Bent E. Sørensen

ECONOMETRICS I, Spring 2025

Bias of the OLS estimator when the regressor is measured with error.

Consider a regression model of form

$$y_i \;=\; \alpha \;+\; \beta x_i \;+ u_i \;.$$

Under the standard OLS assumptions (x_i fixed, $Eu_i = 0$, $Eu_iu_j = 0$ when $i \neq j$ and constant variance of the u_is) the efficient OLS-estimator of β (based on N observations) is

$$\hat{\beta} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \;.$$

(Note: you can assume the variables are demeaned if you want simpler notation.)
Now, because

$$y_i - \bar{y} = \alpha + \beta x_i + u_i - (\alpha + \beta\bar{x} + \bar{u}) = \beta(x_i - \bar{x}) + (u_i - \bar{u}) \;,$$

we have

$$\hat{\beta} \;=\; \frac{\sum(x_i - \bar{x})(\beta(x_i - \bar{x}) + (u_i - \bar{u}))}{\sum(x_i - \bar{x})^2} \;,$$

or

$$\hat{\beta} - \beta \;=\; \frac{\sum(x_i - \bar{x})(u_i - \bar{u})}{\sum(x_i - \bar{x})^2} = \frac{\frac{1}{N}\sum(x_i - \bar{x})(u_i - \bar{u})}{\frac{1}{N}\sum(x_i - \bar{x})^2}.$$

For $N \to \infty$, we have $\frac{1}{N}\sum(x_i - \bar{x})(u_i - \bar{u}) \to 0$ and $\frac{1}{N}\sum(x_i - \bar{x})^2 \to \text{var}(x)$, so the right hand side converges to zero; i.e., the OLS estimator is *consistent* ($\hat{\beta} \to \beta$).
If x_i is measured with error, this consistency result does not hold. Assume

$$x_i^* = x_i + e_i \;,$$

where e_i is a "classical measurement error" where $Ee_i = 0$, $Ee_ie_j = 0; i \neq j$ and $Ee_iu_j = 0; \forall i, j$. Now, if you regress y on x^* using the OLS formula, $\hat{\beta}$ will be biased towards zero; i.e. $E|\hat{\beta}| < E|\beta|$.

This is easy to demonstrate: We have

$$\hat{\beta} \;=\; \frac{\sum(x_i^* - \bar{x^*})(\beta(x_i - \bar{x}) + (u_i - \bar{u}))}{\sum(x_i^* - \bar{x^*})^2}$$

$$= \frac{\beta \frac{1}{N}\sum(x_i - \bar{x})(x_i - \bar{x}) + \beta \frac{1}{N}\sum(e_i - \bar{e})(x_i - \bar{x}) + \frac{1}{N}\sum(x_i^* - \bar{x^*})(u_i - \bar{u})}{\frac{1}{N}\sum(x_i - \bar{x} + e_i - \bar{e})^2} ,$$

where the second and third terms in the numerator converges to 0 by the law of large numbers. We then have

$$\hat{\beta} \approx \beta \frac{\frac{1}{N}\sum(x_i - \bar{x})^2}{\frac{1}{N}\sum(x_i - \bar{x})^2 + \frac{1}{N}\sum(e_i - \bar{e})^2 + \frac{1}{N}\sum((x_i - \bar{x})(e_i - \bar{e}))} \rightarrow \beta \frac{var(x)}{var(x) + var(e)} .$$

This demonstrates that $\hat{\beta}$ converges to the true β times a term numerically smaller than 1. We say that (classical) measurement error leads to "bias towards zero." Notice, that if your main task is to show that some β is non-zero and the coefficient is significant, then it is usually safe to assume that the numerically larger coefficient you would get is significant. (In small samples, I should say"would most likely get" because random variables are random.)

The result here is much more applicable than it looks at first blush. We usually estimate multiple regression models, but if your variable of interest is measured with error and the controls are not, then the Frisch-Waugh residual $M_W X$ from regression on the controls W only has the measurement error coming from X and the result from the univariable regression is back. If there is measurement error in W instead, there will be measurement error in $M_W X$ but it won't be classical measurement error, so you need to convince yourself, if you can, that the problem is minor, for example because X is almost orthogonal to the W (although this is usually not the case). Or maybe you have an idea of the variance of the measurement error in W and you can do a little simulation study to evalue how much it may affect X. Or...use common sense, but educated common sense using the insights from this note.g