A Note on Measurement Error and Proxy Variables

Bronwyn H. Hall

It is very common for researchers to include multiple indicators for the same underlying concept in their regressions. It is equally common for them to discover that one variable enters with a positive sign and the second with a negative sign. Most students who have completed a basic regression course understand that obtaining opposite signs on the coefficients in a two variable regression implies positive correlation for the two independent variables, as in this case. What is less obvious is that when the measurement error in the two proxies is positively correlated, the same result (opposite signs) can be obtained even if the true relation between the dependent variable and the underlying concept was positive. The purpose of this note is to simply to point out this fact and indicate how the bias varies with respect to different properties of the measurement error.

The true model is assumed to be the following:

$$y_i = \beta x_i^* + \varepsilon_i \qquad i = 1, ..., N$$

The researcher has two proxies available for x^*, x_1 and x_2. These proxies are measured with error, and the measurement errors may be correlated with each other, but not with the underlying disturbance ε:

$$x_{1i} = x_i^* + w_{1i}$$
$$x_{2i} = x_i^* + w_{2i}$$
$$Cov(w_i) = \begin{bmatrix} \sigma_1^2 & \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

In general I will assume that x_1 is a better proxy than x_2, that is, that σ_1 is smaller than σ_2.[1] The researcher estimates the following regression:

$$y_i = \gamma_1 x_{1i} + \gamma_2 x_{2i} + u_i$$

What will be the resulting estimates for γ_1 and γ_2? The conditional expectation of y given x_1 and x_2 depends on the conditional expectation of x^* given x_1 and x_2:

$$E[y \mid x_1, x_2] = \beta E[x^* \mid x_1, x_2] = \beta \frac{Cov(x^*, x)}{Var(x)}$$

[1] For example, a common application is to use both R&D and patents to proxy for the innovative activity of a firm. It is well-known that R&D is a "better" measure than counting patents in most relationships (see Griliches, Hall, and Pakes 1987). But the error in the two variables in measuring innovative activity may be related – they both require conscious innovation activities on the part of the firm. In addition, GHP show that the noise to signal ratio in the patent count variable is likely to be about one per cent, which will increase the bias even for small amounts of correlation.

Where I have conditioned on the actual x_1 and x_2 that were observed. These covariances and variances are easily computed give the assumptions of the model above.

$$Cov(x^*, x) = \begin{bmatrix} \sigma_x^2 \\ \sigma_x^2 \end{bmatrix}$$

$$Var(x) = \begin{bmatrix} \sigma_x^2 + \sigma_1^2 & \\ \sigma_x^2 + \rho\sigma_1\sigma_2 & \sigma_x^2 + \sigma_2^2 \end{bmatrix}$$

Using these formulas, one can show that the estimated γ has the following form:

$$E\begin{bmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_2 \end{bmatrix} = \begin{bmatrix} \dfrac{\beta\sigma_x^2(\sigma_2^2 - \rho\sigma_1\sigma_2)}{|Var(x)|} \\ \dfrac{\beta\sigma_x^2(\sigma_1^2 - \rho\sigma_1\sigma_2)}{|Var(x)|} \end{bmatrix}$$

The condition for the expected value of the estimated γ_2 to be negative is easy to derive:

$$E[\hat{\gamma}_2] < 0 \Leftrightarrow \rho > \sigma_1 / \sigma_2$$

Thus we can expect a negative coefficient on the proxy that is poorly measured when correlation of the measurement error is high OR when the variance ratio is small, that is, when one variable is measured much worse than the other.

Simulation results also reveal that for samples of any size, the estimated γ_1 in this regression will be positive (and less than unity) and the estimated γ_2 will be negative if the correlation coefficient ρ is high enough. As the sample size grows, things get worse, not better, since the measurement error bias becomes better and better determined. See Figures 1 and 2, which show the t-statistic on γ_2 as a function of ρ for two variance ratios (10 and 1 per cent) and 4 sample sizes.

As usual with measurement error, the sum of the two coefficients is a biased estimate of the true β:

$$E[\hat{\gamma}_1 + \hat{\gamma}_2] = \frac{\beta(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}{\left(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2 + (1-\rho^2)\dfrac{\sigma_1^2\sigma_2^2}{\sigma_X^2}\right)}$$

Surprisingly, when the measurement error is perfectly correlated (postiively or negatively), that is, when $|\rho|=1$, the bias in the sum is zero, even though the bias in each individual coefficient may be large. The estimated sum as a function of ρ is shown in Figure 3.

Figure 1

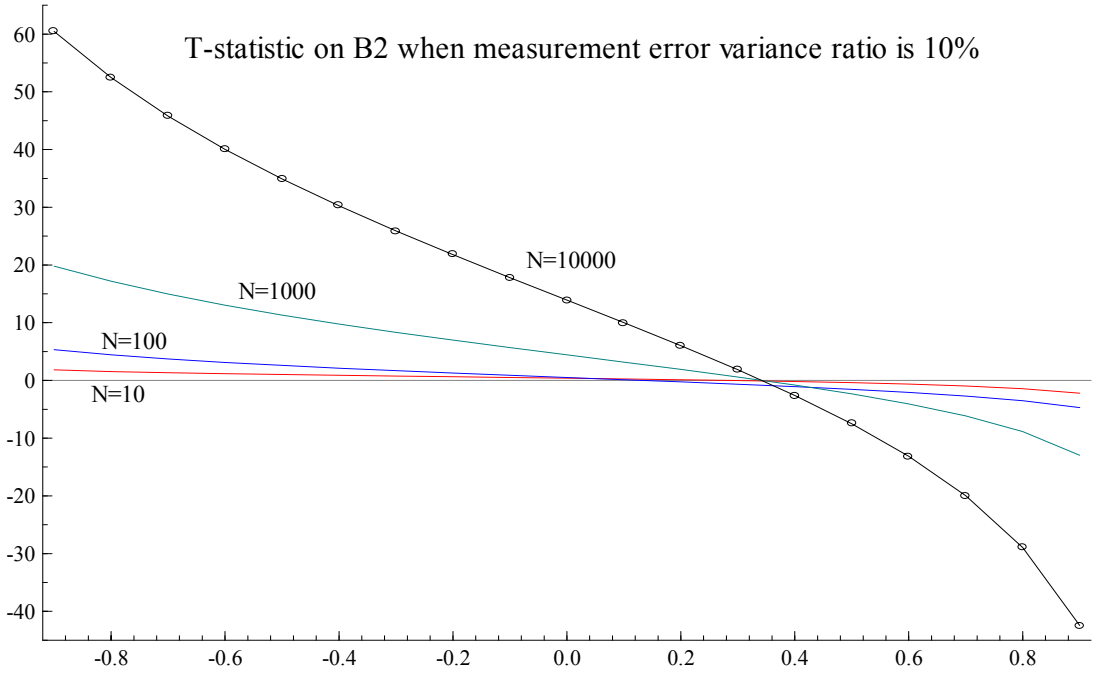


Figure 2

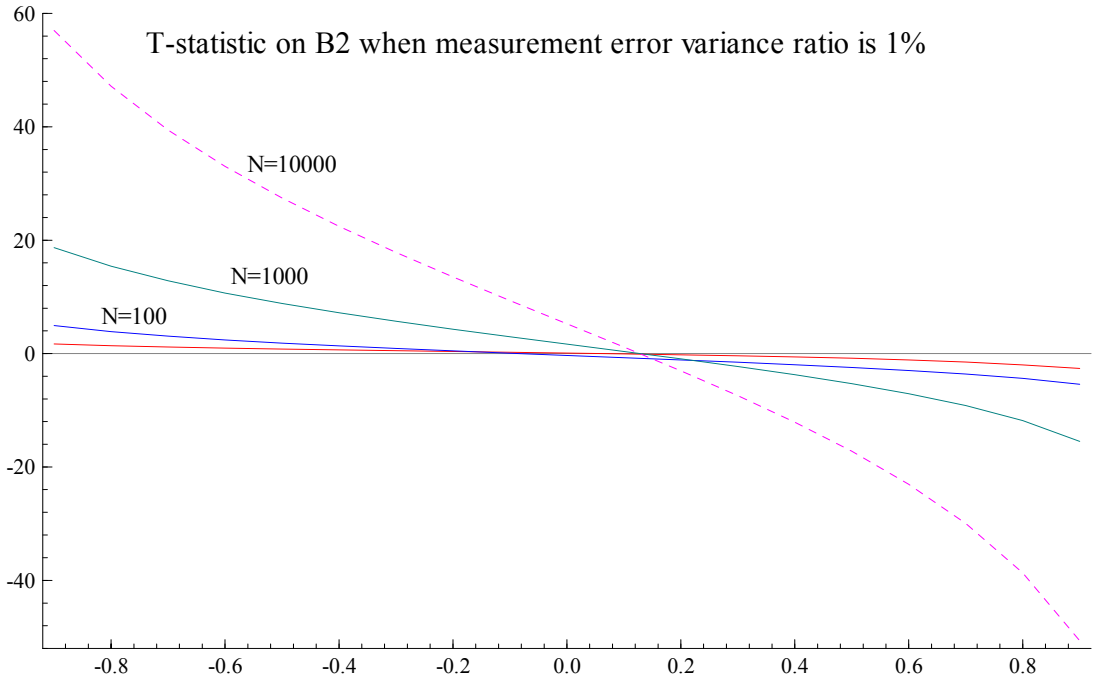
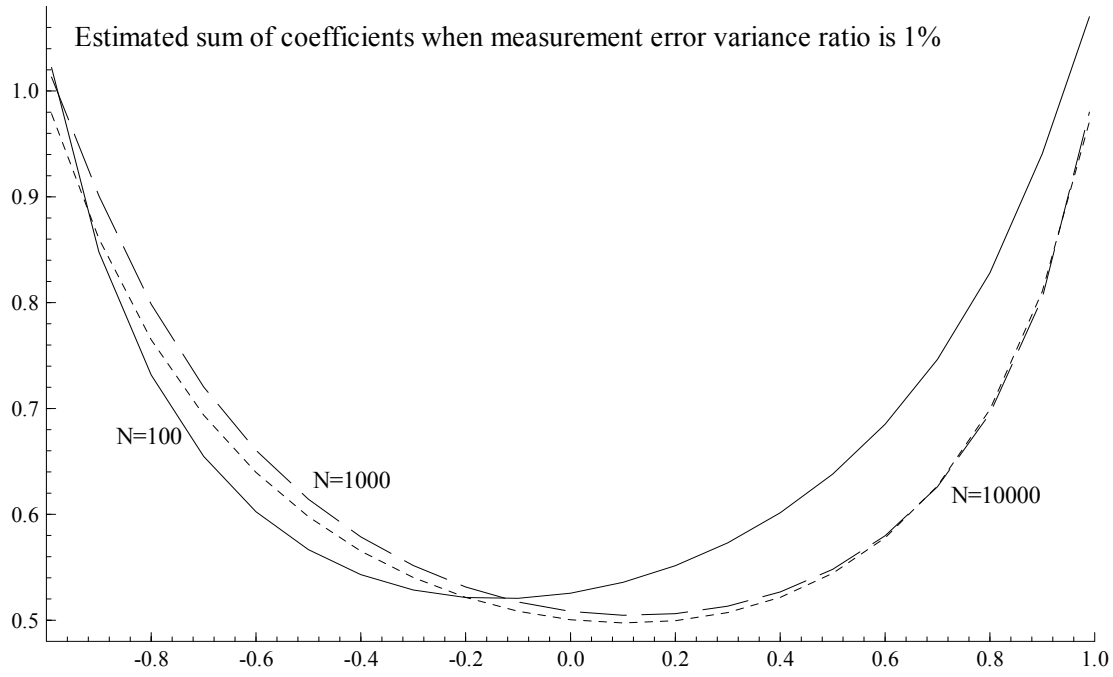


Figure 3


Estimated sum of coefficients when measurement error variance ratio is 1%
1.0
0.9
0.8
0.7
0.6
0.5
N=100
N=1000
N=10000
-0.8
-0.6
-0.4
-0.2
0.0
0.2
0.4
0.6
0.8
1.0