2) Measurement Error

Vitor Kamada

July 2018

Measurement Error in the Dependent Variable

$$y^* = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + v$$

$$e_0 = y - y^*$$

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + v + e_0$$

$$Var(v + e_0) = \sigma_v^2 + \sigma_0^2 > \sigma_v^2$$

Measurement Error in Firm Scrap Rates

$$log(scrap^*) = \beta_0 + \beta_1 grant + v$$

$$log(scrap) = log(scrap^*) + e_0$$

$$log(scrap) = \beta_0 + \beta_1 grant + v + e_0$$

If a firm receiving a grant is more likely to underreport its scrap rate:

$$Cov(v + e_0, grant) < 0$$

Measurement Error in an Explanatory Variable

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k^* + v$$

$$e_k = x_k - x_k^*$$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (v - \beta_k e_k)$$

$$Cov(x_k, e_k) = 0$$

$$Var(v - \beta_k e_k) = \sigma_v^2 + \beta_k^2 \sigma_{e_k}^2$$

Classical Errors-in-Variables (CEV)

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + (v - \beta_k e_k)$$

$$Cov(x_k^*, e_k) = 0$$

$$Cov(x_k, e_k)$$

$$E(x_k e_k)$$

$$E(x_k^* e_k) + E(e_k^2) = \sigma_{e_k}^2$$

CEV - Special Case

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

$$plim \hat{\beta}_1 = \beta_1 + \frac{Cov(x_1, u - \beta_1 e_1)}{Var(x_1)}$$

$$= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}}$$

$$= \beta_1 (1 - \frac{\sigma^2_{e_1}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}})$$

$$= \beta_1 (\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_{e_1}})$$

CEV - General Case

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + (v - \beta_k e_k)$$

$$x_k^* = \delta_0 + \delta_1 x_1 + \ldots + \delta_{k-1} x_{k-1} + r_k^*$$

$$Corr(x_k^*, x_j) = 0 \text{ for } j \neq k$$

$$plim \hat{\beta}_k = \beta_k \left(\frac{\sigma^2_{r_k^*}}{\sigma^2_{r_k^*} + \sigma^2_{e_k}} \right)$$

Measurement Error in Family Income

$$ColGPA = \beta_0 + \beta_1 faminc^* + \beta_2 hsGPA + \beta_3 SAT + v$$

$$faminc = faminc^* + e_1$$

$$H_0 : \beta_1 = 0$$

Type II error

$$smoked = smoked^* + e_1$$

People who do not smoke marijuana: $smoked^* = 0$ and $smoked = 0$

When $smoked^* > 0$, it more likely that someone miscounts

$$Corr(smoked^*, e_1) \neq 0$$

Simulation

$$y = \beta x^* + u$$

$$x^* \sim N(0,9); \qquad u \sim N(0,1)$$

$$x = x^* + v; \qquad v \sim N(0,1)$$

$$s = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_{x^*}^2} = \frac{1}{1+9} = 0.1$$

$$plim \hat{\beta} = \beta - s\beta$$

$$1 - 0.1 \times 1 = 0.9$$

quietly set obs 10000

set seed 10101

matrix mu = (0,0,0)

matrix sigmasq = (9,0,0\0,1,0\0,0,1)

drawnorm xstar u v, means(mu) cov(sigmasq)

generate y = 1*xstar + u

generate x = xstar + v

regress y x, noconstant

y	Coef.	Std. Err.	t
x	.899366	.0043202	208.17

Errors-in-Variables Regression (EIV)

$$\hat{\beta}_{EIV} = (Q'Q - C)^{-1}Q'Y$$

eivreg y x, r(x .9)

y	Coef.	Std. Err.	t	P>|t|
x	.9992973	.0034565	289.11	0.000
_cons	.0027368	.0099535	0.27	0.783

$$Var(\hat{\beta}_{EIV}) = Var(E[\hat{\beta}_{EIV}|D]) + E[Var(\hat{\beta}_{EIV}|D)]$$

Lockwood et al. (2017) : eivreg ignores the first term

Structural Equation Model

sem (x<-X) (y<-X), reliability(x .9)

	OIM Coef.	Std. Err.	z	P>|z|
Measurement				
x				
X	1	(constrained)		
_cons	-.0082131	.031996	-0.26	0.797
y				
X	.9992973	.0050504	197.87	0.000
_cons	-.0054706	.0319238	-0.17	0.864
var(e.x)	1.023744	(constrained)		
var(e.y)	.9905133	.0284656		
var(X)	9.213694	.1447792		