EEP/IAS 118 - Introductory Applied Econometrics, Section 9A

Pierre Biscaye and Jed Silver

November 2021

Measurement Error

- Measurement error in y_i
- Measurement error in x_i (uncorrelated with measurement)
- Measurement error in x_i (uncorrelated with the truth)
- Proxy variables

Measurement error in y_i

Suppose the true model is

$$y_i^* = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_k + u_i$$

and SLR 1-4 hold, but we measure y_i with error: $y_i = y_i^* + e_i$. What else do we need to assume about e_i to obtain unbiased estimates of β_j? Why?

Measurement error in y_i

Suppose the true model is

$$y_i^* = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_k + u_i$$

and SLR 1-4 hold, but we measure y_i with error: $y_i = y_i^* + e_i$. What else do we need to assume about e_i to obtain unbiased estimates of β_j? Why?

$E[e_i|x_1 \ldots x_k]$=0

Measurement error in y_i

Suppose the true model is

$$y_i^* = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_k + u_i$$

and SLR 1-4 hold, but we measure y_i with error: $y_i = y_i^* + e_i$. What else do we need to assume about e_i to obtain unbiased estimates of β_j? Why?

$E[e_i|x_1 \ldots x_k]$=0

If so (as in the case of classical measurement error in y)

$$E[y_i|x_1 \ldots x_k] = E[y_i^* + e_i|x_1 \ldots x_k] = E[y_i^*|x_1 \ldots x_k] + \underbrace{E[e_i|x_1 \ldots x_k]}_{=0}$$

$$= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Measurement error in y_i

Suppose the true model is

$$y_i^* = \beta_0 + \beta_1 x_i + \ldots + \beta_k x_k + u_i$$

and SLR 1-4 hold, but we measure y_i with error: $y_i = y_i^* + e_i$. What else do we need to assume about e_i to obtain unbiased estimates of β_j? Why?

$E[e_i|x_1 \ldots x_k]$=0

If so (as in the case of classical measurement error in y)

$$E[y_i|x_1 \ldots x_k] = E[y_i^* + e_i|x_1 \ldots x_k] = E[y_i^*|x_1 \ldots x_k] + \underbrace{E[e_i|x_1 \ldots x_k]}_{=0}$$

$$= \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

But our SEs will be larger because $\sigma_{u+e} > \sigma_u$

Measurement error in x_i

Suppose the true model is (with SLR 1-4 holding)

$$y_i = \beta_0 + \beta_1 x_i^* + u_i$$

but we observe $x_i = x_i^* + e_i$

When we estimate

$$y_i = \beta_0 + \beta_1 x_i + u_i - \beta_1 e_i = \beta_0 + \beta_1 x_i + \nu_i$$

when do we recover an unbiased estimate of β_1?

Two special cases: Case 1

Case 1: $\text{cov}(e_i, x_i)=0$

- Here $E[v_i|x_i] = E[u_i - \beta_1 e_i|x_i] = E[u_i|x_i] - \beta_1 E[e_i|x_i] = 0$
- By the SLR4 for the "true" model the first term is 0
- Since $\text{cov}(e_i, x_i)=0$ the second term is 0
  $\implies$ We get the equivalent of SLR 4 for our estimated model
- Again, our SEs will be larger because $\sigma_v > \sigma_u$

Case 2: Classical measurement error

Case 2: $\text{cov}(e_i, x_i^*) = 0$

- By construction, $\text{cov}(e_i, x_i) > 0$ (Why?)
- We therefore have an omitted variable in our error term
- Recall formula for sign of OVB: $E[\hat{\beta}_1] = \beta_1 + \beta_2\delta_1$
  - Here $\beta_2 = -\beta_1$
  - $0 < \delta_1 = \frac{cov(e_i, x_i)}{var(x_i)} < 1$
- $\implies$ Bias takes the opposite sign of β_1 (attenuates it towards 0)

Other cases

- If neither $\text{cov}(e_i, x_i) = 0$ nor $\text{cov}(e_i, x_i^*) = 0$ (most likely cases) can't say what will happen
- Same goes with complex correlations between e_i and multiple x_i's
- Even when we get unbiased estimates, standard errors increase

Proxy variables

- We can't always measure everything we think belongs in a model (e.g., ability), but in some cases we can observe something that comes reasonably close (e.g., IQ test or SAT score)
- Can include these in model as **proxies** for the variable we want
- What does including a proxy imply for our estimation? Example

Proxy variable example

Suppose the true population model is:

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 ability_i + e_i$$

But we measure ability using IQ as a proxy

$$ability_i = \delta_0 + \delta_1 IQ_i + v_i$$

We end up estimating

$$\log(wage_i) = \underbrace{\beta_0 + \beta_3\delta_0}_{\text{constant}} + \beta_1 educ_i + \beta_2 exper_i + \beta_3\delta_1 IQ_i + \underbrace{\beta_3 v_i + e_i}_{\text{new error term}}$$

Proxy variable and estimation

$$\log(wage_i) = (\beta_0 + \beta_3\delta_0) + \beta_1 educ_i + \beta_2 exper_i + \beta_3\delta_1 IQ_i + (\beta_3 v_i + e_i)$$

- To obtain unbiased estimates, MLR4 becomes $\mathbb{E}[\beta_3 v_i + e_i | educ, exper, IQ] = 0$
- Key new piece with proxy variable is $\mathbb{E}[v_i | educ_i, exper_i] = 0$
  - Part of ability that is not correlated with IQ is also not correlated with education and experience
- If true, get unbiased estimates and coefficient on IQ is effect of ability on wages scaled by relationship between ability and IQ
- If false, still have an omitted variable - the part of ability not explained by IQ

Exploring measurement error with data

Click this to go to Jupyter!