A Reconsideration of Consistent Estimation of a Dynamic Panel Data Model in the Random Effects (Error Components) Framework

Richard A. Ashley*
Department of Economics
Virginia Polytechnic Institute and State University

April 19, 2010†

Abstract

It is widely believed that the inclusion of lagged dependent variables in a panel data model necessarily renders the Random Effects (RE) estimators, based on OLS applied to the quasi-differenced variables, inconsistent. It is shown here that this belief is incorrect under the usual assumption made in this context — i.e., that the other regressors are strictly exogenous. This result follows from the fact that lagged values of the deviation of the quasi-differenced dependent variable from its mean can be written as a weighted sum of the past values of the quasi-differenced model error term, whereas these quasi-differenced errors are serially uncorrelated by construction. The RE estimators are therefore consistent. Thus, since instrumental variables methods – e.g., Arellano and Bond (1991) – clearly provide less precise estimates, the RE estimates are preferable if a Hausman test is unable to reject the null hypothesis that the parameter estimates of interest from both methods are equal.

Keywords: Panel data, random effects model, error components model.

JEL Classification: C23.

*Richard A. Ashley, Department of Economics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0316. Phone: 540-231-6220, Fax: 540-231-5097, E-mail: ashleyr@vt.edu.

†This version is preliminary; please consult the author prior to quoting. Updated versions will be posted at http://ashleymac.econ.vt.edu/ashleyprofile.htm.

1 Introduction

It is widely held that the usual estimator of the parameters in the Random Effects (RE) error components model yields inconsistent estimates in dynamic panel data models, due to presumed correlation between quasi-differenced values of the lagged dependent variable and the quasi-differenced model error term. For example, Baltagi (2008, p. 148) states, "The random effects GLS estimator is also biased in a dynamic panel data model. In order to apply GLS, quasi-demeaning is performed ... and $(y_{i,t-1} - \theta \bar{y}_{i.,-1})$ will be correlated with $(u_{i,t-1} - \theta \bar{u}_{i.,-1})$." Fortunately, this presumption is incorrect: it is shown below that quasi-differenced lags in the dependent variable are in fact uncorrelated with the (quasi-differenced) model errors in this setting, under the usual assumption that the other regressors are strictly exogenous. This result implies that the RE estimator, based on the quasi-differenced model, is still consistent for the model parameters in a dynamic panel data model.

2 Proof

Let $Y_{i,t}$ be generated by the usual linear error components model, with a single lagged dependent variable and $X_{1,t} \ldots X_{k,t}$ strictly exogenous. That is,

$$Y_{i,t} = \alpha + \rho Y_{i,t-1} + \sum_{j=1}^{k} \beta_j X_{j,t} + \nu_i + \varepsilon_{i,t}, \tag{1}$$

where

$$(\nu_i, \varepsilon_{i,t})^t \sim i.i.d. \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\nu^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{pmatrix} \right] \tag{2}$$

and where $\text{corr}(X_{j,t}, \nu_i + \varepsilon_{i,\tau})$ is zero for all values of j, t, i, and τ. A single lag in $Y_{i,t}$ is posited – with $|\rho| < 1$, so that the lag operator $(1 - \rho B)^{-1}$ exists – but the proof extends in an obvious way to multiple lags in the dependent variable. The usual RE quasi-difference of $Y_{i,t}$ is denoted $\tilde{Y}_{i,t}$, where

$$\tilde{Y}_{i,t} = Y_{i,t} - \theta_i \bar{Y}_{i.} \tag{3}$$

and where $\bar{Y}_{i.}$ is the time-average of $Y_{i,t}$ over the T_i observations available for unit i; the quasi-difference of each other variable and error term is similarly defined and notated. Note that θ_i is the usual GLS value,[1] which is precisely that function of σ_ν^2, σ_ϵ^2, and T_i which forces

$$cov(\tilde{\nu}_i + \tilde{\varepsilon}_{i,t}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t-\lambda}) = 0 \tag{4}$$

for all non-zero values of the relative lag λ, even though the model errors in [1] – i.e., $\nu_i + \varepsilon_{i,t}$ – are, of course, serially correlated.

Then, from [1] and [3], $\tilde{Y}_{i,t}$ must satisfy the equation

$$(1 - \rho B)\tilde{Y}_{i,t} = \alpha(1 - \theta_i) + \sum_{j=1}^{k} \beta_j \tilde{X}_{j,t} + \tilde{\nu}_i + \tilde{\varepsilon}_{i,t}, \tag{5}$$

where B is the lag operator. The inverse of (1 - ρB) is $\sum_{s=0}^{\infty} \rho^s B^s$, so [5] implies that $\tilde{Y}_{i,t}$ can be expressed

[1] I.e., θ_i is $1 - \sqrt{\frac{\sigma_\varepsilon^2}{T_i \sigma_\nu^2 + \sigma_\epsilon^2}}$. This is, of course, replaced by a consistent estimate in FGLS.

as

$$\tilde{Y}_{i,t} = \frac{\alpha(1-\theta_i)}{1-\rho} + \sum_{j=1}^{k} \beta_j \sum_{s=0}^{\infty} \rho^s \tilde{X}_{j,t-s} + \sum_{s=0}^{\infty} \rho^s (\tilde{\nu}_i + \tilde{\varepsilon}_{i,t-s}) \tag{6}$$

Hence, the covariance of the *lagged* dependent variable ($\tilde{Y}_{i,t-1}$) with the *current* value of the quasi-differenced model error ($\tilde{\nu}_i + \tilde{\varepsilon}_{i,t}$) is

$$cov(\tilde{Y}_{i,t-1}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t}) = \sum_{j=1}^{k} \beta_j \sum_{s=0}^{\infty} \rho^s cov(\tilde{X}_{j,t-1-s}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t}) + \sum_{s=0}^{\infty} \rho^s cov(\tilde{\nu}_i + \tilde{\varepsilon}_{i,t-1-s}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t}) \tag{7}$$

$$= \sum_{j=1}^{k} \beta_j \sum_{s=0}^{\infty} \rho^s cov(\tilde{X}_{j,t-1-s}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t}) \tag{8}$$

where the last equality follows from [4].

If $X_{1,t} \ldots X_{k,t}$ are strongly exogenous, then so are the quasi-differenced explanatory variables, $\tilde{X}_{1,t} \ldots \tilde{X}_{k,t}$. Consequently, the covariance of $\tilde{X}_{j,t-1-s}$ with $\tilde{\nu}_i + \tilde{\varepsilon}_{i,t}$ is zero for all non-zero values of s; thus, [8] implies that $cov(\tilde{Y}_{i,t-1}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t})$ is zero.

This completes the proof that the lagged value of the quasi-difference dependent variable is uncorrelated with the quasi-differenced model error.[2] Hence, if $X_{1,t} \ldots X_{k,t}$ are strongly exogenous — as is most typically assumed — then the usual RE estimator provides consistent estimators of the parameters in [1].

If, however, there is feedback between, say, $X_{j,t}$ and $Y_{i,t}$, then $X_{j,t}$ is only weakly exogenous. The quasi-differenced lagged values of $X_{j,t}$ will in that case be correlated with the quasi-differenced current error term, leading to a non-zero value of $cov(\tilde{Y}_{i,t-1}, \tilde{\nu}_i + \tilde{\varepsilon}_{i,t})$ in [8] and, consequently, to inconsistency in the RE estimator. This inconsistency could be significant if the feedback is strong and the panels are short, in which case one would be better off using the Arellano and Bond (1991) or Keane and Runkle (1992) estimators; this proposition could be tested by comparing the RE estimator to one of these estimators on a generalized Hausman test.[3]

3 Implications

The RE estimator is shown above to be consistent for the parameters in the dynamic panel data model [1] if the regressors are strongly exogenous. And the RE estimator is obviously more efficient than the instrumental variables (IV) estimators usually used in estimating dynamic panel data models - e.g. Arellano and Bond (1991). Hence, if a Hausman test cannot reject the null hypothesis that the RE parameter estimates of interest are equal to the IV estimates, then the strong exogeneity assumption is itself not rejected and the RE estimator is preferable.

[2] This proof would be essentially identical with a larger number of lags in the dependent variable. For example, if one replaces $\rho Y_{i,t-1}$ in [1] by $\phi_1 Y_{i,t-1} + \phi_2 Y_{i,t-2} + \cdots + \phi_p Y_{i,t-p}$ then the same proof goes through, so long as the inverse of the lag operator $1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ exists.

[3] Such a test is not hard to implement. In Stata, for example, one can obtain both an IV estimator on the first differenced model and the RE estimator in a *do* file using the *regress* command, combine the stored estimates using the *suest* command, and then directly test the coefficients of interest for equality across the two methods.

References

[1] Arellano, T. W. and S. Bond, 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations," *Review of Economic Studies*, 58 pp. 277-297.

[2] Baltagi, B. H., 2008. *Econometric Analysis of Panel Data*, Wiley: New York.

[3] Keane, M. P. and D. E. Runkle, 1992. "On the Estimation of Panel-Data Models with Serial Correlation When Instruments Are Not Strictly Exogenous," *Journal of Business and Economic Statistics*, 10 pp. 1-9.