Least Squares Dummy Variable in Determination of Dynamic Panel Model Parameters

Joseph U. Okeke and Evelyn N. Okeke

Abstract—This paper investigates the small sample performance of the Least Squares Dummy Variable (LSDV) estimator of the dynamic panel data models for period, T , greater than the cross sections, N and its large sample performance in the direction of T as N remains finite, and compares it with the performance of the instrumental variable-generalized method of moments (IV-GMM) estimators using the properties of root mean squares error (RMSE) of the model, root mean squares error of the autoregressive term γ ($RMSE_\gamma$), the bias of γ ($bias_\gamma$) and the Akaike Information Criterion (AIC) with the motive of ascertaining the usefulness of the LSDV estimator in determining the parameters of a dynamic panel model as $T \rightarrow \infty$ and finite N , for which it is regarded as consistent.

Index Terms—Autoregressive term; dynamic panel model; consistent; least squares dummy variable.

I. INTRODUCTION

A Panel dataset aside from providing a cure to autocorrelation in time series dataset also provides the milieu for studying the behavior of cross-sectional units (N) observed over time periods (T). The dynamic panel data model provides a solution for accommodating the joint occurrence of dynamics and unobserved individual heterogeneity in the phenomena of interest. The inconsistency of the Least Squares methods: Ordinary Least Squares (OLS), Least Squares Dummy Variable (LSDV) and; the Maximum Likelihood Estimators (MLE), in determining the parameters of the dynamic panel data model has led to the development of a range of new estimators notably the Instrumental Variable- Generalized Method of Moments Estimators, that is, IV-GMM methods notably: Anderson-Hsaio (AH), Arellano and Bond (AB) and, Blundell and Bond (BB) methods.

The IV-GMM methods provides consistent and asymptotically efficient estimators for dynamic panel models [1] and [2]. However, The IV-GMM estimators are biased and inefficient in small samples [3], making them not to be the panacea to the problem of modeling a dynamic panel data. By small sample, we mean samples with finite time and cross-sections or long time and finite cross sections. The biasness and inconsistency of the LSDV estimator is occasioned by the introduction of a lagged endogenous explanatory variable which introduces autocorrelation into the model, a condition that violates the assumptions necessary for the LSDV estimator to be consistent. The biasness and inconsistency of a model would

be a great problem if the aim of the model is statistical inference rather than forecasting. Forecasting model is a potentially important tool in enhancing early warning capacity [4]. The desire to achieve unbiasedness and consistency leads to the tradeoff of efficiency for unbiasedness since LSDV estimator have minimum variance and the cost is in terms of reduced efficiency and forecast accuracy [5] and [6].

The inconsistency of the LSDV estimator is established when the cross-section size is large and the time is finite [5]. [7] and [8] derives an approximation for the inconsistency of the LSDV as $N \rightarrow \infty$ which is bounded of order T^{-1} , $N^{-1}T^{-1}$ and $N^{-1}T^{-2}$ respectively. [9] analyzed the accuracy of [8] approximation then [3] extended the inconsistency approximation in [9] to accommodate unbalanced panels. All efforts to derive an expression for the bias of the LSDV estimator is borne out of the desire for remedy since the LSDV estimator though inconsistent in large number of cross sections is consistent as time becomes infinitely large.

This study compares the LSDV estimator and the IV-GMM estimators in determining the parameters of the dynamic panel model when there exists cross section heterogeneity which is fixed for each individual as an extension to [9], [10] and [11].

II. THEORETICAL BACKGROUND

A. Panel Data Models

A panel is a cross-section or group that are surveyed over a given period of time. It is two dimensional: space dimension and time dimension. A panel data comprises a cross-sectional data that are repeated over some time which may include longitudinal data on same individual across time or data representing observation with three or more dimensions (e.g. temporal, spatial and individual). A panel data with more than two dimensions is known as multi-dimension panel data.

In panel studies, the same individuals or units are observed at a series of discrete points in time [6] and this gives rise to panel data. Other names for panel data are pooled data, micropanel data, longitudinal data (a study made over time of variables – or group of subjects), event history analysis (e.g. studying the movement over time of subjects through successive states or conditions), cohort analysis (e.g. following the health development of children born in a certain year). In marketing, panel data refers to data collected at the point of sales and is known as Scanner data.

In this research, a panel data would be viewed as a two-dimensional data with time series and cross-sectional dimension. A panel data could be balanced or unbalanced.

A panel data is balanced if the same length of periods is observed for every unit of the cross-sections while it is unbalanced if varied length of periods are observed for some units of the cross-sections.

From the basic framework stated in (1) below, if α_i is observed for all individuals, then the model can be treated as an ordinary linear model and fit by least squares as a pooled regression model. However, if it were unobserved, various cases would be considered such as: Fixed effect model, random effect model and Random parameter model.

B. Fixed Effect

If α_i is unobserved, but correlated with the regressor, then the least squares estimator of the parameter is biased and inconsistent as a consequence of any omission of variable (s).

However, given

$$Y_{it} = X'_{it}\beta + \alpha_i + \varepsilon_{it} \quad (1)$$

where α_i in the basic frame work, embodies all the unobservable effects and specifies an estimable conditional mean. α_i is a group specific constant term which does not vary over time but varies from individual to individual or groups,

we may write the fixed effect model as

$$Y_i = X_i\beta + i_h\alpha_i + \varepsilon_i, i_h = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix} \quad (2)$$

or more compactly using dummy variable as

$$Y = [X \ d_1 \ d_2 \ \dots \ d_N] \begin{bmatrix} \beta \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{bmatrix} + \varepsilon \quad (3)$$

where, for $i=1,2,\dots,N$, d_i is a dummy variable that indicates the i th individual.

If we let the $NT \times N$ matrix $D = [d_1 \ d_2 \ \dots \ d_N]$ and assemble all NT rows we shall obtain,

$$y = x\beta + D\alpha + \varepsilon \quad (4)$$

If N is small enough, then the model can be estimated by OLS with k regressors in X and N columns of D ; as a multiple regression with $k+N$ parameters. Then the least square estimator of β will be

$$\hat{\beta} = (X' W_D X)^{-1} (X' W_D y) \quad (5)$$

$$X_a = W_D X \text{ and } y_a = W_D y \quad (6)$$

where, $W_D = 1 - D(D'D)^{-1}D'$

(5) is the least squares regression model on the transformed data [6]

$$\text{Or } W_D = \begin{bmatrix} w^0 & 0 & 0 & \dots & 0 \\ 0 & w^0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & w^0 \end{bmatrix} \quad (7)$$

where $w^0 = I_T + \frac{1}{T} i_h i_h'$

C. Random Effect

A model is a random effect model if the unobserved individual heterogeneity, can be assumed to be uncorrelated with the regressors, then the model may be formulated as

$$y_{it} = X'_{it}\beta + E[\alpha_i] + [\alpha_i E[\alpha_i]] + \varepsilon_{it} \quad (8)$$

$$= X'_{it}\beta + \alpha + \mu_i + \varepsilon_{it}$$

where the compound disturbance term $(\mu_i + \varepsilon_{it})$ may be consistently but inefficiently estimated by the least squares. μ_i is a group specific random factor similar to ε_{it} except that for each group an identical draw is made in each period of time. By implication, the group specific effect terms are random variables under random effect.

D. Random Parameters

Extending our idea of random effect model as a model with random constant terms, we may have

$$y_{it} = X_{it}(\beta + h_i) + (\alpha + \mu_i) + \varepsilon_{it} \quad (9)$$

where h_i is a random vector which induces variation of parameter across individuals. Hence, with a sufficiently rich data set, the model coefficient may vary across individuals giving rise to random parameter models.

E. Dynamic Panel Model

The dynamic panel data model to be considered as in [12] is

$$y_{it} = \gamma y_{it-1} + X'_{it}\beta + \alpha_i + \varepsilon_{it}, \quad \gamma < 1, i = 1, 2, \dots, N \text{ and } t = 1, 2, \dots, T \quad (10)$$

where: y_{it} is the endogenous response variable; y_{it-1} is the immediate past value of y_t i.e. lag one period;

X_{it} is a $((k-1) \times 1)$ vector of strictly exogenous explanatory variables; α_i is the unobserved group effects and; ε_{it} is the unobserved random error terms.

Least Squares Dummy Variable (LSDV)

In dynamic panel data models, dummy variables may be introduced to the least squares to explain the effect of each individual unit of a cross section which is unobserved but correctly specifies the model of relation.

Just like the OLS, the LSDV is also applied to the equations in level form and all the T cross –section are applied in the actual estimation. [4] and [6]. It can give estimates of variances of α_i and ε_{it} separately.

In the LSDV estimation the individual effect is assumed to be fixed over time in each individual. The fixed effects model is a useful specification for explaining cross section heterogeneity in panel data. However, in small sample case i.e. short time period, the LSDV estimator is inconsistent owing to the incidental parameters problem. ‘The seriousness of this problem in practical terms remains to be established as there exist only a very small amount of received evidence but the theoretical result is unambiguous’ [13]. The LSDV is generally implemented by the insertion of relevant dummies but being mindful of the dummy

variable trap and application of OLS on the enlarged model. Computationally, it is simpler to obtain LSDV through within estimation. [6].

Estimation of LSDV by Within – Group Variation

This involves averaging the original equation

$y_{it} = X_{it} \beta + \gamma y_{it-1} + \alpha_i + \varepsilon_{it}$ over time and centering the observation such that

$$y_{it} - \bar{y}_i = \gamma(y_{it-1} - \bar{y}_{i-1}) + (X_{it} - \bar{X}_i)' \beta + (\varepsilon_{it} - \bar{\varepsilon}_i).$$

11

$$\text{where } \bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}, \quad \bar{y}_{i-1} = \frac{1}{T} \sum_{t=2}^T y_{it-1} \quad \bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it} \quad \text{and} \quad \bar{\varepsilon}_i = \frac{1}{T} \sum_{t=1}^T \varepsilon_{it}$$

OLS estimation of the above equation amount to the LSDV estimation of the original model. To derive the estimate of this model, we denote

$$\tilde{y}_i = (y_{it} - \bar{y}_i), \quad \tilde{y}_{i-1} = (y_{it-1} - \bar{y}_{i-1}) \quad \text{and} \quad \tilde{x}_{it} = (x_{it} - \bar{x}_i)$$

Then,

$$\begin{aligned} \tilde{Y}_{it} &= [\tilde{y}_{11}, \tilde{y}_{21}, \dots, \tilde{y}_{N1}, \dots, \tilde{y}_{1T}, \tilde{y}_{2T}, \dots, \tilde{y}_{NT}]' \\ \tilde{Y}_{-1} &= [\tilde{y}_{10}, \tilde{y}_{20}, \dots, \tilde{y}_{N0}, \tilde{y}_{11}, \tilde{y}_{21}, \dots, \tilde{y}_{N1}, \dots, \tilde{y}_{1T-1}, \tilde{y}_{2T-1}, \dots, \tilde{y}_{NT-1}]' \\ \tilde{X}_{it} &= [\tilde{x}_{11}, \tilde{x}_{21}, \dots, \tilde{x}_{N1}, \tilde{x}_{12}, \tilde{x}_{22}, \dots, \tilde{x}_{N2}, \dots, \tilde{x}_{1T}, \tilde{x}_{2T}, \dots, \tilde{x}_{NT}]' \end{aligned}$$

Therefore for $\tilde{W} = [\tilde{Y}_{-1} \tilde{X}_{it}]$, the LSDV estimation is given by

$$\delta = (\tilde{W}' \tilde{W})^{-1} \tilde{W}' y \quad (12)$$

Then

$$\text{cov}[(y_{i-1} - \bar{y}_{i-1})(\varepsilon_{it} - \bar{\varepsilon}_i)] \neq 0 \quad (13)$$

where $\tilde{Y}_{-1}$ is the matrix of the lagged variable.

The bias of LSDV estimation which is inconsistent in large N but consistent for $T \rightarrow \infty$ is thereby established by (13).

[14], [13] and [10]. Under the assumption of normality of the random error term and homoscedasticity, the standard errors of the LSDV estimates are obtained from

$$\text{Var}(\delta) = \tilde{S}^2 (\tilde{W}' \tilde{W})^{-1} \quad (14)$$

where $\tilde{S}^2 = \tilde{\varepsilon}' \tilde{\varepsilon} / (NT - N - 2)$

and the bias in the direction of N is

$$\tilde{\varepsilon} = (\tilde{y} - \tilde{W} \delta). \quad (15)$$

The bias approximation of the LSDV as estimated by [7] and [8] culminated to the emanation of the following as in [9]: $\beta_1 = c_1 \bar{T}^{-1}$, $\beta_2 = \beta_1 + c_2 N^{-1} \bar{T}^{-1}$ and $\beta_3 = \beta_2 + c_3 N^{-1} \bar{T}^{-1}$, which depended upon the unobserved parameters σ_ε^2 and γ . A consistent estimator of σ_ε^2 can then be derived from e_q from initial estimators of choice vis-a-vis: AH, AB and BB which are cumbersome, except with the aid of software packages, to obtain

$$\hat{\sigma}_q^2 = \frac{e_q' A_s e_q}{N-K-T} \quad (16)$$

where $e_q = y - W\delta_q$, and $q = \text{AH, AB or BB}$.

[10] established that the particular form of the inconsistency of LSDV estimator does not change in case of cross sectional heteroscedasticity and hence paying special attention to the estimation errors of γ and β we obtained their bias as:

$$\tilde{y} - \gamma = (\tilde{y}'_{-1} M \tilde{y}_{-1})^{-1} \tilde{y}'_{-1} M \tilde{\varepsilon} \quad (17)$$

$$\tilde{\beta} - \beta = -(\tilde{X}' \tilde{X})^{-1} \tilde{X}' \tilde{y}_{-1} (\tilde{y} - \gamma) + (\tilde{X}' \tilde{\beta})_{-1} \tilde{X}' \tilde{\varepsilon} \quad (18)$$

where $M = 1 - (\tilde{X}' \tilde{X})_{-1} \tilde{X}'$

III. SAMPLING DESIGN OF EXPERIMENTS

The sampling design of the experiment for the estimation and comparisons of the LSDV estimator and the IV-GMM methods adheres strictly to the specification of the dynamic panel data model given below:

$$Y_{it} = \gamma Y_{it-t} + X_{it}' \beta + \alpha_i + \varepsilon_{it} \quad (19)$$

$$X_{it} = \rho X_{it-1} + e_t, \quad e_t \sim N(0, 1), \quad |\rho| < 1 \quad (20)$$

$Y_{i0} = \eta_0 + \eta_1 \alpha_i + \eta_2 \varepsilon_{i0}$, $X_{i0} = \lambda_0 + \lambda_1 \varepsilon_{i0}$, where $\varepsilon_{it} = \rho \varepsilon_{it-1} + e_t$, $i=1,2,\dots,N$ and $e_t \sim N(0,1)$

The exogenous variable x_{it} with the specification $X_{it} = \rho X_{it-1} + e_t$, $e_t \sim N(0, 1)$, $|\rho| < 1$ was generated as in [12] and [15] by specifying $\rho = 0.8$. Y_{i0} and X_{i0} are generated using the [16] procedures.

The model has two error components: the individual effect term α_i and the random error component ε_{it} .

The individual effect term is specified as $\alpha_i = Q_1 \bar{X}_{it} + e_t$, $e_t \sim N(0,1)$, we specify $Q_1 = 0.8$ as in [17]. Using the average X_{it} for each individual ($\bar{X}_i$) and the standardized error term generated for each individual, we generate the individual effect term α_i , an $(N \times 1)$ vector. [11] has shown that the specification of the individual effect term as $\alpha_i = Q_1 \bar{X}_{it} + e_{it}$, $e_{it} \sim N(0,1)$ or $\alpha_i \sim \text{iid}(0, \sigma_\alpha^2)$, $\sigma_\alpha = \sigma_\varepsilon(1 - \gamma)$ does not affect the ranking of estimators performance. The random error term ε_{it} is generated by $\varepsilon_{it} = \rho \varepsilon_{it-1} + \varepsilon_t$, $\varepsilon_t \sim N(0,1)$. The values of γ, ρ and β are fixed at 0.8, 0.8 and 0.2 respectively to provide a basis for comparison since the IV-GMM methods are deemed efficient at high values of the autoregressive coefficients of the dynamic panel model. The experiment is replicated two hundred and fifty times.

IV. RESULTS AND DISCUSSIONS

Since the IV-GMM estimators are specifically designed for highly persistent series, we restrict our comparisons on the estimators to $\gamma=0.8$ and $\rho=0.8$ in line with Bruno (2004). Results for RMSE, AIC, RMSE(γ) and Bias(γ) are presented in table 1 below at $(T,N) = (20,10)$ and $(50,10)$. At this point we tried to show the performances of the LSDV, AH, AB and BB in terms of these parameters when T and N are small but $T \geq N$ and when T is slightly large ($T=50$) and N is finite ($N=10$).

In the general assessment of the models measured by the RMSE and AIC, the BB showed the least score compared with the others at RMSE=0.3656 and AIC=0.3148. However, the literature does not give the BB the chance of converging asymptotically at finite N. The LSDV estimator showed the least score in terms of their RMSE(γ)=0.0016 and bias(γ)=-0.0261 as T becomes larger(T=50) and N remains small (N=10) while AB showed the least score of

RMSE(γ)=0.0006 and bias(γ)=-0.0250 at small T (T=20) and small N(N=10) but $T \geq N$. The results of the performance of the LSDV in the direction of T seem to agree with [10], [12] and [6], but the concern is on the high value of the RMSE=0.9255 of the LSDV model relative to the least score of RMSE=0.1348 recorded by the BB estimator, though BB is inconsistent as $T \rightarrow \infty$.

TABLE IA: VALUES OF ESTIMATION PARAMETERS AT VALUES OF T AND N FOR $\Gamma=0.8$ AND $\rho=0.8$ AND $\beta=0.2$

ESTIMATORS	SPEC.	RMSE	AIC		
	T	20	50	20	50
	N	10	10	10	10
LSDV		0.1525	0.0982	0.0237	0.0925
AH		0.3844	0.1508	0.1510	0.0899
AB		17.9622	0.0368	322.58	0.0137
BB		0.4252	0.0366	0.1846	0.0135

TABLE IB: CONTINUATION OF TABLE IA

ESTIMATORS	SPEC.	RMSE(γ)	BIAS(γ)		
	T	20	50	20	50
	N	10	10	10	10
LSDV		0.0280	0.0016	0.0237	0.0925
AH		0.0113	0.5616	0.1510	0.0899
AB		0.0006	0.1192	322.58	0.0137
BB		0.0377	0.1351	0.1846	0.0135

ACKNOWLEDGMENT

We wish to extend our sincerely appreciation to everyone that contributed by way of questions or suggestions towards the success of this research and most especially to all the authors whose works formed the bedrock of this work as contained in the references for their comments and contributions in the subject matter that provoked our attention in this direction. All your guidance is gratefully acknowledged.

REFERENCES

- [1] M. Arellano and S. Bond, "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations," *Review of Economic Studies*, vol. 58, pp. 277-297, 1991.
- [2] J. Hahn and G. Kuersteiner, "Asymptotically unbiased inference for a dynamic panel model with fixed effects when both N and T are large," *Econometrica*, vol. 70, pp. 1639-1657, 2002.
- [3] G.S.F. Bruno, G "Approximating the Bias of the LSDV Estimator for Dynamic Unbalanced Panel Data Models," *Economics Letters*, Milan, vol. 87, pp. 361-366, 2005.
- [4] N. Islam, "Small sample performance of dynamic panel data estimators: a Monte Carlo study on the basis of growth data," Seminar paper at Harvard University and Emory University, 1998.
- [5] D.N. Gujarati, *Basic of Econometrics*. 4th edition. New York. Tata McGraw-Hill, 2003, 636-719, ch.
- [6] W. H. Greene, *Econometric Analysis*. 4th edition. Engle-wood cliffs. Prentice Hall, 2003, 230-310, ch.
- [7] S. Nickel, "Biases in dynamic models with fixed effects," *Econometrica*, Vol. 49, pp. 1417-1426, 1981.
- [8] J. F. Kiviet, "On bias, inconsistency and efficiency of various estimators in dynamic panel data models," *Journal of Econometrics*, vol. 68, pp. 53-78, 1999
- [9] M.J.G. Bun and J. F. Kiviet, "On the diminishing returns of higher order terms in asymptotic expansions of bias," *Economics Letters*, vol. 79, pp. 145-152, 2003
- [10] M. J. G. Bun and M. A. Carree, "Bias-corrected Estimation in Dynamic Panel Data Models," *Journal of Business & Economic Statistics*, Amsterdam vol. 23, pp. 200-210, 2005.

- [11] J. U. Okeke, J. C. Nwabueze and E. N. Okeke, "Individual effects in first order autoregressive dynamic panel data models," *International Journal of Science: Basic and Applied Research*, vol.19, no. 1, pp. 279-284, 2015
- [12] G. S. F. Bruno, "Estimation, inference and Monte Carlo analysis in dynamic panel data models with a small number of individuals," *Istituto di Economia Politica, Università Bocconi, Milan*, 2004
- [13] W. Greene, "Estimating econometric models with fixed effects," *Department of Economics, Stern School of Business. New York University*, 2001.
- [14] T. Amemiya, "A note on the estimation of Balestra-Nerlove's model," technical reports no. 4. Institute of Mathematical Studies in Social Science, Stanford University, Stanford, 1967.
- [15] J. F. Kiviet, "Judging contending estimation by simulation: tournaments in dynamic panel data models," Tinbergen Institute discussion paper, University of Amsterdam, Netherlands, 2005
- [16] A. I. McLeod, and K. W. Hipel, "Simulation procedure for Box-Jenkins models," *Water Resources Research*, vol. 14, pp. 969-975, 1978.
- [17] J. J. Spitzer, "Small sample properties of nonlinear least squares and maximum likelihood estimators in the context of auto-correlated errors," *Journal of American Statistical Association*, vol. 74, pp. 126-138, 1979.

J. U. Okeke was born in Asaba, Delta State, Nigeria, on the 14th May, 1971. He holds: PhD Statistics (2011) from ABSU, Abia, Nigeria; M.Sc. Statistics (2005) from NAU, Anambra, Nigeria and B.Sc. Statistics (1997). His major field of study is Econometric statistics with stint in multivariate statistics which was his area of research at his master's thesis.

He was a LECTURER at the Anambra State University now Chukwuemeka Odumegwu Ojukwu University (2007-2013). Presently, he lectures in the Department of Mathematics and Statistics of the Federal University Wukari, Taraba State, Nigeria. He has published in both local and foreign reputable journals. His research interests are in the areas of Econometric dynamic modeling and Multivariate classification modeling.

Dr. Okeke is a member of the Nigerian Statistical Association, a consultant with the United Nations Development Program, the Secretary, Anambra West Elite Club (2012 to date), the seminar coordinator, Faculty of Pure and Applied Sciences, Federal University Wukari, Nigeria.

E. N. Okeke was born in Obeledu, Anambra State, Nigeria, on the 16th July, 1971. She holds: PhD Statistics (2011) from ABSU, Abia, Nigeria; M.Sc. Statistics (2002) from NAU, Anambra, Nigeria and B.Sc. Statistics (1997) from NAU, Anambra, Nigeria. Her major field of study is Multivariate statistics(Discrimination and classification) and also has interest in Econometric modeling.

She was a LECTURER at the Nnamdi Azikiwe University (NAU, Awka) 2001-2013. Presently, she lectures in the Department of Mathematics and Statistics of the Federal University Wukari, Taraba State, Nigeria. She has published in both local and foreign reputable journals. Her research interests is in the area of Discriminant Analysis.

Dr.Mrs. Okeke is a member of the Nigerian Statistical Association, a consultant with the United Nations Development Program, the Head, Department of Mathematics and Statistics, Federal University Wukari, Nigeria. Chairperson Welfare committee, Faculty of Pure and Applied Sciences, Federal University Wukari, Nigeria.