International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 51–62

Composite Imputation Method for the Multiple Linear Regression with Missing at Random Data

Thidarat Thongsri[1,2], **Klairung Samart**[1,2]

[1]Division of Computational Science
Faculty of Science
Prince of Songkla University
Songkhla, Thailand

[2]Statistics and Applications Research Unit
Faculty of Science
Prince of Songkla University
Songkhla, Thailand

email: 6010210112@psu.ac.th, klairung.s@psu.ac.th

(Received May 16, 2021, Accepted July 6, 2021)

Abstract

Missing data is a common occurrence in the data collection process. If this problem is ignored it can lead to unreliable conclusions. Our research objective is to develop a method for handling missing data in multiple linear regression at random on both response and independent variables and to compare its efficiency with existing techniques. For handling missing data, five so-called techniques were employed; namely, listwise deletion (LD), hot deck imputation (HD), predictive mean matching imputation (PMM), stochastic regression imputation (SR), and random forest imputation (RF). We compare them with the following proposed composite imputation method: stochastic regression random forest with equivalent weight (SREW).

SREW is derived from a combination of stochastic regression and random forest methods weighted by the equivalent weighted method. In

Key words and phrases: Composite method, imputation method, missing data, missing at random, multiple linear regression.

AMS (MOS) Subject Classifications: 62J05.

Klairung Samart is the corresponding author (klairung.s@psu.ac.th).

ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

this study, the Monte Carlo simulations were done under the sample sizes of 30, 60, 90, 120 and 150 along with the missing percentages of 10%, 20%, 30% and 40% and the standard deviations of error of 1, 3 and 5. The criterion to compare the efficiency is the average mean square error (AMSE). The results show that the SREW is most efficient in all situations whereas the hot deck gives the highest AMSE in almost all cases, especially when the missing percentage is high.

1 Introduction

Multiple linear regression is a statistical tool used to investigate the relationship between response and independent variables. A problem that usually occurs during data collection is missing data. This is an important issue to pay attention to because ignoring it can lead to unreliable conclusions. Missing data can occur because of nonresponse by refusals, miscommunication, lack of information, privacy, loss of questionnaires or other reasons and sometimes are caused by the researchers themselves.

Different types of missing data need different techniques. Rubin [16] and Little and Rubin [12] prescribed three missing data mechanisms:
Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). If a missing value is totally random and does not depend on any variable, then it is classified as MCAR. This type increases standard errors since the sample size decreases but does not cause bias [9]. On the other hand, if a missing value of any of the variable in the dataset depends on observed values of other variables in the dataset, then it is defined as MAR. However, if missing data depends on unobserved data rather than observed data, then it is known as MNAR [8].

In the literature, several techniques have been proposed to handle missing data. Among these, there are two strategies for handling missing data [6]; namely, ignoring missing values and imputation of missing values. The first method simply excludes the cases that contain missing data. As it is the simplest method for handling missing data without deep knowledge in statistics, it is widely used in general [8]. However, the disadvantage of this method is that the sample size decreases and therefore results in biased estimate and loss of precision [18]. Thus, this method is appropriate for dataset with small proportion of missing values.

A more attractive approach is imputation of missing values. This method is merely estimation of plausible values to substitute the missing spots. The advantage of the imputation method is to reduce the bias due to missingness rather than deleting incomplete cases [8]. Researchers have developed several imputation techniques such as hot deck (HD), predictive mean matching (PMM), stochastic regression (SR), and random forest (RF) [17], [10], [11], [3], [12], [8].

Imputation methods have been studied extensively in several research articles. For example, Tang and Ishwaran [20] revealed that random forest (RF) imputation performs better with increasing correlation. Performance was good under moderate to high missing data, and when data was missing not at random. Aguilera, Guardiola-Albert and Serrano-Hidalgo [1] compared three techniques; namely, spatio-temporal kriging (STK), predictive mean matching (PMM), and the random forest for imputing large amounts of missing daily precipitation data. The results showed that both STK and RF can handle extreme missing data, while PMM requires larger observed sample sizes whereas RF is efficient under random missing patterns. Moreover, it turned out that the polytomous regression and hot deck imputations are also more effective than other simple methods [4], [13].

In this paper, we propose a composite imputation method: stochastic regression random forest with equivalent weight (SREW) which is a combination of stochastic regression and random forest methods with equivalent weight. We compare this composite method with five methods; namely, listwise deletion (LD), hot deck imputation (HD), predictive mean matching imputation (PMM), stochastic regression imputation (SR), and random forest imputation (RF). Our results will help practitioners to make decision in selecting an appropriate method for missing values imputation.

The rest of this paper is organized as follows: In Section 2, we present the five techniques used for comparison and introduce the composite imputation method. In Section 3, we reveal the performance of all methods via a simulation study under different conditions. In Section 4, we conclude our paper with a brief discussion of the results.

2 Techniques for handling missing Data

Let Y_i be a random variable, where $i = 1, 2, ..., n$ with m missing values and $X_{i1}, X_{i2}, ..., X_{iq}$ be independent variables. The methods used in this work for handling missing data are as follows:

2.1 Listwise deletion

The Listwise deletion technique simply omits cases with missing data; i.e., analysis of the data is restricted to the complete cases. This technique has been widely used in practice since it is easy to implement and is the default in many statistical packages [14]. If the data are MCAR, then this technique will give unbiased parameter estimates. However, the standard errors will be larger since the sample size decreases. When data are MAR, listwise deletion may lead to biased estimates; that is, regression coefficients might be too large or too small. Moreover, this technique can mislead the results if a large proportion of the missing values is excluded [15].

2.2 Hot deck imputation

Hot deck imputation is a method for handling missing data in which each missing value (the recipient) is replaced with an observed response from a similar unit (the donor) [8]. There are two different forms of the hot deck method. If the donor is selected randomly from a set of potential donors, then it is called "random hot deck method". If a single donor is identified base on some metric and values are imputed from that case, then it is called "deterministic hot deck method" [8].

There are some attractive features of the hot deck. First, it does not rely on model fitting for the variable to be imputed, and thus is potentially less sensitive to model misspecification than an imputation method based on a parametric model [8]. Another advantage is that only the plausible values are imputed since they come from observed responses in the donor pool [8]. This feature cannot be guaranteed by most of the other methods. As a result, various government statistics agencies and survey organizations such as the U.S. and the British census, the current population survey, the Canadian census of Construction, the U.S. Annual survey of Manufacturers, and the U.S. National Medical Care utilization and Expenditure survey use this technique in practice [14].

2.3 Predictive mean matching imputation

Predictive mean matching calculates missing values of a particular variable by sampling from the observed values of all complete cases (donors) of that variable and match predicted values as closely as possible [8]. One donor is then randomly drawn from the candidates, and the missing value is replaced by the observed value of the donor. The assumption of this method is the distribution of the missing cell has to be the same as the observed data of the candidate donors (van Buuren, 2018). As a result, the imputed values by this method are obviously realistic since they are based on only observed values. Imputations outside the observed data range will not definitely occur, therefore avoiding problems with meaningless imputations (van Buuren, 2018).

The predictive mean matching algorithm can be divided into 6 steps as follows [22].

1. Estimate a linear regression model from complete cases of all variables. The variable we want to impute is treated as Y and the other variables are treated as X.
2. Draw randomly from the posterior predictive distribution of $\hat{\beta}$ obtained in step 1 and form a new set of coefficients β^*.
3. Use $\hat{\beta}$ to calculate predicted values for observed y and β^* to calculate predicted values for missing y.
4. For each missing $y_i, i = 1, ..., m$, find the closest (typically three) predicted values among the observed $y_j, j = m + 1, ..., n$.
5. Choose randomly one of the three close case and impute the missing value y_i with the observed value of this close case.
6. Repeat steps 1-5 several times for multiple imputation.

2.4 Stochastic regression imputation

Stochastic regression imputation is an extension of regression imputation, where the imputed value is estimated from a regression equation obtained from the observed values. That is, a regression equation is constructed using the observed data y^* on x^* and then will be used to predict the missing values on y. The regression equation is given by

$$\hat{y}^* = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + ... + \hat{\beta}_p x_p^* \tag{2.1}$$

From (2.1), it is obvious that the imputed values will fall on a regression line implying that a correlation between the predictors and the missing outcome variable is one which is impossible in reality. To make it possible, a residual term is added to the imputed value. This is called a stochastic regression imputation.

The residual that is added to the predicted value is drawn from a normal distribution with a mean of zero and a variance equal to the residual variance from the regression of the predictor on the outcome. This is to preserve the variability in the data. Also, parameter estimates are unbiased with MAR data [5].

2.5 Random forest imputation

Random forest is an imputation method based on machine learning algorithms designed for mixed continuous and/or categorical data in the presence of complex interactions and non-linearity without requiring to specify the distributions of the variables [19], [21], [7]

According to Hong and Lynn [7], the random forest algorithm starts with replacing the missing values with the mean (continuous variable) or mode (categorical variable) of that variable. Then the observations in the dataset are divided into two parts, the observed and missing observations. The observed observations are used as the training set, and the missing observations are used as the prediction set. The random forest model is then constructed using the variable under imputation as the response. The missing part of the variable under imputation is then replaced by predictions from random forest models.

This process of imputation repeats several times to obtain better data and will stop once it reaches a stopping criteria or after a certain number of iterations has elapsed. Generally, datasets become well imputed after four to five iterations. However, that might depend on the size and amount of missing data as well [7].

3 Stochastic regression random forest with equivalent weight (SREW)

SREW is a proposed method which is a combination of stochastic regression and random forest imputations. The two methods are chosen from our preliminary study and perform well compared to the other methods. Let $\hat{y}_{SR}$ denote the stochastic regression imputed value and let $\hat{y}_{RF}$ denote the random forest imputed value. Then the SREW imputed value is given by

$$\hat{y}_{SREW} = W_M(\hat{y}_{SR} + \hat{y}_{RF}), \tag{3.2}$$

where the equivalent weight $W_M = \dfrac{1}{M}$ and M is the number of combination methods.

4 Simulation study

In this section, we illustrate the simulation studies used to compare the performances of the six methods for handling missing data described in the previous section. The assessment of the six methods was based on average mean squared error (AMSE) of the estimates from multiple linear regressions.

In the simulation study, a random sample of sizes n =30, 60, 90, 120 and 150 was generated and the values of two independent variables were independently drawn from a uniform distribution where $X_1 \sim U(3,5)$ and $X_2 \sim U(1,10)$. The corresponding values of Y are then given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i; \quad i = 1, 2, ..., n,$$

where the true value of the coefficients $\beta_0 = \beta_1 = \beta_2 = 1$ and the random error term ε_i were set to be randomly generated from a normal distribution with the mean of 0 and three different standard deviations (1,3 and 5) .

For each sample size, the missing data was generated on the dependent variable Y_i and independent variables X_{i1} and X_{i2} using the MAR mechanism. The proportions of missing data were $10\%, 20\%, 30\%$ and 40%. Then, the six methods were applied to handle the missing data and the regression coefficient estimates were obtained. The simulation process was replicated N = 5,000 times. The average mean squared error (AMSE) of the six methods were then computed. All simulations were conducted by using R software.

Tables 1-3 show the AMSEs of the estimates obtained by different methods described in the previous section varied in different sample sizes, missing percentages, and the standard deviations of the error term, respectively.

Overall, the results set out in Tables 1-3 reveal that the estimates obtained by the SREW method outperform other methods in all situations whereas the hot deck method does not perform well especially when the missing proportion is large. By considering Table 1, we can see that when the variability of the data is low (sd=1) with small sample size (less than 100), the stochastic regression method also performs well next below the SREW. However, when the sample size is large, the random forest imputation seems to do better than the stochastic regression imputation. This result also appears in

Tables 2-3 where the variability of the data gets higher, the random forest imputation achieves well compared to the other methods except the SREW.

Table 1: AMSE of the six methods when the standard deviation of the error term is 1

Sample size	Missing data (%)	Methods					
		LD	HD	PMM	SR	RF	SREW
30	10	0.9940	1.1949	0.9249	0.8974	0.9349	**0.8952**
	20	0.9942	1.5662	0.9664	0.9088	0.9821	**0.9021**
	30	0.9946	1.9877	1.0155	0.9166	1.0309	**0.9058**
	40	0.9881	2.4053	1.0650	0.9195	1.0797	**0.9083**
60	10	0.9902	1.1780	0.9561	0.9454	0.9575	**0.9131**
	20	0.9933	1.5233	0.9742	0.9539	0.9744	**0.9039**
	30	0.9966	1.9156	0.9819	0.9472	0.9831	**0.9307**
	40	0.9972	2.3571	1.0058	0.9574	1.0088	**0.9353**
90	10	0.9918	1.1818	0.9733	0.9672	0.9681	**0.9091**
	20	0.9945	1.5056	0.9776	0.9663	0.9694	**0.9237**
	30	0.9957	1.9224	0.9870	0.9680	0.9711	**0.9368**
	40	0.9969	2.3318	0.9930	0.9688	0.9752	**0.9531**
120	10	0.9872	1.1749	0.9786	0.9738	0.9598	**0.9095**
	20	0.9896	1.5129	0.9830	0.9749	0.9624	**0.9250**
	30	0.9902	1.9008	0.9909	0.9760	0.9652	**0.9414**
	40	0.9915	2.3566	0.9978	0.9778	0.9696	**0.9575**
150	10	0.9925	1.1715	0.9819	0.9767	0.9404	**0.9021**
	20	0.9941	1.5129	0.9854	0.9786	0.9525	**0.9237**
	30	0.9946	1.9035	0.9891	0.9794	0.9621	**0.9426**
	40	0.9955	2.3243	0.9934	0.9796	0.9709	**0.9609**

Table 2: AMSE of the six methods when the standard deviation of the error term is 3

Sample size	Missing data (%)	Methods					
		LD	HD	PMM	SR	RF	SREW
30	10	8.8069	8.5440	8.1599	8.0828	8.0785	**7.9706**
	20	8.8381	9.2470	8.3181	8.1522	8.1359	**7.9897**
	30	8.8786	9.8819	8.4089	8.1774	8.1365	**8.0299**
	40	8.8998	10.6732	8.5935	8.2274	8.1578	**8.0367**
60	10	8.7949	8.9059	8.5491	8.4966	8.3580	**8.2745**
	20	8.8159	9.5395	8.6341	8.5611	8.4330	**8.3520**
	30	8.8800	10.2313	8.6886	8.5722	8.4681	**8.4191**
	40	8.9165	10.9531	8.7321	8.5928	8.4751	**8.4348**
90	10	8.8164	9.0611	8.7333	8.6950	8.4506	**8.3902**
	20	8.8384	9.6505	8.7558	8.7074	8.5199	**8.4736**
	30	8.8589	10.3971	8.7892	8.7114	8.5829	**8.5431**
	40	8.9183	11.1400	8.8310	8.7216	8.6595	**8.6365**
120	10	8.7575	9.1501	8.8032	8.7746	8.4700	**8.4343**
	20	8.7723	9.7186	8.8189	8.7815	8.5674	**8.5323**
	30	8.8196	10.4540	8.8395	8.7822	8.6413	**8.6146**
	40	8.8799	11.2191	8.8678	8.7993	8.7370	**8.7224**
150	10	8.8039	9.1289	8.8014	8.7252	8.4801	**8.4604**
	20	8.8249	9.7490	8.8398	8.7865	8.5606	**8.5333**
	30	8.8442	10.4686	8.8415	8.8124	8.6665	**8.6474**
	40	8.8914	11.3047	8.8782	8.8172	8.7265	**8.7110**

Table 3: AMSE of the six methods when the standard deviation of the error term is 5

Sample size	Missing data (%)	Methods					
		LD	HD	PMM	SR	RF	SREW
30	10	24.3017	23.2817	22.6515	22.5194	22.4622	**22.3145**
	20	24.3751	24.4563	23.0131	22.7171	22.5454	**22.4092**
	30	24.5802	25.6639	23.1792	22.7549	22.5887	**22.4121**
	40	24.6583	26.7162	23.3654	22.7764	22.5183	**22.5044**
60	10	24.3033	24.4647	23.8863	23.8062	23.4926	**23.3807**
	20	24.3054	25.4943	23.9366	23.8220	23.5066	**23.3839**
	30	24.5029	26.5842	23.9627	23.7071	23.6602	**23.6004**
	40	24.6887	27.9810	24.1488	23.8430	23.7396	**23.7029**
90	10	24.3193	24.7608	24.1851	24.0959	23.8762	**23.8038**
	20	24.3150	25.7386	24.2093	24.1233	23.9371	**23.8243**
	30	24.4703	27.0880	24.3382	24.1670	23.9852	**23.8820**
	40	24.6561	28.5191	24.4936	24.2783	24.0475	**24.0195**
120	10	24.1289	24.9352	24.3823	24.3032	23.9749	**23.9543**
	20	24.1863	25.9339	24.3886	24.3090	23.9964	**23.9762**
	30	24.3297	27.2534	24.4034	24.3155	24.1224	**24.0818**
	40	24.5526	28.7173	24.5761	24.4150	24.2436	**24.2150**
150	10	24.2532	25.0226	24.4584	24.4240	23.9655	**23.9267**
	20	24.2956	26.1673	24.5741	24.4317	24.1944	**24.1574**
	30	24.4042	27.4523	24.5760	24.4989	24.3118	**24.2781**
	40	24.5980	28.7679	24.6159	24.5122	24.3513	**24.3303**

5 Conclusions

The purpose of this study was to analyze the performances of six techniques for handling missing data on dependent and independent variables in multiple regression analysis via simulation studies. We compared five available techniques with the proposed composite imputation method. Our simulation results revealed that the estimates obtained by the SREW method outperform other methods in all situations whereas the hot deck method did not perform well especially when the missing proportion was large. Moreover, we saw that when the variability of the data was low with small sample size, the stochastic regression method also performed well next below the SREW. However, when the sample size was large, the random forest imputation seemed to perform better than the stochastic regression imputation. This result also appeared when the variability of the data was higher, the random forest imputation achieved well compared to the other methods except the SREW.

Acknowledgment. We would like to thank the referees for their comments and suggestions on the manuscript. We would also like to acknowledge Prince of Songkla University for support with all facilities.

References

[1] H. Aguilera, C. Guardiola-Albert, C. Serrano-Hidalgo, Estimating extremely large amounts of missing precipitation data, Journal of Hydroinformatics, **22**, (2020), 578–592.

[2] R. R. Andridge, R. J. Little, A Review of hot deck imputation for survey non-response, International Statistical Review, **78**, (2010), 40–64.

[3] L. Breiman, Random forests, Machine Learning, **45**, (2001), 5–32.

[4] P. Elliott, G. Hawthorne, Imputing missing repeated measures data: How should we proceed?, Australian & New Zealand Journal of Psychiatry, **39**, (2005), 575–582.

[5] C. K. Enders, Applied missing data analysis, The Guilford Press, New York, 2010.

[6] J. Han, M. Kamber, Data mining: Concepts and techniques, Morgan Kaufmann Publishers, San Francisco, 2012.

[7] S. Hong, H. S. Lynn, Accuracy of random-forest-based imputation of missing data in the presence of non-normality, non-linearity, and interaction, BMC Medical Research Methodology, **20**, (2020), 199.

[8] A. Jadhav, D. Pramod, K. Ramanathan, Comparison of performance of data imputation methods for numeric dataset, Applied Artificial Intelligence, **33**, (2019), 913–933.

[9] J. C. Jakobsen, C. Gluud, J. Wetterslev, P. Winkel, When and how should multiple imputation be used for handling missing data in randomized clinical trials a practical guide with flowcharts, BMC Medical Research Methodology, **17**, (2017), 162.

[10] R. J. Little, Missing-data adjustments in large surveys, Journal of Business & Economic Statistics, **6**, (1988), 287–296.

[11] R. J. Little, D. B. Rubin, The analysis of social science data with missing values, Sociological Methods & Research, **18**, (1989), 292–326.

[12] R. J. Little, D. B. Rubin, Statistical analysis with missing data, John Wiley & Sons, Hoboken, 2002.

[13] T. Mungua, J. Armando, Comparison of imputation methods for handling missing categorical data with univariate pattern, Journal of Quantitative Methods for Economics and Business Administration, **17**, (2014), 101–120.

[14] T. A. Myers, Goodbye listwise deletion: Presenting hot deck imputation as an easy and effective tool for handling missing data, Communication Methods and Measures, **5**, (2011), 297–310.

[15] P. A. Patrician, Focus on research methods: Multiple imputation for missing data, Research in Nursing & Health, **25**, (2002), 76–84.

[16] D. B. Rubin, Inference and missing data, Biometrika, **63**, (1976), 581–592.

[17] D. B. Rubin, Statistical matching using file concatenation with adjusted weights and multiple imputations, Journal of Business & Economic Statistics, **4**, (1986), 87–94.

[18] J. L. Schafer, J. W. Graham, Missing data: Our view of the state of the art, *Psychological Methods*, **7**, (2002), 147–177.

[19] D. J. Stekhoven, P. Buhlmann, MissForest–non-parametric missing value imputation for mixed-type data, Bioinformatics, **28**, (2012), 112–118.

[20] F. Tang, H. Ishwaran, Random forest missing data algorithms, Statistical Analysis and Data Mining, **10**, (2017), 363–377.

[21] S. van Buuren, Flexible imputation of missing Data, Chapman and Hall/CRC, Boca Raton, 2018.

[22] G. Vink, L. E. Frank, J. Pannekoek, S. van Buuren, Predictive mean matching imputation of semicontinuous variables, Statistica Neerlandica, **68**, (2014), 61–90.