Estimating Nonlinear Models with Multiple Fixed Effects: A Computational Note*

Laura Hospido
Bank of Spain and IZA
laura.hospido@bde.es

This draft: February 2010
First draft: March 2006

Abstract

In this paper we consider estimation of nonlinear panel data models that include multiple individual fixed effects. Estimation of these models is complicated both by the difficulty of estimating models with possibly thousands of coefficients and also by the incidental parameters problem; that is, noisy estimates of the fixed effects when the time dimension is short contaminate the estimates of the common parameters due to the nonlinearity of the problem. We propose a bias corrected likelihood estimator which can exploit the additivity of the effects for numerical optimization, thereby avoiding the calculation of estimates of the effects for given values of the common parameters. We exhibit the performance of this new estimator in simulations.

JEL Code: C23.

Keywords: Panel data, non-linear models, multiple fixed effects, incidental parameters, bias reduction.

*This paper started in collaboration with Manuel Arellano, whose suggestions and encouragement are gratefully acknowledged. I would also like to thank Stéphane Bonhomme and seminar participants at the SAEe in Granada, the 15th International Conference on Panel Data in Bonn and the EEA-ESEM meetings in Barcelona, for useful comments. All errors are mine. The views in this paper are those of the author and do not represent the views of the Bank of Spain or the Eurosystem.

1 Introduction

In a typical nonlinear micropanel data model with fixed effects there are hundreds or thousands of individual coefficients to estimate together with a relatively small number of common parameters. A well known computational simplification in the linear model is to obtain first the maximum likelihood (ML) estimates of the common parameters from a regression on the data in deviations from individual means, and secondly retrieve ML estimates of the effects from averaged residuals one by one. A similar computational simplification is available for Newton-Raphson and related algorithms for nonlinear fixed effects models, which exploits the block-diagonal structure of the Hessian. This simplification has been discussed in Hall (1978), Chamberlain (1980), and Greene (2004) for nonlinear models with a scalar fixed effect. The first purpose of this work is to show how to use an iterated algorithm of this type in a nonlinear model with multiple fixed effects.

As first noted by Neyman and Scott (1948), when the time series dimension T is small relative to the cross-sectional dimension n, ML estimates of the common parameters can be severely biased, especially in dynamic models. This Incidental Parameters problem arises because the unobserved individual characteristics are replaced by noisy estimates, which bias estimates of model parameters. In particular, the bias of the MLE is of order $1/T$. In some special cases it is possible to obtain fixed T large n consistent estimators of certain common parameters, but these situations are more the exception than the rule. Alternatively, a number of additional approaches have been proposed to obtain approximately unbiased estimators as opposed to estimators with no bias at all[1]. One of these approaches consists of estimation from an analytically bias corrected objective function relative to some target criterion[2]. In this paper we discuss the application of computationally efficient algorithms to modified concentrated likelihoods of this type to obtain estimators without bias to order $1/T$ in nonlinear panel models with multiple fixed effects.

The paper is organized as follows. Section 2 introduces the model and notation. Section 3 explains how the iterated algorithm works. Section 4 discusses its application to bias corrected concentrated likelihoods. Section 5 presents some simulation results. Finally, Section 6 concludes. Detailed derivations are given in the Appendix.

[1] See Arellano and Hahn (2007) for a review of this literature on bias-adjusted estimation methods for nonlinear panel data models with fixed effects.

[2] See Pace and Salvan (2006) for adjustments of this type for a generic concentrated likelihood with independent observations, Arellano and Hahn (2007) for static nonlinear panel models and Arellano and Hahn (2006), Bester and Hansen (2009), and Hospido (2010), for the dynamic case. For an automatic way of correcting the bias of the concentrated likelihood see Dhaene and Jochmans (2009).

2 Model and Notation

Let us consider the following model for the joint density of T random vectors conditioned on initial observations, strictly exogenous variables, and fixed effects:

$$f\left(y_{i1},...,y_{iT}\mid y_{i0},x_{i1},...,x_{iT},\alpha_{i0}\right)=\prod_{t=1}^{T}f\left(y_{it}\mid y_{i(t-1)},x_{it},\alpha_{i0},\theta_0\right)$$

where θ_0 is a vector of common parameters and α_{i0} is a vector of fixed effects. We observe the random sample $\{y_{i0},...,y_{iT},x_{i0},...,x_{iT}\}_{i=1}^{n}$ and we denote $\alpha_0=(\alpha_{10}',...,\alpha_{n0}')'$ and $\delta_0=\left(\theta_0',\alpha_0'\right)'$. Let the log likelihood of one observation be

$$\ell_{it}\left(\theta,\alpha_i\right)=\ln f\left(y_{it}\mid y_{i(t-1)},x_{it},\alpha_i,\theta\right)$$

and let $\ell_i\left(\theta,\alpha_i\right)=\sum_{t=1}^{T}\ell_{it}\left(\theta,\alpha_i\right)$.

3 Efficient Newton-Raphson iteration

Let us consider the estimator

$$\begin{pmatrix}\widehat{\theta}\\ \widehat{\alpha}\end{pmatrix}=\arg\max_{\theta,\alpha}\sum_{i=1}^{n}\ell_i\left(\theta,\alpha_i\right)$$

and let first and second derivatives be denoted by

$$\begin{aligned}d_{\theta i}&=\frac{\partial\ell_i\left(\theta,\alpha_i\right)}{\partial\theta},\qquad d_{\alpha i}=\frac{\partial\ell_i\left(\theta,\alpha_i\right)}{\partial\alpha_i}\\ H_{\theta\theta i}&=\frac{\partial^2\ell_i\left(\theta,\alpha_i\right)}{\partial\theta\partial\theta'},\qquad H_{\alpha\alpha i}=\frac{\partial^2\ell_i\left(\theta,\alpha_i\right)}{\partial\alpha_i\partial\alpha_i'},\qquad H_{\theta\alpha i}=\frac{\partial^2\ell_i\left(\theta,\alpha_i\right)}{\partial\theta\partial\alpha_i'}.\end{aligned}$$

The Kth step of the iteration of a computationally efficient algorithm for obtaining $\widehat{\theta}$ and $\widehat{\alpha}$ takes the form

$$\theta_{[K]}-\theta_{[K-1]}=-\left[\sum_{i=1}^{n}\left(H_{\theta\theta i}-H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}H_{\alpha\theta i}\right)\right]^{-1}\sum_{i=1}^{n}\left(d_{\theta i}-H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}d_{\alpha i}\right)\tag{1}$$

$$\alpha_{i[K]}-\alpha_{i[K-1]}=-H_{\alpha\alpha i}^{-1}\left[d_{\alpha i}+H_{\alpha\theta i}\left(\theta_{[K]}-\theta_{[K-1]}\right)\right],\qquad(i=1,...,n)\tag{2}$$

where all derivatives are evaluated at $\theta_{[K-1]}$ and $\alpha_{i[K-1]}$.

This result can be easily proved using partitioned inverse formulae (a detailed derivation is in the Appendix A). It is a standard result in nonlinear estimation of models with many group effects.[3]

[3] An alternative Gauss–Newton algorithm which leads to a regression-based iteration is discussed in Appendix B.

4 Analytically Adjusted Concentrated Likelihood

When T is short we may be interested to consider an estimator that maximizes a bias corrected concentrated likelihood of the type reviewed in Arellano and Hahn (2007):

$$\widehat{\theta}^{AH} = \arg\max_{\theta} \sum_{i=1}^{n} \left[\ell_i\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right) + \beta_i\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right]$$

where

$$\widehat{\alpha}_i\left(\theta\right) = \arg\max_{\alpha} \ell_i\left(\theta, \alpha\right)$$

and $\beta_i\left(\theta, \alpha_i\right)$ is an adjustment term.

As long as the adjustment term depends on α, the iterated algorithm discussed above cannot be directly used for estimating $\widehat{\theta}^{AH}$. Note that

$$\begin{pmatrix} \widehat{\theta}^{AH} \\ \widehat{\alpha}^{AH} \end{pmatrix} = \arg\max_{\theta,\alpha} \sum_{i=1}^{n} \left[\ell_i\left(\theta, \alpha_i\right) + \beta_i\left(\theta, \widehat{\alpha}_i\left(\theta\right)\right)\right]$$

where $\widehat{\alpha}^{AH} = \widehat{\alpha}\left(\widehat{\theta}^{AH}\right)$. Thus, if we use the analysis of covariance algorithm discussed in the previous section we still need to calculate $\widehat{\alpha}_i\left(\theta\right)$ for given values of θ.

A Computationally Effective Estimator Alternatively, we can consider an estimator of the form

$$\begin{pmatrix} \widetilde{\theta}^{AH} \\ \widetilde{\alpha}^{AH} \end{pmatrix} = \arg\max_{\theta,\alpha} \sum_{i=1}^{n} \left[\ell_i\left(\theta, \alpha_i\right) + \beta_i\left(\theta, \alpha_i\right)\right]$$

for which the iterated algorithm can be used. This is equivalent to:

$$\widetilde{\theta}^{AH} = \arg\max_{\theta} \sum_{i=1}^{n} \left[\ell_i\left(\theta, \widetilde{\alpha}_i\left(\theta\right)\right) + \beta_i\left(\theta, \widetilde{\alpha}_i\left(\theta\right)\right)\right]$$

where

$$\widetilde{\alpha}_i\left(\theta\right) = \arg\max_{\alpha} \left[\ell_i\left(\theta, \alpha\right) + \beta_i\left(\theta, \alpha_i\right)\right].$$

The statistic $\widetilde{\alpha}_i\left(\theta\right)$ can be regarded as a Bayesian estimator that uses $e^{\beta_i\left(\theta,\alpha_i\right)}$ as the prior distribution of α_i for a given value of θ. Thus, under general conditions, $\widetilde{\alpha}_i\left(\theta\right)$ will be asymptotically equivalent to $\widehat{\alpha}_i\left(\theta\right)$, and $\widetilde{\theta}^{AH}$ will have similar (bias reducing) properties as $\widehat{\theta}^{AH}$ (see Severini, 1998, section 4, for a discussion on the use of adjusted concentrated likelihoods using alternative estimates of nuisance parameters).

It appears that $\widetilde{\theta}^{AH}$ is not only computationally convenient, but it may also exhibit improved finite sample properties in certain situations due to the replacement of $\widehat{\alpha}_i\left(\theta\right)$ by $\widetilde{\alpha}_i\left(\theta\right)$ (for instance, in regressions of individual effects estimates on strictly exogenous regressors).

Estimation of the Bias The form of the approximate bias (Arellano and Hahn, 2007) is

$$\beta_i(\theta) \approx \frac{1}{2} tr\left(H_i^{-1}(\theta,\alpha)\,\Upsilon_i(\theta,\alpha)\right), \tag{3}$$

where

$$H_i(\theta,\alpha) \equiv -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 \ell_{it}(\theta,\alpha)}{\partial\alpha\partial\alpha'},$$

$$\Upsilon_i(\theta,\alpha) \equiv \sum_{l=-m}^{m}\omega_{T,l}\Gamma_l(\theta,\alpha),$$

and

$$\Gamma_l(\theta,\alpha) \equiv \frac{1}{T}\sum_{t=\max(1,l+1)}^{\min(T,T+l)}\left[\frac{\partial \ell_{it}(\theta,\alpha)}{\partial\alpha_i}\cdot\frac{\partial \ell_{it-l}(\theta,\alpha)}{\partial\alpha_i'}\right].$$

The quantity m is a bandwidth parameter and $\omega_{T,l}$ denotes a weight that guarantees positive definiteness of $\Upsilon_i(\Gamma,\Theta_i)$.

Automatically Adjusted Concentrated Likelihood The half-panel split jackknife provides an automatic way of correcting the bias of the MLE (Dhaene and Jochmans, 2009). The bias corrected estimator is defined as

$$\widehat{\theta}^{DJ} = 2\widehat{\theta} - \overline{\theta}_{1/2},$$

where $\widehat{\theta}$ is the MLE from the full panel, and $\overline{\theta}_{1/2}$ is the average of the two half-panel MLEs, each using $T/2$ time periods and all n cross-sectional units.

5 Monte Carlo Study

We consider fixed-effect static and dynamic probit models to test the small-sample performance of the estimator. When T is small, the MLE is severely biased in these models. We present results for different models, keeping the simulation design as consistent as possible across them[4].

5.1 Data Generating Processes

Four probit models are considered:

$$y_{it} = \mathbf{1}\left[w_{it} + \epsilon_{it} \geq 0\right], \quad \text{and} \quad \epsilon_{it} \sim N(0,1),$$

- Static model with scalar fixed effects:

$$w_{it} = \alpha_{1i0} + \theta_{10}x_{it} + \theta_{20}d_{it}; \quad \left(\theta_0 = [\theta_{10},\theta_{20}]'; \alpha_{i0} = \alpha_{1i0}\right),$$

[4] Other studies, that consider nonlinear designs with scalar fixed effects (Carro, 2007, and Fernández-Val, 2009), show that the bias in the ML estimator is similar in magnitude for the logit and the probit models and that bias corrections also perform similarly. Here, we focus on probit designs and extend the analysis to consider multiple fixed effects.

- Static model with multiple fixed effects:

$$w_{it} = \alpha_{1i0} + \theta_{10}x_{it} + \alpha_{2i0}d_{it}; \quad \left(\theta_0 = \theta_{10}; \alpha_{i0} = [\alpha_{1i0}, \alpha_{2i0}]'\right),$$

- Dynamic model with scalar fixed effects:

$$w_{it} = \alpha_{1i0} + \theta_{10}x_{it} + \theta_{20}y_{it-1}; \quad \left(\theta_0 = [\theta_{10}, \theta_{20}]'; \alpha_{i0} = \alpha_{1i0}\right),$$

- Dynamic model with multiple fixed effects:

$$w_{it} = \alpha_{1i0} + \theta_{10}x_{it} + \alpha_{2i0}y_{it-1}; \quad \left(\theta_0 = \theta_{10}; \alpha_{i0} = [\alpha_{1i0}, \alpha_{2i0}]'\right).$$

The data were generated with $x_{it} \sim N(0,1)$, $d_{it} = \mathbf{1}[x_{it} + h_{it} > 0]$, and $h_{it} \sim N(0,1)$. For the dynamic designs, the data were generated with $y_{i0} = \mathbf{1}[\alpha_{1i0} + \theta_{10}x_{it} + \epsilon_{i0} > 0]$, and $x_{i0} \sim N(0,1)$, $\epsilon_{i0} \sim N(0,1)$. We set $n = 100$; $T = \{6, 8, 12, 20\}$; $\alpha_{1i0} = 0$, $\forall i$; $\theta_{10} = 1$, and $\theta_{20} = 0.5$ or $\alpha_{2i0} = 0.5$, $\forall i$; and ran 100 Monte Carlo replications for each design, with just ϵ_{it} redrawn in each replication.

Table 1. DGP summary

$y_{it} = \mathbf{1}[\alpha_{1i0} + \theta_{10}x_{it} + r_{it} + \epsilon_{it} > 0]$; $\epsilon_{it} \sim N(0,1)$			
T	$\{6, 8, 12, 20\}$	α_{1i0}	$0, \forall i$
n	100	θ_{10}	1
rep	100	x_{it}	$N(0,1)$
Static Scalar: $r_{it} = \theta_{20}d_{it}$		Dynamic Scalar: $r_{it} = \theta_{20}y_{it-1}$	
θ_{20}	0.5	θ_{20}	0.5
d_{it}	$\mathbf{1}[x_{it} + h_{it} > 0]$	y_{i0}	$\mathbf{1}[\alpha_{1i0} + \theta_{10}x_{i0} + \epsilon_{i0} > 0]$
h_{it}	$N(0,1)$	m	1
Static Multiple: $r_{it} = \alpha_{2i0}d_{it}$		Dynamic Multiple: $r_{it} = \alpha_{2i0}y_{it-1}$	
α_{2i0}	$0.5, \forall i$	α_{2i0}	$0.5, \forall i$
d_{it}	$\mathbf{1}[x_{it} + h_{it} > 0]$	y_{i0}	$\mathbf{1}[\alpha_{1i0} + \theta_{10}x_{i0} + \epsilon_{i0} > 0]$
h_{it}	$N(0,1)$	m	1

5.2 Simulation results

We estimate the common parameter θ_0 by maximum likelihood, MLE; applying the analytically bias-corrected estimator of Arellano and Hahn (2006, 2007), AH; and the automatically bias-corrected estimator of Dhaene and Jochmans (2009), DJ; both using the usual NR algorithm and the Efficient NR algorithm.

Tables 2 to 5 report the effective computation time (in seconds) for each design, along with the median absolute errors and root mean squared errors. Failure refers to the percentage of cases of divergence or failure to converge in the nonlinear solution over the 100 Monte Carlo replications.

Tables 2 and 3 report the results corresponding to the DGP with scalar fixed effects and $\theta = (\theta_1, \theta_2)'$.

Table 2. Static Probit with Scalar Fixed Effects

			NR			ENR			Time
			MAE	RMSE	Time	MAE	RMSE	Time	NR/ENR
T=6	θ_1	MLE	0.300	0.325	5.00	0.300	0.325	0.57	8.77
		AH	0.162	0.201	7.95	0.162	0.200	0.97	8.19
		DJ	0.223	0.309	7.76	0.215	0.289	1.10	7.05
	θ_2	MLE	0.153	0.248		0.158	0.249		
		AH	0.116	0.194		0.122	0.195		
		DJ	0.168	0.229		0.171	0.247		
T=12	θ_1	MLE	0.116	0.143	4.14	0.116	0.155	0.74	5.59
		AH	0.057	0.079	7.09	0.050	0.091	1.24	5.72
		DJ	0.079	0.113	6.89	0.092	0.110	1.50	4.59
	θ_2	MLE	0.090	0.124		0.096	0.120		
		AH	0.067	0.102		0.081	0.100		
		DJ	0.075	0.119		0.085	0.112		
T=20	θ_1	MLE	0.054	0.083	5.66	0.056	0.086	1.02	5.55
		AH	0.032	0.052	10.00	0.038	0.054	1.77	5.65
		DJ	0.052	0.063	9.09	0.045	0.055	1.94	4.68
	θ_2	MLE	0.062	0.082		0.059	0.076		
		AH	0.046	0.073		0.049	0.073		
		DJ	0.068	0.096		0.055	0.081		

Note: MAE=median absolute error, RMSE = root mean squared error.

In the static probit, the MLE of both θ_1 and θ_2 are seriously biased even for $T = 12$. After applying the corrections, the estimates are closer to the true value of the parameters, especially for the AH estimator. In addition, we can see that the ENR algorithm provides a significant computational time improvement with respect to the NR algorithm (from 4.59 to 8.77 times faster).

Table 3. Dynamic Probit with Scalar Fixed Effects

			NR			ENR			Time
			MAE	RMSE	Time	MAE	RMSE	Time	NR/ENR
T=6	θ_1	MLE	0.245	0.287	2.29	0.246	0.286	0.46	4.98
		AH	0.168	0.215	3.86	0.167	0.214	0.95	4.06
		DJ	0.322	0.485	3.17	0.308	0.463	0.75	4.20
	θ_2	MLE	0.455	0.477		0.457	0.479		
		AH	0.210	0.252		0.211	0.253		
		DJ	0.178	0.293		0.211	0.310		
T=12	θ_1	MLE	0.101	0.122	3.93	0.100	0.122	0.71	5.53
		AH	0.049	0.075	6.49	0.049	0.075	1.34	4.84
		DJ	0.090	0.109	6.33	0.083	0.107	1.38	4.59
	θ_2	MLE	0.205	0.230		0.209	0.232		
		AH	0.081	0.119		0.083	0.120		
		DJ	0.093	0.135		0.076	0.119		
T=20	θ_1	MLE	0.060	0.078	5.35	0.060	0.076	1.02	5.24
		AH	0.029	0.047	9.10	0.029	0.047	1.87	4.87
		DJ	0.037	0.053	8.85	0.035	0.052	1.99	4.44
	θ_2	MLE	0.120	0.128		0.124	0.131		
		AH	0.051	0.071		0.051	0.073		
		DJ	0.061	0.087		0.048	0.074		

Note: MAE=median absolute error, RMSE = root mean squared error.

In the dynamic probit, the MLE of θ_2 is more heavily biased than the one of θ_1. Once again, after applying the corrections, the estimates are closer to the true value of the parameters, but now *AH* estimator does not always dominate *DJ*. Also, we can see that the ENR algorithm still provides a substantial improvement in terms of computational time (from 4.06 to 5.53 times faster).

Tables 4 and 5 report the results corresponding to the DGP with multiple fixed effects and $\theta = \theta_1$.

Table 4. Static Probit with Multiple Fixed Effects

			NR			ENR			Time
			MAE	RMSE	Time	MAE	RMSE	Time	NR/ENR
T=6	θ_1	MLE	0.532	0.619	15.15	0.535	0.620	0.54	28.06
		AH	0.386	0.481	49.15	0.390	0.485	1.83	26.85
		DJ	0.430	0.814	17.62	0.360	0.573	0.64	27.53
	Failure (%)			0			1		
T=8	θ_1	MLE	0.377	0.440	39.05	0.375	0.439	0.80	48.81
		AH	0.250	0.312	110.47	0.249	0.312	2.54	43.49
		DJ	0.164	0.268	50.02	0.155	0.246	1.05	47.64
	Failure (%)			0			0		
T=12	θ_1	MLE	0.249	0.276	29.05	0.249	0.275	1.20	24.21
		AH	0.146	0.178	83.39	0.146	0.178	3.68	22.66
		DJ	0.138	0.168	40.36	0.082	0.122	1.73	23.33
	Failure (%)			0			0		
T=20	θ_1	MLE	0.127	0.154	43.02	0.127	0.154	1.84	23.38
		AH	0.060	0.094	122.95	0.059	0.094	5.62	21.88
		DJ	0.083	0.108	60.6	0.050	0.072	2.88	21.04
	Failure (%)			0			0		

Note: MAE=median absolute error, RMSE = root mean squared error.

Table 5. Dynamic Probit with Multiple Fixed Effects

			NR			ENR			Time
			MAE	RMSE	Time	MAE	RMSE	Time	NR/ENR
T=6	θ_1	MLE	0.570	0.633	19.67	0.568	0.629	0.64	30.73
		AH	0.509	0.569	85.40	0.502	0.569	2.83	30.18
		DJ	0.338	0.583	24.11	0.308	0.530	0.83	29.05
	Failure (%)			0			1		
T=8	θ_1	MLE	0.452	0.490	37.08	0.442	0.471	0.85	43.62
		AH	0.362	0.399	130.56	0.349	0.381	4.05	32.24
		DJ	0.213	0.288	49.75	0.146	0.246	1.21	41.11
	Failure (%)			0			0		
T=12	θ_1	MLE	0.270	0.295	34.27	0.261	0.292	1.22	28.09
		AH	0.186	0.211	114.59	0.178	0.209	4.38	26.16
		DJ	0.115	0.155	47.84	0.076	0.115	1.84	26.00
	Failure (%)			0			0		
T=20	θ_1	MLE	0.133	0.145	38.05	0.133	0.145	1.73	21.99
		AH	0.075	0.093	129.02	0.074	0.092	5.84	22.09
		DJ	0.053	0.077	55.00	0.034	0.058	2.81	19.57
	Failure (%)			0			0		

Note: MAE=median absolute error, RMSE = root mean squared error.

As expected, with multiple fixed effects the incidental parameter problem gets worse, both for the static and the dynamic probit. Now, the MAE of the MLE is sizable even for values of T such as 12 or 20. Again, the bias-corrected estimators can remove a substantial part of that bias. Interestingly, in this case, the improvements in terms of computational time are very large. These results are encouraging because, in many empirical studies that consider complicated models, the goal is not only to obtain an estimator with a good finite sample performance, but also in a reasonable computing time, especially when bootstrap methods are used for inference.

6 Conclusions

In this paper we consider estimation of nonlinear panel data models that include multiple individual fixed effects. Estimation of these models is complicated both by the difficulty of estimating models with possibly thousands of coefficients and also by the incidental parameters problem; that is, noisy estimates of the fixed effects when the time dimension is short contaminates the estimates of the common parameters due to the nonlinearity of the problem. We show how to use an iterated algorithm which simplifies estimation in a nonlinear model with multiple fixed effects and we also discuss its application to bias corrected concentrated likelihoods.

Simulations show that the estimator proposed is not only computationally convenient but it is also as good as others in a variety of probit designs. Different adjustments of the likelihood function result in bias corrected estimators that perform comparably to other bias corrections proposed in the literature. We can think in many microeconometric applications that use nonlinear panel data models. The results of the paper suggest that bias corrected estimates will be very useful in relevant empirical settings given the sample sizes of the panels more often used by researchers and, moreover, because they allow us to introduce more individual heterogeneity to address endogeneity concerns in a robust way.

A Newton-Raphson iteration

The Kth step of the Newton-Raphson iteration takes the form

$$\delta_K = \delta_{K-1} - \left(\frac{\partial^2 L\left(\delta_{K-1}\right)}{\partial \delta \partial \delta'} \right)^{-1} \frac{\partial L\left(\delta_{K-1}\right)}{\partial \delta},$$

or for shortness

$$\Delta \delta = - \left(\frac{\partial^2 L}{\partial \delta \partial \delta'} \right)^{-1} \frac{\partial L}{\partial \delta}$$

where $L\left(\delta\right) = \sum_{i=1}^n \ell_i \left(\theta, \alpha_i\right)$ and

$$\frac{\partial L}{\partial \delta} = \begin{pmatrix} \frac{\partial L}{\partial \theta} \\ \frac{\partial L}{\partial \alpha_1} \\ \vdots \\ \frac{\partial L}{\partial \alpha_n} \end{pmatrix} = \sum_{t=1}^T \begin{pmatrix} \sum_{i=1}^n \frac{\partial \ell_{it}(\theta,\alpha_i)}{\partial \theta} \\ \frac{\partial \ell_{1t}(\theta,\alpha_1)}{\partial \alpha_1} \\ \vdots \\ \frac{\partial \ell_{nt}(\theta,\alpha_n)}{\partial \alpha_n} \end{pmatrix} = \begin{pmatrix} d_\theta \\ d_\alpha \end{pmatrix}$$

$$\frac{\partial^2 L}{\partial \delta \partial \delta'} = \sum_{t=1}^T \begin{pmatrix} \sum_{i=1}^n \frac{\partial^2 \ell_{it}(\theta,\alpha_i)}{\partial \theta \partial \theta'} & \frac{\partial^2 \ell_{1t}(\theta,\alpha_1)}{\partial \theta \partial \alpha_1'} & \dots & \frac{\partial^2 \ell_{nt}(\theta,\alpha_n)}{\partial \theta \partial \alpha_n'} \\ \frac{\partial^2 \ell_{1t}(\theta,\alpha_1)}{\partial \alpha_1 \partial \theta'} & \frac{\partial^2 \ell_{1t}(\theta,\alpha_1)}{\partial \alpha_1 \partial \alpha_1'} & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell_{nt}(\theta,\alpha_n)}{\partial \alpha_n \partial \theta'} & 0 & \dots & \frac{\partial^2 \ell_{nt}(\theta,\alpha_n)}{\partial \alpha_n \partial \alpha_n'} \end{pmatrix} = \begin{pmatrix} H_{\theta\theta} & H_{\theta\alpha} \\ H_{\theta\alpha}' & H_{\alpha\alpha} \end{pmatrix}$$

and

$$d_\alpha = \begin{pmatrix} d_{\alpha 1} \\ \vdots \\ d_{\alpha n} \end{pmatrix}, \quad H_{\alpha\alpha} = \begin{pmatrix} H_{\alpha\alpha 1} & & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & H_{\alpha\alpha n} \end{pmatrix}, \quad H_{\theta\alpha} = \begin{pmatrix} H_{\theta\alpha 1} & \dots & H_{\theta\alpha n} \end{pmatrix}$$

so that $d_\theta = \sum_{i=1}^n d_{\theta i}$ and $H_{\theta\theta} = \sum_{i=1}^n H_{\theta\theta i}$, and

$$H_{\theta\alpha} H_{\alpha\alpha}^{-1} = \begin{pmatrix} H_{\theta\alpha 1} H_{\alpha\alpha 1}^{-1} & \dots & H_{\theta\alpha n} H_{\alpha\alpha n}^{-1} \end{pmatrix}$$

$$H_{\theta\alpha} H_{\alpha\alpha}^{-1} H_{\alpha\theta} = \sum_{i=1}^n H_{\theta\alpha i} H_{\alpha\alpha i}^{-1} H_{\alpha\theta i}$$

Letting

$$\begin{pmatrix} H_{\theta\theta} & H_{\theta\alpha} \\ H_{\theta\alpha}' & H_{\alpha\alpha} \end{pmatrix}^{-1} = \begin{pmatrix} H^{\theta\theta} & H^{\theta\alpha} \\ H^{\theta\alpha\prime} & H^{\alpha\alpha} \end{pmatrix}$$

where

$$\begin{aligned} H^{\theta\theta} &= \left(H_{\theta\theta} - H_{\theta\alpha} H_{\alpha\alpha}^{-1} H_{\alpha\theta}\right)^{-1} \\ H^{\theta\alpha} &= -H^{\theta\theta} H_{\theta\alpha} H_{\alpha\alpha}^{-1} \\ H^{\alpha\alpha} &= H_{\alpha\alpha}^{-1} + H_{\alpha\alpha}^{-1} H_{\alpha\theta} H^{\theta\theta} H_{\theta\alpha} H_{\alpha\alpha}^{-1} \end{aligned}$$

the partitioned formula gives:

$$\begin{pmatrix} \Delta\theta \\ \Delta\alpha \end{pmatrix} = - \begin{pmatrix} H^{\theta\theta} & H^{\theta\alpha} \\ H^{\theta\alpha\prime} & H^{\alpha\alpha} \end{pmatrix} \begin{pmatrix} d_\theta \\ d_\alpha \end{pmatrix}$$

or

$$\begin{aligned}\Delta\theta &= -\left(H^{\theta\theta}d_\theta + H^{\theta\alpha}d_\alpha\right)\\ \Delta\alpha &= -\left(H^{\theta\alpha\prime}d_\theta + H^{\alpha\alpha}d_\alpha\right).\end{aligned}$$

We have

$$\begin{aligned}H^{\theta\theta} &= \left[\sum_{i=1}^{n}\left(H_{\theta\theta i} - H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}H_{\alpha\theta i}\right)\right]^{-1}\\ H^{\theta\alpha}d_\alpha &= -H^{\theta\theta}H_{\theta\alpha}H_{\alpha\alpha}^{-1}d_\alpha = -H^{\theta\theta}\sum_{i=1}^{n}H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}d_{\alpha i}\end{aligned}$$

and

$$\begin{aligned}-\Delta\theta &= H^{\theta\theta}d_\theta + H^{\theta\alpha}d_\alpha = H^{\theta\theta}d_\theta - H^{\theta\theta}\sum_{i=1}^{n}H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}d_{\alpha i}\\ &= H^{\theta\theta}\left(d_\theta - \sum_{i=1}^{n}H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}d_{\alpha i}\right)\end{aligned}$$

so that

$$\Delta\theta = -\left[\sum_{i=1}^{n}\left(H_{\theta\theta i} - H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}H_{\alpha\theta i}\right)\right]^{-1}\sum_{i=1}^{n}\left(d_{\theta i} - H_{\theta\alpha i}H_{\alpha\alpha i}^{-1}d_{\alpha i}\right).$$

Similarly, we have[5]

$$\Delta\alpha = -H_{\alpha\alpha}^{-1}\left(d_\alpha + H_{\alpha\theta}\Delta\theta\right)$$

so that

$$\Delta\alpha_i = -H_{\alpha\alpha i}^{-1}\left(d_{\alpha i} + H_{\alpha\theta i}\Delta\theta\right), \qquad (i = 1, ..., n)$$

[5] Note that

$$\begin{aligned}-\Delta\alpha &= H^{\theta\alpha\prime}d_\theta + H^{\alpha\alpha}d_\alpha\\ &= -H_{\alpha\alpha}^{-1}H_{\alpha\theta}H^{\theta\theta}d_\theta + \left(H_{\alpha\alpha}^{-1} + H_{\alpha\alpha}^{-1}H_{\alpha\theta}H^{\theta\theta}H_{\theta\alpha}H_{\alpha\alpha}^{-1}\right)d_\alpha\\ &= H_{\alpha\alpha}^{-1}\left(d_\alpha + H_{\alpha\theta}H^{\theta\theta}H_{\theta\alpha}H_{\alpha\alpha}^{-1}d_\alpha - H_{\alpha\theta}H^{\theta\theta}d_\theta\right)\\ &= H_{\alpha\alpha}^{-1}\left(d_\alpha - H_{\alpha\theta}H^{\theta\alpha}d_\alpha - H_{\alpha\theta}H^{\theta\theta}d_\theta\right)\\ &= H_{\alpha\alpha}^{-1}\left[d_\alpha - H_{\alpha\theta}\left(H^{\theta\alpha}d_\alpha + H^{\theta\theta}d_\theta\right)\right]\\ &= H_{\alpha\alpha}^{-1}\left(d_\alpha + H_{\alpha\theta}\Delta\theta\right).\end{aligned}$$

B A regression-based iteration

Alternatively, we may consider a Gauss–Newton approach after enforcing block diagonality. The motivation is the same as in Berndt, Hall, Hall, and Hausman (1974) in that the nonzero components of the Hessian are approximated by outer product terms. The advantages of this procedure are that it only requires first derivatives and that it leads to a regression-based iteration.

Let us introduce the notation

$$\begin{aligned} d_{\theta it} &= \frac{\partial \ell_{it}(\theta,\alpha_i)}{\partial\theta}, \qquad d_{\alpha it} = \frac{\partial \ell_{it}(\theta,\alpha_i)}{\partial\alpha_i} \\ \Psi_{\theta\theta i} &= \sum_{t=1}^{T} d_{\theta it} d'_{\theta it}, \qquad \Psi_{\alpha\alpha i} = \sum_{t=1}^{T} d_{\alpha it} d'_{\alpha it}, \qquad \Psi_{\theta\alpha i} = \sum_{t=1}^{T} d_{\theta it} d'_{\alpha it}. \end{aligned}$$

The Kth step of the iteration of the Gauss–Newton algorithm for obtaining $\widehat{\theta}$ and $\widehat{\alpha}$ takes the form

$$\theta_{[K]} - \theta_{[K-1]} = -\left[\sum_{i=1}^{n}\left(\Psi_{\theta\theta i} - \Psi_{\theta\alpha i}\Psi_{\alpha\alpha i}^{-1}\Psi_{\alpha\theta i}\right)\right]^{-1}\sum_{i=1}^{n}\left(d_{\theta i} - \Psi_{\theta\alpha i}\Psi_{\alpha\alpha i}^{-1}d_{\alpha i}\right)$$

$$\alpha_{i[K]} - \alpha_{i[K-1]} = -\Psi_{\alpha\alpha i}^{-1}\left[d_{\alpha i} + \Psi_{\alpha\theta i}\left(\theta_{[K]} - \theta_{[K-1]}\right)\right], \qquad (i = 1, ..., n)$$

where all derivatives are evaluated at $\theta_{[K-1]}$ and $\alpha_{i[K-1]}$.

Thus,

$$\theta_{[K]} - \theta_{[K-1]} = -\left(\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{d}_{\theta it}\widetilde{d}'_{\theta it}\right)^{-1}\sum_{i=1}^{n}\sum_{t=1}^{T}\widetilde{d}_{\theta it}$$

where

$$\widetilde{d}_{\theta it} = d_{\theta it} - \widetilde{\Pi}_i d_{\alpha it}$$

and

$$\widetilde{\Pi}_i = \Psi_{\theta\alpha i}\Psi_{\alpha\alpha i}^{-1},$$

so that the $\widetilde{d}_{\theta it}$ are the residuals of individual-specific regressions of $d_{\theta it}$ on $d_{\alpha it}$. Next, $\theta_{[K]} - \theta_{[K-1]}$ can be calculated as a pooled regression of minus one on $\widetilde{d}_{\theta it}$. Finally, $\alpha_{i[K]} - \alpha_{i[K-1]}$ can be obtained as a regression of $-\left[1 + d'_{\theta it}\left(\theta_{[K]} - \theta_{[K-1]}\right)\right]$ on $d_{\alpha it}$:

$$\alpha_{i[K]} - \alpha_{i[K-1]} = -\left(\sum_{t=1}^{T} d_{\alpha it}d'_{\alpha it}\right)^{-1}\sum_{t=1}^{T} d_{\alpha it}\left[1 + d'_{\theta it}\left(\theta_{[K]} - \theta_{[K-1]}\right)\right], \qquad (i = 1, ..., n).$$

References

[1] Arellano, M. and J. Hahn (2006): "A likelihood-based approximate solution to the incidental parameter problem in dynamic nonlinear models with multiple effects", unpublished manuscript.

[2] Arellano, M. and J. Hahn (2007): "Understanding Bias in Nonlinear Panel Models: Some Recent Developments". In: R. Blundell, W. Newey, and T. Persson (eds.): Advances in Economics and Econometrics, Ninth World Congress, Volume III, Cambridge University Press, 381-409.

[3] Berndt, E. K., B. H. Hall, R. E. Hall, and J. A. Hausman (1974): "Estimation and Inference in Nonlinear Structural Models", *Annals of Economic and Social Measurement*, 3/4, 653–665.

[4] Bester, A. and C. Hansen (2009): "A Penalty Function Approach to Bias Reduction in Nonlinear Panel Models with Fixed Effects", *Journal of Business and Economic Statistics*, 27 (2), 131-148.

[5] Carro, J. (2007): "Estimating Dynamic Panel Data Discrete Choice Models with Fixed Effects", *Journal of Econometrics*, 140, 503-528.

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

[7] Dhaene, G., and K. Jochmans (2009): "Split-panel jackknife estimation of fixed effects models", unpublished manuscript.

[8] Fernández-Val, I. (2009): "Fixed effects estimation of structural parameters and marginal effects in panel probit models", *Journal of Econometrics,* 150, 71-85.

[9] Greene, W. (2004): "The Behaviour of the Maximum Likelihood Estimator of Limited Dependent Variable Models in the Presence of Fixed Effects", *The Econometrics Journal*, 7, 98–119.

[10] Hall, B. H. (1978): "A General Framework for Time Series–Cross Section Estimation", *Annales de l'INSEE*, 30–31, 177–202.

[11] Hospido, L. (2010): "Modelling Heterogeneity and Dynamics in the Volatility of Individual Wages", IZA Discussion Paper 4712.

[12] Neyman, J. and E. L. Scott (1948): "Consistent Estimates Based on Partially Consistent Observations", *Econometrica*, 16, 1-32.

[13] Pace, L. and A. Salvan (2006): "Adjustments of the profile likelihood from a new perspective", *Journal of Statistical Planning and Inference*, 136, 3554-3564.

[14] Severini, T. A. (1998): "Likelihood Functions for Inference in the Presence of a Nuisance Parameter", *Biometrika,* 85, 507–522.