LECTURE 12

Introduction to Econometrics

Endogeneity

December 6, 2016

A little revision: OLS classical assumptions

1. The regression model is linear in coefficients, is correctly specified, and has an additive error term
2. The error term has a zero population mean
3. Observations of the error term are uncorrelated with each other
4. The error term has a constant variance
5. All explanatory variables are uncorrelated with the error term
6. No explanatory variable is a perfect linear function of any other explanatory variable(s)
7. The error term is normally distributed

On previous lectures

- We discussed what happens if some of the assumptions are violated
- Linearity of coefficients and no perfect multicollinearity are essential for the definition of OLS estimator
- Zero mean of the error term is always ensured by the inclusion of intercept
- Normality of the error term is needed for statistical inference, but it can be shown that if the number of observations is sufficiently high, the OLS estimate will have asymptotically normal distribution even if the stochastic error term is not normal
- Heteroskedasticity and serial correlation lead to incorrect statistical inference, but we have studied a set of techniques to overcome this problem

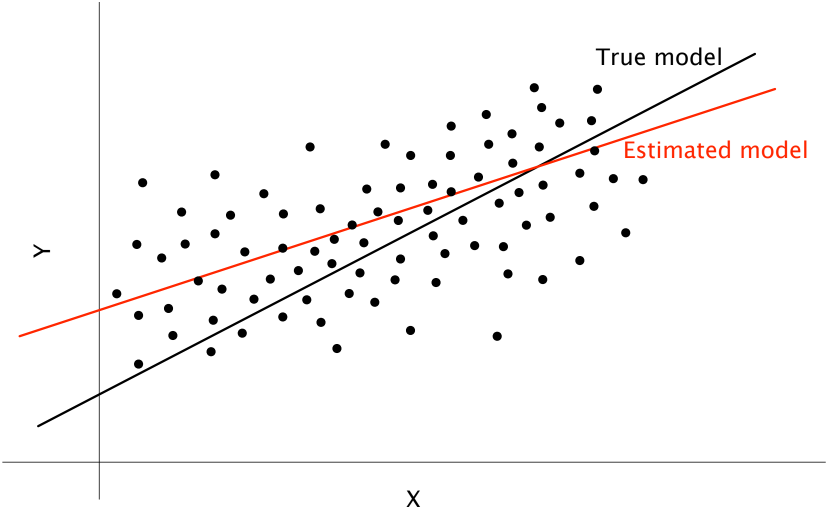
On today's lecture

- The assumption of no correlation between explanatory variables and the error term is crucial
- Variables that are correlated with the error term are called *endogenous variables* (as opposed to *exogenous variables*)
- We will show that the estimated coefficients of endogenous variables are inconsistent and biased
- We will explain in which situations we may encounter endogenous variables
- We will define the concept of instrumental variables
- We will derive the 2SLS technique to deal with endogeneity

Endogenous variables

- Notation: $E[x_i\varepsilon_i] = Cov(x_i, \varepsilon_i) \neq 0$ or $E[\mathbf{X}'\boldsymbol{\varepsilon}] \neq \mathbf{0}$
- Intuition behind the bias:
  - If an explanatory variable x and the error term ε are correlated with each other, the OLS estimate attributes to x some of the variation in y that actually came form the error term ε
- Example: Analysis of household consumption patterns
  - Households with lower income may indicate higher consumption (because of shame)
- Leads to inconsistent estimates

Graphical representation



TYPICAL CASES OF ENDOGENEITY

1. Omitted variable bias
   - An explanatory variable is omitted from the equation and makes part of the error term
2. Selection bias
   - An unobservable characteristic has influence on both dependent and explanatory variables
3. Simultaneity
   - The causal relationship between the dependent variable and the explanatory variable goes in both directions
4. Measurement error
   - Some of the variables are measured with error

- In all 4 cases, the sign of the bias is given by the sign of $Cov(\varepsilon_i, x_i)$

Omitted variable bias

- Studied on lecture 7
- True model: $y_i = \beta x_i + \gamma z_i + u_i$
- Model as it looks when we omit variable z:

$$y_i = \beta x_i + \tilde{u}_i \qquad \text{implying} \qquad \tilde{u}_i = \gamma z_i + u_i$$

- This gives

$$Cov(\tilde{u}_i, x_i) = Cov(\gamma z_i + u_i, x_i) = \gamma Cov(z_i, x_i) \neq 0$$

- It can be remedied by including the variable in question, but sometimes we do not have data for it
- We can include some proxies for such variable, but this may not reduce the bias completely and some endogeneity remains in the equation

Selection bias

- Very similar to omitted variable bias
- We suppose there is some unobservable characteristic that influences both the level of the dependent variable y and of the explanatory variable x
- This unobservable characteristic forms part of the error term ε, causing $Cov(\varepsilon, x) \neq 0$ (in the same manner as an omitted variable)
- Example: unobserved ability in the regression estimating the impact of education on wages

Simultaneity

- Occurs in models where variables are jointly determined

$$\begin{aligned} y_{1i} &= \alpha_0 + \alpha_1 y_{2i} + \varepsilon_{1i} \\ y_{2i} &= \beta_0 + \beta_1 y_{1i} + \varepsilon_{2i} \end{aligned}$$

- Intuitively: change in y_{1i} will cause a change in y_{2i}, which will in turn cause y_{1i} to change again
- Technically:

$$\begin{aligned} Cov(\varepsilon_{1i}, y_{2i}) &= Cov(\varepsilon_{1i}, \beta_0 + \beta_1 y_{1i} + \varepsilon_{2i}) \\ &= \beta_1 Cov(\varepsilon_{1i}, y_{i1}) \\ &= \beta_1 Cov(\varepsilon_{1i}, \alpha_0 + \alpha_1 y_{2i} + \varepsilon_{1i}) \\ &= \beta_1 \left(\alpha_1 Cov(\varepsilon_{1i}, y_{2i}) + Var(\varepsilon_{1i})\right) \\ Cov(\varepsilon_{1i}, y_{2i}) &= \frac{\beta_1}{1 - \alpha_1 \beta_1} Var(\varepsilon_{1i}) \neq 0 \end{aligned}$$

Simultaneity

- Example:

$$\begin{aligned} Q_{Di} &= \alpha_0 + \alpha_1 P_i + \alpha_2 I_i + \varepsilon_{1i} \\ Q_{Si} &= \beta_0 + \beta_1 P_i + \varepsilon_{2i} \\ Q_{Di} &= Q_{Si} \end{aligned}$$

where

Q_D	...	quantity demanded
Q_S	...	quantity supplied
P	...	price
I	...	income

- Endogeneity of price: it is determined from the interaction of supply and demand

Measurement error I

- Measurement error in the dependent variable
- Measurement error is correlated with an explanatory variable

$$y_i^* = y_i + \nu_i \quad \text{where} \quad Cov(\nu_i, x_i) \neq 0$$

- True regression model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Estimated regression: $y_i^* = \beta_0 + \beta_1 x_i + u_i$ where

$u_i = \varepsilon_i + \nu_i$ and so

$$Cov(x_i, u_i) = Cov(x_i, \varepsilon_i + \nu_i) = Cov(\nu_i, x_i) \neq 0$$

- Example: analysis of household consumption patterns (above)

Measurement error II

- Classical measurement error in the explanatory variable

$$x_i^* = x_i + \nu_i \quad \text{where} \quad Cov(\nu_i, x_i) = 0$$

- True regression model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

- Estimated regression: $y_i = \beta_0 + \beta_1 x_i^* + u_i$

  where $u_i = \varepsilon_i - \beta_1 \nu_i$ and so

$$Cov(x_i^*, u_i) = Cov(x_i + \nu_i, \varepsilon_i - \beta_1 \nu_i) = -\beta_1 Var(\nu_i) \neq 0$$

- Causes attenuation bias (estimated coefficient is smaller in absolute value than the true one)

INSTRUMENTAL VARIABLES (IV)

- Answer to the situation when $Cov(x, \varepsilon) \neq 0$

- Instrumental variable (or instrument) should be a variable z such that

  1. z is uncorrelated with the error term: $Cov(z, \varepsilon) = 0$
  2. z is correlated with the explanatory variable x: $Cov(x, z) \neq 0$

- Intuition behind instrumental variables approach:
  - project the endogenous variable x on the instrument z
  - this projection is uncorrelated with the error term and can be used as an explanatory variable instead of x

Instrumental variables

- Suppose the equation we want to estimate is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\eta}$$

  - We can have several instruments for several endogenous variables - we will use the matrix notation **Z** and **X**
  - **X** denotes endogenous variable(s)
  - **Z** denotes instrumental variable(s)
  - Assume that we have at least as many instruments as endogenous variables

Two Stage Least Squares

- 2SLS is a method of implementing instrumental variables approach

- Consists of two steps:

  1. Regress the endogenous variables on the instruments

     $$\mathbf{X} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\nu} \quad ,$$

     get predicted values

     $$\widehat{\mathbf{X}} = \mathbf{Z}\widehat{\boldsymbol{\delta}} = \mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X} \quad ,$$

  2. Use these predicted values instead of $\mathbf{X}$ in the original equation:

     $$\mathbf{y} = \widehat{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\eta}$$

Two Stage Least Squares

- The estimate is

$$\widehat{\boldsymbol{\beta}}^{2SLS} = \left(\widehat{\mathbf{X}}'\widehat{\mathbf{X}}\right)^{-1}\widehat{\mathbf{X}}'\mathbf{y}$$
$$= \left(\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}\left(\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{y}$$

- This estimate is consistent, but it has higher variance than OLS (it is not efficient)
- Intuitively:
  - Only part of the variation in X that is uncorrelated with the error term is used for the estimation.
  - This ensures consistency (X that is uncorrelated with error term).
  - But it makes the estimate less precise (higher variance of β), because not all variation in X is used.

Example

- Estimating the impact of education on the number of children for a sample of women in Botswana
- OLS:

Source	SS	df	MS
Model	12243.0295	3	4081.00985
Residual	9284.14679	4357	2.13085765
Total	21527.1763	4360	4.93742577

Number of obs	=	4361
F(3, 4357)	=	1915.20
Prob > F	=	0.0000
R-squared	=	0.5687
Adj R-squared	=	0.5684
Root MSE	=	1.4597

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.0905755	.0059207	-15.30	0.000	-.102183	-.0789679
age	.3324486	.0165495	20.09	0.000	.3000032	.364894
agesq	-.0026308	.0002726	-9.65	0.000	-.0031652	-.0020964
_cons	-4.138307	.2405942	-17.20	0.000	-4.609994	-3.66662

Example

- Education may be endogenous - both education and number of children may be influenced by some unobserved socioeconomic factors
  - Omitted variable bias: family background is an unobserved factor that influences both the number of children and years of education
- Finding possible instrument:
  - Something that explains education
  - But is not correlated with the family background
- A dummy variable

$$frsthalf = \begin{cases} 1 & \text{if the woman was born in the first} \\ & \text{six months of a year} \\ 0 & \text{otherwise} \end{cases}$$

Example

- Intuition behind the instrument:
- The first condition - instrument explains education:
  - School year in Botswana starts in January
    $\Rightarrow$ Thus, women born in the first half of the year start school when they are at least six and a half.
  - Schooling is compulsory till the age of 15
    $\Rightarrow$ Thus, women born in the first half of the year get less education if they leave school at the age of 15.
- The second condition - instrument is uncorrelated with the error term:
  - Being born in the first half of the year is uncorrelated with the unobserved socioeconomic factors that influence education and number of children (family background etc.)

Example

First-stage regressions

Number of obs	=	4361
F(3, 4357)	=	175.21
Prob > F	=	0.0000
R-squared	=	0.1077
Adj R-squared	=	0.1070
Root MSE	=	3.7110

educ	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	-.1079504	.0420402	-2.57	0.010	-.1903706	-.0255302
agesq	-.0005056	.0006929	-0.73	0.466	-.0018641	.0008529
frsthalf	-.8522854	.1128296	-7.55	0.000	-1.073489	-.6310821
_cons	9.692864	.5980686	16.21	0.000	8.520346	10.86538

EXAMPLE

```
Instrumental variables (2SLS) regression          Number of obs =     4361
                                                  Wald chi2(3)  =  5300.22
                                                  Prob > chi2   =   0.0000
                                                  R-squared     =   0.5502
                                                  Root MSE      =     1.49
```

children	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	-.1714989	.0531553	-3.23	0.001	-.2756813	-.0673165
age	.3236052	.0178514	18.13	0.000	.2886171	.3585934
agesq	-.0026723	.0002796	-9.56	0.000	-.0032202	-.0021244
_cons	-3.387805	.5478988	-6.18	0.000	-4.461667	-2.313943

Instrumented: educ
Instruments: age agesq frsthalf

2SLS

- Note that the endogenous variable has to be instrumented by the instrument and by all other exogenous variables included in the regression

- Think about why:
  - In the first stage, we run $\mathbf{X} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\nu} = \widehat{\mathbf{X}} + \widehat{\boldsymbol{\nu}}$,
  - True model: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \left(\widehat{\mathbf{X}} + \widehat{\boldsymbol{\nu}}\right)\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - Model estimated in the second stage: $\mathbf{y} = \widehat{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\eta}$
  - This implies: $\boldsymbol{\eta} = \widehat{\boldsymbol{\nu}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

- Including all exogenous variables in the first stage make them orthogonal to the residual $\widehat{\boldsymbol{\nu}}$ and hence uncorrelated to the error term $\boldsymbol{\eta}$ in the second stage

Back to the example

- Compare the estimates from OLS and 2SLS:
- OLS:

children	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.0905755	.0059207	-15.30	0.000	-.102183	-.0789679

- 2SLS:

children	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	-.1714989	.0531553	-3.23	0.001	-.2756813	-.0673165

- Is the bias reduced by IV?
- Are these results statistically different?

Summary

- We showed that the estimated coefficients of endogenous variables are inconsistent and biased
- In which situations we may encounter endogenous variables
  - Omitted variable (omitting important variable which is correlated to independent variable)
  - Selection bias (unobserved factors influencing both dependent and independent variable)
  - Simultaneity (causality goes both ways)
  - Measurement error (in either dependent or independent variable)
- We can deal with endogeneity by using instrumental variables (2SLS technique)