ECON-122
Introduction to Econometrics

Agnieszka Postepska

August 8th, 2011

Omitted variable bias

- recall assumption A3: $E(u|x) = 0$
- recall that omitting a relevant variable from the regression is a source of violation of this assumption (why?)
- the broader class of problems that involve violation assumption A3 is called *endogeneity*
- other sources of endogeneity include selection and measurement error

Endogeneity

- selection occurs when there is some underlying selection process that determines whether an individual is in the population of interest (example: wage equation - we only observe wages for working individuals, so the selection problem in this case would be selection to employment)
- selection leads to violation of assumption A3 because many of the regressors that determine individual's wage, influence the selection into employment too, which means that they contain some information about the error term
- example of measurement error: whenever we are using an imperfect data set - information on the variable of interest was gathered incorrectly
- when we include an incorrectly measured regressor into the regression model, we are adding another disturbance which is correlated with the regressor itself

Simplest solution to omitted variable problem: using a proxy for unobserved variable

- Consider the wage equation once again:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ability + u$$

- this model shows explicitly that we would like to hold ability fixed when we are measuring the impact of education on wage
- what happens if we cannot control for ability and we have to leave it out when we are estimating the model?

Proxy solution to omitted variable bias

- the simplest solution to this problem is to find a *proxy* for the omitted variable (ability)
- *proxy* variable is something that is related to the omitted variable
- let's use IQ measures - note that IQ and ability are not the same things - all we require from IQ is to be correlated with ability
- do we satisfy all assumptions now?
- what is the interpretation of the coefficients?
- important: we cannot get unbiased estimates of the coefficient on the omitted variable- we get coefficient on the proxy

Instrumental variable estimation (IV)

- now we will move towards a much more general method
- while the proxy method works and we can get unbiased estimates of the other coefficients, it is often hard to find a proxy
- now we will see an estimation method that recognizes the problem with omitted variable bias and corrects for it while estimating the coefficients - *IV estimation*
- consider once again the wage equation:

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + e$$

- and suppose that we do not have data on IQ and no other proxy is available

IV cont.

- in this case we put ability in the error term and we estimate the following model:

$$log(wage) = \beta_0 + \beta_1 educ + u$$

- what are the properties of the OLS estimator under the conditions above?
- IV will allow us to obtain unbiased estimator for the parameter on education using the regression model above
- which variable is *endogenous* in this model?

Method of instrumental variables

- in order to obtain unbiased estimator of the parameters we need to have some additional information
- we saw earlier that given the information we have on hand, OLS estimator is biased
- this new information comes from a new variable (or variables) that satisfies certain properties
- two assumptions that the new variable, z, must satisfy are:
    1. *Instrument relevance*: $Cov(z, x) \neq 0$ - z and x are correlated
    2. *Instrument exogeneity*: $Cov(z, u) = 0$ - z and u are uncorrelated

The two properties of an instrument in detail

- exogeneity implies that z should have no effect on y after x and omitted variables are controlled for and z should be uncorrelated with the omitted variables
- relevance implies that z must be related to the endogenous variable
- note that while it is possible to test relevance, we cannot test exogeneity of the instrument, and it is the exogeneity that is harder to satisfy (argue that it holds using common sense...)
- to test relevance just run simple regression of x on z and test whether the coefficient on z is significantly different from zero

The two properties of an instrument in detail cont.

- these two assumption imply that we need a variable that is correlated with the endogenous variable but uncorrelated with the omitted variable - not an easy task.... think about education and ability - what kind of instruments can we think of in this setting?
- in many cases this just boils down to selling your story...
- so in our example z must be such that it is (1) correlated with education and (2) uncorrelated with ability
- notice the crucial difference with proxies - proxy must be as high correlated with the omitted variable as possible, whereas IV just the opposite
- this highlights the purpose of IV: we want to get unbiased estimators for all other parameters in the model - so we want to remove the effect of the variable that is omitted but we are not trying to control for it

IV in wage equations in practice

- labor economists have used family background variables to *instrument for education*
- one example is mother's and father's education
- it certainly is a relevant instrument as typically children's education is related to parent's education (more educated parents have better educated children)
- however, it almost certainly fails the exogeneity assumption - parents education is almost surely correlated with parents ability and parents ability is correlated with children's ability
- nevertheless, many papers have been published using this instrument
- another instrument for education is number of siblings - usually having more siblings is associated with lower average levels of education
- it is much more likely that number of siblings is uncorrelated with ability (is it?)

IV in equations

- now we will show why is it possible to identify the parameters in the original model (and primarily the parameter on the endogenous variable) using this extra variable z
- identification in this case means that we can write $\hat{\beta}_1$ as a function of z,x and y
- we can show that β_1 can be written as:

$$\beta_1 = \frac{Cov(z,y)}{Cov(z,x)}$$

- in a simple regression model the **instrumental variable estimator** is:

$$\hat{\beta}_1 = \frac{\sum_i (z_i - \bar{z})(y_i - \bar{y})}{(z_i - \bar{z})(x_i - \bar{x})}$$

- notice that x is not ignored in the estimation - in the denominator we are dividing by the covariance between $x's$ and $z's$
- this is why we cannot just use instrument as any other regressor and run OLS!!! (that would be equivalent to which method?)

Testing with IV estimation

- first we need to obtain the estimator for the standard deviation - the standard errors of the IV estimator
- as in case of OLS, we need to make the assumption of homoskedasticity
- notice though that this assumption is placed on the exogenous variable, z, not $x's$:

$$Var(u^2|x) = Var(u) = \sigma^2$$

- the formula for the variance in this scenario is slightly more difficult than in OLS:

$$Var(\hat{\beta}_1|z) = \frac{\sigma^2}{n\sigma_x^2\rho_{xz}^2}$$

where σ_x^2 is the population variance of x (can be estimated using sample variance of x), σ^2 is the population variance of u (can be estimated using IV residuals) and ρ_{xz}^2 is the population correlation between x and z (can be obtained by getting the R^2 from regressing x_i on z_i)

Testing with IV estimation cont.

- so the estimated variance is:

$$\hat{Var}(\hat{\beta}_1|z) = \frac{\hat{\sigma^2}}{SST_x R^2_{xz}}$$

- notice that in case of no endogeneity IV will have higher variance than OLS because only if $R^2_{xz} = 1$ are these variances equal and $R^2_{xz} = 1$ only if $x = z$
- in any other case is OLS variance lower

IV in multiple regression model

- now consider the following multiple regression model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + ... + \beta_k x_{k-1} + u$$

  where y_2 is an endogenous variable, so that $E(u|y_2) \neq 0$

- also let z be such that
  1. $Cov(z, y_2) \neq 0$
  2. $Cov(z, u) = 0$
- thus z is a good instrument for y_2
- notice that z cannot be the same as any of the other $x's$ - it has to be a variable that is not in the original model

Standard errors in the multiple regression model

- now the assumption of homoskedasticity is placed on all of the exogenous variables, $z's$ and $x's$ and not y_2:

$$Var(u^2|z,x) = Var(u) = \sigma^2$$

- the formula for the variance in this scenario is even more difficult so we will skip it
- the important thing is that homoskedasticity is placed on the instrument an other exogenous $x's$ in the model, and not the endogenous variable

What happens if we have weak instruments

- first consequence of weak correlation between x and z is that IV estimates will have large standard errors - if correlation is low then there is very little information to estimate the coefficient
- recall the formula for the $Var(\hat{\beta}_1)$. The estimated variance is then:

$$\hat{Var}(\hat{\beta}_1|x,z) = \frac{\hat{\sigma}^2}{SST_x R^2_{xz}}$$

- in case of no correlation (instrument doesn't satisfy the relevance property) the regression can produce strange results
- on the other hand we don't want out instrument too highly correlated with the endogenous variable as then it is almost sure that the exogeneity condition is not satisfied
- therefore it is crucial to test relevance every time IV is conducted

Two stage least squares: 2SLS

- consider the following model:

  $$y_i = \beta_0 + \beta_1 x_i + \beta_2 y_{2i} + u_i$$

  where x is exogenous and y_2 is endogenous
- suppose when we have more than one instrument for y_2: z_1 and z_2
- now, we need a slightly different approach - we cannot just insert all of them into the estimation equation instead of the endogenous regressor (if we just use one of the two we have an inefficient estimator - recall the formula for the IV estimator variance)
- first, we need to decide on the best linear combination of the instruments - best linear combination is one that exhibits the highest correlation with the endogenous variable

2SLS

- the correct approach in this situation is 2SLS estimation
- first we find the best linear combination of the instruments - by regressing the endogenous variable on the instruments and the exogenous variables in the model:

  $$y_2 = \pi_0 + \pi_1 x_1 + \pi_2 z_2 + \pi_3 z_3 + v$$

  then the best IV for y_2 becomes:

  $$y_2^* = \pi_0 + \pi_1 x_1 + \pi_2 z_2 + \pi_3 z_3$$

  s.t. at least one of the coefficients on the instruments is different than zero (exclusion restrictions)

2SLS cont.

- in practice we estimate the equation for y_2 by OLS, obtain the coefficients for $\pi's$, test whether $\pi_2 = 0$ **and** $\pi_3 = 0$ and compute $\hat{y}^*$
- if both $\pi_1 = 0$ and $\pi_2 = 0$ then both instrumenst are not valid and we cannot identify parameters in the main equation
- once we have $\hat{y}^*$, we run OLS on the main equation using $\hat{y}^*$ instead of y_2, so we run regression of y_1 on $\hat{y}_2^*$ and x
- however, one must remember that standard errors are calculated using the original $y_2's$ not the $\hat{y}_2^*$ - therefore when doing 2SLS by hand one must account for this, statistical packages do this correction automatically

Testing for endogeneity

- as IV delivers unbiased estimates even in the absence of endogeneity why is it important to test whether endogeneity really occurs?
- to illustrate the test, consider the following model:

  $$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_3 + u$$

  where y_2 is the endogenous regressor and x_1 and x_2 are exogenous
- assume we have two additional variables -potential instruments: z_1 and z_2
- we know that if y_2 is uncorrelated with u we should estimate our model using OLS

Testing for endogeneity: Hausman test

- to test whether this is really the case we can use the *Hausman test*:
- write down the model for the endogenous variable:

$$y_2 = \pi_0 + \pi_1 x_1 + \pi_2 x_2 + \pi_3 z_1 + \pi z_2 + v$$

- y_2 is uncorrelated with u (no endogeneity) iff v is uncorrelated with u
- as we do not observe v, we can use residuals, $\hat{v}$
- then we can include the $\hat{v}$ in the main equation as additional regressor:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + \beta_3 x_3 + \delta \hat{v} + error$$

- now we can perform a simple t test on δ: H_0: $\delta = 0$ so under the null there is no endogeneity in the model
- if however, we will reject the null at a small significance level, we conclude that y_2 is endogenous (in which case we use IV to estimate the model parameters)