Instrument Variables

Daniel Cullen

October 18, 2017

One method to deal with the problem of endogenity ($Cov(x_i, \varepsilon_i) \neq 0$) is an Instrument Variable approach. There are several reasons why the error term may be correlated with a regressor: omitted variables, measurement error in the regressor, and simultaneity.

What we need for an instrument variable:

- An "exogenous" factor (something outside the model) that shifts x_i in such a way that ε_i is not affected.
- Alternatively, something randomly determined that affects x_i

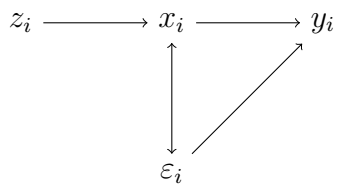
Suppose we have the following model

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i,$$

and we are concerned that x_i and ε_i are correlated (i.e. x_i is endogenous). We can use an instrument variable z_i to "instrument" for x_i. There are two conditions that must be met for a variable, z_i, to be a valid instrument.

1. $Cov(x_i, z_i) \neq 0$ (The instrument is relevant or the first stage exists)
2. $Cov(z_i, \varepsilon_i) = 0$ (exclusion restriction)

The exclusion restriction can be thought of another way. The instrument, z_i, does not directly influence the dependent variable, y_i, its only affect is indirectly through x_i.



Two Staged Least Squares

Casual Relationship of interest:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 w_i + \varepsilon_i,$$

First Stage:

$$x_i = \alpha_0 + \alpha_1 z_i + \alpha_2 w_i + u_i,$$

Predicted First Stage:

$$\widehat{x}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 z_i + \widehat{\alpha}_2 w_i$$

Second Stage:

$$y_i = \beta_0 + \beta_1 \widehat{x}_i + \beta_2 w_i + \varepsilon_i,$$

Note: You need at least as many instruments as endogenous right hand side variables in equation being estimated.

To test whether the instrument, z_i affects x_i, do a t-test of the coefficient on z_i. If the t-stat is less than 3.5, the instrument is a weak instrument.

By the instrument exogeneity assumption $cov(z_i, u_i) = 0$ and the instrument relevance assumption $cov(z_i, x_i) \neq 0$

$$\begin{aligned} cov(z_i, y_i) &= cov(z_i, \beta_0 + \beta_1 x_i + u_i) \\ &= \beta_1 cov(z_i, x_i) + cov(z_i, u_i) \\ \implies \beta_1 &= \frac{cov(z_i, y_i)}{cov(z_i, x_i)} \end{aligned}$$

In practice it is often difficult to find convincing instruments (in particular because many potential IVs do not satisfy the exclusion restriction). An example of a paper utilizing IV is "Children and their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size" by Joshua Angrist and William Evans, American Economic Review, 1996. It turns out, parents typically have strong preferences for mixed-gender children. What this means is that parents of two same-sex children are more likely to have a third child than parents of mixed-sex children (about 6 percentage points more likely). Instrument, z_i, is a dummy indicating first two children are the same sex. We can only look at parents with at least 2 children, instrument shifts the probability of having a third child. y_i is labor supply and x_i is the number of children. Two stage least squares results show that a third child reduces hours per week by 4.5 hours.