APPROXIMATING THE BIAS OF THE LSDV ESTIMATOR FOR DYNAMIC UNBALANCED PANEL DATA MODELS

Giovanni S.F. Bruno

EEA   2004 - 1

Approximating the Bias of the LSDV Estimator for Dynamic Unbalanced Panel Data Models

Giovanni S.F. Bruno[*]
Università Bocconi
Istituto di Economia Politica
Via U. Gobbi, 5, 20136 Milano

May 3, 2004

Abstract

This paper extends the LSDV bias approximations in Bun and Kiviet (2003) to unbalanced panels. The approximations are obtained by modifying the within operator to accommodate the dynamic selection rule. They are accurate, with higher order terms bringing only decreasing improvements. This removes an important cause for limited applicability of bias corrected LSDV estimators.

Keywords: Bias approximation; Unbalanced panels; Dynamic panel data; LSDV estimator; Monte Carlo experiment

JEL classification: C23

---

[*]Tel.: +39 02 5836 5411; Fax: +39 02 5836 5438; e-mail: giovanni.bruno@uni-bocconi.it

1. Introduction

It is well known that the LSDV estimator for dynamic panel data models is not consistent for N large and finite T. Nickell (1981) derives an expression for the inconsistency for $N \to \infty$, which is $O\left(T^{-1}\right)$. Kiviet (1995) uses asymptotic expansion techniques to approximate the small sample bias of the LSDV estimator to also include terms of at most order $N^{-1}T^{-1}$, so offering a method to correct the LSDV estimator for samples where N is small or only moderately large. In Kiviet (1999) the bias approximation is more accurate, including also terms of at most order $N^{-1}T^{-2}$. Bun and Kiviet (2003) analyze the accuracy of Kiviet's (1999) approximation using simpler formulas.

Monte Carlo evidence in Judson and Owen (1999) strongly supports the corrected LSDV estimator (LSDVC) compared to more traditional GMM estimators when N is only moderately large. They point out, however, that "a method for implementing LSDVC for an unbalanced panel has not yet been implemented," which is clearly an important cause for their limited applicability.

This paper extends the bias approximation formulas in Bun and Kiviet (2003) to accommodate unbalanced panels with a strictly exogenous selection rule. Monte Carlo experiments are carried out to assess how unbalancedness affects the LSDV

2

bias and the bias approximations of various order.

2. Bias approximations

Kiviet (1995), (1999) and Kiviet and Bun (2003) consider the standard dynamic panel data model

$$y_{it} = \gamma y_{i,t-1} + x_{it}'\beta + \eta_i + \epsilon_{it}, \ i = 1, ..., N \text{ and } t = 1, ..., T. \qquad (2.1)$$

where y_{it} is the dependent variable; x_{it} is the $((k-1) \times 1)$ vector of strictly exogenous explanatory variables; η_i is an unobserved individual effect; and ϵ_{it} is an unobserved white noise disturbance. Collecting observations over time and across individuals gives

$$y = D\eta + W\delta + \epsilon,$$

where y is the $(NT \times 1)$ vector of observations for the dependent variable; $D = I_N \otimes \iota_T$ is the $(NT \times N)$ matrix of individual dummies, with ι_T being the $(T \times 1)$ vector of all unity elements; η is the $(N \times 1)$ vector of individual effects; $W = \left[y_{-1} \vdots X \right]$ is the $(NT \times k)$ matrix of explanatory variables; y_{-1} is y lagged one time; X is the $(NT \times (k-1))$ matrix of strictly exogenous explanatory variables; ϵ is

3

the $(NT \times 1)$ vector of white noise disturbances; and $\delta = \left[\gamma \vdots \beta' \right]'$ is the $(k \times 1)$ vector of coefficients.

It has been long recognized that the LSDV estimator for model (2.1) is not consistent for finite T. Nickell (1981) derives an expression for the inconsistency for $N \to +\infty$, which is $O\left(T^{-1}\right)$. The following is the bias approximation obtained by Kiviet (1999), which contains terms of higher order than T^{-1} :

$$
\begin{aligned}
E(\delta_{LSDV} - \delta) &= E\left[\left(W'AW \right)^{-1} W'A\epsilon \right] = \qquad\qquad\qquad (2.2) \\
&= QE\left(W'A\epsilon \right) - QE\left(W'AWQW'A\epsilon \right) + \\
&\quad + QE\left(W'AWQW'AW \right) QE\left(W'A\epsilon \right) + o\left(N^{-1}T^{-1} \right) \\
&= c_1\left(T^{-1} \right) + c_2\left(N^{-1}T^{-1} \right) + c_3\left(N^{-1}T^{-2} \right) + o\left(N^{-1}T^{-1} \right).
\end{aligned}
$$

where expectations are to be meant conditional on strictly exogenous regressors, individual effects and start-up values for y, $A = I_{NT} - D\left(D'D\right)^{-1} D'$ is the within operator and $Q = \left[E\left(W'AW \right) \right]^{-1}$. The bias approximation in (2.2) is more accurate than Kiviet's (1995), where the remainder is $O\left(N^{-1}T^{-1} \right)$. Bun and Kiviet (2003) consider (2.2) but with simpler formulas for each component.

Consider, now, model (2.1), but with some observation missing in the interval $[0, T]$ for some individuals. Define a selection indicator r_{it} such that $r_{it} = 1$ if

4

(y_{it}, x_{it}) is observed and $r_{it} = 0$ otherwise. From this define the dynamic selection rule $s(r_{it}, r_{i,t-1})$ selecting only the observations that are usable for the dynamic model, namely those for which both current values and one-time lagged values are observable:

$$s_{it} = \begin{cases} 1 & \text{if } (r_{i,t}, r_{i,t-1}) = (1,1) \\ 0 & \text{otherwise} \end{cases} , \; i = 1, ..., N \text{ and } t = 1, ..., T$$

Thus, for any i the number of usable observations is given by $T_i = \sum_{t=1}^{T} s_{it}$. The total number of usable observations is given by $n = \sum_{i=1}^{N} T_i$, while $\overline{T} = n/N$ denotes the average group size.

The unbalanced dynamic model can then be written as

$$s_{it} y_{it} = s_{it} \left(\gamma y_{i,t-1} + x'_{it}\beta + \eta_i + \epsilon_{it} \right), \; i = 1, ..., N \text{ and } t = 1, ..., T \tag{2.3}$$

with the unbalanced LSDV estimator given by

$$\delta_{LSDV} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \left(w_{it} - \overline{w}_i \right) \left(w_{it} - \overline{w}_i \right)' \right)^{-1} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \left(w_{it} - \overline{w}_i \right) \left(y_{it} - \overline{y}_i \right) \right), \tag{2.4}$$

where w'_{it} denotes the row of W for individual i at time t, $\overline{w}_i = (1/T_i) \sum_{t=1}^{T} s_{it} w_{it}$

5

and $\overline{y}_i = (1/T_i) \sum_{t=1}^{T} s_{it} y_{it}$ (see Wooldridge (2001)).

More conveniently, we can derive (2.4) in matrix form. For each i define the $(T \times 1)$-vector $s_i = [s_{i1}..., s_{iT}]'$ and the $T \times T$ diagonal matrix S_i having the vector s_i on its diagonal. Define also the $(NT \times NT)$ block-diagonal matrix $S = diag(S_i)$. Then, the following formulation is equivalent to model (2.3)

$$Sy = SD\eta + SW\delta + S\epsilon. \tag{2.5}$$

The LSDV estimator is given by

$$\delta_{LSDV} = (W'A_sW)^{-1}W'A_sy, \tag{2.6}$$

where

$$A_s = S\left(I - D(D'SD)^{-1}D'\right)S$$

is the symmetric and idempotent $(NT \times NT)$ matrix wiping out individual means and also selecting usable observations.

Initial times may vary across individuals and are given by $t_0^i = \min\{t : s_{i,t+1} = 1$ and $t = 0, 1, ..., T-1\}$, so that y_{i,t_0^i} indicates the start-up value for i. Then, let y_{t_0} denote the $(N \times 1)$-vector of start-up values.

We make the following assumption: a) $\gamma < 1$ and for each i the variables in X are stationary over time; b) $\epsilon_{it}|X, S, \eta, y_{t_0} \sim i.i.d.N\left(0, \sigma_\epsilon^2\right) \forall i, t$.

Considering all expectations below as conditional on (X, S, η, y_{t_0}), the LSDV bias is given by

$$E\left(\delta_{LSDV} - \delta\right) = E\left[\left(W'A_sW\right)^{-1}W'A_s\epsilon\right]. \tag{2.7}$$

Under our assumption, S is strictly exogenous, so that all the properties of normally distributed variables can be used as in Kiviet (1995) and (1999) to derive the bias approximations. These will differ from expression (2.2) and the approximation formulas in Bun and Kiviet (2003) only for A_s replacing A (of course, the special form of A_s matters for the order of the approximation terms):

$$
c_1\left(\overline{T}^{-1}\right) = \sigma_\epsilon^2 tr\left(\Pi\right) q_1; \tag{2.8}
$$

$$
c_2\left(N^{-1}\overline{T}^{-1}\right) = -\sigma_\epsilon^2\left[Q\overline{W}'\Pi A_s\overline{W} + tr\left(Q\overline{W}'\Pi A_s\overline{W}\right)I_{k+1} + 2\sigma_\epsilon^2 q_{11}tr\left(\Pi'\Pi\Pi\right)I_{k+1}\right]q_1;
$$

$$
c_3\left(N^{-1}\overline{T}^{-2}\right) = \sigma_\epsilon^4 tr\left(\Pi\right)\left\{2q_{11}Q\overline{W}'\Pi\Pi'\overline{W}q_1 + \right.
$$

$$
\left. \left[\left(q_1'\overline{W}'\Pi\Pi'\overline{W}q_1\right) + q_{11}tr\left(Q\overline{W}'\Pi\Pi'\overline{W}\right) + 2tr\left(\Pi'\Pi\Pi'\Pi\right)q_{11}^2\right]q_1\right\};
$$

where $Q = \left[E\left(W'A_sW\right)\right]^{-1} = \left[\overline{W}'A_s\overline{W} + \sigma_\epsilon^2 tr\left(\Pi'\Pi\right)e_1e_1'\right]^{-1}$; $\overline{W} = E\left(W\right)$; $e_1 =$

7

$(1, 0, ..., 0)'$ is a $(k \times 1)$ vector; $q_1 = Qe_1$; $q_{11} = e_1' q_1$; L_T is the $(T \times T)$ matrix with unit first lower subdiagonal and all other elements equal to zero; $L = I_N \otimes L_T$; $\Gamma_T = (I_T - \gamma L_T)^{-1}$; $\Gamma = I_N \otimes \Gamma_T$; and $\Pi = A_s L \Gamma$.

The following three possible bias approximations emerge

$$B_1 = c_1 \left(\overline{T}^{-1} \right); \ B_2 = B_1 + c_2 \left(N^{-1} \overline{T}^{-1} \right); \ B_3 = B_2 + c_3 \left(N^{-1} \overline{T}^{-2} \right).$$

In the next Section we shall evaluate their performance in approximating the LSDV bias as estimated by Monte Carlo simulations.

3. Monte Carlo Experiments

Our Monte Carlo experiments closely follows Kiviet (1995) and Bun and Kiviet (2003), with the difference that a strictly exogenous selection rule is included. Data for y_{it} are generated by model (2.1) with $k = 2$ and for x_{it} by

$$x_{it} = \rho x_{i,t-1} + \xi_{it}, \ \xi_{it} \sim N\left(0, \sigma_\xi^2\right), \ i = 1, ..., N \text{ and } t = 1, ..., T$$

Initial observations y_{i0} and x_{i0} are generated following a procedure that avoids the waste of random numbers and small sample non-stationary problems (see

Kiviet (1986)) and are kept fixed across replications. The long-run coefficient $\beta / (1 - \gamma)$ is always kept fixed to unity, so $\beta = 1 - \gamma$; σ_ϵ^2 is normalized to unity; γ and ρ alternate between 0.2 and 0.8 and σ_s^2 alternates between 2 and 9. The individual effects η_i are generated by assuming $\eta_i \sim N\left(0, \sigma_\eta^2\right)$ and $\sigma_\eta = \sigma_\epsilon \left(1 - \gamma\right)$. Kiviet (1995) finds that the signal to noise ratio of the regression, σ_s^2, is a key determinant of the relative bias of estimators and therefore needs to be controlled in the simulation. Thus, once fixing σ_s^2 (along with γ, β and ρ) σ_ξ^2 gets uniquely determined by

$$\sigma_\xi^2 = \beta^{-2} \left[\sigma_s^2 - \frac{\gamma^2}{1 - \gamma^2} \sigma_\epsilon^2\right] \left[1 + \frac{(\gamma + \rho)^2}{1 + \gamma\rho} (\gamma\rho - 1) - (\gamma\rho)^2\right].$$

We consider two different sample sizes, $\left(N, \overline{T}\right) = (20, 20)$ and $\left(N, \overline{T}\right) = (10, 40)$. Then, following Baltagi and Chang (1994), we control for the extent of unbalancedness as measured by the Ahrens and Pincus (1981) index:

$$\omega = \frac{N}{\overline{T} \sum_{i=1}^{N} (1/T_i)}$$

with $0 < \omega \leq 1$ ($\omega = 1$ when the panel is balanced). For each sample size we analyze a case of mild unbalancedness ($\omega = 0.96$) and a case of severe unbalanced-

9

ness ($\omega = 0.32$). Although not strictly required by our formulas, we exclude gaps and set $t_0^i = 0$ for each i. Individuals are then partitioned into two sets of equal dimension: one set contains the first $N/2$ individuals with the last h observations discarded for each i in the set, so $T_i = T - h$; the other contains the remaining $N/2$ individuals with $T_i = T$ for each i in the set. For each sample size we fix T and h so that ω takes on the desired value. The details of the four panel designs are summarized in Table 1.

-Table 1 approximately here-

To carry out the Monte Carlo experiments and calculate the bias approximations we have developed two codes in Stata, version 8 (available on request). Table 2 presents the results of our simulations for the unbalanced panels. Columns 1 to 5 show the various parametrizations for each panel design. Columns 6 and 10 show the actual LSDV biases for γ and β, respectively, as estimated by 20000 Monte Carlo replications. As expected from results of the preceding section, the bias for both γ and β is decreasing in $\overline{T}$. Interestingly, the bias is also decreasing in the degree of unbalancedness for given sample size. With respect to the other parameter of interest, σ_s^2, γ and ρ, the patterns found by Bun and Kiviet (2003) are all confirmed.

10

Columns 7 to 9 and 11 to 13 in Table 2 present bias approximations for γ and β, respectively. Regardless of the degree of unabalancedness, they are accurate, with the approximations including higher order terms being equal to the true bias in a vast majority of cases. In addition, as it happens for the balanced designs studied by Bun and Kiviet (2003), the leading term of the approximations already accounts for a predominant portion of the true bias (90% on average).

-Table 2 approximately here-

4. Conclusion

This paper has derived approximations of various order to the bias of the LSDV dynamic estimator for unbalanced panel data. The approximations are obtained by modifying the within operator to accommodate the dynamic selection rule. Monte Carlo experiments confirm all results by Bun and Kiviet (2003) for balanced panels. In particular we find that the bias approximations are accurate with a decreasing contribution to the bias of the higher order terms. We also find that the bias is decreasing in $\overline{T}$ and in the degree of unbalancedness. Our results, therefore, suggest that 1) the derived bias approximations can be used to construct LSDVC estimators for unbalanced panels, removing an important cause for their

11

limited applicability; 2) Bun and Kiviet's (2003) finding that bias corrections can be based on the simple leading term of the approximation carries over into unbalanced panels; 3) while increasing $\overline{T}$ is always beneficial in reducing the LSDV bias, reducing unbalancedness at the expenses of time observations, for given N and $\overline{T}$, may instead exacerbate the bias.

References

1. Ahrens, H., Pincus, R., 1981. On Two Measures of Unbalancedness in a One-way Model and Their Relation to Efficiency. Biometric Journal 23, 227-235.

2. Baltagi, B.H., Chang, Y.J., 1995. Incomplete Panels. Journal of Econometrics. 67-89.

3. Bun, M.J.G., Kiviet, J.F., 2003. On the diminishing returns of higher order terms in asymptotic expansions of bias. Economics Letters, 79, 145-152.

4. Judson, R.A., and Owen, A.L., 1999. Estimating dynamic panel data models: a guide for macroeconomists. Economics Letters, 65, 9-15.

5. Kiviet, J.F., 1986. On the Rigour of Some Misspecification Tests For Modelling Dynamic Relationships. Review of Economic Studies LIII, 241-261.

6. Kiviet, J.F., 1995. On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models. Journal of Econometrics, 68, 53-78.

7. Kiviet, J.F., 1999. Expectation of Expansions for Estimators in a Dynamic Panel Data Model; Some Results for Weakly Exogenous Regressors. In: Hsiao, C., Lahiri, K., Lee, L.-F., Pesaran, M.H. (Eds.), Analysis of Panel

Data and Limited Dependent Variables. Cambridge University Press, Cambridge.

8. Nickell, S.J., 1981. Biases in Dynamic Models with Fixed Effects. Econometrica, 49, 1417-1426.

9. Wooldridge, J.M., 2001. The Econometric Analysis of Cross Section and Panel Data. The MIT Press, Cambridge

Table 1

Panel designs

N	$\overline{T}$	T	T_i	ω
20	20	24	16 ($i \leq 10$), 24 ($i > 10$)	0.96
		36	4 ($i \leq 10$), 36 ($i > 10$)	0.32
10	40	48	32 ($i \leq 5$), 48 ($i > 5$)	0.96
		72	8 ($i \leq 5$), 72 ($i > 5$)	0.32

Table 2

Actual LSDV bias and bias approximations for unbalanced panels

σ_s^2	$\overline{T}$	γ	ρ	ω	Bias γ	$B_{1,\gamma}$	$B_{2,\gamma}$	$B_{3,\gamma}$	Bias β	$B_{1,\beta}$	$B_{2,\beta}$	$B_{3,\beta}$
2	20	0.2	0.2	0.96	-0.021	-0.020	-0.021	-0.021	0.002	0.002	0.002	0.002
				0.36	-0.019	-0.018	-0.018	-0.018	0.003	0.003	0.003	0.003
			0.8	0.96	-0.038	-0.036	-0.038	-0.038	0.026	0.024	0.025	0.025
				0.36	-0.034	-0.032	-0.034	-0.034	0.024	0.022	0.023	0.024
		0.8	0.2	0.96	-0.102	-0.098	-0.100	-0.102	0.003	0.002	0.003	0.003
				0.36	-0.072	-0.067	-0.070	-0.072	0.001	0.001	0.001	0.001
			0.8	0.96	-0.108	-0.101	-0.105	-0.108	0.022	0.021	0.022	0.022
				0.36	-0.076	-0.069	-0.074	-0.076	0.020	0.018	0.020	0.020
	40	0.2	0.2	0.96	-0.011	-0.010	-0.011	-0.011	0.002	0.001	0.001	0.001
				0.36	-0.011	-0.010	-0.010	-0.010	0.002	0.002	0.002	0.002
			0.8	0.96	-0.020	-0.018	-0.019	-0.020	0.014	0.012	0.013	0.013
				0.36	-0.019	-0.017	-0.018	-0.019	0.014	0.012	0.014	0.014
		0.8	0.2	0.96	-0.051	-0.046	-0.050	-0.051	0.001	0.001	0.001	0.001
				0.36	-0.040	-0.036	-0.039	-0.040	0.001	0.000	0.001	0.001
			0.8	0.96	-0.054	-0.048	-0.052	-0.054	0.015	0.013	0.015	0.015
				0.36	-0.043	-0.036	-0.042	-0.043	0.011	0.010	0.012	0.012
σ_s^2	$\overline{T}$	γ	ρ	ω	Bias γ	$B_{1,\gamma}$	$B_{2,\gamma}$	$B_{3,\gamma}$	Bias β	$B_{1,\beta}$	$B_{2,\beta}$	$B_{3,\beta}$
9	20	0.2	0.2	0.96	-0.004	-0.004	-0.004	-0.004	0.000	0.000	0.000	0.000
				0.36	-0.004	-0.004	-0.004	-0.004	0.001	0.001	0.001	0.001
			0.8	0.96	-0.013	-0.012	-0.013	-0.013	0.009	0.009	0.009	0.009
				0.36	-0.012	-0.011	-0.012	-0.012	0.009	0.008	0.009	0.009
		0.8	0.2	0.96	-0.006	-0.006	-0.006	-0.006	0.000	0.000	0.000	0.000
				0.36	-0.004	-0.004	-0.004	-0.004	0.000	0.000	0.000	0.000
			0.8	0.96	-0.034	-0.032	-0.033	-0.033	0.012	0.011	0.012	0.012
				0.36	-0.019	-0.017	-0.019	-0.019	0.008	0.007	0.008	0.008
	40	0.2	0.2	0.96	-0.003	-0.003	-0.003	-0.003	0.000	0.000	0.000	0.000
				0.36	-0.003	-0.002	-0.002	-0.002	0.000	0.000	0.000	0.000
			0.8	0.96	-0.008	-0.007	-0.008	-0.008	0.006	0.005	0.005	0.005
				0.36	-0.007	-0.006	-0.007	-0.007	0.005	0.005	0.005	0.005
		0.8	0.2	0.96	-0.007	-0.007	-0.007	-0.007	0.000	0.000	0.000	0.000
				0.36	-0.004	-0.003	-0.004	-0.004	0.000	0.000	0.000	0.000
			0.8	0.96	-0.020	-0.018	-0.020	-0.020	0.007	0.005	0.006	0.006
				0.36	-0.014	-0.012	-0.013	-0.014	0.006	0.005	0.006	0.006