USC, Fall 2016, Economics 513

Lecture 8: Instrumental variables

Endogenous and exogenous independent variables

In lecture 7 we found that if relevant independent variables that are correlated with included independent variables are omitted, then the OLS estimator of the regression coefficients of the included independent variables is biased. The OLS estimator converges in probability to the sum of the partial/direct effect and the indirect effect (through the omitted variable) of a change in an independent variable.

If x is correlated with the error term, then we say that x is *endogenous*. We call x *exogenous* if this variable is not correlated with the random error. This terminology derives from simultaneous equations econometric models in which the variables determined by the system of equations (endogenous variables) are correlated with the random errors.

Endogeneity of x is common if the regression model describes the relation between an outcome variable y and an independent variable x that is under control of an agent while some other variables that are not observed by us, but known to the agent, say a affect both the outcome and the choice of x. You can call this a case of asymmetric information between us and the agent.

Example 1: Earnings and education. Here y is earnings, x is years of education and a is ability. It is obvious that x is chosen to (among other objectives) increase y. If the return to education depends on a, then x is chosen to depend on a.

Example 2: Production function and labor input. Here y is firm output, x is labor input and a is management quality. If the firm maximizes profits and the labor productivity depends on management quality, then the choice of x depends on a. Management quality also has a direct effect on output and therefore x is endogenous.

In assignment 2 you study the use of proxy variables for the omitted variable(s) to eliminate or reduce the bias of the OLS estimator.

In this lecture we consider another method to deal with the omitted variable bias. This method requires that we have a special variable z.

Instrumental variables

Initially we consider the relation between a dependent variable y and a single independent variable x. For instance, y is (log) weekly earnings and x is years of education. The possible omitted variable is ability (but there may be more).

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \quad i = 1, \ldots, n$$

The independent variable x is potentially correlated with ε that contains ability. Therefore it is an endogenous variable.

The special variable z satisfies two conditions:

1. The variable z does not affect the dependent variable directly. It is also not correlated with the omitted variables in the relation under consideration.

2. The variable z is correlated with x.

Condition 1 implies that z is not an omitted variable and therefore not in ε and z is also not correlated with the omitted variables, e.g. ability, in the relation. Therefore this condition implies

$$\mathrm{E}(z\varepsilon) = 0$$

and this is often taken as the condition instead of condition 1. However to gauge the validity of this condition keep condition 1 in mind.

A variable that satisfies conditions 1 and 2 is called an *instrumental variable*. Because condition 1 involves unobserved variables, it can not be tested. Instead we have to make a convincing case that this condition holds.

Example: Quarter of of birth. Angrist and Krueger (QJE, 1991) argue that due to compulsory schooling laws (that in the US differ by state) individuals who are born in the first quarter have on average fewer years of education than individuals born in later quarters. The reason is that a child starts school in the year that she/he turns 5 or 6, but the laws specify that compulsory schooling ends at the day that she/he turns 15 or some later age.

Condition 1 is satisfied if quarter of birth has no effect on weekly earnings and is not correlated with the omitted variables in the relation, e.g. ability. Bound, Jaeger and Baker (JASA, 1995) mention that there is some evidence of (weak) relations between quarter of birth and school attendance, mental health, and parental income and this would invalidate condition 1.

The data that they use are 329509 observations from the census. The variables observed are weekly earnings (dependent variable) and years of education, year of birth, state of birth (independent variables).

Condition 2 can be checked by a regression of years education on quarter of birth. We can do this in two ways: quarter of birth coded as 1 to 4 or using quarter of birth indicators (dummy variables).

$$\begin{array}{rcccc} \widehat{\text{educ}}_i & = & 12.6407 & + & 0.0516\cdot\text{qob}_i \\ & & (0.00141) & & (0.0051) \end{array}$$

or

$$\begin{array}{rcccccccc} \widehat{\text{educ}}_i & = & 12.6881 & + & 0.0566\cdot\text{qob}_{2i} & + & 0.1173\cdot\text{qob}_{3i} & + & 0.1514\cdot\text{qob}_{4i} \\ & & (0.0115) & & (0.0163) & & (0.0160) & & (0.0163) \end{array}$$

Note that the effect is small: with a school year of 200 days it is about 12 days. It is highly significant in the (asymptotic) t-test because we have a very large sample.

The instrumental variable estimator

We use the conditions $\text{E}(\varepsilon) = 0$ (why is this a free assumption in this case?) and $\text{E}(z\varepsilon) = 0$ to define an estimator $\hat{\beta}$ by replacing the population mean by a sample average and ε by an equation residual $e = y - \hat{\beta}_1 - \hat{\beta}_2 x$

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i) &= 0 \\ \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)z_i &= 0 \end{aligned}$$

The solution is

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(z_i - \overline{z})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z})}$$

$$\hat{\beta}_1 = \overline{y} - \hat{\beta}_2\overline{x}$$

If we divide numerator and denominator by $n-1$ $\hat{\beta}_2$ is the ratio of the sample covariances of x and y and that of x and z. If $n \to \infty$ by the LLN

$$\hat{\beta}_2 \xrightarrow{p} \frac{\text{Cov}(z,y)}{\text{Cov}(z,x)} = \beta_2$$

because

$$\text{Cov}(z,y) = \beta_2\text{Cov}(z,x)$$

If $\text{Cov}(z,x) = 0$, then this result does not hold and it is not clear what the probability limit of $\hat{\beta}_2$ is in this case. This is the reason we need condition 2.

Also we have by substituting the relation in $\hat{\beta}_2$ that

$$\hat{\beta}_2 = \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \overline{x})(z_i - \overline{z})}$$

Note that if we make the stronger assumption $\mathrm{E}(\varepsilon_i|z_i) = 0$, then this is not sufficient to show that this estimator is unbiased. We can only prove consistency.

This estimator of β is called the *instrumental variable* estimator.

We now consider a relation that has a single endogenous independent variable and in addition an intercept and $K-2$ exogenous independent variables. Let x_K be the endogenous independent variable. Note that $\mathrm{E}(x_k\varepsilon) = 0$ for $k = 1, \ldots K-1$ (as usual we take $x_1 \equiv 1$). Exogenous variables are valid instrumental variables. However $\mathrm{E}(x_K\varepsilon) \neq 0$. We (note the change in notation) have an instrumental variable w. We define the vector of instrumental variables

$$z_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{i,K-1} \\ w_i \end{pmatrix}$$

We have

$$\mathrm{E}(z_i\varepsilon_i) = 0$$

and using this to define $\hat{\beta}$ we find

$$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_i'\hat{\beta}) = 0$$

or

$$\left(\sum_{i=1}^{n} z_i x_i'\right)\hat{\beta} = \sum_{i=1}^{n} z_i y_i$$

Defining the $n \times K$ matrices (x_i is the $K \times 1$ vector of independent variables for observation i)

$$X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix} \qquad Z = \begin{pmatrix} z_1' \\ \vdots \\ z_n' \end{pmatrix}$$

we can express this as

$$Z'X\hat{\beta} = Z'y$$

To solve this equation we need to invert the $K \times K$ matrix $Z'X$. Therefore we need to assume

$$\mathrm{rank}(Z'X) = K$$

This generalizes condition 2.

With this condition we can solve for the instrumental variables (IV) estimator

$$\hat{\beta} = (Z'X)^{-1}Z'y$$

The asymptotic distribution of the instrumental variables estimator

We have

$$\hat{\beta} = \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} z_i y_i \right) = \beta + \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} z_i \varepsilon_i \right)$$

Now by the LLN

$$\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \xrightarrow{p} \begin{pmatrix} 1 & \mu_x' \\ \mu_z & \Omega_{zx} \end{pmatrix}$$

with Ω_{zx} the $(K-1) \times (K-1)$ matrix with j, k-th component $\mathrm{E}(z_j x_k)$. Also by the LLN

$$\frac{1}{n} \sum_{i=1}^{n} z_i \varepsilon_i \xrightarrow{p} 0$$

Combining these results we conclude that the IV estimator is consistent.

Also

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \varepsilon_i \right)$$

Using the Cramér-Wold device we conclude from the CLT

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} z_i \varepsilon_i \xrightarrow{d} N(0, \mathrm{E}(\varepsilon^2 z z'))$$

so that the large sample variance matrix of $\hat{\beta}$ is

$$V = \begin{pmatrix} 1 & \mu_x' \\ \mu_z & \Omega_{zx} \end{pmatrix}^{-1} \mathrm{E}(\varepsilon^2 z z') \begin{pmatrix} 1 & \mu_x' \\ \mu_z & \Omega_{zx} \end{pmatrix}^{-1}$$

In the derivation we do not make an assumption about the (conditional) variance of the random error. If we assume

$$\mathrm{Var}(\varepsilon_i | z_i) = \sigma^2$$

then

$$V = \sigma^2 \begin{pmatrix} 1 & \mu_x' \\ \mu_z & \Omega_{zx} \end{pmatrix}^{-1} \mathrm{E}(z z') \begin{pmatrix} 1 & \mu_x' \\ \mu_z & \Omega_{zx} \end{pmatrix}^{-1}$$

Obvious consistent estimators for V are

$$\hat{V} = \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} e_i^2 z_i z_i' \right) \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1}$$

and

$$\hat{V} = \hat{\sigma}^2 \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} z_i z_i' \right) \left(\frac{1}{n} \sum_{i=1}^{n} z_i x_i' \right)^{-1}$$

if $\mathrm{Var}(\varepsilon_i|z_i) = \sigma^2$. We estimate

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{i=1}^{n} e_i^2$$

We could also divide by n.

The approximate sampling distribution of the IV estimator is

$$\hat{\beta} \approx N(\beta, \frac{1}{n}\hat{V})$$

The general case: Two-stage Least Squares (2SLS)

We consider a relation with K_1 exogenous variables in the vector x_1 and K_2 endogenous variables in the vector x_2, i.e.

$$y_i = x_{i1}'\beta_1 + x_{i2}'\beta_2 + \varepsilon_i$$

Because x_1 is exogenous we have $\mathrm{E}(x_{i1}\varepsilon_i) = 0$. We have L_1 instrumental variables in the vector w. Therefore $\mathrm{E}(w_i\varepsilon_i) = 0$. For $L = (K_1 + L_1)$ we define the $L \times 1$ vector z by

$$z_i \equiv \begin{pmatrix} x_{i1} \\ w_i \end{pmatrix}$$

We have $\mathrm{E}(z_i\varepsilon_i) = 0$. This is called an (unconditional) moment condition.

To obtain an estimator we replace the population expectation by the sample average and ε_i by $e_i = y_i - x_{i1}'\hat{\beta}_1 - x_{i2}'\hat{\beta}_2$

$$\frac{1}{n}\sum_{i=1}^{n} z_i(y_i - x_{i1}'\hat{\beta}_1 - x_{i2}'\hat{\beta}_2) = 0$$

This is called an sample (unconditional) moment condition. The estimator that solves the sample moment condition is called the Generalized Method of Moments (GMM) estimator. The instrumental variable estimator is a GMM estimator (as is OLS).

This are L equations in K unknowns. There may be a solution if $L = K$. What if $L < K$? And if $L > K$?

We have $L \gtreqless K$ if the number of instrumental variables is greater than, equal to or less than the number of endogenous variables.

As before X is the $n \times K$ matrix ($K = K_1 + K_2$) with the observations on the exogenous and endogenous independent variables and Z is $n \times L$. Therefore the system of equations in matrix notation is

$$Z'(y - X\hat{\beta}) = 0$$

If $L > K$ then there is in general no solution the the system of equations. Instead we minimize

$$S(\hat{\beta}) = (y - X\hat{\beta})' ZWZ'(y - X\hat{\beta})$$

The $L \times L$ matrix W is called the weighting matrix and be chosen by us (it has to be symmetric and positive definite). For instance we can choose $W = I$ the $L \times L$ identity matrix. An obvious idea is to choose W such that the variance of the estimator $\hat{\beta}$ is minimal.

The first order condition for this minimization problem is

$$2(y - X\hat{\beta})' ZWZ'X = 0$$

or

$$X'ZWZ'X\hat{\beta} = X'ZWZ'y$$

with solution

$$\hat{\beta} = (X'ZWZ'X)^{-1}X'ZWZ'y$$

This is the GMM estimator for the population moment conditions given above.

Substitution of $y = X\beta + \varepsilon$ gives

$$\sqrt{n}(\hat{\beta}-\beta) = \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right) W \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right) W \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i \varepsilon_i\right)$$

By the same argument as for the IV estimator we find

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, V)$$

where V can be consistently estimated by the rather long matrix expression

$$\hat{V} = \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right) W \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)\right]^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right) W \left(\frac{1}{n}\sum_{i=1}^{n} e_i^2 z_i z_i'\right) \cdot$$

$$W \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right) \left[\left(\frac{1}{n}\sum_{i=1}^{n} x_i z_i'\right) W \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)\right]^{-1}$$

so that the variance matrix of the approximate distribution of $\hat{\beta}$ is

$$\frac{1}{n}\hat{V} = (X'ZWZ'X)^{-1}X'ZW\hat{\Sigma}WZ'X(X'ZWZ'X)^{-1}$$

with

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} e_i^2 z_i z_i'$$

This is the heteroskedasticity-robust variance matrix.

If we assume homoskedasticity the consistent estimator of the variance matrix is

$$\hat{V} = \hat{\sigma}^2 (X'ZWZ'X)^{-1} X'ZWZ'ZWZ'X(X'ZWZ'X)^{-1}$$

with

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} e_i^2$$

$$e_i = y_i - x_{i1}' \hat{\beta}_1 - x_{i2} \hat{\beta}_2$$

The choice of W that minimizes the (asymptotic) variance of the GMM estimator is

$$W = (Z'Z)^{-1}$$

The formula for the GMM estimator for this choice of W is

$$\hat{\beta} = (X'Z(Z'Z)^{-1}Z'X)^{-1} X'Z(Z'Z)^{-1}Z'y$$

This is not much simpler, but the expression has an interesting interpretation.

Note that if $L = K$, i.e. the number of instrumental variables is equal to the number of endogenous variables, then

$$\hat{\beta} = (Z'X)^{-1} Z'y$$

i.e. it is equal to the IV estimator with matrix of observations on the instrumental variables Z.

If X were a vector, then the OLS estimator with X as the vector of observations on the dependent variable and Z the matrix with observations on the independent variable is $(Z'Z)^{-1}Z'X$ so that

$$\hat{X} = Z(Z'Z)^{-1}Z'X$$

If X is a matrix then $\hat{X}$ is the matrix with columns that are equal to the vectors of the predicted values if we regress a column of X on Z.

Note that there are two cases: the column of X corresponds to an exogenous variable and the column of X corresponds to an endogenous variable. What is the column of $\hat{X}$ in the first case?

We can rewrite the GMM estimator for $W = (Z'Z)^{-1}$ as

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

This is the IV estimator with instrument matrix $\hat{X}$.

The GMM estimator with $W = (Z'Z)^{-1}$ is called the Two-Stage Least Squares (2SLS) estimator. The name derives that we can think of the estimator having two steps or stages:

1. Regress the columns of X on Z and compute $\hat{X}$.

2. Use $\hat{X}$ as the instrumental variables matrix in the IV estimator of β.

Because

$$X'Z(Z'Z)^{-1}Z'X = \hat{X}'\hat{X}$$

we could even replace the second step by: Regress y on $\hat{X}$. Although formally correct it can be misleading when we compute the variance of the estimators. In particular, the 2SLS residuals are

$$e_i = y_i - x_{i1}'\hat{\beta}_1 - x_{i2}\hat{\beta}_2$$

and not

$$e_i \neq y_i - \hat{x}_{i1}'\hat{\beta}_1 - \hat{x}_{i2}\hat{\beta}_2$$

The heteroskedasticity-consistent variance matrix of the 2SLS estimator does not simplify of we substitute $W = (Z'Z)^{-1}$, but the homoskedastic variance matrix does: the estimator of the variance matrix is

$$\hat{V} = \hat{\sigma}^2(X'Z(Z'Z)^{-1}Z'X)^{-1} = \hat{\sigma}^2(\hat{X}'\hat{X})^{-1}$$

This seems the variance matrix of the OLS estimator of y on $\hat{X}$, except that in the estimator of $\hat{\sigma}^2$ we cannot use the residuals from that regression.

Finite sample properties of the IV and 2SLS estimators

If there are omitted variables correlated with the included variables then

1. OLS is biased.

2. If we have an instrumental variable, then IV/2SLS is consistent (but potentially biased in finite samples).

3. The variance of IV/2SLS is larger than that of OLS (see below).

What is worse: the asymptotic bias in OLS or the finite sample bias in IV/2SLS? Is the sampling variance of IV/2SLS much larger?

To develop intuition we consider the case with a single endogenous variable and a single instrument. In that case for the coefficient on x

$$\sqrt{n}(\hat{\beta}_2 - \hat{\beta}_2) \xrightarrow{d} N(0, V)$$

with

$$V = \frac{\sigma^2}{\sigma_x^2 \rho_{xz}^2}$$

Note that for the OLS estimator

$$V = \frac{\sigma^2}{\sigma_x^2}$$

so that the variance of the IV/2SLS estimator is always larger than that of the OLS estimator.

How much greater depends on ρ_{xz}^2, i.e. on the correlation between x and z. If the correlation between x and z is small we say that z is a weak instrument. If the correlation is large we say that it is a strong instrument.

Note that we can check whether the instrument is weak or strong by regressing x on the instrument and the exogenous variables, i.e. the first stage of the 2SLS estimator and looking at the size of the t-statistic of the F-statistic of the test that all coefficients except the intercept are 0.

We study this in the Angrist-Krueger data.

First we estimate the relation by 2SLS using either QOB or the QOB dummies as instruments.

$$\begin{array}{rcccc} \log(\widehat{\text{earnings}})_i & = & 4.5898 & + & 0.1026 \cdot \text{educ}_i \\ & & (0.2490) & & (0.0195) \end{array}$$

The first stage t-statistic is 10.03 and F-statistic 100.7 (square of t-statistic).

$$\begin{array}{rcccc} \log(\widehat{\text{earnings}})_i & = & 4.6329 & + & 0.09922 \cdot \text{educ}_i \\ & & (0.2501) & & (0.0196) \end{array}$$

The first stage F-statistic is 34.0.
We can compare this to the OLS estimates

$$\begin{array}{rcccc} \log(\widehat{\text{earnings}})_i & = & 4.9952 & + & 0.0709 \cdot \text{educ}_i \\ & & (0.0045) & & (0.0003) \end{array}$$

Note that the IV estimator is larger than the OLS estimator. Also note that the IV standard errors are much larger than the OLS ones.

For the finite sample behavior we draw samples (without replacement) of size 5000 (65 datasets), 10000 (32 datasets) and 20000 (16 datasets) from the 329507 observations. For each dataset we compute the OLS and 2SLS/IV estimates using QOB as instrument.

	$n = 5000$	$n = 10000$	$n = 20000$
EDUC OLS	.0708	.0709	.0709
Sampling std. dev.	.0031	.0022	.0018
Av. std. error	.0027	.0019	.0014
First stage t-stat	1.24	1.75	2.47
EDUC 2SLS	.3367	.1550	.1055
Sampling std. dev.	1.8983	1.5578	.3099
Av. std. error	9.8379	6.1271	.3496

How do we find instrumental variables

In Angrist and Krueger the instrument was plausible, but arguing that is sometimes hard. More convincing are (i) random assignment, (ii) instruments derived from economic theory.

For (i) consider an intervention that is randomly assigned to members of a population. Those members can either chose to participate in the intervention (comply) or chose not to participate (not comply). Members not selected for the intervention can choose to participate (not comply) or not to participate (comply) in the intervention. The 0-1 variable z that is 1 if selected and 0 if not is an instrumental variable. The indicator x for the intervention is endogenous.

For (ii) consider the production function example. If the firm is a price taker on input markets, then the wage rate would be an instrumental variable. Note that this requires that there is variation of the wage rate between firms, e.g. because they are at different locations (on the assumption that the production function is independent of location).