ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 11(16), DOI: 10.17485/ijst/2018/v11i16/118701, April 2018*

The Efficiency of Multiple Imputation and Maximum Likelihood Methods for Estimating Missing Values

Tlhalitshi Volition Montshiwa*, Ntebo Moroke and Elias Munapo

Department of Statistics and Operations Research, North West University Mafikeng Campus, South Africa; volition.montshiwa@nwu.ac.za, ntebo.moroke@nwu.ac.za, emunapo@gmail.com

Abstract

Objectives: This study investigated the efficiency of Multiple Imputation (MI) and Maximum Likelihood (ML) methods for estimating missing values. The study was set to use the findings to make recommendations for future studies about the impact of missing data imputation on the accuracy of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). **Methods**: The completedset (with no missing values) used in this study was collected in 2010/11 through the Income and Expenditure Survey (IES) and had 25328 observations. Missing data were generated by randomly deleting 10%, 20%, 30%, 40% and 50% of the values from the complete dataset. The missing values in each of the five datasets were imputed using MI and ML methods. Subsequently, absolute error values of AIC and BIC from multiple regression analysis were computed for each dataset. The study then compared the absolute errors for each missing value imputation method. **Findings**: The findings of the study revealed that AIC and BIC are more accurate when missing values are estimated by the Full Information Maximum Likelihood (FIML) of the ML algorithm, provided 10% of the data are missing. For all datasets, AIC and BIC were least accurate when missing values were imputed by Expectation Maximisation (EM) of the ML algorithm. The findings also showed that AIC and BIC are more accurate when the rate of MISSINGNESS gets large provided missing values were estimated using either the Fully Conditional Specification (FCS) or Markov Chain Monte Carlo (MCMC), MI algorithms. **Application**: When the rate of MISSINGNESS is small (at most 10%), FIML should be used to handle missing data if AIC and BIC are going to be used. Also both FCS and MCMC should be considered over EM algorithms when the rate of MISSINGNESS is high (at least 40% missing).

Keywords: Maximum Likelihood Imputation, Multiple Imputation, AIC, BIC

1. Introduction

Missing values occur in real life data analysis and this is due to several reasons which include lack of subject participation during data collection, deletion of outliers from the dataset and loss of documents due to poor handling. Missing data may affect the strength, reliability, integrity and validity of causal inference[1]. This nuisance may occur in Multiple Linear Regression (MLR) analysis where there is a multi-causal relationship between the dependent and independent variables. More specifically, in MLR analysis, a missing value from one variable leads to the exclusion of all other values for that particular observation. Exclusion of values results in the existing information from other variables not being utilised[2]. As such, it is important to address missing values in MLR analysis. In order to parsimoniously utilise the available values in an observation with missing values, analysts often resort to estimating

**Author for correspondence*

the missing data by assigning arbitrary values to them. However, previous studies have revealed that results are degraded by simply assigning arbitrary values to missing data elements[3].

The arbitrary values assigned to the missing values are imputed using numerous traditional methods such as mean substitution and regression imputation[4]. However, several studies revealed that the use of these traditional single imputation methods may bias measures such as parameter estimates, Standard Errors (SE), Confidence Intervals (CI), Odds Ratios (OR) and the coefficient of determination ()[5–7]. Theseprevious studies advocate the use of novel imputation methods namely: Multiple Imputation (MI) and Maximum Likelihood (ML). Several studies found that these methods can provide unbiased estimates of SE, CI, OR and as opposed to the traditional single imputation methods[8–10]. For these reasons, MI and ML methods are becoming the typical modus operandi for handling missing data. On the other hand, although the effects of using MI and ML in imputing missing data on SE, CI, OR and are known, previous studies have not focused on the impact of these imputation methods on the accuracy of model selection criteria.

Model selection criteria are mostly useful in determining the best MLR model from numerous models with the same dependent variable but different number of predictor variables. Model selection is usually performed to ensure that the MLR model is correctly specified. Model misspecification (under-fitting or over-fitting) have serious impacts on the MLR analysis. For instance, over-fitting the model inflates the variance[11]. The Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are commonly used in selecting MLR models[12]. It is therefore important to ensure that AIC and BIC are accurate so that the rightful MLR model is chosen. However, the effect of estimating missing data using MI and ML methods on the accuracy of AIC and BIC has not been of interest in previous studies. As such, this study generally seeks to assess the effect of estimating missing data using MI and ML methods on the accuracy of AIC and BIC. This study contributes a new idea to the application of missing data imputation methods. The findings of this study may assist researchers in selecting the most appropriate method for imputing missing data when conducting MLR analysis while ensuring that the accuracy of AIC and BIC is not hindered by imputed values.

Literature shows that various studies have explored the effect of the application of missing data imputation methods on the accuracy of diverse statistics except for AIC and BIC. The studies by[13–15] is examples of such studies and is discussed next. In[14] investigated the accuracy of the mean, SE, Pearson correlation coefficient (r), percent misclassified and the K statistic when missing values were imputed using MI, single regression, individual mean, overall mean, the subject's preceding response and random selection of a value from 1 to 4. The results of the study by[14] showed that the accuracy of the statistics under study was best when missing data were addressed using MI than other missing data handling methods. In[15] conducted a study to assess the accuracy of SE and parameter estimates of a logistic regression model when using datasets with imputed missing values. The authors compared the bias in SE and parameter estimates computed from datasets which had missing values addressed by mean substitution, MI and Full Information Maximum likelihood (FIML). The results of the study by[15] revealed that SE and parameter estimates were more accurate when missing data have been handled using either MI or FIML but the bias was severe when mean substitution has been used in addressing the missing values. evaluated the accuracy of SE and parameter estimates of regression analysis for the actual (complete dataset), the dataset with missing values addressed using mean substitution and the datasets with MI- and ML-imputed values. The study by[13] revealed that SE and parameter estimates were more accurate when mean substitution was applied than when MI and the ML were used for imputing the missing values. However, the study showed that the accuracy of SE and parameter estimates does not differ reasonably between MI and ML. The study by[16] assessed the effect of imputing missing data on the accuracy of standard deviation (SD), means, the K statistic and spearman's rank correlation () values. Accuracy was measured by comparing the statistics computed from the complete dataset (actual data)

to the ones computed from the datasets with imputed missing values. The results of the study by[16] revealed that the SD, , K and values from the dataset with MI imputed missing values were more accurate than the ones from the datasets to whichsingle regression, individual mean and overall mean were applied. It is evident from literature that the accuracy of AIC and BIC has not been of interest to previous studies on missing value handling methods. As such, it was important to conduct this study.

2. Study Objectives

This study was set to:

- Assess the impact of imputing missing values with MI and ML on the accuracy of AIC and BIC
- Use the findings of the study to make recommendations for future studies about the impact missing data imputation on AIC and BIC

3. Data Analysis and Results

3.1 Data

The complete dataset used in this study were collected in 2010/11 by Statistics South Africa (Stats SA) through the Income and Expenditure Survey (IES). The following link may be used to access the data: "www.datafirst.uct.ac.za". This study used the following variables in fitting the MLR model: bedrooms (X_1), living rooms ($X_{2)}$, dining rooms (X_3), multipurpose rooms (X_4), bathrooms (X_5) and other bedrooms (X_6) as independents and the value of dwelling (X_7) as the dependent variable. That is, the model was fit to predict the value of dwelling based on its number of rooms. However, prediction of value of dwelling and the regression model were not the focus of this study, but the interest was on AIC and BIC. The data comprised 25328 observations. Microsoft Excel 2013 and the Statistical Analysis Software (SAS) version 9.3, registered to the SAS Institute Inc. Cary, NC, USA were used for data analysis.

4.2 Simulation of Missing Data

An important aspect to consider when dealing with missing values is the MISSINGNESS mechanisms. Missing values may be Missing Not at random (MNAR), Missing at Random (MAR) or Missing Completely at Random (MCAR)[17] explains that values that are MCAR (ignorable MISSINGNESS) are a random subset of the total data. This means that there is no pattern to MISSINGNESS and the missing data are not related to any other variable. In[17] describe MAR as missing data that are related to other variables in the dataset, but not to participants' responses to the question(s). The authors further define MNAR (also known as non-ignorable) to mean that there is either a relationship between the missing data and the independent variable or between the missing data and the participants' responses to the question (s). These MISSINGNESS mechanisms are key assumptions that guide the choice of the imputation methods. The missing data imputation methods considered in this study, MI and ML, assume that the data are MAR[18,19]. Therefore, this study only focused on the data that are MAR.

This study was limited to one of the two MAR mechanisms presented in the study by[20]. To aid in simulating this MAR mechanism, the variable "response_rate_category" was added to the dataset, by dividing the provinces shown in Table 1 into two categories namely: low_response_rate=1 and high_response_rate=2.

The low_response_rate category comprises six provinces. This category makes up 59% of the whole dataset hence it is large enough to include the highest rate of MISSINGNESS of 50%. In simulating the MAR values, the MISSINGNESS was set to be related to the variable "response_rate_category" but was not related to the value of dwelling (). This practice was adopted from[20]. MAR values were simulated by randomly deleting values in the low_response_rate category. The rate of MISSINGNESSS were set to vary from 10% to 50% yielding five datasets which are referred to as MAR-10 to MAR-50 in this study. The random selection was performed using a Microsoft Excel add-in known as Kutools (7.81), and then the selected values were deleted to simulate the MAR mechanism.

Table 1. Percentage response rate by province

Response rate category	Province	Percent
1	Gauteng	80.8
1	Western Cape	91.3
1	Northern Cape	94.1
1	Northern Cape	97.0
1	Free State	97.3
1	Mpumalanga	97.6
2	Eastern Cape	98.9
2	Limpopo	99.1
2	KwaZulu-Natal	99.2

3.3 Preliminary Analysis

3.3.1 Test for the Multivariate Normality of Data

In[21] explain that the violation of the assumption of multivariate normality when using MI or ML can lead to doubtful imputed values that will not work at all for certain kinds of analyses[22] that when each variable is normally distributed multi variate normality can be assumed.

As such, this study tested u normality for all variables (X_1 to X_7) using the Smirnov (K-S) test which is recommended by[23]for sample sizes of at least 50. The K-S test is set to test the following hypothesis:

H_0: The residuals are normally distributed

H_1: The residuals are not normally distributed

The K-S statistic was computed using the following formula which was adopted and Foreman[24]:

$$Z = \sqrt{n}\max\left(|D|, |\widetilde{D}|\right), \quad (1)$$

Where n=25328, which is the sample size,

$$D = \left|\hat{F}_{x_i} - \hat{S}_{x_{i-1}}\right| \quad (2)$$

And

$$\widetilde{D} = \left|\hat{F}_{x_i} - \hat{S}_{x_i}\right|, \quad (3)$$

Which are frequency distributions and $\hat{F}_{x_i}$, and $\hat{S}_{x_{i-1}}$ are computed by converting relative imperial- and relative observed- frequency distributions into cumulative frequency distributions.

For all the six datasets (the observed complete dataset) and MAR-10 to MAR-50, the p-values for the Kolmogorov tests for all the variables were significant at 5% significance level. Therefore, the null hypothesis that the data were collected from a normally distributed population was rejected[1].

3.3.2 Remedies for Violations of the Multivariate Normality Assumption

This study used the Box-Cox multivariate transformations recommended by[25] as a remedy for departures from multivariate normality. The multivariate Box-Cox transformations were achieved using Equations (4), (5) and (6), which are adopted from[25]. The aim was to find the cluster of transformations defined by:

$$\lambda = \lambda_1 \ldots \lambda_p, \quad (4)$$

Which maximises the following function?

$$\ell(\lambda) = -\frac{1}{2}\ln|S_n| + \sum_{j=1}^{p}\left[(\lambda_j - 1)\sum_{i=1}^{n}(X_{ij})\right], \quad (5)$$

Where X_{ij} is the i[th] observation on the j[th] variable and S_n denotes the maximum likelihood estimate of the covariance of the transformed data and is calculated as follows:

$$S_n = \frac{1}{n}\sum_{n}^{1}\left(X_{ij}^{\lambda_j} - X_j^{(\bar{\lambda}_j)}\right)\left(X_{ik}^{\lambda_k} - X_k^{(\bar{\lambda}_k)}\right), \quad (6)$$

Where $X_k^{(\bar{\lambda}_k)}$ denotes the sample mean of n transformed observation on the k[th] variable.

Following the application of the various Box-Cox transformations, the K-S tests for normality were performed again and the p-values for the tests were insignificant at 5% significance level, hence the null hypothesis that the data are normal was not rejected. All variables were normally distributed, hence the multivariate normality was assumed and MI and ML imputations could then be applied.

3.4 Application of MI and ML Imputation Methods to Address Missing Values

4.4.1 ML Imputation Using the EM Algorithm

The EM algorithm was performed using the steps in Figure 1 which are adopted from[26].

The steps summarised in Figure 1 are explained as follows[26]:

Let X_i contain observed values for the i[th] case, let Z_i to have missing values and i = 1, 2...n, and then it is explained that if Z_i did not have missing values, the log likelihood function could be given by:

$$\ell_0(\theta) = \sum_{i=1}^{n}\log f(x_i, z_j; \theta) \quad (7)$$

However, it is explained that because only X_i are fully observed, the log likelihood function is defined by:

$$\ell_0(\theta) = \sum_{i=1}\log f(x_i; \theta), \quad (8)$$

where

$$f(x_i; \theta) = \int f(x_i, z_j; \theta)\, dz \quad (9)$$

The aim is to find

$$\hat{\theta} = (\mu, \Sigma), \quad (10)$$

Which maximises $\ell_0(\theta)$, where μ and Σ in Equation (10) denote the mean vector and the variance-covariance matrix respectively. Using the initial values μ^0 and Σ^0, the E-step utilises $\ell(\theta)$ to obtain $Q(\theta)$ as follows:

$$Q(\theta) = E\left\{\sum_{i=1}^{n}\log f(x_i, z_i, \theta^{(j)})\right\} \quad (11)$$

The M-step maximises $Q(\theta)$ to yield $\theta^{(j+1)}$ which is further used in the next E-step. By letting $y_i = (x'_i, z'_i)$ to denote the complete data for the i[th] case, the M step for maximising $Q(\theta)$ is given by:

$$\mu^{(j+1)} = \frac{1}{n}\sum_{i=1}^{n}E(y_i, |x_i, \theta^{(j)}) \quad (12)$$

and

$$\Sigma^{(j+1)} = \frac{1}{n}\sum_{i=1}^{n}E(y_i y'_i, |x_i, \theta^{(j)}) - \mu^{(j+1)}\mu^{(j+1)}, \quad (13)$$

Alternating between the E and M steps to the value of $\hat{\theta}$ maximises $\mathcal{L}_0(\theta)$ at convergence.

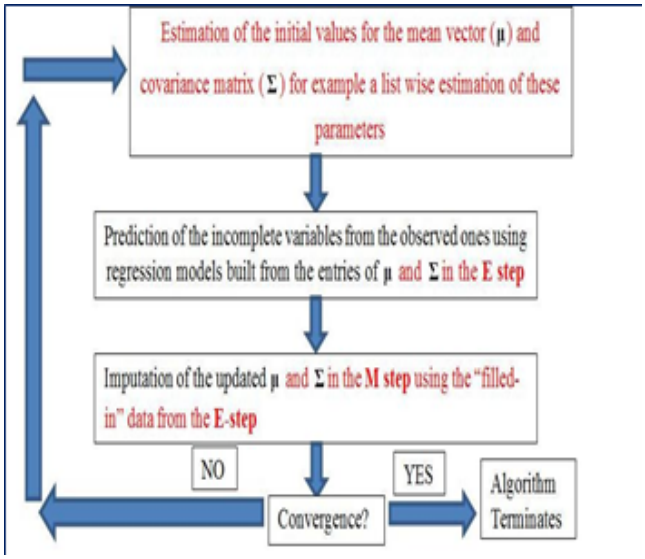

Figure 1. Summary of the EM algorithm steps for handling missing data.

3.4.2 ML Imputation Using the FIML Algorithm

As[4] explains, FIML aims to compute the likelihood function for each observation using only the variables with observed values for that case. The author explains that if variables have observed values for case then the dimension of $\boldsymbol{\Sigma_i(\theta)}$ is $\mathbf{v \times v}$ and for $\boldsymbol{\mu_i(\theta)}$ as well as $\mathbf{z_i}$ is $\mathbf{v \times 1}$.

The methodology of the FIML algorithm used in this study is adopted from[4]. The likelihood function for each variable is computed as follows:

$$\ln \ell_i(\theta) = K_i - \frac{1}{2}\ln|\Sigma_i(\theta)| - \frac{1}{2}[z_i - \mu_i(\theta)]^{\Sigma^{-1}(\theta)[z_i - \mu_i(\theta)]}, \quad (14)$$

Where $\mathbf{z_i}$ denotes a vector of observed variables for the $\mathbf{i^{th}}$ observation and $\mathbf{K_i}$ is a constant unrelated to $\boldsymbol{\theta}$.

The total likelihood is therefore given by the following formula:

$$\ln \ell(\theta) = \sum_{i=1}^{n} \ln \ell_i(\theta) \quad (15)$$

The aim of FIML is to select the value of $\boldsymbol{\theta}$ that maximises $\ln \mathcal{L}(\theta)$.

3.4.3 MI Using the MCMC Algorithm

This study adopted the following methodology of the MCMC algorithm from[6]. Let θ, Y_{obs} and Y_{mis} denote the distributional parameters, the observed values and the missing values respectively, then the posterior predictive distribution $f(y_{mis}|y_{obs},\theta)$ can be predicted in the posterior step (P-step) of MCMC. In the imputation step (I-step), MCMC draws the respective missing values and model parameters at the iteration as follows:

First $\mathbf{y_{mis}^{(l)}}$ from $\mathbf{(y_{mis}|y_{obs}, \theta^{l-1})}$, then the algorithm estimates $\boldsymbol{\theta^{(l)}}$ from $\mathbf{f\left(\theta, y_{obs}|y_{mis}^{(l)}\right)}$. After burn-in (the early part of the chain that is removed before the estimation of the function to minimise bias)[6], MCMC generates m samples from the posterior function $\mathbf{f(y_{mis}|y_{obs}, \theta)}$, then imputed datasets are formed from the simulated copies of $\mathbf{y_{mis}}$. For each imputed sample, complete case analysis is then used to estimate the model parameters $\bar{\theta}$ for $\mathbf{l = 1 \ldots m}$.

To account for variations, the corresponding sample co-variances are also computed. These co-variances are calculated as follows:

the within-imputation variance denoted:

$$\mathbf{W_t = Var(\theta)}$$

for $\mathbf{l = 1 \ldots m}$. (16)

The average of Equation (16) reflects the sampling variation and is calculated using:

$$\bar{W} = \frac{1}{m}\sum_{i=1}^{m} W_i \quad (17)$$

The between-imputation variance explains the variation due to imputation and is computed using:

$$B = \frac{1}{m-1}\sum_{i=1}^{m}(\theta_1 - \bar{\theta})(\hat{\theta} - \bar{\theta})', \quad (18)$$

Where

$$\bar{\theta} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}. \quad (19)$$

The total variance of $\bar{\theta}$ is then computed using:

$$T_i = [\bar{W}]_{ii} + \left(1 + \frac{1}{M}\right)[B]_{ii}, \quad (20)$$

Where $[\bar{W}]_{ii}$ and $[B]_{ii}$ denote the i^{th} diagonal elements of matrices W and B respectively.

Explain that $T_i^{-1/2}(\bar{\theta}_i) - \theta_i \sim t_{v_i}, M$ as $M \to \infty$,

where

$$v_i, M = (M-1)\left[1 + \frac{[\bar{W}]_{ii}}{(1+M^{-1})[B]_{ii}}\right]^2 \quad (21)$$

3.4.4 MI Using the FCS Algorithm

The methodology for FCS applied in this study was adopted from[27] and is explained as follows:

Let X_j denote the variable whose missing values are to be imputed and since there are seven variables in this study, j can take any integer in the interval from 1 to 7. The P-step begins by regressing the observed values of X_j over the observed values of the other six variables using the following general linear regression model:

$$X_j = \beta_0 + \beta_1 X_i + \varepsilon, \quad (22)$$

Where i represent all the remaining six variables after j is chosen. For instance, if FCS starts with X_1 then the regression model is:

$$X_1 = \beta_0 + \beta_1 X_2 + \beta_2 X_3 \ldots + \beta_6 X_7 + \varepsilon, \quad (23)$$

The general model will therefore yield the parameter estimates $\hat{\beta}_j$, the error variance $\hat{\sigma}_j^2$ and the inverse of Sum of Squares and Cross Products (SSCP) matrix from the regression, V_j. In the I-step, X_j^* is imputed by first drawing random parameter estimates from their joint posterior estimates. First the error variance is drawn from its posterior as follows:

$$\sigma_{*i}^2 = \frac{\hat{\sigma}_i^2(n-k-1)}{g}, \quad (24)$$

Where

n_j is the number of observed values from the variable being imputed, $k = 6$, which is the number of parameters (the intercept excluded) in the regression model and is a random draw from the central chi-square distribution χ^2_{n-k-1}. Next, the regression parameters are drawn from their conditional posterior distribution as:

$$\beta_* = \hat{\beta} + \sigma_{*j}\sqrt{V}(Z), \quad (25)$$

Where

$\sqrt{V}$ is defined as the upper triangle in the square root of $V = (X'X)^{-1}$, Z denotes $k+1$ dimensional vector of independent normal $(N(0,1))$ variates. The final phase in the I-step is to impute the missing values of X_j using the succeeding regression equation:

$$Y_{*j} = \beta_0 + \beta_{*0}Y_i + x\sigma_{*j}, \quad (26)$$

where x is defined as a random draw of a $N(0,1))$ deviate. FCS then goes to the P-step for the next X_j using linear regression to define $p(Y_{j,mis}|Y_{i,obs}Y_{j,obs}, \theta_4)$.

The algorithm alternates between the P-step and the I-step until all the missing values are imputed.

3.4.5 *Detection and Removal of Redundant and Highly Collinear Variables from the Complete Dataset*

An important assumption in MLR analysis is that the independent variables should not be highly collinear (multicollinearity). Also, there should be no redundant independent variables, that is, variables which are not related to the dependent variable X_7. Stepwise regression, a method recommended by[28] was used in detecting and removing the highly collinear variables. The SAS default significance level of 0.15 was used in determining variables which are retainable in the final regression model. This was achieved by testing the hypothesis $H_0: \hat{\beta}_i = 0$ where a p-value less than 0.15 rejects the null hypothesis and retains the significant variable. The results of the stepwise regression method are presented in Table 2.

Table 2 shows that and were excluded from the regression model. This is because X_4 was found to be a linear combination of X_3, such that $X_3 = X_4$. Also X_5 was found to be a linear combination of X_6, such that. For comparison reasons, X_3, X_4 and X_6 werealso excluded from all the datasets with simulated missing values (MAR-10 to MAR-50).

3.5 Assessment of the Effect of Analysing Data with MI- and ML- Imputed Values on the Accuracy of AIC and BIC

Following the imputation of missing values with EM, FIML, MCMC and FCS, this study computed AIC and BIC for the observed dataset with no missing values and the five datasets which had imputed values. The AIC and BIC were computed within the stepwise regression method using the following formulas which were adopted from[28,29]:

$$AIC = -2\ell\ell + 2k, \quad (27)$$

Where $\ell\ell$ he log-likelihood is function and k is the number of parameters in the model.

Table 2. Summary of the stepwise selection method

					* Optimal Value Of Criterion				
Parameter	DF	Estimate	Standard Error	t Value	Step	Effect	Number	PRESS	F Value
						Entered	Effects In		
Intercept	1	115912	4488.866142	25.82					
X_1	1	-34975	2272.846602	-15.39	**0**	**Intercept**	1	7.15864E15	0.00
X_2	1	116198	6768.236105	17.17	**1**	**X_5**	2	5.39877E15	8270.83
X_5	1	346588	4837.646896	71.64	**2**	**X_2**	3	5.36628E15	157.34
X_7	1	68782	6581.615795	10.45	**3**	**X_1**	4	5.32452E15	200.57
					4	**X_7**	5	5.30252E15*	109.22

$$BIC = -2\ell\ell + k\log n, \quad (28)$$

Where n is the sample size.

Subsequent to computing AIC and BIC values, this study calculated absolute errors of these selection criteria. This was achieved by taking the absolute values of the difference between the AIC and BIC from the actual dataset without missing values and the ones from the datasets with MI- and ML-imputed values. These absolute errors were then used in assessing the accuracy of AIC and BIC. The assessment criteria was such that, lower absolute errors indicated high accuracy of AIC and BIC under a particular MI and ML algorithm mover five rates of MISSINGNESSS namely: 10%, 20%, 30%, 40% and 50%. Needle plots were plotted in order to summarise the absolute errors and to better understand the relationships between the rate of MISSINGNESSS and the accuracy of AIC and BIC. These needle plots are presented in Section 4.6.

3.6 Assessment of the Accuracy of AIC and BIC under MAR-10 to MAR-50

Figure 2 shows that the EM algorithm (represented by blue squares) has the highest absolute error of AIC across all rates of MISSINGNESSS except for when 50% of values were missing. When 10% of the values were missing, FIML (represented by maroon triangles) had the lowest absolute error but the error increased gradually until it was highest when 50% of the values were missing. It can be observed from Figure 2 that the absolute errors of AIC increases as the rate of MISSINGNESS increases for both ML algorithms (EM and FIML). Figure 2 also shows that the MI algorithms (MCMC and FCS) had exactly equal absolute errors of AIC with FCS (represented by red crosses) being completely equal (inside) MCMC (represented by green dots). The absolute error of AIC for MI algorithms falls in between that of EM and FIML when 10% values are missing, but it is lowest when 20% to 50% of the data are missing. In general, the accuracy of AIC increases as the rate of MISSINGNESS increases provided the missing values were computed using MI algorithms (MCMC and FCS).

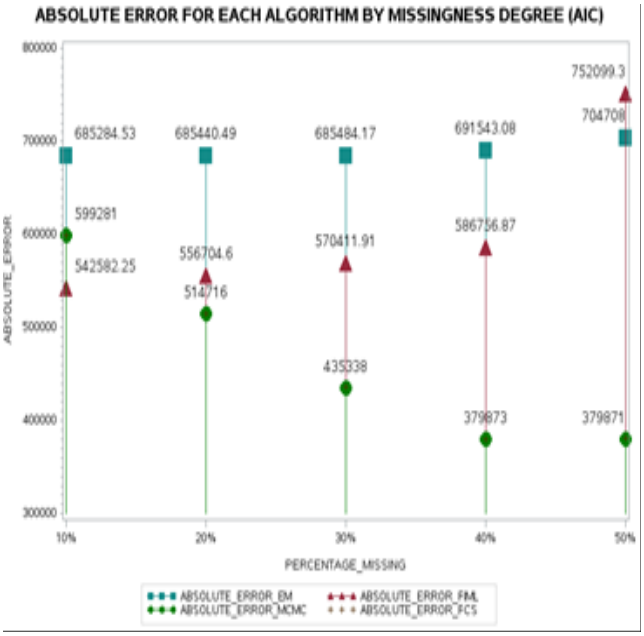

Figure 2. Absolute errors for AIC.

Figure 3 shows that the EM algorithm (represented by blue squares) has the highest absolute error of BIC across all rates of MISSINGNESSS except for when 50% of values were missing. When 10%, 20% and 30% of values were missing, FIML (represented by maroon triangles) had the lowest absolute error but the error increased gradually until it was higher than both MI algorithms (MCMC and FCS) when 40% of the values were missing. The data with FIML imputed values has the highest error rate of BIC when 50% of the values were missing and it is higher than all the other three algorithms at that rate of MISSINGNESS. It can be observed from Figure 3 that the absolute errors of BIC increases as the rate of MISSINGNESS increase for both ML algorithms (EM and FIML). Figure 3 also shows that the MI algorithms (MCMC and FCS) had exactly equal absolute errors of BIC with FCS (represented by red crosses) being completely equal (inside) FCS (represented by green dots). The absolute error of BIC for MI algorithms falls in between that of EM and FIML when 10%, 20% and 30% of the values were missing, but it is lowest when 40% and 50% of the data are missing. Generally MI algorithms (MCMC

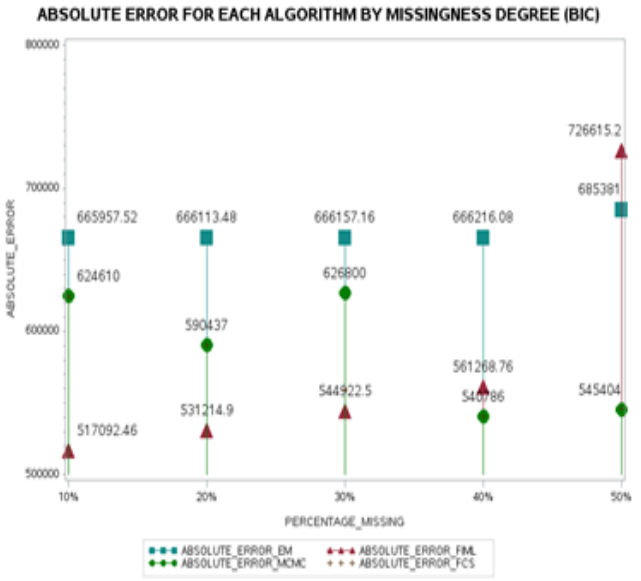

Figure 3. Absolute errors for BIC.

and FCS) performs well than both ML algorithms (FIML and EM) when the rate of MISSINGNESS is high (40% and 50% of the values missing).

5. Summary and Conclusions

The main objective of this study was to assess the accuracy of AIC and BIC using data that had missing values imputed using MI and ML methods. This objective was achieved by applying FIML and EM algorithms for ML; and MCMC and FCS algorithms for MI. The algorithms were implemented on datasets with some of the values simulated to be Missing at Random (MAR). Following imputation of missing values, MLR analysis was performed and absolute errors of AIC and BIC were used as criteria for determining the accuracy of AIC and BIC. The accuracy of AIC and BIC was assessed for five rates of amusingness (10% to 50% missing values). The results of this study revealed that generally, the accuracy of AIC and BIC is poor when missing values have been imputed using the EM algorithm of the ML imputation method. In addition, the accuracy of AIC and BIC becomes poorer as the rate of MISSINGNESS increases when the data used had missing values imputed using FIML of the ML imputation method. The accuracy of AIC and BIC were found to increase as the rate of MISSINGNESS increases provided the missing values were imputed using MI algorithms (FIML and MI). It is therefore evident that the accuracy of AIC and BIC may be affected by both the type of missing data imputation method applied and the rate of MISSINGNESS.

Based on the findings of the study, it is recommended that when the rate of MISSINGNESS is small (at most 10%), FIML should be used to handle missing data if AIC and BIC are going to be used in selecting the best model. MI algorithms (both FCS and MCMC) should be considered over EM algorithms when the rate of MISSINGNESS is high (at least 40% missing) if the objective of the study is to use the dataset with imputed values for computing AIC and BIC. It is recommended that further research may consider other imputation methods and other algorithms when assessing the accuracy of AIC and BIC using data with imputed values. Eminent research may consider other rates of MISSINGNESS than the ones that were considered in this study. This study considered the MAR mechanism only; therefore further research may interrogate other MISSINGNESS mechanisms such as MNAR and MCAR. Generally, this study showed the importance of choosing the rightful missing data imputation method depending on the objective of the study, for example, when the objective is to use model selection criteria in comparing and selecting regression models.

6. References

1. Elliott RJ, Morrell CH. Learning SAS in the computer lab. 3rd ed. Boston: CENGAGE Learning; 2009.
2. Schlomer G, Bauman S, Card N. Best practices for missing data management in counseling psychology. Journal of Counseling Psychology. 2010; 57(1):1–10. Crossref. PMid:21133556
3. Matignon R. Data mining using SAS Enterprise miner. 1st ed. Hoboken, New Jersey: Wiley-Interscience; 2007. p. 1–580. Crossref.
4. Enders C. Applied missing data analysis. New York: Guilford Press; 2009. p. 1–401.
5. Acock AC. Working with missing values. Journal of Marriage and family. 2005; 67(4):1012–28. Crossref.
6. Heijden VDGJ, Donders ART, Stijnen T, Moons KG. Imputation of missing values is superior to complete case

analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. Journal of Clinical Epidemiology. 2006; 59(10):1102–9. Crossref. PMid:16980151
7. White IR, Carlin JB. Bias and efficiency of multiple imputations compared with complete-case analysis for missing covariate values. Statistics in Medicine. 2010; 29(28):2920–31. Crossref. PMid:20842622
8. Stamatis HD. Six sigma and beyond. Boca Raton: Statistical Process Control St. Lucie Press; 2003. p. 1–520.
9. Heiberger RM, Holland B. Statistical analysis and data display. Berlin: Springer; 2004. p. 1–730. Crossref.
10. Feigelson E, Babu G. Modern statistical methods for astronomy. Cambridge: Cambridge University Press; 2012. p. 1–13. Crossref.
11. Brown T. Confirmatory factor analysis for applied research. New York: Guilford Press; 2006. p. 1–462.
12. Gordon R. Regression analysis for the social sciences. New York: Routledge; 2010. p. 1–632. PMCid:PMC2901108
13. Fitzmaurice G, Laird N, Ware J. Applied longitudinal analysis. Hoboken, New Jersey: Wiley-Interscience; 2004 Jun.
14. Shrive FM, Stuart H, Quan H, Ghali WA. Dealing with missing data in a multi-question depression scale: a comparison of imputation methods. BMC Medical Research Methodology. 2006; 6(1):1. Crossref. PMid:17166270 PMCid:PMC1716168
15. Schlomer G, Bauman S, Card N. Best practices for missing data management in counseling psychology. Journal of Counseling Psychology. 2010; 57(1):1–10. Crossref. PMid:21133556
16. Svolba G. Data quality for analytics using SAS. Cary, North Carolina: SAS Institute; 2013 Nov. p. 1–29.
17. Weisberg S, Fox J. An R companion to applied regression. Thousand Oaks, CA: SAGE; 2011. p. 472.
18. Ravishanker N, Dey D. A first course in linear model theory. Boca Raton: Chapman and Hall/CRC. 2002. p. 1–496.
19. MacFarlane I, Veach P, LeRoy B. Genetic counselling research. New York: Oxford University Press; 2014. p. 1–304.
20. Zaidman-Zait A, Zumbo BD. Can multilevel (HLM) models of change over time adequately handle missing Data? Journal of Educational Research and Policy Studies. 2013; 13(1):18–31.
21. McKnight P, McKnight K. Missing data. New York: Guilford Publications; 2007. p. 861–3.
22. Wilks D. Statistical methods in the atmospheric sciences. Amsterdam: Academic Press; 2006. p. 1–704.
23. Rovai A, Baker J, Ponton M. Social science research design and statistics. Chesapeake, VA: Water tree Press; 2013. p. 1–630.
24. Corder G, Foreman D. Nonparametric statistics for non-statisticians. New Jersey: Wiley; 2009. p. 1–264. Crossref.
25. Rencher AC. Methods of multivariate analysis. New York: John Wiley and Sons; 2003. p. 1–727.
26. Malthouse E. Segmentation and lifetime value models using SAS. Cary, North Carolina: SAS Institute; 2013. p. 1–182.
27. Heeringa S, West BT, Berglund PA. Applied survey data analysis. Boca Raton, FL: Chapman & Hall. 2010. p. 1–53. Crossref.
28. Frees E, Derrig R, Meyers G. Predictive modeling applications in actuarial science. New York: Cambridge University Press; 2014. p. 1–20. Crossref.
29. Hilbe JM. Modelling count data. New York: Cambridge University Press; 2014 Aug.