Measurement error

Otto Toivanen

2017

Measurement error in cross-section OLS

- Simplest case: one variable cross-section regression

$$y_i = \alpha + \beta x_i^* + \varepsilon_i$$

- where we observe x_i^* with measurement error

$$x_i = x_i^* + r_i$$

- such that $Cov(x_i^*, r_i) = 0$ and $\mathrm{Var}(r) = \sigma_r^2$ (classical errors-in-variables, CEV).
- The estimated model becomes

$$y_i = \alpha + \beta x_i + \eta_i$$
$$\eta_i = -\beta r_i + \varepsilon_i$$

Measurement error in cross-section OLS

- The estimated $\hat{\beta}_{OLS}$ becomes:

$$p\lim_{N\to\infty}\hat{\beta}_{OLS} = \frac{\text{Cov}(y_i, x_i)}{\text{Var}(x_i)} = \frac{\text{Cov}(\alpha + \beta x_i - \beta r_i + \varepsilon_i, x_i)}{\text{Var}(x_i)}$$

- Notice that the first part in the covariance has four items: α, βx_i, $-\beta r_i$, and ε.

Measurement error in cross-section OLS

- Let's go through the relation between each of these 4 parts and the second part in the covariance, x_i:

1. α: this is a constant, so its covariance with anything is zero.
2. βx_i: This consists of a constant β times the same variable with which we measure the covariance. Thus $\text{Cov}(\beta x_i, x_i) = \beta \text{Cov}(x_i, x_i) = \beta \text{Var}(x_i)$.
3. $-\beta r_i$: Again there is a constant, $-\beta$. To understand the covariance between r_i and x_i, recall that $x_i = x_i^* + r_i$ and that x_i^* is uncorrelated with r_i. Thus we get $\text{Cov}(-\beta r_i, x_i) = \text{Cov}(-\beta r_i, x_i^* + r_i) = \text{Cov}(-\beta r_i, r_i) = -\beta \text{Var}(r_i) = -\beta \sigma_r^2$.
4. ε_i: This is uncorrelated with x_i^* and r_i and therefore also with x_i. Thus the covariance between ε_i and x_i is zero.

Measurement error in cross-section OLS

- The estimated $\hat{\beta}_{OLS}$ becomes:

$$p\lim_{N\to\infty}\hat{\beta}_{OLS} = \frac{\text{Cov}(\alpha+\beta x_i-\beta r_i+\varepsilon_i, x_i)}{\text{Var}(x_i)}$$

$$= \frac{0+\beta\text{Var}(x_i)-\beta\text{Var}(r_i)+0}{\text{Var}(x_i)} = \beta-\beta\frac{\text{Var}(r_i)}{\text{Var}(x_i)}$$

- This can be rewritten to yield

$$= \beta-\beta\frac{\sigma_r^2}{\sigma_x^2} = \beta-\beta\frac{\sigma_r^2}{\sigma_{x^*}^2+\sigma_r^2} = \beta\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2+\sigma_r^2}$$

- where $\sigma_x^2 = \sigma_{x^*}^2+\sigma_r^2$ and
- $\sigma_r^2/\sigma_x^2 =$ noise-to-signal ratio.
- $\sigma_{x^*}^2/(\sigma_{x^*}^2+\sigma_r^2) =$reliability ratio.
- $-\beta\sigma_r^2/(\sigma_{x^*}^2+\sigma_r^2) =$ attenuation bias.