Measurement Error in X
ECON 30331

Bill Evans
Fall 2008

Model: (1) $y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$

We are going to simplify the model and assume that β_0=0. Therefore, we can write the model as

(2) $y_i = x_i\beta_1 + \varepsilon_i$

You can easily demonstrate for yourself that the estimate for $\hat{\beta}_1$ will be

$$(3) \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} x_i^2}$$

We demonstrated this on problem set 3.

Now, assume that x is measured with some error. Let x_i represent the true value of x and let x_i^* be the measured value of x where $x_i^* = x_i + v_i$. The variable v_i is a random error with E[v_i]=0, V[v_i]=σ_v^2 and v_i is uncorrelated with both x and ε, so $\text{cov}(v_i, x_i) = \text{cov}(v_i, \varepsilon_i) = 0$.

If we use x_i^* in the regression instead of x_i the OLS estimate for $\hat{\beta}_1$ will now be

$$(4) \qquad \hat{\beta}_1^* = \frac{\sum_{i=1}^{n} y_i x_i^*}{\sum_{i=1}^{n} \left(x_i^*\right)^2}$$

To find the true underlying properties of the estimate, we must substitute two values in equation (4). First, in the numerator, we must substitute in the true value for y, given by equation (2). Next, we must substitute the true value for x_i^* given by $x_i^* = x_i + v_i$

$$(5) \qquad \hat{\beta}_1^* = \frac{\sum_{i=1}^{n} (\beta_1 x_i + \varepsilon_1)(x_i + v_i)}{\sum_{i=1}^{n} (x_i + v_i)^2}$$

Complete the squares in the numerator and in the denominator

(6) $$\hat{\beta}_1^* = \frac{\sum_{i=1}^{n}(\beta_1 x_i^2 + \beta_1 x_i v_i + \varepsilon_i x_i + \varepsilon_i v_i)}{\sum_{i=1}^{n}(x_i^2 + 2x_i v_i + v_i^2)}$$

And breaking apart the terms in the numerator and denominator

(7) $$\hat{\beta}_1^* = \frac{\beta_1\sum_{i=1}^{n} x_i^2 + \beta_1\sum_{i=1}^{n} x_i v_i + \sum_{i=1}^{n}\varepsilon_i x_i + \sum_{i=1}^{n}\varepsilon_i v_i}{\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i v_i + \sum_{i=1}^{n} v_i^2}$$

Note that we can divide each term in the numerator and denominator by (n-1).

(7) $$\hat{\beta}_1^* = \frac{\left[\beta_1\sum_{i=1}^{n} x_i^2 + \beta_1\sum_{i=1}^{n} x_i v_i + \sum_{i=1}^{n}\varepsilon_i x_i + \sum_{i=1}^{n}\varepsilon_i v_i\right]/(n-1)}{\left[\sum_{i=1}^{n} x_i^2 + 2\sum_{i=1}^{n} x_i v_i + \sum_{i=1}^{n} v_i^2\right]/(n-1)}$$

Now let's make some substitutions. Recall that the definition of the sample variance of x is

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

But recall also that we assumed that $\overline{x} = 0$ therefore

$$\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} x_i^2$$

Likewise, recall that the sample covariance between x and ε is by definition

$$\hat{\sigma}_{x\varepsilon} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})(\varepsilon_i - \overline{\varepsilon})$$

Which using the properties of summations we can write as

$$\hat{\sigma}_{x\varepsilon} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\varepsilon_i$$

But recall also that we assumed that $\overline{x} = 0$ therefore

$$\hat{\sigma}_{x\varepsilon} = \frac{1}{n-1}\sum_{i=1}^{n} x_i \varepsilon_i$$

Using these same arguments, it is easy to show that

$$\frac{1}{n-1}\sum_{i=1}^{n} v_i^{\,2} = \hat{\sigma}_v^2 \;\text{ and }\; \frac{1}{n-1}\sum_{i=1}^{n} x_i v_i = \hat{\sigma}_{xv} \;\text{ and }\; \frac{1}{n-1}\sum_{i=1}^{n} \varepsilon_i v_i = \hat{\sigma}_{\varepsilon x}$$

Substituting all these values into equation (7), we obtain

$$(8) \qquad \hat{\beta}_1^* = \frac{\beta_1 \hat{\sigma}_x^2 + \beta_1 \hat{\sigma}_{vx} + \hat{\sigma}_{x\varepsilon} + \hat{\sigma}_{v\varepsilon}}{\hat{\sigma}_x^2 + 2\hat{\sigma}_{v\varepsilon} + \hat{\sigma}_v^2}$$

Note one thing about equation (8). There are two random variables in the model – ε and v. We cannot take the expected value of ratios of random variables, so we cannot identify whether $\hat{\beta}_1$ is unbiased. Therefore, we can only examine the consistency of $\hat{\beta}_1$.

Now, let's take the plim of $\hat{\beta}_1$. When the sample size grows ($n \to \infty$), we know that each of the variances and covariances in (8) are consistent estimates, and therefore

$$p\lim(\hat{\sigma}_x^2) = \sigma_x^2 \;\text{ and }\; p\lim(\hat{\sigma}_{xv}) = \sigma_{xv} \;\text{ and }\; p\lim(\hat{\sigma}_{x\varepsilon}) = \sigma_{x\varepsilon}\text{, etc.}$$

Therefore,

$$(9) \qquad p\lim(\hat{\beta}_1^*) = \frac{\beta_1 \sigma_x^2 + \beta_1 \sigma_{vx} + \sigma_{x\varepsilon} + \sigma_{v\varepsilon}}{\sigma_x^2 + 2\sigma_{v\varepsilon} + \sigma_v^2}$$

Recall from above that we assumed $\sigma_{xv} = \sigma_{v\varepsilon} = 0$ and we always assume $\sigma_{x\varepsilon} = 0$ which means (9) reduces to

$$(10) \qquad p\lim(\hat{\beta}_1^*) = \frac{\beta_1 \sigma_x^2}{\sigma_x^2 + \sigma_v^2} = \beta_1 \left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2} \right)$$

The ratio $\left(\frac{\sigma_x^2}{\sigma_x^2 + \sigma_v^2} \right) = \theta$ is called the reliability ratio. It represents the fraction of the variance in x_i^* that is due to the true variance in x. In equation (10) notice that with any measurement error $0 \le \theta \le 1$ and $p\lim(\hat{\beta}_1^*) = \beta_1 \theta < \beta_1$. Therefore, as σ_v^2 increases, the measurement error in x_i^* increases and $p\lim(\hat{\beta}_1^*)$ declines – in the limit, as ($n \to \infty$) $\hat{\beta}_1$ will not converge to the true value when there is measurement error in x.