ECON 4151
Lab Session: Measurement Error

Wonjin Lee


Washington University in St. Louis

October 9, 2020

Classical measurement error

- A well-specified causal regression:

$$y_i = \alpha + \beta x_i + \epsilon_i$$

- Observe x imperfectly: $\tilde{x} = x + \eta$ with $\eta \sim N(0, \sigma_\eta^2)$.
  - Assume that the measurement error is independent of everything else in our model.
- Running a regression on $\tilde{x}$ instead of x will generally lead to a coefficient that is biased toward zero (attenuation bias).

$$\begin{aligned}
\tilde{\beta} &= \frac{\text{Cov}(\tilde{x}, y)}{\text{Var}(\tilde{x})} \\
&= \frac{\text{Cov}(x + \eta, \alpha + \beta x + \epsilon)}{\text{Var}(x) + \sigma_\eta^2} \\
&= \beta \frac{\text{Var}(x)}{\text{Var}(x) + \sigma_\eta^2}
\end{aligned}$$

- As $\sigma_\eta^2 \to \infty$, the population parameter we would try to estimate would be zero.

Classical measurement error

- Now the true DGP is

$$y_i = \beta x_i + \underbrace{\alpha_i + \epsilon_i}_{} \quad \text{and} \quad \text{Cov}(x, \alpha) \neq 0$$

- Two problems: measurement error and an omitted variable bias.

$$\tilde{\beta} = \frac{\beta \, \text{Var}(x) + \text{Cov}(x, \alpha)}{\text{Var}(x) + \sigma_\eta^2}$$

- Use the within estimator (FE model) to eliminate the omitted variable bias.

$$\begin{aligned}
\tilde{\beta}^{FE} &= \frac{\text{Cov}((\tilde{x} - \bar{\tilde{x}}), (y - \bar{y}))}{\text{Var}(\tilde{x} - \bar{x})} \\
&= \frac{\text{Cov}((\tilde{x} - \bar{\tilde{x}}), (\beta(x - \bar{x}) + (\epsilon - \bar{\epsilon})))}{\text{Var}(\tilde{x} - \bar{\tilde{x}})} \\
&= \beta \frac{\text{Var}(x - \bar{x})}{\text{Var}(x - \bar{x}) + \text{Var}(\eta - \bar{\eta})}
\end{aligned}$$

Classical measurement error

- The within estimator still has problems:
  1. The share of the denominator due to measurement error is likely to increase (attenuation bias).
  2. Switching to a "within" estimator will frequently increase your standard errors.

Exercise

Suppose you are a nutritionist analyzing the relationship between BMI (body mass index) and calorie consumption. The formula for computing BMI is weight (in kg) divided by height squared (height is measured in meters). The true mode is given by

$$y_i = b_0 + b_1 x_i + u_i$$

where x_i is calorie consumption. You are concerned, however, that weight is self-reported and people who are overweight report lower values while others report accurately. Therefore, instead of weight, you observe BMI with a noise

$$y_i^* = y_i + v_i.$$

Exercise

1 Analyze the implications of this type of reporting error on the consistency of, $\hat{\beta}_1$, the OLS estimator for b_1. To do that write down the formula for the OLS estimator, and write down the probability limit of the estimator.

ANS We can rewrite the equation as $y^* = b_0 + b_1 x + u + v$. Then,

$$\begin{aligned}\hat{\beta}_1 &= \frac{\text{Cov}\left(y^*, x\right)}{\text{Var}(x)} \\ &= \frac{\text{Cov}\left(b_0 + b_1 x + u + v, x\right)}{\text{Var}(x)} \\ &= b_1 + \frac{\text{Cov}(x, u) + \text{Cov}(x, v)}{\text{Var}(x)}.\end{aligned}$$

As long as $\text{Cov}(x, u) = 0$ and $\text{Cov}(x, v) = 0$, $\hat{\beta}_1$ is unbiased. From the true model, $\text{Cov}(x, u) = 0$, but it is likely that $\text{Cov}(x, v) \neq 0$ due to the misreport.

Exercise

ANS For consistency, note that

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(y_i^*-\bar{y}^*\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \\ &= b_1+\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(v_i-\bar{v}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}+\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\left(u_i-\bar{u}\right)}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}.\end{aligned}$$

Thus,

$$p\lim_{n\to\infty}\hat{\beta}_1 = b_1+\frac{\mathrm{Cov}(x,u)+\mathrm{Cov}(x,v)}{\mathrm{Var}(x)}.$$

As long as $\mathrm{Cov}(x,u)=0$ and $\mathrm{Cov}(x,v)=0$, $\hat{\beta}_1$ is consistent.

Exercise

2 Suppose someone proposed to you to use a variable that is correlated with the true BMI, but is uncorrelated with both the reporting error, v, and the error term, u. How would you make use of this information? Discuss how would you use it, and relate your discussion to estimating b_1 consistently.

ANS Let z denote the new variable. Note that $y^* = b_0 + b_1 x + u + v$ has an endogeneity issue because $\text{Cov}(x, u+v) \neq 0$. WTS: z is an IV for x. Since since z is correlated with y, $\text{Cov}(z, y) \neq 0$.

$$\begin{aligned}
\text{Cov}\left(z_i, y_i\right) &= \text{Cov}\left(z_i, b_0 + b_1 x_i + u_i\right) \\
&= b_1 \,\text{Cov}\left(z_i, x_i\right) + \underbrace{\text{Cov}\left(z_i, u_i\right)}_{=0} \\
&= b_1 \,\text{Cov}\left(z_i, x_i\right) \neq 0.
\end{aligned}$$

Thus, $\text{Cov}\left(z_i, x_i\right) \neq 0$ (relevance condition). Also z is uncorrelated with u and v, $\text{Cov}(z, u+v) = 0$ (exclusion condition). Therefore, z is an IV for x so do IV estimation!