Endogeneity

Taro Mieno

AECN 896-003: Applied Econometrics

Endogeneity

Endogeneity

$E[u|x_k] \neq 0$

Endogeneous independent variable

If u (the error term) is, for whatever reason, correlated with the independent variable x_k, then we say that x_k is an endogenous independent variable.

- Functional form misspecification
- Measurement error

Functional form misspecification

True model

$$\begin{aligned} log(wage) =& \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \beta_4 female \\ &+ \beta_5 female \cdot educ + u \end{aligned}$$

Incorrectly specified model

$$\begin{aligned} log(wage) =& \beta_0 + \beta_1 educ + \beta_2 exper + \beta_4 female \\ &+ \beta_5 female \cdot educ + v \;\; (u + \beta_3 exper^2) \end{aligned}$$

- Often times, adding a quadratic term helps capture non-linear relationship between the dependent and an independent variable
- Whether quadratic or interaction terms should be included can be tested using F-test

Regression Specification Error Test (RESET)

Idea

If the original model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$$

is correct, then no non-linear functions of the independent variables should be significant when added to equation (9.2)

Regression Specification Error Test (RESET)

RESET Steps

1. estimate the original linear model and get $\hat{y}$ (predicted y)
2. estimate the following model:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \sigma_1 \hat{y}^2 + \sigma_2 \hat{y}^3 + u$$

3. test the joint significance of σ_1 and σ_2 (F-test)

Notes

RESET is a test against the general form of non-linearity (Not against a specific functional form like quadratic or log)

Test against nested alternatives

Nested-models

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 \cdot x_2 + u
\end{aligned}
$$

These models are nested because the first model is a special case of the second model when $\beta_3 = 0$ and $\beta_4 = 0$

Testing if the second is appropriate

F-test with the null hypothesis: $\beta_3 = 0$ and $\beta_4 = 0$

- Restricted model: the first model
- Full model: the second model

Test against non-nested alternatives

Non-nested models

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
y =& \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + u
\end{aligned}
$$

These models are non-nested because neither of them is a special case of the other

Testing

F-test cannot be used because one of the model model cannot be a restricted version of the other model

Test against non-nested alternatives

Davidson-MacKinnon test: Idea

Consider the following alternatives:

$$
\begin{aligned}
y =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u \\
y =& \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + u
\end{aligned}
$$

If the first linear model is true, then the fitted values from the other model should be insignificant in the first model.

DM test: the first against the second

1. estimate the second model (non-linear model) to obtain the fitted values, denoted as $\hat{y}$
2. estimate the first model with $\hat{y}$ added:

$$
y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{y} + u
$$

3. conduct a two-sided t-test on $\hat{\theta}_1$

DM test implementation

Non-nested testing

$$H_0: \; log(wage) = \beta_0 + \beta_1 log(educ) + u$$

$$H_1: \; log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ^2 + u$$

```
#--- load the data ---#
wage <- readRDS('wage1.rds')
wage <- data.table(wage)
wage <- wage[educ!=0,]

#--- estimate the null model ---#
null_lm <- lm(wage~log(educ),data=wage)
wage[,y_hat:=null_lm$fitted.value]

#--- estimate the alternative model with y_hat---#
alt_lm <- lm(wage~educ+educ^2+y_hat,data=wage)
summary(alt_lm)$coef
```

```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -1.539045  0.7075019 -2.175323 3.005481e-02
educ         1.337554  0.2090854  6.397164 3.527297e-10
y_hat       -1.595780  0.4204018 -3.795845 1.644405e-04
```

Test against non-nested alternatives

Complications

- Both models could be rejected: there are other functional forms that are more appropriate than the two you tested
- Neither model could be rejected: we could use the adjusted R^2
- Rejection of the second (first) model against the first (second) model does not mean that the first (second) model is correct

Omitted (Unobserved) Variables

Omitted Variable

True Model

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ablility + u$$

Incorrectly Specified Model

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + v(u + \beta_3 ablility)$$

Use of proxy variables to mitigate bias

One way to mitigate the omitted variable bias is to include a proxy variable

Use of proxy variables to mitigate bias

One way to mitigate the omitted variable bias is to include a proxy variable

Proxy Variable

a variable that is related to the unobserved variable that we would like to control

An Example

True Model

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ablility + u$$

ability : not observable

IQ : observed, but does not perfectly capture ability

An Example

True Model

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ablility + u$$

ability : not observable

IQ : observed, but does not perfectly capture ability

Plug-in method

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 IQ + v$$

An Example

True Model

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 ablility + u$$

ability : not observable

IQ : observed, but does not perfectly capture ability

Plug-in method

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 IQ + v$$

Question

When does this plug-in method work (unbiased estimation of the coefficient on education)?

General Framework

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

- x_3^*: unobserved
- x_3: observed

When does the plug-in method work?

True Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

1st Condition (obvious)

Proxy is related to the variable omitted

$$x_3^* = \sigma_0 + \sigma_3 x_3 + v$$

- v is the error term (error exists because they are not the same)
- The parameter σ_3 measures the relationship between x_3^* (ability) and x_3 (IQ).
- If $\sigma_3 = 0$, then x_3 (IQ) is not a good proxy for x_3^* (ability).

When does the plug-in method work?

True Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

Relationship between x_3^* and x_3

$$x_3^* = \sigma_0 + \sigma_3 x_3 + v$$

When does the plug-in method work?

True Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

Relationship between x_3^* and x_3

$$x_3^* = \sigma_0 + \sigma_3 x_3 + v$$

True Model Re-written

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(\sigma_0 + \sigma_3 x_3 + v) + u \\ &= (\beta_0 + \beta_3 \sigma_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \sigma_3 x_3 + (\beta_3 v + u) \\ &= \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + \varepsilon \end{aligned}$$

- $\alpha_0 = \beta_0 + \beta_3 \sigma_0$
- $\alpha_3 = \beta_3 \sigma_3$
- $\varepsilon = \beta_3(\sigma_0 + v) + u$

When does the plug-in method work?

True Model Re-written

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + \varepsilon$$

- $\alpha_0 = \beta_0 + \beta_3 \sigma_0$
- $\alpha_3 = \beta_3 \sigma_3$
- $\varepsilon = \beta_3 v + u$

When does the plug-in method work?

True Model Re-written

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + \varepsilon$$

- $\alpha_0 = \beta_0 + \beta_3 \sigma_0$
- $\alpha_3 = \beta_3 \sigma_3$
- $\varepsilon = \beta_3 v + u$

Question

When you regress y on x_1, x_2, and x_3, what are the conditions for unbiased estimation of the coefficients on the independent variables?

When does the plug-in method work?

True Model Re-written

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + \varepsilon$$

- $\alpha_0 = \beta_0 + \beta_3 \sigma_0$
- $\alpha_3 = \beta_3 \sigma_3$
- $\varepsilon = \beta_3 v + u$

Question

When you regress y on x_1, x_2, and x_3, what are the conditions for unbiased estimation of the coefficients on the independent variables?

Answer

$E[\varepsilon | x_1, x_2, x_3] = 0$

Investigation the condition

Conditions for unbiasedness

$$
\begin{aligned}
&E[\varepsilon|x_1, x_2, x_3] = 0 \\
\Rightarrow &E[\beta_3 v + u|x_1, x_2, x_3] = 0
\end{aligned}
$$

Breaking into two conditions,

$$
\begin{aligned}
&E[v|x_1, x_2, x_3] = 0 \\
&E[u|x_1, x_2, x_3] = 0
\end{aligned}
$$

When does the plug-in method work?

True Model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

2nd Condition: $E[u|x_1, x_2, x_3] = 0$

x_1, x_2, and x_3 are not correlated with u

- $E[u|x_1, x_2] = 0$: (standard condition for unbiasedness)
- $E[u|x_3] = 0$: x_3 does not belong in the true model after you control for x_1, x_2, and x_3^*.
  - This is essentially true by definition, since x_3 (IQ) is a proxy variable for x_3^* (ability): it is x_3^* (ability) that directly affects y (log(wage)), not x_3 (IQ)

When does the plug-in method work?

Relationship between x_3^* and x_3

$$x_3^* = \sigma_0 + \sigma_3 x_3 + v$$

3rd Condition: $E[v|x_1, x_2, x_3] = 0$

x_1, x_2, and x_3 are not correlated with v

- $E[v|x_1, x_2] = 0$: Once x_3 (IQ) is partialled out, x_3^* (ability) is uncorrelated with x_1 (education) and x_2 (experience)
- $E[v|x_3] = 0$: always satisfied by construction

When does the plug-in method work?

3rd Condition

$E[v|x_1, x_2] = 0$: Once x_3 (IQ) is partialled out, x_3^* (ability) is uncorrelated with x_1 (education) and x_2 (experience)

Put it differently,

Does something in ability other than IQ determine education or experience?

-

When does the plug-in method work?

3rd Condition

$E[v|x_1, x_2] = 0$: Once x_3 (IQ) is partialled out, x_3^* (ability) is uncorrelated with x_1 (education) and x_2 (experience)

Put it differently,

Does something in ability other than IQ determine education or experience?

- Probably yes. But, since the IQ part is taken out of ability in the error term, $educ$ and $exper$ should be less correlated with the error term now!

Table

	Dependent variable:	
	log(wage)	
	(1)	(2)
educ	0.065^{***} (0.006)	0.054^{***} (0.007)
exper	0.014^{***} (0.003)	0.014^{***} (0.003)
tenure	0.012^{***} (0.002)	0.011^{***} (0.002)
married	0.199^{***} (0.039)	0.200^{***} (0.039)
south	-0.091^{***} (0.026)	-0.080^{***} (0.026)
black	-0.188^{***} (0.038)	-0.143^{***} (0.039)
urban	0.184^{***} (0.027)	0.182^{***} (0.027)
IQ		0.004^{***} (0.001)
Constant	5.395^{***} (0.113)	5.176^{***} (0.128)

Measurement Errors

Measurement Errors (ME)

Inaccuracy in the values observed as opposed to the actual values

Examples

- reporting errors (any kind of survey has the potential of mis-reporting)
  - household survey on income and savings
- the use of estimated values
  - spatially interpolated weather conditions (precipitation)
  - imputed irrigation costs

Question

What are the consequences of having measurement errors in variables you use in regression?

ME in the dependent variable

True model

Consider the following general model

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{1}$$

with MLR.1 through MLR.6 satisfied.

Measurement errors

The difference between the observed and actual values

$$e = y - y^* \tag{2}$$

ME in the dependent variable

True model

Consider the following general model

$$y^* = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \tag{1}$$

with MLR.1 through MLR.6 satisfied.

Measurement errors

The difference between the observed and actual values

$$e = y - y^* \tag{2}$$

Re-write the true model

Plugging (2) into (3),

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v(u + e)$$

ME in the dependent variable

Re-write the true model

Plugging (2) into (3),

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v, \text{ where } v = (u + e) \tag{3}$$

Question

What are the conditions under which OLS estimators are unbiased?

ME in the dependent variable

Re-write the true model

Plugging (2) into (3),

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + v, \text{ where } v = (u + e) \tag{3}$$

Question

What are the conditions under which OLS estimators are unbiased?

Answer

$$E[v|x_1, \ldots, x_k] = 0$$

So, as long as the measurement error is uncorrelated with the independent variables, OLS estimators are still unbiased.

ME in an independent variable

True Model

Consider the following general model

$$y = \beta_0 + \beta_1 x_1^* + u \tag{4}$$

with MLR.1 through MLR.6 satisfied.

Measurement errors

The difference between the observed and actual values

$$e_1 = x_1 - x_1^* \tag{5}$$

ME in an independent variable

True Model

Consider the following general model

$$y = \beta_0 + \beta_1 x_1^* + u \tag{4}$$

with MLR.1 through MLR.6 satisfied.

Measurement errors

The difference between the observed and actual values

$$e_1 = x_1 - x_1^* \tag{5}$$

Re-write the true model

Plugging (5) into (4),

$$y = \beta_0 + \beta_1 x_1 + v, \;\; \text{where} \;\; v = (u - \beta e_1) \tag{6}$$

ME in an independent variable

Re-write the true model

Plugging (5) into (4),

$$y = \beta_0 + \beta_1 x_1 + v, \quad \text{where} \quad v = (u - \beta e_1) \tag{7}$$

Question

What are the conditions under which OLS estimators are unbiased?

ME in an independent variable

Re-write the true model

Plugging (5) into (4),

$$y = \beta_0 + \beta_1 x_1 + v, \text{ where } v = (u - \beta e_1) \tag{7}$$

Question

What are the conditions under which OLS estimators are unbiased?

Answer

$$E[v|x_1] = 0$$

ME in an independent variable

Classical errors-in-variables (CEV)

The correctly observed variable (x_1^*) is uncorrelated with the measurement error (e_1):

$$Cov(x_1^*, e_1) = 0$$

Under CEV

The incorrectly observed variable (x_1) must be correlated with the measurement error (e_1):

$$\begin{aligned}
Cov(x_1, e_1) &= E[x_1 e_1] - E[x_1]E[e_1] \\
&= E[(x_1^* + e_1)e_1] - E[x_1^* + e_1)]E[e_1] \\
&= E[x_1^* e_1 + e_1^2] - E[x_1^* + e_1)]E[e_1] \\
&= \sigma_{e_1}^2 = \sigma_{e_1}^2
\end{aligned}$$

ME in an independent variable

Classical errors-in-variables (CEV)

The correctly observed variable (x_1^*) is uncorrelated with the measurement error (e_1):

$$Cov(x_1^*, e_1) = 0$$

Under CEV

The incorrectly observed variable (x_1) must be correlated with the measurement error (e_1):

$$\begin{aligned} Cov(x_1, e_1) &= E[x_1 e_1] - E[x_1]E[e_1] \\ &= E[(x_1^* + e_1)e_1] - E[x_1^* + e_1)]E[e_1] \\ &= E[x_1^* e_1 + e_1^2] - E[x_1^* + e_1)]E[e_1] \\ &= \sigma_{e_1}^2 = \sigma_{e_1}^2 \end{aligned}$$

So, the mis-measured variable x_1 is always endogenous

ME in an independent variable

The direction of the bias

In general, for the following model:

$$y = \beta_0 + \beta_1 x_1 + u$$

$$sign(bias) = sign(Cov(x, u))$$

ME in an independent variable

The direction of the bias

In general, for the following model:

$$y = \beta_0 + \beta_1 x_1 + u$$

$$sign(bias) = sign(Cov(x, u))$$

ME in an independent variable

$$y = \beta_0 + \beta_1 x_1 + v, \text{ where } v = (u - \beta e_1)$$

$$\begin{aligned} sign\Big(Cov(x_1, v)\Big) &= sign\Big(Cov(x_1, u - \beta e_1)\Big) \\ &= sign\Big(-\beta Cov(x_1, e_1)\Big) \\ &= -sign(\beta) sign\Big(Cov(x_1, e_1)\Big) \\ &= -sign(\beta) \end{aligned}$$

ME in an independent variable

ME in an independent variable

$$y = \beta_0 + \beta_1 x_1 + v, \text{ where } v = (u - \beta e_1)$$

$$\begin{aligned} sign\Big(Cov(x_1, v)\Big) &= sign\Big(Cov(x_1, u - \beta e_1)\Big) \\ &= sign\Big(-\beta Cov(x_1, e_1)\Big) \\ &= -sign(\beta) sign\Big(Cov(x_1, e_1)\Big) \\ &= -sign(\beta) \end{aligned}$$

Attenuation Bias

Under CEV, the direction of bias is always the negative of the sign of the coefficient, which leads to a coefficient estimate closer to 0 than it truly is.