Instrumental Variables

Christopher Taber

Department of Economics
University of Wisconsin-Madison

April 12, 2010

Lets go back to thinking about the causal model with the simple regression model

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

To remind you what we did we started with the condition

$$E\left(u_i \mid x_i\right) = 0$$

which yielded the two conditions

$$\begin{aligned} 0 &= E(x_i u_i) \\ &= E\left(x_i\left(y_i - \beta_0 - \beta_1 x_i\right)\right) \\ 0 &= E\left(u_i\right) \\ &= E\left(y_i - \beta_0 - \beta_1 x_i\right) \end{aligned}$$

We then wrote down the sample analogue of these equations:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N} x_i\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) &= 0\\ \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) &= 0\end{aligned}$$

and derived the estimator.

However, as we have talked about throughout the course, we are very nervous about the assumption that $E(u_i \mid x_i) = 0$.

Lets think about some examples

Returns to schooling again

$$\log(W_i) = \beta_0 + \beta_1 S_i + A_i + u_i$$

where A_i is ability

What are the issues with this?

Openess and Inflation

Romer thinks more open economies have lower inflation rates

$$inf_i = \beta_0 + \beta_1 open_i + \beta_2 log(pcinc_i) + u_i$$

Problem is that openness is endogenous and could depend on inflation

Smoking and Birth Weight again

Lets go back to the smoking and birth weight example

$$bwght_i = \beta_0 + \beta_1 packs_i + u_i$$

Does it make sense to assume that packs is uncorrelated with u_i?

We controlled for a bunch of stuff, but was that enough?

Fixing this problem is the most important thing in econometrics. So what can we do?

Suppose we have some other variable, call it z_i and we believe that

$$E[u_i \mid z_i] = 0$$

This is called an **instrumental variable**

It turns out that we can do essentially the same thing that we did before.

We get the two equations

$$
\begin{aligned}
0 &= E(z_i u_i) \\
&= E\left(z_i\left(y_i - \beta_0 - \beta_1 x_i\right)\right) \\
0 &= E\left(u_i\right) \\
&= E\left(y_i - \beta_0 - \beta_1 x_i\right)
\end{aligned}
$$

So the intuitive way to estimate this model would be to use the two equations

$$
\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} z_i\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) &= 0 \\
\frac{1}{N}\sum_{i=1}^{N} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) &= 0
\end{aligned}
$$

Lets see where this goes.

The second equation is the same as before:

$$\begin{aligned}
0 &= \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} y_i - \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta}_0 - \frac{1}{N}\sum_{i=1}^{N}\widehat{\beta}_1 x_i
\end{aligned}$$

$$\frac{1}{N}\sum_{i=1}^{N}\widehat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i - \widehat{\beta}_1 \frac{1}{N}\sum_{i=1}^{N} x_i$$

or

$$\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1 \bar{x}$$

Just as before.

Now the new equation

$$
\begin{aligned}
0 &= \frac{1}{N}\sum_{i=1}^{N} z_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} z_i \left(y_i - \left[\bar{y} - \widehat{\beta}_1 \bar{x}\right] - \widehat{\beta}_1 x_i\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} z_i \left([y_i - \bar{y}] - \widehat{\beta}_1 [x_i - \bar{x}]\right) \\
&= \frac{1}{N}\sum_{i=1}^{N} z_i (y_i - \bar{y}) - \frac{1}{N}\sum_{i=1}^{N} z_i \left(\widehat{\beta}_1 [x_i - \bar{x}]\right)
\end{aligned}
$$

This can be written as

$$\frac{1}{N}\sum_{i=1}^{N} z_i\left(y_i-\bar{y}\right)=\widehat{\beta}_1\frac{1}{N}\sum_{i=1}^{N} z_i\left(x_i-\bar{x}\right)$$

or

$$\begin{aligned}\widehat{\beta}_1 &= \frac{\frac{1}{N}\sum_{i=1}^{N} z_i\left(y_i-\bar{y}\right)}{\frac{1}{N}\sum_{i=1}^{N} z_i\left(x_i-\bar{x}\right)}\\ &= \frac{\frac{1}{N}\sum_{i=1}^{N}\left(z_i-\bar{z}\right)\left(y_i-\bar{y}\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(z_i-\bar{z}\right)\left(x_i-\bar{x}\right)}\end{aligned}$$

Note that this is just like OLS when

$$z_i = x_i$$

Look at the bottom term $\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})$

Notice that this is basically the sample covariance between z_i and x_i

If this is close to zero we are going to be dividing by something close to zero which is not a good idea

Thus we need our instrument to really have two properties:

- It should be uncorrelated with u_i
- It should be correlated with z_i and the higher correlated it is, the better

Lets think about whether this estimator is consistent

First notice that we can rewrite

$$
\begin{aligned}
\widehat{\beta}_1 &= \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\left[\beta_0 + \beta_1 x_i + u_i - \beta_0 - \beta_1 \bar{x} - \bar{u}\right]}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} \\
&= \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})\beta_1(x_i - \bar{x})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} + \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(u_i - \bar{u})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})} \\
&= \beta_1 + \frac{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(u_i - \bar{u})}{\frac{1}{N}\sum_{i=1}^{N}(z_i - \bar{z})(x_i - \bar{x})}
\end{aligned}
$$

But now lets apply the Law of large numbers to the numerator and denominator

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})(u_i-\bar{u}) &\approx cov(z_i,u_i) \\ &= 0 \end{aligned}$$

but

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})(x_i-\bar{x}) &\approx cov(z_i,x_i) \\ &\neq 0 \end{aligned}$$

Thus this model is consistent since

$$\begin{aligned} \widehat{\beta}_1 &\approx \beta_1+\frac{cov(z_i,u_i)}{cov(z_i,x_{i)}} \\ &\approx \beta_1 \end{aligned}$$

One can calculate standard errors in manner similar to what we did before.

Doing this in stata is straight forward we say

ivreg y (x=z)

Lets return to the examples

Returns to schooling

$$\log(W_i) = \beta_0 + \beta_1 S_i + A_i + u_i$$

Mroz uses number of siblings arguing that father's education should be correlated with schooling but not with wages

Openess and Inflation

Romer uses land as an instrument: smaller countries are more likely to be open

No reason for land to directly influence inflation

Smoking and Birth Weight

$$bwght_i = \beta_0 + \beta_1 packs_i + u_i$$

Key new potential instruments:

- cigtax: cig. tax in home state, 1988
- cigprice: cig. price in home state, 1988

Now what if we have other regressors in the model?

I want to distinguish between different kinds of right hand side variables so I will use different notation.

I also don't particularly like Wooldridge's notation, so I will use my own.

Lets define

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_i + u_i.$$

The difference between w_i and x_i is that we will assume that

$$E\left(u_i \mid x_i\right) = 0$$

but we are not willing to assume that $E(u_i \mid w_i) = 0$.

Instead we assume that

$$E(u_i \mid z_i) = 0$$

and

$$cov(z_i, w_i) \neq 0.$$

The difference between all of these kinds of variables:

Variable	Enters Model	Correlated with u_i
w_i	yes	yes
x_i	yes	no
z_i	no	no

To implement this we use the three equations

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i &= 0 \\
\frac{1}{N}\sum_{i=1}^{N} z_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i\right) &= 0 \\
\frac{1}{N}\sum_{i=1}^{N} x_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i\right) &= 0
\end{aligned}$$

This gives us three equations in the three unknowns $\widehat{\beta}_0, \widehat{\beta}_1$, and $\widehat{\beta}_2$.

In stata we can estimate this using the command

```
ivreg y (w=z) x
```

Lets see how this works

There is nothing special about only having one x_i.

We can easily generalize this to

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 x_{2i} + ... + \beta_k x_{ki} + u_i.$$

with

$$E(u_i \mid x_{ji}) = 0$$

for $j = 2, ..., k$

We then implement this solving

$$\begin{aligned}
\frac{1}{N}\sum_{i=1}^{N} y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_{2i} - ... - \widehat{\beta}_i x_{ki} &= 0 \\
\frac{1}{N}\sum_{i=1}^{N} z_i \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i - ... - \widehat{\beta}_i x_{ki} \right) &= 0 \\
\frac{1}{N}\sum_{i=1}^{N} x_{2i} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i - ... - \widehat{\beta}_i x_{ki} \right) &= 0 \\
&\vdots \;\; \vdots \\
\frac{1}{N}\sum_{i=1}^{N} x_{ki} \left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 w_i - \widehat{\beta}_2 x_i - ... - \widehat{\beta}_i x_{ki} \right) &= 0
\end{aligned}$$

Now we have $k+2$ equations and $k+2$ unknowns so this will work

We implement in stata as

ivreg y (w=z) $x_2 ... x_k$

Lets look at a number of examples

We can also generalize this to have multiple endogenous variables

$$y_i = \beta_0 + \beta_1 w_{1i} + \beta_2 w_{2i} + ... + \beta_s w_{si} + \beta_{s+1} x_{s+1i} + ... + \beta_{s+k} x_{s+ki} + u_i.$$

Now how do we do this?

If we only have one z_i we have k+2 equations but $s + k + 1$ parameters.

This isn't going to work.

We need more equations, how do we get them?

We need more $z's$.

In particular we need at least s z's.

Lets assume we have ℓ z's.

In the simple world $\ell = s$ and things are as before

If $\ell > s$ it is slightly more complicated.

You implement this in stata saying

ivreg y $(w_1 ... w_s = z_1 ... z_\ell) x_{s+1} ... x_{s+k}$

Lets see how to do this.

Simultaneous Equations

Quite often in economics more than one variable is being determined at a time

Classic Example: Supply and Demand, prices and quantities being determined simultaneously

A given consumer takes price as given and decides how much to buy gives demand curve:

$$\log(Y_t) = \alpha_d \log(P_t) + \beta X_t + u_{dt}$$

Firms set price given demand for product yielding supply curve:

$$\log(P_t) = \alpha_s \log(Y_t) + \gamma Z_t + u_{st}.$$

Note that we are assuming that we only have one variable (X_t) that affects demand and another (Z_t) that affects supply. I could easily add more but that doesn't change any of the basic points.

Can we just run OLS to estimate α_d?

Problem: We can substitute first equation into second:

$$\begin{aligned}\log(P_t) =&\alpha_s(\alpha_d \log(P_t) + \beta X_t + u_{dt}) + \gamma Z_t + u_{st}.\\ =&\alpha_s(\alpha_d \log(P_t) + \beta X_t) + \gamma Z_t + \alpha_s u_{dt} + u_{st}.\end{aligned}$$

Notice that $\log(P_t)$ depends on u_{dt} directly

OLS for demand curve

$$\log(Y_t) = \alpha_d \log(P_t) + \beta X_t + u_{dt}$$

clearly won't work because the idea that $E(u_{dt} \mid P_t) = 0$ seems crazy.

This problem is much more general than this, there are many cases in which variables are determined simultaneously:

- GNP,Inflation
- Schooling,Earnings
- Firm A price, Firm B Price
- Husband Labor Supply, Wife Labor Supply
- Production function inputs
- Consumption goods (e.g. Peanut Butter and Jelly)

Lets think more generally of the equations as:

$$\begin{aligned} Y_{0i} &= \alpha_0 Y_{1i} + \beta X_i + \delta_0 W_i + u_{0i} \\ Y_{1i} &= \alpha_1 Y_{0i} + \gamma Z_i + \delta_1 W_i + u_{1i} \end{aligned}$$

The distinction between W, X and Z is that

- X_i only affects Y_{0i}
- Z_i only affects Y_{1i}
- W_i affects both Y_{0i} and Y_{1i}

I will write everything as if there is only one element of each, but it is easy to generalize beyond that

We will assume that W_i, X_i and Z_i are "exogenous" or uncorrelated with (u_{0i}, u_{1i})

(Y_{0i}, Y_{1i}) are endogenous or correlated with (u_{0i}, u_{1i})

Our goal is estimation of α_0 (or α_1, or both).

OLS will not work because u_{0i} helps determine Y_{1i} therefore they are correlated

The primary assumption underlying OLS is violated because Y_{1i} is correlated with u_{0i}

It turns out that there is a nice solution to this problem

Notice that we can write

$$
\begin{aligned}
Y_{1i} &= \alpha_1\left(\alpha_0 Y_{1i} + \beta X_i + \delta_0 W_i + u_{0i}\right) + \gamma Z_i + \delta_1 W_i + u_{1i} \\
&= \alpha_1\alpha_0 Y_{1i} + \alpha_1\beta X_i + \gamma Z_i + \left(\alpha_1\delta_0 + \delta_1\right) W_i + \left(\alpha_1 u_{0i} + u_{1i}\right) \\
&= \frac{1}{1-\alpha_1\alpha_0}\left(\alpha_1\beta X_i + \gamma Z_i + \left(\alpha_1\delta_0 + \delta_1\right) W_i + \left(\alpha_1 u_{0i} + u_{1i}\right)\right) \\
&\equiv \beta^* X_i + \gamma^* Z_i + \delta^* W_i + u_i^*
\end{aligned}
$$

But now comes the part that will turn out nice,

plug this into the other equation to get

$$
\begin{aligned}
Y_{0i} &= \alpha_0 Y_{1i} + X_i'\beta + W_i'\delta_0 + u_{0i} \\
&= \alpha_0\left(\beta^* X_i + \gamma^* Z_i + \delta^* W_i + u_i^*\right) + \beta X_i + \delta_0 W_i + u_{0i} \\
&= \alpha_0\left(\beta^* X_i + \gamma^* Z_i + \delta^* W_i\right) + \beta X_i + \delta_0 W_i + u_{0i} + \alpha_0 u_i^*
\end{aligned}
$$

If we knew $(\beta^*, \gamma^*, \delta^*)$ we could run a regression of Y_{0i} on $(\beta^* X_i + \gamma^* Z_i + \delta^* W_i)$, X_i and W_i.

As long as u_{0i} or u_{1i} are uncorrelated with (X_i, Z_i, W_i) we can get consistent estimates of the parameters from this regression

The coefficient on the big term is α_0 so we would be done.

The problem is that we don't know $(\beta^*, \gamma^*, \delta^*)$ but we can estimate them

- Regress Y_{1i} on X_i, Z_i and W_i to get estimates of β^*, γ^*, and δ^*.
- use these estimates (hats mean estimates) to form

$$\hat{Y}^*_{1i} = \widehat{\beta^*} X_i + \widehat{\gamma^*} Z_i + \widehat{\delta^*} W_i$$

- Regress Y_{0i} on X_i, W_i and $\hat{Y}^*_{1i}$.
- The coefficient on $\hat{Y}^*_{1i}$ gives us a consistent estimate of α_0

This is called “Two staged least squares”

It gives consistent estimates of the parameters

However standard errors are not right if you do it directly because it doesn’t deal with the fact that $\widehat{\beta^*}$ is estimated.

Lets dig a little deeper into this problem

Under what conditions can I run the second regression?

It has to be the case that I have a Z_i that affects Y_{1i} but not Y_{0i}, otherwise there will be a collinearity problem.

This is not a small problem since this is fundamental for identification.

To see this suppose $\gamma^* = 0$ then our regression would be

$$Y_{0i} = \alpha_0 \left(\beta^* X_i + \delta^* W_i\right) + \beta X_i + \delta_0 W_i + u_{0i} + \alpha_0 u_i^*$$

we would have a perfect multicollinearity problem

Relationship between 2SLS and IV

The fact that you need a variable that is correlated with Y_{1i} and doesn't affect Y_{0i} makes 2SLS sound a lot like IV

In fact they are the same thing when there are exactly as many instruments as endogenous variable

I am not going to show this mathematically, but I'll show you in stata

Now lets look at a number of example of simultaneous equation models