Endogeneity

Fall 2008

Definition

- **Endogeneity** is said to occur in a multiple regression model if

$$E\left(X_j u\right) \neq 0, \text{ for some } j = 1, ..., k$$

Definition

- **Endogeneity** is said to occur in a multiple regression model if

$$E\left(X_j u\right) \neq 0, \text{ for some } j = 1, ..., k$$

- Examples:

Definition

- **Endogeneity** is said to occur in a multiple regression model if

$$E\left(X_j u\right) \neq 0, \text{ for some } j = 1, ..., k$$

- Examples:
  - Omitted variables

Definition

- **Endogeneity** is said to occur in a multiple regression model if

$$E(X_j u) \neq 0, \text{ for some } j = 1, ..., k$$

- Examples:
  - Omitted variables
  - Measurement error

Definition

- **Endogeneity** is said to occur in a multiple regression model if

$$E\left(X_j u\right) \neq 0, \text{ for some } j = 1, ..., k$$

- Examples:
  - Omitted variables
  - Measurement error
  - Simultaneity in simultaneous equations models

Omitted Variable and Proxy Variable

- Suppose that a regression model excludes a key variable, due to data unavailability.
- For example, consider a wage equation explicitly recognizing that ability affects wage:

$$\log(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 Abil_i + u_i.$$

- Our primary interest is to measure the effects of education and job experience on wage, holding the effect of ability constant. But ability is usually not available in the data.
- One remedy is to obtain a **proxy variable** that is correlated to the omitted variable.
- In the wage equation, we may want to use the intelligence quotient (IQ) as a proxy for ability:

$$\log(Wage_i) = \beta_0 + \beta_1 Educ_i + \beta_2 Exper_i + \beta_3 IQ_i + v_i.$$

Example: Wage Equation

- The data contains 935 men in 1980 from the Young Men's Cohort of the National Longitudinal Survey (NLSY), USA.

Example: Wage Equation

- The data contains 935 men in 1980 from the Young Men's Cohort of the National Longitudinal Survey (NLSY), USA.
- 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

The estimated return to education is 7.8%.

Example: Wage Equation

- The data contains 935 men in 1980 from the Young Men's Cohort of the National Longitudinal Survey (NLSY), USA.
- 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

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

Measurement Error

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

Measurement Error in an Independent Variable

- Consider a simple regression model:

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

- X_i is measured with errors. That is, we observe $\widetilde{X}_i = X_i + e_i$ instead of X_i.
- We assume that e_i is uncorrelated with X_i, $E(X_i e_i) = 0$.
- Then, the regression equation we use is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_i + u_i - \beta_1 e_i \\ &= \beta_0 + \beta_1 \widetilde{X}_i + v_i. \end{aligned}$$

- It can be seen that the problem of endogeneity occurs:

$$\begin{aligned} E\left(\widetilde{X}_i v_i\right) &= E\left((X_i + e_i)(u_i - \beta_1 e_i)\right) \\ &= -\beta_1 Var(e_i) \neq 0. \end{aligned}$$

Attenuation Bias

- If we perform the regression of Y_i on $\widetilde{X}_i$, then the measurement error leads to a biased OLS estimate towards zero. This is called *attenuation bias*.

Attenuation Bias

- If we perform the regression of Y_i on $\widetilde{X}_i$, then the measurement error leads to a biased OLS estimate towards zero. This is called *attenuation bias*.
- The OLS estimator of β_1 is

$$\begin{aligned}\widehat{\beta}_1 &= \beta_1 + \frac{\sum_{i=1}^{N}\left(\widetilde{X}_i - \overline{\widetilde{X}}\right)\left(u_i - \beta_1 e_i\right)}{\sum_{i=1}^{N}\left(\widetilde{X}_i - \overline{\widetilde{X}}\right)^2} \\ &\longrightarrow {}^p \beta_1 - \beta_1 \frac{Var\left(e\right)}{Var\left(X\right) + Var\left(e\right)}.\end{aligned}$$

Attenuation Bias

- If we perform the regression of Y_i on $\widetilde{X}_i$, then the measurement error leads to a biased OLS estimate towards zero. This is called *attenuation bias*.
- The OLS estimator of β_1 is

$$\widehat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{N}\left(\widetilde{X}_i - \overline{\widetilde{X}}\right)(u_i - \beta_1 e_i)}{\sum_{i=1}^{N}\left(\widetilde{X}_i - \overline{\widetilde{X}}\right)^2}$$

$$\longrightarrow {}^{p}\beta_1 - \beta_1 \frac{Var(e)}{Var(X) + Var(e)}.$$

- Thus, the OLS estimator is inconsistent

$$p\lim\left(\widehat{\beta}_1\right) = \beta_1 \frac{Var(X)}{Var(X) + Var(e)} \leq \beta_1.$$

Simultaneity

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

Simultaneous Equations Model

- Suppose that the equilibrium relation between X and Y is expressed by the following simultaneous equations:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + u_i, \\ X_i &= \alpha_0 + \alpha_1 Y_i + v_i. \end{aligned}$$

- Each one is called a *structural equation* since it has a ceteris paribus, causal interpretation.
- By solving two equations, we have

$$\begin{aligned} Y_i &= \frac{\beta_0 + \beta_1 \alpha_0}{1 - \alpha_1 \beta_1} + \frac{\beta_1 v_i + u_i}{1 - \alpha_1 \beta_1}, \\ X_i &= \frac{\alpha_0 + \alpha_1 \beta_0}{1 - \alpha_1 \beta_1} + \frac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1}. \end{aligned}$$

- These are called the *reduced form* of the model. It is easy to see that, if one perform the regression with just one equation, it will lead to a biased OLS estimator, called **simultaneity bias**.

$$
\begin{aligned}
Cov\left(X_i, u_i\right) &= Cov\left(\frac{v_i + \alpha_1 u_i}{1 - \alpha_1 \beta_1}, u_i\right) \\
&= \frac{\alpha_1}{1 - \alpha_1 \beta_1} Var\left(u_i\right).
\end{aligned}
$$

Endogenous and Exogenous Variables

- Suppose a more general model:

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + u_i \\ X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2 Z_i + v_i \end{cases}$$

- We have two kinds of variables:
  - **Endogenous** variables (X_i and Y_i) are determined within the system.
  - **Exogenous** variables (T_i and Z_i) are exogenously given outside of the model.

- Example: wage and labor supply for married women

$$\begin{cases} \log(Hours_i) = \beta_0 + \beta_1 \log(wage_i) + \beta_2 Educ_i \\ \qquad +\beta_3 Age_i + \beta_4 Kidslt6_i + \beta_5 Nwinc_i + u_i \\ \log(wage_i) = \alpha_0 + \alpha_1 \log(Hours_i) + \alpha_2 Educ_i \\ \qquad +\alpha_3 Exper_i + \alpha_4 Exper_i^2 + v_i \end{cases}$$

Identification I

- The reduced form of the model is

$$\begin{cases} Y_i = \frac{\beta_0+\beta_1\alpha_0}{1-\alpha_1\beta_1} + \frac{\beta_1\alpha_2}{1-\alpha_1\beta_1} Z_i + \frac{\beta_2}{1-\alpha_1\beta_1} T_i + \widetilde{u}_i \\ \quad = B_0 + B_1 Z_i + B_2 T_i + \widetilde{u}_i \\ X_i = \frac{\alpha_0+\alpha_1\beta_0}{1-\alpha_1\beta_1} + \frac{\alpha_2}{1-\alpha_1\beta_1} Z_i + \frac{\alpha_1\beta_2}{1-\alpha_1\beta_1} T_i + \widetilde{v}_i \\ \quad = A_0 + A_1 Z_i + A_2 T_i + \widetilde{v}_i \end{cases}$$

- We can OLS estimate both equations of the reduced form to get consistent estimates of the recuded form parameters: B_0, B_1, B_2, A_0, A_1, and A_2.
- Note that

$$\begin{aligned} \frac{B_1}{A_1} &= \beta_1,\ B_2\left(1 - \frac{B_1 A_2}{A_1 B_2}\right) = \beta_2 \\ \frac{A_2}{B_2} &= \alpha_1,\ A_1\left(1 - \frac{B_1 A_2}{A_1 B_2}\right) = \alpha_2 \end{aligned}$$

Identification II

- Thus, we can back out the estimates of structural parameters from the reduced form coefficients. In this case it is said that the model is *identified*.

Rules for Identification I

- $M\,(K)$ is the number of endogenous (exogenous) variables in the model. $m\,(k)$ is the number of endogenous (exogenous) variables in a given equation.
- **Order Condition** (necessary but not sufficient): In order to have identification in a given model, we must have

$$K - k \geq m - 1$$

- Example1: $M = 2, K = 0$

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + u_i & m = 2, k = 0 \quad \text{not identified} \\ X_i = \alpha_0 + \alpha_1 Y_i + v_i & m = 2, k = 0 \quad \text{not identified} \end{cases}$$

- Example2: $M = 2, K = 1$

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + \beta_2 T_i + u_i & m = 2, k = 1 \quad \text{not identified} \\ X_i = \alpha_0 + \alpha_1 Y_i + v_i & m = 2, k = 0 \quad \text{identified} \end{cases}$$

Estimation of an Identified Equation I

- Consider the following system of equations.

$$\begin{cases} Y_i = \beta_0 + \beta_1 X_i + u_i, \\ X_i = \alpha_0 + \alpha_1 Y_i + \alpha_2 Z_i + v_i. \end{cases}$$

- Note that the first equation is identified. Thus, we are interested in estimating β_1.
- The reduced form is

$$\begin{cases} Y_i = B_0 + B_1 Z_i + \widetilde{u}_i, \\ X_i = A_0 + A_1 Z_i + \widetilde{v}_i, \end{cases}$$

where $B_1 / A_1 = \beta_1$.

Estimation of an Identified Equation II

- The OLS from the reduced form model gives us

$$\widehat{B}_1 = \frac{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2}, \ \widehat{A}_1 = \frac{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)}{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)^2}$$

- Hence, the estimator of β_1 is

$$\widehat{\beta}_{1,IV} = \frac{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N}\left(Z_i - \overline{Z}\right)\left(X_i - \overline{X}\right)} = \frac{\widehat{Cov\left(Z, Y\right)}}{\widehat{Cov\left(Z, X\right)}}.$$

In fact, this is the instrumental variable (IV) estimator, which can be obtained in just one step.

Instrumental Variables (IVs)

- Definition: An instrument for the model, $Y_i = \beta_0 + \beta_1 X_i + u_i$, is a variable Z_i such that

$$Cov(Z, X) \neq 0 \text{ and } Cov(Z, u) = 0.$$

- The IV estimation can be seen as a two step estimator within a simultaneous equations model as seen just before.
- Another way of deriving an IV estimator is from its definition:

$$\begin{aligned} 0 &= Cov(Z, u) = Cov(Z, Y - \beta_0 - \beta_1 X_i) \\ &= Cov(Z, Y) - \beta_1 Cov(Z, X) \end{aligned}$$

And so

$$\widehat{\beta}_{1,IV} = \frac{\widehat{Cov(Z, Y)}}{\widehat{Cov(Z, X)}}.$$

Properties of IV Estimator

- Under the maintained assumptions, the IV estimator is **consistent**:

$$\widehat{\beta}_{1,IV} = \beta_1 + \frac{\sum_{i=1}^{N} \left(Z_i - \overline{Z}\right) u_i}{\sum_{i=1}^{N} \left(Z_i - \overline{Z}\right) \left(X_i - \overline{X}\right)}$$

Since $\sum_{i=1}^{N} \left(Z_i - \overline{Z}\right) u_i / N \longrightarrow^{p} 0$ as $N \longrightarrow \infty$,

$$p\lim \left(\widehat{\beta}_{1,IV}\right) = \beta_1$$

- The IV estimator can have a substantial bias in small samples and thus large samples are preferred.
- The asymptotic variance of the IV estimator is

$$Var\left(\widehat{\beta}_{1,IV}\right) \approx^{p} \sigma_u^2 \frac{Var\left(Z\right)}{N \cdot Cov\left(Z, X\right)^2}$$

Another Example: Lagged dependent variable

- Consider the following time-series model:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 X_t + u_t,$$

where $u_t = v_t + \lambda v_{t-1}$ and v_t is a iid noise and $E\left(u_t | X\right) = 0$.

- It can be easily seen that

$$Cov\left(Y_{t-1}, u_t\right) = \lambda Var\left(v_{t-1}\right) \neq 0.$$

- A valid instrument is X_{t-1} since it is correlated with Y_{t-1} but not with u_t.
- Therefore, the IV estimator is

$$\widehat{\beta}_{1,IV} = \frac{\sum_{t=2}^{T}\left(X_{t-1} - \overline{X}\right)\left(Y_t - \overline{Y}\right)}{\sum_{t=2}^{T}\left(X_{t-1} - \overline{X}\right)\left(X_t - \overline{X}\right)}.$$

More Than One Instrument

- So far we showed how to use one variable as an instrument. Sometimes, we can think more than one variable as an instrument.
- Suppose that Z_1 and Z_2 are two possible instruments for a variable X.

$$\begin{aligned} Cov\left(Z_1, u\right) &= 0 = Cov\left(Z_2, u\right) \\ Cov\left(Z_1, X\right) &\neq 0 \text{ and } Cov\left(Z_2, X\right) \neq 0. \end{aligned}$$

- Rather than using just one instrument, it will be more efficient to use two instruments at the same time. How?

Two-stage Least Squares (2SLS)

- We can use a linear combination of both instruments:

$$Z_i = \alpha_1 Z_{1i} + \alpha_2 Z_{2i},$$

which is still a valid instrument since $Cov(Z, u) = 0$.

- In order to choose α_1 and α_2 so that the correlation between Z_i and X_i is maximal, we perform the OLS from the regression equation:

$$Z_i = \alpha_0 + \alpha_1 Z_{1i} + \alpha_2 Z_{2i} + w_i.$$

- Once we have obtained the fitted value, $\widehat{Z}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 Z_{1i} + \widehat{\alpha}_2 Z_{2i}$, we are back to the case with a single IV:

$$\widehat{\beta}_{1,2SLS} = \frac{\sum_{i=1}^{N} \left(\widehat{Z}_i - \overline{\widehat{Z}}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{N} \left(\widehat{Z}_i - \overline{\widehat{Z}}\right)\left(X_i - \overline{X}\right)}.$$

- This entire procedure is called *two-stage least squares (2SLS) estimation*.

Example: Wage and Labor Supply of Married Woman

- Suppose that the wage and labor supply are determined by

$$\begin{cases} \log(Hours_i) = \beta_0 + \beta_1 \log(wage_i) + \beta_2 Educ_i \\ \qquad +\beta_3 Age_i + \beta_4 Kidslt6_i + \beta_5 Nwinc_i + u_i \\ \log(wage_i) = \alpha_0 + \alpha_1 \log(Hours_i) + \alpha_2 Educ_i \\ \qquad +\alpha_3 Exper_i + \alpha_4 Exper_i^2 + v_i \end{cases}$$

- Is each equation in the model identified?

Example: Wage and Labor Supply of Married Woman

- Using the 2SLS estimation, we have the following results:

Log(hours)	Coeff.	Std. Err.	Log(wage)	Coeff.	Std. Err.
Log(wage)	1.994	0.564	Log(hours)	0.060	0.146
Educ	-0.235	0.071	Educ	0.110	0.016
Age	-0.014	0.011	Exper	0.036	0.018
Kidslt6	-0.465	0.219	(Exper)^2	-0.0007	0.0005
Nwinc	-0.014	0.008	Constant	-0.929	1.003
Constant	8.370	0.689			

Example: Wage and Labor Supply of Married Woman I

- For comparison, we perform the OLS regression for the model:

Log(hours)	Coeff.	Std. Err.		Log(wage)	Coeff.	Std. Err.
Log(wage)	0.043	0.067		Log(hours)	-0.019	0.035
Educ	-0.025	0.022		Educ	0.107	0.014
Age	-0.004	0.006		Exper	0.043	0.014
Kidslt6	-0.621	0.124		(Exper)^2	-0.0008	0.0004
Nwinc	-0.009	0.004		Constant	-0.394	0.310
Constant	7.536	0.373				

- The coefficient on log(wage) is statistically insignificant in OLS, while significant in 2SLS.

Exogeneity Test

- When the independent variables are exogenous, the 2SLS is less efficient than OLS since the 2SLS estimates can have very large standard errors.
- Hauseman's exogeneity test is as follows

$$H_0 : Cov\left(X, u\right) = 0, \;\; H_1 : Cov\left(X, u\right) \neq 0$$

- An idea is to compare both the OLS estimator, $\widehat{\beta}_{1,OLS}$, and the 2SLS estimator, $\widehat{\beta}_{1,2SLS}$. To test whether the differences are statistically significant, it is easier to use the following regression test:
  - First, regress X_i on Z_i and get the residual $\widehat{v}_i$:

$$X_i = \alpha_0 + \alpha_1 Z_i + v_i.$$

  - Regress

$$Y_i = \beta_0 + \beta_1 X_i + \gamma \widehat{v}_i + u_i.$$

  - Test for $\gamma = 0$. If γ is statistically different from zero, then we conclude that X_i is endogeneous.

Example: Wage and Labor Supply of Married Women

- In the first equation (labor supply), we want to test whether log(wage) is endogenous.
- First, we regress the following equation to get the residual $\widehat{v}_i$:

$$\log(wage_i) = \alpha_0 + \alpha_1 Exper_i + \alpha_1 Exper_i^2 + v_i$$

- Then add $\widehat{v}_i$ in the first equation and do OLS:

$$\log(Hours_i) = \beta_0 + \beta_1 \log(wage_i) + \beta_2 Educ_i + \beta_3 Age_i + \beta_4 Kidslt6_i + \beta_5 Nwinc_i + \gamma \widehat{v}_i + u_i$$

- The estimation and t-statistic on $\widehat{v}_i$ are as follows:

	Coeff.	Std. Err.	t-statistic
$\widehat{v}_i$	-1.995	0.322	-6.20