Communications for Statistical Applications and Methods
2023, Vol. 30, No. 6, 577–587

https://doi.org/10.29220/CSAM.2023.30.6.577
Print ISSN 2287-7843 / Online ISSN 2383-4757

Two-stage imputation method to handle missing data for categorical response variable

Jong-Min Kim[a], Kee-Jae Lee[b], Seung-Joo Lee[1,c]

[a]Statistics Discipline, Division of Science and Mathematics, University of Minnesota at Morris, MN, USA; [b]Department of Information Statistics, Korea National Open University, Seoul, Korea; [c]Department of Data Science, Cheongju University, Chungbuk, Korea

Abstract

Conventional categorical data imputation techniques, such as mode imputation, often encounter issues related to overestimation. If the variable has too many categories, multinomial logistic regression imputation method may be impossible due to computational limitations. To rectify these limitations, we propose a two-stage imputation method. During the first stage, we utilize the Boruta variable selection method on the complete dataset to identify significant variables for the target categorical variable. Then, in the second stage, we use the important variables for the target categorical variable for logistic regression to impute missing data in binary variables, polytomous regression to impute missing data in categorical variables, and predictive mean matching to impute missing data in quantitative variables. Through analysis of both asymmetric and non-normal simulated and real data, we demonstrate that the two-stage imputation method outperforms imputation methods lacking variable selection, as evidenced by accuracy measures. During the analysis of real survey data, we also demonstrate that our suggested two-stage imputation method surpasses the current imputation approach in terms of accuracy.

Keywords: missing data, variable selection, imputation, categorical variable

1. Introduction

When collecting data through survey methods for politics, economics, education and health, it is expected to encounter missing data in categorical variables. Missing data can arise due to non-response, which is a common issue encountered in survey methods. Ignoring missing data can have negative consequences, including reduced statistical power and unreliable results and interpretations. Recently, text based data analysis methods have been a popular research area due to the rapid development of modern technologies, such as artificial intelligence, semiconductor, smartphone, and ChatGPT. Currently, we are facing more issues with missing data for categorical variables rather than missing data for quantitative variables. To handle missing data for categorical variable, we need to know some key considerations and approaches for missing data mechanisms because the choice of appropriate methods for handling missing data can affect the result of missing data analysis. Generally, when there is missing data, researchers choose easy way to exclude any observations with missing values and analyze only the complete cases. However, this easy approach may lead to biased results if missingness is related to the variables of interest. So imputation techniques are commonly used to fill in missing values with plausible estimates to ensure reliable and valid statistical analysis and interpretation of results in various research domains. The most common mechanisms to impute missing

[1] Corresponding author: Department of Data Science, Cheongju University, 298 Daeseong-ro, Cheongwon-gu, Cheongju-si, Chungcheongbuk-do 28503, Korea. E-mail: access@cju.ac.kr

Published 30 November 2023 / journal homepage: http://csam.or.kr

data include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Rubin (1976) differentiated between three types of missingness mechanisms. First is MCAR, where cases with missing values are considered a random sample of all cases. Second is MAR, which does not depend on missing variables and can be modeled using observed data. Specialized missing data analysis methods can then be applied to correct for the effects of missingness. Third is MNAR, which occurs when data is neither MCAR nor MAR. Multiple imputation methods, such as multiple imputation by chained equations (MICE), are often employed for handling missing categorical data. These methods create multiple imputed datasets based on the observed data and impute missing values using appropriate models for each variable. In addition, advanced imputation methods such as latent variable models, pattern mixture models, or selection models have been developed to handle missing categorical data. But these advanced methods need to understand underlying missing data mechanisms comprehensively to perform sophisticated modeling. Pioneering research works dealing with missing continuous data have been studied. Rubin (1976) noticed that when making sampling distribution inferences about the parameter of the data, θ, it is appropriate to ignore the process that causes missing data if the missing data are 'missing at random' and the observed data are 'observed at random', but these inferences are generally conditional on the observed pattern of missing data. Sande (1979) described that a Hot Deck imputation procedure is defined to be one where an incomplete response is completed by using values from one or more other records on the same file and the choice of these records varies with the record requiring imputation. Sande (1979) mentioned that Hot Deck procedures should be evaluated to study the bias and reliability of the estimates, donor usage and frequency of imputation failure in terms of a variety of conditions of the data and variations of the imputation procedure. When a data set contains imputed values, special care is needed in studying the interrelationships between variables, whether the interrelationships are examined in terms of cross-tabulations, regression analyses or other forms of multivariate analysis Kalton and Kasprzyk (1982). Kalton and Kish (1984) examined the effect of imputation variance on the precision of the mean, and propose four procedures for sampling the respondents that reduce this additional variance. A compromised imputation procedure has been suggested by Singh and Horn (2000). The estimator of mean obtained from compromised imputation remains better than the estimators obtained from ratio method of imputation and mean method of imputation. Singh and Deo (2003) proposed a new power transformation estimator of population mean in the presence of non-response. The estimator of mean obtained from proposed technique remains better than the estimators obtained from ratio or mean methods of imputation. An alternative estimator of population mean in the presence of non-response has been suggested which comes in the form of Walsh's estimator by Singh (2009). The estimator of mean obtained from the proposed technique remains better than the estimators obtained from ratio or mean methods of imputation. Kim and Shao (2013) explained statistical methods for handling incomplete data well. So extensive research has been conducted on imputation methods over the past several decades. However, imputation for missing categorical data has received less attention among survey statistics researchers. The most common used imputation method for both continuous and categorical variables is mean, median, mode imputation methods. The mean, median and mode imputation methods provide a fast and easy way of obtaining complete datasets. Especially, when a large number of missing values exist, mode imputation may lead to an over-representation of the most frequent label. To deal with categorical missing data, multinomial logistic regression imputation is the preferred method whenever it is computationally feasible. However, if the variable has too many categories, using this method may be impossible due to computational limitations. To rectify this computation feasibility issue, we propose a two-stage imputation method combining the Bourta algorithm variable selection method by Kursa and Rudnicki (2010) and multivariate imputation via chained equations (MICE) method by van Buuren and Groothuis-Oudshoorn (2011).

The remainder of this paper is structured as follows: In Section 2, we provide a description of the variable selection method (Bourta algorithm), the MICE method, and our proposed two-stage imputation approach that integrates both the Bourta algorithm for variable selection and MICE. Moving on to Section 3, we will apply the Clayton copula asymmetric simulated data to both the MICE imputation and the two-stage imputation method. This is done to demonstrate the superiority of our proposed imputation approach through a simulation study. In Section 4, we will further illustrate the superiority of the two-stage imputation method using real-world data. Finally, Section 5 will present our conclusions.

2. Statistical methods

This research proposes a two-stage imputation method for a categorical response variable with multiple explanatory variables. In the first stage, we apply the variable selection method (Bourta Algorithm) to the entire dataset, including the missing data. After identifying the important variables through this selection, we proceed to the second stage where we apply multivariate imputation via chained equations (MICE) to handle the missing data. Thus, in this section, we introduce the variable selection method (Bourta Algorithm) and the multivariate imputation via chained equations (MICE) method.

We use the Bourta variable selection in this research because 'Boruta' R Package takes into account multi-variable relationships and it considers all features which are relevant to the outcome variable. The Boruta algorithm is a wrapper built around the random forest classification algorithm to automatically perform feature selection on a dataset. It was born as a 'Bourta' package for R and a version of Boruta for Python called BorutaPy. Boruta algorithm performs shuffling of predictors' values and join them with the original predictors and then build random forest on the merged dataset. In the next step, Boruta algorithm compares the original variables with the randomised variables to measure variable importance. Only variables having higher importance than that of the randomised variables are considered important, this is called a hit. The idea is that a variable is useful only if it's capable of doing better than the best randomized variable. It tries to capture all the important, interesting variables we might have in our dataset with respect to an outcome variable. In wrapper methods, we try to use a subset of variables and train a model using them. Based on the inferences that we draw from the previous model, we decide to add or remove features from the subset. Forward Selection, Backward elimination are some of the examples for wrapper methods. So we will apply the Bourta variable selection method to simulated data and real data in the following Sessions.

In the second stage for the two-stage imputation for categorical response variable, we will employ MICE after the Bourta variable selection method in the first stage. MICE is one of the most popular R package which can impute mixes of continuous, binary, unordered categorical and ordered categorical data. In this research, we only consider MAR case which is the probability that a value is missing depends only on observed value and can be predicted using them. MICE can create multiple imputations for multivariate missing data and is based on fully conditional specification, where each incomplete variable is imputed by a separate model.

van Buuren (2018) summarized categorical data and imputation. Imputation of missing categorical data is possible under the broad class of generalized linear models in McCullagh and Nelder (1989). For incomplete binary variables we employ logistic regression, where the outcome probability is modeled as

$$P\left(y_i = 1 \mid X_i, \beta\right) = \frac{\exp\left(X_i\beta\right)}{1 + \exp\left(X_i\beta\right)}.$$

A categorical variable with K unordered categories is imputed under the multinomial logit model

$$P\,(y_i = k \mid X_i, \beta_k) = \frac{\exp\,(X_i\beta_k)}{\sum_{k=1}^{K} 1 + \exp\,(X_i\beta_k)},$$

for $k = 1, \ldots, K$, where β_k varies over the categories and where $\beta_1 = 0$ to identify the model. A categorical variable with K ordered categories is imputed by the ordered logit model, or proportional odds model

$$P\,(y_i \leq k \mid X_i, \beta, \tau_k) = \frac{\exp\,(\tau_k - X_i\beta)}{1 + \exp\,(\tau_k - X_i\beta)},$$

with the slope β is identical across categories, but the intercepts τ_k differ. For identification, we set $\tau_1 = 0$. The probability of observing category k is written as

$$P\,(y_i = k \mid X_i) = P\,(y_i \leq k \mid X_i) - P\,(y_i \leq k - 1 \mid X_i)\,,$$

where the model parameters β, τ_k, and τ_{k-1} are suppressed for clarity.

A mice command in 'mice' R pacakge has the predictive mean matching method in case of numeric data, logistic regression imputation in case of binary data or a factor with 2 levels, and polytomous regression imputation for unordered categorical data in case that factor has more than 2 levels. We also set the number of multiple imputations to be five for data analysis in this paper.

So we propose the two-stage imputation method. In the first stage, we perform Boruta variable selection method with the incomplete full data to select important variables for the targeted categorical variable. In the second stage, we apply MICE method to the targeted categorical response variable with the important selected variables. In detail, we impute missing data in a binary variable using logistic regression for a targeted binary variable, impute missing data in a targeted categorical variable using polytomous regression, and impute missing data in a targeted quantitative variable using predictive mean matching method with the selected important variables to the targeted categorical or quantitative variables.

Our proposed two-stage imputation method involves using the Boruta variable selection method and multiple imputation by chained equations (MICE) to handle missing data in a categorical response variable. Here's a breakdown of the steps involved in each stage:

In the first stage (Boruta variable selection method), we start with the incomplete full data, which contains missing values for the categorical response variable. We apply the Boruta variable selection method to identify important variables related to the targeted categorical variable. Boruta is a feature selection algorithm that uses a random forest model and a permutation-based feature importance measure. The Boruta method ranks the variables based on their importance and selects the significant variables for further analysis.

In the second stage (MICE), we take the important variables selected in the first stage, along with the incomplete data. We perform MICE, a multiple imputation technique, to impute the missing values in the targeted categorical response variable. For binary variables, impute missing data using logistic regression. For categorical variables with more than two categories, impute missing data using polytomous regression. For quantitative variables, impute missing data using the predictive mean matching method. This involves finding similar observations with complete data and using their values to impute the missing values. The two-stage imputation method combines variable selection with imputation techniques to handle missing data in a targeted categorical response variable. The Boruta method helps identify important variables, and then MICE is used to impute the missing values based on the selected variables and the type of variable being imputed (binary, categorical, or quantitative).

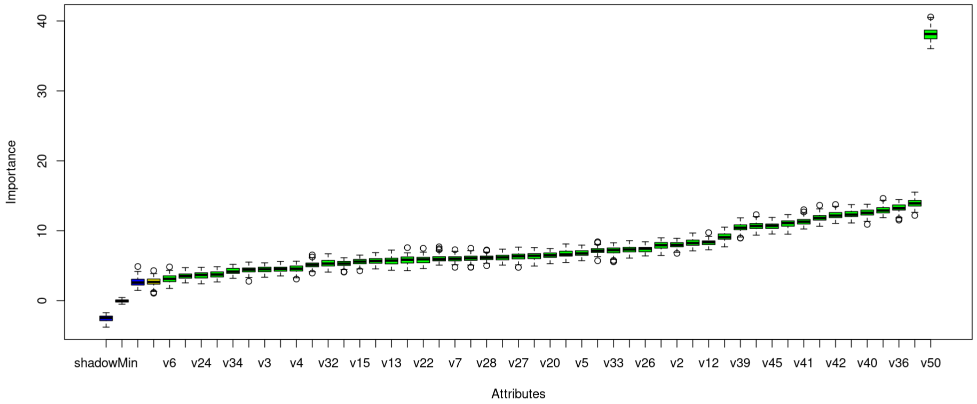

Figure 1: *Variable selection with 50 variables and 40% missing rate.*

In other words, Kim, Lee, and Lee's two-stage imputation method for a categorical response variable with multiple explanatory variables will be verified by performing the procedures 1,000 times. This process will produce a confusion matrix, which calculates a cross-tabulation of observed and predicted classes along with associated statistics. We will then calculate the average accuracy from 1,000 repetitions using this confusion matrix.

3. Simulation study

In a simulation study session, we aim to generate categorical incomplete data that is non-normal, highly correlated, and mimics real categorical data. To achieve this, we use the Clayton copula, which allows for the generation of highly correlated and non-normally distributed input variables and count response variables. The use of Clayton copula helps relax the assumptions of normality, linearity, and independence, allowing for more realistic modeling of the marginal behavior and joint dependence of random variables.

Here's a brief overview of the steps involved in generating data using the Clayton copula: We determine the number and types of input variables we want to generate. These could be categorical variables or variables that can be transformed into categorical variables. We specify the desired correlation structure among the input variables. The Clayton copula allows you to model dependence while considering the skewness and tail behavior typically found in categorical data. We generate random observations from the Clayton copula distribution. The generated values will represent the joint dependence structure of the input variables. We transform the generated observations into categorical data as needed. we can assign categories based on thresholds or other criteria, depending on the nature of your categorical data. We repeat the above steps to generate multiple instances of the incomplete categorical data for your simulation study. By employing the Clayton copula, we can generate highly correlated and non-normally distributed categorical data that better captures the characteristics of real-world categorical data, including left-skewness and lower tail dependence.

For any chosen copula, the MLE that maximizes the preceding formula is denoted by $(\hat{\mu}, \hat{\sigma}, \hat{\alpha})$.

Table 1: Summary statistics of accuracy with (Var.Sel.W) and without variable selection (Var.Sel.WO) (simula ted data)

Missing rate	40 Percent		50 Percent		60 Percent	
Variable selection	Var.Sel.WO	Var.Sel.W	Var.Sel.WO	Var.Sel.W	Var.Sel.WO	Var.Sel.W
Mean	0.9985	0.9985	0.9981	0.9981	0.9978	0.9978
Standard error	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Median	0.9985	0.9985	0.9981	0.9981	0.9978	0.9978
Mode	0.9985	0.9985	0.9981	0.9981	0.9978	0.9978
Standard deviation	0.0002	0.0002	0.0002	0.0002	0.0002	0.0002
Sample variance	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Skewness	−0.1693	−0.0743	−0.2143	−0.1410	−0.0247	0.0158
Range	0.0011	0.0009	0.0012	0.0011	0.0014	0.0012
Minimum	0.9979	0.9980	0.9975	0.9976	0.9970	0.9972
Maximum	0.9990	0.9989	0.9987	0.9986	0.9985	0.9984

Under the Clayton copula, the log-copula density is

$$\begin{aligned}\log c\left(u_1, u_2; \alpha\right) &= \log\left(1+\alpha\right) - \left(1+\alpha\right)\log u_1 - \left(1+\alpha\right)\log u_2 \\ &\quad - \left[\frac{1}{\alpha} + 2\right]\log\left(u_1^{-\alpha} + u_2^{-\alpha} - 1\right).\end{aligned}$$

So we used the following R commands in copula R package such as claytonCopula (θ = 8, dim = p) for correlated and non-normal simulated dataset, and normalCopula (param = 0.8, dim = p) for correlated and normal simulated dataset, where θ is clayton copula parameter, and p = 50 is the total number of variables. For converting thirteen uniform random variables to binary or categorical data where each variable has thousand data size, we set up the first twenty variables to be binary data by splitting less than 0.5 to be zero and more than 0.5 to be one, the next fifteen variables to be categorical data by splitting less than 0.333 to be zero, more than 0.333 and less than 0.666 to be one, and more than 0.666 to be two, and the remaining fifteen variables to be categorical data by splitting less than 0.25 to be zero, more than 0.25 and less than 0.50 to be one, more than 0.50 and less than 0.75 to be two, and more than 0.75 to be three. To get robust simulation result, we do one thousand number of iterations to simulate data in this setup.

With 'mice' R package, we generate MAR missing values with 'ampute' command with missing proportions, 40%, 50%, and 60% from the asymmetric simulated data generated by Clayton copula function. In order to test their predictive ability, we used 'caret' R packages and created confusion matrix by using 'confusionMatrix' command. It was used to generate 1,000 data sets randomly. The prediction accuracy of these regular mice imputation method and two stage imputation method was then compared by finding their average accuracy across all 1,000 data sets. These values were found using the following formulas:

$$\begin{aligned}\text{Accuracy} &= \tfrac{\text{TP+TN}}{\text{TP+TN+FP+FN}}, \\ \text{Sensitivity} &= \tfrac{\text{TP}}{\text{TP+FN}}, \\ \text{Specificity} &= \tfrac{\text{TN}}{\text{TN+FP}},\end{aligned} \tag{3.1}$$

where TP (true positive) = the number of cases correctly predicted 'Yes', TN (true negative) = the number of cases correctly predicted 'No', FP (false positive) = the number of cases incorrectly predicted 'Yes', and FN (false negative) = the number of cases incorrectly predicted 'No'.

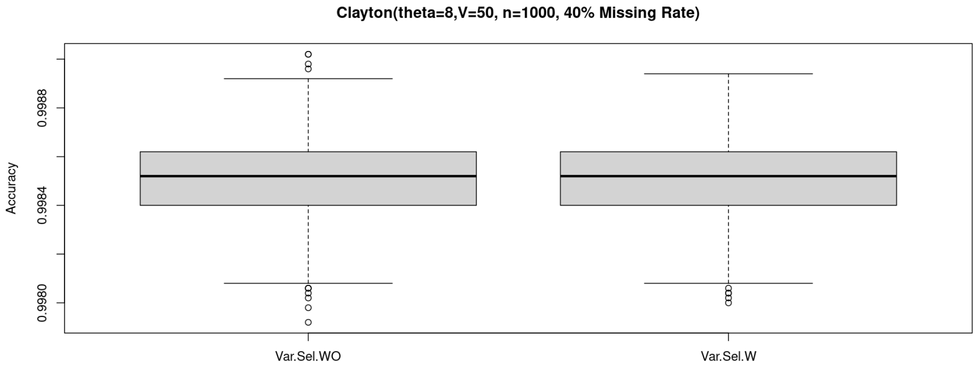

Figure 2: *Accuracy with and without variable selection method for 40% missing rate.*

Since we are imputing a count response, measuring accuracy is challenging. We present confusion matrices to summarize overall performance. We use accuracy, sensitivity and specificity to compare imputed values to actual values for each model.

Figure 1 is variable selection with 50 explanatory variables and 40% missing rate.

Table 1 shows that the two-stage imputation method is roughly equivalent or slightly superior to the mice imputation method in the case of non-normally distributed lower-tailed asymmetric simulation data with the Clayton copula. This comparison is based on key accuracy metrics such as the maximum, median, mean, and interquartile range (IQR). Figure 2 confirms the result of Table 1.

4. Illustrated data analysis

The national center for health statistics (NCHS) of the centers for disease control and prevention (CDC) collects, analyzes, and disseminates data on the health status of U.S. residents. The national health and nutrition examination survey (NHANES) is a periodic survey conducted by NCHS. The third national health and nutrition examination survey (NHANES III), conducted from 1988 through 1994, was the seventh in a series of these surveys based on a complex, multi-stage sample plan.

Since each individual in the sample design does not share the same probability of selection, various aspects of the NHANES design must be considered during data analysis. This includes accounting for sample weights and the complexities of the survey design. Appropriate sample weights are needed to estimate prevalence, means, medians, and other statistics. Sample weights incorporate varying probabilities of selection and encompass adjustments for noncoverage and nonresponse. For a detailed survey description, we recommend visiting the website `https://www.cdc.gov/nchs/data/nhanes/nhanes3/nh3gui.pdf`.

The first stage of the design consisted of selecting a sample of 81 primary sampling unit (PSU)'s that were mostly individual counties. The PSU's were stratified and selected with probability proportional to size (PPS).

The 89 locations were randomly divided into two groups, one for each phase. The first group consisted of 44 and the other of 45 locations. One set of PSU's was allocated to the first three-year survey period (1988-91) and the other set to the second three-year period (1991-94). Therefore, unbiased estimates of health and nutrition characteristics can be independently produced for both Phase 1 and Phase 2 as well as for both phases combined.

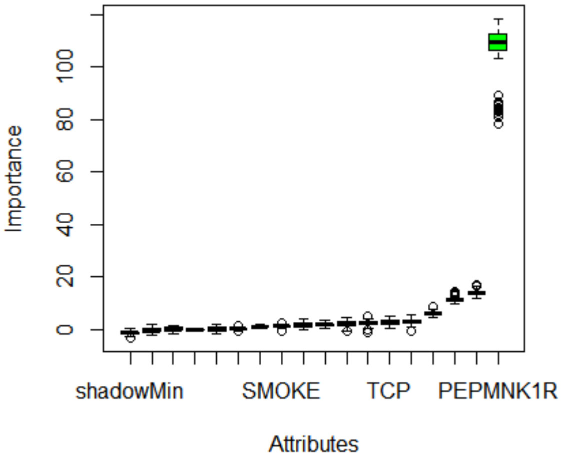

Figure 3: *Bourta variable selection for high blood pressure with 10% missing rate.*

We downloaded the NHANES III data with 17,030 observations and 16 variables from Hosmer and Lemeshow (2020)'s textbook. A subset of data from the National Health and Nutrition Examination Study (NHANES) III. Subjects age ≥ 20 are included. 49 pseudo strata were created with 2 pseudo-PSU's in each stratum. In this dataset, there are about 17,030 individuals, which are described by 16 variables. But 9,103 were excluded because they were missing one or more points for 16 variables. So 7,927 individuals are the number of complete dataset. In this paper, we treat 7,927 individuals as a sample for the population. This sample classifies people described by a set of attributes as whether smoking influence people or not i.e whether the individual got high blood pressure or not caused by smoking. In this research, we select fifteen variables including SDPPSU6, SDPSTRA6, HSSEX, DMARACER, WTPFHX6, HSAGEIR, BMPWTLBS, BMPHTIN, PEPMNK1R, PEPMNK5R, HAR1, TCP, HAR3, SMOKE, HBP in Table 2 By applying MAR 10%, 20%, 30%, 40% missing rates to the complete dataset, we intentionally make missing dataset and then impute missing variables by using 'logreg', 'polyreg', 'pmm' commands in mice R package.

Figures 3, 4 and 5 show the bourta variable selection plots for high blood pressure (HBP) response variable with fourteen explanatory variables (SDPPSU6, SDPSTRA6, HSSEX, DMARACER, WTPFHX6, HSAGEIR, BMPWTLBS, BMPHTIN, PEPMNK1R, PEPMNK5R, HAR1, TCP, HAR3, SMOKE) in the three 10%, 20%, 30 % missing rate cases. And then we performed the accuracies of two imputation techniques with or without variable selection method. By looking at Table 3, the imputation technique with variable selection method had better accuracy than imputation technique without variable selection method. In real data analysis, we have included continuous and categorical variables for explanatory variable and binary variable for response variable, which is different from the previous simulation study session where response and explanatory variables are binary and categorical variables.

This real data analysis serves as an illustration of the superiority of the two-stage imputation method over the mice imputation method. In our future studies, we plan to conduct extensive real data

Table 2: NHANES III Data set

Variable	Description	Codes/Values	Name
1	Respondent ID	Number	SEQN
2	Pseudo-PSU	1,2	SDPPSU6
3	Pseudo-stratum	01 - 49	SDPSTRA6
4	Statistical weight	225.93 - 139744.9	WTPFHX6
5	Age	years	HSAGEIR
6	Sex	0 = Female, 1 = Male	HSSEX
7	Race	1 = White, 2 = Black, 3 = Other	DMARACER
8	Body Weight	pounds	BMPWTLBS
9	Standing Height	inches	BMPHTIN
10	Average Systolic BP	mm Hg	PEPMNK1R
11	Average Diastolic BP	mm Hg	PEPMNK5R
12	Has respondent smoked > 100 cigarettes in life	1 = Yes, 2 = No	HAR1
13	Does repondent smoke cigarettes now?	1 = Yes, 2 = No	HAR3
14	Smoking	1 = if HAR1 = 2 2 = if HAR1=1 & HAR3=2 3 = if HAR1=1 & HAR3=1	SMOKE
15	Serum Cholesterol	mg/100ml	TCP
16	High Blood Pressure	0 if PEPMNK1R ≤140 1 if PEPMNK1R > 140	HBP

Table 3: Real data prediction accuracy of binary variable (HBP) with variable selection (Var.Sel.W) and without variable selection (Var.Sel.WO)

Missing Rate		**10%**	**20%**	**30%**	**40%**
Accuracy	**Var.Sel.WO**	0.9508	0.9163	0.8909	0.8619
	Var.Sel.W	0.9646	0.9404	0.9153	0.8717

analyses to further demonstrate the effectiveness of our two-stage imputation method.

Through simulation and real data analysis, we can say that our proposed imputation method with variable selection works well for the case of a categorical response variable of the mixed continuous and categorical explanatory variables.

5. Conclusion

The simulation studies demonstrated that our proposed two-stage imputation method exhibited slightly better accuracy compared to the current imputation method when applied to the Clayton copula asymmetric simulated dataset. Additionally, we conducted a comparison of imputation techniques with and without the variable selection method, utilizing data from the National Health and Nutrition Survey. This comparison was performed considering missing-at-random (MAR) rates of 10%, 20%, 30%, and 40%. The results of our real data analysis revealed that our proposed two-stage imputation method significantly outperformed the MICE imputation in terms of accuracy when dealing with real datasets.

In particular, our approach demonstrated the ability to reduce computation time for handling missing data in categorical response variables. This efficiency was achieved by employing a reduced set of important variables for imputing the target categorical variable. Looking ahead, our future research aims to extend the application of the two-stage imputation method to longitudinal (panel) data analysis

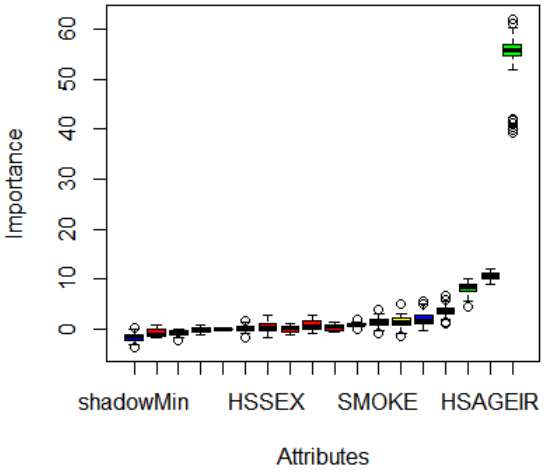

Figure 4: *Bourta variable selection for high blood pressure with 20% missing rate.*

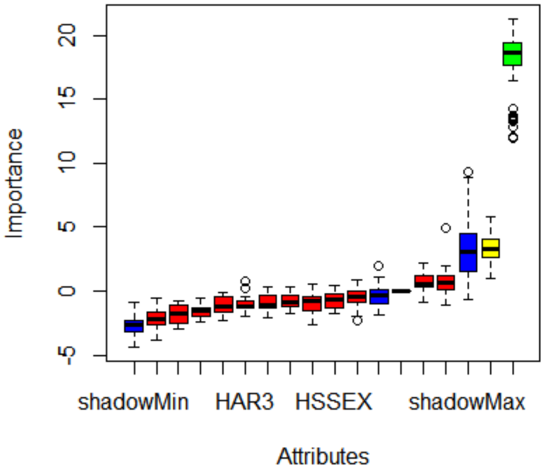

Figure 5: *Bourta variable selection for high blood pressure with 30% missing rate.*

in fields such as clinical studies or labor economics studies. By doing so, we anticipate further reductions in computation time and resource utilization through the utilization of our two-stage imputation method.

Acknowledgment

We thank the two respected referees, Associated Editor and Editor for constructive and helpful suggestions which led to substantial improvement in the revised version.

References

Hosmer DW and Lemeshow S (2020). *Applied Logistic Regression*, John Wiley & Sons, Inc Hoboken, New Jersey.

Kalton G and Kasprzyk D (1982). Imputation for missing surveys responses, *Proceedings of the Survey Research Methods Section, American Statistical Association*, 22–31.

Kalton G and Kish L (1984). Some efficient random imputation methods, *Communications in Statistics-Theory and Methods*, **13**, 1919–1939.

Kim JK and Shao J (2013). *Statistical Methods for Handling Incomplete Data*, CRCpress, Boca Raton, FL.

Kursa MB and Rudnicki WR (2010). Feature Selection with the Boruta Package, *Journal of Statistical Software*, **36**, 1–13.

McCullagh P and Nelder JA (1989). *Generalized Linear Models* (2nd ed), Chapman & Hall, New York.

Rubin RB (1976). Inference and missing data, *Biometrika*, **63**, 581–592.

Sande IG (1979). A personal view of hot-deck imputation procedures, *Survey Methodology Statistics Canada*, **5**, 238–247.

Singh S and Horn S (2000). Compromised imputation in survey sampling, *Metrika*, **51**, 267–276.

Singh S and Deo B (2003). Imputation by power transformation, *Statistical Papers*, **44**, 555–579.

Singh S (2009). A new method of imputation in survey sampling, *Statistics*, **43**, 499–511.

van Buuren S and Groothuis-Oudshoorn K (2011). Mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software*, **45**, 1–67.

van Buuren S (2018). *Flexible Imputation of Missing Data* (2nd ed), Chapman & Hall/CRC, Boca Raton, FL.

Received August 26, 2023; Revised September 16, 2023; Accepted September 18, 2023