MONTE CARLO EXPERIMENT ON THE PERFORMANCES OF PANEL DATA ESTIMATORS IN A ONE WAY ERROR COMPONENT

Supported by tetfund

A. H. Bello
Department of Statistics, Federal University of Technology, PMB 704, Akure, Ondo State, Nigeria
habello@futa.edu.ng

Received: December 29, 2019 **Accepted:** February 20, 2020

Abstract: This research is on the behavior of panel data estimators in a simulated study. The major objective is to investigate these estimators on individual effect and other remaining disturbance term in one way error component. The estimators considered were pooled OLSE, within effect and between effect estimators. The data was simulated using R Statistical software after considering all statistical properties with sample size of N=20, N=50 and fixed time T=10. The methods of validation of result include R-squared, Root Mean Square Error (RMSE) and Variance. The results of the analysis reveals that at sample size N=20 and T=10, the absolute bias of the estimators indicate that the pooled OLSE is better than all other estimators. It also reveals that within effect estimator better explain the fitness of the model due to the highest R-squared value. The between effect estimator perform better than other estimators because it has the least Variance values. The between estimator is also more efficient than other methods of estimation as its value of variance and RMSE of 0.9710397 and 0.9854 are respectively low. As the sample size increase to N=50, the results still remain the same. In conclusion the between estimator is the best and most efficient estimator among all other estimators considered.

Keywords: Absolute bias, RMSE, OLSE, R^2-Coefficient of determination, panel data

Introduction

Panel Data

Panel data is a kind of data in which observations are obtained on the same set of entities at several periods of time. A panel dataset is one where there are repeated observations on the same units. The units may be individuals, households, firms, regions or countries. It has the combination of the characteristics of both time Series and cross-sectional data. There are two kinds of information in cross-sectional time-series data: the cross-sectional information reflected in the differences between subjects, and the time-series or within-subject information reflected in the changes within subjects over time. Panel data regression techniques allow you to take advantage of these different types of information.

The use of panel data in applied research is increasingly gaining relevance as follows:

1. Panel data provide sufficient observations and, consequently, more sample variability, less collinearity, more degrees of freedom, and more accurate inference of model parameters. However, in the case of panel data, like-wise providing more observations and more sample variability than either cross- sectional data or time series data alone
2. In connection with (1), panel data models better capture the complexity of human behavior than a single cross-section or time series data. For example, consider a cross- sectional sample of university students with an average grade of 50% in all the courses for the same period in time. This suggests that every student has the chances of having 50% grade based on information obtained from their performance at a particular year/level of academic study. Thus, current information about a student's academic performance is a perfect predictor of his/her future performance. However, sequential observations for students contain information about their academic performance in different years/levels of academic study that are captured in the cross-sectional framework. With panel data models, performance of each student can be observed over time and more informed judgments can be made.
   Similarly, consider a time series sample of a student's academic performance. Generalizing with the information obtained from the student will lead to unbiased and inefficient estimates. Therefore, by pooling a cross-sectional sample of students over time, variations in each unit over time are captured.
3. In connection with (2), panel data models are better able to capture the heterogeneity inherent in each individual unit because the structure of panel data suggests that the cross- sectional units whether individuals, firms, states or countries are heterogeneous. In empirical modelling, ignoring these heterogeneous effects when in fact they exist leads to biased and inefficient results. We can illustrate this with an empirical example using Baltagi and Levin (1992) paper. They consider cigarette demand across 46 American states for the years 1963-88. Consumption is modelled as a function of lagged consumption, price and income. They however note that there are a lot of variables that may be state-invariant or time-invariant that may affect consumption. Examples of these state invariant variables are advertisement on nationwide television and radio and national policies while time-invariant variables are religion and education.

The purpose of this paper is to undertake an extensive investigation of three different estimation methods for panel data selection models by a Monte Carlo Experiment. The three estimators considered are; Pooled OLSE, Within Estimator and Between Estimator. It will also give the opportunity to assess the performances of estimators in one-way error component on both individual effects and other remainder disturbance term under repetitive sampling distribution properties.

Although, random coefficients can be used in the estimation and specification of panel data models, see Swamy (1971), Hisao (1986) and Dielman (1989), most panel data applications have been limited to a simple regression with error components disturbances;

$y_{it} = x'_{it}\beta + \mu_i + \lambda_t + v_{it} \quad i = 1 \ldots\ldots\ldots, N;\ t = 1, \ldots, T$ ……eqn (1)

where i denotes individuals and t denotes time, x_{it} is a vector of observations on k explanatory variables, β is a k vector of unknown coefficients, μ_i is an unobserved individual specific effect, λ_t is an unobserved time specific effect v_{it} is a zero mean random disturbance with variance $\sigma^2 v$. The error components disturbances follow a two-way analysis of variance (ANOVA). If μ_i and $\lambda_t$denote fixed

parameters to be estimated, this model is known as the fixed effects (FE) model. The $x'_{it}s$ are assumed independent of the v_{it} for all i and$t$. Inference in this case is conditional on the particular N individuals and over the specific time-periods observed. Estimation in this case amounts to including $(N-1)$ individual dummies and $(T-1)$ time dummies to estimate these time invariant and individual effects. This leads to an enormous loss degrees. In addition, this attenuates the problem of multicollinearity among the regressors. Furthermore, this may not be computationally feasible for N or T. In this case, one can eliminate the $\mu_i's$ and $\lambda_t's$ and estimate β by running least squares of y_{it} =$y_{it} - y_{i.} - y_{.t} + y_{..}$ on the $x'_{it}s$ similarly defined, where the dot indicates summation over that index and the bar denotes averaging. This transformation is known as the within transformation and the corresponding estimator of β is called the within estimator or the Fixed effect (FE) estimator. Note that the FE estimator cannot estimate the effect of any time invariant variable like sex, race, religion, or union participation. Nor can it estimate the effect of any individual invariant variable like price, interest rate, etc., that varies only with time. These variables are wiped out by the within transformation.

If μ_i and λ_t are random variables with zero means and constant variances σ^2_{μ} and σ^2_{λ}, this model is known as the random effects (RE) model. The preceding moments are conditional on the $x'_{it}s$. In addition, μ_i,λ_t and v_{it} are assumed to be conditionally independent. The random effects (RE) model can be estimated by GLS which can be obtained using a least squares regression of $y_{it} = y_{it} - \theta 1 y_{i.} - \theta 2 y_{.t} + \theta 3 y_{..}$ on x^*_{it} similarly defined, where$\theta 1$, $\theta 2$ and $\theta 3$ are simple functions of the variances components σ^2_{μ},σ^2_{λ} and $\sigma^2 v$. Fuller and Battese (1974). The corresponding GLS estimate, the estimator of β is known as the RE estimator. Note that for this RE model, one can estimate the effects of time invariant and individual invariant variables. The Best Quadratic Unbiased (BQU) estimators of the variance components are ANOVA type estimators based on the true disturbances and these are Minimum Variable Unbiased (MVU) under normality of the disturbances. One can obtain feasible estimates of the variances components by replacing the true disturbances by OLSE residuals, see Wallace and Hussain (1969). Alternatively, one could substitute the fixed effects residuals, as proposed by Amemiya (1971). In fact, Amemiya (1971) shows that the Wallace and Hussain (1969) estimate of the variance components have the same asymptotic distribution as that knowing the true disturbances. Other estimators of the variance components were proposed by Swamy and Arora (1972) and Fuller and Battese (1974).

Methodology

A general panel data model is given as;

$\boldsymbol{Y_{it} = X'_{it}\beta + \beta_0 + u_{it}}$;$i = 1,2,\dots,n$; $t = 1,2,\dots,T$ …..(2)

Where Y_{it} – the response for unit i at time t, X'_{it} - contains k regressors for unit i at time t,

β - a vector of k regression coefficients to be estimated,

u_{it} - the error component for unit i at time t.

Specifically, we considered the panel data model with three (3) exogenous and one (1) endogenous variables as shown below;

$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it}}$ ……(3)

The model therefore becomes

$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + u_{it}}$ …… (4)

$\boldsymbol{\varepsilon_i}$ is the individual specific error component with variance σ^2_{ε}, u_{it} is the combined

time – series and cross – sectional error component with variance σ^2_u .

Assume;$\boldsymbol{w_{it} = \varepsilon_i + u_{it}}$

Then the model becomes;

$\boldsymbol{Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it}}$ ……(5)

(1) The Pooled OLS:

$\boldsymbol{y = X'\beta + w}$……(6)

Where $\boldsymbol{y}$ is an $\boldsymbol{nT \times 1}$ column vector response variable, $\boldsymbol{X}$ is an $\boldsymbol{nT \times k}$ matrix of regressors, $\boldsymbol{\beta}$ is a$\boldsymbol{(k+1) \times 1}$column vector of regression coefficients, $\boldsymbol{w}$ is an $\boldsymbol{nT \times 1}$column vector of the combined error terms.

Hence, we have the model;

$\boldsymbol{y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + w}$ …… (7)

The pooled estimator is given as; $\boldsymbol{\hat{\beta}_{pooled} = (X'X)^{-1}X'y}$

(2) Between Estimator:

This estimator is quite intuitive as one performs OLS on a 'collapsed' data set where all data are converted into individual specific averages $\boldsymbol{\bar{y}}$ and $\boldsymbol{\bar{X}}$**.** The resulting between estimator is given by;

$\boldsymbol{\hat{\beta}_B = (X'P_D X)^{-1}X'P_D y}$ …… (8)

Where $\boldsymbol{P_D = D(D'D)^{-1}D'}$and $\boldsymbol{D = I_n \otimes i_T}$, i.e. a$\boldsymbol{nT \times n}$ matrix of n dummy variables corresponding to each cross-section unit, that is, each individual. Note that if OLS on the pooled sample is consistent, the between estimator $\hat{\beta}_B$ is also consistent, though not efficient.

(3) Within Estimator:

The data is pre - multiplied by a matrix $\boldsymbol{M_D}$, where $M_D = I_{nT} - \boldsymbol{D(D'D)^{-1}D'}$and OLS is then computed and then transformed. The following estimator, the within estimator, then is

$\boldsymbol{\hat{\beta}_w = [(M_D X)'(M_D X)]^{-1}(M_D X)'(M_D y)}$

$\boldsymbol{= (X'M_D X)^{-1}X'M_D y}$ …… (9)

If the assumptions underlying the random effects model are correct, the within estimator $\boldsymbol{\hat{\beta}_w}$is, like the between estimator, consistent, but not efficient.

Evaluation of Results

Absolute Bias: This implies the absolute difference between the estimated value and actual value of parameters of a model.

$$AB = \left|\beta_i - \hat{\beta}_i\right| ; i = 1, \dots, k$$

Relative Efficiency: The Relative Efficiency of two estimators is given by the ratio of their efficiencies. It can be expressed mathematically as;

$$\frac{RMSE_1}{RMSE_2} \times 100$$

Simulation Scheme

This work considers one way error component model with three (3) exogenous and one (1) endogenous variables. Though there was no multicollinearity as there was no strong linear relationship between the exogenous variables. We simulated the three exogenous variables and the error terms using *R statistical software package* (*www*.cran.org) following a normal distribution. We specify arbitrarily the parameters (coefficients), we derived the endogenous variable from the simulated data .In our analysis, we checked for the bias and consistency of each Estimator. We compared their RMSE (Root Mean Square Error) values, R-squared values and Variances from different sample sizes (20 and 50).

The datasets used for this work were simulated using Monte Carlo experiments in the environment of R statistical package (***www.cran.org***). Three exogenous variables were simulated across 20 individuals (i.e., n=20) and over 10 years' time period (i.e., T=10). We made a replication of this simulation in the form (n=50, T=10) i.e. we simulated across 50 individuals over 10 years' time period. This is a good case of balanced panel data.

Case a: (Normally Distributed Exogenous Variables)
Exogenous variable was generated using the equation provided by Ayinde, 2012
$X_1 \sim N(\mu_1, \sigma_1^2)$, $X_2 \sim N(\mu_2, \sigma_2^2)$ and $X_3 \sim N(\mu_3, \sigma_3^2)$ ..... (10)
Case b: We choose arbitrarily the coefficients: $\beta_0 = 101$, $\beta_1 = 2.5$, $\beta_2 = 1.8$, $\beta_3 = 0.9$
Case c: The individual error term follows normal distribution with mean (0) and variance (1).
i.e $u_i \sim N(0, 1)$
The remainder error terms follow a normal distribution with mean (0) and variance (1).i.e $w_{it} \sim N(0, 1)$
Case d: We combined cases A, B and C together to establish the endogenous variable.

Data analysis
Results of the analysis
We compared the results of the estimators in the Tables below using their estimates, RMSE (Root Mean Square Error) Values, R–Squared Values, and Variances.

Table 1: R-squared, MSE, and variance of the estimators

Parameter	n=20, T=10					
	β_1	β_2	β_3	R-Squared	Variance	RMSE
Pooled OLS	2.563603 (0.07096)*	1.789910 (0.04742)*	0.980401 (0.06046)*	0.93464	1.9418423	1.3935
Within Estimator	2.568909 (0.057069)*	1.825416 (0.037868)*	0.904605 (0.046898)*	0.96084	1.0697014	1.0343
Between Estimator	2.42100 (0.44019)*	1.87402 (0.35017)*	1.89614 (0.55305)*	0.84271	0.9710397	0.9854
	n=50, T=10					
Pooled OLS	2.500672 (0.040036)*	1.776819 (0.028445)*	0.900508 (0.031844)*	0.94648	1.9418	1.3935
Within Estimator	2.467997 (0.032561)*	1.784177 (0.022517)*	0.909592 (0.025318)*	0.96872	0.963667	0.9817
Between Estimator	2.69753 (0.25179)*	1.70955 (0.22393)*	0.83028 (0.24053)*	0.78098	0.8556407	0.9250

*These are the value for Standard Errors of the estimators

Table 2: Absolute bias values

Parameter	n=20, T=10		
	β_1	β_2	β_3
Pooled OLS	0.063603	0.010090	0.080401
Within Estimator	0.068909	0.025416	0.004605
Between Estimator	0.079000	0.074020	0.996140
	n=50, T=10		
	β_1	β_2	β_3
Pooled OLS	0.000672	0.023181	0.000508
Within Estimator	0.032003	0.015823	0.009592
Between Estimator	0.19753	0.090450	0.069720

Table 3: Relative efficiency using the RMSE values

n=20, T=10	
Pooled OLS vs Within Estimator	134.7288
Pooled OLS vs Between Estimator	141.4147
Within Estimator vs Between Estimator	104.9625
n=50, T=10	
Pooled OLS vs Within Estimator	141.9476
Pooled OLS vs Between Estimator	150.6486
Within Estimator vs Between Estimator	106.1297

Analysis of Results
The Pooled OLSE has R – squared value of 0.93464 when n = 20, and T =10, and when n = 50, T = 10, it has R-squared value of 0.94648 which indicate a good fit of the model through different sampling distributions. The Within Estimator is better than pooled OLS as n increases. The values of β_1, β_2, and $\beta_3$as 2.57, 1.825 and 0.904, respectively. When n=20, T=10 and when n=50, T=10, the Beta values are 2.468, 1.7842 and 0.9096, respectively. The Within Estimator has a relatively low Variance and RMSE values compared to Pooled OLSE hence, it is more efficient than Pooled OLSE. The Between Estimator is more efficient compared to Within Estimator and pooled OLS even as n increases. The values of β_1, β_2, and β_3 are 2.42, 1.874 and 1.896, respectively when n=20, T=10, while when n=50 and T=10, the Beta values are 2.698, 1.7096 and 0.8303, respectively. The Within Estimator has a relatively low Variance and RMSE values compared to between and Pooled OLSE

Conclusion
The result of the analysis after considering some methods of validation (RMSE, R-squared, Variance) reveals that the absolute bias considered for individual parameters shows that OLSE has the least value compare to other estimators. The within effect estimator better explain the fitness of the model due to the highest R-squared value. The between effect estimator perform better than other estimators because it has the least RMSE and Variance values. It can be concluded that the between effect estimator is more efficient and better than the pooled OLSE and within effect estimator in a panel data because of its relatively low RMSE and Variance value. Also under the sampling distribution properties as $n \to \infty$ the RMSE for all estimators also decreases.

Conflict of Interest
Author declares that there is no conflict of interest related to this study.

References
AmirKhalkhali S & AmirKhalkhali S 2013. Predictive efficiency of random effects approach: A real model simulation study. *J. Bus. & Econ. Res.*, 11: 497-500.

Arlleno M 1993. On the testing of correlated effects with panel data. *Journal of Econometrics*, 59: 87-97.

Arlleno M 2003 Panel Data Econometric. Oxford University Press, Oxford, England.

Ayinde K 2012. Generation of Equation of Normal distribution Random Variable

Ayoola FJ 2003. On the performances of panel data estimators in the presence of heteroscedaticity. Ph.D Thesis Submitted to the Department of Statistics, University of Ibadan, Nigeria.

Baltagi BH 2005. Econometric Analysis of panel data, England, John Wiley and Sons Ltd.

Baltagi BH 1985. Pooling Cross – section with unequal time series length. *Economic Letters*, 18: 133 – 136.

Baltagi BH & Li Q 2002. On instrument variable estimation of semi parametric dynamic panel data models. *Economics Letter*, 76: 1 - 9.

Baillie RT & Baltagi BH 1995. Prediction from Regression Model with One-way Errorcomponents, Chapter 10.

Baltagi BH 1981a. Pooling: An experimental study of alternative testing and estimation procedures in a two way error components models. *Journal of Econometrics*, 17: 21-49.

Baltagi BH & Li Q 1992. A monotonic property of iterative GLS in the two-way random effects model, *Journal of Econometrics*, 53: 45-51.

Breush TS 1987. Maximum Likelihood estimation of random effect models. Journal of Econometrics, 36: 383-389.

Baltagi BH 2005. Estimating an econometric model of crime using panel data from North Carolina. *Journal of Applied Econometrics*

Belsley DA, Kuh E & Welsch RE 1980. Regression Diagnostics: Identifying Influential Data and Source of Collinearity (John Wiley, New York).

Fuller WA & Battese GE 1974. Estimation of linear models with cross-error structure. *Journal of Econometric*, 2: 67-78.

Jin L & Jin JC 2014. Internet Education and Economic Growth: Evidence from Cross-Country Regressions, pp. 78-94.

Hausman JA & Taylor WE 1981. Panel data and unobserved individual effects. *Econometrica*, 49: 1377 – 1398.

Nihat T, Emrah O & Ali H 2013. Determinants of economic growth in G20 countries: A panel data approach. *Int. J. Latest Trends in Fin & Eco Soc.*, 523-581.

Reza F & Widodo T 2013. The impact of education on economic growth. *J. Indonesian Econ. and Bus.*, 28(1): 23-44.

Stefan C 2016. Human capital as a determinant of the economic growth – A panel data approach. *Int. J. Econ., Commerce and Mgt., United Kingdom*, 4(5): 28.

Taylor WE 1980. Small sample considerations in estimation from panel data. *Journal of Econometrics*, 13: 203-223.