CAUCHY –Jurnal Matematika Murni dan Aplikasi
Volume 7(4) (2023), Pages 641-648
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Comparing Several Missing Data Estimation Methods in Linear Regression;Real Data Example and A Simulation Study

Anwar Fitrianto[1,2*], Yap Ee Jia[3], Budi Susetyo[1] , La Ode Abdul Rahman[1]

[1]Department of Statistics, IPB University, Bogor, West Java, Indonesia
[2]Institute of Engineering Mathematics, Universiti Malaysia Perlis, Malaysia
[3]Department of Mathematics and Statistics, Universiti Putra Malaysia, Serdang, Malaysia

Email: anwarstat@gmail.com

ABSTRACT

One of the causes of bias in parameter estimation is incomplete data when analyzed using standard statistical procedures. In addition, if the analysis is performed on missing data, the researcher may not have sufficient observations necessary for the analysis. For this reason, a method is needed to estimate the missing data. Until now, there are many methods for estimating lost data.. The main objective of the study is to compare the performance between listwise deletion (LD), mean substitution (MS) and multiple imputation (MI) methods in estimating parameters. The performance will be measured through bias, standard error and 95% confidence interval of interested estimates for handling missing data with 10% missing observations. A complete empirical data set was used and assumed as population data. Ten percent of total observations in the population ere set as missing arbitrarily by generating random numbers from a uniform distribution, . Then, bias of parameter estimates and confidence interval of parameter estimates are calculated to compare the three methods. A Monte Carlo simulation was carried out to know the properties of missing data estimation methods. Simulation of 1000 sampled data with 20, 50, and 100 observations and each sample is set to have 10% missing observations. Standard statistical analyses are run for each missing data and get the average of parameter estimates to calculate the bias and standard error of parameter estimates for every missing data method. The analysis was conducted by using SAS version 9.2. It was found that the MI method provided the smallest bias and standard error of parameter estimates and a narrower confidence interval compared to the LD and MS methods. Meanwhile, the LD method gives a smaller bias of parameter estimates and standard error for small sample size of missing data. And, MS method is strongly recommended not to use for handling missing data because it will result in large bias and standard error of parameter estimates.

Keywords: incomplete, mcar, missing, regression, simulation

INTRODUCTION

Many researchers try to solve the problems of missing data analysis by using last observation carried forward (LOCF), complete case, available case, and single imputation method [1]. The LOCF is a method on substituting the last available measurement whenever there is a missing value [2]. This method was popular for treating data with monotone and non-monotone missing patterns and is valid for longitudinal studies. Beside that, many people use complete case methods such as the Listwise Deletion (LD)

Submitted: February 27, 2023 Reviewed: April 01, 2023 Accepted: April 23, 2023
DOI: http://dx.doi.org/10.18860/ca.v7i4.20548

method to solve the missing data problem, [3]. LD is the simplest and easiest method among conventional methods of dealing with missing data, [4]. Although LD is the simplest method, the number of deleted cases will increase as the values are missing arbitrarily. The major advantage of listwise deletion is that it produces a complete data set, which in turn allows for the use of standard analysis techniques, [5]. However, there is a loss in power using LD method even though the data is MCAR, especially since a large number of subjects have been deleted.

Furthermore, pairwise deletion (PD) is another case-based method. It uses all available information in statistical analysis. At the same time, the mean substitution (MS) method is also popular since it is one of the single imputation methods. A study by [6] found that overall mean computed from mean substitution (MS) is equal to the complete case values but the variance of the same variable is underestimated. Meanwhile, [7] defined the MS method as an approach to substitute the missing items with the mean of the non-missing items. It could be problematic when applying to categorical data. Although the MS method can be used to analyze a complete case, it will reduce the variability and weaken covariances and correlation estimates in the data, [8].

Another method is Multiple Imputation (MI) proposed by [3]. It is conducted by replacing each missing value with a set of plausible values. In the MI method, missing values for any variable are predicted using existing values from other variables. The predicted values, called "imputes", are substituted for the missing values, resulting in a full data set called an imputed data set, [9]. [10] provided three phases for MI inference.

Conventional statistical methods and analysis tools presume that all variables in a specified model are measured for all cases. The default method for all statistical software is simply deleting the cases with missing value on the interesting variables such as LD method. [11] supported that if the standard approach analysis is used to analyze incomplete data, the estimation of such analysis can be biased. Complete case analysis by ignoring the missing data will lead to an inefficient, biased, unreliable result. Missing data results in information loss and statistical power, [12]. Meanwhile, for a small data set that contains a relatively large number of missing observations, many cases will be simply deleted by the default method. Based on that fact, the research aims to compare the performance of Listwise Deletion (LD) and Mean Substitution (MS) methods on bias, standard error and 95% confidence interval.

METHOD

Data

The data set involved in this study is about human resources obtained from Human Development Reports of United Nationals Development Programme, [13]. Several variables are involved, such as the human development index, life expectancy at birth and gross national income variable. Human development enlarges people's choices. The most critical of these wide-ranging choices are to live a long and healthy life, be educated and access the resources needed for a decent living standard. The measure includes life expectancy, literacy, and a modified measure of income, [14].

Meanwhile, life expectancy at birth reflects the overall mortality level of a population. It summarizes the mortality pattern across all age groups from children to the elderly, [15]. Life expectancy relates to the value people attach to living long, healthy, and well. The life expectancy at birth component in the HDI data is calculated using a minimum value of 20 years and a maximum value of 85 years. The other variable included in the study is gross national income. According to [16], the gross national income is the total

domestic and foreign output claimed by residents of a country, consisting of the gross domestic product plus factor incomes earned by foreign residents minus income earned in the domestic economy by non-residents. In the data, the standard of living dimension is measured by gross national income per capita. The human development index is expected to be strongly affected by life expectancy at birth and gross national income. Thus, human development index is response variable, denoted by , life expectancy at birth and gross national income per capita are the predictors.

Simulation Designs

Monte Carlo simulation was conducted to compare performance between missing data estimation methods in estimating missing values. According to [17], the researcher begins the simulation by creating a model with known population parameters that the researcher sets the values. In the Monte Carlo simulation, B samples of N size are drawn from the population and, for each sample, estimates the interested parameter. Next, a sampling distribution is estimated for each population parameter by collecting the parameter estimates from all the samples. The properties of that sampling distribution, such as its mean and variance, come from this estimated sampling distribution.

SAS version 9.2 was used for the simulation. There are few steps involved in the simulation:

Step 1: Develop a simple linear regression model with known parameters.

The simulation is begun with a simple linear regression model $y = \beta_0 + \beta_1 x + \varepsilon$ with arbitrary known parameters β_0=5 and β_1=10. The independent variable X is generated as $2 \times$ observation number. In this simulation, the number of observations was set equal to 20. Then random errors were generated from a normal distribution $\varepsilon \sim N(0, \sigma = 2)$,

Step 2: Generate the missing data set

Let's d is the number of variable observations in the generated data set that are missing arbitrarily and $d < n$. The missing measurements are selected by generating a random number from a uniform distribution $U(0,1)$. The measurements of the variable X_1 with the first *d* smallest random numbers are made to be missing by denoted as dot "." . Now, a missing data set are created,

Step 3: Repeat B samples from a population

Repeat step 2 for 1000 times to generate 1000 missing data sets with 20 observations,

Step 4: Missing data analysis using LD, MS and MI method

For each missing data set, statistical analysis are carried out to calculate the parameter estimates of the linear regression model by using LD, MS and MI method,

Step 5: Calculate the standard error (SE) of parameter estimates and confidence intervals

After obtaining 1000 parameter estimates for the linear regression model based on each missing data estimation method, bias, standard errors and confidence intervals of the parameters are calculated as follows:

$$Bias = \bar{b}_i - \beta_i = i = 1, 2, ..., 1000. \quad (1)$$

$$SE = \sqrt{\frac{\sum \bar{b}_i^2 - \frac{(\sum \bar{b}_i)^2}{n}}{n-1}} \quad (2)$$

where $\beta_0 = 5$ and $\beta_0 = 10$. Meanwhile, the 95% confidence interval of β_i is calculated as:

$$\bar{b}_i \pm t_{0.05/2;df} \sqrt{\frac{\sum \bar{b}_i^2 - \frac{(\sum \bar{b}_i)^2}{n}}{n-1}}, \quad (3)$$

Step 6: Repeat Step 1 to step 5 for 50 and 100 observations, respectively.

Mechanism of Missing Data

[3] had developed a taxonomy and terminology of missing data mechanisms. The missing data mechanisms include missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). It is essential to understand the mechanism because the problems caused by the missing data and the solutions to these problems are different.

Regarding tests for missing data, [18] proposed an MCAR test that is distributed as a χ^2 under the H_0 and called d^2. The Little's MCAR test was conducted using a custom SAS macro. Little's MCAR test is a chi-square test to determine whether data is MCAR, [19]. Most researchers use Little's MCAR test to test MCAR. The d^2 sums the squared standardized mean differences across the J missing data patterns. The d^2 is Mahalanobis distances which is written as:

$$d^2 = \sum_{j=1}^{J} m_j \left(\bar{Y}_{obs.j} - \hat{\mu}_{obs.j}\right) \hat{\Sigma}_{obs.j}^{-1} \left(\bar{Y}_{obs.j} - \hat{\mu}_{obs.j}\right), \quad (4)$$

where m_j is the number of cases in pattern j, $df = \sum p_j - p$, where p_j is the number of complete variables for pattern j. The estimates $\hat{\mu}$ and $\hat{\sum}$ shown in Equation (4) are obtained via maximum likelihood estimation using Expectation Maximization (EM) algorithm. It is an iterative procedure that produces Maximum Likelihood (ML) estimates of a covariance matrix and mean vector, assuming MAR. The data is considered as MCAR if the p value from Little's MCAR test is insignificant. Unfortunately, no standard statistical test determines if the missing data is MAR, [20].

The three common missing data mechanisms are as follows:

Missing Completely At Random

According to [11], MCAR mechanism indicates that the probability of a missing value is unrelated to any observed and unobserved values. [21] explained the MCAR by denoting a single variable with missing data as W. Suppose a set of variables is represented by vector Z. Now, let R_W be a dummy variable with a value 1 if W is missing and 0 if W is observed. The expression of MCAR mechanism is written as:

$$Pr(R_W = 1|W, Z) = Pr(R_W = 1) \quad (5)$$

The probability that W is missing depends neither on the observed variables in Z vector nor on the missing values W itself. Complete case analysis will only give unbiased result if the missing data is assumes MCAR.

Missing At Random

MCAR mechanism is a special case of MAR mechanism. For the data to be MAR, the probability of $W$is missing depends on observed variables but independent on Witself. [22] claimed that the data is MAR when the missingness is correlated with other observed variables in the analysis. The MAR mechanism is expressed as follows:

$$Pr(R_W = 1|W, Z) = Pr(R_W = 1|Z) \quad (6)$$

Since missing $W$is arising from the same distribution as the observed Z, thus missing Wcan be predicted by using the observed Z distribution, [11]. MCAR or MAR data are sometimes considered ignorable missingness, [23]. This is because for those data still can produce unbiased estimates without needing a model to explain the missingness.

Missing Not At Random

If the data are not MCAR or MAR, then the data would be MNAR, where the probability of $W$is missing depends on both observed and missing variables. The data for MNAR is a case of non-ignorable missingness. The missing data mechanism must be modeled as part of the estimation process for a valid estimation. Unfortunately, MCAR mechanism is fraught with difficulty. The model for the missing data mechanism must be carefully tailored with each situation because every MNAR situation is different, [21]. She also stated that there is no information in data to help choose the appropriate model and no statistic to tell how well a chosen model fits the data.

RESULTS AND DISCUSSION

Bias of Parameter Estimates

As shown in Table 1, MI results in the least bias in parameter estimates. For b_0, there is only 0.0047 bias from the true parameter estimates from the original data. The b_1 can be said to be unbiased parameter estimates since the bias value is small enough, which is only -0.0001. While b_2 have biased downward with 0.0038. The three-parameter estimates using the MI method yield good parameter estimates where their biases are nearly zero.

Table 1. Bias of Parameter Estimates of LD, MS and MI Method

Methods	Bias of Parameter Estimates		
	b_0	b_1	b_2
LD	-0.0170	0.0002	0.0202
MS	0.2716	-0.0020	-0.4992
MI	0.0047	-0.0001	-0.0038

The MI was expected to have the smallest bias compared to the other two methods. Although the missing data is MCAR, since the human development index and life expectancy at birth are highly correlated, it helps predict the missing value nearest to the original value. In the MS method, each missing value is substituted by the mean of the observed variable. It means that all missing values are replaced with same value. It causes the substituted values are not similar or nearer to the original value since a single mean value only substitutes every missing value. That is the reason MS method yields the most extensive bias among these three methods.

Table 2. 95% Confidence Interval and Corresponding Length of Parameter Estimates of LD, MS and MI Methods

Estimator	b_0			b_1			b_2		
True Value		0.8252			0.00736			-2.5594	
Method	LCI	UCI	Length	LCI	UCI	Length	LCI	UCI	Length
LD	0.6739	0.9426	0.2687	0.0065	0.0086	0.0021	-2.7980	-2.2802	0.5178
MS	0.9738	1.2199	0.2461	0.0044	0.0064	0.0020	-3.2985	-2.8186	0.4800
MI	0.7015	0.9583	0.2567	0.0063	0.0083	0.0020	-2.8077	-2.3187	0.4889

Meanwhile, the performance of each estimation method of missing data about 95% confidence interval is displayed in Table 2. The MS method gives the smallest confidence interval length based on the table. But, true parameter values from MS method did not fall in the interval. Hence, the MS method was not suitable for estimating the true value. But, the LD and MI methods can help estimate true value since true value falls in the confidence interval. MI method was preferable than the LD method in estimating parameters since it can produce a narrower confidence interval.

Results of The Simulation Study

When there are 10% missing observations in a small data set (n=20), for after 1000 simulation runs, the MS method provides the least bias of parameter estimates. However, the bias of parameter estimates for MS method increases as the sample size increases. In other words, the MS method is not good enough to handle medium or large numbers of missing data since it will produce large bias.

Table 3. Bias of Parameter Estimates from 1000 Simulation Runs for LD, MS and MI with 10% Missing Data in 20, 50, and 100 Observations, respectively

Number of Simulation	**Number of Observations**	**Number of Missing Observations**	**Methods**	**Bias of Parameter Estimates**	
				b_0	b_1
1000	20	2	LD	-1.0013	0.0125
			MS	0.0969	0.0070
			MI	-1.0950	0.0141
	50	5	LD	-0.8805	0.0093
			MS	2.0675	-0.0024
			MI	-0.7981	0.0071
	100	10	LD	-0.8420	0.0083
			MS	6.3112	-0.0095
			MI	-0.7075	0.0062

Meanwhile, the LD and MI methods are better for handling many missing data. Bias of b_0 produced by LD method for small number of missing data is negative with -1.0013 and b_1 is 0.0125. When n is increases to 50, bias of parameter estimates reduces. Table 3 shows that the bias for b_0 and b_1 in LD method decrease as the number of observations increases. As we see in Table 3, MI method follows the trends as LD method that the bias of parameter estimates is reduced as n increases. However, we notice that the bias from MI method reduces more drastically compared to LD method as n increases.

The results from the simulation study follow the results from the empirical data where the MI method gives the least bias for a large number of missing data.

Table 4. Standard Error of Parameter Estimates from 1000 Simulation Runs for LD, MS and MI with 10% Missing Data in 20, 50, and 100 Observations, respectively.

Number of Simulation	Number of Observations	Number of Missing Observations	Methods	Standard Error of Parameter Estimates	
				b_0	b_1
1000	20	2	LD	0.2368	0.0109
			MS	10.3413	0.0763
			MI	0.2411	0.0110
	50	5	LD	0.2305	0.0086
			MS	14.6124	0.0744
			MI	0.1466	0.0045
	100	10	LD	0.1838	0.0062
			MS	20.4287	0.0647
			MI	0.1110	0.0010

Table 4 shows the standard error of parameter estimates for LD, MS and MI method. The standard error of b_0 from MS method is increase as n increases. And, although the standard error of b_1 decreases as n increases, but the standard error of b_1 is the largest compared to the standard error of parameter estimates from LD and MI methods. Overall, the MS method is not suggested to handle missing data because it will result in the largest standard error of parameter estimates if we compare to the LD and MI methods.

On the other hand, the LD and MI method provide a small standard error of parameter estimates. The LD method provides the smallest standard error of parameter estimates when n =20. However, the MS method gives the smallest standard error of parameter estimates, increasing to 50 and 100.

CONCLUSIONS AND FUTURE WORKS

Missing data may result in biased parameter estimates. However, a different sample size of missing data is recommended by various methods. The LD method is more appropriate for treating small missing data because it gives negligible bias and standard error of parameter estimates. However, the MS method is strongly recommended not to use for handling missing data due to large bias and standard error of parameter estimates. From the study result, MI method can reduce the bias of parameter estimates of missing data. Based on that fact, it is strongly recommended to analyze missing data using MI method rather than LD method and MS method because the MI method results in smaller bias and standard error of parameter estimates.

In the future, analyses of data with missing values in two or more variables are suggested. In the statistical analysis, other than standard error and bias of parameter estimates, analysts can also compare the mean and efficiency of different missing data estimation methods at different missing percentages. Furthermore, analyzing missing data with different mechanisms or patterns can also be a research topic.

REFERENCES

[1] S. Ghosh and P. Pahwa, "Assessing bias associated with missing data from Joint Canada," in *United States Survey of Health: an application, paper presented at the Joint Statistical Meetings, Denver, CO, USA*, 2008.

[2] G. Molenberghs and G. Verbeke, "Multiple imputation and the expectation-maximization algorithm," *Models for discrete longitudinal data*, pp. 511–529, 2005.
[3] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, vol. 81. John Wiley & Sons, 2004.
[4] R. L. Carter, "Solutions for missing data in structural equation modeling.," *Research & Practice in Assessment*, vol. 1, pp. 4–7, 2006.
[5] A. N. Baraldi and C. K. Enders, "An introduction to modern missing data analyses," *J Sch Psychol*, vol. 48, no. 1, pp. 5–37, 2010.
[6] R. J. A. Little, "Regression with missing X's: a review," *J Am Stat Assoc*, vol. 87, no. 420, pp. 1227–1237, 1992.
[7] P. R. de Gil and J. D. Kromrey, "Missing_Items: A SAS® Macro for Missing Data Imputation in Summative Response Scales".
[8] K. F. Widaman, "Best practices in quantitative methods for developmentalists: III. Missing data: What to do with or without them.," *Monogr Soc Res Child Dev*, 2006.
[9] J. C. Wayman, "Multiple imputation for missing data: What is it and how can I use it," in *Annual Meeting of the American Educational Research Association, Chicago, IL*, 2003, vol. 2, p. 16.
[10] Y. C. Yuan, "Multiple imputation for missing data: Concepts and new development (Version 9.0)," *SAS Institute Inc, Rockville, MD*, vol. 49, no. 1–11, p. 12, 2010.
[11] T. E. Raghunathan, "What do we do with missing data? Some options for analysis of incomplete data," *Annu. Rev. Public Health*, vol. 25, pp. 99–117, 2004.
[12] C.-Y. J. Peng, M. Harwell, S.-M. Liou, and L. H. Ehman, "Advances in missing data methods and implications for educational research," *Real data analysis*, vol. 3178, p. 102, 2006.
[13] UNDP, "United Nations Development Programme: Human Development Reports 2015." Oxford University Press Oxford, 2014.
[14] G. Ranis, F. Stewart, and E. Samman, "Human development: beyond the human development index," *Journal of Human Development*, vol. 7, no. 3, pp. 323–358, 2006.
[15] W. H. Organization, *The world health report 2006: working together for health*. World Health Organization, 2006.
[16] G. Chamberlin, "Gross domestic product, real income and economic welfare," *Economic & Labour Market Review*, vol. 5, pp. 5–25, 2011.
[17] P. Paxton, P. J. Curran, K. A. Bollen, J. Kirby, and F. Chen, "Monte Carlo experiments: Design and implementation," *Structural Equation Modeling*, vol. 8, no. 2, pp. 287–312, 2001.
[18] R. J. A. Little, "A test of missing completely at random for multivariate data with missing values," *J Am Stat Assoc*, vol. 83, no. 404, pp. 1198–1202, 1988.
[19] T. G. Morrison, M. A. Morrison, and J. M. McCutcheon, "Best practice recommendations for using structural equation modelling in psychological research," *Psychology*, vol. 8, no. 09, p. 1326, 2017.
[20] T. Schwartz and R. Zeig-Owens, "Knowledge (of your missing data) is power: handling missing values in your SAS dataset," in *SAS Global Forum*, 2012, pp. 1–8.
[21] P. D. Allison, *Missing data*, vol. 200210, no. 9781412985079.31. Sage Thousand Oaks, CA, 2010.
[22] D. C. Howell, "The treatment of missing data," *The Sage handbook of social science methodology*, vol. 208, p. 224, 2007.
[23] T. D. Pigott, "A review of methods for missing data," *Educational research and evaluation*, vol. 7, no. 4, pp. 353–383, 2001