Estimation and Specification Testing of Panel Data Models with Non-Ignorable Persistent Heterogeneity

by
Vassilis Hajivassiliou
Department of Economics and Financial Markets Group,
London School of Economics

December 2009

Abstract

This paper shows how a simple modification of estimators based on the Random Effects principle can preserve the consistency and asymptotic efficiency of the method in panel data despite non-ignorable persistent heterogeneity driven by correlations between the heterogeneity and the regressors. The approach is extremely easy to implement and allows straightforward tests of the significance of such correlations that lie behind the non-ignorable persistent heterogeneity. The method applies to linear as well as nonlinear panel data models, static or dynamic.

Keywords: Panel Data, Correlated Random Effects, Fixed Effects, Unobserved Heterogeneity

JEL Classifications: C51, C52

*I have benefited from constructive comments by participants at seminars at the London School of Economics, the University of Toulouse, and the World Bank Development Research Group, as well as at the 2006 Conference on Panel Data at the University of Cambridge. A preliminary version of this paper was circulated as LSE working paper Hajivassiliou (2006). I would like to thank especially Martin Ravallion and Yannis Ioannides for their remarks and suggestions. The usual disclaimer applies for any remaining errors and shortcomings.

Address correspondence to:
E-mail: vassilis@lse.ac.uk
Address: Department of Economics
London School of Economics
London WC2A 2AE, England

Estimation and Specification Testing of Panel Data Models with Non-Ignorable Persistent Heterogeneity

1 The Problem

Consider three classic cases of panel data models with time-varying and time-invariant regressors x and z respectively:

A. Linear Static:

$$y_{it} = x'_{it}\beta + z'_i\gamma + \epsilon_{it} \tag{1}$$

B. Linear Dynamic:

$$y_{it} = \delta y_{i,t-1} + x'_{it}\beta + z'_i\gamma + \epsilon_{it} \tag{2}$$

C. Nonlinear with nonadditive errors:

$$y_{it} = h\left(x'_{it}\beta + z'_i\gamma + \epsilon_{it}\right) \tag{3}$$

where $h(\cdot)$ is a known function, allowed to be nondifferentiable and discontinuous. Limited Dependent Variable (LDV) models are clearly a special version of this. For simplicity, we assume a balanced data set indexed by $i = 1, \cdots, N, \quad t = 1, \cdots, T$. We concentrate on the common situation of large N, and small to moderately large T.[1] In each case, suppose that ϵ_{it} follows the one-factor error components structure $\epsilon_{it} = \alpha_i + \nu_{it}$, with $E(\nu_{it}|X, Z) = 0$ and α and ν independent for any i,t.

A usual problem in many practical cases is that α_i may be believed to be correlated with one or more of the regressors (x'_{it}, z'_i). We define this problem as "Non-Ignorable Persistent Heterogeneity," which results in inconsistency of estimators based on the Random-Effects (RE) principle. This problem very frequently leads applied researchers to adopt Fixed-Effects type estimators (FE), which are not affected by such random effects-regressors correlations. These decisions are predicated on the well-known fact that such correlations normally wreak havoc to estimators that are based on the standard RE principle of accounting for the non-sphericality of the error term distribution through suitable GLS and MLE methods.

Estimators based on the FE principle either eliminate or condition upon the persistent heterogeneity term α_i and are thus consistent irrespective of any regressor-heterogeneity correlations. These estimators for (1) yield Ordinary Least Squares estimation after applying either first-differencing $(w_{it} - w_{i,t-1})$ or the within transformation $(w_{it} - \frac{1}{T}\sum_{t=1}^{T} w_{it})$, where w_{it} stands in for the dependent variable y_{it} and all

[1] Exogenously unbalanced data sets can be accommodated readily. In case the causes of unbalancedness are endogenously determined, all models become of category C, since a valid probability model characterizing the data availability necessarily introduces a nonlinearity of type (3). We let X and Z denote the matrices of the complete sample data on the time-varying and time-invariant regressors respectively.

the regressors x_{it}^j and z_{it}^l; for (2) they yield Instrumental Variables estimation using sufficiently older lags of the dependent variable ($y_{i,t-l}$, $l > 1$) [see Arellano and Bond (1991)]; and for (3) they are in general inconsistent due to the incidental parameters problem.[2]

It is our view that abandoning RE estimation in favour of FE in such situations is premature, unnecessary, and likely to have rather unfortunate consequences. This is because well-understood shortcomings of estimators based on the FE principle include, *inter alia*: (a) FE-type methods provide no estimates in general for the time-invariant coefficients γ; (b) since N α_i parameters are implicitly or explicitly estimated, such methods suffer substantial efficiency losses as compared to methods based on the RE principle; and (c) the within and first-differencing transformations typically reduce very significantly the signal-to-noise ratio of the time-varying regressors, thus resulting in serious inconsistencies in FE-based methods.[3]

2 Modified Random Effects Estimation

We show how a simple modification of estimators based on the RE principle, following ideas of Mundlak (1978) and Chamberlain (1984), can preserve the consistency and asymptotic efficiency of the RE methodology.

Our approach models explicitly the suspected non-ignorable persistent heterogeneity by characterizing its correlation with the regressors as:

$$E(\alpha_i|X,Z) = \mu_i = g(X,Z) \tag{4}$$

and considering specific functions $g(\cdot)$. For example for the case without time-invariant regressors z_i, Mundlak (op.cit.) proposed $\mu_i = \bar{x}_{i\cdot}'\xi$ where $\bar{x}_{i\cdot} \equiv \frac{1}{T_i}\sum_{t=1}^{T_i} x_{it}$ is the time average of the regressor vector.[4] An alternative proposal was Chamberlain (op.cit.) who modelled instead this conditional mean as $E(\alpha_i|X) = \sum_{t=1}^{T_i} r_t x_{it}$ where r_t are period-specific weights.

We introduce three assumptions concerning the conditional mean function $g(\cdot)$ characterizing the correlation between the unobserved persistent heterogeneity α_i and regressors x and z:

Assumption 1: $g(\cdot)$ is a linear *function of the regressors;*

Assumption 2: $g(\cdot)$ depends only on the regressor data for individual i; and

Assumption 3: $g(\cdot)$ only depends on the regressors in a time-invariant *way.*

[2] In very specific cases, consistent FE estimators exist for (3), e.g., the conditional logit model of Chamberlain (1980).

[3] These shortcomings can be explained in an intuituive way by noting that the FE-based methods sweep away also *ignorable* heterogeneity (that is uncorrelated with regressors). Hence, they clean out "too much" and make it harder to precisely identify the effects of main interest (β).

[4] Hajivassiliou (1985) used a similar approach for deriving formal tests of regressor-heterogeneity correlations in a switching regressions framework.

Assumptions (1)-(3) are satisfied by the Mundlak error model after extending it for the presence of invariant regressors, by defining:

$$E(\alpha_i|X,Z) = g(X,Z) = \bar{x}_{i\cdot}'\xi + z_i'\zeta \tag{5}$$

If we now write

$$\alpha_i^* \equiv \alpha_i - \bar{x}_{i\cdot}'\xi - z_i'\zeta \tag{6}$$

this new persistent heterogeneity term has by construction conditional mean zero. We can thus substitute out α_i from (1), (2), and (3) in each of the three classic cases considered and collect terms.

Specifically, for each of the canonical models above we obtain:

A. Modified Linear Static:

$$y_{it} = x_{it}'\beta + \bar{x}_{i\cdot}'\xi + z_i'(\gamma + \zeta) + \alpha_i^* + \nu_{it} \tag{7}$$

B. Modified Linear Dynamic:

$$y_{it} = \delta y_{i,t-1} + x_{it}'\beta + \bar{x}_{i\cdot}'\xi + z_i'(\gamma + \zeta) + \alpha_i^* + \nu_{it} \tag{8}$$

C. Modified Nonlinear with nonadditive errors:

$$y_{it} = h\left(x_{it}'\beta + \bar{x}_{i\cdot}'\xi + z_i'\left(\gamma + \zeta\right) + \alpha_i^* + \nu_{it}\right) \tag{9}$$

Since by construction $E(\alpha_i^*|X,Z) = 0$ and $E(\nu_{it}|X,Z) = 0$ by assumption, this approach results in modified models with well-behaved random persistent heterogeneity effects that do not pose consistency problems for GLS/MLE estimation: *the solution proposed here thus involves simply adding the time-averages of the time-varying regressors as additional regressors in the RHS of the respective panel data model and proceeding with the RE estimator that is appropriate for each case.*[5] Consequently, our modified RE estimators will have the usual optimality properties: for case A the optimal RE/GLS estimator corresponds to OLS of the model (7) made spherical by applying the transformation $(w_{it} - \lambda\bar{w}_{i\cdot})$; for case B, optimal RE corresponds to FIML and 3SLS applied to (8) written as a cross-sectional simultaneous equations system of T equations, one per period [see Barghava and Sargan (1982)]; and for case C, efficient estimation is achieved through MLE, possibly with the aid of simulation-based inference in case likelihood contributions involve high dimensional integrals [see inter alia Hajivassiliou (1993)].[6]

[5] If it is believed that the correlation function $g(X,Z)$ should allow for nonlinearities in the regressors, our method can accomodate this by expanding (5) to contain polynomials in $\bar{x}_{i\cdot}$ and z_i.

[6] For nonlinear *dynamic* models, the methods of Wooldridge (2005) are useful for handling the initial conditions problem inherent in such models.

3 Testing for Non-Ignorable Persistent Heterogeneity

Our approach enables also straightforward testing of the significance of correlation between the regressors and the persistent heterogeneity, which would render it non-ignorable: under the maintained hypothesis of this paper, a classical test (by employing any of the traditional methods of Lagrange Multiplier, Likelihood Ratio, or Wald) of the time-averages $\bar{x}_{i.}$ when entered as additional regressors, provides a formal test as to whether the conditional mean function $E(\alpha_i|X,Z)$ indeed depends on the X regressors. To the extent that the conditional mean model is only an approximation, such significance tests should be viewed as omnibus specification tests of the presence of important Regressor-Heterogeneity correlations that are modelled less precisely.

Finally, specification tests in the Wu-Hausman mould can be constructed by comparing alternative estimates of the β parameters. In particular consider traditional FE estimates $\hat{\beta}_{FE}$ that are consistent irrespective of Heterogeneity-Regressor correlations; traditional $\hat{\beta}_{RE}$ estimates that are consistent and efficient under the assumption of no Regressor-Heterogeneity correlations $E(\alpha_i|X,Z)=0$; and our modified RE $\hat{\beta}_{MRE}$.estimator that is consistent and efficient under the correlation model $E(\alpha_i|X,Z)=\bar{x}'_{i.}\xi+z'_i\zeta$. Constructing Wu-Hausman quadratic forms based on pairing $\hat{\beta}_{MRE}$ with $\hat{\beta}_{FE}$ on one hand and with $\hat{\beta}_{RE}$ on the other yields straightforward specification tests in this context.

4 Interpreting Coefficients of Time-Invariant Regressors $\bar{x}_{i.}$ and z_i

It is a direct consequence of our approach that the time-invariant regressor coefficients γ are not identifiable separately from parameter vector ζ, as can be seen from equations (7)-(9). At first glance this may appear as a limitation of the approach we propose. Upon further reflection, however, one realizes that our approach actually yields the correct marginal effects with respect to changes in regressor variables, taking into account both the direct as well as the indirect effects of such changes. To illustrate, consider a change in time-varying regressor j, say Δx^j_{it} and a change in a time-invariant regressor m, say Δz^m_i. Given that we focus on the case $E(\alpha_i|X,Z)=g(X,Z)$ where we assume specifically that $g(X,Z)$ is well modelled by $\bar{x}'_{i.}\xi+z'_i\zeta$, it follows that for panel data Model A the expected marginal effect of a change Δx^j_{it} that is relevant for policy-making purposes is:[7]

$$\Delta E(y_{it}|X,Z)/\Delta x^j_{it}=\beta^j+\frac{1}{T}\xi^j$$

[7] This formula needs to be adjusted accordingly in case the change in x^j is assumed to persist for longer than one period.

while for a change Δz_i^m it is:

$$\Delta E(y_{it}|X,Z)/\Delta z_i^m = \gamma^m + \zeta^m$$

Our method provides estimates of both marginal effects as derived here, since it yields separately parameter vectors β and ξ as well as the combined vector $\gamma + \zeta$. Similar logic gives also the marginal effects for cases B and C, *mutatis mutandis*.[8]

5 Conclusions

This paper proposed a simple modification of estimators based on the Random Effects principle that preserves the consistency and asymptotic efficiency of the method in panel data despite the presence of non-ignorable persistent heterogeneity. It thus overcomes the well-known shortcomings of FE-type estimators. The approach is extremely easy to implement and entails simply adding the time-averages of the time-varying regressors as additional regressors. The method applies to linear as well as nonlinear panel data models, that are static or dynamic. We also showed how to construct classical and Wu-Hausman specification tests of correlations between the regressors and the persistent heterogeneity, which would render it non-ignorable.

References

[1] Arellano, M. and S. Bond (1991): "Test of Specification for Panel Data: Monte-Carlo Evidence and an Application to Employment," *Review of Economic Studies* **58**, 277-297.

[2] Barghava, A. and J. Sargan (1982): "Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods, *Econometrica* **51**, 1635-1660.

[3] Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, **47**, 225-238.

[4] Chamberlain, G. (1984): "Panel Data", in Z. Griliches and M. D. Intriligator (eds.), Handbook of Econometrics, Vol. 2, Elsevier Science.

[5] Hajivassiliou, V. (1985): "Disequilibrium Modelling in Economics and Related Limited Dependent Variables Models," M.I.T. *unpublished doctoral dissertation.*

[8] Note that under certain scenarios (e.g., Hausman and Taylor (1981)) it may be possible to extend the FE approach to recover estimates of the time-invariant parameters γ. That would allow one to identify separately the indirect effect vector ζ from the combined estimate generated by our modified RE method. In general, whether one desires the combined *direct plus indirect* $\gamma + \zeta$ or the two parameters separately will depend on the specific policy analysis one has in mind.

[6] Hajivassiliou, V. (1993): “Simulation Estimation Methods for Limited Dependent Variable Models.” In *Handbook of Statistics*, **11** (Econometrics), G.S. Maddala, C.R. Rao and H.D. Vinod (eds.). Amsterdam: North-Holland, 519–543.

[7] Hajivassiliou (2006): “A Modified Random Effects Estimator for Linear Panel Data Models with Regressor-Heterogeneity Correlations,” LSE working paper.

[8] Hausman, J. and W. Taylor (1981): “Panel Data and Unobservable Individual Effects," *Econometrica* **49**, 1377-1398.

[9] Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica* **46**, 69-85.

[10] Wooldridge (2005): “Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity”, *Journal of Applied Econometrics* **20**, 39-54.