Afrika Statistika
Vol. 4 (1), 2017, pages 241–258.
DOI: http://dx.doi.org/10.16929/ajas/241.213

Afrika Statistika
ISSN 2316-0861
ISSN 2316-090X

Comparison of Imputation Methods for Missing Values in Longitudinal Data Under Missing Completely at Random (MCAR) Mechanism

Anani Lotsi[1,*], **Louis Asiedu** [2] **and Johnson Katsekpor**[3]

[1,2,3]Box 115 LG Department of Statistics, School of Physical and Mathematical Sciences, College of Basic and Applied Sciences

Received August 01, 2017; Accepted November 01, 2017; Published online February 07, 2017

Abstract. This paper compared the performance of five (5) techniques of imputing missing values under the assumptions of MCAR mechanisms. The study compared the techniques for solving missing values using the Generalized Estimating Equation (GEE) model for the complete dataset, the coefficient of determination, mean squared error (MSE) and root mean squared error (RMSE). The pairwise deletion is the best under MCAR mechanism. Listwise deletion and the hot deck imputation methods performed poor under the MCAR mechanism.

Key words: Longitudinal data; Missing data; GEE; Coefficient of determination.
AMS 2010 Mathematics Subject Classification : Primary 62-07, Secondary, 62-J05

Presented by Professor Abdou K Diongue
University Gaston Berger, Saint-Louis (Sénégal)
Member of the Editors Board.

[*]alotsi@ug.edu.gh : Email Corresponding Author
lasiedu@ug.edu.gh
prophesy20012001@yahoo.co.uk

Résumé. Dans cet article, nous comparons les performances des cinq (5) techniques d'imputation de valeurs sous les hypothèses de données manquantes de manière complètement aléatoirement (MCAR). La comparaisonse fait la base du modèle des Equations Généralisées d'Estimation (GEE) pour la base complète, le ceofficient de détermination, de l'erreur quadratique moyenne (MSE) et du coefficient RMSE. Notre étude conclut que la méthode d'élimination appairée est la meilleure sous l'hypothèse (MCAR). Les performance des méthodes Liswise et hot deck se révèlent relativement faibles.

1. INTRODUCTION

According to Hedeker and Gibbons (2006), longitudinal study is an observational research technique in which data is collected for the same subjects repeatedly over a time period. Longitudinal studies are now regularly used in biology, psychology, social, public health and clinical research (Singer and Willett (2002)). There are two main types of longitudinal research designs, these include: prospective and retrospective longitudinal study design. The prospective longitudinal research design is used to collect data on subjects going forward in time. In prospective study, subjects are follow from enrollment to end of the study. However, subjects are sampled with and without risk factors, they are monitored over period of time to repeatedly measure a defined outcome variable. The retrospective longitudinal study is used to collect data on subjects going backwards in time where the outcome variable for both cases (those already known to have disease based on their outcome) and controls (those already known to not have the disease) is repeatedly collected backwards in time. Retrospective studies collect data at various time point in the past. For example, a researcher may look for a trend when he finds the medical records of previous years.

Longitudinal research has many benefits over cross-sectional studies (See Hedeker and Gibbons (2006)). First, in order to achieve the same statistical power, smaller subjects are needed in longitudinal studies. This is because more information is deliver for repeated measurement from a single subject than a single measurement of a single subject. Second, in a longitudinal research, each subject can assist as his or her own control. Generally, intra-subject variability is much less than inter-subject variability. Third, investigators are able to separate timing effects from cohort effects. Finally, longitudinal studies can give information on individual change, which could not be provided by cross-sectional studies. However, longitudinal studies are also having their own challenges. There are several reasons that is both practical and theoretical, which make the longitudinal analysis very difficult. Such reasons include, but are not limited to, between repeated outcome measurements, missing data, irregularly timed data, mixture of static and time varying covariates, and availability of software for model fitting. This paper focuses on missing data in longitudinal study. Practically all methods of statistical analysis are affected by problems with missing values. It is well known that the use of wrong methods for handling missing data can lead to bias in parameter estimates (Jones (1996)), bias in standard errors and test statistics (Glasser (1964)), and unproductive use of the data (Afifi and Elashoff (1996)). There are many reasons why data may be missing from a complete dataset, for instance, unable to find certain characteristics (Hulse Van and Khoshgoftaar (2008)). The most common and

simple technique of handling missingness in a datasets is to overlook either the projects or the features with missing observations. When this techniques not used, contributes to loss of important information and result to imprecise cost estimation models. According to Cohen *et al.* (2003), when researchers use conventionally proper approaches for dealing with missingness in the datasets, different methodologies may result to different conclusions. Gad (2006) claimed that disregarding the missing values in this case leads to biased conclusions. Furthermore, when an attribute has a missing value in a test case, it may or may not be meaningful to take the extra effort in order to achieve a value for that attribute(s). There are many approaches of solving the problem of missing data.

Schmitt *et al.* (2015) compared six (6) different imputation methods: Mean, K-nearest neighbors(KNN), fuzzy K-means (FKM), singular value decomposition (SVD), Bayesian principal component analysis (bPCA) and multiple imputations by chained equations (MICE). Comparison was performed on four real datasets of various sizes, under a missing completely at random (MCAR) assumption, and based on four evaluation criteria: Root mean squared error (RMSE), unsupervised classification error (UCE), supervised classification error (SCE) and execution time. Their results suggest that bPCA and FKM are two imputation methods of interest which deserve further consideration in practice.

Niass *et al.* (2015) also compared six methods to handle missing values in longitudinal data under MCAR mechanism. These include: Complete-case (CC) analysis so-called listwise deletion, mean substitution, k-nearest neighbors (knn), multiple imputation using the expectation- maximization (EM), predictive mean matching (pmm) and regression. Incomplete dataset with percentage of missing values varying between 5% to 50% were created from complete dataset. They used the Root mean square error (RMSE), Mean absolute error (MAE), p.value, multiple R-square, AIC and BIC criteria to compare the aforementioned imputation approaches. The results demonstrate that multiple imputation using the predictive mean matching (MI.pmm) and the k-nearest neighbor (knn) methods were the best when the missing data percentage was larger than 5 percent. The listwise deletion approach produces the most inaccurate result.

Garcia-Laencina *et al.* (2009) claimed Missing data is a common draw back in many real-life pattern classification scenarios and the most popular solutions is missing data imputation by the K nearest neighbours (KNN) algorithm. In their article, they proposed a novel KNN imputation procedure using a feature-weighted distance metric based on mutual information (MI). This method provides a missing data estimation aimed at solving the classification task, that is, it provides an imputed dataset which is directed toward improving the classification performance.

Little (1992) proposed that missing data have three significant implications or longitudinal data analysis. First, when data are missing in longitudinal study, the data set is certainly not balanced over the time period since not all individuals have similar number of repeated measurements at a common set of occasions. This imbalance data let the methods of analysis change from the one of balanced data. Secondly, there must be some loss of information and also reduction in the sample size when there are missing data. The missing values spread

sporadically over several subjects and how highly correlated the missing data are with the observed data will affect loss of precision. Finally, under certain circumstances, missing data can contributes to bias and thereby lead to misleading inferences about changes in the mean response. The higher attrition is likely to have bias and the potential for serious bias makes the longitudinal analysis more complicated. Selecting the most suitable technique to solve the problem of missing data during analyses is one of the most difficult decisions researchers go through. Most often, missing values are ignored rather to use suitable imputation technique to replace them. In view of the above mentioned problems, the following data imputation methods such as Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, and Last observation carried forward (LOCF) are based on the assumption that data are MCAR will be compared to know the best imputation method to solve the problem of missing data. It is against this background that this study is being undertaken to research and compare the best imputation technique for missingness in longitudinal data.

1.1. Missing Data Mechanisms

The missing data mechanism defines the association between the missing values of the data and the values of the variables in the data matrix, i.e. whether the missing values depend on the underlying values of the variables in the data set. Gelman and Hill (2007) posit several reasons why data may be missing. There are various assumptions concerning missing data mechanisms:

1.1.1. Missing completely at random (MCAR)

The probability of dropout is independent of the observed data and the missing data. That is $f(R_i/Y_i, X_i)$=$f(R_i)$. A typical example is that a subject moved to a different location where the treatment cant be continued. MCAR happens when any data of a variable have the same likelihood of being missing. It also when the data values in the dataset will be randomly missing and there will be no reason why a specific value is missing.

1.1.2. Missing at random (MAR)

The likelihood of dropout is only dependent on the observed data but not dependent on missing data. That is $f(R_i/Y_i, X_i)$=$f(R_i/Y_{i(obs)}, X_i)$. Where the observed dependent response vector is $Y_{i(obs)}$ and the observed covariate vector is X_i. Example: The MAR assumption would be satisfied if the probability of missing data on income depended on a persons age, but within age group the probability of missing income was unrelated to income.

1.1.3. Missing not at random (MNAR)

The probability of dropouts is dependent on the unobserved data and also the observed data or missing values do depend on unobserved values. That is $f(R_i/Y_i, X_i)$=$f(R_i/Y_{i(mis)}, X_{i(obs)})$. Under MNAR, the dropout procedure is also dependent on the missing values given observed measures. Example: to achieve the assumption of MNAR individual in high income class are less likely to report their income.

1.2. Patterns of Missingness

The missing values pattern defines which values in the data matrix that are actually missing, and can help in the choice of method for handling the missing values. Missing data patterns are usually divided into monotone missing patterns (MMP) and arbitrary missing patterns (AMP). Data are missing monotone often happens due to attrition in longitudinal studies, where dropping out by definition means that all the following observations will be missing. A special case of MMP is the univariate missing data pattern (UMP) where only one variable in the data set suffer from missing observations, (see Table 1). An AMP on the other hand arises when the data matrix cannot be ordered as in MMP, (see Table 1). One example of AMP is item non response in surveys where respondents for some reason have failed to answer one or more questions, but missing values in one variable does not necessarily implies that all following variables are missing. (Little and Rubin (2002)).

Table 1. Patterns of Missingness

Missing Monotone				Missing Arbitrarily			
K1	K2	K3	K4	K1	K2	K3	K4
√	√	√	√	√	√	NA	√
√	√	√	√	NA	√	√	NA
√	√	√	√	√	NA	√	NA
√	√	√	√	√	√	NA	NA
√	√	√	NA	√	√	NA	√
√	√	NA	NA	√	√	√	NA
√	NA	NA	NA	NA	√	√	√

NA	Missing Values
√	Not missing

Assumptions and patterns of missing values helped to decide the methods that can be used to deal with missing data.

2. Techniques for Handling Missing Data under MCAR

There are so many methods in handling missing values under MCAR. Many methods have been suggested and developed to handle missing values in longitudinal clinical trials under the MCAR assumptions. However, there are few methods that are actually used in real trials with missing values. This paper will find the best method for handling missing data when considering listwise deletion, pairwise deletion, mean substitution, hotdeking, and LOCF methods. The Methods used in this work to deal with missing data under MCAR are described as follows:

2.1. Listwise Deletion

The most common and easiest technique of dealing with missing values is listwise deletion, Schafer and Graham (2002). When listwise deletion is used, the computer program auto-

matically deletes any item that has missing data for any bivariate or multivariate analysis. Even though each variable may be missing only a small percentage of responses, collectively a large portion of the data may not be used as cases are deleted.

2.2. Pairwise Deletion

Pairwise deletion, also called available case analysis is a common alternative to listwise deletion in linear models. Pairwise deletion, uses all existing data to obtain parameters of the model. When a researcher looks at univariate descriptive statistics of a data set with missing observations, he or she is using available case analysis, inspecting the means and variances of the variables observed throughout the data set. Pairwise deletion is a technique that focuses on the variance-covariance matrix. Each element of that matrix is estimated from all data available for that element.

2.3. Mean Substitution

The technique of mean imputation imputes the missing values using the mean of the available observed values. This method has the potential of giving biases as well as underestimating variability (Carpenter *et al.* (2004)).

2.4. Hotdecking

Hotdecking, detects a person in the data set with complete data who is similar on an identified correlated characteristic to a person with incomplete data and uses that person's score to substitute the missing value. This method works well when the variable used to sort the data is highly predictive of the variable with the missing data and when there is a large sample so that a similar case is easily recognized (**?**).

2.5. Last Observation Carried Forward (LOCF)

The simplest imputation method is the LOCF method that substitutes every missing value with its corresponding last observed value. The LOCF technique is often used in longitudinal studies of continuous outcomes under MCAR mechanism. This method assumes that the result would not change after the last observed value. Thus, there is no time effect since the last observed value.

3. Measures of Performance for Imputation Methods

The Mean squared error (MSE), the root mean squared error (RMSE) and the coefficient of determination will be used as criteria to assess the performance of the best imputation methods.

3.1. Mean squared error (MSE)

Mean square error (MSE) is the mean of the squared differences. It is the average squared difference between the estimated parameters ($\hat{\beta}_0$ and $\hat{\beta}_1$) and the corresponding true pa-

rameters (β_0 and β_1) derived from the original data set. It shows how the estimator is close to the true value. MSE is also equal to the sum of variance and the squared bias of the estimated parameters.

3.2. The Root Mean Squared Error (RMSE)

The Root Mean squared error (RMSE) is defined as the square root of the MSE. The RMSE is a valuable measure of total precision or accuracy and can help to know how each imputation technique is performing. In general, the more efficient method would have a lower RMSE (Huang andCarriere (2006)).

3.3. Coefficient of Determination (R^2)

The coefficient of determination is used to measures how well a model explains and predicts future outcomes. The coefficient of determination in statistical analysis, also known as R-squared, is used as a guideline to assess the accuracy of the model. It is the degree of variability in factor or variable that is explained by another variable done with the variable.

4. Methodology

4.1. Data Description

The National Income Dynamics Study (NIDS) Data from South Africa will be used to assess the aforementioned imputation methods. In South Africa, the first national panel research conducted was the NIDS. The Southern Africa Labour and Development Research Unit (SALDRU) in the School of Economics at the University of Cape Town is responsible in executing this survey. The research took a national sample of 28,000 respondents from closely 7,300 households across the entire country when it started in 2008. In every two years, the survey is repeated with these same household members and it observes the livelihoods of individuals and households over time period. The NIDS shows South Africa dynamic household structure, changes in people living situations and the well-being of members in the household in a way that no other study in South Africa has been able to do.

The main characteristic of the research is its ability to follow respondents as they relocate to different households. NIDS is a programme to compile comprehensive longitudinal information on respondents selected for the study and to find out who is moving ahead and who is falling behind. This data is also key for research and policy makers. The NIDS data constitute areas such as health, education, labour market and birth history. The 2008 data was compiled into the Wave1 dataset. The second, Wave 2 dataset was compiled after the second visit was made to the same group of people between 2010 -2011. This paper focused on the work status of persons selected for the research. The binary response variable measured therefore was whether an individual was employed or not at the time of visit. Specifically, (Employed=0, Not employed=1)

Covariates chosen for this exercise include:

- Gender of respondent (Male= 0, Female = 1)

– Education (Educated= 0, Not educated = 1)
– Age (18-30 years = 0, 31-57 years = 1) and
– Marital Status (Married=0, Not married=1)

The covariates chosen were tested to find out whether there is significant impact for conditional, marginal models and joint models.

4.2. Models Used For Analysis

The ordinary least squares (OLS) is a technique for estimating the unknown parameters in a linear regression model, with the goal of minimizing the sum of the squares of the differences between the observed responses in the given dataset and those predicted by a linear function of a set of explanatory variables. The individual national income from South Africa (i.e. wave 1 and wave 2) was correlated, hence the assumption under ordinary least squares (OLS) regression was violated. As such these correlations required to be taken into account in modeling; otherwise the standard errors of the estimates would be underestimated for the between-subject and overestimated for the within-in-subject effects. Generalized estimating equations (GEE) were introduced by Liang and Zeger (1986) as an extension of generalized linear models (GLM) to analyze discrete and correlated data.

4.2.1. Generalized Estimating Equations (GEE) Models

The GEE is a semi-parametric regression approach which uses moment-based inference, it was first introduced by Liang and Zeger (1986). It is an extension of generalized linear models that account for correlated responses. Instead of attempting to specify a model for the whole multivariate distribution of a data vector, GEE only models the first moment, specifically the mean response $E(Y_{it})$ at each visit t for the i^{th} subject. The Generalized Estimating Equation for estimating β is an extension of the independence estimating equation to correlated data and is given by:

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta} V_i^{-1}(Y_i - \mu_i(\beta)) = 0. \tag{1}$$

The GEE specifications entail those of GLM with one addition. So, first, the linear predictor is given as:

$$\eta_{ij} = x_{ij}^{'}\beta, \tag{2}$$

where x_{ij} is the covariate vector for subject name i at time j. we then consider the link function as:

$$g(\mu_{ij}) = \eta_{ij}. \tag{3}$$

Mean response:

$$E(y_{ij}) = \mu_{ij}. \tag{4}$$

As in GLMs, the general choices here are the identity, logit, and log link for continuous, binary, and count data, respectively. The variance is then described as a function of the mean, namely,

$$V(\mu_{ij}) = \varphi v(\mu_{ij}). \tag{5}$$

Where $v(\mu_{ij})$ is a known variance function and φ is a scale parameter that may be known or estimated.

4.2.2. The GEE Estimation (Working Correlations)

If A_i to be the $n_i \times n_i$ diagonal matrix with $V(\mu_{ij})$ as the j^{th} diagonal element, as specified above, we define $n_i \times n_i$ working correlation matrix (of the n repeated measures) for the i^{th} subject (i.e. Y_i) as $R(\alpha)$. Hence, the working variance-covariance matrix for Y_i will be:

$$V(\alpha) = \varphi A_i^{1/2} R_i(\alpha) A_i^{1/2}. \tag{6}$$

For the case of outcomes that are normally distributed with homogeneous variance across time, is given as:

$$V(\alpha) = \varphi R_i(\alpha). \tag{7}$$

In the case of normal outcomes, Park and Davis improves this to heterogeneous variance across time by making the scale parameter φ_j to change across time period $(j = 1, ..., n)$. The GEE estimator of β is the solution of :

$$\sum_{i=1}^{N} D_i^{'}[V(\hat{\alpha})]^{-1}(y_i - \mu_i) = 0, \tag{8}$$

where $\hat{\alpha}$ is a consistent estimate of α and $D_i = (\frac{\partial \mu_i}{\partial \beta})$ and therefore equation (8) becomes:

$$\sum_{i=1}^{N} \left(\frac{\partial \mu_i}{\partial \beta}\right) (V(\hat{\alpha}))^{-1}[y_i - \mu_i] = 0. \tag{9}$$

This is an improvement on estimating equation for β in any GLM, which is given in (9). Therefore, the GEE solution can be seen as a natural generalization of the GLM solution for correlated data. As an example, in the normal case, for equation (9), that is:

$$\begin{gathered} U(\beta) = \textstyle\sum_{i=1}^{N} (\frac{\partial \mu_i}{\partial \beta})^{'} (V(y_i))^{-1}[y_i - \mu_i] = 0, \\ \mu_i = X_i \beta, \\ D_i = X_i, \\ V(\alpha) = R_i(\hat{\alpha}). \end{gathered} \tag{10}$$

The solution for the parameter β (by making β a subject) gives;

$$\beta = [\sum_{i=1}^{N} X_i^{'}[R_i(\hat{\alpha})]^{-1} X_i]^{-1} \sum_{i=1}^{N} X_i^{'}[R_i(\hat{\alpha})]^{-1} y_i]^{-1}]. \tag{11}$$

Equation (11) depends on the mean and variance of y and is called quasi-likelihood estimates. Working the GEE includes iterating between the quasi-likelihood solution for estimating β and a robust technique of finding α as a function of β.

4.3. Missing Data Mechanism Test

To properly analyze missing data sets requires the knowledge of how the data is missing (i.e in a random way or non-random way). This will help us to classify missing data under the assumptions of various missing mechanism. In this study, percentage of missing values such as 5%, 10%, 15%, 20% and 30% were artificially created in a random and non-random way using complete life data (NIDS data) from south Africa. In order to use the required imputation method to handle each percentage of missingness, little's test of MCAR was employed.

4.4. Testing the Missing Data Mechanism

In this paper, we will adopt the littles test MCAR to check whether a dataset with missing values is MCAR or MAR. Little's test of MCAR provides tests for the MCAR and MAR assumption. If we failed to reject the null hypothesis under the little test of MCAR, then we can conclude that imputation methods such as Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, and Last observation carried forward (LOCF) depend on the assumption that the pattern of missing values does not depend on the data values. (This condition is known as missing completely at random, or MCAR.). Violation of the MCAR assumption can lead to biased estimates produced by the methods of handling missing data. When the p value is less than the alpha value, we reject the null hypothesis under the little test of MCAR and say that imputation methods such as Multiple imputation and Expectation maximization depend on the assumption that the pattern of missing data is associated to the observed data only (This condition is called missing at random, or MAR).

5. Results of Data Analysis

5.1. Missing Data Mechanism Test

To properly analyze missing data sets requires the knowledge of how the data is missing (i.e in a random way or non-random way). This will help us to classify missing data under the assumptions of various missing mechanism. In this study, percentage of missing values such as 5%, 10%, 15%, 20% and 30% were artificially created in a random and non-random way using complete life data (NIDS data) from south Africa. In order to use the required imputation method to handle each percentage of missingness, little's test of MCAR was employed. Table 2 shows the output of littles MCAR test on the percentages of missing values artificially created. Table 2 shows that the significant values

Table 2. Output of Little's MCAR test: under MCAR

Percentages of missing values created					
MCAR test	5%	10%	15%	20%	30%
Chi-sq	53.16	51.31	53.43	62.03	54.32
D.f	48	58	61	65	69
Sig	0.282	0.721	0.744	0.582	0.902

for all the percentages of missingness are greater than the alpha values of 0.05, hence

we fail to reject the null hypothesis. We conclude that the data values in the dataset was randomly missing and there is no specific reason for missingness. This condition is known as missing completely at random (MCAR). This means that the various imputation methods for handling these missingness depend on the assumption that the pattern of missing values is independent of the data. In addition, all methods for estimation give consistent and unbiased estimates of the correlations and covariances when the data are MCAR.

After the little MCAR test, Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking and Last observation carried forward (LOCF) will be used to replace missing values created in the complete dataset under MCAR mechanism.

5.2. Marginal model-GEE

Using the marginal model, from table 3 below, four covariates such as Gender, Marital status, Age and Educational status were significant, meaning that they contribute significantly to the state of employment status. The fitted marginal model is

$$logit\,(\hat{p}_1) = log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) = -0.041 + 0.124X_1 - 0.146X_2 + 0.081X_3 + 0.602X_4, \qquad (12)$$

where X_1= Gender, X_2= Marital status, X_3 = Age, X_4 = Educational status, $\hat{p}_1$ represents the estimated probability of recording an "employed" response at the second measurement and $1-\hat{p}_1$ represents the estimated probability of not recording an "employed" response at the second measurement. The estimated intercept is -0.041 representing the estimated logit when a respondent's gender is male, married, belongs the age bracket $18-30$ and educated. This means that the respondent had no age group, no gender, no marital status and no educational status which is impossible in this particular study.

Table 3. Fitted models using data from NIDS: GEE Model

Estimate	β values	Odds Ratio	Std.err	Wald	Pr($>\|W\|$)	
(Intercept)	-0.04090	0.95993	0.10032	0.166	0.68353	
Gen	0.12411	1.13214	0.05402	5.278	0.0216	*
Mari	-0.14627	0.86392	0.05056	8.369	0.00382	**
Age	0.08058	1.08392	0.05123	2.474	0.03574	*
Edu	0.60229	1.82630	0.09171	43.132	5.12E-11	***

Note: $R^2 = 0.00835$ * Significant at 5%, ** Significant at 1%, *** Significant at 0.1%

5.3. Comparison of imputation methods for handling missing values under GEE model

In order to compare various imputation methods and know the best, we compare each imputation method used to handle percentage of missingness to the general GEE model for the complete dataset, which is;

$$logit\,(\hat{p}_1) = log\left(\frac{\hat{p}_1}{1-\hat{p}_1}\right) = -0.041 + 0.124X_1 - 0.146X_2 + 0.081X_3 + 0.602X_4. \qquad (13)$$

The procedures for the comparison are stated below:

- We compare the model for each methods used to handle the percentage of missing values to the general GEE model. This is done by finding the coefficient difference for each imputation methods used to handle percentage of missing value.
- Coefficient difference is calculated by subtracting each coefficient from the coefficients of the general model.
- We compute the average of the coefficients difference for each imputation method.
- We sum the average coefficients difference for all the percentage of missingness for each imputation methods.
- To compare, we select the best imputation methods by picking the smallest average number.

5.4. Comparison of methods for handling missing values under MCAR mechanism

Missing data can frequently occur in a longitudinal data analysis. In a real-world data analysis, the missing data can be MCAR, MAR, or MNAR depending on the reasons that lead to data missing. In this paper, method for handling missing data such as Listwise deletion, Pairwise deletion, Mean substitution, Hotdecking, and Last observation carried forward (LOCF) under MCAR mechanism were compared. To evaluate the performance of these five imputation methods, we first use the total average coefficient difference for each imputation method and adjudge the smallest values as the best method. Table 4 shows how each methods for handling missing values under MCAR mechanism performed.

Table 4. Performance of methods for handling missing values under MCAR mechanism

Percentages	**Listwise**	**Pairwise**	**MeanSubs**	**HotDeck**	**LOCF**
	Averages	Averages	Averages	Averages	Averages
5	0.01154	0.009105	0.00726	0.01526	0.01182
10	0.03046	0.009101	0.009105	0.01741	0.01036
15	0.0298	0.008105	0.01001	0.02009	0.01461
20	0.11236	0.007705	0.01316	0.02497	0.01697
30	0.238305	0.23831	0.238192	0.23692	0.211332
Overall	**0.422465**	**0.272328**	**0.277727**	**0.31465**	**0.265092**

From Table 4, the best imputation method under MCAR is Last observation carried forward (LOCF), which recorded the minimum average coefficient difference of 0.265092. Among the five imputation methods compared under MCAR mechanism using the average coefficient, listwise deletion is the poorest method. Mean substitution and pairwise deletion performed well when small percentage of missing values occurred in a dataset. In concluding, when small percentages such as 5% and 10% of missingness occurred in the dataset under MCAR mechanism, it is advisable to use the mean substitution or pairwise imputation methods to replace missing values in the dataset. Again, when large percentage of missingness occurred in a dataset under MCAR mechanism, the Last observation carried forward (LOCF) method gave comparatively consistent estimates, hence LOCF imputation method is preferred when the percentage of missingness is large under MCAR mechanism.

5.5. Comparison of Imputation Methods Using the Coefficient of Determination (R^2)

From Table 3, the R^2 for the complete dataset is 0.00835, meaning 0.84% of the total variation in employment status was explained by the regression model. This study seeks to compare imputation methods, hence individual percentage of missingness of the various imputation methods may be doing well if their coefficient of determination values are closer to the R^2 value of the complete dataset which is 0.84%. To also identify the best imputation methods, average coefficients of determination of the various imputation methods will be compared and the best selected. The higher the average coefficient of determination, the better the method for handling missingness in the data.

Table 5 shows how each imputation methods for handling missing values under MCAR mechanism performed using the coefficient of determination (R^2).

Table 5. Performance of imputation methods using the R^2 under MCAR

Percentages	Listwise	Pairwise	MeanSubs	HotDeck	LOCF
5	0.007943	0.98354	0.008136	0.008092	0.007684
10	0.006582	0.95453	0.007798	0.007766	0.008189
15	0.006534	0.87438	0.007681	0.007791	0.007586
20	0.006732	0.94231	0.007713	0.007373	0.008218
30	0.500048	0.87398	0.534737	0.531458	0.581234
Overall	**0.527839**	**4.62874**	**0.566065**	**0.56248**	**0.612911**
Averages	**0.105568**	**0.925748**	**0.113213**	**0.112496**	**0.122582**

From Table 5, when small percentage of values (5% or 10%) are missing complete at random (MCAR) from a dataset, the mean imputation or the LOCF methods will be preferred. This is because under 5% missing values using the mean imputation to replace missingness in the data, the total variation in employment status that was explained by the regression model is 0.81% which is closer to the coefficient of determination for the complete dataset value of 0.84%. In other hand, when 10% of missing values were replaced by the LOCF imputation method, the coefficient of determination recorded was 0.82% also closer to the complete dataset value of 0.84%. Replacing large sets of missing values in a dataset, it is important to use the pairwise imputation method to replace missing values in the dataset under MCAR mechanism. From the above table, when 30% values were missing, the pairwise imputation method recorded R^2 of 87%. Thus 87% of the total variation of employment status was explained by the regression model. In all, pairwise method performed well in replacing missing values under MCAR mechanism. To achieve proper statistical inference, it is advisable to use either the mean imputation or the LOCF to replace missing values when the lost in the dataset is small (5% or 10%) and the pairwise imputation when the missing values in the dataset is large. Listwise and hot deck imputation methods performed poor under MCAR mechanism. For this reason, they may lead to bias in parameter estimates and improper statistical interpretation of the analysis.

Figure 1 is a pictorial representation of the Performance of imputation methods using the R^2 under MCAR.

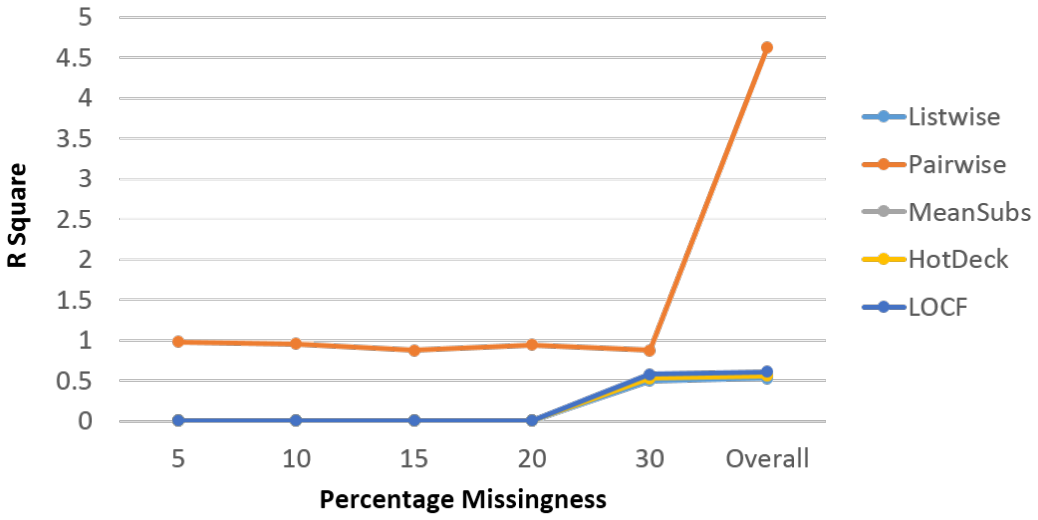

Fig 1: Performance of imputation methods using the R^2 under MCAR

5.6. Comparison of Imputation Methods Using the Mean Square Error (MSE)

Table 6 below shows how each imputation methods for handling missing values under MCAR mechanism performed using the MSE.

Table 6. Performance of imputation methods using the MSE under MCAR

Percentages	Listwise	Pairwise	MeanSubs	HotDeck	LOCF
5	0.2296888	0.0105362	0.2272594	0.2295869	0.2295212
10	0.2304968	0.01066695	0.2261406	0.2299805	0.230001
15	0.231021	0.01058949	0.2238092	0.2300539	0.2300619
20	0.2307411	0.01054023	0.2216802	0.2305831	0.230191
30	2.02E-28	2.22E-16	1.29E-26	1.29E-26	1.29E-26
Overall	**0.9219477**	**0.04233287**	**0.8988894**	**0.9202044**	**0.9197751**

From Table 6, the pairwise deletion yields smaller mean square error (MSE). The pairwise deletion performed well in replacing missing values for both small and large percentage of values lost in the dataset. The mean substitution and last observation carried forward

(LOCF) also did creditably well after assessed with the mean square error (MSE). This means that, when values are missing complete at random in a dataset, it is paramount to use pairwise deletion to replace the missing values in order to achieve proper statistical inference. The mean substitution and last observation carried forward (LOCF) can also be used to replace missingness in the dataset. Moreover, listwise deletion and the hot deck imputation methods performed poor under the MCAR mechanism, hence it should not be encouraged in replacing missing values in a dataset. This may lead to bias in parameter estimates of the analysis.

5.7. Comparison of Imputation Methods Using the Root Mean Square Error (RMSE)

Table 7 below shows how each imputation methods for handling missing values under MCAR mechanism performed using the RMSE.

Table 7. Performance of imputation methods using the RMSE under MCAR

Percentages	Listwise	Pairwise	MeanSubs	HotDeck	LOCF
5	0.4792586	0.102646	0.4767173	0.4791523	0.4790837
10	0.4801008	0.1032809	0.4755424	0.4795628	0.4795842
15	0.4806464	0.1029053	0.4730848	0.4796394	0.4796476
20	0.4806464	0.1026656	0.4708293	0.4801907	0.4797823
30	1.42E-14	1.49E-08	1.14E-13	1.14E-13	1.14E-13
Overall	**1.9206522**	**0.411497815**	**1.8961738**	**1.9185452**	**1.9180978**

Table 7 indicates the performance of imputation methods using the RMSE. Listwise and hotdeck imputation have the RMSE value of 1.9206522 and 1.9185452 respectively, which is the worst performance compared to the various imputation methods under MCAR mechanism. Pairwise deletion performed well in all the percentages of missingness artificially created. This means that, pairwise deletion will do well in both small and large amount of missing values in a dataset under MCAR mechanism. The table also reveals that, the mean substitution and the last observation carried forward (LOCF) did well under the MCAR mechanism. It is important to use either pairwise deletion, the mean substitution or the last observation carried forward (LOCF) imputation methods to replace missing values in a dataset under MCAR mechanism. There is no extreme different using either the MSE or the RMSE to assess the performance of various imputation methods under MCAR mechanism.

6. Conclusion

Comparing imputation methods under MCAR mechanism, the analysis shows that pairwise deletion is the best. The mean substitution and last observation carried forward (LOCF) also did creditably well. This shows that, when large percentage of missing values occurred in a dataset under MCAR mechanism, the pairwise deletion, last observation carried forward (LOCF) or mean substitution method will give consistent and unbiased estimates, hence it will be significant to go for either one of these imputation methods when the percentage of missingness is large or small under MCAR mechanism. Listwise deletion and the hot deck imputation methods performed poor under the MCAR mechanism, for this reason, it

should not be encouraged in replacing missing values in a dataset. This may lead to bias in parameter estimates of the analysis.

6.1. Recommendations

The following recommendations are made both in the area of policy formulation and future studies based on the findings and conclusions made from the study.

(1) The study recommends that when data are missing complete at random (MCAR), the pairwise deletion, or mean substitution or last observation carried forward (LOCF) is recommended to replace either small or large amount of missing values in the dataset. This will help to achieve proper statistical inference in data analysis.

(2) Researchers must determine whether the cause and pattern of the missing data will seriously weaken the quality of the inferences derived and which procedure is most suitable for handling missing data. Examining of factors causing missing data and the missing data pattern carefully, allows researchers to decide if and how to best deal with missing values in a study.

(3) All research report must report the reasons for and the amount of missing data as well as what data imputation method was used during the analysis.

(4) The study also recommended that future investigations find a better approach for imputing missing not at random (MNAR) with multiple imputation. This is vital due to the hope many researchers have in this method because of the advantages that multiple imputations have among the other imputation methods compared in this research.

(5) This paper focused on missing values in a longitudinal dataset. However, an extension of this strategy in the case of categorical data deserves further research.

Acknowledgment

The authors of the paper wish to express their sincere gratitude to the presenter of the paper for his immense and valuable suggestions and constructive comments. We owe a debt of gratitude at his time for his assistance in bringing this paper to publication. We recognize the fact his comments and suggestions were particularly appreciated and very useful to us. He has made very careful reading of the manuscript and pointed out many ways in which it could be improved.

References

Afifi, A. A., Elashoff, R. M. (1966). Missing observations in multivariate statistics: Review of the literature. *Journal of the American Statistical Association*, 61, 595-604.

Carpenter, J., Kenward, M.G., Evans, S. White, I. (2004) Last Observation Carry-Forward and Last Observation Analysis. *Statistics in Medicine*, 23, 3241-3242.

Cohen, J., Cohen, P., West, S. and Aiken, L. (2003). *Applied multiple regression / correlation analysis for the behavioral sciences* (3rd Ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Gad, A. M., Ahmed, A. S. (2006). Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm, *Computational Statistics and Data Analysis*, 50(10), 2702-2714.

Garcia-Laencina, P. J., Sanch-Gomez, J. L., Figueiras-Vidal, A.R., Verleysen, M. (2009), K neares neighbours with mutual information for simulations classification and missing data imputation. *Neurocomputing*, vol. 72, pp. 1483-1493.

Gelman, A., Hill, J. (2007). *Data analysis using regression and multilevel / hierarchical models*. Columbia University, NY: Cambridge University Press.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, 59, 834-844

Hedeker, D., Gibbons, R.D. (2006). *Longitudinal Data Analysis*, Wiley Publications

Huang, R., Carriere, K.C. (2006). Comparison of Methods for Incomplete Repeated Measures Data Analysis in Small Samples. *Journal of Statistical Planning and Inference*, 136, 235-247.

Hulse Van, J., Khoshgoftaar, T.M. (2008). A comprehensive empirical evaluation of missing value imputation in noisy software measurement data, *The Journal of Systems & Software*, 81(5), 691-708.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association*, 91,222-230.

Liang, K., Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13-22.

Little, R. J., & Rubin, D. B. (2002). *Analysis with Missing Data*. Hoboken, New Jersey: Wiley.

Little, R.J.A. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association*, 87, 1227 - 1237.

Niass, O., Diongue, A.K., Touré, A. (2015), *Analysis of missing data in sero-epidemiological studies*, African Journal of Applied Statistics, Vol. 2, pp. 29-37.

Park, T., Davis, C. (1993). A test of the missing data mechanism for repeated categorical data. Biometrics, 49 (2), 631- 638.

Schafer, J. L., Graham, J.W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147-177.

Schmitt, P., Mandel, J., & M. Guedj, M. (2015), A *comparison of six methods for missing data imputation*, Journal of Biometrics and Biostatistics, vol. 6 pp. 1-6.

Singer, J.D., J. B. Willett, J.B. (2002). *Applied Longitudinal Data Analysis*: Modeling Change and Event Occurrence, Oxford University.

Streiner, D. L. (2002).The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry*, 47, 68-75.