Performance Evaluation of Some Estimators under Unbalanced Panel Data Models

O. P. Balogun[1]; W. B. Yahya[2]; A. Umar-Mann[3]

[1,3]Department of Statistics,
Federal Polytechnic, Bida, Nigeria.

[2]Department of Statistics,
, University of Ilorin, Nigeria.
E-mail: omoshadebalogun@gmail.com [1]

***Abstract* — This work investigates the efficiencies of five estimators of panel data models under unbalanced data structure triggered by the presence of missing values. The methods considered are the Between, Random (Swamy-Arora), First Difference, Pooling, and Within estimators. In the Monte-Carlo experiments, unbalanced sample panel data were generated with 5% missingness at random using a published balanced panel dataset with five sample units (n) each measured at an equal time interval of five (t). This was done to inject missingness in time (t) or in sample unit (n) or both, thereby creating an unbalanced data structure. The performances of these five estimators were evaluated using the Mean Square Error (MSE) and the Mean Absolute Error (MAE). The results showed that Between estimator, with the least values of MSE and MAE, proved to be the best estimator for Panel data model under an unbalanced data structure. In terms of the order of performances, further results showed that the Within estimator was the second-best followed by the Random estimator with the Pooling estimator at the First Difference having the least performance for estimating unbalanced panel data model, especially under the small sample size situations. This study recommends that the Between estimator should be adopted for fitting the panel data models when evidence of missingness is apparent in the data, especially when the number of sample units is very small.**

Keywords - *Panel data, missingness, unbalanced panel data, mean square error, mean absolute error.*

I. INTRODUCTION

Panel data refers to data sets consisting of multiple (repeated) observations on each sampling unit. A panel data set is one where observations are obtained on the same set of entities at several periods. This could be generated by pooling time-series observations across a variety of cross-sectional units including countries, states, regions, firms, or randomly sampled individuals or households. Panel data set covers a much larger sample and is representative of all demographic groups. Baltagi (2014)

Missing data constitute a major problem in the behavioral sciences, particularly when data collection is costly or involves destructions. Rubin et. al (2007). The general approach that is often adopted by researchers is to delete cases with missing observation(s). This approach can result in biased estimates and reduce power of the statistical tests used to analyze the data. Trying to avoid the deletion of a case because of a missing data point can be conducted, but implementing a naïve missing data method can result in distorted estimates and incorrect conclusions.

Many literatures established various estimators for analyzing panel data under unbalanced panel data for error component models. Error component model is synonymous or a byword for random effect. These considered unbalanced one-way model, unbalanced two-way model and unbalanced nested model. Estimating panel data with missing values is a growing interest of many researchers particularly in Medicine and Economics. Few of these literatures are, Mayer (2010); Graham (2009) Young and Johnson (2015); Kang (2015); Cottrell (2017) and Lee et. al (2021).

Also, literature has shown that 'missingness' in panel data set leads to unbalanced panel data set. Bruno (2013). This literature also, reveals that a theoretical analysis of the algorithm has been
proposed in literature to estimate Error Component Models (ECM) for unbalanced panel data.

Estimating missingness could be under different types of missingness, that is Missing at Random (MAR), Not Missing at Random (NMAR) and Missing completely At Random (MCAR). An interesting point is that a data could be removed due to missing values in panel data. This attrition leads to bias estimates of panel data parameters as it reduced the number of observations in the studies. Some

literature that reviewed the mechanism of types of missingness are: Hedeker and Gibbons (2006); Nijman and Verbeek (1992); Schafer (1997); Rubin et al (1981).

This work is to estimate panel data set under unbalanced data set using existing panel data models in R-package. Two error criteria, Mean Square Error and Mean Absolute Error are employed to investigate the most efficient estimator among the five estimators considered..

II. METHODOLOGY

This work focus on estimating panel data set under unbalanced data set. Previous studies considered estimating unbalanced panel models under error component models (EMC). The EMC is a byword for "random effect" , Croissant and Millo (2018). In general, parameter estimation in the regression analysis with cross-section data is done by estimating the least squares method, the Ordinary Least Square (OLS), Zulfikar et al. (2019). The method of estimating the regression model in panel data is done using three approaches, these are the common effect model or the Pooled Least Square which uses OLS, the Fixed Effect Model, and the Random Effect Model.

The methods of estimating the coefficients and EMC (random effect model) of unbalanced data set among others are ANOVA which is best quadratic unbiased estimator (BQU), quadratic unbiased estimator (QUE), ANOVA-type feasible GLS, MLE, ML, MINQUE and MIVQUE. Mayer (2010). This work is to examine the coefficients of the parameters of five panel data models. The Between Estimator, the Within Estimator, the Random (Swamy-Arora's) Estimator, the Pooling Estimator and First Difference Estimator; and to investigate the performances of these five estimators using the Mean Square Error and Mean Absolute Error criteria.

In general, the panel data model is represented as follows

$$y_{it} = \beta_0 + \beta^T x_{it} + \varepsilon_{it} \quad (1)$$

$$= \beta_0 + \beta^T x_{it} + \rho_i + \mu_t + \gamma_{it} \quad (2)$$

where i and t are the individual and time indexes, y the response, x a vector of covariates, β_0 the overall intercept andβ^T the vector of parameters of interest that we are willing to estimate.
Equation (2) is the unbalanced panel data model where the error term ε_{it} is decomposed into three elements (in the two-way case)

ρ_i is the individual effect,
μ_t is the time effect,
γ_{it} is the idiosyncratic error.

2.1 Five Panel Data Estimators under studies

The theoretical review of the five estimators considered in this study is presented in this section.

2.1.1 Pooling Estimator: This is an OLS estimation equivalent to lm and "walhus". Wallace and Hussain (1969); Baltagi (2010)

$$y_{it} = \beta_0 + \beta^T x_{it} + \varepsilon_{it}; \quad t = 1,2,\dots T\,; i = 1,2,\dots,N \quad (3)$$

where:
y_{it} is the observation on the dependent variable for the ith individual at the tth time period,
x_{it} is ith observation on a vector of k nonstochastic regressors
$\beta^T$is a $kx1$ vector of regression coefficients
β_0 is the intercept.
The pooled estimator is given as:

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

where,
y is an $nT \times 1$ column vector of a dependent variable,
X is an $nT \times k$ square matrix of regressors,
β is a (k+1) × 1column vector of regression coefficients,
w is an $nT \times 1$ column vector of the combined error terms (i. e. ε_i+μ_{it}). Garba et al. (2013)

2.1.2 Between Estimator: This estimator performs the estimation on the individual or time mean. It explicitly converts all the observations into individual-specific averages and performs OLS on the transformed data. Averaging model (3) over t gives:

$$\bar{Y}_{i.} = \alpha + \beta_1 \bar{X}_{1i.} + \beta_2 \bar{X}_{2i.} + w_{it} \quad (5)$$

Generally,

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

The Between estimator ignores all of the individual-specific variation in y and X that is considered by the Within estimator, replacing each observation for an individual with their mean behavior. Baum (2013).

2.1.3 Within Estimator: This is equivalent to "amemiya". T. Amemiya (1971); Matyas and Sevestre (1992). This regresses on the deviations from the individual or/and time mean.

$$y_{it} = X_{it}^{*}\beta^{*} + Z_{j\alpha} + \varepsilon_{jt} \quad (6)$$

where,

$$\varepsilon_{jt} = \alpha_i + u_{it} \quad (7)$$

The X_{it}^{*} matrix does not contain a unit's 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's vector, then pooled OLS is a consistent and efficient estimator of $[\beta^{*}\ \alpha]$.

2.1.4 First Difference Estimator: This model is equivalent to "fd". This model regresses on the first differences of the mean of an individual unit i over time t. (Arellano, 2003; Baltagi, 2005)

The First Difference model is given as:

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

where,

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

and

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

2.1.5 Random Estimator: This is equivalent to "swar" models. Swamy and Arora (1972); Cottrell (2017). " According to Hauser, the estimator follows the underlying model expression:

$$\propto_i \sim iid\ N(0, \delta_{\propto}^2)$$

$$y_{it} = \beta_0 + X_{it}'\beta + \propto_i + u_{it}, u_{it} \sim iid\ (0, \delta_u^2) \quad (9)$$

The $\propto_i$'s are random variables with the same variance. The value $\alpha_i$is specific for individual i. The $\propto$'s of different individuals are independent, have a mean of zero, and their distribution is assumed to be closed o normal. The overall mean is captured in β_0. $\propto_i$ is time-invariant and homoscedastic across individuals. There is only one additional parameter $\delta_{\propto}^2$. Only $\propto_i$ contributes to Corr $(\epsilon_{i,s}, \epsilon_{i,t})$. $\propto_i$ determines both $\epsilon_{i,s}$ *and* $\epsilon_{i,t}$

The switch between OLS and FE is anchored on the covariance between the alpha and the independent variable(s). If the covariance is zero (i.e., very small) it means that there is no correlation and OLS is preferred, however, if the covariance is not zero or greater than zero or large there is correlation and FE should be preferred. Brugger (2021)

$$y_{it} = \beta_0 + X_{it}'\beta + \alpha_i + u_{it}, u_{it} \sim iid\ (0, \delta_u^2) \quad (10)$$

where $t = \lambda, \ldots, T$ and $i = 1, \ldots, N$

$$Cov\ (\alpha_i, X_{it}) \neq 0 \sim FE - model \quad (11)$$

$$Cov\ (\alpha_i, X_{it}) = 0 \sim FE - OLS \quad (12)$$

Also, if

$$\lambda = 1 - \left(\frac{\delta_\mu^2}{\delta_\mu^2 + T.\ \delta_\alpha^2}\right), \quad (13)$$

$$\lambda = 1 \sim FE\ model \quad (14)$$

$$\lambda = 0 \sim OLS \text{ model.} \quad (15)$$

2.2. Simulation

This project work adapts the scheme adopted by Reed and Ye (2011) with some little modifications. The R-package was used for the simulation and analysis.

The following settings were used in the simulation task: For a total of $n = 5$ subjects were studied over T = 5 times. Thus, a total of N = 25 $(n \times T)$ observations were generated for the data. The panel data model considered is of the form:

$$y_{it} = \beta_0 + \beta_1 X_{it} + e_{it}, \quad (16)$$

where:

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

The X_{it} was simulated from Gaussian population with the following: $X_{it} \sim iid\ N(20, 1)$. The error term e_{it}, was simulated from $e_{it} \sim iid\ N(0, 1)$. The parameters β_0 and β_1 in model (16) were set at: $\beta_0 = 20$ and $\beta_1 = 3$.

The vectors y_{it} and X_{it} values are then used to obtain estimates of β_x for each of the estimators under study.

Unbalance time intervals were infused into the data by randomly removing 5% of the total sample from the data. Population parameter values for the DGPs in the Monte Carlo experiments.

2.3 Measurement Criteria

2.3.1 Mean Square Error: The MSE of the residuals of each Model Estimator considered in this study was calculated. The behavior of each estimate was observed for both Balanced and Unbalanced panel data to know how they weigh on the error.

The MSE equation is given as:

$$MSE = \frac{\sum_i^n (y_i - \hat{y}_i)^2}{N} \quad (17)$$

where *N* is the number of samples we are testing against. Seif (2019).

If $MSE(\hat{\theta}_1) < MSE(\hat{\theta}_2)$ (18)

then $\hat{\theta}_1$ *is* said to be more efficient than $\hat{\theta}_2$. Veroniki and Salanti (2013)

2.3.2. Mean Absolute Error: To calculate the MAE, you take the difference between your model's predictions and the ground truth, apply the absolute value to that difference, and then average it out across the whole dataset. The behavior of the MAE was also observed. The MAE, like the MSE, will never be negative since in this case, we are always taking the absolute value of the errors.

The MAE equation is given as:

$$MAE = \frac{\sum_i^n |y_i - \hat{y}_i|}{n} \quad (19)$$

III. RESULT AND DISCUSSION

Table 1: Results for Balance Panel Data Estimators

		POOLING	WITHIN	RANDOM (Swamy-Arora's)	FD	BETWEEN
BALANCE PANEL: n = 5, T = 5, N = 25						
Coefficients	Intercept	17.5506		17.2781	-0.2253	20.2117
	X	3.5340	3.15704	3.1433	2.9369	2.9959
Std. Error	Intercept	4.9661		4.8353	0.2791	16.8946
	X	0.2494	0.25912	0.2426	0.1870	0.8491
	X	12.5500	12.184	12.9585	15.7043	3.5284
Pr(>|t|)	Intercept	**0.00177****		**0.0003525 *****	0.4301	0.3175
	X	**8.99e-12 ***	**2.003e-10 *****	**< 2.2e-16 *****	**5.968e-12 *****	**0.0387 ***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Table 2: Results for Unbalanced Panel Data Estimators

		POOLING	WITHIN	RANDOM (Swamy-Arora's)	FD	BETWEEN
UNBALANCED PANEL: n = 5, T = 3-5, N = 20						
Coefficients	Intercept	17.8785		19.4729	-0.7065	16.1866
	X	3.1258	3.0048	3.0440	2.9167	3.1910
Std. Error	Intercept	5.1787		4.4755	0.2750	25.8119
	X	0.2598	0.2333	0.2241	**0.1807**	1.2971
t value	Intercept	3.4520		4.3510	-0.1943	0.6271
	X	12.0310	12.8770	13.5860	16.1411	2.4600
Pr(>|t|)	Intercept	**0.00284** **		**1.355e-05** ***	0.4923	0.6434
	X	**4.84e-10** ***	**3.761e-09** ***	**< 2.2e-16** ***	**5.569e-10** ***	0.2458

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

The result of the five estimators considered for balance pane models under the unbalanced panel models are discussed here. The discussion of the result from the Monte-Carlo study; the results of the parameters of the five regression models analyzed, the performances of the estimators assessed via the mean square error and the mean absolute error. The estimates were ranked from 1st, 2nd, 3rd, 4th and 5th with 1st ranked attributed to the most efficient estimator that has the lowest value of the mean square error and the absolute mean square error. 2nd rank is assigned to the second to performed best and so on.

Table 1 and Table 2 show the result of the regression analysis for Pooling, Within, Random, First Difference, and Between Estimators for both Balance panel and Unbalanced panel data.

For the balance data set outputs, Random (Swamy-Arora's) estimate rank first in performance, followed by Within Estimator, followed by First Difference Estimator, followed by Pooling Estimator, and lastly, the Between Estimator rank last in terms of the P-value and their significance in respect to the independent variables.

Similarly, the estimators employed for a 5% missingness attributable for unbalanced data and, the outcomes follow the same trend as when it was for the balance panel data. Besides the ranking, it is interesting to note that the p-value estimate for the dependent variable for Swamy- Arora's model in both panel data set and unbalanced panel data are the same. Also, the Between model coefficient estimate is not significant.

Similar to the MSE outcomes, Table 4 shows the estimates of the Mean Absolute Error (MAE) for the Estimators considered for balance panel models and their corresponding estimate for unbalanced panel data models. The results also follow the same pattern in ranking as MSE. Between Estimator rank first while Within Estimator followed, Random (Swamy-Arora's) Estimator was next, followed by Pooling Estimator and lastly First Difference Estimator rank last. Here there was no negative estimate because MAE is taking the absolute values of the residual of the Estimators. Figure 1 and Figure 2 also reflect the ranking of the MAE results similar to the presentation in Table 4.

Table 3: Mean Square Error (MSE)

	POOLING (OLS)	WITHIN	RANDOM (Swamy-Arora's)	FD	BETWEEN
BALANCE PANEL: n = 5, T = 5, N = 25					
MSE	1.1287	0.8354	0.9687	1.3986	0.2904
UNBALANCED PANEL: n = 5, T = 3-5, N = 20					
MSE	0.8924	0.4754	0.5919	0.9803	0.2465

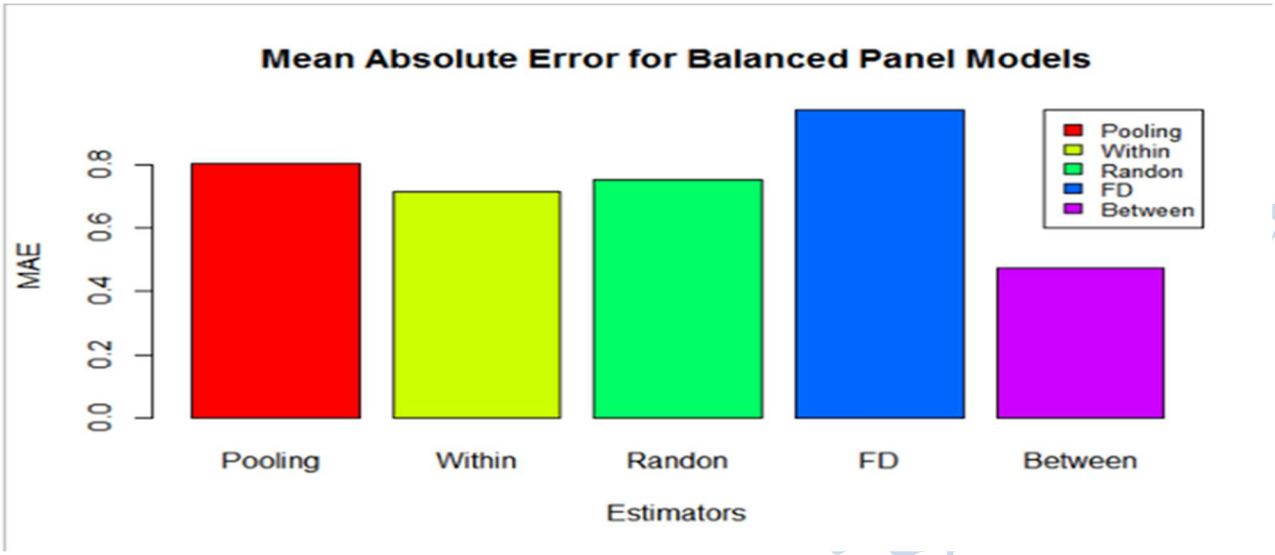

Figure 1: Plot of Mean Absolute Error of Balance Data

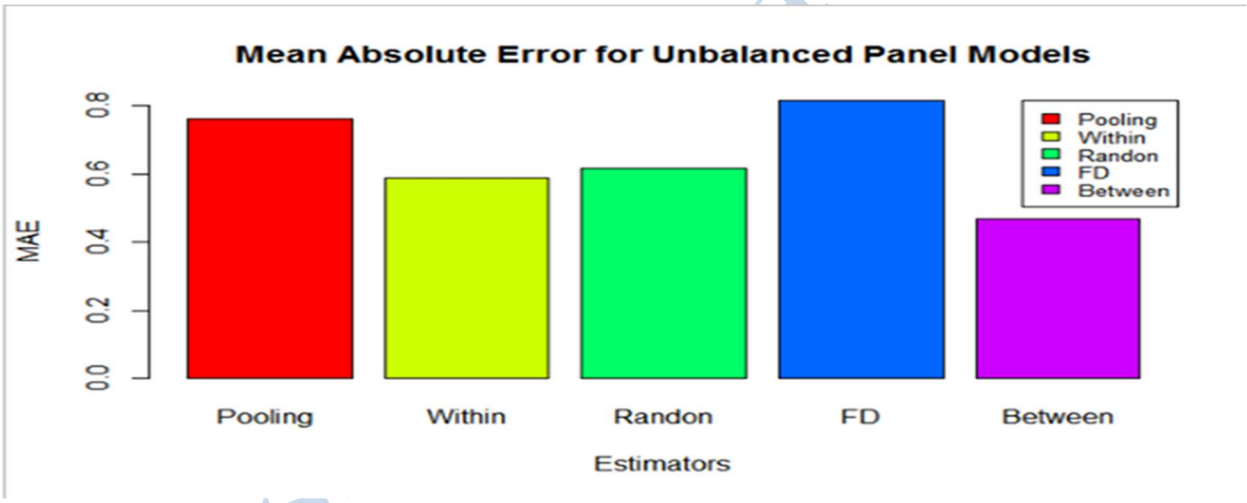

Figure 3: Plot of Mean Absolute Error of Unbalanced Data

Table 4: Mean Absolute Error (MAE)

MAE	POOLING (OLS)	WITHIN	RANDOM (Swamy-Arora's)	FD	BETWEEN
BALANCE PANEL: n = 5, T = 5, N = 25					
MAE	0.8023	0.7148	0.7541	0.9706	0.4724
UNBALANCED PANEL: n = 5, T = 3-5, N = 20					
MAE	0.7619	0.5891	0.6165	0.8140	0.4677

The MSE is always positive (and not zero) because of randomness. The MSE is the second moment (about the origin) of the error and thus incorporates both the variance of the estimator (how widely spread the estimates are) and its bias (how far off the average estimated value is from the true value), For an unbiased estimator, the MSE is the variance of the estimator. Like the variance, MSE has the same units of measurement as the square of the quantity being estimated. An analogy to standard deviation. If MSE is greater than zero which means that is $\lambda \neq 0$ therefore, the Random effect is FE, however, this is less efficient than Within and Between. Brugger (2021). This assumption makes Random estimator less efficient in this study.

Pooling (OLS) Estimator ranked 4th. OLS ignores time and individual characteristics and focuses only on dependencies between the individual. It is characterized by no correlation between the unobserved, independent variable(s) and the independent variables (i.e., exogeneity) for the same individual. This assumption on the error terms is very strong or unrealistic. This accounts for its high estimate compared to other Estimators as seen in Table 3 and Table 4.

First difference Estimator performs poorly among the five Estimators considered as it ranked 5th. Correlation between X_{it} and $w_{j.t-2}$. equation (8) an assumption on exogeneity makes it less demanding for FD than Within estimator and other Estimators. It is also less efficient than other Estimators because W_{it} is serially correlated and even if $W_{it}{}'s$ is uncorrelated. Therefore, FD does not violate this assumption in this project.

IV. CONCLUSION

Findings in this work show the ranking in terms of p-value and significance of the regression coefficients in respect to the independent variables for the five estimators considered for the balance panel data set; Random (Swarmy-Arora's) estimator rank 1st in performance, Within Estimator ranked 2nd, First Difference Estimator ranked 3rd, Pooling Estimator ranked 4th and lastly, the Between Estimator ranked 5th.

Similarly, the estimators employed for a 5% missingness attributable for unbalanced data and, the outcomes follow the same trend as when it was for the balance panel data.

Besides the ranking, it is interesting to note that the p-value estimate for the dependent variable for Swamy-Arora's model in both panel data set and unbalanced panel data are the same. Also, the Between model coefficient estimate is not significant.

Following the Monte-Carlo studies to investigate the performances of the five different estimators of balance panel data models under unbalanced panel data models. The results for the criteria for the performances 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 while Within Estimator ranked 2nd, Random (Swamy-Arora's) Estimator ranked 3rd, Pooling Estimator ranked 4th and lastly First Difference Estimator ranked 5th.

In general, the result shows that the Between estimator performed better than the other four estimators considered for panel data set under the unbalanced data set for small sample size(n). As signal in the study, it is recommended that the Between estimator should be adopted for fitting the panel data models when evidence of missingness is observable in the data, especially when the number of sample units is very small

REFERENCES

Amemiya, T. 1971: The estimation of the variances in a variance-components model, International Economic Review 12, 1-13.

Arellano, M. 2003: Panel Data Econometrics.

Baltagi, B. H. 2005: Econometric Analysis of Panel Data, John Wiley and Sons, England.

Baltagi, B.H. 2010: Panel Data Inference under Spatial Dependence. Syracuse University, Center for Policy Research Alain Pirotte Université Panthéon-Assas Paris II.

Baltagi, B. H. 2014: Panel Data and Difference-in-Differences Estimation Syracuse University, Syracuse, NY, USA. Elsevier Inc.

Brugger, B. 2021: A Guide to Panel Data Regression: Theoretics and Implementation with Python (towards Data Science blog).

Croissant, Y. and Millo, G. (2018): Panel Data Econometrics with R: The Error Component Model. pp23–51.

Cottrell, A. 2017: Random effects estimators for unbalanced panel data: a Monte Carlo analysis using gretL.

Garba M. K., Oyejola, B.A. and Yahya, W. B. 2013: Investigations of Certain Estimators for Modeling Panel Data Under Violations of Some Basic Assumptions. 3(10).

Graham, J. W. 2009: Missing Data Analysis: Making It Work in the Real World., 60(1), 549–576.

Hedeker, D. and Gibbons, R.D.: 2006: Longitudinal Data Analysis [Wiley Series in Probability and Statistics] *(Hedeker/Longitudinal)*.

Kang, W. 2015: "Missing-Data Imputation in Nonstationary. Panel Data Models" *In* Missing Data Methods: Time-Series Methods and Applications. pp 235-251

Lee, K. J., Tillingc, K. M., Cornishc, R. P., Littled, R. J. A., Belle, M. L., Goetghebeurf, E., Hogang, J. W. and Carpenterh, J. R. 2021: Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. Journal of Clinical Epidemiology 134 (2021). Pp79-88.

Matyas, L. and Sevestre, P. 1992: The Econometrics of Panel Data, 46-71. Kluwer Academic Publish

Mayer M. 2010: Unbalanced Panel Data Model. Department of Economics, University of Vienna, U.S.A.

Nijman, T. and Verbeek, M. (1992): Nonresponse in panel data: The impact on estimates of a life cycle consumption function.
Journal of Applied Econometrics, *Vol. 7, 243-257*

Rubin, l. H., Witkiewitz, K., St. Andre, J. and Reilly, S. 2007: Methods for Handling Missing Data in the ehavioral Neurosciences: Don't Throw the Baby Rat Out With The Bath Water

Rubin, D. B., Stroud, T. W. F. and Thayer, D. T. A. 1981: fitting additive models to unbalanced two-way data. *Journal of Educational and Behavioral Statistics.*

Schafer, J. L. 1997: The Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

Swamy. P. A. V. B. and 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.

Wallace, T.D. and Hussain, A. 1969. The use of error component models in combining cross section with time series data, Econometrica 37. 5-72.

Young, R. and Johnson, D.R. 2015: Handling Missing Values in Longitudinal Panel Data with Multiple Imputation. Journal of Marriage and Family. 77(1). pp 277-294.