Sarah Hamersma
1/19/05; revised 1/23/06 for use in ECO 7424

Instrumental Variables Lecture

These notes are drawn from Angrist and Krueger (JEP, 2001), Angrist and Krueger (Handbook of Labor Economics, 1999), the Wooldridge text (2002), and the Cameron & Trivedi text (2005).

The goals:

You will know when you should look for an IV
You will know what an IV looks like (i.e. how to judge if it's a good one)
You will see how an IV fixes the endogeneity problem (it makes a lot of sense)
You will see what dangers there are in doing IV estimation
All of this will be imbedded in your economic intuition from now on!

When should you look for an IV?

When you are trying to establish a causal link (how does X affect Y?) you may be concerned that X is not exogenous, i.e. X is endogenous (correlated with the error).

(1) $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

You are concerned that x_2 is correlated with u. There are 4 possible sources of endogeneity:
1) simultaneity (things "cause each other", like wages and hours, or crime and police force)
2) measurement error
3) omitted variables
4) selection bias

Note that this means β is defined "structurally" (it has meaning; causality) – there's no such thing as endogeneity if we are just doing a statistical projection (to generate predictions w/o claiming causality).

We will focus on (3). For example,
y = wages, x_2 = education (maybe the highly educated have higher wages b/c of unobs. ability)
y = wages, x_2 = veteran status (entering military is a decision, maybe based on unobservables)

The second one will be our running example for the lecture. It is based on Angrist 1990.
Assume x_1 is something exogenous that affects wages, like race or gender.

If we only have info on y, x_1, and x_2 we are sunk. All βhats will be inconsistent. This is when we need an IV.

What does an IV look like?

We need a variable z_1 that can "represent" our endogenous variable, but without the parts of our endogenous variable that are picking up an omitted variable (since they don't represent causal effects of the endogenous variable (x_2) itself). We'll assume a single instrument for a single endogenous variable for now (the "just-identified" case), but the rules are the same for multiple instruments.

There are two important features an instrument MUST have:

1) z_1 should be correlated with x_2

More specifically, if we run the linear projection ("reduced form equation")

(2) $$x_2 = \delta_0 + \delta_1 x_1 + \theta_1 z_1 + r$$

we had better expect θ_1 to be *nonzero*. In other words, after the other variables are taken into account (x_1) the IV (z_1) had better have some **power in acting as a representative** for x_2. Otherwise it will be a lousy IV.

Notes: by definition of a projection, E(r) = 0 and r is uncorrelated with x_1 and z_1. Also, it doesn't matter if x_2 or z_1 are continuous or discrete.

2) z_1 should be uncorrelated with u

In other words, $\text{cov}(z_1,u) = 0$, i.e. z_1 should be exogenous like x_1. The intuition here is that z_1 should not have its own independent influence on y (i.e. it shouldn't be an omitted variable in (1) – if it were, it should have been in the model in the first place). The requirement is that any relationship that z_1 has to y is *only* through the way z_1 influences x_2, which then influences y. Otherwise we won't be able to sort out how much of the influence of z_1 on y is due to its direct influence vs. its indirect influence (via x_2). We want it to be **a representative *only***.

Q: Are these testable?
A: (1) is testable, but (2) must be based on economic reasoning/common sense (can't be tested)

How does IV fix the endogeneity problem?

We somehow want to integrate z_1 into our structural equation of interest, (1), since x_2 is problematic. Substitute (2) in for x_2 to create a new linear projection:

$$y = \beta_0 + \beta_1 x_1 + \beta_2\{ \delta_0 + \delta_1 x_1 + \theta_1 z_1 + r \} + u$$

Collect terms so that each variable has just one coefficient:

(3) $$y = [\beta_0 + \beta_2 \delta_0] + [\beta_1 + \beta_2\delta_1] x_1 + [\beta_2\theta_1] z_1 + [\beta_2 r + u]$$

This looks like it could just be estimated with a regression of y on x_1 and z_1 – and so it can! The new error term is uncorrelated with x_1 and z_1 so this can be consistently estimated with OLS. Obviously, though, ***not all separate parameters will be identified***. For example, the coefficient on z_1 (call it λhat) will be an estimate of the *product* $\beta_2\theta_1$. We might be interested in that – it describes the influence of the instrument on the outcome; for example, the influence of being draft-eligible on your future wages.

But what we're really interested in is β_2 – we think that veteran status is really the causal variable in the wage equation, and we think that the variable identifying those who are draft-eligible is just picking up that effect (if it's a good instrument, that is exactly what it should be doing).

How do we estimate β_2 itself?
The answer is straightforward: if we can divide our estimate of $\beta_2\theta_1$ by the value of θ_1, then we'll be all set. In practice, this means first estimating equation (2), the reduced form, to get an estimate of θ_1 and second, estimating (3) to get our estimate of $\beta_2\theta_1$. If we divide these estimates, we get an estimate of β_2.

The intuition: Continue with our example on military issues. If we get a value of $\theta_1 = .5$, it means that every unit increase in z_1 is "representing" a half-unit increase in the variable of interest, x_2. Suppose we then find that z_1 has a coefficient of - $5000 in equation (3). This means for every unit increase in z_1, wages fall by $5000. But since a unit change in z_1 only represents HALF a unit change in x_2, then a unit change in x_2 must cause a $10,000 loss in wages (i.e. $5000/.5).

An equivalent way to think about this (in the just-identified case) is:

$$\hat{\beta}_{IV} = \frac{Cov(z, y)}{Cov(z, x)}$$

This can be written as a familiar looking formula:

$$\hat{\beta}_{IV} = (z'X)^{-1} z'y$$

With variance estimate: $$\hat{V}(\hat{\beta}_{IV}) = (z'X)^{-1} z' \hat{\Omega} z (z'X)^{-1}$$

where $\hat{\Omega} = Diag(\hat{u}_i^2)$

What if you have more than one instrument for the endogenous regressor?

This is called the "overidentified" case -- you can't use the same simple process when you have multiple instruments. One solution is to drop some instruments, but this can hurt efficiency.

You can use multiple instruments with a procedure called two-stage least squares, or 2SLS (and it also works for the just-identified case). 2SLS works as follows:

Stage 1: regress the endogenous variable on all of the exogenous regressors and instruments:

$x_2 = \delta_0 + \delta_1 x_1 + \theta_1 z_1 + \theta_2 z_2 + r$

Calculate the predicted value of x_2 for each observation, based on the coefficients found

$x_2\text{hat} = (\delta_0\text{hat}) + (\delta_1\text{hat})x_1 + (\theta_1\text{hat})z_1 + (\theta_2\text{hat})z_2$

Stage 2: regress y on the exogenous variables and the **predicted** $x_2$hat variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2(x_2\text{hat}) + u$$

This identifies β_2 (and the other β's) directly!

This can again be written in a way that helps compare it to OLS:

$$\hat{\beta}_{IV} = \left[X'Z\,(Z'Z)^{-1}Z'X \right]^{-1} \left[X'Z\,(Z'Z)^{-1}Z'y \right]$$

Be careful though – if you use this procedure directly, your standard errors will not be right. (they will be based on residuals defined as "y – xhat'βhat" rather than the correct calculation "y – x'βhat"). Stata has a 2SLS command that will fix these for you.

IV Pitfalls

1) BAD instruments

If your instrument has some of its own predictive power beyond its work as a representative of the endogenous variable (i.e. it is correlated with the error term in the structural equation) you have big problems.

Q: Why?
A: The coefficient on the instrument is going to be MULTIPLIED when generating an estimate of the effect of the endogenous variable. If any part of the coefficient should NOT be attributed to the endogenous regressor, that error is being multiplied, potentially a lot. Remember: "a little omitted variable bias can go a long way"

Ex. In a returns to education estimation, one could propose using mother's educ as an instrument for one's own education. Though it's certainly correlated w/ the endogenous variable, it might be a bad instrument.

There's no really good way to test this assumption (and not ANY way when just-identified).

2) Weak instruments

These instruments are properly exogenous, but they just aren't too powerful. This is problematic because IV is not unbiased, and the bias tends to be larger when instruments are weak. Moreover, it tends to bias the results toward the OLS results. The same happens if we use lots and lots of instruments.

-> we should always check the strength of our instruments in the first stage (partial R^2 or F-test)

Ex. Angrist and Krueger's quarter-of-birth IV

3) Need to implement it properly

One should note that the identification relies on a series of linear regressions. If the endogenous regressor is a dummy variable, people sometimes want to do the first stage with a logit or probit. If you look at our derivation of the identification of β_2, you can see that this could really mess things up. It turns out that the only way this will work is if your nonlinear functional form is exactly right. You are always safe using OLS – the second stage will be consistent.

Also, you need to be sure to get the right standard errors, as mentioned earlier.

4) Need to interpret it properly

If we don't think treatment would necessarily impact everyone in a population in the same way, what we identify here may be a "local" average treatment effect (LATE) – i.e. what is the effect of the treatment on those who could be induced to treatment via the instrument?

Finally, be thinking about the distinction between an IV and a proxy. You will be addressing this issue in the next homework assignment.