Intermediate Econometrics 06A

Chen Zhu

College of Economics and Management
China Agricultural University

Endogeneity

Problem of endogenous explanatory variables

1. Endogeneity is said to occur in a multiple regression model if

$$E(u|X) \neq 0 \quad \text{for some j} = 1, \ldots, \text{k}$$

   - when there is a correlation between your X variable and the error term in your model
   - when there is something that is related to your Y variable that is also related to your X variable, and you do not have that something in your model

2. Sources:
   - omitted variables
   - measurement error
   - simultaneity (i.e. X causes Y but Y also causes X; self-selection)

3. OLS estimates will be biased and inconsistent.

Omitted variables in a simple regression model

When faced with the prospect of omitted variables bias (or unobserved heterogeneity), we have options:

1. Use a suitable proxy variable for the unobserved variable
2. Instrumental variable method

Omitted variables in a simple regression model

Consider the problem of unobserved ability in a wage equation for working adults

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

e is the error term.

1. Ability is unobservable
2. Under certain assumptions, a proxy variable such as IQ can be substituted for ability
3. Then a consistent estimator of β_1 is available from the regression of

$$\log(wage) \text{ on } educ, IQ$$

Example

1. The data contains 935 men in 1980 from the Young Men's Cohort of the National Longitudinal Survey (NLSY), USA
2. The results from the regression with omitting ability variable are

Log(wage)	Coeff.	Std. Err.
Education	0.078	0.007
Experience	0.020	0.003
Constant	5.503	0.112

3. The estimated return to education is 7.8%

Example

1. The results from the regression with the proxy variable IQ for ability are

Log(wage)	Coeff.	Std. Err.		Coeff.	Std. Err.
Education	0.078	0.007		0.057	0.007
Experience	0.020	0.003		0.020	0.003
IQ	—	—		0.006	0.001
Constant	5.503	0.112		5.198	0.122

2. The estimated return to education changes from 7.8% to 5.7%

The instrumental variables approach

1. Suppose, however, that a proxy variable is not available
2. Then, we put *abil* into the error term, and we are left with the simple regression model

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

u contains *abil*; β_1 from OLS is biased and inconsistent because *educ* and *abil* are correlated.

The instrumental variables approach

1. We can still use the above specification as the basis for estimation, by finding an instrumental variable for *educ*
2. To describe this approach, the simple regression model is written as

$$y = \beta_0 + \beta_1 x + u, \tag{1}$$

where we think that x and u are correlated:

$$\text{Cov}(x, u) \neq 0.$$

Definition

1. In order to obtain consistent estimators when x and u are correlated, we need some additional information
2. Suppose that we have an observable variable z that satisfies these two assumptions:
   - z is uncorrelated with u

$$\text{Cov}(z, u) = 0; \tag{2}$$

   - z is correlated with x

$$\text{Cov}(z, x) \neq 0. \tag{3}$$

3. We call z an instrumental variable (IV) for x, or sometimes simply an instrument for x

The instrumental variables approach

1. For the log(*wage*) equation, an instrumental variable z for *educ* must be
   - uncorrelated with ability (and any other unobserved factors affecting wage)
   - correlated with education
2. How about the last digit of an individual's mobile phone number? Can it be a valid IV for *educ*?

The instrumental variables approach

1. No
   - the last digit of an individual's mobile phone number satisfies the first requirement: it is uncorrelated with ability because it is determined randomly
   - but it is not correlated with education, either $\rightarrow$ poor IV for *educ*
2. How about a proxy variable for the omitted variable? Can *IQ* be a valid IV for *educ*?

The instrumental variables approach

1. No
   - the proxy variable *IQ* is highly correlated with *abil* in $u \rightarrow$ poor IV
2. Remember that an instrumental variable must be uncorrelated with *abil* in u

Difference between IV and Proxy

1. With IV we will leave the unobserved variable in the error term but use an estimation method that recognizes the presence of the omitted variable
2. With a proxy we were trying to remove the unobserved variable from the error term
3. Need something correlated with education but uncorrelated with ability

Identification

1. Identification of a parameter means that we can write β_1 in terms of population moments that can be estimated using a sample of data
2. The availability of an instrumental variable can be used to estimate consistently the parameters
3. From equation (1), we have

$$\text{Cov}(z, y) = \beta_1 \text{Cov}(z, x) + \text{Cov}(z, u) \tag{4}$$

Since $\text{Cov}(z, u) = 0$ and $\text{Cov}(z, x) \neq 0$, we can solve for β_1 as

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

Identification

1. After canceling the sample sizes in the numerator and denominator, we get the IV estimator of β_1

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

2. Given a sample of data on x, y, and z, it is simple to obtain the IV estimator in equation (6)
3. The IV estimator of β_0 is $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
4. If either assumption of IV fails, the IV estimators are not consistent

Statistical inference with the IV estimator

1. To perform inference on β_1, we need a standard error that can be used to compute t statistics and confidence intervals
2. The homoskedasticity assumption is stated conditional on the instrumental variable z, not the endogenous explanatory variable x
3. Along with the previous assumptions on u, x, and z, we add

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

Statistical inference with the IV estimator

1. Under (2), (3), and (7), the asymptotic variance of $\hat{\beta}_1$ is

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2} \tag{8}$$

   where σ_x^2 is the population variance of x, σ^2 is the population variance of u, and $\rho_{x,z}^2$ is the square of the population correlation between x and z
2. Equation (8) provides a way to obtain a standard error for the IV estimator
3. All quantities in (8) can be consistently estimated given a random sample
4. Cost: when x and u are uncorrelated, the asymptotic variance of the IV estimator is always larger, and sometimes much larger, than the asymptotic variance of the OLS estimator

Statistical inference with the IV estimator

1. To estimate σ_x^2, we simply compute the sample variance of x_i
2. To estimate $\rho_{x,z}^2$, we can run the regression of x_i on z_i to obtain the R-squared, $R_{x,z}^2$
3. To estimate σ^2, we can use the IV residuals

$$\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, \quad i = 1, 2, ..., n, \tag{9}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the IV estimates

Statistical inference with the IV estimator

1. A consistent estimator of σ^2 looks just like the estimator of σ^2 from a simple OLS regression:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \hat{u}_i^2 \tag{10}$$

2. The asymptotic standard error of $\hat{\beta}_1$ is the square root of the estimated asymptotic variance, which is given by

$$\frac{\hat{\sigma}^2}{SST_x \times R^2_{x,z}} \tag{11}$$

Example: estimating the return to education

1. Suppose we use the data to estimate the return to education for married women in the simple regression model

$$\log(wage) = \beta_0 + \beta_1 educ + u, \tag{12}$$

2. We first obtain the OLS estimates:

$$\begin{aligned} \widehat{\log(wage)} &= -0.185 + 0.109 educ, \\ &\quad (0.185) \quad (0.014) \\ n &= 428, \ R^2 = 0.118. \end{aligned}$$

   It implies an almost 11% return for another year of education.

3. But education is potentially endogenous

Return to education

1. We use father's education (*fatheduc*) as an instrumental variable for *educ*
2. We have to maintain that *fatheduc* is uncorrelated with *u*
3. The second requirement is that *educ* and *fatheduc* are correlated
4. We can check this using a simple regression of *educ* on *fatheduc*:

$$\begin{aligned} \hat{educ} =& 10.24 + 0.269 fatheduc, \\ & (0.280) \quad (0.029) \\ n = 428,\ & R^2 = 0.173. \end{aligned}$$

The *t* statistic on *fatheduc* is 9.28, which indicates that *educ* and *fatheduc* have a statistically significant positive correlation.

Return to education

1. Using *fatheduc* as an IV for educ gives:

$$\begin{aligned}\widehat{\log(wage)} =& 0.441 + 0.059 educ, \\ & (0.446) \quad (0.035) \\ n = 428,\ & R^2 = 0.093.\end{aligned}$$

The IV estimate of the return to education is 5.9%.

Other possible IVs for *educ*

1. Number of siblings (negative correlation)
2. Month of born (Angrist and Krueger, 1991)
   - Let *frstqrt* be equal to one if the man was born in the first quarter of the year, and zero otherwise.
   - It seems that ability should be unrelated to quarter of birth.
   - *frstqrt* also needs to be correlated with *educ* (Years of education do differ systematically in the population based on quarter of birth in sample).
   - The compulsory school attendance laws: students born early in the year typically begin school at an older age. Therefore, they reach the compulsory schooling age with somewhat less education than students who begin school at a younger age.
   - For students who finish high school, Angrist and Krueger verified that there is no relationship between years of education and quarter of birth.

Multiple IV Regression Model

1. The IV estimator for the simple regression model is easily extended to the multiple regression case
2. We begin with the case where only one of the explanatory variables is correlated with the error
3. Consider a standard linear model with two explanatory variables

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1 \tag{13}$$

- We call this a structural equation to emphasize that we are interested in the β_j, which simply means that the equation is supposed to measure a causal relationship
- y_1 is the dependent variable. y_2 and z_1 are the explanatory variables. u_1 is the error term
- We assume z_1 is exogenous, and y_2 is endogenous (i.e. correlated with u_1)

Multiple IV Regression Model

1. If equation (13) is estimated by OLS, all of the estimators will be biased and inconsistent
2. Thus, we need to seek an instrumental variable for y_2
   - Since z_1 itself appears as an explanatory variable in (13), it cannot serve as an instrumental variable for y_2
   - We need another exogenous variable that does not appear in (13), denoted by z_2
3. Key assumptions:
   - $E(u_1) = 0$
   - $Cov(z_1, u_1) = 0 \implies E(z_1 u_1) = 0$
   - $Cov(z_2, u_1) = 0 \implies E(z_2 u_1) = 0$

IV Estimator

1. $\hat{\beta}$s can be solved by the sample counterparts of the key assumptions

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

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

$$\sum_{i=1}^{n} z_{i2}(y_{i1} - \hat{\beta}_0 - \hat{\beta}_1 y_{i2} - \hat{\beta}_2 z_{i1}) \tag{16}$$

2. This is a set of three linear equations in the three unknowns, and it is easily solved given the data on y_1, y_2, z_1, and z_2
3. The estimators are called instrumental variables estimators

Tests of IV Assumptions

1. We need the instrumental variable z_2 to be correlated with y_2
2. We can write the endogenous explanatory variable as a linear function of the exogenous variables and an error term

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \nu_2 \tag{17}$$

where, by construction, E(ν_2)=0, Cov(z_1, ν_2) =0, and Cov(z_2, ν_2)=0, and the π_j are unknown parameters

3. Thus, we need to show that

$$\pi_2 \neq 0 \tag{18}$$

- i.e. after partialling out z_1, y_2 and z_2 are still correlated
- OLS+t test

Tests of IV Assumptions

1. Equation (17) is an example of a reduced form equation, which means that we have written an endogenous variable in terms of exogenous variables
2. The name helps distinguish it from the structural equation (13)
3. ▶▶▶ Notice that we CANNOT test that z_1 and z_2 are uncorrelated with u_1.Instead, we have to rely on common sense and economic theory to decide if it makes sense

Other sources of endogeneity: measurement error

1. Data is often measured with error:
   - reporting errors
   - coding errors
2. When the measurement error is in the dependent variable, the zero conditional mean assumption is not violated and thus no endogeneity
3. In contrast, when the measure error is in the independent variable, the problem of endogeneity arises

Measurement error

1. Consider a simple regression model:

$$y = \beta_0 + \beta_1 x + u$$

2. Suppose x is measured with errors. That is, we observe $\tilde{x} = x + \epsilon$ instead of x
3. We assume that ϵ is uncorrelated with x, $E(x\epsilon) = 0$
4. Then, the regression equation we use is

$$y = \beta_0 + \beta_1 \tilde{x} + u - \beta_1 \epsilon \quad (19)$$

$$= \beta_0 + \beta_1 \tilde{x} + v \quad (20)$$

Measurement error

1. It can be seen that the problem of endogeneity occurs:

$$E(\tilde{x}v) = E((X+\epsilon)(u-\beta_1\epsilon)) \quad (21)$$

$$= -\beta_1 Var(\epsilon) \neq 0 \quad (22)$$

2. The measurement error leads to a biased OLS estimate towards zero. This is called attenuation bias.
3. The OLS estimator will be inconsistent

Simultaneity

1. Simultaneity arises when one or more of the independent variables is jointly determined with the dependent variable, typically through an equilibrium mechanism.
2. This arises in many economic contexts:
   - quantity and price by demand and supply
   - investment and productivity
   - sales and advertisement
   - police and crime

Two Stage Least Squares (2SLS)

1. It often happens that we have more than one exogenous variable that is excluded from the structural model and might be correlated with y_2, i.e. multiple IVs
2. Consider again the structural model (13), which has one endogenous and one exogenous explanatory variable. Suppose now that we have two exogenous variables excluded from (13): z_2 and z_3
3. Our assumptions that z_2 and z_3 do not appear in (13) and are uncorrelated with the error u_1 are known as exclusion restrictions
4. Since each of z_1, z_2, and z_3 is uncorrelated with u_1, any linear combination is also uncorrelated with u_1, and therefore any linear combination of the exogenous variables is a valid IV

2SLS

1. To find the best IV, we choose the linear combination that is most highly correlated with y_2
2. This turns out to be given by the reduced form equation for y_2

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \nu_2 \tag{23}$$

where $\mathrm{E}(\nu_2)=0$, $\mathrm{Cov}(z_1, \nu_2)=0$, $\mathrm{Cov}(z_2, \nu_2)=0$, and $\mathrm{Cov}(z_3, \nu_2)=0$

2SLS

1. Then, the best IV for y_2 is the linear combination of the z_j in (23), which we call y_2^*

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 \tag{24}$$

   For this IV not to be perfectly correlated with z_1, we need at least one of π_2 or π_3 to be different from zero

2. Given data on the z_j, we can compute y_2^* for each observation, i.e. $\hat{y}_2$

2SLS

1. Once we have $\hat{y}_2$, we can use it as the IV for y_2
2. With multiple instruments, the IV estimator using $\hat{y}_2$ as the instrument is also called the two stage least squares estimator