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

Giovanni SF Bruno, Università 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 $N \rightarrow \infty$, 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; $\eta_i$ is an unobserved
individual effect; and $\epsilon_{it}$ 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 $\iota_T$ being the $(T \times 1)$ vector of all unity elements;
$\eta$ 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;
$\epsilon$ is the $(NT \times 1)$ vector of white noise disturbances;
$\delta = [y:\beta']'$ is the $(k \times 1)$ vector of coefficients.

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

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:

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

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 $n = \sum_{i=1}^{N} T_i$, $ar{T} = n/N$ 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 $S = \text{diag}(S_i)$.

Thus,

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

The LSDV estimator is:

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

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.

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

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

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:
\[ c_1 \left( \bar{T}^{-1} \right) = \sigma^2 \text{tr}(\Pi)q_1; \]
\[ c_2 \left( N^{-1} \bar{T}^{-1} \right) = -\sigma^2 \left[ Q\bar{W}'\Pi A_s\bar{W} + \text{tr}(Q\bar{W}'\Pi A_s\bar{W})I_{k+1} + 2\sigma^2 q_{11} \text{tr}(\Pi'\Pi\Pi)I_{k+1} \right]q_1; \]
\[ c_3 \left( N^{-1} \bar{T}^{-2} \right) = \sigma^4 \text{tr}(\Pi) \left\{ 2q_{11} Q\bar{W}'\Pi\Pi'\bar{W}q_1 + \left[ (q_1'\bar{W}'\Pi\Pi'\bar{W}q_1) + q_{11} \text{tr}(Q\bar{W}'\Pi\Pi'\bar{W}) + 2\text{tr}(\Pi'\Pi\Pi'\Pi)q_{11}^2 q_1 \right] \right\}; \]

where \( Q = [E(W'A_sW)]^{-1} = \left[ \bar{W}'A_s\bar{W} + \sigma^2 \text{tr}(\Pi'\Pi)e_1 e_1' \right]^{-1}; \bar{W} = E(W); \)
\( e_1 = (1, 0, \ldots, 0)' \) 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 \( \Pi = A_sL\Gamma. \)
The following three possible bias approximations emerge

\[ B_1 = c_1 \left( \frac{1}{T} \right); \quad B_2 = B_1 + c_2 \left( \frac{1}{N^1} \frac{1}{T} \right); \quad B_3 = B_2 + c_3 \left( \frac{1}{N^1} \frac{1}{T}^2 \right). \]

**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 \( \gamma \) and \( \sigma^2_{\varepsilon} \).
- To make correction feasible, estimates from a consistent estimator should replace \( \gamma \) and \( \sigma^2_{\varepsilon} \) into the bias approximation 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:
\[ \text{LSDVC}_i^j = \text{LSDV} - \hat{B}_i^j, \quad i = 1, 2, 3, \quad j = ah, \ ab, \ bb. \]

\( \text{LSDVC}_i^j \) 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)).
help for xtlsdvc

Corrected LSDV dynamic panel data estimator

\texttt{xtlsdvc} depvar [varlist] [if exp] [in range], \texttt{initial(estimator)} [bias(#)] 
\texttt{lsdv first}

where \texttt{estimator} is
\begin{align*}
\texttt{ah} & \quad \text{Anderson-Hsiao} \\
\texttt{ab} & \quad \text{Arellano-Bond} \\
\texttt{bb} & \quad \text{Blundell-Bond}
\end{align*}

Options

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

\texttt{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).

\texttt{first} requests that the first-stage regression results be displayed.

\texttt{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

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

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 $\beta = 1 - \gamma$; $\sigma_{\xi}^2$ is normalized to unity; $\gamma$ and $\rho$ alternate between 0.2 and 0.8 and the signal to noise ratio $\sigma_s^2$ 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:

\[
\omega = \frac{N}{\bar{T} \sum_{i=1}^{N} (1/T_i)}
\]

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 \(\omega\) (along with \(\bar{T}\) and \(T_i\)) for the relevant estimation sample.

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

<table>
<thead>
<tr>
<th>$N$</th>
<th>$T$</th>
<th>$T_i$</th>
<th>$\omega$</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>20</td>
<td>24</td>
<td>16 ($i \leq 10$), 24 ($i &gt; 10$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>36</td>
<td>4 ($i \leq 10$), 36 ($i &gt; 10$)</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>48</td>
<td>32 ($i \leq 5$), 48 ($i &gt; 5$)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>72</td>
<td>8 ($i \leq 5$), 72 ($i &gt; 5$)</td>
</tr>
</tbody>
</table>
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 $\gamma$ and $\beta$, respectively, as estimated by 20000 Monte Carlo replications. The bias for both $\gamma$ and $\beta$ is decreasing in $\bar{T}$. Interestingly, the bias for $\gamma$ 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 $\gamma$ and $\beta$, respectively. Regardless of the degree of unbalancedness, 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

<table>
<thead>
<tr>
<th>$\sigma^2$</th>
<th>$T$</th>
<th>$\gamma$</th>
<th>$\rho$</th>
<th>$\omega$</th>
<th>Bias $\gamma$</th>
<th>$B_{1,\gamma}$</th>
<th>$B_{2,\gamma}$</th>
<th>$B_{3,\gamma}$</th>
<th>Bias $\beta$</th>
<th>$B_{1,\beta}$</th>
<th>$B_{2,\beta}$</th>
<th>$B_{3,\beta}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>20</td>
<td>0.2</td>
<td>0.96</td>
<td>0.021</td>
<td>-0.020</td>
<td>-0.021</td>
<td>-0.021</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.019</td>
<td>-0.018</td>
<td>-0.018</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.038</td>
<td>-0.038</td>
<td>-0.038</td>
<td>0.026</td>
<td>0.024</td>
<td>0.025</td>
<td>0.025</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>0.2</td>
<td>0.96</td>
<td>-0.0102</td>
<td>-0.098</td>
<td>-0.100</td>
<td>-0.102</td>
<td>0.003</td>
<td>0.002</td>
<td>0.003</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.072</td>
<td>-0.067</td>
<td>-0.070</td>
<td>-0.072</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.108</td>
<td>-0.101</td>
<td>-0.105</td>
<td>-0.108</td>
<td>0.022</td>
<td>0.021</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.076</td>
<td>-0.069</td>
<td>-0.074</td>
<td>-0.076</td>
<td>0.020</td>
<td>0.018</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>0.2</td>
<td>0.96</td>
<td>-0.0111</td>
<td>-0.010</td>
<td>-0.011</td>
<td>-0.011</td>
<td>0.002</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.020</td>
<td>-0.018</td>
<td>-0.010</td>
<td>-0.010</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.019</td>
<td>-0.017</td>
<td>-0.018</td>
<td>-0.019</td>
<td>0.014</td>
<td>0.012</td>
<td>0.013</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.046</td>
<td>-0.050</td>
<td>-0.051</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.054</td>
<td>-0.048</td>
<td>-0.052</td>
<td>-0.054</td>
<td>0.015</td>
<td>0.013</td>
<td>0.015</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.043</td>
<td>-0.036</td>
<td>-0.042</td>
<td>-0.043</td>
<td>0.011</td>
<td>0.010</td>
<td>0.012</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>0.2</td>
<td>0.96</td>
<td>-0.004</td>
<td>-0.004</td>
<td>-0.004</td>
<td>-0.004</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.013</td>
<td>-0.012</td>
<td>-0.013</td>
<td>-0.013</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.012</td>
<td>-0.011</td>
<td>-0.012</td>
<td>-0.012</td>
<td>0.009</td>
<td>0.008</td>
<td>0.009</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>0.2</td>
<td>0.96</td>
<td>-0.006</td>
<td>-0.006</td>
<td>-0.006</td>
<td>-0.006</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.034</td>
<td>-0.033</td>
<td>-0.033</td>
<td>-0.033</td>
<td>0.012</td>
<td>0.011</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.019</td>
<td>-0.017</td>
<td>-0.019</td>
<td>-0.019</td>
<td>0.008</td>
<td>0.007</td>
<td>0.008</td>
</tr>
<tr>
<td></td>
<td></td>
<td>40</td>
<td>0.2</td>
<td>0.96</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.003</td>
<td>-0.003</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.007</td>
<td>-0.007</td>
<td>-0.007</td>
<td>-0.007</td>
<td>0.006</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.004</td>
<td>-0.003</td>
<td>-0.004</td>
<td>-0.004</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.8</td>
<td>-0.020</td>
<td>-0.018</td>
<td>-0.020</td>
<td>-0.020</td>
<td>0.007</td>
<td>0.005</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.36</td>
<td>-0.014</td>
<td>-0.012</td>
<td>-0.013</td>
<td>-0.014</td>
<td>0.006</td>
<td>0.005</td>
<td>0.006</td>
</tr>
</tbody>
</table>
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 $\gamma$ 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


Section and Panel Data. The MIT Press, Cambridge