Tatra Mt. Math. Publ. **39** (2008), 183–191

TATRA MOUNTAINS
Mathematical Publications

ANALYSIS OF PANEL DATA MODELS WITH GROUPED OBSERVATIONS

Carlos Rivero — Teófilo Valdés

ABSTRACT. We present an iterative estimation procedure to estimate panel data models when some observations are missed or grouped with arbitrary classification intervals. The analysis is carried out from the perspective of panel data models, in which the error terms may follow an arbitrary distribution. We propose an easy-to-implement algorithm to estimate all of the model parameters and the asymptotic stochastic properties of the resulting estimate are investigated as the number of individuals and the number of time periods increase.

1. Introduction

Panel data sets have become increasingly available in many scientific areas. One reason for this is that with panel data models we can better explain the complexity of some real life processes, since they present major advantages when compared to cross-sectional data models (Hsiao, 2003).

Many of the available panel data sets have both non-grouped, grouped or missing data (Verbeek and Nijman, 1992, present several types of non-response that can occur in panel data sets). The existence of grouped or missing data disables the usual parameter estimators of covariance panel data models, since the exact values are not available. When the percentage of grouped or missing data is significant, to ignore this data or to assign particular values to it may yield undesirable biases of the parameter estimates or may reduce their efficiency (Little and Rubin, 1987; McLachlan and Krishnan, 1997).

2000 Mathematics Subject Classification: 62J10, 62N01, 62F12.
Keywords: panel data, grouped or missed data, mean-based imputations, asymptotic results, European Community Household Panel.
Running title: Panel data with grouped observations. Acknowledgements: This paper springs from research partially funded by Spanish Ministry of Education and Science under grant MTM2004-05776 and by EUROSTAT under contract number 9.242.010.

Grouped or missing data may distort inference and it is likely that this distortion is more severe in panel data than in cross-sectional data (Verbeek and Nijman, 1992). For example, the Spanish part of the European Community Household Panel (ECHP) has a non-response rate of 20.95 % from 1997 to 2001 and in the 15 countries contained in the ECHP the cumulative non-response rate may be even higher. Similar non-response rates appear in many other panel data sets (Kalton, Kasprzyk and McMillen, 1989).

In this paper we propose an iterative algorithm to estimate the parameters of covariance panel data models with grouped or missing data. This algorithm is based on conditional expectation imputations and it reduces the undesirable effect of the information loss.

Even though the proposed algorithm is similar to the EM type algorithm, it has some notable differences which make the implementation easier. The algorithm we propose consists of a first step that fills in the grouped data using conditional expectations given the available information. The second step updates the current estimate by Ordinary Least Square (OLS) projections. The EM algorithm adopts this form only when errors are normally distributed (Tanner, 1993; McLachlan and Krishnan, 1997). However, with non-normal errors the EM algorithm could involve awkward integrations and optimizations on several variables, since the number of parameters in a panel data model increases as the number of individuals or time periods increases. Notwithstanding this, since the procedure we propose is based on OLS estimates after single imputations, it maintains an easy-implementation form whatever distribution the errors have. Moreover, for the general class of symmetric and strongly unimodal error distribution the estimate that we propose satisfies good asymptotic stochastic properties under weak conditions.

In section 2 we introduce the panel data model and the notation used. In section 3 we describe the estimation algorithm and we state the main convergence theorems. In section 4 we present a simulation study which shows the good performance of the algorithm. Finally, some concluding remarks are mentioned in section 5.

2. Panel data model with grouped data

Let us consider the panel data model

$$y_{it} = \mu_i + x'_{it}\beta + u_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T, \tag{1}$$

where the exogenous variables $x'_{it} = (x_{1it}, \ldots, x_{kit})$ are observed on the individ-

ual i at time t, $\beta = (\beta_1, \ldots, \beta_k)'$ is an unknown vector parameter and μ_i is the unknown ith individual effect. We assume that the errors u_{it} are i.i.d. following a known mean-zero density function $f > 0$.

The model may be rewritten in matrix form as

$$Y = X\varphi + U, \tag{2}$$

where

$$\begin{aligned}
&\varphi = \big(\mu_1, \ldots, \mu_N,\ \beta_1, \ldots, \beta_k\big)', \\
&Y = (y_1, \ldots, y_N)', \quad y_i = (y_{i1}, \ldots, y_{iT})', \\
&U = (u_1, \ldots, u_N)', \quad u_i = (u_{i1}, \ldots, u_{iT})',
\end{aligned}$$

$$X = \begin{pmatrix} e & 0 & \cdots & 0 & X_1 \\ 0 & e & \cdots & 0 & X_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & e & X_N \end{pmatrix},$$

$$X_i = \begin{pmatrix} x_{1i1} & \cdots & x_{ki1} \\ \vdots & \ddots & \vdots \\ x_{1iT} & \cdots & x_{kiT} \end{pmatrix} = \begin{pmatrix} x'_{i1} \\ \vdots \\ x'_{iT} \end{pmatrix}$$

and $e = (1, \ldots, 1)'$. Let us define

$$\overline{x}_i = \frac{1}{T}\sum_{t=1}^{T} x_{it}\,, \qquad \overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it}\,.$$

It is known that β can be unbiased and consistently (as $NT \to \infty$) estimated from OLS by means of

$$\widehat{\beta} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \overline{x}_i)\,(x_{it} - \overline{x}_i)'\right]^{-1} \left[\sum_{i=1}^{N}\sum_{t=1}^{T} (x_{it} - \overline{x}_i)\,(y_{it} - \overline{y}_i)\right], \tag{3}$$

which only involves the inversion of a matrix of order $k \times k$.

For every fixed $i = 1, \ldots, N$, the parameter μ_i can be unbiased and consistently (as $T \to \infty$) estimated, by means of

$$\widehat{\mu}_i = \overline{y}_i - \overline{x}'_i\widehat{\beta}, \qquad i = 1, \ldots, N. \tag{4}$$

Throughout the rest of the paper, we will assume that some values of the dependent variable y_{it} are grouped following different criteria. This means that when a datum y_{it} is grouped, the exact value is missed, but we know a grouping interval which contains it, $y_{it} \in (l_{it}, d_{it}]$. Consequently, the usual estimation

of β and μ_i through (3) and (4) is impracticable. Whether or not we observe y_{it} depends on the missing data mechanism. In this paper, the missing data mechanism is *ignorable* (as defined in Little and Rubin, 1987), hence this mechanism and y_{it} are independent. The index set

$$I = \{it | i = 1, \ldots, N,\, t = 1, \ldots, T\}$$

can be partitioned in two sets:

$$I_g = \{it | y_{it} \text{ it is grouped}\} \quad \text{and} \quad I_o = \{it | y_{it} \text{ is observed}\}.$$

3. Estimation algorithm

We propose to estimate the true vector parameter $\varphi = (\mu_1, \ldots, \mu_N, \beta')'$ in the fixed effects panel data model (2), by means of an iterative algorithm based on least squares estimates and conditional expectation imputations.

Initialization. Fix an arbitrary vector φ^0 as the initial estimate of the true vector parameter φ. $p = 0$.

Step 1 (Conditional expectation imputations). For every index $i = 1, \ldots, N$ and $t = 1, \ldots, T$, evaluate an imputation which depends on the former estimate of the parameter, φ^p, in the following way

$$y_{it}(\varphi^p) = \begin{cases} y_{it} & it \in I_o \\ \widehat{y}_{it} & it \in I_g \end{cases},$$

where

$$\begin{aligned} \widehat{y}_{it} &= E_{\varphi^p}(y_{it} | y_{it} \in (l_{it}, d_{it}]) \\ &= \mu_i^p + x_{it}'\beta^p + \widehat{u}_{it} \end{aligned}$$

and

$$\widehat{u}_{it} = E\big(u | u \in (-\mu_i^p - x_{it}'\beta^p + l_{it}, -\mu_i^p - x_{it}'\beta^p + d_{it}]\big).$$

Step 2 (OLS estimate). Update the estimation of the parameter φ^p by means of the following expressions

$$\beta^{p+1} = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)'\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(x_{it} - \overline{x}_i)\left(y_{it}(\varphi^p) - \overline{y}_i(\varphi^p)\right)\right]$$

$$\mu_i^{p+1} = \overline{y}_i(\varphi^p) - \overline{x}_i'\beta^p, \qquad i = 1, \ldots, N,$$

where

$$\overline{y}_i(\varphi^p) = \frac{1}{T}\sum_{t=1}^{T} y_{it}(\varphi^p).$$

Step 3. $p \leftarrow p+1$ and return to Step 1, until the convergence is achieved, in accordance with a usual stop criterion.

This algorithm is applicable for arbitrary error distributions. In the Gaussian error distribution case the algorithm agrees with the EM algorithm (McLachlan and Krishnan, 1997; Lange, 1999). However, with non-Gaussian error distributions our algorithm differs from the EM algorithm and our resulting parameter estimates of model (2) do not agree with the maximum likelihood estimates. Notwithstanding this, the estimates that we propose satisfy good asymptotic stochastic properties under very weak conditions.

Note that the proposed algorithm only involves the inversion of a $k \times k$ matrix and elemental OLS calculations, although $\dim(\varphi) = N + k \to \infty$, as $N \to \infty$.

Assuming that the errors have a symmetric and strongly unimodal distribution, the following theorem shows that the sequences $\{\beta^p\}$ and $\{\mu_i^p\}$, generated by the algorithm, converge to a unique point (which does not depend on the initial points β^0 and μ_i^0), as $p \to \infty$.

Theorem 1. *Under weak assumptions regarding the regressors x_{it}, for any starting vector $\varphi^0 = \left(\mu_1^0, \ldots, \mu_N^0, \beta^{0\prime}\right)'$, the sequence $\{\varphi^p\}$ generated by the iteration of Steps 1 to 4 converges, as $p \to \infty$, to a unique vector $\widehat{\varphi} = \left(\widehat{\mu}_1, \ldots, \widehat{\mu}_N, \widehat{\beta}'\right)'$, which does not depend on the initial point φ^0. Furthermore, $\widehat{\varphi}$ is the unique vector which satisfies the implicit equation $\widehat{\varphi} = \left(X'X\right)^{-1} X'Y\left(\widehat{\varphi}\right)$, where $Y\left(\widehat{\varphi}\right) = \left(y_{11}\left(\widehat{\varphi}\right), \ldots, y_{NT}\left(\widehat{\varphi}\right)\right)'$.*

The point $\widehat{\varphi}$ is proposed to be the estimate of the true vector parameter φ, based on the sample of size NT. The following theorem states the limit behaviour of this estimate as $N \to \infty$ and/or $T \to \infty$.

Theorem 2. *Under the same assumptions of* Theorem 1 *if $\varphi = (\mu_1, \ldots, \mu_N, \beta')'$ denotes the true vector parameter, it holds that*

(i) *$\widehat{\beta} \to \beta$ almost surely and in L_2, as $NT \to \infty$.*

(ii) *For every $i = 1, \ldots, N$, $\widehat{\mu}_i \to \mu_i$ almost surely and in L_2, as $T \to \infty$.*

(iii) *There exists a $k \times k$ non-null covariance matrix Γ, such that*

$$\sqrt{NT}\left(\widehat{\beta} - \beta\right) \xrightarrow{D} N(0, \Gamma), \quad \text{as} \quad NT \to \infty.$$

(iv) *For every $i = 1, \ldots, N$, there exists a positive constant γ_i, such that*

$$\sqrt{T}\left(\widehat{\mu}_i - \mu_i\right) \xrightarrow{D} N(0, \gamma_i), \quad \text{as} \quad T \to \infty.$$

Furthermore, a consistent estimation of the asymptotic covariance matrix Γ *and the asymptotic variances* γ_i *can be proposed.*

4. Simulations and numerical results

We present some simulations that allow us to analyze the performance of the proposed algorithm presented in this paper. We consider the following covariance panel data model

$$y_{it} = \mu_i + x'_{it}\beta + u_{it}, \qquad i = 1, \ldots, N, \qquad t = 1, \ldots, T,$$

for which we have fixed the parameter $\beta = (5, -10)'$, μ_i $(i = 1, \ldots, N)$ are selected uniformly on $[-5, 5]$ and the independent variables x_{it} $(i = 1, \ldots, N,$ $t = 1, \ldots, T)$ are selected uniformly in the rectangle $[0, 1] \times [0, 1]$. We consider the errors u_{it} to be distributed as

(i) *Laplace*, with density function $f(u) = \frac{1}{2}\exp(-|u|)$,

(ii) Standard Normal,

(iii) *Logistic*, with density function $f(u) = e^{-u}(1 + e^{-u})^{-2}$.

The dependent variable y_{it} is grouped with probability 0.6, in which case the value y_{it} is classified in one of the intervals

$$(-\infty, -10], (-10, -5], (-5, 0], (0, 5] \quad \text{and} \quad (5, \infty).$$

For the purpose of showing the efficiency of the algorithm, 300 replications of the former model have been simulated, for sample sizes $N = 10, 50$ and $T = 10, 30$. The corresponding estimate of the parameter will be denoted by $\widehat{\beta}^{(j)}_{NT}$. We have empirically estimated the mean square error by

$$E\left|\widehat{\widehat{\beta}_{NT} - \beta}\right|^2 = \frac{1}{300}\sum_{j=1}^{300}\left|\widehat{\beta}^{(j)}_{NT} - \beta\right|^2. \tag{5}$$

These estimated MSE are shown in Tables 1 for Laplace, standard normal and logistic error distributions. The parameter estimation methods used are: (1) the algorithm proposed in this paper; (2) OLS discharging the grouped data (denoted by "ols"); (3) OLS with a single imputation of the grouped data (denoted by "OLS imputed"); (4) OLS using the complete non-grouped sample values (denoted by "OLS"). As Table 1 shows, the method proposed in the paper is advantageous with respect to the OLS estimate when the grouped data is simply discharged and not considered and, also, with respect to the OLS estimate imputing an arbitrary value when a datum is grouped.

The asymptotic normality of the estimates of Theorem 2, has been tested using the 300 replicated simulations. For $N = 50$ and $T = 30$, the normality of each component of $\widehat{\beta}_{NT}$ was accepted at the significance level of 0.10, using the usual Kolmogorov-Smirnov test. Also, the 300 replications have been used to empirically estimate the covariance matrix Γ of Theorem 2. The empirical estimate is

$$\Gamma_e = \begin{pmatrix} 30.33 & 0.53 \\ 0.53 & 26.21 \end{pmatrix}, \quad \Gamma_e = \begin{pmatrix} 26.19 & 0.13 \\ 0.13 & 18.76 \end{pmatrix}, \quad \Gamma_e = \begin{pmatrix} 40.30 & 0.83 \\ 0.83 & 35.00 \end{pmatrix}$$

when the errors follow Laplace, standard normal and logistic distribution, respectively. Using the consistent estimate mentioned in Theorem 2, based on a sample of size $N = 50$ and $T = 30$, we obtain the covariance matrices

$$\widehat{\Gamma} = \begin{pmatrix} 27.45 & 0.36 \\ 0.36 & 25.65 \end{pmatrix}, \ \widehat{\Gamma} = \begin{pmatrix} 21.09 & 0.23 \\ 0.23 & 20.01 \end{pmatrix} \text{ and } \widehat{\Gamma} = \begin{pmatrix} 36.22 & 1.06 \\ 1.06 & 33.77 \end{pmatrix}.$$

5. Concluding remarks

We have presented an algorithm to exploit the information of grouped data or even to handle missing data. The algorithm is simply based on two steps: conditional expectation imputations of the errors u_{it} and ordinary least squares estimates. Both steps notably reduce the computational requirements of other alternative methods. The proposed algorithm converges to a unique fixed point, which does not depend on the initial selected point. The limit point is taken as the estimate of the true vector parameter of the model. Under weak conditions, this estimate is consistent and asymptotically normal, centered on the real parameters.

The numerical results show the advantages of the proposed algorithm when compared with other alternative procedures. As the number of individuals increases, the alternative procedures (the EM algorithm among them) become more and more unwieldy, due to the increase in the number of individual parameters. On the contrary, this pernicious effect does not manifest itself in our algorithm, as shown in Table 1.

"Algorithm" refers to the estimate obtained when the algorithm proposed in Section 3 is used; "OLS" refers to the ordinary least squares estimate computed using the complete simulated data; "ols" refers to the ordinary least squares estimate discharging the grouped data; "OLS imputed" refers to the ordinary least squares estimate when the group data are imputed by an arbitrary value.

TABLE 1. Empirical mean square error of estimation of the parameter $\beta = (\beta_1, \beta_2)' = (5, -10)'$. Laplace, Standard Normal and Logistic error distribution.

	$N = 10$, $T = 10$	$N = 50$, $T = 10$
	Laplace/Normal/Logistic	Laplace/Normal/Logistic
Algorithm	0.88431/0.51990/1.22818	0.16237/0.10108/0.23915
OLS	0.60886/0.26049/0.99884	0.10170/0.05036/0.17031
ols	1.37971/0.86502/2.13619	0.24710/0.18949/0.39404
OLS imputed	1.18261/0.95775/1.67479	0.99436/0.60481/1.05784

	$N = 10$, $T = 30$	$N = 50$, $T = 30$
	Laplace/Normal/Logistic	Laplace/Normal/Logistic
Algorithm	0.23095/0.19732/0.33970	0.04332/0.03032/0.05516
OLS	0.18307/0.07863/0.29630	0.02552/0.01674/0.04282
ols	0.46889/0.33286/0.71995	0.06961/0.05105/0.11095
OLS imputed	0.36418/0.54596/0.48258	0.68546/0.50689/0.63929

REFERENCES

[1] HSIAO, C.: *Analysis of Panel Data*, University Press, Cambridge, 2003.

[2] KALTON, G.—KASPRZYK, D.—MCMILLEN, D. B.: *Nonsampling Errors in Panel Surveys*, Panel Surveys, (D. Kasprzyk, G. Duncan, G. Kalton, M. P. Singh, eds.), John Wiley, New York, 1989.

[3] LANGE, K.: *Numerical Analysis for Statisticians*, Statistics and Computing, Springer-Verlag, New York, 1999.

[4] LITTLE, R. J. A.—RUBIN, D. B.: *Statistical Analysis with Missing Data*, Wiley Series in Probab. and Math. Statist.: Applied Probability and Statistics, John Willey & Sons, New York, 1987.

[5] MCLACHLAN, G. J.—KRISHNAN, T.: *The EM Algorithm and Extensions*, Wiley Series in Probab. and Math. Statist., Applied Probability and Statistics, John Willey & Sons, New York, 1997.

[6] TANNER, M. A.: *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, Springer Series in Statist., Springer-Verlag, New York, 1993.

[7] VERBEEK, M.—NIJMAN, T.: *Incomplete Panels and Selection Bias*, The Econometrics of Panel Data: Handbooks of Theory and Applications, (L. Matyas, P. Sevestre, eds.), Kluwer Academic Publishers. Dordrecht, London, 1992.

Received September 25, 2006

Carlos Rivero
Departamento de Estadística e Investigación Operativa II
Facultad de Ciencias Económicas y Empresariales
Universidad Complutense de Madrid
28223 Madrid
SPAIN
E-mail: crivero@estad.ucm.es

Teófilo Valdés
Departamento de Estadística e Investigación Operativa I
Facultad de Ciencias Matemáticas
Universidad Complutense de Madrid
28040 Madrid
SPAIN
E-mail: tevaldes@mat.ucm.es