A Note on Bias in Conventional Standard Errors under Heteroskedasticity

Joshua Angrist and Jörn-Steffen Pischke

November 29, 2010

The conventional standard error for the slope parameter in the bivariate regression model

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

is given by

$$\left[\sigma^2_{\widehat{\beta}}\right]_{conv} = \frac{1}{n}\frac{\sigma^2_e}{Var(x_i)}.$$

The true sampling variance is given by

$$\sigma^2_{\widehat{\beta}} = \frac{1}{n}\frac{Var\left[e_i\left(x_i - \overline{x}\right)\right]}{\left[Var(x_i)\right]^2}.$$

Notice that

$$\left[\sigma^2_{\widehat{\beta}}\right]_{conv} > \sigma^2_{\widehat{\beta}} \Longleftrightarrow \sigma^2_e > \frac{Var\left[e_i\left(x_i - \overline{x}\right)\right]}{Var(x_i)}.$$

Using the fact that

$$\begin{aligned} Var\left[e_i\left(x_i - \overline{x}\right)\right] &= E\left[e_i^2\left(x_i - \overline{x}\right)^2\right] \\ &= E\left[e_i^2\right]E\left(x_i - \overline{x}\right)^2 + Cov\left[e_i^2, \left(x_i - \overline{x}\right)^2\right] \end{aligned}$$

we get

$$\begin{aligned} \frac{Var\left[e_i\left(x_i - \overline{x}\right)\right]}{Var(x_i)} &= \frac{\sigma^2_e Var(x_i) + Cov\left[e_i^2, \left(x_i - \overline{x}\right)^2\right]}{Var(x_i)} \\ &= \sigma^2_e + \frac{Cov\left[e_i^2, \left(x_i - \overline{x}\right)^2\right]}{Var(x_i)} \end{aligned}$$

Hence we have

$$\left[\sigma^2_{\widehat{\beta}}\right]_{conv} > \sigma^2_{\widehat{\beta}} \Longleftrightarrow Cov\left[e_i^2, (x_i - \overline{x})^2\right] < 0.$$

Conventional standard errors are biased up whenever observations on x_i far from the mean (observations with high leverage) are associated with lower variance residuals.

An equivalent result holds in multivariate regressions. This follows easily by first partialling out any additional regressors from both x_i and y_i. All the results above also hold for these residuals.