Lecture 7

Endogeneity 1 of 4: IV & 2SLS (i)

Michael Curran

Trinity College Dublin

JS Econometrics

Lecture 7 Outline

Overview: endogenous explanatory variables

Week 4: chapter 15

- Background: omitted variable bias – derivation of omitted variable bias (chapter 3); OLS is inconsistent under omitted variables (chapter 5); omitted variable bias can be eliminated (or at least mitigated) when a suitable proxy variable is given for an unobserved explanatory variable (chapter 9); unfortunately, suitable proxy variables are not always available.
- Different approach to endogeneity here: **instrumental variables** (**IV**) – solve endogeneity in one or more explanatory variables and **2SLS**, which is second in popularity only to OLS for estimating linear equations in applied econometrics.
  1. IV yields consistent estimators under omitted variables.
  2. IV used to solve **errors-in-variables** problem.
- Not covering: IV applied to time series and panel data just like OLS.
- Next week, chapter 16: use IV to estimate simultaneous equation models.

Motivation: Omitted Variables

Omitted variable bias (unobserved heterogeneity): excluding a relevant variable $\equiv$ underspecifying the model.

1. Ignore problem and suffer consequences of biased and inconsistent estimators.
2. Find suitable proxy for unobserved variable.
3. Assume omitted variable time invariant and use panel data (fixed effects) / time series (first-differencing methods).

IV: leaves unobserved variable in error term but recognizes presence of omitted variable. Example:

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

Proxy variable such as IQ can be substituted for ability and then (under assumptions) a consistent estimator of β_1 is available from the regression of $\log(wage)$ on $educ, IQ$.

Motivation: Omitted Variables

- What if no proxy variable is available or does not have properties needed to produce a consistent estimator of β_1?
- Put *abil* into error term and left with simple regression model:

$$\log(wage) = \beta_0 + \beta_1 educ + u \tag{1}$$

where u contains *abil*.

- OLS yields biased and inconsistent estimator of β_1 if *educ* and *abil* are correlated.
- With an IV for *educ*, can still use equation (1) as basis for estimation.

$$y = \beta_0 + \beta_1 x + u \tag{2}$$

$Cov(x, u) \neq 0$.

- IV works whether or not x and u are correlated, but OLS should be used if x is uncorrelated with u.

Motivation: Omitted Variables

To obtain consistent estimates of β_0 and β_1 when x and u are correlated, we need some extra info, via an observable variable z such that

1. z is uncorrelated with u, i.e. $Cov(z, u) = 0$. 'z is exogenous in equation (2)' so z should have no partial effect on y (once x and the omitted variables in u are controlled for), and z should not be correlated with unobserved factors that affect y. Generally untestable.
2. z is correlated with x, i.e. $Cov(z, x) \neq 0$. z must be related (positively or negatively) to the endogenous explanatory variable x. Testable given a random sample from the population.

Then, call z an **instrumental variable** for x.

Motivation: Omitted Variables

- Testing $Cov(z, x) \neq 0$:

$$x = \pi_0 + \pi_1 z + v$$

- Since $\pi_1 = Cov(z, x) / Var(z)$, $Cov(z, x) \neq 0 \iff \pi_1 \neq 0$, thus we should be able to *reject* null hypothesis:

$$H_0 : \pi_1 = 0 \tag{3}$$

  against two-sided alternative $H_0 : \pi_1 \neq 0$ at a sufficiently small significance level (say 5% or 1%).
- If this is the case, then we can be fairly confident that $Cov(z, x) \neq 0$.

Motivation: Omitted Variables

Example

- For $\log(wage)$ equation in (1), IV z for educ must be (i) uncorrelated with ability (and any other unobservable factors affecting wage) and (ii) correlated with education.
- Example: last digit of an individual's social security number satisfies first requirement: it is uncorrelated with ability because it is determined randomly; however, this is not correlated with education so it makes a poor IV for educ.
- What we have called a *proxy variable* for the omitted variable makes a poor IV for the opposite reason.
- E.g. in $\log(wage)$ example with omitted ability, proxy variable for *abil* must be as highly correlated as possible as *abil*.
- IV variable must be uncorrelated with *abil*, so while *IQ* is a good candidate as a proxy variable for *abil*, it is not a good IV for *educ*.

Motivation: Omitted Variables

Example

- Less clear whether other possible IV candidates satisfy exogeneity requirement $Cov(z, u) = 0$.
- In wage equations, used family background variables as IVs for education, e.g. mother's education positively correlated with child's education as can be seen by collecting a sample of data on working people and running a simple regression of *educ* on *motheduc*. Thus, *motheduc* satisfies $Cov(z, x) \neq 0$.
- Problem: mother's education might be correlated with child's ability (through mother's ability and perhaps quality of nurturing at an early age) so $Cov(z, u) = 0$ fails.
- Another IV choice for *educ* in (1) is number of siblings while growing up (*sibs*).
- Typically: more siblings associated with lower average levels of education, so if number of siblings is uncorrelated with ability, it can act as an IV for *educ*.

Motivation: Omitted Variables

Another Example

Estimating the causal effect of skipping classes on final exam score:

$$score = \beta_0 + \beta_1 skipped + u \tag{4}$$

skipped could be correlated with other factors in u so a simple regression of *score* on *skipped* may not give a good estimate of the causal effect of missing classes. What might be a good IV for *skipped*? Need something that has no direct effect on *score* and is not correlated with student ability and motivation. Also, IV must be corr with *skipped*. e.g. distance between living quarters and campus. So, *skipped* may be positively correlated with distance – check by regressing *skipped* on *distance* and doing a t test, as before. Is *distance* uncorrelated with u? In model (4), some factors in u may be correlated with *distance*, e.g. students from low-income families may live off campus; if income affects student performance, this could cause *distance* to be correlated with u. An IV approach might not be necessary if a good proxy exists for student ability, e.g. cumulative GPA prior to the semester.

Motivation: Omitted Variables

Availability of IV can be used to consistently estimate the parameters in equation (2). Specifically, $Cov(z, u) = 0$ & $Cov(z, x) \neq 0$ (equivalently, $Cov(z, u) = 0$ and (3)) serve to *identify* the parameter β_1.

Identification of a parameter in this context means that we can write β_1 in terms of population moments that can be estimated using a sample of data.

To write β_1 in terms of population covariances:

$$Cov(z, y) = \beta_1 Cov(z, x) + Cov(z, u)$$

Under $Cov(z, u) = 0$ & $Cov(z, x) \neq 0$:

$$\beta_1 = \frac{Cov(z, y)}{Cov(z, x)} \tag{5}$$

Note: algebra fails if z and x are uncorrelated, i.e. if $Cov(z, x) = 0$.

(5) shows that β_1 is the population covariance between z and y divided by the pop cov between z and x, which shows that β_1 is identified.

Motivation: Omitted Variables

Given a random sample, we estimate the population quantities by the sample analogs (Analogy Principle).

Instrumental variables (IV) estimator of β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})} \tag{6}$$

Given sample of data on x, y and z, simple to obtain IV estimate in (6). IV est of β_0 is simply $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, which looks just like the OLS intercept estimator except that the slope estimator $\hat{\beta}_1$ is now the IV estimator. When $z = x$ we obtain OLS est of β_1, i.e. when x is exogenous, it can be used as its own IV, and the IV estimator is then identical to the OLS estimator. LLN shows that IV estimator is consistent for β_1 : $\text{plim}(\hat{\beta}_1) = \beta_1$ provided assumptions $Cov(z, u) = 0$ and $Cov(z, x) \neq 0$ are satisfied. If either assumption fails, the IV estimators are not consistent. When $Corr(x, u) \neq 0$, so IV estimation is actually needed, it's never unbiased so in small samples IV estimators can have a substantial bias – one reason why large samples are preferred.

IV: Statistical Inference

IV estimators have approximately Normal distributions in large samples. For inference, need standard errors to compute t statistics and confidence intervals. As for homogeneity assumption, now state conditional on IV z, not on endogenous variable x:

$$E(u^2|z) = \sigma^2 = Var(u) \tag{7}$$

$$AVar(\hat{\beta}_1) = \frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{8}$$

As with OLS estimator, asymptotic variance of IV estimator decreases to 0 at rate of $\frac{1}{n}$ where n is the sample size.

1. (8) provides a way to obtain standard errors for IV estimation: all quantities in (8) can be consistently estimated given a random sample; any modern metrics package computes standard errors after any IV estimation.
2. (8) allows us to compare asymptotic variances of IV and OLS est (when x and u are uncorrelated).

IV: Statistical Inference

- More on the second point from last slide: under Gauss-Markov assumptions, variance of OLS estimator is σ^2/SST_x, while that for IV is $\sigma^2/(SST_x \cdot R^2_{x,z})$; these differ only in that $R^2_{x,z}$ appears in the denominator of the IV variance. Since $R^2 < 1$, IV variance is always larger than OLS variance (when OLS is valid). If $R^2_{x,z}$ is small, then the IV variance can be much larger than the OLS variance. If x and z are only slightly correlated, $R^2_{x,z}$ can be small, and this can translate into a very large sampling variance for the IV estimator. The more highly correlated z is with x, the closer $R^2_{x,z}$ is to one and the smaller is the variance of the IV estimator. In the case that $z = x$, $R^2_{x,z} = 1$ and we get the OLS variance, as expected. Thus an important cost of performing IV estimation when x and u uncorrelated: *AVar* of IV estimators is always larger and sometimes much larger than the *AVar* of OLS estimators.
- See examples 15.1 & 15.2.

IV: Statistical Inference

Qualitative Variables

Angrist & Krueger (1991): clever binary IV for *educ*. Quarter of birth (1 if first, 0 otherwise) should be unrelated to error / ability, but years of education vary systematically in population based on quarter of birth due to compulsory school attendance laws. Students born early in year start school older so reach compulsory age with less education but no relationship for students who finish high school. $R^2_{x,z}$ is very small since years of education varies only slightly across quarter of birth, so needed a very large sample size (247199 men born between 1920 and 1929) to get a reasonably precise IV estimates. Econometric critique: Bound, Jaeger & Baker (1995): not obvious that season of birth is unrelated to unobserved factors that affect wage. Even small amount of correlation between z and u can cause serious problems for IV estimator.

IV: Statistical Inference

Qualitative Variables

For policy analysis, endogenous explanatory variable is often binary, e.g. Angrist (1990) studied effect that being a veteran in Vietnam War had on lifetime earnings:

$$\log(earns) = \beta_0 + \beta_1 veteran + u$$

where veteran is binary variable. Problem with OLS est here is there may be *self-selection* problem (chapter 7): perhaps people who get most out of military choose to join or decision to join is correlated with other characteristics affecting earnings causing veteran and u to be correlated. Angrist: Vietnam draft lottery provided **natural experiment** (chapter 13) that created IV for veteran. Random assignment: plausible that draft lottery number is uncorrelated with error term u, but those with a low enough number had to serve in Vietnam so probability of being a veteran is correlated with lottery number. If both assertions are true, then draft lottery number is a good IV candidate for veteran. Also possible to have a binary endogenous explanatory variable and binary IV.

Properties of IV with a Poor IV

While IV is consistent when z and u are uncorrelated and z and x are positively/negatively correlated, IV suffers from large standard errors especially if z and x are weakly correlated. Weak correlation (**weak identification**) between z and x can have even more serious consequences: IV estimates can have a large asymptotic bias even if z and u are only moderately correlated. We can see this by studying the plim of IV estimates when z and u are possibly correlated. We can derive this in terms of population correlations and standard deviations as

$$\text{plim}\,\hat{\beta}_1 = \beta_1 + \frac{Corr(z,u)}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x} \tag{9}$$

Even if $Corr(z,u)$ is small, the inconsistency in the IV estimator can be very large if $Corr(z,x)$ is also small. So, even if we focus only on consistency, it's not necessarily better to use IV than OLS if the correlation between z and u is smaller than that between x and u.

Properties of IV with a Poor IV

Using the fact that $Corr(x, u) = Cov(x, u)/(\sigma_x \sigma_u)$ along with $Cov(x, u) \neq 0$, we can write the plim of OLS estimator ($\tilde{\beta}_1$) as:

$$\text{plim}\, \tilde{\beta}_1 = \beta_1 + Corr(x, u) \cdot \frac{\sigma_u}{\sigma_x}$$

Comparing these formulae shows that IV is preferred to OLS on asymptotic bias grounds when $Corr(z, u)/Corr(z, x) < Corr(x, u)$. Recall Angrist & Krueger (1991) where x is years of schooling and z is binary indicator of quarter of birth, $Corr(z, x)$ very small. Bound, Jaeger & Baker (1995) discussed reasons why quarter of birth and u might be somewhat correlated. From equation (9), we see that this can lead to a substantial bias in the IV estimator. When z and x are not correlated at all, things are especially bad whether or not z is uncorrelated with u. Always check to see if the endogenous explanatory variable is correlated with the IV candidate – example 15.3.

IV: R-squared

- Stata: $R^2 = 1 - SSR/SST$ can be negative (if $SSR > SST$).
- Not very useful to report R-squared for IV estimation.
- When x and u are correlated, can't decompose variance of y into $\beta_1^2 Var(x) + Var(u)$ so R-squared has no natural interpretation.
- These R-squareds can't be used in usual way to compute F statistics of joint restrictions.
- Goodness of fit is not a factor – goal of IV is to provide better estimates of *cet. par.* effect of x on y when x and u are correlated.

Lecture 7 Outline

IV Estimation in Multiple Regression

- Initially, only one explanatory variable is correlated with the error:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{10}$$

  called a **structural equation**: emphasis on interest in the β_j, i.e. equation is supposed to measure a causal relationship.
- Notation to distinguish endogenous from **exogenous variables**.
- Assume $E(u_1) = 0$. Use z_1 to indicate that this variable is exogenous in (10) (z_1 is uncorrelated with u_1).
- Use y_2 to indicate that this variable is suspected of being correlated with u_1.
- Don't specify why y_2 and u_1 are correlated, but think for now of u_1 as containing an omitted variable correlated with y_2.
- Notation originates in SEM (chapter 16).

IV Estimation in Multiple Regression

Example

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + u_1 \tag{11}$$

i.e. assume *exper* is exogenous in (11) but allow *educ* to be correlated with u_1 for usual reasons. As z_1 uncorrelated with u_1, can we use z_1 as an instrument for y_2 assuming y_2 and z_1 are correlated? No.
Key ass are that z_1 and z_2 uncorrelated with u_1, $E(u_1) = 0$ (WLOG) when equation contains an intercept:

$$E(u_1) = 0, Cov(z_1, u_1) = 0, Cov(z_2, u_1) = 0 \tag{12}$$

Latter two assumptions $\equiv E(z_1 u_1) = E(z_2 u_1) = 0$ so solve:

$$\sum_{i=1}^{n} (y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0 \tag{13}$$

$$\sum_{i=1}^{n} z_{i1}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0 \tag{14}$$

$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) = 0 \tag{15}$$

Estimators are called *instrumental variables estimators*.

IV Estimation in Multiple Regression

Example

If think y_2 exogenous and choose $z_2 = y_2$, these equations are same as FOC for OLS. Still need IV z_2 correlated with y_2, but correlation is complicated by presence of z_1 in equation (10). Need to state assumptions in terms of *partial* correlation. Easiest: write endogenous explanatory variable as a linear function of the exogenous variable and an error term:

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2 \tag{16}$$

$E(v_2) = 0$, $Cov(z_1, v_2) = 0$ and $Cov(z_2, v_2) = 0$ and π_j are unknown parameters. Key identification condition along with (12) is:

$$\pi_2 \neq 0 \tag{17}$$

i.e. after partialling out z_1, y_2 and z_2 are still correlated. Correlation can be positive or negative but not 0.

IV Estimation in Multiple Regression

Example

Test (17): estimate (16) by OLS and use t test. Unfortunately, can't test that z_1 and z_2 are uncorrelated with u_1; hopefully, we can make the case based on economic reasoning or introspection. Equation (16) is an example of a **reduced form equation**, i.e. endogenous variables in terms of exogenous variables. Name comes from SEM and distinguishes it from structural equation (10). Question 15.2. Adding more **exogenous explanatory variables** to the model is straightforward, structural model:

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \ldots + \beta_k z_{k-1} + u_1 \tag{18}$$

where y_2 is thought to be corr with u_1. Let z_k be a variable not in (18) that's also exogenous. So, we assume that:

$$E(u_1) = 0, Cov(z_j, u_1) = 0, j = 1, \ldots, k \tag{19}$$

IV Estimation in Multiple Regression

Example

Under (19), $z_1, \ldots, z_{k-1}$ are exogenous variables appearing in (18). These act as their own IV in estimating β_j in (18). Special case $k = 2$ given in equations (13)- (15); along with z_2, z_1 appears in set of moment conditions used to obtain IV estimates. More generally, $z_1, \ldots, z_{k-1}$ used in moment conditions along with IV for y_2, z_k. Reduced form for y_2 is:

$$y_2 = \pi_0 + \pi_1 z_1 + \ldots + \pi_{k-1} z_{k-1} + \pi_k z_k + v_2 \tag{20}$$

partial correlation between z_k and y_2:

$$\pi_k \neq 0 \tag{21}$$

Under (19) and (21), z_k is a valid IV for y_2. Don't care about remaining π_j in (20); some or all of them could be zero. Minor additional assumption is that there are no perfect linear relationships among the exogenous variables; this is analogous to the assumption of no perfect collinearity in context of OLS. Homogeneity of u_1. Example 15.4.

Lecture 7 Outline

Summary

- IV yields consistent estimators under omitted variables.
- Weak identification: weak correlation between z and x.
- Structural equation measures a causal relationship (can have endogenous variables on both sides).
- Reduced form equations express endogenous variables in terms of predetermined (exogenous and / or lagged endogenous) variables.
- Final form equations express endogenous variables in terms of purely exogenous variables.

References

- IV & 2SLS (i): Wooldridge 15.1-2.