Croon's Measurement Error Correction

Jorge Luis García[*,†]

First Draft: October 16, 2014
This Draft: October 24, 2014

Abstract

I summarize the procedure in Croon (2002) to correct for measurement error and I accompany it with an implementation in Stata.

In the following lines we explain very briefly how to correct for measurement error estimates from standard linear regressions estimated through OLS.

Let y_i be a (demeaned) outcome of interest, X_i be a (demeaned) vector of characteristics or controls, and θ_i be a (demeaned) vector of factors. X_i is of dimension $1 \times K$ and θ_i is of dimension $1 \times K$. The objective is to estimate the coefficients in the following model

$$y_i = \mathbf{X_i}\beta + \theta_i\gamma + \varepsilon_i \tag{1}$$

where β (vector) and γ (vector) are estimands and $\mathbb{E}[\varepsilon_i|\mathbf{X_i}, \theta_i] = 0$.

Plausibly, θ_i is measured with error, since it is a factor score based on multiple measures arbitrarily chosen by the the researcher.

Croon (2002) proposes a correction. Let the relationship between the "true factor scores" and the "estimated" factor scores be given by

$$\hat{\theta}_\mathbf{i} = \theta_\mathbf{i} + V_i \tag{2}$$

where $\mathbb{E}\left[V_i\right] = 0, \mathbb{E}[V_i|\mathbf{X_i}, \theta_i] = 0$. Let the convariance matrix of $\mathbf{R_i}$ and $\mathbf{S_i}$ be given by

$$\Sigma = \begin{bmatrix} \Sigma_{\mathbf{R_i},\mathbf{R_i}} & \Sigma_{\mathbf{R_i},\mathbf{S_i}} \\ \Sigma_{\mathbf{S_i},\mathbf{R_i}} & \Sigma_{\mathbf{S_i},\mathbf{S_i}} \end{bmatrix}. \tag{3}$$

If we assume (4) is the data generating process, the OLS estimator is inconsistent. To see this substitute (2) into (4) and obtain the following:

$$y_i = \beta_0 + \mathbf{X_i}\beta + \theta_\mathbf{i}\gamma + \varepsilon_i - \gamma V_i \tag{4}$$

so that

$$\text{plim} \begin{bmatrix} \hat{\beta}^{OLS} \\ \hat{\gamma}^{OLS} \end{bmatrix} = A \begin{bmatrix} \beta \\ \gamma \end{bmatrix} \tag{5}$$

where $A = \begin{bmatrix} \Sigma_{\hat{\theta}_\mathbf{i},\hat{\theta}_\mathbf{i}} & \Sigma_{\theta_\mathbf{i},\mathbf{X_i}} \\ \Sigma_{\mathbf{X_i},\theta_\mathbf{i}} & \Sigma_{\mathbf{X_i},\mathbf{X_i}}. \end{bmatrix}^{-1} \begin{bmatrix} \Sigma_{\theta_\mathbf{i},\theta_\mathbf{i}} & \Sigma_{\theta_\mathbf{i},\mathbf{X_i}} \\ \Sigma_{\mathbf{X_i},\theta_\mathbf{i}} & \Sigma_{\mathbf{X_i},\mathbf{X_i}}. \end{bmatrix}$. Importantly

[*]Department of Economics, the University of Chicago (jorgelgarcia@uchicago.edu).
[†]I thank Sneha Elango for helpful comments.

$$\Sigma_{\hat{\theta}_\mathbf{i},\hat{\theta}_\mathbf{i}} = \Sigma_{\theta_\mathbf{i},\theta_\mathbf{i}} + \Sigma_{V_i,V_i}. \tag{6}$$

Now note that $\Sigma_{\theta_\mathbf{i},\mathbf{X_i}} = \Sigma_{\hat{\theta}_\mathbf{i},\mathbf{X_i}}$, so that we can estimate a. To get it straight, $\theta_\mathbf{i}$ is the vector of factor scores in the data. Thus, it's easy to obtain all the components in A except for the component in (6). They come from computing the empirical covariance matrices between the factor scores, $\theta_\mathbf{i}$, and the rest of the data in the estimation (controls), $\mathbf{X_i}$. To obtain $\Sigma_{\hat{\theta}_\mathbf{i},\hat{\theta}_\mathbf{i}}$ calculate the variance of the factor scores, which depends on the factor extraction method. Once all the components of A have been estimated, premultiplying the OLS estimates by its inverse yields consistent estimates:

$$A^{-1}\,\mathrm{plim}\begin{bmatrix} \hat{\beta}^{OLS} \\ \hat{\gamma}^{OLS} \end{bmatrix} = \begin{bmatrix} \beta \\ \gamma \end{bmatrix}. \tag{7}$$

Put differently, premultiplying the OLS estimates by A^{-1} is Croon's correction.

References

Croon, M. (2002). Using Predicted Latent Scores in General Latent Structure Models. *Latent Variable and Latent Structure Models*, 195–223.