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

Giovanni SF Bruno, Universita' Bocconi, Milano

giovanni.bruno@uni-bocconi.it

10th London Stata User Group Meeting, 28-29 June 2004

Outline of the presentation

- Introduction
- Bias approximations
- The Stata program: xtlsdvc
- Monte Carlo results

Introduction

- The Least Square Dummy Variable (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 $O(T^{-1})$.
- IV-GMM estimators: Anderson-Hsiao (1982); Arellano-Bond (1991); Blundell-Bond (1998)
- 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. However "a method for implementing LSDVC for an unbalanced panel has not yet been implemented"

- Bruno (2004) extends the bias approximation formulas in Bun and Kiviet (2003) to accommodate unbalanced panels with a strictly exogenous selection rule, and carry out Monte Carlo experiments to assess how unbalancedness affects the LSDV bias and the bias approximations of various order.
- For this talk, I have gone a step forward, implementing a Stata code for the LSDVC estimator. Its performance has been evaluated via Monte Carlo experiments.

Bias approximations

Consider the standard autoregressive panel data model

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

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

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

y is the $(NT \times 1)$ vector of obs. 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;

$W = [y_{-1} : X]$ 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 the $(NT \times 1)$ vector of white noise disturbances;

$\delta = [\gamma : \beta']'$ is the $(k \times 1)$ vector of coefficients.

Kiviet (1995) obtains a bias approximation that contains terms of

higher order than T^{-1}. In Kiviet (1999) a more accurate bias approximation is derived. Bun and Kiviet (2003) reformulate the approximation in Kiviet (1999) with simpler formulas for each term.

In Bruno (2004) I extend the autoregressive model to allow missing observations. Define a selection indicator r_{it} such that $r_{it} = 1$ if (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 obs. for which both current values and one-time lagged values are observable:

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

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 $n = \sum_{i=1}^{N} T_i$,

$\overline{T} = n/N$ denotes the average group size.

The unbalanced dynamic model can then be written as

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

To formulate this in matrix form,

- for each i define the $(T \times 1)$-vector $s_i = [s_{i1} \ldots, s_{iT}]'$ and the $T \times T$ diagonal matrix S_i having the vector s_i on its diagonal; and
- define the $(NT \times NT)$ block-diagonal matrix $S = diag(S_i)$.

Thus,

$$Sy = SD\eta + SW\delta + S\epsilon.$$

The LSDV estimator is:

$$\delta_{LSDV} = (W' A_s W)^{-1} W' A_s y,$$

where

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

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

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

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

If S is strictly exogenous, the same approach as in Kiviet (1995) and (1999) can be followed to derive the bias approximations. These will differ from the approximation formulas in Bun and Kiviet (2003) only for A_s replacing the within operator:

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

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

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

where $Q = [E(W'A_sW)]^{-1} = \left[\overline{W}'A_s\overline{W} + \sigma_\epsilon^2 tr(\Pi'\Pi)e_1e_1'\right]^{-1}$; $\overline{W} = E(W)$; $e_1 = (1,0,\ldots,0)'$ 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

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

The Stata program calculating the LSDVC estimator: xtlsdvc

- The LSDV estimator may be corrected by subtracting the bias terms from it.
- The foregoing bias approximations, however, depend on the unknown population parameters γ and σ_ϵ^2.
- To make correction feasible, estimates from a consistent estimator should replace γ and σ_ϵ^2 into the bias apprximation terms.
- Three natural options for the initial consistent estimator are: Anderson-Hsiao (*ah*); Arellano-Bond (*ab*); and the Blundell-Bond system estimator (*bb*)

There is a different corrected estimator for each order of bias approximation and choice of initial estimator:

$LSDVC_i^j = LSDV - \widehat{B}_i^j,\ i = 1,2,3,\ j = ah,\ ab,\ bb.$

$LSDVC_i^j$ is implemented by my Stata code -xtlsdvc-, for the three levels of approximation accuracy and with three alternative initial estimators: Anderson-Hsiao (option: initial(ah)); Arellano-Bond (option: initial(ab)); Blundell-Bond, through David Doorman's Stata routine -xtabond2- (option: initial(bb)).

Statistics/Data Analysis

help for **xtlsdvc**

Corrected LSDV dynamic panel data estimator

xtlsdvc *depvar* [*varlist*] [**if** *exp*] [**in** *range*] **, initial(***estimator***)** [**bias(#)** **lsdv** **first**]

where *estimator* is

ah	Anderson-Hsiao
ab	Arellano-Bond
bb	Blundell-Bond

Options

initial(*estimator***)** specifies which consistent estimator among Anderson-Hsiao (**ah**), Arellano-Bond (**ab**), and Blundell-Bond (**bb**) is to initialize the bias correction.

bias(#) determines the accuracy of the approximation: up to O(1/T) (**1**); up to O(1/NT) (**2**); up to O(1/NT^2) (**3**).

first requests that the first-stage regression results be displayed.

lsdv requests that the lsdv regression results be displayed.

Monte Carlo Experiments

We follow 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 the autoregressive model with $k = 2$ and for x_{it} by

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

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 the signal to noise ratio σ_s^2 alternates between 2 and 9.

Two different sample sizes are considered, $(N, \overline{T}) = (20, 20)$ and $(N, \overline{T}) = (10, 40)$. Then, following Baltagi and Chang (1994), I 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 I analyze a case of mild unbalancedness ($\omega = 0.96$) and a case of severe unbalancedness ($\omega = 0.32$). My Stata code -xtdes2- calculates ω (along with $\overline{T}$ and T_i) for the relevant estimation sample.

The details of the four panel designs are summarized in Table 1.

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

To carry out the Monte Carlo experiments and calculate the theoretical bias approximations I have developed do files that generates the data according to the DGP described above.

Table 2 presents the results of my simulations for the bias approximations. 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. The bias for both γ and β is decreasing in $\overline{T}$. Interestingly, the bias for γ is also decreasing in the degree of unbalancedness for given sample size.

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 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, on

average, for 90% of the true bias.

Table 2
Actual LSDV bias and bias approximations for unbalanced panels

σ_s^2	$\overline{T}$	γ	ρ	ω	Bias γ	$B1_{,\gamma}$	$B2_{,\gamma}$	$B3_{,\gamma}$	Bias β	$B1_{,\beta}$	$B2_{,\beta}$	$B3_{,\beta}$
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	$B1_{,\gamma}$	$B2_{,\gamma}$	$B3_{,\gamma}$	Bias β	$B1_{,\beta}$	$B2_{,\beta}$	$B3_{,\beta}$
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

Monte Carlo experiments have been also carried out to compare the performance of the LSDVC estimator (initialised by AH) against Anderson-Hsiao, Arellano-Bond and the LSDV estimators. There are the following results:

1) LSDVC estimators and AH have smaller bias than AB and LSDV, with $LSDVC_3$ performing slightly better than $LSDVC_1$ and $LSDVC_2$;

2) The LSDVC estimators have always the smallest RMSE (with almost no difference among the three versions);

3) Similarly to what found for the LSDV estimator (Bruno 2004), the AB bias for γ is always negative, and it is increasing in absolute value from severe unbalancedness to mild unbalancedness for given sample size.

Conclusion

Based on the bias approximation formulas for the LSDV estimator, a corrected LSDV estimator suitable for unbalanced panels has been obtained and implemented through my Stata routine -xtlsdvc-.

Monte Carlo experiments show that the LSDVC estimator, in small samples, outperforms consistent IV-GMM estimators such as Anderson-Hsiao and Arellano-Bond. This occurs in terms of both bias and RMSE and regardless of the degree of unbalancedness. These results confirms the findings by Judson and Owen (1999).

Limits:

1) strict exogeneity of S and X;

2) white noise disturbances;

3) analytical standard errors for the LSDVC estimator break down quite often. Solution: bootstrap.

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. Bruno G.S.F., 2004. Approximating the Bias of the LSDV Estimator for Dynamic Unbalanced Panel Data Models. Uiversità Bocconi w.p. 2004-1.
4. Bun, M.J.G., Carree, M.A., 2003. Bias-corrected estimation in dynamic panel data models. METEOR research memorandum RM/02/025, University of Maastricht.
5. 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.
6. Judson, R.A., and Owen, A.L., 1999. Estimating dynamic panel

data models: a guide for macroeconomists. Economics Letters, 65, 9-15.

7. Kiviet, J.F., 1986. On the Rigour of Some Misspecification Tests For Modelling Dynamic Relationships. Review of Economic Studies LIII, 241-261.
8. Kiviet, J.F., 1995. On Bias, Inconsistency and Efficiency of Various Estimators in Dynamic Panel Data Models. Journal of Econometrics, 68, 53-78.
9. 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.
10. Nickell, S.J., 1981. Biases in Dynamic Models with Fixed Effects. Econometrica, 49, 1417-1426.
11. Wooldridge, J.M., 2001. The Econometric Analysis of Cross

Section and Panel Data. The MIT Press, Cambridge