Econometrics

Endogenous Regressors and Instrumental Variables

Anna Donina

Lecture 7

Endogeneity Problem

- An ***endogenous*** variable is one that is correlated with u
- An ***exogenous*** variable is one that is uncorrelated with u
- Intuition behind 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
- In IV regression, we focus on the case that X is endogenous and there is an instrument, Z, which is exogenous.

Digression on terminology: "Endogenous" literally means "determined within the system." If X is jointly determined with Y, then a regression of Y on X is subject to simultaneous causality bias. But this definition of endogeneity is too narrow because IV regression can be used to address OV bias and errors-in-variable bias. Thus we use the broader definition of endogeneity above.

Endogeneity Problem

- *Omitted variable bias* from a variable that is correlated with X but is unobserved and for which there are inadequate control variables;
- *Selection bias*: an unobservable characteristic has influence on both dependent and independent variables;
- *Measurement error bias* (X is measured with error)
- *Simultaneous causality bias* (X causes Y, Y causes X);

All three problems cause X to be **endogenous**, $E(u|X) \neq 0$

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: surveying only non-smoking mothers when inferring the impact of the number of prenatal visits on the birth weight of children.
  - Smoking affects both the number of prenatal visits and the birth weight

Simultaneity

- Occurs in models where variables are jointly determined:

$$y_{1i} = \alpha_0 + \alpha_1 y_{2i} + \varepsilon_{1i}$$
$$y_{2i} = \beta_0 + \beta_1 y_{1i} + \varepsilon_{2i}$$

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

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

Endogeneity Problem

- **The endogeneity problem is endemic in social sciences/economics**
  - In many cases important personal variables cannot be observed (examples?)
  - These are often correlated with observed explanatory information
  - In addition, measurement error may also lead to endogeneity
  - Solutions to endogeneity problems:
    - *Proxy variables method for omitted regressors*
    - *Fixed effects methods if: 1) panel data is available, 2) endogeneity is time-constant, and 3) regressors are not time-constant*
- **Instrumental variables method (IV)**
  - IV is the most well-known method to address endogeneity problems

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

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- IV regression breaks X into two parts: a part that might be correlated with u, and a part that is not. By isolating the part that is not correlated with u, it is possible to estimate β_1.

- This is done using an ***instrumental variable***, Z_i, which is correlated with X_i but uncorrelated with u_i.

Instrumental Variables

- **Properties of IV with a poor instrumental variable**
  - IV may be much more inconsistent than OLS if the instrumental variable is not completely exogenous and only weakly related to x

There is no problem if the instrumental variable is really exogenous. If not, the asymptotic bias will be the larger the weaker the correlation with x.

$$plim\ \widehat{\beta}_{1,OLS} = \beta_1 + Corr(x,u) \cdot \frac{\sigma_u}{\sigma_x}$$

$$plim\ \widehat{\beta}_{1,IV} = \beta_1 + \frac{\boxed{Corr(z,u)}}{Corr(z,x)} \cdot \frac{\sigma_u}{\sigma_x}$$

IV worse than OLS if: $\dfrac{Corr(z,u)}{Corr(z,x)} > Corr(x,u)$ e.g. $\dfrac{0.03}{0.2} > 0.1$

- **Variance of IV estimator is always (!) greater than variance of OLS estimator!**

Instrumental Variables

- **IV estimation in the multiple regression model**

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 \boxed{z_1} + \cdots + \beta_k \boxed{z_{k-1}} + u_1$$

endogenous

exogenous variables

- **Conditions for instrumental variable**
  - **1) Does not appear in regression equation**
  - 2) Is uncorrelated with error term
  - 3) Is partially correlated with endogenous explanatory variable

In a regression of the endogenous explanatory variable on all exogenous variables, the instrumental variable must have a non-zero coefficient.

$$y_2 = \pi_0 + \pi_1 z_1 + \cdots + \pi_k z_{k-1} + \pi_k z_k + v_2$$

This is the so called <u>„reduced form regression“</u>

Two Stage Least Squares: 2SLS

As it sounds, TSLS has two stages – two regressions:

1. Isolate the part of X that is uncorrelated with u by regressing X on Z using OLS:

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

- Because Z_i is uncorrelated with u_i, $\pi_0 + \pi_1 Z_i$ is uncorrelated with u_i. We don't know π_0 or π_1 but we have estimated them, so…
- Compute the predicted values of X_i,

2. Replace X_i by $\hat{X}_i$ in the regression of interest:

regress Y on $\hat{X}_i$ using OLS:

$$Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i \qquad (2)$$

Two Stage Least Squares: 2SLS

- **Because $\hat{X}_i$ is uncorrelated with u_i, the first least squares assumption holds for regression (2).** (This requires n to be large so that π_0 and π_1 are precisely estimated.)
- Thus, in large samples, β_1 can be estimated by OLS using regression (2)
- The resulting estimator is called the *Two Stage Least Squares (TSLS)* estimator, $\hat{\beta}_1^{TSLS}$.

Two Stage Least Squares: 2SLS

Suppose Z_i, satisfies the two conditions for a valid instrument:

1. ***Instrument relevance***: $\text{corr}(Z_i, X_i) \neq 0$
2. ***Instrument exogeneity***: $\text{corr}(Z_i, u_i) = 0$

<u>Two-stage least squares:</u>

Stage 1: Regress X_i on Z_i (including an intercept), obtain the predicted values, $\hat{X}_i$

Stage 2: Regress Y_i on $\hat{X}_i$ (including an intercept); the coefficient on $\hat{X}_i$ is the TSLS estimator, $\hat{\beta}_1^{TSLS}$.

$\hat{\beta}_1^{TSLS}$ is a consistent estimator of β_1.

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

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

Example: Intuition behind the IV

- The first condition - instrument explains education:
  - School year in Botswana starts in January
    - 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
    - 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 - the 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 the number of children (family background etc.)

Example: 2SLS

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: 2SLS

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

Example

- Compare the estimates:
- 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

Two Stage Least Squares: 2SLS

- **Why does Two Stage Least Squares work?**
  - All variables in the second stage regression are exogenous because endogenous variable has been replaced by a prediction based on only exogenous information;
  - By using the prediction based on exogenous information, endog. variable is purged of its endogenous part (the part that is related to the error term)
- **Properties of Two Stage Least Squares**
  - The standard errors from the OLS second stage regression are wrong. However, it is not difficult to compute correct standard errors.
  - If there is one endogenous variable and one instrument then 2SLS = IV
  - The 2SLS estimation can also be used if there is more than one endogenous variable and at least as many instruments

Two Stage Least Squares: 2SLS

Statistical properties of 2SLS/IV-estimation

- Under assumptions completely analogous to OLS, but conditioning on $\mathbf{z}_i$ rather than on $\mathbf{x}_i$, 2SLS/IV is consistent and asymptotically normal

- 2SLS/IV is typically much less precise because there is more multicollinearity and less explanatory variation in the second stage regression

- Corrections for heteroscedasticity analogous to OLS

- 2SLS/IV easily extends to time series and panel data situations