TI 2002-008/2
Tinbergen Institute Discussion Paper

Nearly Unbiased Estimation in Dynamic Panel Data Models

Martin A. Carree

Department of General Economics, Faculty of Economics, Erasmus University Rotterdam,
Tinbergen Institute, and
Faculty of Economics and Business Administration, Maastricht University

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Keizersgracht 482
1017 EG Amsterdam
The Netherlands
Tel.: +31.(0)20.5513500
Fax: +31.(0)20.5513555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Rotterdam
The Netherlands
Tel.: +31.(0)10.4088900
Fax: +31.(0)10.4089031

Nearly Unbiased Estimation in Dynamic Panel Data Models

by

Martin A. Carree

Faculty of Economics
Erasmus University Rotterdam
and
Faculty of Economics and Business Administartion
Maastricht University

Abstract

This paper introduces two easy to calculate estimators with desirable properties for the autoregressive parameter in dynamic panel data models. The estimators are (nearly) unbiased and perform satisfactorily even for small samples in either the time-series or cross-section dimension.

Version October 2001

Keywords: dynamic panel data; Nickell bias; bias-correction

JEL-code: C23

email: m.carree@mw.unimaas.nl or carree@few.eur.nl

dr. Martin A. Carree

MW-ORG

Maastricht University

PO Box 616

6200 MD Maastricht

The Netherlands

1. Introduction

The field of dynamic panel data models has received considerable attention in the last decade. A large part of this attention has been devoted to the development of improved GMM-estimators (see e.g. Blundell and Bond, 1998). However, despite the increasing sophistication of the GMM-estimators, at least two important problems remain. First, there is an important upward bias of the GMM-estimator in case the autoregressive parameter becomes close to unity (see Collado, 1997, Blundell and Bond, 1998, Kitazawa, 2001). Second, the performance of the GMM-estimators depends strongly upon the ratio of the variance of the fixed effects across individuals and the variance of the error term. In case the variance of fixed effects is much larger than that of the error term variance, the GMM-estimators perform poorly (Kitazawa, Table 3).

Kiviet (1995) chose a different approach trying to correct the "Nickell bias" of the well-known least-squares dummy variable (LSDV) approach. This paper is in line with this second approach but takes a different approach to removing the bias. An important advantage of making use of bias-correcting the LSDV estimator is that the performance of the estimators is independent of the ratio of the variance of the fixed effects and the error term variance. I develop two estimators, a linearly and a quadratically corrected LSDV estimator. Monte Carlo experiments show that they are (nearly) unbiased even for small *N* and *T*. In addition they outperform the GMM-estimators in terms of root mean squared errors.

2. Nearly unbiased estimators for the autoregressive parameter

The dynamic panel data model with h_i as the fixed-effects and g as the AR(1)-parameter equals

(1) $y_{it} = g\, y_{i,t-1} + h_i + n_{it}$ for $i = 1,\ldots,N$ and $t = 1,\ldots,T$.

We assume that this process has been going on for a long time and that, hence, the initial condition is

(2) $y_{i0} = \frac{h_i}{1-g} + \frac{n_{i0}}{\sqrt{1-g^2}}$ for $i = 1,\ldots,N$.

The disturbance term n_{it} has zero mean and constant variance s_n^2 for all observations. In addition, $E[n_{it} n_{js}] = 0$ in case $i \neq j$ or $s \neq t$. Equation (2) requires that $|g| < 1$. The least-squares dummy-variable (LSDV) estimator for the parameters in equation (1) can be derived by first eliminating the unknown individual effects:

(3) $\tilde{y}_{it} = g\, \tilde{y}_{i,t-1} + \tilde{n}_{it}$ for $i = 1,\ldots,N$ and $t = 1,\ldots,T$

with $\tilde{y}_{it} = y_{it} - \bar{y}_i$, $\tilde{y}_{i,t-1} = y_{i,t-1} - \bar{y}_{i,-1}$ and $\tilde{n}_{it} = n_{it} - \bar{n}_i$. The LSDV-estimators for g and the individual effects are equal to

(4) $\hat{g} = \sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{y}_{it}\tilde{y}_{i,t-1} / \sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{y}_{i,t-1}^2$ and $\hat{h}_i = \bar{y}_i - \hat{g}\,\bar{y}_{i,-1}$.

The LSDV-estimators are biased because $\tilde{y}_{i,t-1}$ and $\tilde{n}_{it}$ are correlated. Only when *T* tends to infinity does this correlation disappear and in many practical applications where the time period is relatively short the LSDV-estimators suffer from severe bias. Nickell (1981) has made an important contribution by deriving the bias of the LSDV-estimator for g when *N* tends to infinity. The asymptotic bias of the LSDV-estimator for a given *T*, also known as the "Nickell-bias" equals (see Nickell, 1981, Hsiao, 1986, p.74):

(5) $$\underset{N\to\infty}{plim}(\hat{g} - g) = -\frac{1+g}{T-1}\left\{1 - \frac{1-g^T}{T(1-g)}\right\}\left\{1 - \frac{2g}{(1-g)(T-1)}\left[1 - \frac{1-g^T}{T(1-g)}\right]\right\}^{-1}$$

Therefore, the asymptotic value of the LSDV-estimator $\hat{g}$ can be expressed as a function of g and the number of periods *T*, say $g(g, T)$. For example, for *T*=2 we find that $plim\,\hat{g} = (g-1)/2$. Hence, for *T*=2 and large *N* we may use $2\hat{g}+1$ as a (nearly) unbiased estimate for g. In general, it is possible to express g as a function of $plim\,\hat{g}$ and *T*, say $g = f(\underset{N\to\infty}{plim}\,\hat{g}, T)$, but with the function *f* unknown. We will show that this function *f* can be approximated very well by a linear or quadratic specification. Hence, for large *N* we may find the following nearly unbiased estimators as a function of the LSDV-estimate $\hat{g}$:[1]

(6) $\hat{g}_{lc} = a_T + b_T\hat{g}$

(7) $\hat{g}_{qc} = c_T + d_T\hat{g} + e_T\hat{g}^2$

[1] Bias correction of the autoregressive parameter in an AR(1) model has been dealt with in various papers, see e.g. Andrews (1993) and MacKinnon and Smith (1998). Cermeño (1999) proposes to extend the Andrews (1993)-bias correction to the case of panel data. However, no direct use is made of the "Nickell bias" then. In addition, it would require researchers using this correction to first make simulations for a range of values for g for the specific values of *N* and *T* in their samples.

with a_T, b_T, c_T, d_T and e_T as constants different for different values of *T*. We use the term *nearly* unbiased because *N* is finite and because the approximation of the function *f* is not perfect. Values for the constants are given in Table 1. They are computed by taking one thousand values of g from 0.000 to 0.999 with step size 0.001, then calculating the corresponding "Nickell-bias" to find the corresponding value of $\hat{g}$, and then performing a least squares regression of g on a constant and $\hat{g}$ to find a_T and b_T and a least squares regression of g on a constant, $\hat{g}$ and $\hat{g}^2$ to find c_T, d_T and e_T.[2] The R^2 of this regression is 0.9990 or higher for the linear case and 0.9999 or higher for the quadratic case. This very good fit indicates that very little bias-correction is lost when using either a linear or quadratic approximation. Table 1 provides values for the constants up till *T* equal to thirty. For values of *T* in excess of thirty the following approximations can be used:

(8) $$\hat{g}_{lc} = \hat{g} + \frac{0.839 + 1.553\hat{g}}{T - 2.083} \qquad \text{for } T > 30$$

(9) $$\hat{g}_{qc} = \hat{g} + \frac{0.908 + 0.575\hat{g} + 1.256\hat{g}^2}{T - 2.397} \qquad \text{for } T > 30$$

These approximations are found by using the combinations of g and the implied "Nickell bias" for all values of *T* between 10 and 30 and again using (non-linear) regression (number of 'observations' is 21,000). The R^2 of regression outcome equation (8) is 0.9995 and for equation (9) it is 0.9999.

3. Monte Carlo experiments

The linearly and quadratically corrected estimators $\hat{g}_{lc}$ and $\hat{g}_{qc}$ are easy to compute and will have low bias when *N* is not too small. However, it is a question (i) which of the two estimators performs best; (ii) whether the estimators perform better than the recently developed GMM-estimators; (iii) whether the estimators perform satisfactorily in case *N* is relatively small. The questions can be addressed by performing Monte Carlo experiments. The Monte Carlo analysis are carried out using the same framework as used by Blundell and Bond (1998) and Kitazawa (2001).

[2] It is assumed that g is zero or positive. It is also possible to consider negative values of g as well. This would lead to different values of a_T, b_T, c_T, d_T and e_T. The linear and quadratic approximations remain very good. In most applications it can be considered prior knowledge that g is either zero or positive. Also, the extent of the bias of the LSDV-estimators is increasing in g.

The experiments have n_{it} and h_i drawn as mutually independent i.i.d. N(0,1) random variables. That is, $s_n^2 = s_h^2 = 1$. Kitazawa (2001) also uses different values of s_h^2 and finds GMM-estimators performing poorly in case of high s_h^2 / s_n^2-ratios (see his Table 3). Because in many applications the variance of individual effects will exceed that of the disturbance term, the strong upward bias of the GMM-estimators makes them less attractive. The LSDV-estimator is insensitive to the scaling of the individual effects, and, hence its performance is independent of the value of s_h^2. This very desirable property extends to the case of the linearly and quadratically corrected estimators $\hat{g}_{lc}$ and $\hat{g}_{qc}$.

Blundell and Bond (1998) use values of g equal to 0.0, 0.3, 0.5, 0.8 and 0.9 and values of *T* equal to 3 (results reported in their Table 2(a)) and 10 (results reported in their Table 2(c)). They use three different values for *N*, viz. 100, 200 and 500. Kitazawa (2001) uses values of g equal to 0.1, 0.3, 0.5, 0.7 and 0.9, *N* equal to 100 and a value of *T* equal to 6 (results reported in his Table 2). The Monte Carlo exercise will use these same values to allow for the performance of the estimators to be compared with the GMM-estimators of Blundell and Bond and Kitazawa. In addition Monte Carlo results for the case of g equal to 0.95 and for the cases of *N* equal to 5, 10, 20 and 50 are presented. The number of replications is 500 in each of the experiments.

Table 2 shows the mean and root mean squared error (RMSE) of the LSDV, the linearly corrected and quadratically corrected estimators for each of the cases reported in Blundell and Bond, Tables 2(a) and 2(c). The means of $\hat{g}_{lc}$ and $\hat{g}_{qc}$ are very close to the value of g used in the experiments. Each of the thirty experiments has a difference between the means of either of the two estimates and g less than 0.02. This confirms that the estimators are nearly unbiased, at least for large *N*. When we compare the RMSE of $\hat{g}_{lc}$ and $\hat{g}_{qc}$ with the corrsponding values for the Blundell and Bond, GMM2 (ALL)-estimator, we find that (i) the RMSE of $\hat{g}_{qc}$ is lower than that of the GMM2 (ALL)-estimator in all thirty cases; (ii) the RMSE of $\hat{g}_{lc}$ is lower than that of the GMM2 (ALL)-estimator in twenty-five cases, with the main exception being the experiments with g equal to zero; (ii) the relative performance in terms of RMSE of the GMM (ALL)-estimator is relatively poor when compared to $\hat{g}_{lc}$ and $\hat{g}_{qc}$ for values of g close to one.

The first five lines of Table 3 show the mean and root mean squared error of the LSDV and the two corrected estimators for the cases reported in Table 2 of Kitazawa (2001). The various GMM-estimators examined by Kitazawa show strong biases with the exception of the GMM (SYS)-estimator. Both $\hat{g}_{lc}$ and $\hat{g}_{qc}$ have lower RMSE than this GMM (SYS)-estimator. The rest of Table 3 shows Monte Carlo experiments for lower values of *N* and *T* equal to six. The near unbiasedness of the estimators appears to suffer from a decrease in *N* only in a limited way. For the quadratically corrected estimator $\hat{g}_{qc}$ the difference between the average of the estimates and the value of g remains less than 0.03 except for the case of N equal to five and g equal to 0.95. For the linearly corrected estimator $\hat{g}_{lc}$ this difference exceeds 0.03 for g equal to 0.9 or higher already for the case of N equal to 10. However, it is safe to conclude that the near unbiasedness remains present for the full range of (positive) values of g when N is as small as twenty. In addition, in case g is less than 0.9, the near unbiasedness remains present when N is as small as five.

The difference between the performance of the linearly and quadratically adjusted estimators is very limited for most cases. In general, the quadratically adjusted estimator performs somewhat better, in terms of RMSE, for values of g between zero and one half, while the reverse is the case for values of g between one half and unity. In terms of the average bias, the quadratically adjusted estimator performs somewhat better than the linearly adjusted one in case N is relatively small.

4. Conclusion

This paper introduces two easy to calculate estimators with desirable properties for the autoregressive parameter in dynamic panel data models. The estimators are (nearly) unbiased, outperform GMM-estimators and perform satisfactorily even for small samples in either the time-series or cross-section dimension. The paper does not have exogenous variables incorporated into the model and future research should seek to find the most adequate way to introduce them so as to find the nearly unbiased estimators in a more general setting. The current contribution is important for two reasons. First, the model without exogenous variables is of interest of itself because it is used in many empirical research efforts. Hence, it is important to have sound estimators for this case. Second, the very finding of the bias-corrected estimators to outperform the GMM-estimators in the simple case without exogenous variables may indicate that such estimators would also be strong competitors for the GMM-estimators in case of such variables included.

Table 1: Linear and quadratic approximations of the *f*-function.

T	a_T	b_T	R^2_{linear}	c_T	d_T	e_T
3	0.565	1.716	0.9999	0.561	1.726	0.120
4	0.370	1.540	0.9995	0.365	1.508	0.201
5	0.268	1.426	0.9992	0.264	1.358	0.221
6	0.207	1.349	0.9990	0.207	1.259	0.217
7	0.168	1.294	0.9990	0.170	1.193	0.205
8	0.140	1.252	0.9990	0.145	1.147	0.191
9	0.121	1.221	0.9990	0.127	1.115	0.176
10	0.105	1.195	0.9991	0.113	1.091	0.163
11	0.094	1.175	0.9992	0.102	1.074	0.150
12	0.084	1.158	0.9992	0.093	1.060	0.139
13	0.077	1.144	0.9993	0.085	1.050	0.129
14	0.070	1.132	0.9993	0.079	1.042	0.120
15	0.065	1.122	0.9994	0.074	1.036	0.112
16	0.060	1.113	0.9994	0.069	1.031	0.105
17	0.056	1.105	0.9995	0.065	1.027	0.099
18	0.053	1.098	0.9995	0.061	1.024	0.093
19	0.050	1.092	0.9996	0.058	1.021	0.088
20	0.047	1.086	0.9996	0.055	1.019	0.083
21	0.045	1.082	0.9996	0.052	1.017	0.078
22	0.042	1.077	0.9997	0.050	1.015	0.074
23	0.040	1.073	0.9997	0.048	1.014	0.071
24	0.039	1.070	0.9997	0.046	1.013	0.067
25	0.037	1.066	0.9997	0.044	1.012	0.064
26	0.036	1.063	0.9997	0.042	1.011	0.061
27	0.034	1.061	0.9998	0.041	1.010	0.058
28	0.033	1.058	0.9998	0.039	1.009	0.056
29	0.032	1.056	0.9998	0.038	1.009	0.053
30	0.031	1.053	0.9998	0.037	1.008	0.051

Note: values of the constants are based upon the values of g for 0.000 (0.001) 0.999 and the implied $\hat{g}$ based upon "Nickell bias".

Table 2: Monte Carlo results for the LSDV-estimator and the unbiased estimators

T	N	gamma	LSDV mean	LSDV rmse	$LSDV_{lc}$ mean	$LSDV_{lc}$ rmse	$LSDV_{qc}$ mean	$LSDV_{qc}$ rmse
3	100	0.0	-0.3342	0.3390	-0.0086	0.0976	-0.0021	0.0932
3	100	0.3	-0.1622	0.4668	0.2867	0.1130	0.2848	0.1115
3	100	0.5	-0.0336	0.5380	0.5073	0.1178	0.5036	0.1178
3	100	0.8	0.1268	0.6770	0.7827	0.1248	0.7825	0.1279
3	100	0.9	0.1979	0.7054	0.9045	0.1171	0.9078	0.1211
3	200	0.0	-0.3330	0.3351	-0.0064	0.0658	-0.0002	0.0629
3	200	0.3	-0.1551	0.4574	0.2988	0.0788	0.2964	0.0777
3	200	0.5	-0.0339	0.5361	0.5068	0.0840	0.5029	0.0839
3	200	0.8	0.1372	0.6649	0.8004	0.0909	0.8004	0.0933
3	200	0.9	0.1908	0.7111	0.8924	0.0892	0.8949	0.0919
3	500	0.0	-0.3331	0.3340	-0.0066	0.0424	-0.0005	0.0401
3	500	0.3	-0.1562	0.4571	0.2970	0.0488	0.2945	0.0482
3	500	0.5	-0.0379	0.5387	0.5000	0.0505	0.4959	0.0507
3	500	0.8	0.1371	0.6636	0.8003	0.0536	0.8000	0.0549
3	500	0.9	0.1920	0.7088	0.8944	0.0565	0.8969	0.0582
10	100	0.0	-0.0987	0.1034	-0.0129	0.0392	0.0071	0.0336
10	100	0.3	0.1649	0.1392	0.3020	0.0400	0.2975	0.0384
10	100	0.5	0.3354	0.1675	0.5058	0.0375	0.4974	0.0374
10	100	0.8	0.5803	0.2217	0.7985	0.0352	0.8011	0.0377
10	100	0.9	0.6574	0.2443	0.8906	0.0365	0.9009	0.0385
10	200	0.0	-0.0993	0.1018	-0.0136	0.0302	0.0064	0.0247
10	200	0.3	0.1639	0.1380	0.3008	0.0274	0.2963	0.0265
10	200	0.5	0.3385	0.1632	0.5095	0.0289	0.5010	0.0275
10	200	0.8	0.5813	0.2196	0.7997	0.0240	0.8024	0.0258
10	200	0.9	0.6547	0.2461	0.8874	0.0273	0.8972	0.0266
10	500	0.0	-0.0998	0.1008	-0.0142	0.0221	0.0058	0.0161
10	500	0.3	0.1658	0.1350	0.3031	0.0178	0.2984	0.0168
10	500	0.5	0.3381	0.1626	0.5090	0.0198	0.5005	0.0177
10	500	0.8	0.5816	0.2189	0.8000	0.0166	0.8027	0.0180
10	500	0.9	0.6571	0.2433	0.8902	0.0179	0.9002	0.0163

Table 3: Monte Carlo results for the LSDV-estimator and the unbiased estimators

T	N	gamma	LSDV mean	LSDV rmse	$LSDV_{lc}$ mean	$LSDV_{lc}$ rmse	$LSDV_{qc}$ mean	$LSDV_{qc}$ rmse
6	100	0.1	-0.0838	0.1881	0.0939	0.0542	0.1034	0.0489
6	100	0.3	0.0667	0.2373	0.2970	0.0585	0.2924	0.0563
6	100	0.5	0.2258	0.2774	0.5116	0.0575	0.5072	0.0567
6	100	0.7	0.3677	0.3353	0.7030	0.0604	0.6997	0.0635
6	100	0.9	0.5017	0.4006	0.8838	0.0603	0.8936	0.0638
6	100	0.95	0.5389	0.4134	0.9339	0.0599	0.9488	0.0638
6	50	0.1	-0.0833	0.1919	0.0946	0.0768	0.1043	0.0696
6	50	0.3	0.0713	0.2362	0.3032	0.0800	0.2987	0.0765
6	50	0.5	0.2201	0.2867	0.5040	0.0841	0.4955	0.0844
6	50	0.7	0.3613	0.3440	0.6944	0.0812	0.6910	0.0854
6	50	0.9	0.5030	0.4013	0.8856	0.0811	0.8960	0.0874
6	50	0.95	0.5334	0.4210	0.9265	0.0851	0.9410	0.0908
6	20	0.1	-0.0850	0.2058	0.0923	0.1219	0.1033	0.1104
6	20	0.3	0.0759	0.2410	0.3094	0.1198	0.3055	0.1145
6	20	0.5	0.2173	0.2978	0.5002	0.1264	0.4928	0.1269
6	20	0.7	0.3633	0.3496	0.6971	0.1269	0.6950	0.1332
6	20	0.9	0.4967	0.4157	0.8771	0.1380	0.8881	0.1485
6	20	0.95	0.5353	0.4246	0.9291	0.1248	0.9449	0.1355
6	10	0.1	-0.0880	0.2261	0.0882	0.1698	0.1013	0.1536
6	10	0.3	0.0670	0.2671	0.2973	0.1761	0.2960	0.1679
6	10	0.5	0.2177	0.3118	0.5007	0.1788	0.4952	0.1788
6	10	0.7	0.3529	0.3747	0.6831	0.1912	0.6827	0.1992
6	10	0.9	0.4894	0.4325	0.8671	0.1862	0.8791	0.2004
6	10	0.95	0.5213	0.4484	0.9102	0.1818	0.9260	0.1962
6	5	0.1	-0.0779	0.2522	0.1019	0.2411	0.1171	0.2187
6	5	0.3	0.0496	0.3127	0.2739	0.2540	0.2776	0.2401
6	5	0.5	0.2012	0.3626	0.4784	0.2779	0.4782	0.2772
6	5	0.7	0.3447	0.4016	0.6720	0.2540	0.6744	0.2629
6	5	0.9	0.4847	0.4498	0.8608	0.2361	0.8746	0.2532
6	5	0.95	0.4944	0.4979	0.8739	0.2815	0.8912	0.2979

References

Andrews, D.W.K. (1993), Exactly median-unbiased estimation of first-order autoregressive/unit-root models, *Econometrica* 61, 139-165.

Blundell, R., Bond, S., 1998, Initial conditions and moment restrictions in dynamic panel data models, *Journal of Econometrics* 87, 115-143.

Cermeño, R., 1999, Median-unbiased estimation in fixed-effects dynamic panels, *Annales d'Économie et de Statistique* 55-56, 351-368.

Collado, M.D., 1997, Estimating dynamic models from time series of independent cross-sections, *Journal of Econometrics* 82, 37-62.

Hsiao, C., 1986, *Analysis of Panel Data*, Cambridge: Cambridge University Press.

Kitazawa, Y., 2001, Exponential regression of dynamic panel data models, *Economics Letters* 73, 7-13.

Kiviet, J.F., 1995, On bias, inconsistency, and efficiency of various estimators in dynamic panel data models, *Journal of Econometrics* 68, 53-78.

MacKinnon, J.G., Smith, Jr., A.A., 1998, Approximate bias correction in econometrics, *Journal of Econometric*s 85, 205-230.

Nickell, S., 1981, Biases in dynamic models with fixed effects, *Econometrica* 49, 1417-1426.