

Least Squares Estimation of a Panel Data Model with Multifactor Error Structure and Endogenous Covariates*

Matthew Harding[†] and Carlos Lamarche[‡]

October 11, 2009

Abstract

We propose a method for estimating an interactive effects panel data model with endogenous loadings and factors, and endogenous regressors. We show that the slope parameter can be easily estimated by accommodating recent developments.

JEL: C33, C31, I21.

Keywords: Panel data; Instrumental variables; Interactive fixed effects.

1. Introduction

The use of panel data models has long been associated with the opportunity to control for unobserved individual heterogeneity. Recently, attempts have been made to relax the traditional assumption of unique time invariant individual effects using multiple interactive effects (Bai, 2009; Heckman and Navarro, 2007; Pesaran, 2006). The natural extension to the standard panel data model with N cross-sectional units and T time periods, $y_{it} = \beta' \mathbf{x}_{it} + \epsilon_{it}$, imposes a multi-factor error structure on the error term $\epsilon_{it} = \boldsymbol{\lambda}'_i \mathbf{F}_t + u_{it}$, where $\boldsymbol{\lambda}_i$ is an $r \times 1$ vector of factor loadings and \mathbf{F}_t corresponds to the r common factors, and where both $\boldsymbol{\lambda}_i$ and \mathbf{F}_t are unobserved. The classical individual effects model can be obtained by setting \mathbf{F}_t and r equal to one. Note however that the standard differencing methods are inconsistent in this setting since they will not remove the unobserved effects, although quasi-differencing methods such as that of Holtz-Eakin, Newey

*We are grateful to Cecilia Rouse for providing the MPCP data

[†]Corresponding author: Department of Economics, Stanford University, 579 Serra Mall, Stanford, CA 94305; Phone: (650) 723-4116; Fax: (650) 725-5702; Email: mch@stanford.edu

[‡]Department of Economics, University of Oklahoma, 729 Elm Avenue, Norman, OK 73019; Phone: (405) 325-5857; Email: lamarche@ou.edu

and Rosen (1988) may be feasible alternatives. Neither will attempts to introduce time trends be successful in this context.

Bai (2009) considers the estimation of the interactive effects model in large N and T panel data settings and treats both $\boldsymbol{\lambda}_i$ and \mathbf{F}_t as fixed-effects parameters to be estimated, thus allowing for potential correlation between the unobservable interactive effects and the regressors \mathbf{x}_{it} . The paper employs a concentrated least squares estimation which embeds the Principal Components (PCA) approach to the estimation of the unknown factors. The finite sample performance of this method is relatively poor however in micro-panels with small T values even though the method performs very well in large panels.

Pesaran (2006) estimates the model using cross-sectional averages of the data matrix to proxy for the unobserved effects. Once consistent estimates of the coefficients on the observed regressors have been obtained, the model can be transformed into a standard factor model in order to estimate the interactive effects. The method however relies on the assumption that only the latent factors \mathbf{F}_t are allowed to be correlated with the regressors and the method is inconsistent if this assumption is violated. Moreover, Bai (2009) and Pesaran (2006) assume that the covariates \mathbf{x} and the error term are u are stochastically independent.

The aim of this paper is to consider the estimation of the panel data model with interactive fixed effects when some of the regressors are endogenous and T is small. We show in this paper that the approach of Pesaran (2006) can be easily modified to accommodate instrumental variables estimation. Moreover, instrumental variables estimation is also shown to be an easy fix for the case where both factors and loadings are correlated with the regressors.

2. Model and Method

This paper considers the following model for $i = 1 \dots N$ and $t = 1 \dots T$,

$$(2.1) \quad y_{it} = \boldsymbol{\beta}'\mathbf{x}_{it} + \boldsymbol{\gamma}'\mathbf{d}_t + \epsilon_{it},$$

$$(2.2) \quad \epsilon_{it} = \boldsymbol{\lambda}_i'\mathbf{F}_t + u_{it},$$

$$(2.3) \quad \mathbf{x}_{it} = \boldsymbol{\Pi}'\mathbf{w}_{it} + \mathbf{G}'\mathbf{d}_t + \boldsymbol{\Lambda}_i'\mathbf{F}_t + \mathbf{A}'\boldsymbol{\lambda}_i + \mathbf{v}_{it}.$$

We relax Bai and Pesaran assumption by considering that some of the component of \mathbf{v} could be stochastically dependent on u . In this model \mathbf{x}_{it} and \mathbf{d}_t denote the individual and common observed regressors, while \mathbf{F}_t denotes the unobserved common regressors with individual loadings $\boldsymbol{\lambda}_i$. It is

assumed that we additionally observe a vector of instruments \mathbf{w}_{it} which satisfies the appropriate identification conditions. The dimensions of the model are such that it is conformal to a design with k_1 regressors \mathbf{x}_{it} , k_2 common time trends \mathbf{d}_t , r factors \mathbf{F}_t and $m > k_1$ instruments \mathbf{w}_{it} . In what follows we consider for simplicity the case of $k_1 = m$.

Define $\mathbf{z}_{it} = (y_{it}, \mathbf{x}'_{it})'$. Then,

$$(2.4) \quad \mathbf{z}_{it} = \mathbf{C}_1 \mathbf{w}_{it} + \mathbf{C}_2 \mathbf{d}_t + \mathbf{C}_{3i} \mathbf{F}_t + \mathbf{C}_4 \boldsymbol{\lambda}_i + \boldsymbol{\xi}_{it},$$

where $\boldsymbol{\xi}_{it} = (\boldsymbol{\beta}' v_{it} + u_{it}, \mathbf{v}'_{it})'$ and $\mathbf{C}_1 = ((\boldsymbol{\Pi} \boldsymbol{\beta})', \boldsymbol{\Pi}')'$, $\mathbf{C}_2 = ((\mathbf{G} \boldsymbol{\beta} + \boldsymbol{\gamma})', \mathbf{G}')'$, $\mathbf{C}_{3i} = ((\boldsymbol{\Lambda}_i \boldsymbol{\beta} + \boldsymbol{\lambda}_i)', \boldsymbol{\Lambda}'_i)'$ and $\mathbf{C}_4 = ((\mathbf{A} \boldsymbol{\beta})', \mathbf{A}')'$.

Now consider cross-sectional averages,

$$(2.5) \quad \bar{\mathbf{z}}_t = \mathbf{C}'_1 \bar{\mathbf{w}}_t + \mathbf{C}'_2 \mathbf{d}_t + \bar{\mathbf{C}}'_3 \mathbf{F}_t + \mathbf{C}'_4 \bar{\boldsymbol{\lambda}} + \bar{\boldsymbol{\xi}}_t.$$

Hence,

$$(2.6) \quad \mathbf{F}_t = (\bar{\mathbf{C}}_3 \bar{\mathbf{C}}'_3)^{-1} \bar{\mathbf{C}}_3 (\bar{\mathbf{z}}_t - \mathbf{C}'_1 \bar{\mathbf{w}}_t - \mathbf{C}'_2 \mathbf{d}_t - \mathbf{C}'_4 \bar{\boldsymbol{\lambda}} - \bar{\boldsymbol{\xi}}_t)$$

Under the regularity conditions of Pesaran (2006), as $N \rightarrow \infty$, for all t we have $\bar{\boldsymbol{\xi}}_t \rightarrow 0$ and $\bar{\mathbf{C}}_3 \rightarrow \mathbf{C}_3$ (constant). Hence, we can proxy for \mathbf{F}_t by $(\bar{\mathbf{w}}'_t, \mathbf{d}'_t, \bar{\mathbf{z}}'_{it}, \bar{\boldsymbol{\lambda}}')'$. Notice, however that the required proxy $\bar{\boldsymbol{\lambda}}$ is not observed and consistent estimation is not possible unless either $\mathbf{A} = \mathbf{0}$ or $\bar{\boldsymbol{\lambda}} = \mathbf{0}$. Furthermore, a second source of endogeneity is present as long as $E(u_{it} v_{it,k}) \neq 0$. The analysis indicates that if valid instruments \mathbf{w}_{it} are present performing an instrumental variable regression on equation 2.1 augmented by $(\bar{\mathbf{w}}'_t, \mathbf{d}'_t, \bar{\mathbf{z}}'_{it})'$ will lead to consistent estimates of $\boldsymbol{\beta}$. The instruments address both the endogeneity of the regressors \mathbf{x}_{it} due to the correlation between u_{it} and \mathbf{v}_{it} and the correlation of the factor loadings with the regressors.

We propose to estimate the parameter of interest $\boldsymbol{\beta}$ in equation 2.1 by $\hat{\boldsymbol{\beta}} = (\mathbf{W}' \bar{\mathbf{M}} \mathbf{X})^{-1} \mathbf{W}' \bar{\mathbf{M}} \mathbf{y}$, where $\mathbf{X} = [\mathbf{x}_{it} : \mathbf{d}_t : \bar{\mathbf{z}}_t]$ is a $NT \times (1 + 2k_1 + k_2)$ matrix, $\mathbf{W} = [\mathbf{w}_{it} : \mathbf{d}_t : \bar{\mathbf{z}}_t]$ is a $NT \times (1 + 2m + k_2)$ matrix and $\bar{\mathbf{M}} = \mathbf{I} - \bar{\mathbf{H}} (\bar{\mathbf{H}}' \bar{\mathbf{H}})^{-1} \bar{\mathbf{H}}'$ where $\bar{\mathbf{H}} = [\mathbf{d}_t : \bar{\mathbf{z}}_t : \bar{\mathbf{w}}_t]$. Inferential procedures could be implemented by accommodating the variance formulas introduced in Pesaran (2006) or by considering the bootstrap, as in Section 4. Once the coefficients on the observables have been estimated consistently each model can be transformed into a traditional factor model and the factors and factor loadings consistently estimated using maximum likelihood or PCA. If the number of factors is unknown, it can also be estimated consistently using the eigenvalue method of Harding and Nair (2009) even in the presence of factor dynamics.

3. Monte Carlo

We consider the dependent variable considering a design similar to Bai (2009) and Pesaran (2006):

$$\begin{aligned} y_{it} &= \beta_0 + \beta_1 x_{it} + \gamma d_t + \boldsymbol{\lambda}'_i \mathbf{F}_t + u_{it} \\ x_{it} &= \pi_0 + \pi_1 w_{it} + g d_t + l \boldsymbol{\nu}' \mathbf{F}_t + a \boldsymbol{\lambda}'_i \mathbf{F}_t + v_{it} \\ F_{jt} &= \rho_f F_{jt-1} + \eta_{jt} \\ \eta_{jt} &= \rho_\eta \eta_{jt-1} + e_{jt} \end{aligned}$$

for $j = \{1, 2\}, \dots, t = -49, \dots, 0, \dots, T$ in the last two equations. The random variables are $d_t \sim \mathcal{N}(0, 1)$, $\lambda_{i1}, \lambda_{i2} \sim \mathcal{N}(1, 0.2)$, $(u_{it}, v_{it})' \sim \mathcal{N}(0, \boldsymbol{\Omega})$, and e and w are Gaussian random variables. The parameters are assumed to be: $\beta_0 = \pi_0 = l = 2$, $\beta_1 = \gamma = \pi_1 = g = 1$, $\rho_f = 0.90$, $\rho_\eta = 0.25$, and $\Omega_{11} = \Omega_{22} = 1$. We consider three designs: (I) $a = 0$ and $\Omega_{12} = \Omega_{21} = 0$; (II) $a = 2$ and $\Omega_{12} = \Omega_{21} = 0$; (III) $a = 2$ and $\Omega_{12} = \Omega_{21} = 0.5$.

In Table 1 we compare the performance of several estimators for different samples sizes $N \in \{50, 100\}$ and $T \in \{5\}$. We consider the ordinary least squares estimator (OLS), the within estimator (FE), the two-stage least squares estimator (2SLS), the instrumental variables estimator applied to the within transformation (IVFE), the infeasible common correlated effect estimator (ICCE) which observes the factors \mathbf{F}_t , the common correlated effects estimator (CCE) which proxies for the unobserved factors using cross-sectional averages and the instrumental variables common correlated effects estimator (IVCCE) proposed in this paper, which applies instrumental variables estimation to the regression equation augmented by the cross-sectional averages of the observables.

The first MC design corresponds to a situation with interactive effects, where the regressors are exogenous and the factor loadings are not correlated with the regressors. As we would expect IV, IVFE, ICCE, CCE and IVCCE are all unbiased. Notice that both CCE and ICCE have better finite sample performance than IV.

The second MC design corresponds to a model where the regressors are exogenous but the factor loadings are correlated with the regressors. In this case we notice that CCE is biased since it does not fully proxy for the latent factors. Additional evidence for the model with $N = 50$ indicates that the bias is reduced from 16.6% to 7.1% if the equation is also augmented by the time averages of the regressors. IVCCE however automatically corrects this problem. This model could have been estimated using the concentrated least squares approach of Bai (2009), however due to the small T dimension common in micro-econometric applications, the IVCCE estimator has better mean

squared error properties than an approach involving PCA which produces large mean-squared errors (Online Appendix to Bai, 2009).

The final MC design allows for both endogenous regressors and factor loadings correlated with the regressors. In this case the CCE estimator is severely biased while the IVCCE estimator continues to perform very well. Notice that the ICCE estimator is also severely biased due to the endogenous regressors and shows that knowing the true factors is not as important as correcting for endogeneity when it is present.

| Statistics | N | T | Estimators | | | | | | |
|------------------------|-----|---|------------|--------|---------|---------|---------|---------|---------|
| | | | OLS | FE | 2SLS | IVFE | ICCE | CCE | IVCCE |
| Monte Carlo Design I | | | | | | | | | |
| Bias | 50 | 5 | 0.2555 | 0.3150 | -0.0127 | -0.0220 | -0.0006 | 0.0000 | -0.0018 |
| RMSE | 50 | 5 | 0.2806 | 0.3354 | 0.1280 | 0.1732 | 0.0313 | 0.0510 | 0.0776 |
| Bias | 100 | 5 | 0.2520 | 0.3143 | -0.0028 | -0.0057 | 0.0002 | -0.0021 | 0.0020 |
| RMSE | 100 | 5 | 0.2757 | 0.3329 | 0.0774 | 0.0889 | 0.0230 | 0.0379 | 0.0569 |
| Monte Carlo Design II | | | | | | | | | |
| Bias | 50 | 5 | 0.2274 | 0.2157 | 0.0203 | -0.0282 | -0.0006 | 0.1657 | -0.0055 |
| RMSE | 50 | 5 | 0.2312 | 0.2197 | 1.9788 | 0.6528 | 0.0313 | 0.1891 | 0.0805 |
| Bias | 100 | 5 | 0.2258 | 0.2159 | -0.0223 | -0.0357 | 0.0002 | 0.1614 | 0.0000 |
| RMSE | 100 | 5 | 0.2290 | 0.2190 | 0.3654 | 0.5040 | 0.0230 | 0.1840 | 0.0569 |
| Monte Carlo Design III | | | | | | | | | |
| Bias | 50 | 5 | 0.2950 | 0.2553 | 0.0355 | -0.0410 | 0.2567 | 0.3877 | -0.0107 |
| RMSE | 50 | 5 | 0.2974 | 0.2560 | 2.4101 | 0.7013 | 0.2621 | 0.3898 | 0.0882 |
| Bias | 100 | 5 | 0.2941 | 0.2553 | -0.0257 | -0.0433 | 0.2572 | 0.3872 | -0.0059 |
| RMSE | 100 | 5 | 0.2962 | 0.2557 | 0.2749 | 0.5518 | 0.2613 | 0.3888 | 0.0584 |

TABLE 1. *Bias and root MSE of panel data regression estimators. Results are based on 1000 replications.*

4. Empirical Application

We consider data from the Milwaukee Parental Choice program (MPCP), the same data as Rouse (1998). Consider a model with interactive fixed effects:

$$(4.1) \quad T_{it} = \delta c_{it} + \beta' \mathbf{x}_{it} + \boldsymbol{\lambda}'_i \mathbf{F}_t + u_{it}$$

where T measures educational attainment, c is actual attendance to choice school, \mathbf{x} is a vector of exogenous variables that includes grade level of the test, gender, income, application lotteries, indicators for years from application to the program, and a dummy variable for whether the test was imputed. The variable w is an indicator variable for whether the students was randomly selected to attend choice schools. We allow for possible correlation between c and $(\boldsymbol{\lambda}', \mathbf{F}', u)$, and by definition, $w \perp u$.

| Treatment | Estimators | | | | | |
|---------------------------|------------------|-------------------|------------------|------------------|------------------|------------------|
| Variable | OLS | FE | CCE | 2SLS | IVFE | IVCCE |
| Enrolled in choice school | 0.699 (1.175) | -0.772 (1.264) | 0.720 (1.176) | 3.530 (2.071) | 4.276 (2.265) | 3.585 (2.067) |

TABLE 2. *Estimates of the causal effect of choice schools on math test scores.*

Table 2 presents results. The OLS, FE, and CCE estimates are likely to be biased because the treatment variable is suspected to be endogenous. Additionally the standard within transformation does not get rid of the individual specific effect, and thus $E((x_{it} - \bar{x}_i)\lambda_i(F_t - \bar{F})) \neq 0$ if we consider for simplicity the case of one factor. 2SLS and IVFE might be biased if there are interactive fixed effects also. The instrument could be correlated with the multifactor error term: $E((w_{it} - \bar{w}_i)\lambda_i(F_t - \bar{F})) \neq 0$ in the IVFE case and $E(w_{it}\lambda_i F_t) \neq 0$ in the IV case. These conditions might be interpreted as suggesting that the selection to the program could be related to parents' motivation and time factors like the availability of seats in the schools. The IVCCE suggests that the Milwaukee vouchers benefited students on mathematics, gaining approximately 3.6 additional percentile points per year relative to the students that were not attending the choice schools. This gain is considerably smaller than the 4.3 percentile points gain suggested by the IVFE method.

5. Final Remarks

We show that even for small T , the slope parameter in the more general interactive effects model recently introduced in the literature (e.g., Pesaran 2006, Bai 2009) can be estimated with the same

ease that the standard additive effects panel data model. In a model with interactive effects, if valid instruments are present, performing an instrumental variable regression address the endogeneity of the regressors due to their correlation with the error term, the factors, and the factor loadings. Providing a detailed treatment for inference, the estimation of factors, and the extension to robust estimation appear as the critical next steps.

References

- BAI, J. (2009): “Panel Data Models with Interactive Fixed Effects,” *Econometrica*, 77(4), 1229–1279.
- HARDING, M., AND K. K. NAIR (2009): “Estimating the Number of Factors and Lags in High-Dimensional Dynamic Factor Models,” mimeo.
- HECKMAN, J. J., AND S. NAVARRO (2007): “Dynamic Discrete Choice and Dynamic Treatment Effects,” *Journal of Econometrics*, 136(2), 341–396.
- HOLTZ-EAKIN, D., W. NEWEY, AND H. ROSEN (1988): “Estimating Vector Autoregressions with Panel Data,” *Econometrica*, 56, 1371–1395.
- PESARAN, M. H. (2006): “Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure,” *Econometrica*, 74(4), 967–1012.
- ROUSE, C. (1998): “Private School Vouchers and Student Achievement: An Evaluation of the Milwaukee Parental Choice Program,” *Quarterly Journal of Economics*, pp. 553–602.