International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(6): 144-149
© 2024 Stats & Maths
https://www.mathsjournal.com
Received: 09-09-2024
Accepted: 15-10-2024

OP Balogun
Department of Statistics, Federal Polytechnic, Bida, Nigeria

WB Yahya
Department of Statistics, University of Ilorin, Nigeria

AA Issa
Department of Statistics, Abubakar Tafawa Balewa Bauchi, Nigeria

Corresponding Author:
OP Balogun
Department of Statistics, Federal Polytechnic, Bida, Nigeria

Evaluation of panel data estimators under the unbalanced panel data for small data sizes occasioned by missingness

OP Balogun, WB Yahya and AA Issa

DOI: https://doi.org/10.22271/maths.2024.v9.i6b.1914

Abstract

This paper examines the performances of the developed Between Median estimator (BMd) for balanced panel data under the Unbalanced panel data for small datasets by introducing varying degrees of missingness. This study focuses on 5% increments of missingness (5%- 20%), and sample sizes, N (20 and 30) with a unit beta (slope); β=. Adopting the Monte -Carlo investigation of (Balogun *et al* 2022), where the behaviors of the panel data estimators' using Mean Square Error (MSE) and Mean Absolute Error (MAE) criteria to measure their performance were tested; among the five estimators of balance panel data models under Unbalanced panel data estimators to varying levels of degree of missingness injected for panel data (N=25) shows that the Between estimator performed best among the estimators. In a separate study, (Balogun and Yahya 2024) [3] found that the BMd estimator, developed for small size N (20 and 25) with β=3 and T = 5 panel data under the Unbalanced panel data has the lowest MSE and MAE values, and therefore, outperformed the Between estimator. This work builds on (Balogun and Yahya 2024) [3]; the findings demonstrate that the developed estimator (BMd) performed best of the six estimators tested using the same measurement criteria.

Keywords: Missingness, small sample size, panel data estimators, unbalanced panel data, rank

Introduction

There is a growing interest in the investigation of panel data analysis especially small sample panel data set. There is no rule of thumb for choosing a sample size, however, appropriate sample size is required for validity. If the sample size is too small the outcomes will be invalid, moreover, if the sample is too large it will result into waste of time and money. (VanVoorhis and Morgan 2007, pg. 48; Memon *et al*. 2020, pg. 4) [16, 12] believes regression analysis should have at least 10 observations per variable. (Indrayan and Mishra 2021) [10] stress the importance of small data in medical research.

Researchers are in a cross-road as to what estimator to be used in the analysis for a small sample size. Ordinary least square (OLS) estimator is believed to have small sample properties. (Santos and Barrios 2011) [14] study small and large sample properties of the within-groups (WG) estimator and the first difference generalized method of moments (FD-GMM) estimator of a dynamic panel data (DPD) model. In their study, different panel data samples were examined; a case of small-time dimensions case with T = 3 and T= 5 for small sample or finite sample properties; and a small size for both the cross-section and time dimensions, say N= 20 were considered.

In literatures, the Monte-Carlo studies to investigate the performances of the five different estimators of balance panel data models under unbalanced panel data models show results for the performances for the criteria computed; mean square error (MSE) and mean absolute error (MAE) show that for balance panel models and a respective estimate for unbalanced panel models ranking have the same pattern. That is Between estimator ranked 1st and First Difference estimator ranked last, while Within estimator, Random (Swamy-Arora's) estimator and Pooling estimator do not show a steady pattern of ranking. (Balogun *et al*. 2022) [4].

Using the same procedure and measurement criteria, Balogun and Yahya (2024) [3] developed an estimator, particularly for small sample panel data, that outperformed the trio's between estimator researched. The Between Median estimator (BMd) is an extension of the between estimator. It uses Generalized Least Squares (GLS) in the one-way error component of the between estimator, and the resultant estimation is the regression of the dependent variable Y on the mean of the independent variable(s) X(s). This estimator was modified by regressing the Y dependent variable's median on the median of the X(s) independent variable(s) to obtain a Between Median Estimator. An investigation of the performances of the six various estimators that includes the developed estimator of balance panel data estimators under the Unbalanced panel data where increasing degrees of missingness are added to a panel data size reveals that the BMd estimator outperformed the other estimators under studied. (Balogun *et al*., 2022) [4].

In general, the Between Median estimator outperformed the other five estimators investigated for panel data estimators under the unbalanced data set for a levels of sample sizes (n), as signal in the study. (Balogun and Yahya 2024) [3]. Similar to (Balogun and Yahya 2024) [3], Between Median estimator outperformed the estimators used in this investigation.

Material and Methods

The Classical Panel Data Model

Consider a general panel data model

$$y_{it} = \alpha_i + \beta_1 X'_{it} + \mu_{it}, \qquad (1)$$

where,

y_{it} is the response for unit i at time t,

α_i denotes the individual- specific intercept,

vector X'_{it} contains k regressors for unit i at time t,

vector β contain regression coefficients to be estimated and

μ_{it} is the error component for unit i at time t, $i = 1,2 \ldots, n$ and t = 1,2 …., T

In equation (3.1) considering the panel data model that has two exogenous and one endogenous variable;

$$y_{it} = \alpha_i + \beta_1 X_{1it} + \beta_2 X_{2it}\beta + \mu_{it} \qquad (2)$$

where $\alpha_i = \alpha + \varepsilon_i$, the individual- specific intercept (α_i) includes the effects of those variables that are peculiar to the ith individual and that are time – variant.

This model becomes

$$y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it}\beta + \varepsilon_i + \mu_{it}, \qquad (3)$$

where,

y_{it} is the response of an individual i over period t

ε_i is the individual-specific error component and

μ_{it} is the combined time-series and cross –section error component with variance δ^2_ε and δ^2_μ respectively.

If $w_{it} = \varepsilon_i + \mu_{it}$, then model (3.3) becomes:

$$y_{it} = \alpha + \beta_1 X_{1it} + \beta_2 X_{2it}\beta + w_{it}, \qquad (4)$$

3.2. Estimators of Panel Data Model

The most common estimators of panel data models to be fitted with different conditions in criteria and efficiency are discussed in this section. Some of these estimators are as follow.

3.2.1. Pooled Estimator: This Estimator stacks the data over i and t into one long regression with nT observations and estimates of the parameters are obtained by OLS using the model (Greene 2008) [9]; (Garba *et al.* 2013) [8]; (Balogun and Yahya 2024) [3].

3.2.2. Within Estimator: This is equivalent to "amemiya". (Amemiya 1971) [1]; (Matyas and Sevestre 1992) [11]; (Balogun and Yahya 2024) [3]. This regresses on the deviations from the individual or/and time mean.

3.2.3. Between Estimator (BTW): This regresses the group means of Y on the group means of X’s in a regression of n observations. It uses cross-sectional variation by averaging the observations over period t; (Creel and Tillman 2011) [7]; (Wooldridge 2012) [17]. Explicitly, it converts all the observations into individual-specific average and performs OLS on the transformed data. (Balogun and Yahya 2024) [3].

3.2.4. First- Differenced Estimator (FD)

This is the ordinary least squares estimation of the difference between the original model and its one-period-lagged model. (Arellano 2003) [2]; (Baltagi 2005) [5]; (Balogun and Yahya 2024) [3]

3.2.5. Random Estimator

This is equivalent to “swar” models. (Swamy and Arora 1972) [15]; (Cottrell 2017) [6]. " According to Hauser, the estimator follows the underlying model expression. (Balogun and Yahya 2024) [3].

2.7. Monte-Carlos Procedure

Monte-Carlos is a mathematical technique based on experiment for evaluation and estimation of problems which are intractable by probabilistic or deterministic approach.

2.7.1. The Data Structure

The panel data model considered is of the form:

$$Y_{it} = \beta_{0it} + \beta_{1it}x_{1it} + \varepsilon_{it}., \varepsilon_{1t} \sim N\,(0,1), \qquad (5)$$

$t = 1, \ldots, T;\ i = 1, \ldots, n_k$ and $k = 1, \ldots, 5$.

i) A total of k = 5 subjects (for instance) was studied and simulated over T = 5 times. Thus, a total of n = N ($k \times T$) observations were generated.

ii) The sample sizes considered are n = (20 and 30); at different sample units k = 5, with T = 4 and 6

iii) An independent variable x_{it} was considered and simulated from the Gaussian distribution; $X_{it} \sim iid\ N(20,1)$.

iv) Over fixed time interval of k = 4 and 6

v) The vector of parameters, β is set as $\hat{\beta} = \left(\hat{\beta}_{0it}, \hat{\beta}_{1it}\right)^{/} = (20,1)^{/}$.

vi) The response variable Y_{it} was simulated from (1) accordingly using all the above definitions.

vii) Unbalance time interval was infused into the data by randomly removing ω% of the total sample from the data, where ω take on values 5, 10, 15 and 20. (Balogun *et al.*, 2022) [4].

2.7.2. Models Assessment

The following model assessment criteria shall be employed to determine the relative Absolute efficiencies of the various estimation considered.

i) Mean Absolute Error (MAE):
ii)

$MAE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}|\beta_{it} - \beta|,$ (6

Where,
n is the number of errors,
Σ is the summation symbol, that is, adding individual i over n and time t over T,
$|\beta_{it} - \beta$ | is the absolute errors.

iii) Mean Square Error (MSE):

$$MSE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{\beta}_{it} - \beta)^2, \quad (7)$$

Where,
n is the number of errors,
Σ is the summation symbol, that is, adding individual i over n and time t over T,
$(\beta_{it} - \beta)$ is the square errors. (Robeson and Wilmott 2023) [13]; (Balogun *et al*., 2022) [4].

3. Result and Discussion

Table 1: Mean Square Error (MSE), N=20, β=1, T=4

MSE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0%**	**5%**	**10%**	**15%**	**20%**
N=20, β=1, T= 4	5	20	17	13	10	6
Pooling		0.6872	0.5910	0.6304	0.2247	0.1878
Within		0.1878	0.4739	0.2969	0.1318	0.0331
Random		0.0331	0.5910	0.5238	0.2247	0.0839
First Difference		0.8799	1.1007	0.5609	0.2554	0.0729
Between		0.0165	0.0794	0.2027	0.2027	0.2027
Between Median		0.0000	0.0000	0.0000	0.0000	0.0000

Table 2: Mean Square Error Result in Order of Ranking, N=20, β=1, T=4

MSE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0%**	**5%**	**10%**	**15%**	**20%**
	5	20	17	13	10	6
Pooling		5th	4th	6th	4th	4th
Within		4th	3rd	3rd	2nd	2nd
Random		3rd	4th	4th	4th	4th
First Difference		6th	5th	5th	5th	3rd
Between		2nd	2nd	2nd	3rd	5th
Between Median		1st	1st	1st	1st	1st

Table 3: Mean Square Error (MSE), N=30, β=1, T=6

MSE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0%**	**5%**	**10%**	**15%**	**20%**
N=30, β=1, T=6	5	30	24	18	17	13
Pooling		0.6100	0.6171	0.6374	0.6652	0.3056
Within		0.5816	0.5231	0.5908	0.4648	0.2454
Random		0.6100	0.6171	0.6374	0.6138	0.3056
First Difference		0.9704	1.4265	1.2028	0.8989	0.7826
Between		0.0245	0.0911	0.0625	0.2827	0.0318
Between Median		0.0000	0.0000	0.0000	0.0000	0.0000

Table 4: Mean Square Error Result in Order of Ranking, N=30, β=1, T=6

MSE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0%**	**5%**	**10%**	**15%**	**20%**
	5	30	24	18	17	13
Pooling		4th	4th	4th	5th	4th
Within		3rd	3rd	3rd	3rd	3rd
Random		4th	4th	4th	4th	4th
First Difference		5th	5th	5th	6th	5th
Between		2nd	2nd	2nd	2nd	2nd
Between Median		1st	1st	1st	1st	1st

Table 1 and 3 show the Mean Square Error (MSE) values, while Table 2 and 4 indicate the order of rank of the MSE for N (20 and 30), across T=4 and T=6 respectively. The developed Between Median estimator has the lowest values of MSE and therefore ranks first for the various N values across T values as the amount of missingness increases; the Between estimator rank second, and the Pooling estimator with the highest values ranks last.

Figure 1 illustrates the trend of the MSE of N across different T levels, showing that the MSE converges across N values as the degree of missingness increases.

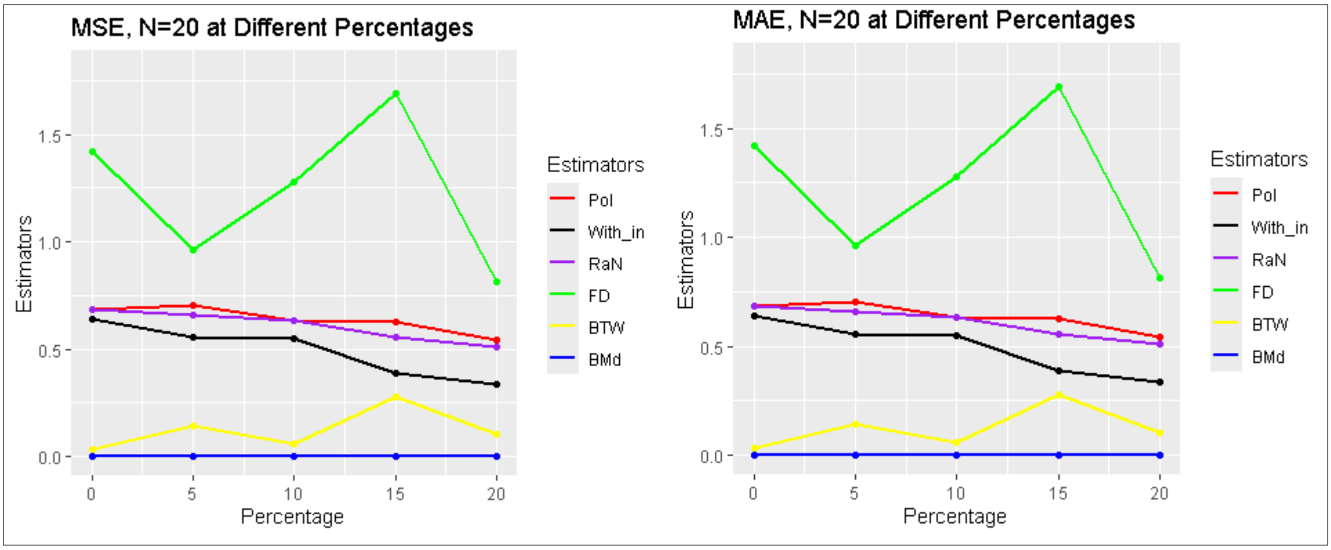

Fig 1: Mean Square Error (MSE) and absolute mean square Error (MAE) for small data size, N = 20, n=5, and at percentage (0, 5, 10, 15, 20)

Table 5: Mean Absolute Error (MAE), N=20, β=1, T=4

MAE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0**	**5**	**10**	**15**	**20**
N=20, β=1, T=4	5	20	17	13	10	6
Pooling		0.6579	0.5829	0.6859	0.3688	0.6579
Within		0.6230	0.5527	0.4061	0.3023	0.1578
Random		0.6579	0.5829	0.6251	0.3688	0.2277
First Difference		0.8288	0.9923	0.6179	0.4878	0.2218
Between		0.1018	0.2586	0.3681	0.2398	0.2274
Between Median		0.0000	0.0000	0.0000	0.0000	0.0000

Table 6: Mean Absolute Error Result in Order of Ranking, N=20, β=1, T=4

MAE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0**	**5**	**10**	**15**	**20**
	5	20	17	13	10	6
Pooling		4th	4th	6th	4th	5th
Within		3rd	3rd	3rd	3rd	2nd
Random		4th	4th	5th	4th	4th
First Difference		5th	5th	4th	5th	3rd
Between		2nd	2nd	2nd	2nd	4th
Between Median		1st	1st	1st	1st	1st

Table 6: Mean Absolute Error (MAE), N=30, β=1, T=6

MAE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0**	**5**	**10**	**15**	**20**
N=30, β=1, T=6	5	30	24	18	17	13
Pooling		0.6566	0.6648	0.6396	0.6423	0.6566
Within		0.6468	0.6208	0.6065	0.5580	0.3821
Random		0.6566	0.6648	0.6396	0.6170	0.3727
First Difference		0.8123	0.9926	0.9139	0.8657	0.7936
Between		0.1373	0.2224	0.2218	0.3797	0.1336
Between Median		0.0000	0.0000	0.0000	0.0000	0.0000

Table 6: Mean Absolute Error Result in Order of Ranking, N=30, β=1, T=6

MAE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0**	**5**	**10**	**15**	**20**
	5	30	24	18	17	13
Pooling		4th	4th	4th	5th	5th
Within		3rd	3rd	3rd	3rd	4th
Random		4th	4th	4th	4th	3rd
First Difference		5th	5th	5th	6th	6th
Between		2nd	2nd	2nd	2nd	2nd
Between Median		1st	1st	1st	1st	1st

Similar to MAE result, Tables 5 and 7 exhibit the Mean Absolute Error (MAE) values, whereas Tables 6 and 8 show the order of rank of the MAE for N (20 and 30) at T=4 and T=6, respectively. The Between Median estimator has the lowest MAE values and hence ranks first for the different N values throughout T values as the degree of missingness grows; the Between estimator ranks second; and the Pooling estimator, with the highest values, ranks last.

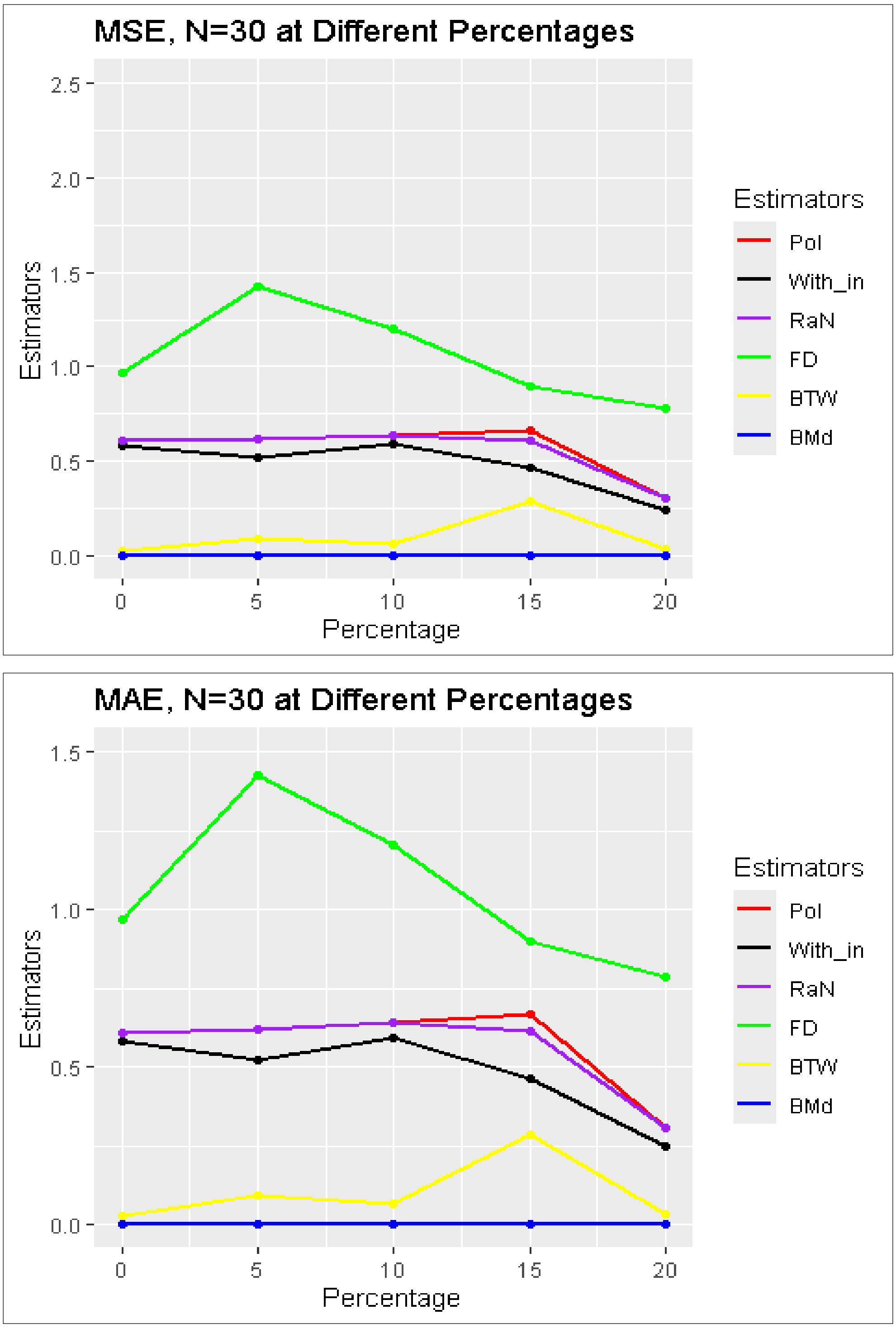

Fig 2: Mean square error (MSE) and absolute mean square error (MAE) for small data size, N =30, n=5, beta=1, and at percentages (0,5,10,15,20)

Also, Figure 2 shows the pattern of the MAE of N across various levels of T, as the level of missingness grows, the MAE converge throughout the N values.

7.2. Conclusion

Finding in this work shows the Between Median estimator (BMd) outperform other panel data estimators considered for small sample size of unbalanced data, when N (20 and 30). Therefore, it is concluded that the Between estimator performed best, this aligns with the findings of (Balogun and Yahya 2024) [3]. They show that, for the data structure considered Between Median estimator is the best estimator.

7.3. Recommendation

It is advised that the Between Median estimator be used to fit the panel data models when there is noticeable missingness in tiny data with varying sample sizes (20 and 30) with varying percentages of missingness (0, 5, 10, 15, and 20).

Limitation

The scope of this work is restricted to the examination of small panel data sizes when $\beta = 1, or$ a unit slope (beta), further investigation can be made with different values of $\beta, or\ \alpha$. Also, the size N across T can also be increase to examine the performance of the developed estimator (BMd)

Ethical Approval

This work was approved by the University Ethical Review Committee (UERC) of University of Ilorin, Nigeria.

Data Available Statement

Data for this work is basic and it is contained in the body of the paper.

References

1. Amemiya T. The estimation of the variances in a variance-components model. International Economic Review. 1971;12:1–13.
2. Arellano M. Panel Data Econometrics. 2003.
3. Balogun OP, Yahya WB. Development of estimation methods for modeling unbalanced panel data for small data occasioned by missingness. Edited Proceedings of the 8th International Conference. Professional Statisticians Society of Nigeria. 2024, 8(1).
4. Balogun OP, Yahya WB, Umar-Mann A. Performance evaluation of some estimators under unbalanced panel data models. Edited Proceedings of the 6th International Conference. Professional Statisticians Society of Nigeria. 2022, 6(1).
5. Baltagi BH. Econometric Analysis of Panel Data. England: John Wiley and Sons; c2005.
6. Cottrell A. Random effects estimators for unbalanced panel data: A Monte Carlo analysis using gretL. 2017.
7. Creel E, Tillman K. Stigmatization of overweight patients by nurses. Qual Rep. 2011;16(5):1330-1351.
8. Garba MK, Oyejola BA, Yahya WB. Investigations of certain estimators for modeling panel data under violations of some basic assumptions. 2013, 3(10).
9. Greene WH. Econometric Analysis; c2008.
10. Indrayan A, Mishra A. The importance of small samples in medical research. J Postgrad Med. 2021;67(4):219-223. DOI:10.4103/jpgm.JPGM_230_21.
11. Matyas L, Sevestre P. The Econometrics of Panel Data. Kluwer Academic Publishers; c1992. p. 46-71.
12. Memon MA, Ting H, Cheah JH, Thurasamy R, Chuah F, Cham TH. Sample size for survey research: Review and recommendations. J Appl Struct Eq Model. 2020;4(2):1-20. DOI:10.47263/jasem.4(2)01.
13. Robeson SM, Willmott CJ. Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PLoS One. 2023;18(2):e0279774. DOI:10.1371/journal.pone.0279774.
14. Santos L, Barrios E. Small sample estimation in dynamic panel data models: A simulation study. Open J Stat. 2011;1(2):58-73. DOI:10.4236/ojs.2011.12007.
15. Swamy PAVB, Arora SS. The exact finite sample properties of the estimators of coefficients in the error components regression models. Econometrica. 1972;40(2):261-275.
16. VanVoorhis CW, Morgan BL. Understanding power and rules of thumb for determining sample sizes. Tutorials Quant Methods Psychol. 2007;3(2):4350.
17. Wooldridge JM. Introductory Econometrics: A Modern Approach. 5th ed. 2012.