*Maejo Int. J. Sci. Technol.* **2024**, *18*(01), 61-74

*Maejo International*
*Journal of Science and Technology*

**ISSN 1905-7873; e- ISSN: 2697-4746**
Available online at www.mijst.mju.ac.th

***Full Paper***

# Imputation of missing dependent variable in binary logistic regression

**Tidarat Thammachoto and Klairung Samart**[*]

Statistics and Applications Research Unit, Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla, Thailand

[*] Corresponding author, e-mail: klairung.s@psu.ac.th

*Received: 18 October 2023 / Accepted: 3 March 2024 / Published: 5 March 2024*

**Abstract:** Missing data are an important issue affecting data analysis. This study develops and compares methods of imputing missing data in binary logistic regression. Seven imputation methods are applied: mode imputation, hot deck imputation, multiple imputation (MI), k-nearest neighbour imputation, random forest imputation, logistic regression imputation (LR), and modified logistic regression imputation (MLR). Missing data are simulated in three conditions: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The simulations were run using sample sizes of 20, 50, 100, 150, 200, 500 and 1,000 and missing percentages of 10%, 20%, 30% and 40%. The simulated missing data in the three conditions were applied to real-life heart disease data and the obtained data sets were analysed using the seven imputation methods. Performance was compared by estimating the mean square error of each analysis. The results reveal that when the missing data condition is either MCAR or MAR, the MLR method gives the best performance with small sample sizes ($n \leq 50$) at most levels of missing data, while the MI method gives the best performance with large sample sizes. For the MNAR condition, the LR method gives the best performance with small sample sizes for all levels of missing data.

**Keywords:** missing data, imputation method, logistic regression, estimated mean square error

## INTRODUCTION

In a data collection process missing data can arise. For example, when collecting information about a medical condition, there is a possibility that some data may be lost. The data could be related to risk factors that affect the condition, or to information that indicates whether a patient has the condition or not. This is a problem that significantly affects the analysis of medical data, especially in a logistic regression analysis, which identifies the relationship between a qualitative dependent variable and independent variables [1]. In binary logistic regression model a dependent variable is binary and independent variables can be either categorical or continuous [2].

Missing data are classified into three categories: those that are missing completely at random (MCAR), those that are missing at random (MAR), and those that are missing not at random (MNAR). MCAR refers to a situation where missing data are unrelated to the observed or unobserved data. In other words there are no apparent differences between participants with missing data and those with complete data. In the MAR category the probability that the data are missing is related to observed data, but not to unobserved data. For example, if women refrained from disclosing their weight, the undisclosed weight would be denoted as MAR. MNAR describes a scenario where missing data are related to unobserved data [3]. For example, individuals with the lowest level of education tend to have incomplete education data. The statistical significance of missing data and the approach to handling them are substantially impacted by their classification.

A simple technique that is widely used as the default method for handling missing data is to omit cases that contain them. A serious problem with this approach is that it reduces the size of the data set. It also limits the analysis to those observations in which all values are observed, which often results in a biased estimate and a loss of precision [4]. To maintain the size and representativeness of the data set, missing data can be replaced with substitute values imputed from available data. Until now there has been little study regarding the handling of missing values, especially in the context of missing dependent variables in binary logistic regression models, and therefore more investigation is needed to develop more efficient imputation methods.

Some researchers have studied the performance of methods of imputing categorical independent variables. For example, Peng and Zhu [5] assessed the imputation of categorical variables in logistic regression using multiple imputation (MI) method and expectation maximisation (EM) method. Results showed that the former performed better than the latter. Waljee et al. [6] compared the performance of four imputation methods for categorical and continuous variables: mean imputation (Mean), MI, k-nearest neighbour imputation (KNN) and random forest imputation (RF). Results showed that RF was a highly accurate method of imputing missing laboratory data. Xu et al. [7] compared the accuracy of four common methods of imputing categorical variables in logistic regression: direct deletion, mode imputation (Mode), hot deck imputation (HD) and MI. Results showed that the MI method produced the best performance. These findings were consistent with those of Tsiampalis and Panagiotakos [8], who compared the performance of seven imputation methods for categorical and continuous variables in both logistic and Poisson regressions. Their results also showed that the MI method performed best. Mohamed et al. [9] reviewed the performance of ten imputation methods for missing data in the binary logistic regression model: Mean, HD, last observation carried forward, stochastic regression imputation, predictive mean matching (PMM), EM, KNN, logistic regression imputation (LR), RF and nearest neighbour hot deck. Results showed that the EM and KNN methods were most appropriate for imputing missing data in the binary logistic regression model.

To the best of our knowledge, the investigation of imputation methods for missing binary outcomes has been limited. Ma et al. [10] compared six MI methods that accounted for intra-cluster correlation for missing binary outcomes in cluster randomised trials with standard imputation methods and complete case analysis under the MCAR missing condition. The within-cluster MI methods applied were logistic regression, propensity score and Markov chain Monte Carlo. Three across-cluster MI methods applied were propensity score, random-effects logistic regression and fixed-effect logistic regression. Results showed that the within-cluster and across-cluster MI methods were more appropriate for handling the missing outcomes from cluster randomised trials. Sullivan et al. [11] investigated MI for relative risk estimation with missing data existing for both outcome and exposure variables induced under the MAR condition. Standard model-based MI approaches imputed missing data using multivariate normal imputation or fully conditional specification in a logistic outcome imputation model. Results showed that the fully conditional specification performed better than the multivariate normal imputation. Mohamed et al. [12] investigated nine imputation approaches in a binary logistic regression model with MAR as the missing condition: KNN, EM, HD, RF, regression imputation, LR, Mean, PMM and a new imputation approach EPK, which was an average of three single imputation methods: EM, PMM and KNN. The EPK and EM methods outperformed other imputation methods by having the lowest Akaike information criterion and Bayesian information criterion values.

Therefore, in this study we focus on imputation methods for missing dependent variables in three missing conditions: MCAR, MAR and MNAR. The modified logistic regression imputation (MLR) method is proposed, which is developed from the LR method by modifying the cut-off point for logistic regression from 0.5 to an optimal cut-off point for a particular data set. The optimal cut-off point is based on the receiver operating characteristic curve [13]. We then compare the performance of this method with that of six other popular methods, viz. Mode, HD, MI, KNN, RF and LR, using a data set with missing quantities of data equivalent to 10%, 20%, 30% and 40% of the total data set. The modified method is then applied to a real-life heart disease data set.

## IMPUTATION METHODS

### Mode Imputation (Mode)

Mode is one of the easiest and most naive methods of imputing missing values of categorical variables [7] by substituting the mode of complete data for missing data in the same variable.

### Hot Deck Imputation (HD)

HD is a method of handling missing data in which each missing value is replaced with an observed response from a similar unit [14]. The HD method uses a completely observed donor case for the imputation of an incomplete case. The missing value is replaced by the corresponding value of the best donor case, which is found by minimising the distance between the donee and all potential donor cases (typically, Euclidean distance computed in the space of covariates) [15]. This method is popular because it does not rely on model fitting to impute the variable and is thus potentially less sensitive to model misspecification than a parametric model-based imputation method [14].

### Multiple Imputation (MI)

Rubin [16] developed a method of averaging the outcome across multiple imputed data. Thus, in MI instead of replacing each missing observation with a single value, multiple plausible values are inputted to reflect the underlying uncertainty around the imputation. The MI method generates ‘m’ different complete data sets with observed and imputed values. It uses the following three steps.

*Imputation*: As in single imputation, missing values are imputed, but imputed values are generated m times rather than just once. So there could be m different complete data sets after imputation.

*Analysis of each data set*: After imputation and the generation of m different data sets, each of the m data sets is analysed.

*Pooling*: The results obtained from each analysed data set are consolidated [17].

MI by chained equations implemented via the 'mice' package available in R software enables an iterative estimation of missing values in multiple variables and provides flexibility in imputing both categorical and continuous variables [18].

### K-nearest Neighbour Imputation (KNN)

KNN becomes more common in models implemented to forecast missing values. The '*k*' samples are identified from the data set to find the estimated value of the missing data. This necessitates the development of a model for each input variable that has missing values [19].

The KNN method implements the following operations [20].

*Step 1*: Define $k$, where $k=\sqrt{c}$ and $c$ is the amount of complete data.

*Step 2*: Calculate the distance between the missing rows and complete rows using the Euclidean distance method [21]:

$$dist(R_i, R_j)=\sqrt{\sum_{q=1}^{p}(X_{iq}-X_{jq})^2} \quad , \tag{1}$$

where $dist(R_i, R_j)$ denotes the distance between a missing row *i* and a complete row *j*, $X_{iq}$ denotes missing data at row $i$ and column $q$, and $X_{jq}$ denotes complete data at row *j* and column $q$.

*Step 3*: When sorting the distances between missing rows and complete rows, consider the *k* sets with the smallest distances. The imputation estimate is then calculated using the average of smallest distances for continuous variables or the mode for categorical variables.

### Random Forest Imputation (RF)

RF is a method that uses random forest algorithm for imputing missing data. The algorithm is a supervised machine learning algorithm which is an ensemble method combining multiple decision trees, and the final prediction is made by aggregating the predictions of individual trees [22].

The RF method is popular because of its ability to handle both continuous and categorical data and can be used for classification and regression [6, 23]. This method was implemented via the 'missForest' package available in R software [24]. The method employs the following operations [25].

*Step 1*: The missing values are replaced with the mean for continuous variables or the mode for categorical variables.

*Step 2*: The imputation process is done sequentially in ascending order of missing observations for each variable. The variable under imputation is used as the response when building the random forest model, which can be a regression or classification model. The observations in the data set are divided into two parts according to whether the variable is observed or missing in the original data set. Observed observations are used as the training set and missing observations are used as the prediction set. The missing part of the variable under imputation is replaced by the prediction from the generated random forest model for that variable.

*Step 3*: After imputing the missing values, the imputation process is iterated until the proportion of falsely classified entries (PFC) for categorical variables between the current and previous imputation results increases. The PFC equation is given by [26]

$$PFC = \frac{\sum_{i=1}^{s} count(y_i \neq \hat{y}_i)}{s} \tag{2}$$

where $s$ represents the number of missing values, $y_i$ denotes the missing $y$ which has been replaced with the mode, $\hat{y}_i$ denotes the estimate $y$ from the random forest model, and therefore $count\left(y_i \neq \hat{y}_i\right)$ denotes the count of incorrectly classified entries.

The PFC ranges from 0 to 1, and the smaller the value, the better the imputation. The RF method outputs the previous imputation as the final result [23]. A user setting for a maximum number of iterations is installed in the process, with a default value of 10 to limit the computational time to a reasonable level.

## Logistic Regression Imputation (LR)

LR is a type of regression imputation. The regression equation derived from the observed data is used to determine the missing data that must be imputed. The LR method is used to impute the binary variable directly [9]. The LR method is applied as follows.

*Step 1*: Only the data set of a dependent variable with no missing data is used to estimate logistic regression coefficients with the maximum likelihood estimation.

*Step 2*: After obtaining the regression coefficient estimates, missing values from the dependent variable are predicted. The logistic regression model is performed [9]:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2)} \quad , \tag{3}$$

where $\hat{\pi}_i$ denotes the estimated probability of events of interest, and $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\beta}_2$ denote the estimated regression coefficient values.

*Step 3*: Probabilities range from 0 to 1. A cut-off point of 0.5 is often used to indicate events [27]. The chosen prediction rule is applied to the estimated probabilities as a predicted value of 1 if $\hat{\pi}_i \geq 0.5$, and 0 if $\hat{\pi}_i < 0.5$ [28].

*Step 4*: After obtaining the prediction value, it is imputed to the dependent variable.

## Modified Logistic Regression Imputation (MLR)

MLR is a method developed from the LR method by modifying the cut-off point from 0.5 to an optimal cut-off point for a particular data set. The optimal cut-off point is based on the receiver operating characteristic curve, which is a graph of the true positive rate (sensitivity) versus the false positive rate (1-specificity). The optimal cut-off point value is defined as the value whose sensitivity and specificity are closest to the value of the area under the receiver operating characteristic curve, and where the absolute value of the difference between the sensitivity and specificity values is minimal [13].

One of the commonly used methods for finding the optimal cut-off point is the Youden index method. This method defines the optimal cut-off point as the point maximising the Youden function, which is the difference between sensitivity and specificity over all possible cut-off points $(c)$ [29]. The Youden index $(J)$ equation [30] is given by

$$J = \max_c \{Sensitivity(c) + Specificity(c) - 1\} \ , \tag{4}$$

where $Sensitivity(c)$ denotes the probability of a positive test result, conditioned on the sample truly being positive, and $Specificity(c)$ denotes the probability of a negative test result, conditioned on the sample truly being negative.

## PERFORMANCE EVALUATION

The objective of imputation is to obtain statistically valid inferences from incomplete data. Consequently, the assessment of the quality of an imputation method should be conducted with respect to this purpose. Several assessments exist for comparing the efficacy of imputation techniques. This study employs the estimated mean square error (EMSE) of the regression coefficients to quantify the mean square difference between the estimated regression coefficients of the entire data and those of the imputed data. The EMSE equation is given by [31]

$$EMSE(\tilde{\beta}) = \frac{1}{1{,}000}\sum_{t=1}^{1{,}000}\sum_{b=0}^{2}(\hat{\beta}_{(b,t)} - \hat{\beta}^{*}_{(b,t)})^2 \ , \tag{5}$$

where $\hat{\beta}_{(b,t)}$ denotes the estimated regression coefficients of the complete sample data in the $t^{th}$ iteration, and $\hat{\beta}^{*}_{(b,t)}$ denotes the estimated regression coefficients of the imputed data in the $t^{th}$ iteration. The imputation method with the lowest EMSE is considered the best performance method.

## SIMULATION STUDY

This section contains the results of a simulation study that compare the performance of methods of imputing missing dependent variables in binary logistic regression models. First, two independent variables were drawn from standard normal distributions $X_1 \square N(0,1)$ and $X_2 \square N(0,1)$. Then a dependent variable was constructed based on the Bernoulli distribution $Y_i \square Ber(\pi_i)$ where

$$\pi_i = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2}} \tag{6}$$

with regression coefficients $\beta_0, \beta_1, \beta_2 = 1$. Next, a sample was selected by simple random sampling, giving sample sizes of 20, 50, 100, 150, 200, 500 and 1000. MCAR, MAR and MNAR missing conditions were imposed on the dependent variable, with percentages of missing data of 10%, 20%, 30% and 40%. To complete the data set, the missing data from the dependent variable were imputed using seven imputation methods. New binary logistic regression coefficients were then generated with the maximum likelihood estimation. The simulation procedure was repeated 1,000 times. After that, the EMSE of each of the seven techniques was obtained. R software was used to run all simulations. The results are shown in Tables 1-3.

Tables 1 and 2 show the EMSEs of the seven methods when the missing conditions for the simulation data set are MCAR and MAR respectively. The results reveal that the MLR method performs better than the others with small sample sizes (n ≤ 50) at most levels of missing data, while the MI method performs better when the sample sizes are large (n > 50) for all levels of missing data. In every situation the MLR method outperforms the LR method.

Table 3 shows the EMSEs of the seven methods when the missing condition for the simulation data set is MNAR. The results show that the LR method performs better than the other methods when the sample sizes are small (n ≤ 50) for all levels of missing data. However, when the sample sizes are large (n > 50), the Mode method produces the best results at small missing data percentages (≤20%), and the LR method performs the best for missing data percentages of 30% and 40%. In every scenario the LR method outperforms the MLR method. Additionally, in all scenarios the EMSE increases with higher percentages of missing data and decreases when sample sizes are bigger.

**Table 1.** EMSEs of seven imputation methods when missing condition assigned to simulation data set is MCAR (Bold EMSE values are lowest values in each case.)

| **n** | **% Missing data** | **Imputation method** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Mode** | **HD** | **MI** | **KNN** | **RF** | **LR** | **MLR** |
| 20 | 10 | 0.5911 | 0.6101 | 0.5024 | 0.457 | 0.4949 | 0.4275 | **0.3665** |
| | 20 | 1.0897 | 1.2101 | 1.3069 | 1.1856 | 1.4505 | 0.9588 | **0.6981** |
| | 30 | 1.8378 | 1.901 | 1.9077 | 2.341 | 2.6995 | 1.8096 | **1.7392** |
| | 40 | 2.4658 | 2.1801 | 2.5308 | 2.7798 | 4.2039 | 2.5269 | **2.1017** |
| 50 | 10 | 0.2041 | 0.2986 | 0.2493 | 0.2317 | 0.2361 | 0.2213 | **0.2021** |
| | 20 | 0.4278 | 0.607 | 0.4574 | 0.4545 | 0.4804 | 0.4716 | **0.3808** |
| | 30 | 0.7703 | 0.946 | **0.7068** | 0.8208 | 0.88 | 0.86 | 0.7595 |
| | 40 | 1.2003 | 1.2932 | **1.1851** | 1.4074 | 1.8442 | 1.8719 | 1.7713 |
| 100 | 10 | 0.0907 | 0.1154 | **0.0534** | 0.0594 | 0.0547 | 0.0593 | 0.0568 |
| | 20 | 0.2483 | 0.2543 | **0.1255** | 0.1842 | 0.1668 | 0.2022 | 0.1891 |
| | 30 | 0.4832 | 0.4494 | **0.2225** | 0.379 | 0.3364 | 0.4659 | 0.4423 |
| | 40 | 0.8189 | 0.6261 | **0.3415** | 0.7002 | 0.5931 | 0.9147 | 0.8598 |
| 150 | 10 | 0.0708 | 0.0872 | **0.0376** | 0.0458 | 0.0412 | 0.049 | 0.0505 |
| | 20 | 0.2137 | 0.2244 | **0.0768** | 0.1429 | 0.1097 | 0.1618 | 0.1542 |
| | 30 | 0.4218 | 0.3904 | **0.129** | 0.2796 | 0.2172 | 0.3725 | 0.3636 |
| | 40 | 0.7516 | 0.5799 | **0.2141** | 0.5904 | 0.4305 | 0.7976 | 0.7421 |
| 200 | 10 | 0.0632 | 0.076 | **0.0253** | 0.0374 | 0.0311 | 0.0412 | 0.0407 |
| | 20 | 0.1969 | 0.1952 | **0.0579** | 0.1304 | 0.0892 | 0.1524 | 0.1471 |
| | 30 | 0.4004 | 0.3544 | **0.101** | 0.2718 | 0.1884 | 0.3483 | 0.3373 |
| | 40 | 0.715 | 0.5325 | **0.1659** | 0.5556 | 0.3631 | 0.756 | 0.7202 |
| 500 | 10 | 0.0479 | 0.0532 | **0.0098** | 0.0255 | 0.0151 | 0.0295 | 0.0289 |
| | 20 | 0.1701 | 0.1661 | **0.0223** | 0.101 | 0.0532 | 0.1249 | 0.1193 |
| | 30 | 0.3732 | 0.3031 | **0.0403** | 0.2534 | 0.1203 | 0.3233 | 0.3061 |
| | 40 | 0.6673 | 0.4653 | **0.0627** | 0.4937 | 0.2142 | 0.6731 | 0.6278 |

**Table 1. (Continued)**

| n | % Missing data | Imputation method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Mode | HD | MI | KNN | RF | LR | MLR |
| 1000 | 10 | 0.0437 | 0.0468 | **0.005** | 0.0237 | 0.0105 | 0.0265 | 0.0255 |
| | 20 | 0.1613 | 0.1541 | **0.0108** | 0.0996 | 0.0389 | 0.116 | 0.1104 |
| | 30 | 0.3548 | 0.2896 | **0.0195** | 0.2421 | 0.0863 | 0.2962 | 0.2799 |
| | 40 | 0.6478 | 0.4565 | **0.0311** | 0.5014 | 0.1706 | 0.6485 | 0.6059 |

**Table 2**. EMSEs of seven imputation methods when missing condition assigned to simulation data set is MAR (Bold EMSE values are lowest values in each case.)

| n | % Missing data | Imputation method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | Mode | HD | MI | KNN | RF | LR | MLR |
| 20 | 10 | 0.4301 | 0.7124 | 0.3918 | 0.3585 | 0.3737 | 0.2969 | **0.256** |
| | 20 | 1.0542 | 1.4064 | 1.2863 | 1.3938 | 1.4202 | 1.0551 | **0.9712** |
| | 30 | 2.1896 | 2.506 | 1.5899 | 1.7556 | 1.918 | 1.4243 | **1.1774** |
| | 40 | 3.4126 | 2.8294 | 2.4708 | 3.0721 | 3.699 | 1.863 | **1.6621** |
| 50 | 10 | 0.1336 | 0.4619 | 0.1247 | 0.0989 | 0.1076 | 0.095 | **0.0831** |
| | 20 | 0.4248 | 0.9198 | 0.2876 | 0.2926 | 0.2845 | 0.2934 | **0.246** |
| | 30 | 0.9642 | 1.2982 | 0.6243 | 0.6465 | 0.6413 | 0.6841 | **0.5283** |
| | 40 | 1.9623 | 1.6311 | 1.0512 | 1.1536 | 1.1559 | 1.3444 | **0.971** |
| 100 | 10 | 0.0528 | 0.1951 | 0.0538 | 0.0513 | 0.0512 | 0.0533 | **0.0481** |
| | 20 | 0.1812 | 0.4574 | **0.1366** | 0.1684 | 0.1479 | 0.1871 | 0.154 |
| | 30 | 0.4921 | 0.7131 | **0.2352** | 0.3376 | 0.3075 | 0.4314 | 0.3332 |
| | 40 | 0.9779 | 1.0158 | **0.3791** | 0.6607 | 0.5612 | 0.8869 | 0.6544 |
| 150 | 10 | 0.0363 | 0.2425 | 0.0346 | 0.0367 | 0.033 | 0.038 | **0.0329** |
| | 20 | 0.1365 | 0.3938 | **0.0838** | 0.1336 | 0.1001 | 0.1483 | 0.1195 |
| | 30 | 0.3239 | 0.666 | **0.1564** | 0.293 | 0.2232 | 0.3889 | 0.2954 |
| | 40 | 0.7651 | 0.9347 | **0.2519** | 0.6051 | 0.4419 | 0.7818 | 0.5855 |
| 200 | 10 | 0.0344 | 0.1492 | **0.0282** | 0.0345 | 0.0288 | 0.0369 | 0.0319 |
| | 20 | 0.1162 | 0.3933 | **0.0629** | 0.1173 | 0.0815 | 0.1324 | 0.1059 |
| | 30 | 0.2951 | 0.6621 | **0.1127** | 0.2656 | 0.1826 | 0.342 | 0.2586 |
| | 40 | 0.6816 | 0.8894 | **0.1948** | 0.6001 | 0.371 | 0.772 | 0.5608 |
| 500 | 10 | 0.0239 | 0.1237 | **0.0098** | 0.0248 | 0.0148 | 0.0265 | 0.0319 |
| | 20 | 0.0995 | 0.3331 | **0.0242** | 0.1011 | 0.0516 | 0.1159 | 0.0925 |
| | 30 | 0.256 | 0.594 | **0.0437** | 0.2758 | 0.1246 | 0.316 | 0.2434 |
| | 40 | 0.5261 | 0.8513 | **0.072** | 0.5594 | 0.2425 | 0.6861 | 0.4989 |
| 1000 | 10 | 0.0209 | 0.1168 | **0.0047** | 0.0225 | 0.0104 | 0.0235 | 0.0193 |
| | 20 | 0.0939 | 0.3254 | **0.0114** | 0.1036 | 0.0394 | 0.1113 | 0.0868 |
| | 30 | 0.2418 | 0.5695 | **0.0204** | 0.268 | 0.0956 | 0.3008 | 0.2237 |
| | 40 | 0.5093 | 0.8245 | **0.035** | 0.5631 | 0.2029 | 0.6716 | 0.4729 |

**Table 3.** EMSEs of seven imputation methods when missing condition assigned to simulation data set is MNAR (Bold EMSE values are lowest values in each case.)

| n | % Missing data | Imputation method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Mode** | **HD** | **MI** | **KNN** | **RF** | **LR** | **MLR** |
| 20 | 10 | 0.3994 | 0.7139 | 0.4798 | 0.5166 | 0.4328 | **0.2206** | 0.3362 |
| | 20 | 1.0047 | 1.4151 | 1.1079 | 1.31 | 1.2588 | **0.6539** | 0.8567 |
| | 30 | 2.0273 | 2.396 | 2.2275 | 2.513 | 2.8411 | **1.4421** | 1.7712 |
| | 40 | 3.9258 | 3.2697 | 3.2232 | 3.5661 | 4.2104 | **1.9326** | 2.9982 |
| 50 | 10 | 0.0841 | 0.3098 | 0.126 | 0.0908 | 0.1161 | **0.0801** | 0.1068 |
| | 20 | 0.3767 | 0.7109 | 0.3467 | 0.3754 | 0.348 | **0.2049** | 0.3192 |
| | 30 | 1.1525 | 1.3832 | 0.7652 | 0.9411 | 0.7569 | **0.5301** | 1.0625 |
| | 40 | 2.8354 | 2.1069 | 1.4433 | 1.7065 | 1.6658 | **1.0871** | 1.6995 |
| 100 | 10 | **0.0206** | 0.1971 | 0.0648 | 0.0498 | 0.0632 | 0.0446 | 0.0682 |
| | 20 | **0.1306** | 0.5444 | 0.184 | 0.1616 | 0.1873 | 0.139 | 0.221 |
| | 30 | 0.7673 | 1.09 | 0.4099 | 0.5434 | 0.451 | **0.3007** | 0.5377 |
| | 40 | 2.7483 | 1.7785 | 0.845 | 1.3566 | 0.9989 | **0.6277** | 1.19 |
| 150 | 10 | **0.0125** | 0.1635 | 0.0459 | 0.0416 | 0.0435 | 0.0322 | 0.0516 |
| | 20 | **0.0802** | 0.5321 | 0.1602 | 0.1531 | 0.1644 | 0.1185 | 0.2124 |
| | 30 | 0.6978 | 1.039 | 0.3707 | 0.5019 | 0.3963 | **0.2654** | 0.5072 |
| | 40 | 2.5076 | 1.7683 | 0.6728 | 0.917 | 0.7969 | **0.5266** | 1.0475 |
| 200 | 10 | **0.008** | 0.155 | 0.0353 | 0.0288 | 0.0359 | 0.0253 | 0.0455 |
| | 20 | **0.0493** | 0.499 | 0.134 | 0.125 | 0.1386 | 0.0961 | 0.1896 |
| | 30 | 0.5171 | 1.0149 | 0.313 | 0.4474 | 0.3618 | **0.2557** | 0.5057 |
| | 40 | 2.6185 | 1.7021 | 0.6168 | 0.8798 | 0.7281 | **0.4933** | 1.0227 |
| 500 | 10 | **0.0033** | 0.0854 | 0.0243 | 0.0136 | 0.0198 | 0.014 | 0.0273 |
| | 20 | **0.0124** | 0.339 | 0.0997 | 0.0848 | 0.0887 | 0.0585 | 0.1193 |
| | 30 | **0.0776** | 0.7386 | 0.2563 | 0.2367 | 0.2395 | 0.1449 | 0.3082 |
| | 40 | 2.1623 | 1.2611 | 0.5197 | 0.6274 | 0.5036 | **0.2955** | 0.6412 |
| 1000 | 10 | **0.0025** | 0.0819 | 0.02 | 0.0126 | 0.0164 | 0.0114 | 0.0247 |
| | 20 | **0.0105** | 0.3216 | 0.0934 | 0.0756 | 0.0784 | 0.0533 | 0.1175 |
| | 30 | **0.0305** | 0.7123 | 0.2422 | 0.2536 | 0.22 | 0.1364 | 0.307 |
| | 40 | 2.0713 | 1.2355 | 0.4879 | 0.6703 | 0.4579 | **0.2753** | 0.6318 |

## APPLICATION TO REAL-LIFE DATA

The actual data set used in the study was a set of heart disease predictions from the online database www.kaggle.com. [32]. This data set includes 1,025 observations and one dependent variable $(y)$, heart disease. A value of 0 represents not having heart disease and a value of 1 represents having heart disease. Two independent variables were used for this study: blood pressure $(x_1)$ and cholesterol level $(x_2)$. Next, samples were selected by simple random sampling, giving sample sizes of 50 and 500, referred to as small and large sample sizes. The MCAR, MAR and MNAR missing conditions were assigned to the dependent variable. The missing data percentages were 10%, 20%, 30% and 40%, which were then estimated using the seven imputation methods. The parameters were estimated by the maximum likelihood estimation method and the efficiency of each imputation method was defined by the EMSE. The results are shown in Tables 4-6.

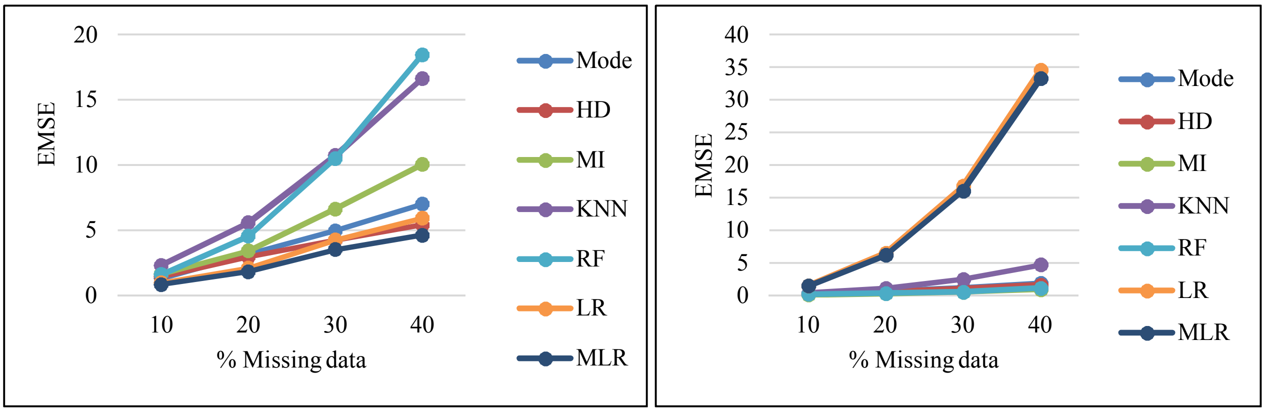
Tables 4 and 5 and Figures 1 and 2 show the EMSEs of the seven methods when the missing conditions are MCAR and MAR. The results reveal that the MLR method performs best when the sample size is 50 while the MI method performs best when the sample size is 500. In every situation, the MLR method outperforms the LR method.

Table 6 and Figure 3 show the EMSEs of the seven methods when the missing condition is MNAR. The results reveal that the MLR method performs best when the sample size is 50 while the Mode method gives the best results when the sample size is 500. Additionally, although the MLR method outperforms the LR method when $n = 50$, the LR method outperforms the MLR method when $n = 500$. These results are consistent with those obtained in the simulation study, especially when the missing conditions are MCAR and MAR.

**Table 4.** EMSEs of seven imputation methods when MCAR missing condition is assigned to heart disease prediction data set

| n | % Missing data | Imputation method | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Mode** | **HD** | **MI** | **KNN** | **RF** | **LR** | **MLR** |
| 50 | 10 | 1.2985 | 1.4545 | 1.6443 | 2.3165 | 1.6019 | 0.9565 | **0.8521** |
| | 20 | 3.1659 | 2.9561 | 3.4293 | 5.5947 | 4.5412 | 2.1009 | **1.8231** |
| | 30 | 4.9746 | 4.2238 | 6.6518 | 10.735 | 10.5311 | 4.2487 | **3.5256** |
| | 40 | 7.0184 | 5.4066 | 10.0569 | 16.6316 | 18.4505 | 5.9324 | **4.6212** |
| 500 | 10 | 0.2402 | 0.2092 | **0.086** | 0.2917 | 0.1475 | 1.501 | 1.4183 |
| | 20 | 0.6162 | 0.5436 | **0.2529** | 1.0377 | 0.3059 | 6.488 | 6.1542 |
| | 30 | 1.0724 | 0.9855 | **0.4875** | 2.4703 | 0.492 | 16.7636 | 16.0368 |
| | 40 | 1.8198 | 1.5662 | **0.8787** | 4.6676 | 1.0745 | 34.563 | 33.3094 |

Note: Bold EMSE values are lowest values in each case.



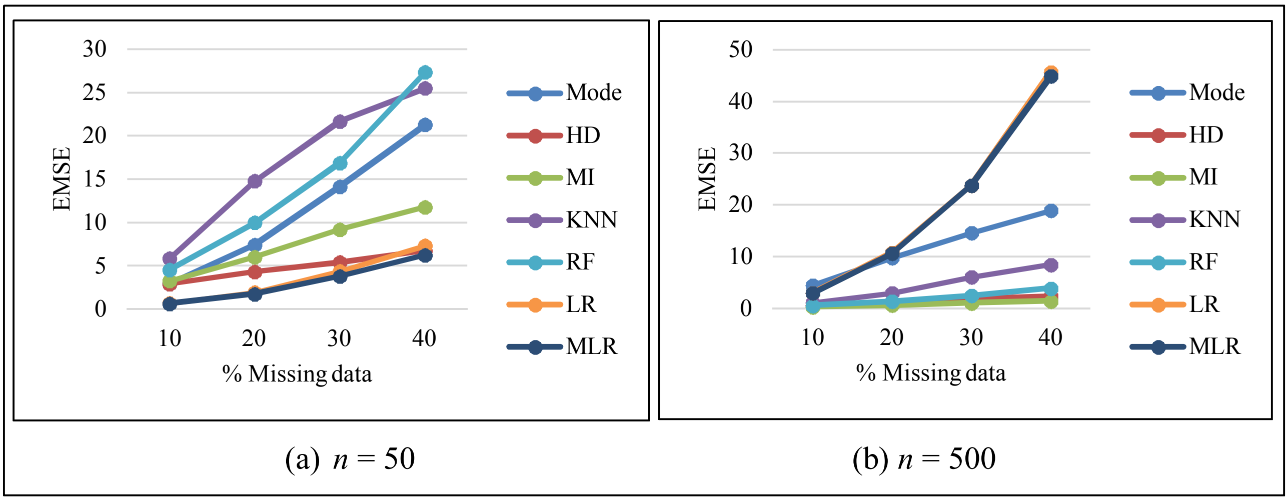
(a) $n = 50$ (b) $n = 500$

**Figure 1.** EMSEs of seven imputation methods when MCAR missing condition is assigned to heart disease prediction data set

**Table 5.** EMSEs of seven imputation methods when MAR missing condition is assigned to heart disease prediction data set

| **n** | **% Missing data** | **Imputation method** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Mode** | **HD** | **MI** | **KNN** | **RF** | **LR** | **MLR** |
| 50 | 10 | 2.9291 | 2.9227 | 3.2495 | 5.8502 | 4.55 | 0.6868 | **0.6363** |
| | 20 | 7.4511 | 4.3272 | 6.0424 | 14.7784 | 9.9751 | 1.929 | **1.7608** |
| | 30 | 14.1573 | 5.4033 | 9.2139 | 21.6973 | 16.9056 | 4.3768 | **3.8069** |
| | 40 | 21.2927 | 6.7895 | 11.7816 | 25.5103 | 27.3465 | 7.3129 | **6.2516** |
| 500 | 10 | 4.3728 | 0.5855 | **0.275** | 1.0421 | 0.4702 | 2.9816 | 2.9114 |
| | 20 | 9.7376 | 1.2011 | **0.5511** | 2.9072 | 1.2992 | 10.7696 | 10.5844 |
| | 30 | 14.5897 | 1.8053 | **0.985** | 6.0274 | 2.4289 | 23.8251 | 23.7744 |
| | 40 | 18.9135 | 2.2727 | **1.3366** | 8.4251 | 3.8492 | 45.7501 | 44.9565 |

Note: Bold EMSE values are lowest values in each case.



(a) *n* = 50

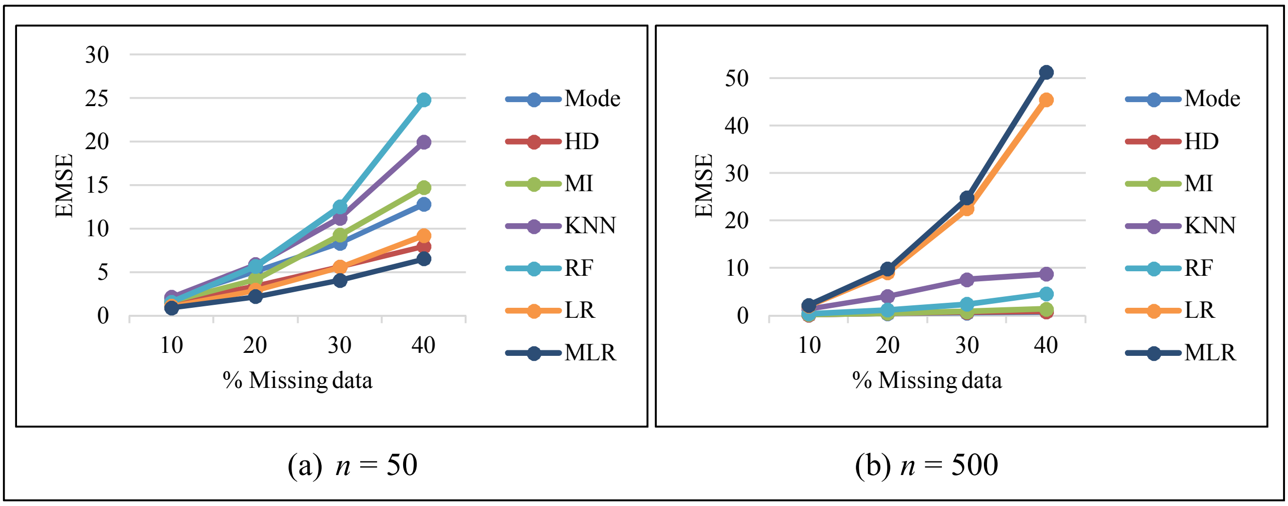
(b) *n* = 500

**Figure 2.** EMSEs of seven imputation methods when MAR missing condition is assigned to heart disease prediction data set

**Table 6.** EMSEs of seven imputation methods when MNAR missing condition is assigned to heart disease prediction data set

| **n** | **% Missing data** | **Imputation method** | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | **Mode** | **HD** | **MI** | **KNN** | **RF** | **LR** | **MLR** |
| 50 | 10 | 2.003 | 1.5238 | 1.6362 | 2.1273 | 1.5676 | 1.073 | **0.8829** |
| | 20 | 5.1217 | 3.426 | 4.1891 | 5.8669 | 5.6942 | 2.9301 | **2.1978** |
| | 30 | 8.3482 | 5.6179 | 9.2561 | 11.2109 | 12.4961 | 5.5865 | **4.0799** |
| | 40 | 12.8224 | 7.9778 | 14.7367 | 19.9994 | 24.8089 | 9.2117 | **6.5458** |
| 500 | 10 | **0.172** | 0.1844 | 0.2519 | 1.3387 | 0.3395 | 2.0271 | 2.1636 |
| | 20 | **0.4177** | 0.4796 | 0.5116 | 4.0507 | 1.1619 | 9.0693 | 9.7956 |
| | 30 | **0.5698** | 0.6684 | 0.8527 | 7.6202 | 2.4013 | 22.5309 | 24.8214 |
| | 40 | **0.7405** | 0.8216 | 1.3661 | 8.7293 | 4.597 | 45.55 | 51.3829 |

Note: Bold EMSE values are lowest values in each case.



(a) $n = 50$ (b) $n = 500$

**Figure 3.** EMSEs of seven imputation methods when MNAR missing condition is assigned to heart disease prediction data set

## DISCUSSION AND CONCLUSIONS

The main aim of this study was to develop and compare the performance of seven different missing data imputation methods applied to the dependent variable of the binary logistic regression model when the missing data were of three different types: MCAR, MAR and MNAR. Logistic regression analysis was applied due to the binary nature of the missing data variable. The MLR method exhibited superior performance in cases where the sample size was small $(n \leq 50)$. When the sample size was large $(n \geq 100)$, the MI method performed best because it uses multiple substitution values and summarises the results obtained from the estimation to get the optimal value, giving reliable and highly accurate results [18]. In the majority of cases, the MLR method outperformed the LR method. For MLR, an optimal cut-off point can be determined so that it is specific to the data set, whereas the LR method relies on a universally accepted cut-off point of 0.5.

Moreover, the outcomes of the simulated study and real-life data were consistent, especially when the missing data condition was either MCAR or MAR. The outcomes for MNAR were different because its condition made parameter estimates more biased than the MCAR and MAR conditions [3]. As the sample size increased, the EMSE values tended to decrease because estimations are more accurate when based on a larger sample size. When the percentage of missing data was higher, the EMSE value tended to increase because the sample size was smaller, making the parameter estimates more biased.

Overall, it is evident that each imputation approach has advantages and disadvantages. The Mode and HD methods are simple techniques for filling in missing data. They are especially well-suited for categorical variables but may not exhibit the same level of efficiency as other methods. The KNN method is a versatile approach that takes into account relationships among variables but is susceptible to outliers because it heavily depends on distance measures. The RF method is well-suited for data sets from large sample sizes that exhibit complex patterns but it is not appropriate for data sets of small sample sizes that do not provide enough data to construct an imputation model. Inaccurate results are produced. The MI approach produces multiple imputed data sets, allowing accurate estimation of parameter values but it is not suitable for small sample sizes or when dealing with the MNAR missing data condition. The LR method is perfectly appropriate for filling in missing data in binary or categorical variables, especially when the missing condition is MNAR with small sample sizes. However, it becomes computationally intricate when dealing with large sample sizes. The MLR technique, derived from the LR method, exhibits superior performance and is particularly well-suited to limited sample sizes when the missing data conditions are MCAR or MAR.

## ACKNOWLEDGEMENTS

We would like to thank the reviewers for their comments and suggestions about the manuscript. This research was supported by the Faculty of Science Research Fund, Prince of Songkla University (Contract No. 1-2565-02-006).

## REFERENCES

1. B. Grofman and C. Q. Schneider, "An introduction to crisp set QCA, with a comparison to binary logistic regression", *Polit. Res. Quart.*, **2009**, *62*, 662-672.
2. J. K. Harris, "Primer on binary logistic regression", *Fam. Med. Commun. Health*, **2021**, *9*, Art.no.e001290.
3. R. J. A. Little and D. B. Rubin, "Statistical Analysis with Missing Data", Wiley, Hoboken, **2002**, pp.11-19.
4. J. L. Schafer and J. W. Graham, "Missing data: Our view of the state of the art", *Psychol. Meth.*, **2002**, *7*, 147-177.
5. C. Y. J. Peng and J. Zhu, "Comparison of two approaches for handling missing covariates in logistic regression", *Educ. Psychol. Meas.*, **2008**, *68*, 58-77.
6. A. K. Waljee, A. Mukherjee, A. G. Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu and P. D. Higgins, "Comparison of imputation methods for missing Laboratory data in medicine", *BMJ Open*, **2013**, *3*, Art.no.e002847.
7. X. Xu, L. Xia, Q. Zhang, S. Wu, M. Wu and H. Liu, "The ability of different imputation methods for missing values in mental measurement questionnaires", *BMC Med. Res. Methodol.*, **2020**, *20*, Art.no.42.
8. T. Tsiampalis and D. B. Panagiotakos, "Missing-data analysis: Socio- demographic, clinical and lifestyle determinants of low response rate on self- reported psychological and nutrition related multi-item instruments in the context of the ATTICA epidemiological study", *BMC Med. Res. Methodol.*, **2020**, *20*, Art.no.148.
9. S. M. Mohamed, M. R. Abonazel and M. G. Ghallab, "A review of ten imputation methods for handling missing values in logistic regression", *Sci. Forum (J. Pure Appl. Sci.)*, **2021**, *21*, 440-451.
10. J. Ma, N. Akhtar-Danesh, L. Dolovich, L. Thabane and CHAT investigators, "Imputation strategies for missing binary outcomes in cluster randomized trials", *BMC Med. Res. Methodol.*, **2011**, *11*, Art.no.18.
11. T. R. Sullivan, K. J. Lee, P. Ryan and A. B. Salter, "Multiple imputation for handling missing outcome data when estimating the relative risk", *BMC Med. Res. Methodol.*, **2017**, *17*, Art.no.134.
12. S. M. Mohamed, M. R. Abonazel and M. G. Ghallab, "Performance evaluation of imputation methods for missing data in logistic regression models: Simulation and application", *Thailand Statist.*, **2023**, *21*, 926-942.
13. I. Unal, "Defining an optimal cut-point value in ROC analysis: An alternative approach", *Comput. Math. Meth. Med.*, **2017**, *2017*, Art.no.3762651.

14. R. R. Andridge and R. J. A. Little, "A review of hot deck imputation for survey non-response", *Int. Stat. Rev.*, **2010**, *78*, 40-64.
15. H. Peyre, A. Leplège and J. Coste, "Missing data methods for dealing with missing items in quality of life questionnaires. A comparison by simulation of personal mean score, full information maximum likelihood, multiple imputation, and hot deck techniques applied to the SF-36 in the French 2003 decennial health survey", *Qual. Life Res.*, **2011**, *20*, 287-300.
16. D. B. Rubin, "Statistical matching using file concatenation with adjusted weights and multiple imputations", *J. Bus. Econ. Stat.*, **1986**, *4*, 87-94.
17. A. Jadhav, D. Pramod and K. Ramanathan, "Comparison of performance of data imputation methods for numeric data set", *Appl. Artif. Intell.*, **2019**, *33*, 913-933.
18. M. J. Azur, E. A. Stuart, C. Frangakis and P. J. Leaf, "Multiple imputation by chained equations: What is it and how does it work?", *Int. J. Meth. Psychiatr. Res.*, **2011**, *20*, 40-49.
19. R. Atiq, F. Fariha, M. Mahmud, S. S. Yeamin, K. I. Rushee and S. Rahim, "A comparison of missing value imputation techniques on coupon acceptance prediction", *Int. J. Inform. Technol. Comput. Sci.*, **2022**, *5*, 15-25.
20. T. Thongsri and K. Samart, "Development of imputation methods for missing data in multiple linear regression analysis", *Lobach. J. Math.*, **2022**, *43*, 3390-3399.
21. M. Bazmara and S. Jafari, "K nearest neighbor algorithm for finding soccer talent", *J. Basic Appl. Sci. Res.*, **2013**, *3*, 981-986.
22. L. Breiman, "Random forests", *Mach. Learn.*, **2001**, *45*, 5-32.
23. A. Liaw and M. Wiener, "Classification and regression by randomForest", *R News*, **2002**, *2*, 18-22.
24. D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data", *Bioinformatics*, **2012**, *28*, 112-118.
25. S. Hong and H. S. Lynn, "Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction", *BMC Med. Res. Methodol.*, **2020**, *20*, Art.no.199.
26. C. Y. Guo, Y. C. Yang and Y. H. Chen, "The optimal machine learning-based missing data imputation for the cox proportional hazard model", *Front. Public Health*, **2021**, *9*, Art.no.680054.
27. E. Y. Boateng and D. A. Abaye, "A review of the logistic regression model with emphasis on medical research", *J. Data Anal. Inform. Process.*, **2019**, *7*, 190-207.
28. Q. M. Abdulqader, "Applying the binary logistic regression analysis on the medical data", *Sci. J. Univ. Zakho*, **2017**, *5*, 330-334.
29. N. J. Perkins and E. F. Schisterman, "The youden index and the optimal cut-point corrected for measurement error", *Biom. J.*, **2005**, *47*, 428-441.
30. W. J. Youden, "Index for rating diagnostic tests", *Cancer*, **1950**, *3*, 32-35.
31. M. R. Özkale and E. Arıcan, "First-order r − d class estimator in binary logistic regression model", *Stat. Prob. Lett.*, **2015**, *106*, 19-29.
32. D. Geb, "Heartdisease predictions", **2023**, https://www.kaggle.com/code/desalegngeb/heart-disease-predictions/input (Accessed: February 2023).