is called a J-statistic (see book for details)

Where do IVs come from?

There is a famous line from a Steve Martin Movie:

“I’m going to show you how to make $1,000,000 tax free!”

“OK, first get $1,000,000....”

This is relevant here. IVs have a nice theory, and they work, but finding a good one is very hard.