Thailand Statistician
October 2023; 21(4): 926-942
http://statassoc.or.th
Contributed paper

Performance Evaluation of Imputation Methods for Missing Data in Logistic Regression Model: Simulation and Application

Salah M. Mohamed, Mohamed R. Abonazel*, Mohamed G. Ghallab
Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research (FSSR), Cairo University, Giza, Egypt.
*Corresponding author; e-mail: mabonazel@cu.edu.eg

Received: 25 January 2021
Revised: 20 February 2022
Accepted: 4 April 2022

Abstract

Missing data is a common phenomenon most analysts have experienced. Even if the dataset includes a significant number of data points, many of the variables of interest will have missing values. The most prevalent method for dealing with such data points is to leave them out of the analysis. This method is not ideal for multiple reasons. One is that unless the data are missing completely at random, leaving out data points with missing values will bias the results of analysis. A second is that it leads to smaller datasets used for analysis. In this paper, we discuss some commonly used imputation methods, such as Expectation-Maximization (EM), multiple imputation by chained equations, and K-nearest neighbor. Furthermore, we propose a new imputation (EPK) method. The Monte Carlo simulation study is conducted to examine the efficiency of nine imputation methods in the binary logistic regression model when the missingness mechanism is missing at random. Moreover, we used a real data on social network advertising, as an empirical study, to examine these methods. The results of our simulation and empirical studies indicated that the EPK and EM methods are more efficient than other imputation methods; where the EPK and EM have smallest values of Akaike information criterion (AIC) and Bayesian information criterion (BIC), whether the missing data is in the independent variables only, the dependent variable only, or in both together.

Keywords: Expectation maximization, imputation by random forests, K- nearest neighbor, multivariate imputation by chained equations, predictive mean matching.

1. Introduction

The logistic regression (LR) is a tool for building models when there is a categorical response variable with two levels. The LR is a type of generalized linear model (GLM) for response variables where regular multiple regression does not work very well. The LR sometimes called the logistic model or logit model analyzes the relationship between multiple independent variables and a categorical dependent variable and estimates the probability of occurrence of an event by fitting data to a logistic curve. There are two models of LR, binary logistic regression (BLR), and multinomial logistic regression (MLR). BLR is typically used when the dependent variable is dichotomous, and the independent variables are either continuous or categorical. When the dependent variable is not dichotomous and is comprised of more than two categories, an MLR can be employed. Many variables of interest are dichotomous, e.g., whether someone voted in the last election, whether someone is a smoker, whether one has a child, whether one is unemployed, etc. These types of variables are often referred to as discrete or qualitative. The goal of the binary logistic analysis is to find the best fitting and most parsimonious, yet reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables (Tranmer and Elliot 2008).

There are various imputation methods for dealing with missing data. According to these methods, the missing values are replaced by estimates obtained from statistical procedures. The problem is that most of the imputation methods produce in general continuous estimates, which are not realistic replacements of the missing values when the variables are categorical. The aim of this paper is to propose a new imputation method, and study the performance of these methods in the BLR model when the missingness mechanism is missing at random by conducting a Monte Carlo simulation study and real data application. In our study, we allow the missing data to be in all the variables, whether in the independent variables only, in the dependent variable only, or in both together.

The paper is organized as follows. Section 2 introduces logistic regression (transformations, model, assumptions, and the maximum likelihood estimator). Section 3 discusses the main methods to deal with missing values (expectation-maximization, K-nearest neighbor, multivariate imputation by chained equations). Section 4 presents the Monte Carlo simulation study. In Section 5, an empirical study has been presented for assessing the performance of different estimation methods under the existence of missing data. Finally, Section 6 offers concluding remarks.

2. The Logistic Regression Model

The logistic (logit) function is a common transformation for linearizing sigmoid distributions of proportions (El-Masry et al. 2021). A sigmoid function is a mathematical function having a characteristic "S"-shaped curve or sigmoid curve. The sigmoid function is used to convert the input into ranges 0 and 1 (Armitage and Berry 1994). The logit function asymptotically approaches 0 as the input approaches negative infinity and 1 as the input approaches positive infinity. Since the results are bounded by 0 and1, it can be directly interpreted as a probability. To achieve this, a regression is first performed with a transformed value of Y, called the "logit function".

$$\log(odds) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{1}$$

where "odds" refers to the odds of Y being equal to 1. In other words, "odds" is defined as the probability of belonging to one group divided by the probability of belonging to the other: $odds = \left(\frac{\pi}{1-\pi}\right)$. The odds are always positive: $odds = \left(\frac{\pi}{1\text{-}\pi}\right) \to [0, \infty[$, this means that the values of odds are always positive. But the $\log(odds)$ is continuous: $\log(odds) = \log\left(\frac{\pi}{1-\pi}\right) \to (-\infty, \infty)$. The LR will model the chance of an outcome based on individual characteristics. Because chance is a ratio, what will be modeled is the logarithm of the chance given by:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k, \tag{2}$$

where π is the probability of an event, $\beta_0, \beta_1, \cdots, \beta_k$ are the regression coefficients associated with the reference group, and $x_1, \ldots, x_k$ are the explanatory variables.

2.1. Logistic regression assumptions

No matter how one goes about selecting the explanatory or independent variables, basic assumptions for conducting LR must always be met. One assumption is the independence of errors, whereby all sample group outcomes are separate from each other. A second assumption is a linearity in the logit for any continuous independent variables. A third assumption is the absence of exact multicollinearity between the independent variables. A final assumption is the lack of strongly influential outliers (Stoltzfus 2011).

2.2. Maximum likelihood estimation

In general, the maximum likelihood estimation (MLE) method yields estimates for the unknown parameters that maximize the probability of obtaining the observed set of data. To apply this method, we must first construct a function, called the likelihood function. This function expresses the probability of the observed data as a function of the unknown parameters. The MLE of the parameters is the values that maximize likelihood function. Thus, the resulting estimators are those that agree most closely with the observed data. A convenient way to express the contribution to the likelihood function for the pair $\left(x_i, y_i\right)$ is through the expression

$$\pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}. \tag{3}$$

As the observations are assumed to be independent, the likelihood function is obtained as the product of the terms given in Equation (3) as follows

$$L(\beta)=\prod_{i=1}^{n}\pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}. \tag{4}$$

To estimate of β that maximizes $L(\beta)$ we differentiate $L(\beta)$ with respect to β_0 and β_1 and set the resulting expressions equal to zero. These equations, known as the likelihood equations (Hosmer et al. 2013).

Although you will probably use a statistical package to compute the estimates, here is a brief description of the underlying procedure. Because LR predicts probabilities, rather than just classes, one can fit it using likelihood. Usually, we will distinguish the log-likelihood concerning the conditions, set the derivatives to zero, and solve to find the MLE. Because this equation is nonlinear in β, certain special methods for obtaining the approximate parameters should be employed. The iterative re-weighted least squares (IRLS) method can be applied to get the solutions, which can be found in Hilbe (2011). The MLE estimator of β can be obtained by using the IRLS algorithm as follows

$$\hat{\beta}_{MLE}=\left(X'\hat{W}X\right)^{-1}\left(X'\hat{W}\hat{Z}\right). \tag{5}$$

where $\hat{W}=diag\left(\hat{\pi}_i(1-\hat{\pi}_i)\right)$ and $\hat{Z}=\left(\hat{z}_1,\ldots,\hat{z}_n\right);\hat{z}_i=\log(\hat{\pi}_i)+\dfrac{y_i-\hat{\pi}_i}{\hat{\pi}_i(1-\hat{\pi}_i)};i=1,\ldots,n.$ The hats in the equations show the iterative process.

3. Missing Data

Missing data typically refers to the absence of one or more values within a study variable(s) contained in a dataset. The development is often the result of a study participant choosing not to provide a response to a survey item. Missing data plagues almost all surveys, and quite a several designed experiments. No matter how carefully an investigator tries to have all questions fully responded to in a survey, or how well designed an experiment is; examples of how this can occur are when a question is unanswered in a survey, or a flood has removed a crop planted close to a river. The problem is, how to deal with missing data, once it has been deemed impossible to recover the actual missing values (Scheffer 2002). Because of these problems, methodologists routinely advise researchers to design studies to minimize the occurrence of missing values. Appearing in the literature from 1985 to 1989, Concato et al. (1993) reported that the LR was the most frequently used procedure comprising an average of 43% of the multivariate methods in the five-year period reviewed. Two reports (Khan et al. 1999) described a significant increase in the use of the LR in the public health, epidemiology, obstetrics, and gynecology research literature. Bender (2009) reviewed the statistical methods reported in a probability sample of 348 articles published between 1970 and 1998 in the American Journal of Public Health and the American Journal of Epidemiology. The study revealed significant increases in the use of LR.

3.1. Missing data mechanisms

Rubin (1976), Little and Rubin (1987), Little (1992), Little and Schenker (1995), and Little and Rubin (2002) established the foundations of missing data theory. Central to missing data theory is his classification of missing data problems into three categories:

- Missing completely at random (MCAR): means that there is no relationship between the missingness of the data and any values, observed or missing.
- Missing at random (MAR): means that there is a systematic relationship between the propensity of missing values and the observed data, but not the missing data.
- Missing not at random (MNAR): means that there is a relationship between the propensity of a value to be missing and its values.

These three classes of missing data are referred to as missing data mechanisms (for a slightly different classification; see Gelman and Hill 2007, Nakagawa 2015). Missing data mechanisms represent the statistical relationship between observations (or the variables) and the probability of missing data.

3.2. Statistical methods for handling the missing data

The concept of missing values is important to understand to successfully manage data. If the missing values are not handled properly by the researcher, then may end up drawing an inaccurate inference about the data, due to improper handling. The researcher may leave the data or do data imputation to replace them. Suppose the number of cases of missing values is extremely small; then, an expert researcher may drop or omit those values from the analysis. In statistical language, if the number of cases is less than 5% of the sample, then the researcher can drop them. The best possible method of handling the missing data is to prevent the problem by well-planning the study and collecting the data carefully; see Tsikriktsis (2005), and Wisniewski et al. (2006). In the next section we will introduce some of the methods used in our paper, you can also find more details about those methods through the following references (e.g., Sentas and Angelis 2006, Peng and Zhu 2008, Meeyai 2016).

Missing data reduces the representativeness of the sample and can therefore distort inferences about the population. There are three main approaches to handle missing data: (1) Omission "deletion": where samples with invalid data are discarded from further analysis and (2) imputation: where values are filled in the place of missing data, (3) analysis: by directly applying methods unaffected by the missing values. In the next lines, we will explain K-nearest neighbor, expectation maximization, and other multivariate important methods.

3.2.1. K-nearest neighbor

It is good practice to identify and replace missing values for each column in your input data prior to modeling your prediction task. This is called missing data imputation or imputing for short. A popular approach to missing data imputation is to use a model to predict the missing values. This requires a model to be created for each input variable that has missing values. Although any one among a range of different models can be used to predict the missing values, the K-nearest neighbor (KNN) algorithm has proven to be generally effective, often referred to as "nearest neighbor imputation (Pan et al. 2015). The KNN algorithm is one of the top ten data mining algorithms and known as an instance-based learning method (Wu et al. 2008).

This method does not perform well when there are large amounts of missing data. To date, there is no theoretical result for selecting the optimal K-value. The most frequent value among K-nearest neighbor and the mean among the KNN can be predicted by KNN imputation. In this method, the main factor is the distance metrics. In the KNN imputation method, we can replace the missing values with the nearest neighbor. But if the value of K is greater than one then replaces the missing values with the mean or weighted average of KNN. By setting K-value between 10 and 20 brings the best results for KNN imputation.

3.2.2. Expectation maximization algorithm

The expectation-maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates, useful in a variety of incomplete data problems, where algorithms such as the Newton-Raphson method may turn out to be more complicated. On each iteration of the EM algorithm, there are two steps-called the expectation step or the E-step and the maximization step or the M-step. Because of this, the algorithm is called the EM algorithm. This name was given by Dempster et al. (1977) in their fundamental paper. The essential idea behind the EM algorithm is to calculate the maximum likelihood estimates for the incomplete data problem by using the complete data likelihood instead of the observed likelihood because the observed likelihood might be complicated or numerically infeasible to maximize. To do this, we augment the observed data with manufactured data to create a complete likelihood that is computationally more tractable. We then replace, at each iteration, the incomplete data, which are in the sufficient statistics for the parameters in the complete data likelihood, by their conditional expectation given the observed data and the current parameter estimates (Expectation step: E-step) The new parameter estimates are obtained from these replaced sufficient statistics as though they had come from the complete sample (Maximization step: M-step) alternating E- and M-steps, the expected log-likelihood is maximized to produce new values of the parameters. The EM algorithm is closely related to the following ad hoc process of handling missing data:

Step 1: Fill in the missing values by their estimated values.

Step 2: Estimate the parameters for this completed dataset.

Step 3: Use the estimated parameters to re-estimate the missing values.

Step 4: Re-estimate the parameters from this updated completed dataset.

We alternate between steps 3 and 4 until convergence of parameter estimates is achieved.

3.2.3. Multiple imputation by chained equations

Just as there are multiple methods of single imputation, there are multiple methods of multiple imputation (MI) as well. One advantage that MI has over the single imputation and complete case methods is that MI is flexible and can be used in a wide variety of scenarios. MI can be used in cases where the data is MCAR, MAR, and even when the data is MNAR. However, the primary method of

MI is multiple imputation by chained equations (MICE). It is also known as "fully conditional specification" (Azur et al., 2011). The MICE has been shown to work very well on MAR. MI is not very difficult to implement. There are a wide range of different statistical packages in different statistical software that readily allow someone to perform MI. For example, the "mice" R-package allows users in R to perform MI using the MICE method.

The MICE algorithm, designed by Buuren and Groothuis-Oudshoorn (2010), is based on a Markov chain Monte Carlo method wherein the state space is the collection of all imputed values. The merit of MICE is that the results are computed after a comparatively few iterations. As per some studies (Buuren and Groothuis-Oudshoorn 2010), in general, five iterations are usually sufficient. In MICE, the user can specify an elementary imputation method for each incomplete data column. The elementary imputation method takes a set of (at that moment) complete predictors and returns a single imputation for each missing entry in the incomplete target column. Multiple imputations are created by repeated calls to the function. The "mice" R-package supplies several built-in elementary imputation models, which are given in Table 1. For more details about this package, see Azur et al. (2011).

Table 1 Some built-in methods in "mice" R-package

Method	Description	Type of target
Mean	Unconditional mean imputation	Numeric
RI	Regression imputation	Numeric
PMM	Predictive mean matching	Any type
LRI	Logistic regression	2 categories
RF	Imputation by random forests	Any type

4. Simulation Study

Monte Carlo simulation is a method of repeating a statistical analysis using multiple iterations, with a subset of data sampled, changed, or deleted at random with each iteration as described in Abonazel (2018) and Abonazel (2020). To explore the performance of imputation methods under different missing data percentages, different sample sizes, and different numbers of independent variables, we used the complete training dataset, to artificially generate missing values by imitating the MAR mechanism. The programming of the simulation study is written in R-software, see Abonazel (2018) and Abonazel (2020).

4.1. Data model

The explanatory variables (X's) are simulated from a multivariate normal distribution $MVN(1, I_k)$. The response variable (y) of the logistic model is obtained by using the Bernoulli distribution (π).

4.2. Missing data methods

Via simulations, we compared the performance of eight imputation methods that have been applied using R-packages for handling missing data with the proposed EPK method. These methods are KNN, EM, hot deck (HD), and using of "mice" package for each of the: mean imputation (Mean), regression imputation (RI), logistic regression imputation (LRI), predictive mean matching (PMM), and random forest (RF). Based on the results of Mohamed et al. (2021), we can suggest a new imputation method (EPK) by combining (simple average) of the imputed values of the three best MI methods (EM, PMM, and KNN) as follows (El-Sheikh et al. 2022):

$$EPK = \frac{1}{3}\left(EM + PMM + KNN\right). \tag{6}$$

4.3. Missing data generation

In our simulation study, 500 Monte Carlo samples were generated at different sample size are chosen random samples of $n = 50$, 100, and 150 observations were withdrawn. As for our selection of the number of independent variables, the simulation was performed with two different independent variables: $k = 2$, and 4. Simulations were started by deleting MAR with two percentages of data missing (10% and 40%). The missing data stage was worked on all the variables, either combined or separately for each case, as it was divided into three cases: i) Missing in independent variables (X's), ii) Missing in dependent variable (y), and iii) Missing in independent and dependent variables (X's, y). To evaluate the performance of methods, we used the following goodness-of-fit criteria: Akaike information criterion (AIC) and Bayesian information criterion (BIC).

4.4. Simulation results

The simulation results are presented in Tables 4-15 in the Appendix. Tables 4-7 presents the values of AIC and BIC of different imputation methods for different k and missingness percentage (%) when the missing data in X's only. While Tables 8-11 present the AIC and BIC values of different imputation methods when the missing data in y only when the same values of k and missingness percentage. While the simulation results for the case of the missing data in X's and y together are presented in Tables 12-15.

The most important conclusions that have been concluded from these results can be summarized as follows:

- In general, the EM and EPK methods performed better than other methods in the three cases, even with a different loss rate was (10%, or 40 %). Since EM and EPK methods have minimum values of AIC and BIC.
- From Tables 4-7, we note that the RI method works well with small samples, while the PMM method is more suitable for large samples when the sample size is greater than 50 and it performed well when increasing the proportion of missing data to 40 % and $k = 2$.
- Methods such as (RI, KNN, PMM) their results were much converged when they were $n = 50$ and $k = 4$, but the superiority of the KNN method appeared when the sample size reached 150, and the values of both (RF, PMM) converged when the sample size is large.
- The results of some methods converge, such as (KNN, PMM, and RF) with the best methods "EM and EPK", when handling missing data in y only.
- The performance of the LRI method is also like the KNN method when the sample size is greater than 100, but its performance remains poor when the loss ratio reaches 40%, $k = 4$; as shown in Tables 8-11. The large similarity condition did not differ in the results between the PMM and RF as in the case of the missing data in X's.
- The methods (Mean, LRI, PMM, RF) converge in the results when handling missing data in y only and when $k = 2$, and the sample size is 50 or 100, the slight difference appears when the sample size reaches 150. But when the percentage of missing data was increased to 40%, the results differed between them.
- The situation did not change much when dealing with the missing data in X's and y together, considering the superiority of the EM and EPK methods in processing the missing data with the logistic model, our focus was on the performance of all methods alongside the EM and EPK

methods, as it was noted that the LRI method performed well in handling the missing data in (X’s, y). Even when the proportion of missing data is high, the KNN method has been shown to be efficient in large samples and high rates of lost data, see Tables 12-15.

- The mean imputation method achieved the highest rates when the missing data in X’s as shown Tables 4-7.
- The results of most methods converge in handling missing data, whether in the independent variables only or the dependent variable only or both, especially when the sample size is at the lowest level.
- The HD method achieves its maximum value for the two criteria (AIC, BIC).

5. Empirical Application: Social Network Advertising

In this section, the real dataset is used as an application on the BLR model in case of the dataset contains missing values. Our dataset contains some information about the users of social network advertising (SNA), including their user gender, age, and estimated salary. The SNA is a group of terms that are used to describe forms of online advertising that focus on social networking services. One of the major benefits of this type of advertising is that advertisers can take advantage of the user's demographic information and target their ads appropriately. Patnana et al. (2020) analyzed this data by machine learning models. This data is also available in “mephas” R-package. The used sample size is 400 users (observations), and the definitions of the used variables are: Gender: Person can male or female, where “0” means Male and “1” means Female. Estimated Salary: Contains the salary of a person as a salary can affect the shopping. Age: Age of the person. Purchased (dependent variable)*:* Contains two numbers 0 or 1, where “0” means not purchased and “1” means purchased.

5.1. Complete data analysis

We have a case study on information about all our users in the SNA, the aim is to understand the characteristics that to identify whether everyone ended up clicking on the advertisement. To learn more about our data, let us look at the most important descriptive statistics, to start with our dependent variable, which ranked not purchased, “0”, 64.25 % compared to 35.75 % for purchased, with 257 and 143 frequencies, respectively. Regarding the independent variables, the variable “gender” contained two numbers (0,1), where the percentage of females occupied the most with 51 % compared to 49 % for males, with a frequency of 204 for females and 196 for males. The second variable of the independent variables “age”, which is one of the continuous variables, came with a mean of 37.66 and a standard deviation of 10.48. While the third variable is the “estimated salary” with two values (69743, 34097) for both the mean and the standard deviation, respectively.

It notes that some independent variables are weakly correlated; since all correlation coefficient less than 0.5. Moreover, the VIF for all independent variables are less than 5. This means that we have not multicollinearity problem and the MLE estimation method is a property for estimating the parameters of logistic regression (Abonazel and Farghali 2019, Dawoud and Abonazel 2021, Awwad et al. 2022, Abonazel et al. 2022, Akram et al. 2022, Farghali et al. 2023). Table 2 gives us some indicators to measure the efficiency of our model (Rady et al. 2021).

Table 2 Model summary and goodness-of-fit tests

Statistic	Value
Cox-Snell R-Square	0.458
Nagelkerke R-Square	0.630
McFadden pseudo R-Square	0.471
AIC	283.84
Area under curve (AUC)	92.7%
Pearson's Chi-squared test	χ^2 = 397.82 with p-value = 0.231
Osius-Rojek's test	Z = 0.54 with p-value = 0.589

From Table 2, we can say that the model is good to fit the data, because p-values of Pearson's Chi-squared and Osius-Rojek's (1992) tests are higher than the usual significance level of 0.05, this means that there is no evidence to reject the null hypothesis, so the fitted model is correct. Moreover, the values of R^2 for Cox-Snell, Nagelkerke, and McFadden are in good range (0.458 to 0.630). Moreover, the higher value of the area under the ROC (AUC = 0.927) curve corresponds to the better quality of the regression model.

Table 3 Maximum likelihood estimation results

Variable	Estimate	Standard error	p-value
Intercept	−12.78	1.3590	2E−16
Gender	0.334	0.3052	0.274
Age	0.237	0.0026	2E-16
Estimated Salary	3.644E−5	5.47E−06	2.77E−11

From Table 3, we note that all variables have a positive impact on the response variable. Gender variable shows up as not statistically significant. Age and estimated salary variables show up as statistically significant. According to the site of "Statista", the statistic shows the age distribution of the SNA audience in Singapore as of January 2020, sorted by gender. As of this date, approximately 18% of the social media advertising audience in Singapore were between 25 and 34 years old and male. New technology increases user participation and real-time content and existing networks enhance their platform and product (e.g., Facebook, Twitter, Pinterest, and Instagram launching 'buy' buttons). If the first era of social was engagement, the new era is acquisition and conversion, where social commerce has been growing over the last few years.

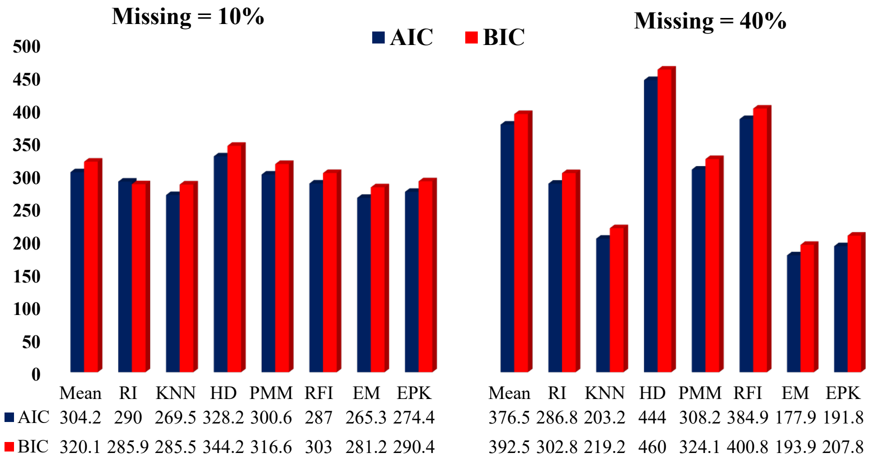
5.2. Incomplete data analysis

In this section, we have focused on discussing dealing with the different methods of missing data that can be used to analyze our dataset with missing data. We assume that some of our data are missing approximately 10% and 40% respectively, the data missing by the mechanism is 'MAR'. It should be noted that we use the same methods that were applied in the simulation study, as the missing data was applied to all variables, whether the independent variables X's only or the dependent variable y only or both X's and y together. Then we compare these methods using two goodness-of-fit criteria (AIC and BIC), see Abonazel and Ibrahim (2018). The data has been handled by the statistical R-software. After the imputation of the missing values based on the nine methods (Note that: there are 7 similar methods for handle missing data in the three cases. In addition, the RI method was used to handle the missing data in the independent variables X’s. Also, the LRI method was used for handling the missing data in other two cases.

5.3. Application results

In our datasets was dealt with in two parts, namely I) Complete data analysis II) Incomplete data analysis, the most important conclusions on our application can be presented as follows:

- In case of complete data analysis, we have shown our data without the missing data issue, the goodness-of-fit of the model was checked using different tests. The results of these tests indicate that the model is fit. And the variables such as age and the estimated salary appeared statistically significant.
- In case of incomplete data analysis, with the same methods that were applied in the simulation study. AIC and BIC criteria have been calculated at each step of this procedure implementation, see Figures 1-3. It was also evident that the simulation study and the application were consistent in concluding the best method. The results shown that the best methods are EM and EPK.



	Mean	RI	KNN	HD	PMM	RFI	EM	EPK	Mean	RI	KNN	HD	PMM	RFI	EM	EPK
■ AIC	304.2	290	269.5	328.2	300.6	287	265.3	274.4	376.5	286.8	203.2	444	308.2	384.9	177.9	191.8
■ BIC	320.1	285.9	285.5	344.2	316.6	303	281.2	290.4	392.5	302.8	219.2	460	324.1	400.8	193.9	207.8

Figure 1 Goodness-of-fit criteria when the missing values in X’s

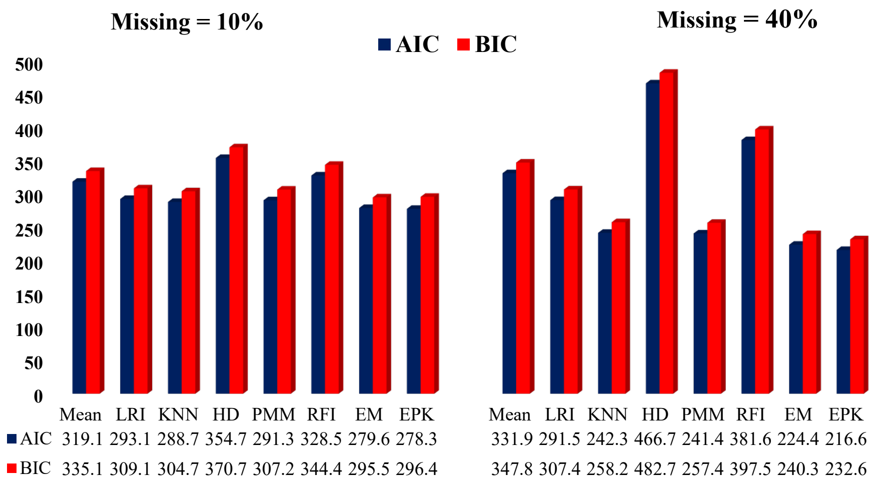


Figure 2 Goodness-of-fit criteria when the missing values in y

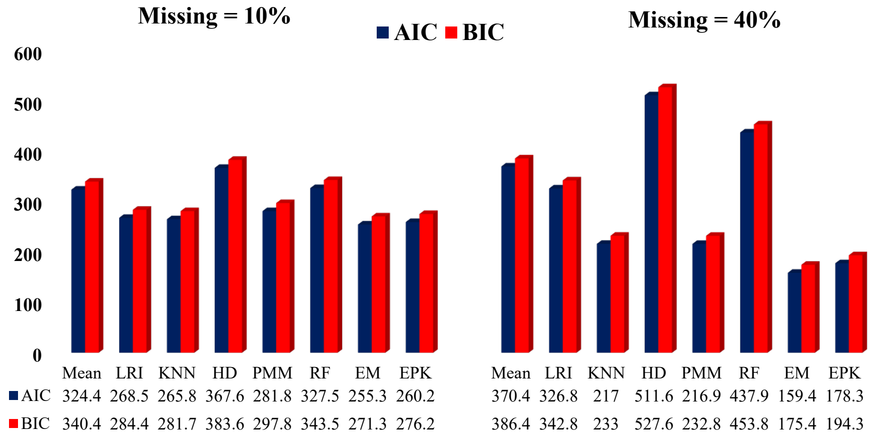


Figure 3 Goodness-of-fit criteria when the missing values in X's and y

6. Conclusion

In this paper, we have studied the performance of eight imputation methods that have been commonly used for handling missing data in the logistic regression model. Furthermore, we have proposed a new multiple imputation (EPK) method. To evaluate the performance of these methods, we have conducted a simulation study with different proportions of the missing data in all variables; whether the missingness is in the independent variables only, the dependent variable only, or both. Moreover, a real data has been used to examine the nine imputation methods and confirm the simulation results. The results of the simulation study and real data application indicated that the expectation-maximization and EPK methods are more efficient than other imputation methods, even if the missingness present in the data up to 40% and whether the missing data is in the independent variables only, the dependent variable only, or both.

References

Abonazel MR. A practical guide for creating Monte Carlo simulation studies using R. Int J Math. Comput Sci. 2018; 4(1): 18-33.

Abonazel MR, Algamal ZY, Awwad FA, Taha IM, A new two-parameter estimator for beta regression model: Method, simulation, and application. Front Appl Math Stat. 2022; 7, https://doi.org/10.3389/fams.2021.780322.

Abonazel MR, Ibrahim MG. On estimation methods for binary logistic regression model with missing values. Int J Math Comput Sci. 2018; 4 (3): 79-85.

Abonazel MR, Farghali RA. Liu-type multinomial logistic estimator. Sankhya B. 2019; 81(2): 203-225.

Abonazel MR. Handling outliers and missing data in regression models using R: simulation examples. Acad J Appl Math Sci. 2020; 6(8): 187-203.

Akram MN, Golam Kibria BM, Abonazel MR, Afzal N. On the performance of some biased estimators in the gamma regression model: simulation and applications. J Stat Comput Simul. 2022; 92(12): 2425-2447.

Armitage P, Berry G. Logistic regression. In Statistical methods in medical research. Oxford: Blackwell Scientific Publications; 1994.

Awwad FA, Dawoud I, Abonazel MR, Development of robust Özkale-Kaçiranlar and Yang-Chang estimators for regression models in the presence of multicollinearity and outliers. Concurr Comput Pract Exp. 2022; 34(6), https://doi.org/10.1002/cpe.6779.

Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work?. Int J Meth Psych Res. 2011; 20(1): 40-49.

Bender R. Introduction to the use of regression models in epidemiology. Methods Mol Biol. 2009;

Buuren SV, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Soft. 2010; 45(i03): 1-68.

Concato J, Feinstein AR, Holford TR. The risk of determining risk with multivariable models. Ann Int Med 1993; 118(3):201-210.

Dawoud I, Abonazel MR, Robust Dawoud-Kibria estimator for handling multicollinearity and outliers in the linear regression model. J Stat Comput Simul. 2021; 91: 3678-3692.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc Series B Stat Methodol. 1977; 39(1): 1-38.

El-Masry AM, Youssef AH, Abonazel MR. Using logit panel data modeling to study important factors affecting delayed completion of adjuvant chemotherapy for breast cancer patients. Commun Math Biol Neurosci. 2021; 48: 1-16

El-Sheikh AA, Alteer FA, Abonazel MR. Four imputation methods for handling missing values in the ARDL model: an application on Libyan FDI. J App Prob. Stat., 2022; 17(3):029-47.

Farghali RA, Qasim M, Kibria BG, Abonazel MR. Generalized two-parameter estimators in the multinomial logit regression model: methods, simulation and application. Commun Stat -Simul Comput. 2023; 52(7): 3327-3342.

Gelman A, Hill J. Data analysis using regression and multilevel hierarchical models. New York: Cambridge University Press; 2007.

Hilbe JM. Negative binomial regression. New York: Cambridge University Press; 2011.

Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. New York: John Wiley & Sons; 2013.

Khan KS, Chien PF, Dwarakanath LS. Logistic regression models in obstetrics and gynecology literature. Obstet Gynecol. 1999; 93(6): 1014-1020.

Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 2002.

Little RJA, Schenker N. Missing data. In: Arminger G, Clogg CC, Sobel ME, editors. Handbook of statistical modeling for the social and behavioral sciences. Boston: Springer; 1995.

Little RJA. Regression with missing X's: a review. J Am Stat Assoc. 1992; 87(420): 1227-1237.

Little RJA, Rubin DB. Statistical analysis with missing data. New York: John Wiley & Sons; 1987.

Meeyai S. Logistic regression with missing data: a comparison of handling methods and effects of percent missing values. J Traffic Logist Eng. 2016; 4(2): 128-134.

Mohamed SM, Abonazel MR, Ghallab MG. A review of ten imputation methods for handling missing values in logistic regression: a medical application. J Pure Appl Sci. 2021; 21(3): 440-451.

Nakagawa S. Missing data: mechanisms, methods, and messages. Ecol Stat Contemp Theory Appl. 2015; 81-105.

Osius G, Rojek D. Normal goodness-of-fit tests for multinomial models with large degrees of freedom. J Am Stat Assoc. 1992; 87(420): 1145-1152.

Pan R, Yang T, Cao J, Lu K, Zhang Z. Missing data imputation by K-nearest neighbours based on grey relational structure and mutual information. Appl Intell. 2015; 43(3): 614-632.

Patnana DS, Hitesh G, Kumar INS. Logistic regression analysis on social networking advertisement. J Crit Rev. 2020; 7(4): 914-917.

Peng CYJ, Zhu J. Comparison of two approaches for handling missing covariates in logistic regression. Educ Psychol Meas. 2008; 68(1): 58-77.

Rady E, Abonazel MR, Metaweâ MH. A comparison study of goodness of fit tests of logistic regression in R: simulation and application to breast cancer data. Acad J Appl Math Sci. 2021; 7(1): 50-59.

Rubin DB. Inference and missing data. Biometrika. 1976; 63(3): 581-592.

Scheffer J. Dealing with missing data. Res Lett Info Math Sci. 2002; 3(1): 153-160.

Sentas P, Angelis L. Categorical missing data imputation for software cost estimation by multinomial logistic regression. J Syst Softw. 2006; 79(3): 404-414.

Stoltzfus JC. Logistic regression: a brief primer. Acad Emerg Med. 2011; 18(10): 1099-1104.

Tranmer M, Elliot M. Binary logistic regression. Cathie Marsh for census and survey research, paper, 20; 2008.

Tsikriktsis N. A review of techniques for treating missing data in OM survey research. J Oper Manag. 2005; 24(1): 53-62.

Wisniewski SR, Leon AC, Otto MW, Trivedi MH. Prevention of missing data in clinical research studies. Biol Psychiatry. 2006; 59(11): 997-1000.

Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, Zhou Z H. Top 10 algorithms in data mining. Knowl Inf Syst. 2008; 14(1): 1-37.

Appendix

Table 4 AIC and BIC values of different imputation methods when missingness = 10 % in X's and $k = 2$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	49.595	55.331	106.830	114.646	162.220	171.252
LRI	48.207	53.943	104.519	112.334	158.704	167.736
KNN	50.335	56.071	108.447	116.263	165.113	174.145
HD	48.321	54.057	103.980	111.796	159.118	168.149
PMM	48.581	54.3177	104.683	112.498	158.731	167.763
RF	48.072	53.808	103.373	111.189	156.976	166.007
EM	47.423	53.159	102.445	110.260	155.643	164.675
EPK	47.212	52.948	101.476	109.292	154.366	163.398

Table 5 AIC and BIC values of different imputation methods when missingness = 40 % in X's and $k = 2$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	52.180	57.916	112.401	120.217	170.456	179.488
LRI	45.806	51.542	102.163	109.978	156.482	165.514
KNN	54.034	59.770	116.835	124.650	177.894	186.926
HD	48.791	54.527	101.995	109.811	156.210	165.241
PMM	47.967	53.703	102.941	110.757	156.684	165.716
RF	46.039	51.775	93.677	101.492	142.388	151.420
EM	39.680	45.416	89.395	97.211	137.072	146.104
EPK	42.392	48.128	88.817	96.633	135.816	144.847

Table 6 AIC and BIC values of different imputation methods when missingness = 10 % in X's and $k = 4$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	54.982	64.542	112.478	125.504	167.469	182.523
LRI	52.879	62.439	109.557	122.583	163.825	178.879
KNN	55.984	65.544	114.330	127.356	170.348	185.401
HD	53.024	62.584	109.934	122.960	163.820	178.873
PMM	54.065	63.625	110.441	123.467	164.459	179.512
RF	52.697	62.257	108.920	121.946	162.322	177.376
EM	52.229	61.790	107.823	120.849	161.022	176.075
EPK	52.256	61.817	107.905	120.931	160.721	175.774

Table 7 AIC and BIC values of different imputation methods when missingness = 40 % in X's and $k = 4$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	58.040	67.600	116.904	129.930	175.937	190.990
LRI	45.098	54.658	103.772	116.798	160.134	175.187
KNN	60.542	70.102	120.694	133.720	182.277	197.330
HD	49.432	58.992	105.494	118.519	160.529	175.582
PMM	54.354	63.915	108.694	121.720	164.368	179.421
RF	51.948	61.5090	106.277	119.303	158.018	173.071
EM	49.275	58.836	100.427	113.453	153.859	168.912
EPK	48.002	57.562	99.846	112.872	151.239	166.292

Table 8 AIC and BIC values of different imputation methods when missingness = 10 % in y and $k = 2$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	47.112	52.848	102.249	110.064	156.082	165.114
LRI	48.633	54.369	103.864	111.679	159.445	168.477
KNN	46.648	52.384	101.030	108.846	154.858	163.890
HD	49.735	55.472	108.093	115.908	164.873	173.905
PMM	47.924	53.660	103.615	111.431	159.950	168.982
RF	47.813	53.549	103.342	111.157	158.048	167.080
EM	46.587	52.323	100.702	108.518	154.275	163.307
EPK	46.259	51.995	100.167	107.983	153.479	162.511

Table 9 AIC and BIC values of different imputation methods when missingness = 40 % in y and $k = 2$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	42.726	48.462	95.317	103.133	146.561	155.593
LRI	47.776	53.512	102.458	110.273	157.448	166.480
KNN	40.131	45.867	90.337	98.152	139.486	148.518
HD	52.471	58.207	115.653	123.468	177.121	186.153
PMM	43.828	49.564	101.884	109.699	155.755	164.787
RF	46.857	52.593	101.078	108.894	154.249	163.281
EM	39.487	45.223	86.848	94.664	132.588	141.620
EPK	38.468	44.204	86.072	93.887	132.347	141.379

Table 10 AIC and BIC values of different imputation methods when missingness = 10 % in y and *k* = 4

Method	*n* = 50		*n* = 100		*n* = 150	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	52.460	62.020	108.269	121.294	162.291	177.344
LRI	52.827	62.387	109.987	123.013	164.699	179.752
KNN	51.321	60.881	107.246	120.272	161.214	176.267
HD	55.133	64.693	113.557	126.583	169.885	184.938
PMM	52.286	61.846	109.586	122.611	164.952	180.005
RF	52.621	62.182	109.413	122.439	164.062	179.115
EM	51.249	60.809	107.317	120.343	160.488	175.541
EPK	50.795	60.355	106.061	119.087	159.152	174.205

Table 11 AIC and BIC values of different imputation methods when missingness = 40 % in y and *k* = 4

Method	*n* = 50		*n* = 100		*n* = 150	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	49.905	59.466	102.105	115.131	150.730	165.783
LRI	51.202	60.763	107.782	120.808	159.969	175.023
KNN	44.386	53.946	97.171	110.197	144.719	159.772
HD	58.550	68.110	120.901	133.927	179.389	194.442
PMM	51.810	61.370	108.108	121.134	166.086	181.139
RF	51.961	61.522	107.071	120.097	158.421	173.474
EM	44.751	54.311	95.346	108.372	141.313	156.366
EPK	43.356	52.916	92.275	105.300	137.853	152.907

Table 12 AIC and BIC values of different imputation methods when missingness = 10 % in (X's, y) and *k* = 2

Method	*n* = 50		*n* = 100		*n* = 150	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	48.375	54.111	103.592	111.407	159.742	168.774
LRI	48.499	54.236	103.524	111.340	160.044	169.076
KNN	46.359	52.095	99.647	107.463	153.828	162.860
HD	51.539	57.275	109.935	117.750	169.552	178.584
PMM	48.248	53.984	103.063	110.879	159.795	168.827
RF	48.481	54.217	103.575	111.391	159.018	168.050
EM	44.319	50.055	95.681	103.496	147.341	156.373
EPK	45.474	51.211	97.778	105.594	150.970	160.002

Table 13 AIC and BIC values of different imputation methods when missingness = 40 % in (X's, y) and $k = 2$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	44.635	50.371	98.996	106.811	150.747	159.779
LRI	45.851	51.587	99.217	107.032	152.208	161.240
KNN	39.238	44.974	79.681	87.497	119.902	128.934
HD	53.600	59.336	118.709	126.524	180.754	189.786
PMM	45.213	50.949	98.356	106.172	150.646	159.678
RF	45.955	51.691	98.846	106.661	152.641	161.673
EM	29.568	35.304	68.161	75.977	106.083	115.115
EPK	36.528	42.264	78.185	86.001	118.775	127.807

Table 14 AIC and BIC values of different imputation methods when missingness = 10 % in (X's, y) and $k = 4$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	53.242	62.803	110.074	123.099	164.975	180.028
LRI	52.152	61.712	109.515	122.541	164.724	179.777
KNN	50.494	60.055	107.143	120.169	160.498	175.551
HD	56.166	65.726	116.118	129.144	174.543	189.597
PMM	51.540	61.100	109.726	122.752	164.323	179.376
RF	53.343	62.903	110.516	123.542	165.112	180.165
EM	47.865	57.425	103.841	116.867	156.084	171.137
EPK	49.713	59.273	105.004	118.031	157.752	172.805

Table 15 AIC and BIC values of different imputation methods when missingness = 40 % in (X's, y) and $k = 4$

Method	$n = 50$		$n = 100$		$n = 150$	
	AIC	BIC	AIC	BIC	AIC	BIC
Mean	51.860	61.420	104.381	117.407	155.611	170.664
LRI	45.426	54.986	99.417	112.442	152.014	167.067
KNN	46.716	56.276	99.660	112.686	148.950	164.003
HD	58.547	68.107	121.603	134.629	183.219	198.272
PMM	44.082	53.642	98.321	111.347	152.407	167.460
RF	53.050	62.610	107.902	120.928	161.492	176.545
EM	36.864	46.424	79.912	92.938	121.268	136.322
EPK	42.440	52.006	90.143	103.168	135.430	150.483