Scientific Research Publishing

Open Journal of Statistics, 2021, 11, 759-777
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Why Can Multiple Imputations and How (MICE) Algorithm Work?

Abdullah Z. Alruhaymi, Charles J. Kim

Department of Electrical Engineering and Computer Science, Howard University, Washington DC, USA
Email: azmotairi@hotmail.com, ckim@howard.com

How to cite this paper: Alruhaymi, A.Z. and Kim, C.J. (2021) Why Can Multiple Imputations and How (MICE) Algorithm Work? *Open Journal of Statistics*, **11**, 759-777.
https://doi.org/10.4236/ojs.2021.115045

Received: July 27, 2021
Accepted: October 11, 2021
Published: October 14, 2021

Abstract

Multiple imputations compensate for missing data and produce multiple datasets by regression model and are considered the solver of the old problem of univariate imputation. The univariate imputes data only from a specific column where the data cell was missing. Multivariate imputation works simultaneously, with all variables in all columns, whether missing or observed. It has emerged as a principal method of solving missing data problems. All incomplete datasets analyzed before Multiple Imputation by Chained Equations (MICE) presented were misdiagnosed; results obtained were invalid and should not be countable to yield reasonable conclusions. This article will highlight why multiple imputations and how the MICE work with a particular focus on the cyber-security dataset. Removing missing data in any dataset and replacing it is imperative in analyzing the data and creating prediction models. Therefore, a good imputation technique should recover the missingness, which involves extracting the good features. However, the widely used univariate imputation method does not impute missingness reasonably if the values are too large and may thus lead to bias. Therefore, we aim to propose an alternative imputation method that is efficient and removes potential bias after removing the missingness.

Keywords

Multiple Imputations, Imputations, Algorithms, MICE Algorithm

1. Introduction

The challenge of missing data values is encountered by researchers in many fields and many types of research. Missingness lowers the quality of a dataset, affecting the analysis carried out using that data. This has led to a lot of research being done to develop practical missing data imputation approaches. Data im-

putation tries to restore the missingness within a dataset by carrying out an analysis of the relationships and characteristics that are hidden within it. However, inapt imputation techniques can lead to incorrect analysis of the dataset, and thus wrongful conclusions and predictions have been drawn from it. This paper attempts to test the multiple imputations by chained equations (MICE) approach as the best imputation alternative to the univariate approach.

The data is the most assets one can have, and is the basis of every decision-making approach and the quality of data if manipulated possess huge backlash in decision outcome from it. So, the attack in data has been huge threat of cyber security. The cyber security attack might lead to multiple losses of data sets which can be instrumental in planning and implementing the policies. The re-collection of such data has a lower probability of being consistent with the existing data, so recollecting might not be feasible and any good in most cases, hence correct prediction of the missing data is a must. The prediction of missing values of data requires robust mechanism that can be trusted. Also, this study is carried out to illustrate the multiple imputation of missing data and use of MICE algorithm for dataset with missing values.

The work of Rubin and Little on missing data and its treatment has been a cornerstone in statistics for a long time. The work of these two continues to influence academics and professionals alike when faced with missing data problems. In general, imputation can be described as the replacement of missing values with new values from given criteria. Multiple imputations follow the same criteria except for multiple iterations and are produced to make an m number of completed datasets with the imputed values. Results from the completed m datasets are then analyzed and pooled together to arrive at a final imputed value. Just like the single imputation approaches, there are different multiple imputation approaches.

2. Multiple Imputation Explained

Multiple imputations use multiple regression models to estimate missing values and it integrates uncertainty by applying the iterative approach. Multiple imputations by chained equations utilize the wide distributions of the observed values to come up with the missing data values. It is related to the NORM program, where key objectives involve the restoring of error-variance that is lost during single-based imputation. Usually, such imputed data values form on the right side of the regression line. The actual data during such cases typically deviate from the regression line to some extent. Restoring lost variance requires the initial imputation to be added to the random error variance; this should be the unexpected standard error. The second step in restoring lost variance is because every amputated value depends on a single regression equation. This is because single draws from target populations define the covariance matrix and the regression equation. Imputations, of course, entail making up data, where plausible values that do not exist undergo plugging. Some researchers criticize mul-

tiple imputations by saying that it merely makes up some numerical values for the missing data. However, this assumption is not correct because the primary usage of it in any statistical analysis is to determine the population estimates parameters through predicting close values to the missing ones.

In multiple imputations, following a typical model tends to preserve variances, means, covariance, linear regression coefficients, and correlation. MI, therefore, was designed to restore variability that is lost during single imputation and consequently correct it. Rounding procedures included increasing the variability of imputed values.

Multiple imputations help to ensure that the missing data is uncertain by generating various reasonable imputed data sets and integrating findings from each of them properly. Multiple imputations employ the imputation, analysis, and pooling processes [1].

In these estimated values, random components are inserted to represent their uncertainty. A collection of parameter estimates is generated and then evaluated independently but identically. Multiple imputations were made under the MAR mechanism in most statistical software to provide impartial and accurate association estimates based on available data. This approach not only co-effective provides estimates with missing variables, but also provides estimates without any missing data for all the other variables. MI could, but standard implementations take MAR, also it can be implemented under MNAR frameworks in a special case. When properly implemented, in both cases it may asymptotically and asymptotically produce accurate estimates and default errors.

According to Kontopantelis [2], multiple imputations are simulation-based, aiming not to re-create the single missing data cell but to handle missing data to deduce proper statistical inferences if the imputation suits with missing mechanism and complete data are valid in the absence of any data missingness.

Several integrations of factor levels of multiple imputations by Rubin [3], Little and Rubin [4], Van Buuren [5], Carpenter and Kenward [6] are to figure out the number of imputations necessary? Is it an unlimited M? and what value should be used in practice? Because valid imputation is dependent on the size of imputations of M. And the closest relative effectiveness of the MI process limited M compared with unlimited M is around 90%, meaning there are only 2 imputations for the rate of missing-information as big as 50% Rubin [7]. According to Rubin 1987, relativeness to one is based on the unlimited number.

The effectiveness of an estimate which is based on m imputation $\left(1+\lambda/m\right)^{-1}=100/1+0.05=95\%$, where λ is the rate of missingness of data as 50% and M is the number of imputations. The MSE can also be calculated, $\mathrm{MSE}=\left(b-\beta\right)^2/N$, where b is regression coefficients, N is random draws from the population, β is the value of regression coefficients.

Multiple imputations are a modern technique, but sometimes researchers object to imputation because imputation is coming up with artificial data, and synthetic data should not be used. Still, the actual data should be used wherever

possible and not make guesses about missing data. These reasons can justify this kind of criticism:

1) The idea of missing data imputation is not to generate valid estimates of specific case values. Instead, this is a tool for calculating relations between variables. Whether the individual projections for specific case values are correct or not is irrelevant as long as the correlation and other statistics association correctly done.

2) While this technique sounds like cheating, it has been proven to be consistent and produce better results than simply using available data and deleting missing entries.

The imputation technique imputes the missing data entries of data sets that are not complete M times. These values are drawn from a statistical distribution. The analysis step assesses each one of the complete M data sets. A dataset that falls within 5 to 10 imputations provides unbiased parameter estimates and full sample size when done well. Lastly, pooling is done using Rubin's rule [8] by integrating m analysis results into a final value. The process can be expressed in the following expression. Let the column vector of the interest estimate be Q, its estimate be $\tilde{Q}$, the number of datasets imputed is m and the estimated (I_{th}) repeated analysis be $\tilde{Q}_l$. The combination of the estimate is expressed as:

$$\bar{Q} = \frac{1}{m}\sum_{l=1}^{m}\tilde{Q}_l \tag{1}$$

Hence, multiple imputations aim to come up with an estimated $\bar{Q}$ that is unbiased and with a valid confidence level. This method is by far the most ideal for imputation because it is easy to visualize and use and unbiased if the imputation model is correct, thus, the imputation model is simple to review and use (**Figure 1**).

1) Stage 1:

Generating multiply imputed data sets by fitting the model $z \sim \theta X$ (means z follow beta X) using individuals with observed $z \to \hat{\theta}, \hat{V}$ repeating following steps m times to draw θ^* from multivariate normal MVN ($\hat{\theta}, \hat{V}$), then from

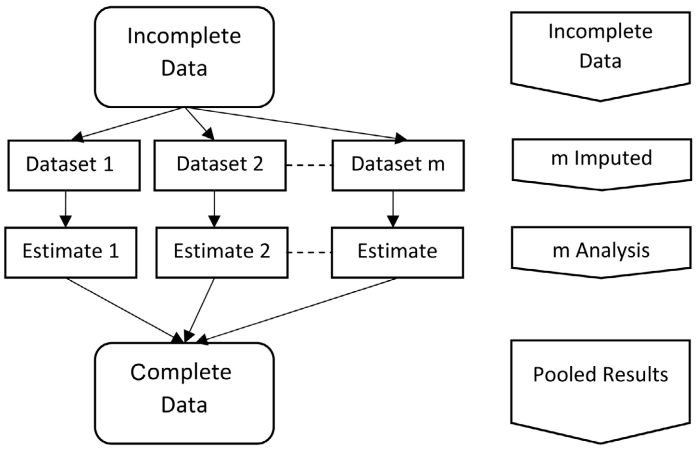

Figure 1. Three stages of multiple imputation process.

the posterior distribution of Z, drawing imputations for z via θ^*. This process is defined as a proper imputation since the alleged values contain all sources of uncertainty and variability (**Figure 2**).

2) Stage 2:

Multiple data sets imputed are analyzed, so that each imputed dataset estimates whatever parameters of interest (e.g., regression coefficients) and its estimated variability. Every data set imputed is individually analyzed after multiple imputations are produced. It is a simple process as the methods for complete data of every data set imputed and their various matrices of variance-covariance are accessible, scientific amounts (normally regression coefficients) are calculated [9]. The results of these m analyze vary since various imputations have substituted the missing values (**Figure 3**).

3) Stage 3:

In this step, along with estimates of many imputed data sets usually five imputed datasets, the integration of m estimates from stage two into a total estimation and covariance matrix is done, while using the Rubin rule. When the in-between estimate and the model are stable and there is no variance in the last two datasets then iteration of the imputation process is stopped. A common mistake that researchers pool imputed dataset warned against Buuren in his book "Flexible Imputation of Missing Data" where he quoted (Researchers tempted to average the multiply imputed data and analyze the average data as if it were complete, this method yields incorrect standard errors).

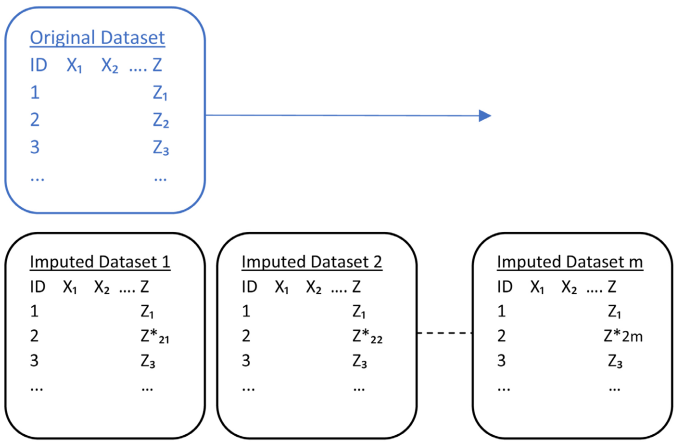

Figure 2. Generating multiple imputed data set.

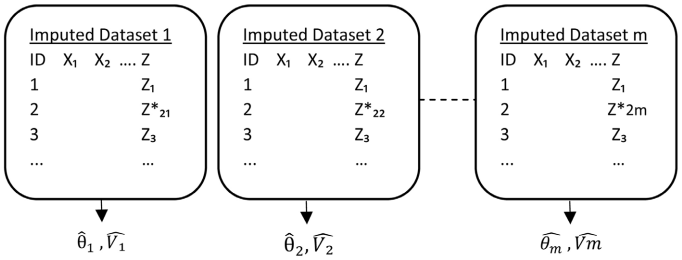

Figure 3. Analysis of imputed data.

The combined matrix of variance-covariance includes within variability imputation and provisional variability imputation (including lack of information). Assume that $\hat{\theta}_j$ approximates a multivariate or univariate interest quantity derived from its jth imputed dataset j and W_j is $\hat{\theta}_j$'s variance of estimate [10]. The combination of the individual estimates is utilized to create the combined estimate $\hat{\theta}$. Expressed as:

$$\hat{\theta} = \frac{1}{m}\sum_{j=1}^{m}\hat{\theta}_j \tag{2}$$

The total due to the $\hat{\theta}$ variance is obtained from the variance found within-imputation expresses as:

$W = (1/m)\sum_{j=1}^{m}\hat{W}_j$ while imputation variance between B is expressed as $(1/m)\sum_{j=1}^{m}\left(\hat{\theta}_j - \hat{\theta}\right)^2$.

$$\text{Variance}\left(\hat{\theta}\right) = B\left(\frac{1}{m}+1\right)+W \tag{3}$$

Missing values in several variables are typical in large datasets. Multiple imputations are a practical way of generating values based on a series of models of imputation, one with missing values for each feature. For any other variables that have missing values, the process is repeated many times. The repetition during the process is termed a cycle. The process that is used to generate a single imputed data set is normally replicated over multiple cycles, and the entire process is used m times to provide datasets imputed. MI is crucial since each variable is imputed to be appropriate for different types of variables by its model imputation. In many MI cases, there are conditions to be satisfied before performing any missing data analysis. First, data must be MAR data that depends on observed variables. Also the appropriate model should match other models for MI's missing data.

3. MICE Approach Highlighted with Examples

The MICE algorithm in its present form was advanced by Buuren [11], but before this time, there were different methods of implementing imputation using the models of conditional specification. It appeared first after the publication of the Statistical Software Journal of Stef Van Buuren and Karin Groothuis-Oudshoorn [12]. The algorithm works by randomly drawing from the observed values and then imputing the missing data values in a variable-by-variable manner. The MICE is a Markov chain Monto Carlo (MCMC) method [13].

Many MICE methods recalled from [methods(mice)] that suit to different data types, two kinds mentioned that used for data in subsequent other section which:

- Multiple Imputation with Predictive Mean Matching

A popular multiple imputation method is the semi-parametric predictive mean matching (PMM) model. First proposed by Rubin, the model seeks to replace missing values using values from observed values with a similar predictive

mean from the dataset. PMM is considered a standard model for handling incomplete continuous data. The PMM used for the dataset imputation had some superiority over other multiple imputation models regarding the preservation of the empirical data when said data deviates from their distributional assumptions. The study found that PMM performed better with smaller donor pools than larger pools.

- Multiple Imputation with Logistic Regression

Logistic regression is a popular classification method used when faced with binary.

MICE is intended to work with MAR data, though it can also work on data that is MNAR.

MICE is an informative and robust reliable approach when handling missing data in any dataset. The procedure imputes missing cells in variables in a dataset through a repeated step of an iterative series of some productive models. In each cycle of iteration, each specific variable in the data set is imputed using other variables.

Two methods of generating multiple imputations are univariate and multivariate missing data. Multivariate imputation is considered a method of identifying relationships and patterns among variables concurrently. Multivariate imputation facilitates the prediction of the consequences the change in one variable has on other variables. This tool for multivariate imputation has been widely accepted as the principal method when addressing missing data. MICE is also called "fully conditional specification" or "sequential regression multiple imputations" and can handle varying missing datasets as well as complex cases.

It uses linear regression to impute null values in continuous data in N iterations to achieve more accurate and stable results by finding the best fit line for the data by utilizing the minimizing error function (MSE) and applies a logistic regression algorithm/log loss using a sigmoid function for discrete data. In the case of the discrete data, logistic regression models are an excellent tool for analyzing binary and categorical data because they allow you to conduct a contextual analysis to understand the relationships between variables, test for differences, estimate impacts, make predictions, and plan for future scenarios. The key benefit of multiple imputations over univariate imputation is that it retains N without presenting any biasness especially if the data in question MAR. It is also offering the correct SEs for uncertainty as a result of missing data values. The process is as follows:

1) Removing every missing value's mean in the dataset, and these imputations can be thought of as "place holders".

2) Variable "var" means imputations are put back as missing.

3) "var" observed values in step 2 above are reverted on the other variables in the model." var" is the dependent variable and all the other variables are predictor variables in the model. The regression models use similar assumptions made when carrying out a linear regression model under normal situations.

4) "var" missing data values are then substituted by predications of the regression model. When "var" is finally made use of as a predictor variable in the model for all the other variables, the observed and the imputed values are utilized.

5) Next, steps 2 to 4 are repeated for every variable with a missing data value. The recycling over all of the variables comprises of one iteration, and after the completion of one cycle, the predictions from the regression model that show the relationships in the observed data have substituted the missing data values.

6) Steps 2 to 4 are again recurrent for several other cycles with the imputation getting updated repeatedly for each cycle.

The benefits of the MICE approach of imputing missing data for a Cybersecurity database of detecting the cyber threat leads to a higher detection rate and a lower rate of false alarms.

MICE work numerically in only one single iteration, then repeat iterations many m times from step 3 to 9 until the model is stable, as the diagram below explains (**Figure 4**).

The method that emerged as the main method of handling missing data used to be called "Fully Conditional Specification" or "Sequential Regression Multiple Imputation" as it is a development of multiple regression process that handle varying missing datasets as well as complex cases, uses linear regression to impute Null values in Continuous data and logistic regression for discrete values in N iteration to achieve more accurate and stable results of the model. It is worth

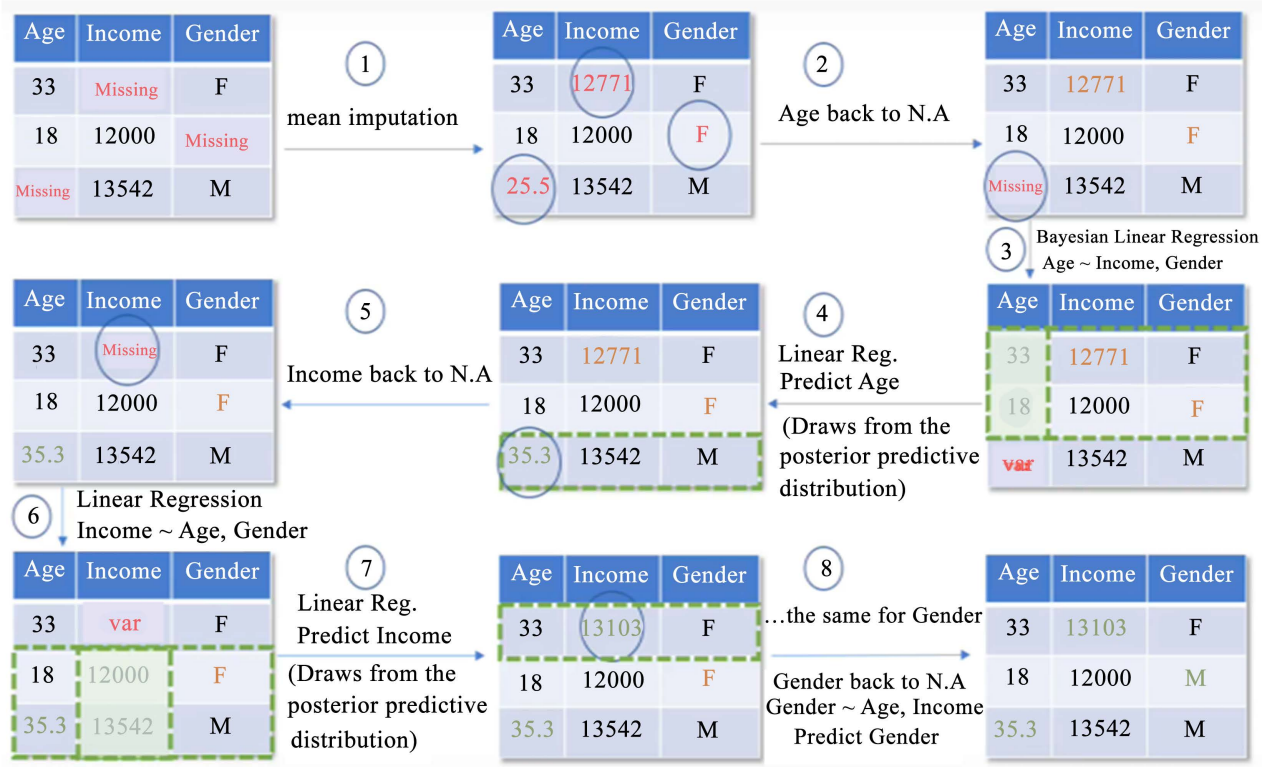

Figure 4. How MICE work numerically? (Philip 9876, 2018. You tube, retrieved September 2021 from https://www.youtube.com/watch?v=zX-pacwVyvU)

thinking that MICE may be used in cybersecurity databases to handle data masking in case of cyber threats. By modifying sensitive user and organization data into a useless form, like making it incomplete with MAR missingness, an attacker cannot use it to harm the data owners and maintain the validity of the data, to help converse data confidentiality of the dataset.

3.1. Example of Multivariate vs. Univariate

In the example below, the easy representing the MICE method is shown:

1) A simulates a real-world scenario by explaining the MICE from a small dataset consisting of eight rows and three variables.

2) Intentionally remove the values of some cells and make incomplete data that missing some elements from it a proportion of 3/24 equivalent to nearly 13% missing rate.

3) The single imputation was done by taking the mean from each column separately, then the means were 29, 7 and 134, which it did not represent actual value.

In the above deterministic example, the work was done on a complete simulated dataset and intentionally removed some data values from some cells to make incomplete dataset mimic real-world situations. As explained above impute by a single imputation the missing values with univariate imputation. Of course, there are many single imputation methods, so the imputation is done using mean from each column and imputing values in place of the missing values. The bias results present here are apparent from strange results of 29, 7 and 134. The results did not represent an actual situation, so the problem is not solved yet because of this univariate imputation. It could be imputed by median or mode or a constant number, but no results are better than the imputed ones. Then using the machine learning regression model and divide the data to train and test data. And set the target is missing in each column. Iterating using the MICE approach, final result is reached in dataset imputed number four. A consequence is the same as the actual values in the real dataset or very close. This proves the use of MICE as an effective solution for the missing data problems and any other method, including deletion, were to cause incorrect inferences and invalid results (Table 1).

3.2. Imputation Using R Code for 8 Rows and Three Variables

```
library ("mice")
data <- read.csv ("C:/Users/azmot/Desktop/8rows.csv")
set.seed (1234)
imp <- mice (data, maxit = 30, method = c ("norm.predict", "norm.predict", "norm.nob"), m = 10, threshold = 2)
imp_data <- complete (imp, 4)
## Alternative
imp1 <- mice (data, maxit = 30, method = "norm.nob", m = 5, threshold = 2)
imp_data1 <- complete (imp1, 4)
```

Table 1. Example of Regression Deterministic model using ML and imputed by MICE.

Example: Regression deterministic model

Real dataset

Age	Experience	Income in K
25	1	50
27	3	80
29	5	110
31	7	140
33	9	170
35	11	200
37	13	230
39	15	260

Missing dataset

Age	Experience	income in K
25	NA	50
27	3	NA
29	5	110
31	7	140
33	9	170
NA	11	200
37	13	230
39	15	260

Mean imputation (Dataset 0)

Age	Experience	Salary in K
25	7	50
27	3	134
29	5	110
31	7	140
33	9	170
29	11	200
37	13	230
39	15	260

Complete data consists of 24 cells, we will impute by cell wise and about 10% from the data 3/24. Simple linear regression used.

Step 2: Set back mean value of age as NA and remaining cells become feature matrix age is the target variable

Age	Experience	Income in K
25	7	50
27	3	80
29	5	110
31	7	140
33	9	170
NA	11	200
37	13	230
39	15	260

Step 3: Run a linear regression on observed to estimate missing age, using missing row as test set

Age	Experience	Income in K
25	7	50
27	3	80
29	5	110
31	7	140
33	9	170
36.3	11	200
37	13	230
39	15	260

Step 4: Now set back experience to NA, remaining rows and features are become feature matrix, and experience is the target variable, run LR on the fully observed row

Age	Experience	Income in K
25	NA	50
27	3	134
29	5	110
31	7	140
33	9	170
36.3	11	200
37	13	230
39	15	260

Step 5: Do same with income run LR to predict income

Age	Experience	Income in K
25	1.85	50
27	3	NA
29	5	110
31	7	140

Continued

33	9	170
36.3	11	200
37	13	230
39	15	260

Step 6: First imputed with regression (step 6 - dataset 0). Dataset 1

Age	Experience	Income K
25	1.85	50
27	3	72.77
29	5	110
31	7	140
33	9	170
36.3	11	200
37	13	230
39	15	260

Goal: convergence is far do more iteration (Dataset 1 - Dataset 0)

Age	Experience	Income in K
0	−5.15	0
0	0	−61.23
0	0	0
0	0	0
0	0	0
7.25	0	0
37	0	0
39	0	0

Convergence is far to reach, and the goal is to minimize the different matrix. Stop iterations when a pre-defined threshold is reached or when the difference between the last two imputed datasets is slight small or none.

Repeat steps from 2 to 6 again and start with Dataset 1, iteration 2

Dataset 1

Age	Experience	Salary in K
25	1.85	50
27	3	72.77
29	5	110
31	7	140
33	9	170
36.3	11	200
37	13	230
39	15	260

Dataset 2 imputed linear regression

Age	Experience	Salary in K
25	0.91	50
27	3	80.7
29	5	110
31	7	140
33	9	170
34.8	11	200
37	13	230
39	15	260

Difference 1 - 2

Age	Experience	Salary in K
0	0.89	0
0	0	7.96
0	0	0
0	0	0
0	0	0
1.38	0	0
0	0	0
0	0	0

Data set 1 already regressed and now repeat the same process and calculate Data set 2

Repeat steps from 2 to 6 again and start with Dataset 2, iteration 3

Dataset 2

Age	Experience	Income in K
25	0.92	50
27	3	80.7
29	5	110
31	7	140
33	9	170
34.87	11	200
37	13	230
39	15	260

Imputed linear regression dataset 3

Age	Experience	Income in K
25	1	50
27	3	79.98

Continued

29	5	110
31	7	140
33	9	170
35	11	200
37	13	230
39	15	260

Difference 2 - 3

Age	Experience	Income in K
0	−0.08	0
0	0	0.72
0	0	0
0	0	0
0	0	0
−0.13	0	0
0	0	0
30	0	0

Repeat steps from 2 to 6 again and start with Dataset 3, iteration 4

Dataset 3

Age	Experience	Income in K
25	1	50
27	3	70.98
29	5	110
31	7	140
33	9	170
35	11	200
37	13	230
39	15	260

Imputed linear regression dataset 4

Age	Experience	Income in K
25	0.99	50
27	3	80
29	5	110
31	7	140
33	9	170
34.9	11	200
37	13	230
39	15	260

Difference 3 - 4

Age	Experience	Income in K
0	0.01	0
0	0	0.02
0	0	0
0	0	0
0	0	0
0.01	0	0
0	0	0
0	0	0

Stop Imputation since convergence is met after 4 imputation times and the difference between the last two imputed is zeros or very small.

For imputation by MICE, there are nearly 22 methods that might be picked to suit the dataset on hand. Based on observations, to impute the missing data, the best fit for imputation would be using linear regression. For example, there might be use of norm.predict for Age and Experience while norm.nob for the Income column. The norm.predict impute the missing values based on the prediction of the model from the data. The "norm" method can also be used, which uses Bayesian linear regression. The "norm.nob" method is similar to norm.predict, but it adds some random noise to the predicted value. Since stochastic regression is being performed norm.nob is the most appropriate method. An alternative method has been put that uses norm.nob for all columns. And the increase in the maximum iteration value can be done to get better results (keeping in mind it increases the computation time too). The maxit is set to 30, so that there are more choices to choose from. This code imputes the three cells missing data using the MICE library and gets results after four imputed datasets.

4. Discussion

Real-world missingness was simulated by intentionally removing some data values from a dataset. The MICE algorithm was then used to remove this missingness by identifying a group of explanatory variables for every variable with incomplete data value. It then populated a matrix with these variables and used it to predict the missing values. Finally, the original dataset was arrived at, proving that this approach is practical with no bias. The same procedure was repeated using the univariate method, and this yielded bias results that were not like the original dataset.

For statistical accuracy and valid conclusion, any gaps need to be correctly adjusted and improved. One of the significant improvements for solving missing data problems is to use multiple imputations by chained equations as this method is a crucial solution and accounted for; many researchers conclude this technique as the best strategy to fill in the missing values in an incomplete data-

set. Moreover, this article contributes to the main Dissertation body, which investigates how to relate MI to handle missingness in datasets, especially Cybersecurity databases, which suffer dropping packets on traffic TCP/IP.

The ability to predict a cyber-attack before it occurs will be a revolutionary event in the fight against cyber threats in the online space. However, there is a challenge of retrieving sufficient and complete data to build robust and valid predictive models. We try to expand more on the machine learning solutions already in place to solve missing signal data values in the cybercrime event datasets and suggests detecting the attacks if they happen.

Over the past few years, cyber-attacks have become diverse and thus becoming a way for criminals to earn money, steal personal intellectual properties and even promote extreme political agendas. These attacks continue to rise even now, and their cost to individuals and corporations has become tremendous. Therefore, it has become necessary for engineers and researchers overall to come up with ways of predicting these attacks before they even occur. However, there has been a challenge of the significantly changing unusual signals of impending attacks, mainly on social media. These signals do not give out values regularly since they can only produce them when these events occur. This gives rise to the problem of incomplete data caused by unrecorded events during specific periods. Furthermore, successes in cyber-crime are now expected to be rare events since organizations are currently not heavily protected against them. This results in data that is not balanced and complete, which causes not to create powerful predictive models of forecasting the same.

Therefore, we seek to address these problems by suggesting a method of dealing with the missing signal values by encompassing the already existing works on approaches of dealing with missing data through imputation. Furthermore, it also suggests coping with the cyber-attacks if they happen through masking through the addition of missingness and its removal using the multiple imputation by chained equations (MICE).

It is vital to ensure the best data imputation approaches are used to uncover the intended missingness initiated by computer systems in many organizations that aim to protect databases against cyber-attacks after detection. It is illustrated that the best method to ensure data masking without losing any information is the multiple imputations by chained equations. Immediately after a cybercriminal tries to hack a system, MAR missingness is added intentionally into the user and company data and masks the original data ensuring that confidential user and company data is kept away from hacking activities.

Afterward an imputation method is used to convert back and unmask the data by removing the missingness out of the data. Assurance that the data before and after masking is the same. It is suggested using multiple imputations by chained equations (MICE) to ensure the validity and accuracy of intentionally removed data.

The results of multiple imputations using MICE algorithm gave imputed values very accurate to the knowingly deleted values. First the values were filled in

empty places by univariate imputation using mean of data of each column. Then the regression model is established based on type of variable throughout the columns of dataset. For continuous variables linear regression was used and for categorical variables logical regression was used. The next step of imputation is multiple imputation by regressing the variables to revise the values of previously univariately regressed data. It is repeated multiple times to decrease error of prediction with multiple iteration.

The results shows that the multivariate approach of imputation yields in more accurate and representative imputed valued compared to univariate model as it uses the repetitive iteration of correcting the values with each step. The MICE approach of multivariate imputation is robust and can be used for handling of missing or lost segment of data from dataset.

5. Conclusions

Missingness in a dataset should not lead to wrongful conclusions about a dataset by using inappropriate handling methods. By recommending using the MICE approach, as shown in the previous sections, imputation leads to a dataset that results in robust prediction models and leads to valid conclusions about the data. The effectiveness of the proposed method is tested by comparing it to the univariate approach, which yields bias results when used on a dataset mimicking the real world. Indeed, the MICE approach, on the other hand, produced valid results and showed no biases. Multivariate imputation by the chained equation is a valuable solution for the single imputation problem. Using a single type of data to get imputed values is called univariate imputation, which means using only one column to impute missing values. The univariate approach is fast and less tasking, but unfortunately, it causes bias and indicates false inferences.

It is upsetting that the univariate brute-force approach is considered a severe issue of single imputation techniques. To solve the problem of univariate imputation is the use of multivariate imputation by chained equations through factoring in other variables in the dataset to make better predictions about the actual potential values of a missing dataset. However, it is impossible to analyze an incomplete dataset or delete missing elements that cause a reduction of the power and yield real problems of wrong decisions that mislead in many fields, and also, using multiple imputation (MI) as a protection tool by masking data utilizing MAR missingness. This essentially means converting complete data to incomplete data MAR mechanism as a security protection method, then reversing it back to complete when needed. Therefore, it is recommended to use the multivariate imputation by chained equations (MICE) approach as imputing missing data and an alternative method to fight against cyber-crime in databases until further research leads to better practices that eliminate biases and maintain consistency unmasking.

Further, the data is the basis of any decision making; any outliers in data sets or missing values during collection and analysis of data sets can be imputed.

Further, the multiple imputation using the MICE algorithm yields the representative data set from uncategorized values missing data sets using the passive imputation technique, which generates consistent imputation between two levels of continuous data.

6. Recommendation

Multivariate imputation for missing data values using the MICE algorithm is the most appropriate form of MAR data until a further breakthrough in missing data handling for MNAR mechanism. Research in multivariate imputation can be done by analyzing accuracy in imputed values with a change in the proportion of missing data in one or multiple columns. Also, the accuracy of imputed values with regression models can be analyzed. More can be carried on the ability of the algorithms to impute between logically regressed variables and linearly regressed variables, to enhance the ways to effectively use multivariate imputations and inspect and improve the quality of imputations using various analysis plots.

Acknowledgements

The authors would like to thank reviewers for their helpful notes and suggestions to this article. Thanks, to Scientific Research/the Open Journal of Statistics for their valuable reviews and publishing.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Huque, M.H., Carlin, J.B., Simpson, J.A. and Lee, K.J. (2018) A Comparison of Multiple Imputation Methods for Missing data in Longitudinal Studies. *BMC Medical Research Methodology*, **18**, 1-16. https://doi.org/10.1186/s12874-018-0615-6

[2] Kontopantelis, E., White, I.R., Sperrin, M. and Buchan, I. (2017) Outcome-Sensitive Multiple Imputations: A Simulation Study. *BMC Medical Research Methodology*, **17**, 1-13. https://doi.org/10.1186/s12874-016-0281-5

[3] Rubin, D.B. (1996) Multiple Imputation after 18+ Years. *Journal of the American Statistical Association*, **91**, 473-489. https://doi.org/10.1080/01621459.1996.10476908

[4] Little, R.J.A. and Rubin, D.B. (2002) Statistical Analysis with Missing Data. 2nd Ed., Wiley Interscience, New York. https://doi.org/10.1002/9781119013563

[5] Van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M. and Rubin, D.B. (2006) Fully Conditional Specification in Multivariate Imputation. *Journal of Statistical Computation and Simulation*, **76**, 1049-1064. https://doi.org/10.1080/10629360600810434

[6] Carpenter, J. and Kenward, M. (2013) Multiple Imputation and Its Application. 1st ed. Wiley, New York.

[7] Rubin, D.B. (1993) Discussion: Statistical Disclosure Limitation. *Journal of Official Statistics,* **9**, 461-468.

[8] Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. Wiley, New York. https://doi.org/10.1002/9780470316696

[9] White, I.R., Royston, P. and Wood, A.M. (2011) Multiple Imputation Using Chained Equations: Issues and Guidance for Practice. *Statistics in Medicine*, **30**, 377-399. https://doi.org/10.1002/sim.4067

[10] Rubin, D.B. (2003) Discussion on Multiple Imputation. *International Statistical Review*, **71**, 619-625. https://doi.org/10.1111/j.1751-5823.2003.tb00216.x

[11] Van Buuren, S. (2010) Multiple Imputation of Multilevel Data. In: Hox, J. and Roberts, K., Eds., *The Handbook of Advanced Multilevel Analysis*, Routledge, Milton Park, UK.

[12] Van Buuren, S. and Oudshoorn, K. (2000) Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual, Volume PG/VGZ/00.038. TNO Prevention and Health, Leiden.

[13] Scheidegger, A. (2012) adaptMCMC: Implementation of a Generic Adaptive Monte Carlo Markov Chain Sampler. R Package Version 1.1.