WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

ON INVESTIGATION OF LONGITUDINAL DATA ESTIMATORS UNDER THE UNBALANCED PANEL DATA FOR SMALL DATA SIZES INDUCED BY MISSINGNESS

O. P. Balogun[1], W. B. Yahya[2], A. A. Issa[3]

[1]Department of Statistics, Federal Polytechnic, Bida, Nigeria
[2]Department of Statistics, University of Ilorin, Nigeria
[3]Department of Statistics, Abubakar Tafawa Balewa Bauchi, Nigeria
e-mail: balogun.omoshadephilomena@fedpolydida.edu.ng[1];
wbyahya@unilorin.edu.ng[2] ; aaissa@atbu.edu.ng[3]

ABSTRACT.

This paper focuses on the performance of longitudinal models under the unbalanced panel data for N (20, 25, and 30) when beta $\beta = 3$. An unbalanced model are time invariant models, where there are missing values in the panels of the data. [1] in their study, evaluated Between Median estimator (BMD) which was developed for modeling a panel data under the unbalanced panel data by assessing their behaviors using Mean Square Error (MSE) and Mean Absolute Error (MAE). Among the six estimators studied Between Median estimator (BMD) has the lowest values of MSE and MAE for N (20 and 25) with β = 1 and T = 4 and 6. This work builds on [2] with different data structures; an additional N across various T and a constant $\beta = 3$. 5% additions at regular interval of missingness (5%- 20%), sample sizes of N =20, 25, and 30; and T= 4, 5, and 6 $with\ \beta = 3$ were considered. A Monte Carlo study of a normal distribution of a panel data model of exogenous x_{it}, and endogenous variable Y_{it}, and the error ε_{1t} show that the Between Median estimator (BMD) performed best of the six estimators tested using MSE, MAE, and RMSE. Application of the real-world panel dataset show that, the Between Median estimator outperformed other five panel data estimators employed using the same measurement criteria.

Keywords: Missingness, small sample size, Longitudinal models, model selection, slope.

1. Introduction

Longitudinal data is an individual cross-sectional data with time invariant T. Investigation of panel data analysis especially small sample panel data set is important. The slope, Beta (β) is a parameter of interest [3]. This paper is a contribution to the effect of β on panel data models; heuristically, a change in X (predictor) of one unit leads to a prediction of an increase or decrease in $\beta\ unit$. The effect of β control the individual and time heterogeneity. [4]; [5]. Therefore, it implies that, an increase in β will increase or decrease y and have a disturbance on the error ε_{1t} .

Slope and intercept are parameters of interest in regression; heuristically, slope is needed for statistical derivation. This cannot be defined independent of the covariate of the individual across the period. For unbiasedness of Beta (β), the estimator of $\hat{\beta}$ is the population slope in addition to the error term. For a simple linear regression, the expected values of beta given the regressor is the slope. $E(\hat{\beta}/X) = \hat{\beta}$. [6].

In their study, [1] observed that the Between Median Estimator (BMD) outperformed the best-performing Between estimator in [7]. The Between Median estimator (BMD) is a variant of the Between estimator. Between estimator employs Generalized Least Squares (GLS) in the two-way error component panel data models, with the final estimation being the mean of the dependent variable Y's regressing on the mean of the independent variable(s) X(s).

This work focuses on the heuristic search technique, building on [1]. In simulation, adjusting parameter variable settings is important for effective runs to attain an optimal output. The adjusted constant $\beta = 3$ is examined for an addition of N=30. The sizes of N across T studied are N (20, 25, and 30).and T varied at T= 4, 5 and 6 respectively. [8-10]. [11]; [12-13]; and [14] all investigated small data. The work was validated using the real -world data of the top five African Countries data ranked by Gross Domestic Product (GDP) at Purchasing Power Parity (PPP), current price.

The Between Median Estimator (BMD) outperformed the other five estimators tested for balance panel models under unbalanced panel by regressing the median of Y dependent variable on the median of the X independent variable. The method of estimation was generalized least square (GLS), this was consistent and efficient estimator [15-16].

2. MATERIALS AND METHODS

2.1. The Panel Data Estimator

panel data model is given as;

$y_{it} = \alpha_i + \beta X'_{it} + \mu_{it}$, (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 u_{it} is the error component for unit i at time t, $i = 1,2 \ldots, n$ and t = 1,2 …., T

2.2. Six Panel Data Estimators Employed

The five common estimators of panel data and the Between Median estimator developed by [1] to be fitted with different conditions and criteria and efficiency were discussed in this section. These estimators are:

2.2.1. *Pooled Estimator:* This is also known as the OLS. It regresses the dependent variable y on the independent variable X. 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 [17]; [18].

$y_{it} = \beta_0 + \beta^T x_{it} + \varepsilon_{it}$, $t = 1,2, \ldots T\,; i = 1, 2, \ldots, N$, (2)

$y = X'\beta + w$, (3)

where, y is an $nT \times 1$ column vector of response variable, X is an $nT \times k$ matrix of regressors,

β is a (k+1) × 1 column vector of regression coefficient, and w is an $nT \times 1$ column vector of the combined error terms (i.e. ε_i+μ_{it}). The pooled estimator is given as

$\widehat{\boldsymbol{\beta}}_{pooled} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$, (4)

2.2.2. Within *Estimator:* This regress on the deviations from the individual or/and time mean. [19]; [20].

$y_{it} = X^*_{it}\beta^* + Z_{j\alpha} + \varepsilon_{jt}$, (5)

where,

$\varepsilon_{jt} = a_i + u_{it}$, (6)

The X^*_{it} matrix does not contain a unit vector. The heterogeneity or individual effect is captured by Z, which contains a constant term and possibly several other individual-specific factors. Likewise, β^*contains $\beta_2, \ldots, \beta_k$, constrained to be equal over i and t. If Z contains only a unit vector, then pooled OLS is a consistent and efficient estimator of $[\beta^* \alpha]$.

2.2.3. *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 [18]. Averaging model (1) over t gives

$\bar{Y}_{i.} = \alpha + \beta_1 \bar{X}_{1i.} + \beta_2 \bar{X}_{2i.} + w_{it}$, (7)

where,

y is an $\bar{Y}_{i.} = T^{-1}\sum_t Y_{it}$, $\bar{X}_{ji.} = T^{-1}\sum_t X_{jt}$, and $\bar{w}_{i.} = T^{-1}\sum_t w_{it}$, for i = 1,2…n and j = 1,2.

2.2.4. *First- Difference Estimator (FD):* This is the ordinary least squares estimation of the difference between the original model and its one-period-lagged model. This estimator is useful in addressing omitted variables problems with panel data. It controls for fixed effects and remove the problem of unobserved heterogeneity. [21]; [22].

$\Delta Y_{it} = \beta_1 \Delta X_{1it} + \beta_2 \Delta X_{2it}\beta + \Delta w_{it}$, (8)

where, $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$, $\Delta X_{1it} = X_{1it} - X_{1i,t-1}$, $\Delta X_{2it} = X_{2it} - X_{2i,t-1}$ and

$\Delta w_{it} = w_{it} - w_{i,t-1}$, for i = 1,2…n and t = 1,2… T

2.2.5. *Random Estimator*: This is the individual specific effect that is unrelated to the explanatory variable [23-24]; [25]; and [26],

$y_{it} = \beta_0 + X'_{it}\beta + \alpha_i + u_{it}$, $u_{it} \sim iid\ (0, \delta_u^2)$, (9)

where,

β_0 is the individual- specific intercept, β is a (k+1) × 1 column vector of regression coefficient, $t = \lambda, \ldots, T$ and $i = 1, \ldots, N$

$Cov\ (\alpha_i, X_{it}) \neq 0 \ \sim FE - model$, (10)

$Cov\ (\alpha_i, X_{it}) = 0 \ \sim OLS - model$, (11)

Also, if

$\lambda = 1 - \left(\frac{\delta_\mu^2}{\delta_\mu^2 + T.\ \delta_\alpha^2}\right)$, (12)

$\lambda = 1 \ \sim FE\ model$, (13)

$\lambda = 0 \ \sim OLS$ model. , (14)

2.2.6. *Between Median Estimator (BMD):* This regresses the group medians of Y on the group medians of X(s) in a regression of n observations. It uses cross-sectional variation by using the median of the observations over period [1]. Let ρ_i represent the group median of the response variable and κ_i the group median of the explanatory variables of the cross-sectional observations n. it follows therefore that for a model with two explanatory variables we have;

$\rho_i = \alpha + \beta_1 \kappa_{1i.} + \beta_2 \kappa_{2i.} + w_{it}$, (15)

where, $\rho_i = \left(\frac{n+1}{2}\right) for\ an\ odd\ observation\ \ and\ \frac{(n/2)+\ (n/2+1)}{2} for\ an\ even\ observation,$

$\kappa_i = for\ an\ odd\ observation\ \ and\ \frac{(n/2)+\ (n/2+1)}{2} for\ an\ even\ observation$

and,

$w_{it} = \left(\frac{n+1}{2}\right) for\ an\ odd\ observation\ \ and\ \frac{(n/2)+\ (n/2+1)}{2} for\ an\ even\ observation$

for i = 1,2…n and j = 1,2.

2.3. 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

Considering the panel data in equation (1)

It follows that $\mu_{it} \sim N(0,1), \quad \mu_{it} = \varepsilon_{it}$

$Y_{it} = \beta_{0it} + \beta_{1it}x_{1it} + \varepsilon_{it}, \qquad \varepsilon_{it} \sim N(0,1)$ (16)

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

Adopting the simulation framework used by [1]. This paper focuses on a Monte Carlo study for a panel dataset of a total of k = 5 subjects over T = 5 and 6 periods and for different sample sizes n (20, 25, and 30). In the balance panel case, GLS is obtained by running the ordinary least square (OLS). This takes care of the autocorrelation or heteroscedasticity in the error term, and the efficiency is obtained by transforming the heteroscedasticity variance-covariance matrix into a homoscedastic.

2.3.1. Assessment Criteria

The following model assessment criteria were employed to determine the relative absolute efficiencies of the various estimation considered.

(i) Mean Absolute Error (MAE):

$MAE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}|\beta_{it} - \beta|$, (17)

where,

n is the number of errors,

Σ is the summation symbol, that is, adding individual i over n and time t over T ,

$|\beta_{it} - \beta|$ are the absolute errors.

[26];

(ii) **Mean Square Error (MSE):**

$MSE = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(\hat{\beta}_{it} - \beta)^2$, (18)

where

n is the number of errors,

Σ is the summation symbol, that is, adding individual i over n and time t over T ,

$(\hat{\beta}_{it} - \beta)^2$ are the square errors.

(iii) **Root Mean Square Error (RMSE):**

$$\left[\frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}(y_{it} - \hat{y}_{it})^2\right]^{1/2} \quad (19)$$

where, $[\,]^{1/2}$ is the root of the mean square error. [27]; [26].

2.4. Theoretical background

2.4.1 Estimation Method for the Balance Panel Data Using Generalized Least Square (GLS)

If the diagonal matrix $\sigma^2\Sigma = V$ is a known n × n matrix, then the parameter estimation for the balanced panel data can be calculated using equation (2). If V has unequal diagonal elements, the observations in y are uncorrelated but have unequal variance, while if V has non-zero off-diagonal elements, the observations are correlated. Estimating β by OLS gives $\hat{\beta} = (X'X)^{-1}y$, however, the estimator is not optimal. The solution is to transform the model to a new set of observations that satisfy the constant variance assumption and use least squares to estimate the parameters.

Since $\delta^2 V$ is a covariance matrix, V is a symmetric non-singular matrix, therefore

$V = K'K = KK$, and K is called the square root of V. [28].

Defining $z = K^{-1}y$, $B = K^{-1}X \; and \; g = K^{-1}\varepsilon \;\Rightarrow z = B\beta + g$, then, using Matrix Algebra,

$E[g] = K^{-1}[\varepsilon] = 0$, (20)

$Var[g] = K^{-1}\varepsilon = K^{-1}Var[\varepsilon]K^{-1} = \delta^2 K^{-1}VK^{-1} = \delta^2 K^{-1}KKK^{-1} = \delta^2 \mathbb{I}$ (21)

Under the assumption of ordinary least square, the least square function is:

$S(\beta) = (z - B\beta)'(z - B\beta) = (K^{-1}y - K^{-1}X\beta)'(K^{-1}y - K^{-1}X\beta) = (y - X\beta)'(K^{-1})'K^{-1}(y - X\beta)$

Since $V = K'K = KK, therefore, (K^{-1})'K^{-1} = V^{-1}$

$S(\beta) = (y - X\beta)'V^{-1}(y - X\beta)$, (22)

$= y'y - y'X\beta - (X\beta)'y + (X\beta)'X\beta)V^{-1}$,

$= y'y - y'X\beta - X'\beta'y + (X\beta)'X\beta)V^{-1}$, (23)

Here, $(X'\beta'y)' = y'X\beta$ is a scalar and equal to its own transpose. Equation (23) becomes,

$= y'yV^{-1} - y'X\beta V^{-1} - X'\beta'yV^{-1} + (X\beta)'X\beta V^{-1}$

$= y'yV^{-1} - 2X'\beta'yV^{-1} + X'\beta'V^{-1}X\beta$, (24)

Taking the partial derivative via matrix calculus with respect to β and setting it to 0 to satisfy the first order condition. Let $(X'V^{-1}X) = A, and \; \beta = x, \; therefore$:

$\frac{\partial S(\beta)}{\partial \beta} = -2X'yV^{-1} + 2(X'V^{-1}X)\beta$, (25)

this gives:

$-2X'yV^{-1} + 2(X'V^{-1}X)\beta = 0$, (26)

$(X'V^{-1}X)\beta = X'V^{-1}y$, (27)

The generalized least squares estimator of β now becomes:

$\hat{\beta} = (X'V^{-1}X)^{-1}X'V^{-1}y = (B'B)^{-1}B'$, (28)

The quantity V^{-1} is known as the precision matrix or dispersion matrix, a generalization of the diagonal weight matrix.

and

$E[\hat{\beta}] = (X'V^{-1}X)^{-1}XV^{-1}E[y] = (X'V^{-1}X)^{-1}XV^{-1}X\beta = \beta$, (29)

and

$\text{Var}[\hat{\beta}] = \delta^2(B'B)^{-1} = \delta^2(X'K^{-1}K^{-1}X)^{-1} = \delta^2(X'V^{-1}X)^{-1}$, (30)

Which leads to the unbiased and consistent estimator of δ^2. [29].

Similarly, under normal theory, the generalized least squares estimators are the maximum likelihood estimators since the log-likelihood function is:

$L\alpha = \ln(\delta^2) - \frac{1}{2}\ln|V| - \frac{1}{2\delta^2}\ln(y - X\beta)'V^{-1}(y - X\beta)$, (31)

2.4.2. Estimation Method for the Unbalanced Panel Data Using Generalized Least Square (GLS)

The modified general case of the generalized least square estimator for unbalanced panel datasets assumed to be a case where the V is unknown. The parameter estimates for the unknown V are obtained using the feasible generalized least square (FGLS) estimator by replacing V by $\hat{V}$. The GLS estimator for the unbalanced case can be interpreted as a matrix-weighted mean, with weight depending on X. In practice, the Vs have to be estimated, which requires estimation of σ^2 *and* σ_α^2.

Analyzing unbalanced panel data becomes more challenging, especially when the values are missing at a higher percentage level. The rule of thumb for missing values in data is 50%; any missing values higher than this are not to be tolerated or accepted for analysis [30]. In this study, missing values were infused into the panel dataset at different degrees, say 5%–20%. The unbalanced data output used Biorn's [31] estimation procedure of the error component of the Between estimator, the generalized least squares (GLS), and the feasible generalized least squares (FGLS) to estimate the parameters of interest.

The one-way error components regression model for unbalanced panel data in which individual i $(i = 1, \ldots, N)$ is observed in T_i periods, and t denotes the observation of numbers that differ from the calendar period if the starting period of the individuals differs or if gaps occur in the time series of some of them. The GLS therefore depends on the T-dimensional relationship in the panel. [32].

3. Results and Discussion

3.1. Results of the Mean Square Error and the Mean Absolute Error

Tables 1, 2, and 3 show the rank of the estimates of the Mean Square Error (MSE), and Tables 4, 5, and 6 the corresponding Mean Absolute Error (MAE) ranks for N across T. The Between Median estimator (BMD) with the lowest values ranked first for different levels of missingness (0%-20%), the Between estimator ranked second, while Pooling estimator with the highest values of MSE ranked last.

Also, for all the levels of missingness employed, it was observed that other estimators have no consistent pattern of ranking as observed in Figures 1and

3.1.1. The behavior of the six estimators using the Mean Square Error when N =20, 25, and 30

Table 1: **Mean Square Error Result in Order of Ranking, N=20, β=3, T=5**

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

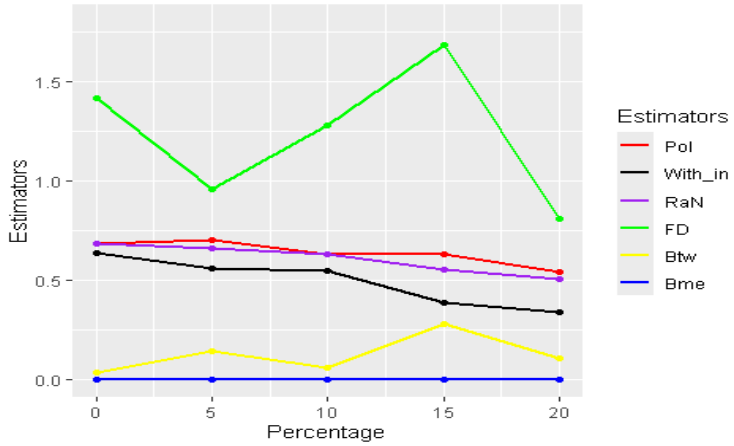


Figure 1: **Mean Square Error (MSE) for panel data size, N =20, n=4, beta=3, and at percentages (0,5,10,15,20)**

Pol-Pooling estimator, With_in –Within estimator, RAN –Random estimator, FD –First Difference, estimator, BTW-Between estimator, Bme –Between Median estimator.

Table 2: **Mean Square Error Result in Order of Ranking, N=25, β=3, T=5**

MSE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0**	**5**	**10**	**15**	**20**
	5	25	20	16	11	11
Pooling		5th	5th	5th	4th	4th
Within		3rd	3rd	3rd	3rd	3rd
Random		4th	4th	4th	4th	4th
First Difference		6th	6th	6th	5th	5th
Between		2nd	2nd	2nd	2nd	2nd
Between Median		1st	1st	1st	1st	1st

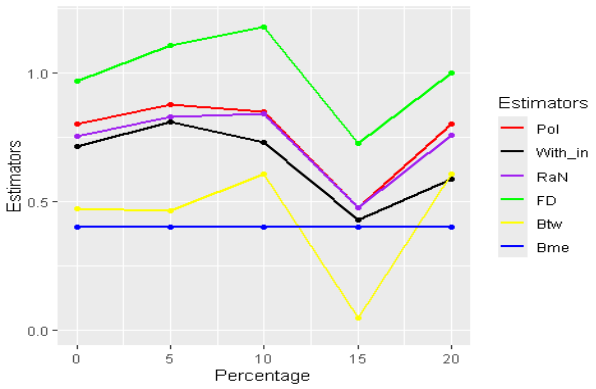


Figure 2: **Mean Square Error (MSE) for panel data size, N =25, n=5, beta=3, and at percentages (0,5,10,15,20)**

Table 3: **Mean Square Error Result in Order of Ranking, N=30, β=3, T=6**

MSE	n	N	N	N	N	N
%		0	5	10	15	20
N=30, β=3, T=6	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

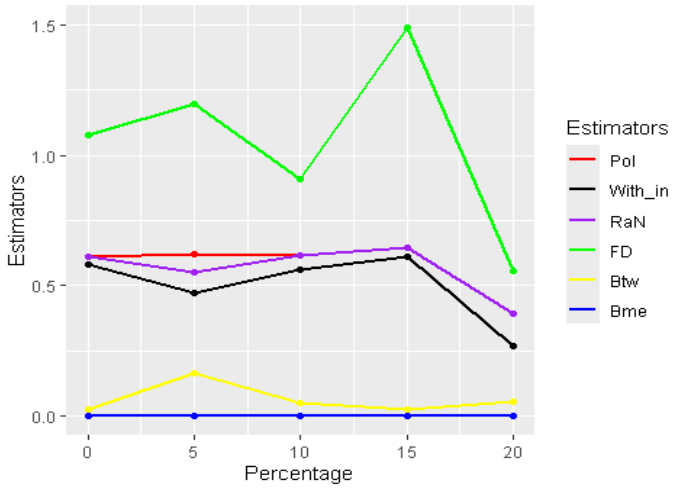


Figure 3: **Mean Square Error (MSE) for panel data size, N =30, n=5, beta=3, at percentages (0,5,10,15,20)**

Similarly, the pattern of ranking shown for MSE in Tables 1, 2, and 3, and the corresponding MAE order of ranking in Tables 4, 5, and 6 for N across T were plotted in the graphs shown in Figures 1, 2, and 3 for MSE and Figures 4, 5, and 6 for MAE. The Between Median estimator (BMd) ranked first for different levels of missingness (0%-20%) with the lowest MSE and MAE values. The Between estimator ranked second, while the Pooling estimator ranks last with the highest MSE and MAE values. Furthermore, other estimators did not show a consistent pattern of ranking when N is (20, 25, and 30).

3.1.2. The behavior of the six estimators using the Mean Absolute Error N =20, 25, and 30

Table 4: **Mean Absolute Error Result in Order of Ranking, N=20, β=3, T=5**

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

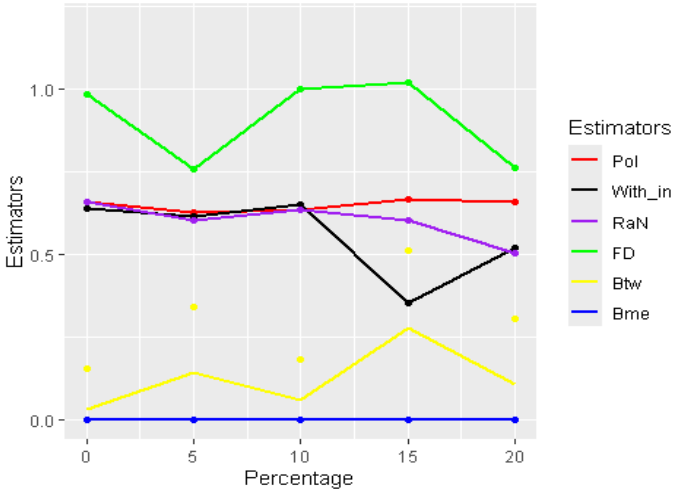


Figure 4: **Absolute Mean Square Error (MAE) for panel data size, N =20, n=4, beta=3, and at percentages (0,5,10,15,20)**

Table 5: **Mean Absolute Error Result in Order of Ranking, N=25, β=3, T=5**

MAE	**n**	**N**	**N**	**N**	**N**	**N**
%		**0**	**5**	**10**	**15**	**20**
	5	25	20	16	11	11
Pooling		5th	5th	5th	4th	5th
Within		3rd	3rd	3rd	3rd	3rd
Random		4th	4th	4th	4th	4th
First Difference		6th	6th	6th	5th	6th
Between		2nd	2nd	2nd	1st	2nd
Between Median		1st	1st	1st	2nd	1st

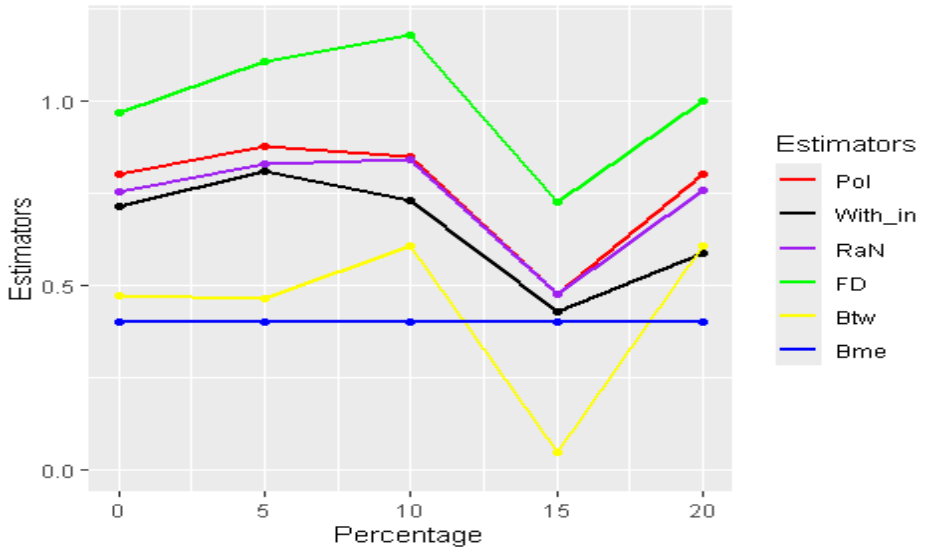


Figure 5: **Absolute Mean Square Error (MAE) for panel data size, N =25, n=5 beta=3, and at percentages (0,5,10,15,20)**

Table 6: **Mean Absolute Error Result in Order of Ranking, N=30, β=3, 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

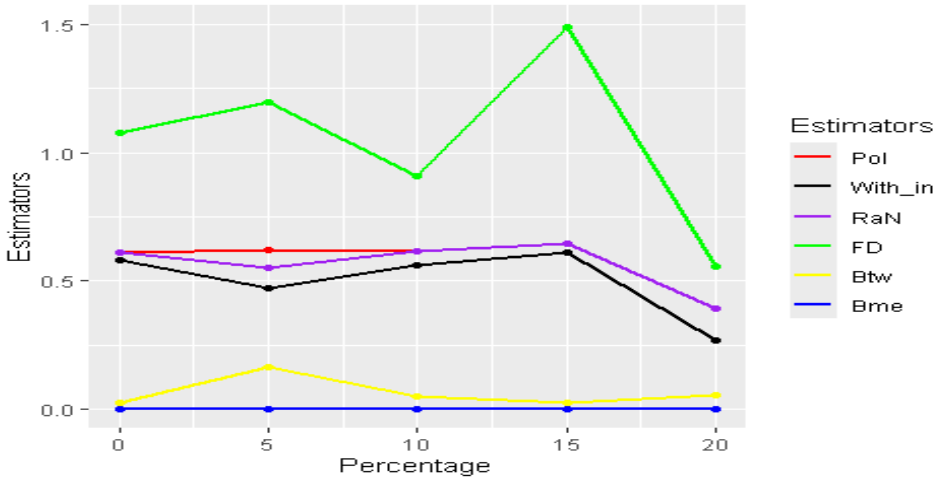


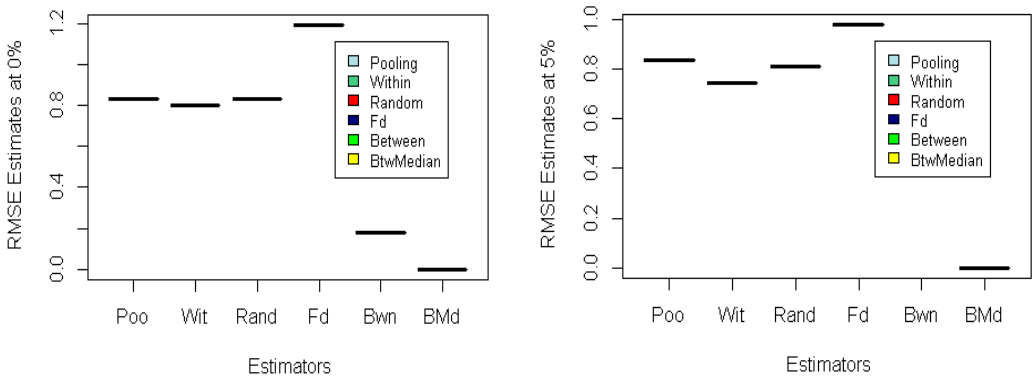
Figure 6: **Mean Absolute Error (MAE) for panel data size, N =30, n=5, beta=3, at percentages (0,5,10,15,20)**

Over all, the graphs plotted show that the Between Median estimator is consistent and efficient with low values of N sizes, N (20, 25, and 30) across T (4,5and 6).

3.1.2. Model Selection

This section demonstrates how effectively the models predicted the variations in responses. Root Mean Square Error (RMSE) was employed to compare the accuracy of the panel data models assessed. The boxplots in figures 7, 8, and 9 clearly illustrate that the Between Median estimator appears to be consistent for different N (20, 25, and 30) over T (4, 5, and 6), with varying degrees of missing values. The metric assessed how well the predicted values matched the actual values. Thus, the Between Median estimator with the lowest value shows how well the model fits the data.

The box-plot legend is described as: Poo -Pooling estimator, Wit –Within estimator, Rand –Random estimator, Bwn –Between estimator, BMd –Between Median estimator.



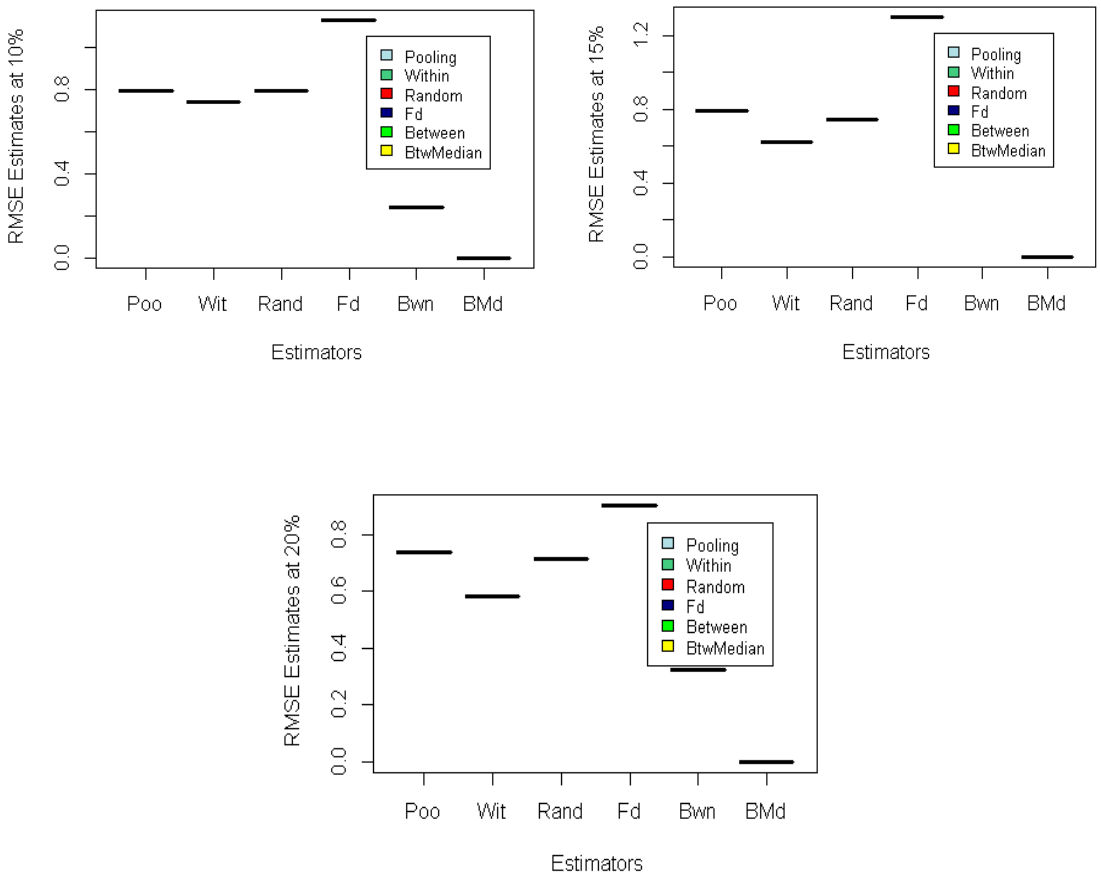
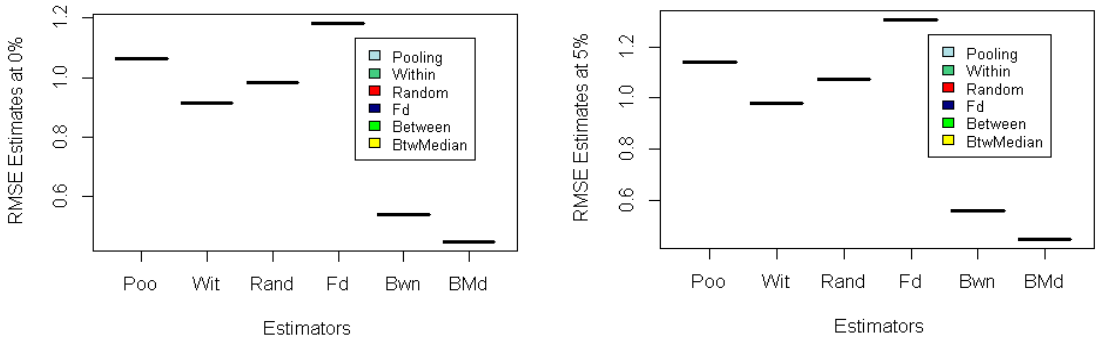


Figure 7: **The Boxplot for Root Mean Square Error (RMSE) of all the six estimators including the proposed BME estimator when the sample size when N = n* T = 20 at 0%- 20%.**



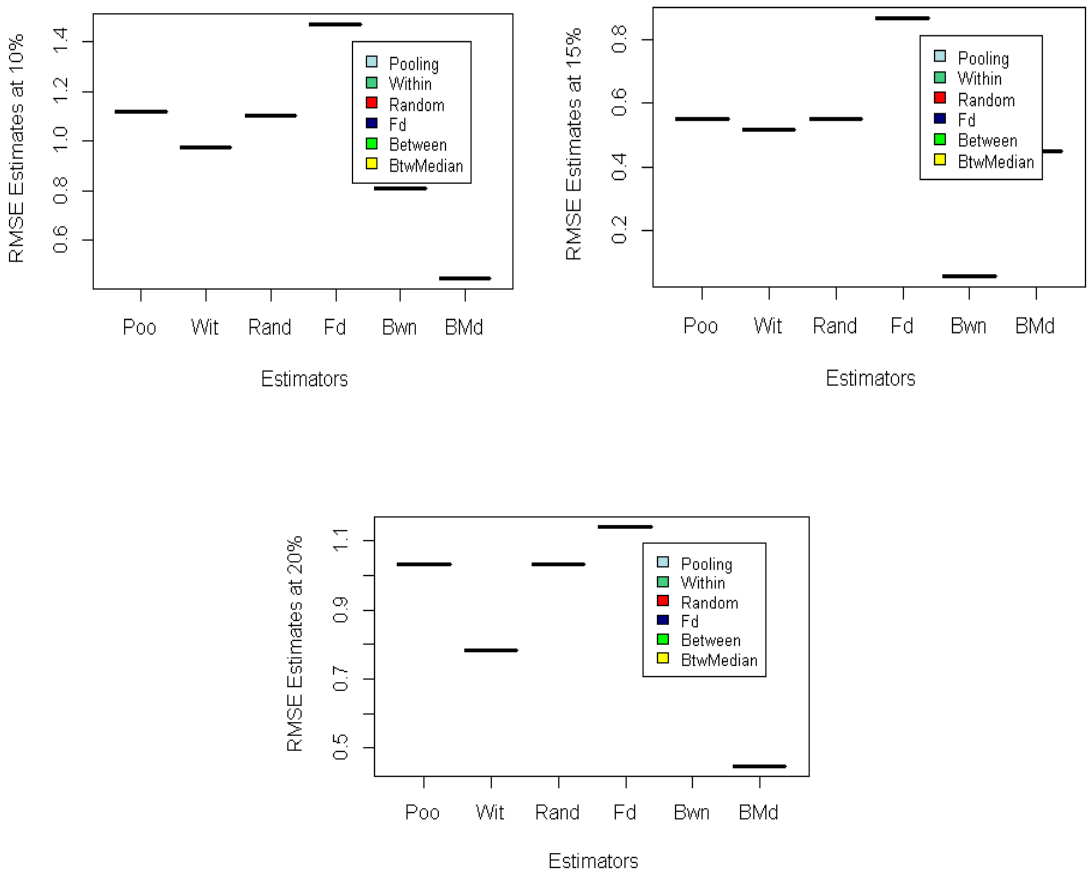
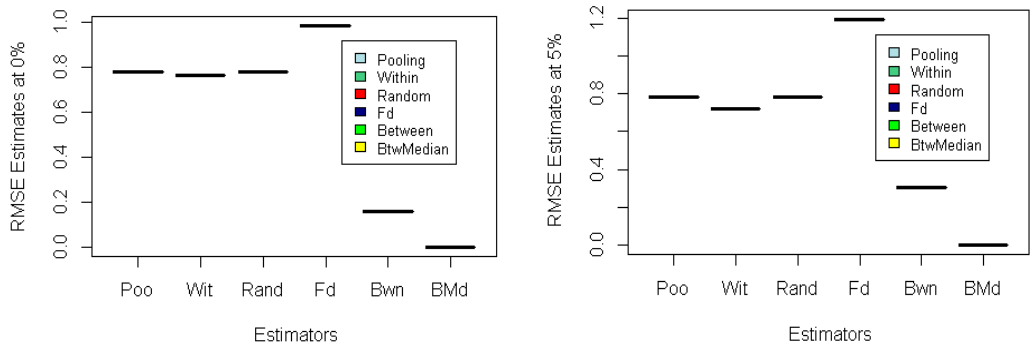


Figure 8: **The Boxplot for Root Mean Square Error (RMSE) of all the six estimators including the proposed BME estimator when the sample size when N = n* T = 25 at 0%- 20%**



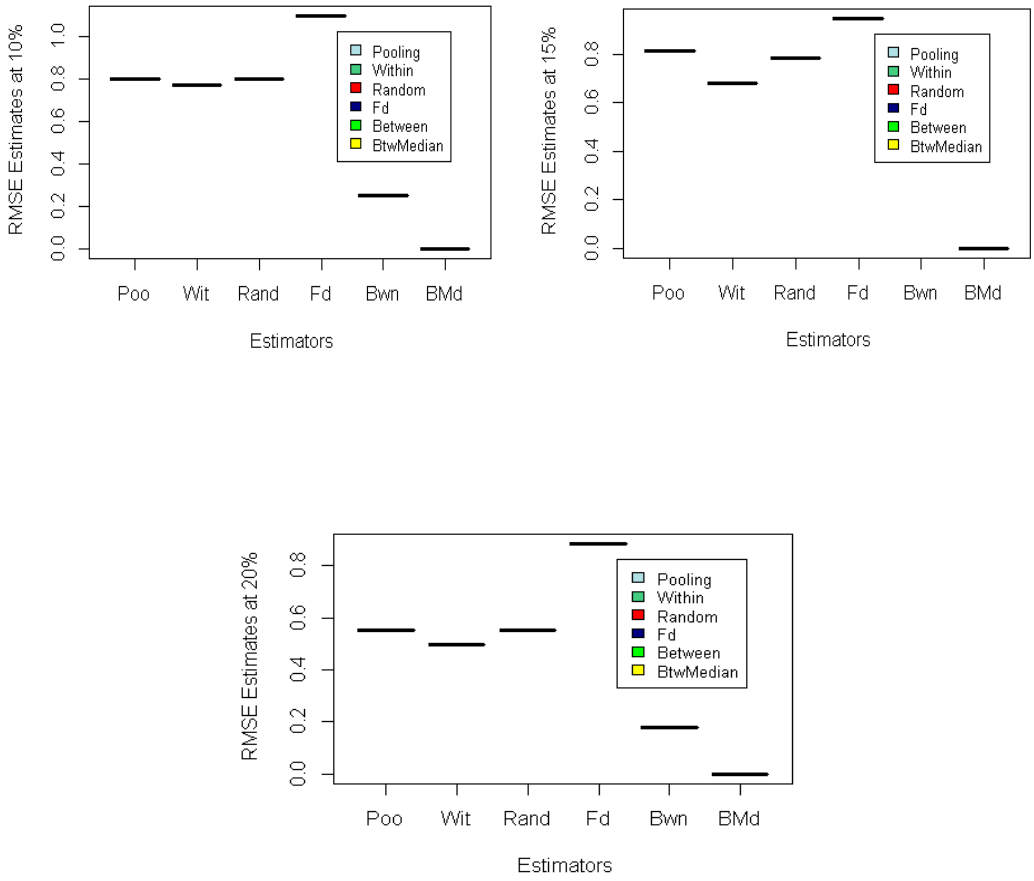


Figure 9: **The Boxplot for Root Mean Square Error (RMSE) of all the six estimators including the proposed BME estimator when the sample size when** N = n*** T = 30 at 0%- 20%**

3.2. Application

The data used for this paper is the top five African Countries data ranked by Gross Domestic Product (GDP) at Purchasing Power Parity (PPP), current price. According to International Monetary Fund (IMF) these countries which were sorted in ascending order are: Egypt, Nigeria, South Africa, and Ethiopia.[33].

The data represents Penn World Table data of the Expenditure-side real GDP at current PPPs (in mil. 2017US$) (RGDP(e)) and the Real GDP at constant 2017 national prices (in mil. 2017US$) RGDP (na)) for these countries between year (1995 - 1999). [34]. The data described in [35] which was adopted by [36] was considered for this study.

Table 7: **Description of Datasets Used to Generate Population Parameters**

Dataset	***Source***	***Dependent Variable***	***Independent Variables***	***N***	***T***
		Groningen Growth and Development Centre			
Penn World Table for Algeria, Nigeria, Egypt, South Africa and Ethiopia		Log of Real GDP at constant 2017 national prices (in mil. 2017US$)	Ratio of Expenditure-side real GDP at chained PPPs (in mil. 2017US$) to Real GDP at constant 2017 national prices (in mil. 2017US$) Nigeria (X)	5, 10, 20, 50, 77	5, 10, 15, 20, 25

Tables 8 and 9 represent the MSE and the MAE estimates derived from the log of Real GDP at constant 2017 national prices in US dollars, which is the dependent variable (y); and the ratio of the Expenditure-side real GDP at chained PPPs at constant 2017 national prices in US dollars, that is the independent variable (x) between the year (1995- 1999)

Table 8: **Mean Square Error for the Real-Life Data N= 25**

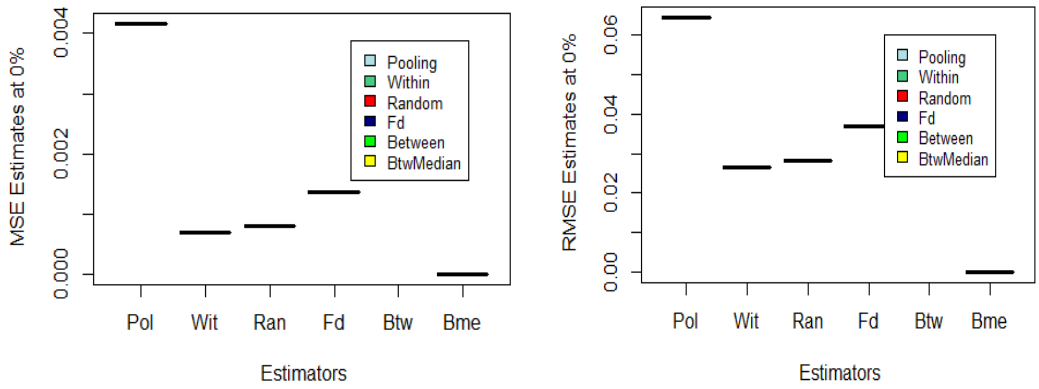
MSE %	N 0	N 5	N 10	N 15	N 20
N=25	20	17	13	10	6
Pooling	4.164053 e-03	4.131529e-03	3.871167 e-03	2.602326 e-03	2.382442 e-03
Within	6.990777 e-04	7.091882 e-04	1.190229 e-03	3.663067 e-04	3.664713 e-04
Random	8.043466 e-04	8.398876 e-04	8.673234 e-04	5.886533 e-04	6.673329 e-04
First Difference	1.363227 e-03	2.004461 e-03	1.957130 e-03	7.973587 e-04	7.971794 e-04
Between	3.119414 e-03	2.874384 e-03	3.630064 e-03	5.236522 e-03	2.377882 e-03
Between Median	3.944305 e-31	3.944305 e-31	3.944305 e-31	3.944305 e-31	3.944305 e-31

Table 9: **Mean Absolute Error for the Real-Life Data N= 25**

MAE %	N 0	N 5	N 10	N 15	N 20
N=25	20	17	13	10	6
Pooling	4.903188 e-02	4.949588 e-02	4.800512 e-02	3.280266 e-02	4.903188 e-02
Within	2.088891 e-02	2.035116 e-02	2.451817 e-02	1.310500 e-02	1.619621 e-02
Random	2.224726 e-02	2.116436 e-02	2.748063 e-02	1.938224 e-02	1.687865 e-02
First Difference	2.594057 e-02	3.119219 e-02	3.325788 e-02	1.842664 e-02	2.269920 e-02
Between	4.043781 e-02	3.804940 e-02	4.434387 e-02	5.626120 e-02	3.266876 e-02
Between Median	3.972055 e-16	3.944305 e-31	3.972055 e-16	3.972055 e-16	3.972055 e-16

3.2.1 Box-plot for the Real-Life Dataset

Figures 47-61 show the box-plots plotted for the MSE, RMSE and the MAE for the real-life dataset. The plot displays the efficiency of the estimators at different degrees of missingness. The new estimator, Between Median Estimator, with the lowest estimate was found to outperformed the other existing panel data estimators considered for this study. It is therefore concluded that the real-life dataset fit the new estimator (BME).



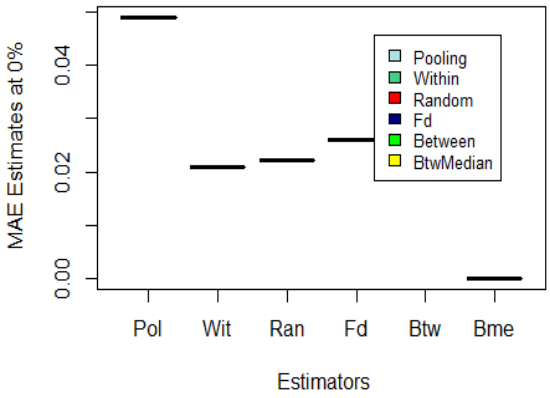
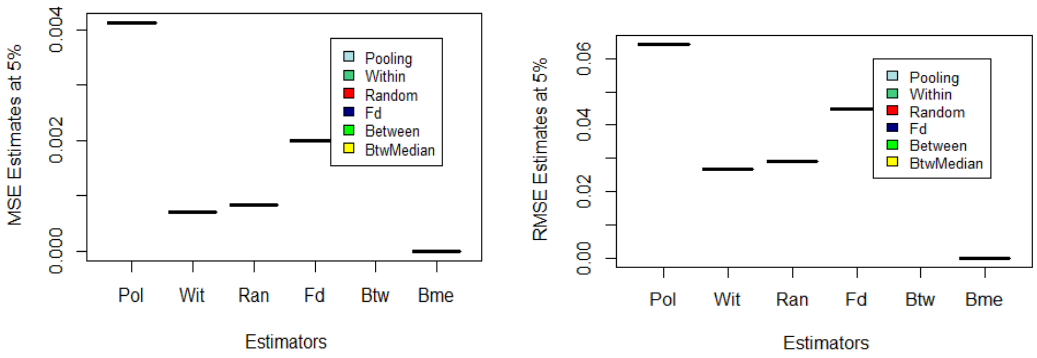


Figure 10: **Box-plot for MSE, RMSE and MAE for the Real-life data at 0% level of missingness**



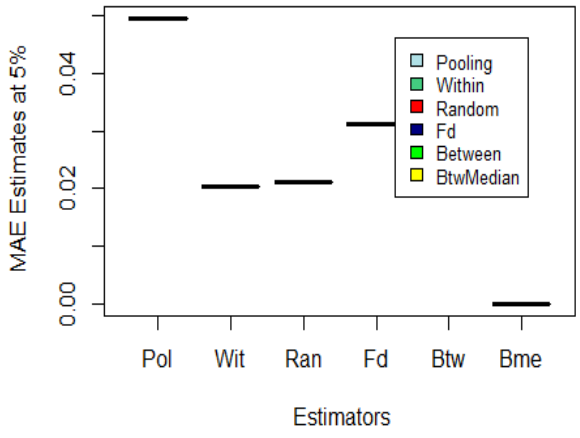
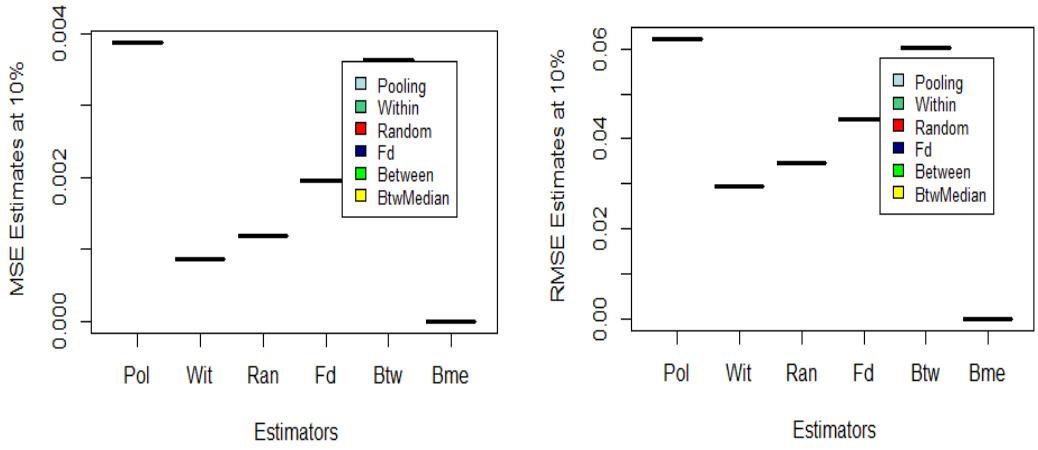


Figure 11: **Box-plot for MSE, RMSE and MAE for the Real-life data at 5% level of missingness**



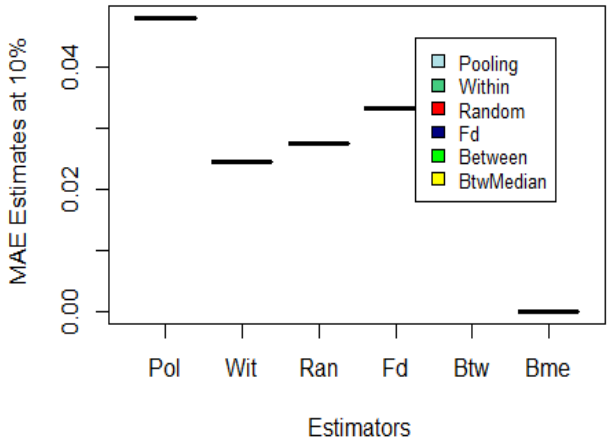
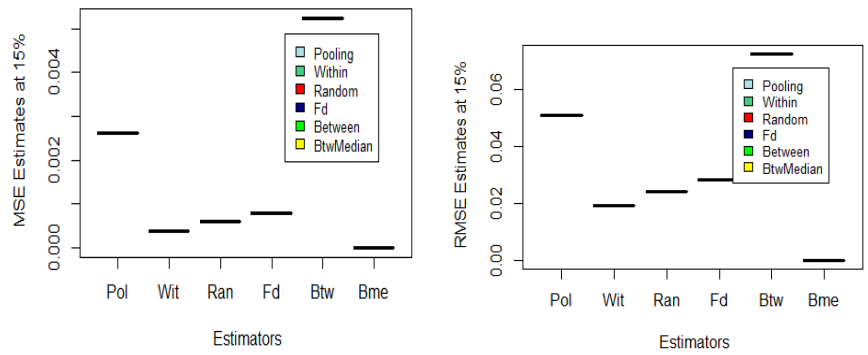


Figure 12: **Box-plot for MSE, RMSE and MAE for the Real-life data at 10% level of missingness**



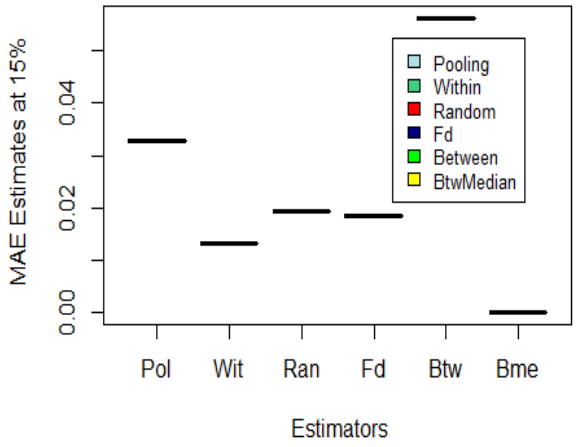


Figure 13: **Box-plot for MSE, RMSE and MAE for the Real-life data at 15% level of missingness**

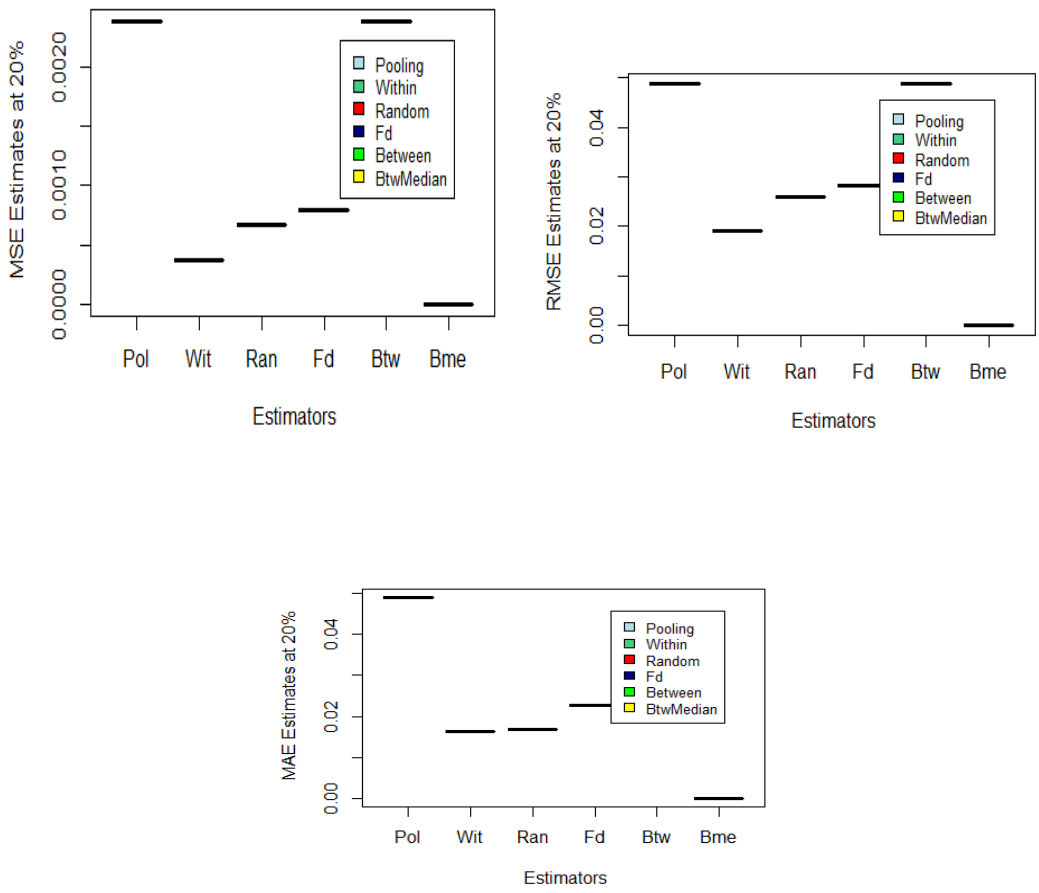


Figure 14: **Box-plot for MSE, RMSE and MAE for the Real-life data at 20% level of missingness**

3.2.2. Plot of Mean Square Error (MSE) and Absolute Mean Square (MAE) for Real data at percentages (0,5,10,15,20)

Figures 15 represents the plot of the MSE and the MAE for the real-life data, the plots show that, the new estimator developed, that is the BMD is consistent at different levels of missingness. The new BMD estimator is conclusively fit for the real-life data.

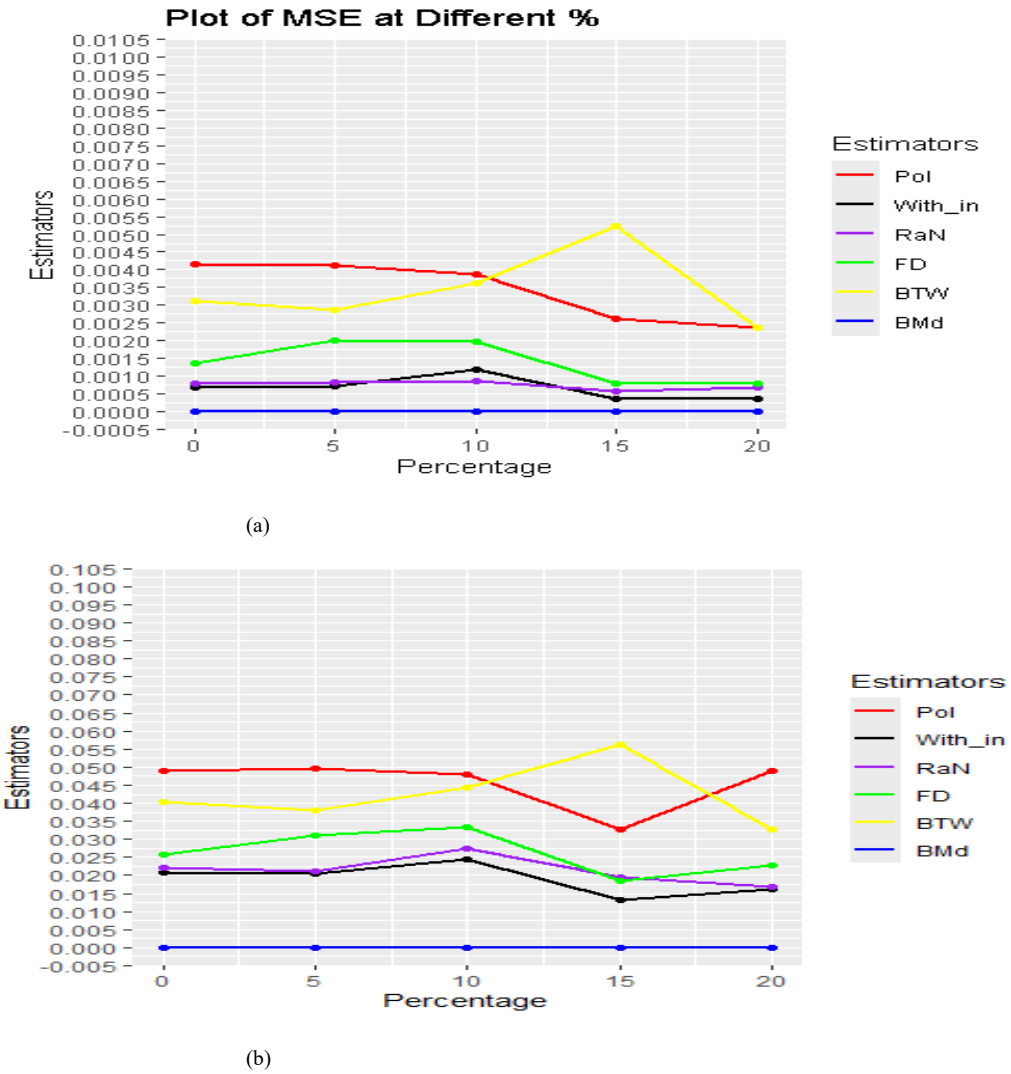


Figure 15: **The MSE (a) and the MAE (b) for data Real-Life data at percentages (0,5,10,15,20)**

4. Conclusions

This study focused on the panel data to use for an unbalanced panel dataset for small sample sizes. The findings based on the results obtained from a Monte Carlo simulation show that among the six estimators examined, *Between* Median (BMD) estimator consistently outperforms its counterparts in managing unbalanced panel data under varying degrees of missingness when the panel data structure was adjusted for N, T and β. The parameters for Generalized Least Square (GLS) are consistent and efficient estimator.

Using the real-world panel data of the top five African Countries data ranked by Gross Domestic Product (GDP) at Purchasing Power Parity (PPP), current price; the Between Median Estimator (BMD) outperformed the other five estimators tested for balance panel models under unbalanced panel by regressing the median of Y dependent variable on the median of the X independent variable.

The findings offer valuable guidance for the application of this estimator in empirical research, reinforcing the significance of model selection and its impact on the validity of conclusions drawn from panel data.

It is therefore concluded that the Between Median estimator was best for the balanced and the unbalanced dataset and its best fitted for the panel data structures considered.

CRediT authorship contribution statement. Author 1: Methodology, Investigation. Author 2: Formal analysis, Supervision. Author 3: Formal analysis, Data curation

Declaration of competing interest. The authors declare that there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

1. Balogun, O. P., & Yahya, W. B. Development of Estimation Methods for Modeling Unbalanced Panel Data for Small Data Occasioned by Missingness. Professional Statisticians Society of Nigeria, Edited Proceedings of 8th International Conference. (2024). 8(1).
2. Balogun, O. P., & Yahya, W. B. & Issa, A. A. Evaluation of panel data estimators under the unbalanced panel datafor small data sizes occasioned by missingness. International Journal of Statistics and Applied Mathematics 2024; 9(6): 144-149. DOI: https://doi.org/10.22271/maths. (2024) .v9.i6b.1914
3. Khan S, Ponomareva M, Tamer E. Identification of panel data models with endogenous censoring. J. 89oEconometr. 194:(1):57–75. (2016).
4. Bartels, B. Beyond Fixed versus Random Effects: A Framework for Improving Substantive and Statistical Analysis of Panel, Time-Series Cross-Sectional, and Multilevel Data. Society for Political Methodology. (2009).
5. de Blok, L. Who Cares? Issue Salience as a Key Explanation for Heterogeneity in Citizens' Approaches to Political Trust. Social Indicators Research. 171. 1-20. (2023). https://doi.org/10.1007/s11205-023-03256-w.
6. Wooldridge. J. M. Introductory Econometrics. A Modern Approach Fifth Edition. (2012).
7. Balogun, O. P., Yahya, W. B., & Umar-Mann A. Performance Evaluation of Some Estimators under Unbalanced Panel Data Models. Professional Statisticians Society of Nigeria, Edited Proceedings of 6th International Conference. (2022). 6(1).
8. Arsham, H. Techniques for Monte Carlo Optimizing. Monte Carlo Methods and Applications, 4(3), 181-230. (1998). https://doi.org/10.1515/mcma.1998.4.3.181
9. Rollans, S., & McLeish, D. L. Estimating the Optimum of a Stochastic System using Simulation. Journal of Statistical Computation and Simulation, 72(5), 357–377. (2002). https://doi.org/10.1080/00949650213533.
10. Tolk, A. et al. Philosophy and Modeling and Simulation. In: Ören, T., Zeigler, B.P., Tolk, A. (eds) Body of Knowledge for Modeling and Simulation. Simulation Foundations, Methods and Applications. Springer, Cham. (2023). https://doi.org/10.1007/978-3-031-11085-6_16
11. Santos, L. & Barrios, E. Small Sample Estimation in Dynamic Panel Data Models: A Simulation Study," Open Journal of Statistics, Vol. 1 No. 2, 2011, pp. 58-73. (2011). https://doi.org/10.4236/ojs.2011.12007.
12. VanVoorhis, C. W., & Morgan, B. L. Understanding power and rules of thumb for determining sample sizes. Tutorials in quantitative methods for psychology, 3, (2), 43-50. (2007).
13. Memon, M. A., Ting, H., Cheah, J. H., Thurasamy, R., Chuah, F., & Cham, T. H. Sample Size for Survey Research: Review and Recommendations. Journal of Applied Structural Equation Modelling, 4(2), 1-20. (2020). https://doi.org/10.47263/jasem.4 (2), 01.
14. Indrayan, A. and Mishra, A. The importance of small samples in medical research. (2021).
15. Hansen, Christian B. Generalized Least Squares Inference in Panel and Multilevel Models with Serial Correlation and Fixed Effects. Journal of Econometrics 140 (2) 670–694 https://doi:10.1016/j.jeconom.2006.07.011. JPostgradMed.67(4):219–223. (2007). https://doi:10.4103/jpgm.JPGM_230_21.
16. Suryanta, B., & Patunru, A. A. Trade Impediments in Indonesia. Journal of Economic Integration, 38(2), 247–277. (2023). https://www.jstor.org/stable/27217155.
17. Greene, W. H. (2008). Econometric Analysis.
18. Garba M. K., Oyejola, B. A., & Yahya, W. B. Investigations of Certain Estimators for Modeling Panel Data Under Violations of Some Basic Assumptions. 3(10). (2013).
19. Amemiya, T. (1971). The estimation of the variances in a variance-components model, International Economic Review 12, 1-13.
20. Matyas, L., & Sevestre, P. (1992). The Econometrics of Panel Data, 46-71. Kluwer Academic Publish
21. Arellano, M. 2003: Panel Data Econometrics.
22. Baltagi, B. H. (2005). Econometric Analysis of Panel Data, John Wiley and Sons, England.
23. Swarmy. P. A. V. B. & Arora. S. S. (1972). The Exact Finite Sample Properties of the Estimators of Coefficients in the Error Components Regression Models. Econometrica, 40(2,), pp. 261-275.
24. Cottrell, A. (2017). Random effects estimators for unbalanced panel data: a Monte Carlo analysis using gretL
25. Verbeek, M. (2021). Panel Methods for Finance: A Guide to Panel Data Econometrics for Financial Applications, 1 of De Gruyter studies in the practice of econometrics, ISSN 2570-0928.
26. Robeson. S. M., Willmott, C. J. Decomposition of the mean absolute error (MAE) into systematic and unsystematic components. PLoS ONE 18(2): e0279774. (2023). https://doi.org/10.1371/journal.pone.0279774.
27. Arslanoglu, N. Empirical modeling of solar radiation exergy for Turkey Faculty of Engineering, Mechanical Engineering Department, Uludag University, Gorukle Campus, TR-16059 Bursa, Turkey (2016).
28. Ruppert, D. & Carroll, R. (1988). Transformation and Weighting in Regression. Chapman and Hall, London and New York.
29. Davidson, R., MacKinnon, J. G. (2004). Econometric Theory and Methods. New York: Oxford University Press.
30. Enders C. K. Using the Expectation Maximization Algorithm to Estimate Coefficient Alpha for Scales with Item-Level Missing Data. Psychol Meth, 8(3):322–337. (2003). doi: 10.1037/1082-989X.8.3.322.
31. Biørn, E. (2010). How is generalized least squares related to Within and Between Estimators in Unbalanced Panel Data? Department of Economics, University of Oslo. P.O. Box 1095. Blindern, 0317 Oslo, Norway. E-mail: erik.biorn@econ.uio.no.
32. Jirata, T. M., Chelule, J. C., & Odhiambo, R. O. Deriving Some Estimators of Panel Data Regression Models with Individual Effects. (2015).
33. World Economic Outlook (2024) - GDP, current price. www.imf.org. Retrieved 2024-11-10.
34. Feenstra, R. C., Inklaar, R., and Marcel, P. T. (2015). The Next Generation of the Penn World Table. American Economic Review, 105(10), 3150-3182, www.ggdc.net/pwt.

35. Reed, W.R. & Ye, H. (2011). Which panel data estimator should I use? *Applied Economics*, 43(8), pp. 985-1000.
36. Balogun, O. P., Yahya, W. B., & Umar-Mann A. (2022) Performance Evaluation of Some Estimators under Unbalanced Panel Data Models. *Professional Statisticians Society of Nigeria, Edited Proceedings of 6th International Conference.* (2022). *6(1).*