Comparison of multiple imputation and complete-case in a simulated longitudinal data with missing covariate

Chin Wan Yoke and Zarina Mohd Khalid

Citation: AIP Conference Proceedings **1605**, 918 (2014); doi: 10.1063/1.4887712
View online: http://dx.doi.org/10.1063/1.4887712
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1605?ver=pdfcov
Published by the AIP Publishing

Comparison of Multiple Imputation and Complete-Case in a Simulated Longitudinal Data with Missing Covariate

Wan Yoke, Chin and Zarina Mohd Khalid

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, 81310 UTM Johor Bahru, Johor, Malaysia.

Abstract. Along a continual process of collecting data, missing recorded datum always a main problem faced by the real application. It happens due to the carelessness or the unawareness of a recorder to the importance of data documentation. In this study, a random-effects analysis which simulates data from a proposed algorithm is presented with a missing covariate. It is an improved simulation method which involves first-order autoregressive (AR(1)) process in measuring the correlation between measurements of a subject across two time sequence. Complete-case analysis and multiple imputation method are comparatively implemented for the estimation procedure. This study shows that the multiple imputation method results in estimations which fit well to the data which are not only missing completely at random (MCAR) but also missing at random (MAR). However, the complete-case analysis results in estimators which fit well to the data which are only MCAR.

Keywords: Complete-case analysis, missing data, multiple imputation, random-effects analysis.
PACS: 02.50.-r

INTRODUCTION

Missing data or losing information from a raw range of records, is a frequent problem happened in human daily working life. A datum may be technically lost due to human carelessness or an object being absence from a follow-up study which occurs occasionally or directly dropout from the study if the nonparticipation is continuous. It looks unimportant and does not affect much to the production line of a factory, or treatment services during hospitalization. However, this incomplete information has caused an extensive large bias and lack of robustness to the mathematical analysis.

Observations may produce a concrete collection along the experiment path while this data can then be modeled cross-sectional or longitudinally. By making the proposed model looks perfect to the real application; a longitudinal model is the preference. Random-effects analysis or the so called longitudinal model by Laird and Ware [1] is a widely applicable mathematical function to formulate these repeated measures. Goldstein [2] thereafter introduced a multilevel model to categorize the full random-effects model into two hierarchical levels such as within-subjects and between-subjects models.

Chin et al. [3] has proposed an algorithm to simulate the repeated measurements from Goldstein [2] multilevel model. However, the respective study was not supported by the correlation across two time sequence. Therefore, this study comes across to the improved proposed algorithm by adding Diggle [4] suggestions, and then makes the analysis run through the complete-case analysis and multiple imputation. Three features of variance matrix have to be followed when duplicating the repeated measurements; which includes (1) random measurements are sometimes imperfectly correlated for the small time distance, (2) the random outcomes are largely diverged and results in a positive correlation across two measurements of an object, and (3) the correlation across follow-up measurements of an object is functionalized in term of monotone decreasing model [4]. As a result, first-autoregressive (AR(1)) is chosen to correlate the relationship between two time sequences.

Next, Diggle and Kenward [5] defined the missing data by three missingness mechanisms, for instance the probability of missing completely at random (MCAR) that autonomous of both observed and losing information while the probability of missing at random (MAR) is defined only from the observed variables whereas the probability of missing not at random (MNAR) is claimed under the effect of both observed information and losing information. In mathematical definition, the probability of missingness, $p(M \mid Y, \varphi)$ is written as $p(M \mid \varphi)$ for all Y, φ under MCAR case; while $p(M \mid Y_{obs}, \varphi)$ for all Y_{mis}, φ under MAR case; and finally $p(M \mid Y_{mis}, \varphi)$ under the MNAR case where p probability function, M missingness, φ probability function parameter set, $Y = (Y_{obs}, Y_{mis})$ in

Proceedings of the 21st National Symposium on Mathematical Sciences (SKSM21)
AIP Conf. Proc. 1605, 918-922 (2014); doi: 10.1063/1.4887712

which Y_{obs} is the observed data value matrix and Y_{mis} is the missing data value matrix [6]. Sources by Presti et al. [6] further specified the suggested method for MCAR is gone well with the deletion-based techniques such as listwise and pairwise methods; MAR is suited with the regression imputation and multiple imputation methods; and MNAR should ensemble under the mixed-models of modeling approaches.

For the procedure that subject to missing values, it can be referred to Mohamad [7] and Ho et al. [8]. The latter authors adjusted the logistic regression model of the former researcher in simulating the missing datum. The missing data simulated in this study therefore follow the adjusted procedure published in the latter article; meanwhile, additional adjustment is applied into the studies, which comprises only MCAR and MAR mechanisms.

METHODOLOGY

Let Y_{obs} denotes the response variable with complete observed elements in the row, Y_{mis} denotes the response variable with any losing information in the row, X_{obs} denotes the observed covariate, and X_{mis} denotes the losing covariate; the summarized data collection are written as in Equation (1)

$$(Y, X) = \left(\begin{bmatrix} Y_{obs} \\ Y_{mis} \end{bmatrix}, \begin{bmatrix} X_{obs} \\ X_{mis} \end{bmatrix} \right). \tag{1}$$

This study provides two estimation approaches to approximate longitudinal model with respect to covariate X which use to subject with missing data. One is the complete-case analysis and another is the multiple imputation. Both methods are accordingly applied into the MCAR and MAR study.

Complete-Case Analysis

Complete-case analysis is also named as listwise deletion method. It has the simplest concept to apply by discarding any object that carries losing information in any variables of collected data. The analysis is then followed by the estimation of Equation (2) via a standard process which assumes

$$Y_{i(obs)} = X_{i(obs)}\beta + Z_{i(obs)}b_{i(obs)} + \varepsilon_{i(obs)} \tag{2}$$

as a longitudinal model with the error term, $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$ and random effects, $b_i \sim N(0, \sum)$.

Regression Multiple Imputation

Regression multiple imputation, on the other hand is a replacement-based technique which includes mean substitution, hot-deck imputation, regression imputation and last observation carried forward (LOCF). At this point, a stochastic regression imputation [9] was selected during the data-augmentation procedure. The method description starts by a single imputation process before proceed into the multiple imputation process.

Given covariate, X_j is the variable with missing datum while $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ are the regression coefficient parameters correspond to $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ repeated response variables. The new regression model to formulate the completely observed variables is assumed as below; with the variables $\mathbf{Y}$ predict the missing datum of X_j [10]:

$$X_j = \beta_0 + \beta_1 Y_1 + \beta_2 Y_2 + \ldots + \beta_k Y_k \,. \tag{3}$$

Let $\mathbf{V}_j$ be the usual $\mathbf{X}'\mathbf{X}$ matrix from the intercept and variables $\mathbf{Y}$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ defines from Equation (3) are the regression coefficient estimators that associate with covariance matrix $\sigma_j^2 \mathbf{V}_j$. The following procedures are the process to imputing the missing values for each imputation:

1. Simulate the new regression coefficient parameters, $\boldsymbol{\beta}_* = (\beta_{*0}, \beta_{*1}, \ldots, \beta_{*k})$ by

$$\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}} + \sigma_{*j} \mathbf{V}'_{hj} \mathbf{Z} \tag{4}$$

where

$\hat{\boldsymbol{\beta}}$ are the first regression coefficient estimators,

$\sigma_{*j}^2 = \hat{\sigma}_j^2 (n_j - k - 1) \big/ \chi_{n_j - k - 1}^2$ is the variance with n_j number of nonmissing observations for X_j,

$\mathbf{V}'_{hj}$ is the upper triangular matrix of the Cholesky decomposition, $\mathbf{V}_j = \mathbf{V}'_{hj} \mathbf{V}_{hj}$, and

$\mathbf{Z}$ is the independent random normal variates of $k + 1$ vector.

2. The missing values are then predicted from

$$\beta_{*0} + \beta_{*1} Y_1 + \beta_{*2} Y_2 + \ldots + \beta_{*k} Y_k + z_i \sigma_{*j} \tag{5}$$

where $\mathbf{Y}$ are the values of the predicting variables and the simulated random error z_i follows a normal deviate.

3. Iterate the imputation procedures for at least 20 times to obtain multiple copies of data sets.

ANALYSIS OF SIMULATION STUDIES

The simulation analysis in this study will start by assuming the true parameter values, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)$ in the full longitudinal model, $y_{ik} = \beta_0 + \beta_1 t_{ik} + \beta_2 x_i + \beta_3 (x_i t_{ik}) + u_{0i} + u_{0i} t_{ik} + \varepsilon_{ik}$, with (5,–2.5,1.125,0.9). A sample of 500 objects for 1000 simulations studies is performed for three circumstances such as complete observations, missing completely at random (MCAR) and missing at random (MAR). Complete observations indicates no missingness values in the collection, however for the mechanism of MCAR and MAR, both are responsible subjected to 30% and 50% missing covariate.

As presented in Table (1), performances of the regression parameters estimate from the method of complete-case analysis and multiple imputation are comparatively summarized in term of standard error (Equation 6), bias (Equation 7), and mean square error (MSE) (Equation 8).

$$SE(\hat{\theta}) = \sigma_\theta \big/ \sqrt{M} \tag{6}$$

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta \tag{7}$$

$$MSE = Var(\hat{\theta}) + Bias(\bar{\hat{\theta}})^2 \tag{8}$$

where given θ as the true parameter, $\hat{\theta}$ is the estimator, $\bar{\hat{\theta}} = \sum_{i=1}^{M} \hat{\theta} \big/ M$ with M total number of simulations is the mean of the estimator, and σ_θ denotes the standard deviation of the estimator. The accessible of standard error is to test the efficiency of the regression estimators while bias is to evaluate the accuracy of the respective estimators, whereas MSE holds both the ability of accuracy and efficiency tests in the investigation.

Estimators from complete observations, which achieve a 0% missing value in the data collection, have performed the best. As expected, the estimators performed the best when there is no missing value, with lowest standard errors, less biases, and smallest MSE values. Unfortunately, none of a set of data collections is perfectly observed in the real application. Therefore, when the study turns into the analysis of missing values, complete-case analysis lead to estimators performs better under the MCAR missing mechanism whereas it fits less steady to the MAR condition in

which a great biases and MSE values have been earned by the MAR. Multiple imputation on the other hand overcomes this problem and satisfies the condition of MAR, therefore is suggested in the analysis.

Next, as the missing data increase, the MSE values observed by complete-case analysis also severely increase especially in MAR incident. Nevertheless, by using multiple imputation, the results performed across both the missingness mechanism and the percentage of missing values, are generally consistent.

TABLE (1). Simulation Results of Missing Covariate X via the Complete-Case Analysis and Multiple Imputation with respect to (i) Complete Observations (0% missingness), (ii) Missing Completely At Random (MCAR), and (iii) Missing At Random (MAR) at 30% and 50% missingness

Study	**Complete-Case Analysis**				**Multiple Imputation**			
$\bar{\hat{\boldsymbol{\beta}}}$	$\bar{\hat{\beta}}_0$	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$\bar{\hat{\beta}}_3$	$\bar{\hat{\beta}}_0$	$\bar{\hat{\beta}}_1$	$\bar{\hat{\beta}}_2$	$\bar{\hat{\beta}}_3$
Complete Observations	***0% missingness***							
Estimators	4.9715	-2.5099	1.1256	0.9003				
Standard Error	0.0376	0.0078	0.0012	0.0003				
Bias	-0.0285	-0.0099	0.0006	0.0003				
MSE	1.4168	0.0605	0.0016	0.0001				
MCAR	***30% missingness***							
Estimators	4.9477	-2.5197	1.1265	0.9006	5.3378	-2.2174	1.1136	0.8907
Standard Error	0.0458	0.0091	0.0015	0.0003	0.0381	0.0099	0.0013	0.0003
Bias	-0.0523	-0.0197	0.0015	0.0006	0.3378	0.2826	-0.0114	-0.0093
MSE	2.0969	0.0837	0.0023	0.0001	1.5693	0.1774	0.0017	0.0002
	50% missingness							
Estimators	4.9880	-2.5182	1.1253	0.9006	5.2898	-2.2562	1.1151	0.8921
Standard Error	0.0550	0.0106	0.0018	0.0004	0.0395	0.0115	0.0013	0.0004
Bias	-0.0120	-0.0182	0.0003	0.0006	0.2898	0.2438	-0.0099	-0.0079
MSE	3.0210	0.1134	0.0033	0.0001	1.6424	0.1919	0.0018	0.0002
MAR	***30% missingness***							
Estimators	7.1055	-2.2562	1.0351	0.8895	5.1655	-2.4211	1.1197	0.8980
Standard Error	0.0475	0.0097	0.0016	0.0003	0.0392	0.0130	0.0013	0.0005
Bias	2.1055	0.2438	-0.0899	-0.0105	0.1655	0.0789	-0.0053	-0.0020
MSE	6.6923	0.1536	0.0108	0.0002	1.5612	0.1753	0.0018	0.0002
	50% missingness							
Estimators	7.5530	-2.2000	1.0064	0.8860	5.0022	-2.5695	1.1254	0.9034
Standard Error	0.0551	0.0114	0.0019	0.0004	0.0397	0.0145	0.0014	0.0006
Bias	2.5530	0.3000	-0.1186	-0.0140	0.0022	-0.0695	0.0004	0.0034
MSE	9.5485	0.2208	0.0178	0.0004	1.5760	0.2158	0.0018	0.0003

As a result, complete-case analysis causes biases against estimators. Therefore, the estimators are less convincing, especially when the missing values are high and results in severe reduction sets of observations. The estimates of complete-case analysis are unbiased if the missingness does not depend on the response variable. Therefore, Allison [11] stated that the complete-case analysis only applies appropriately under the condition of MCAR. For the next missingness mechanism, MAR, it is preferred under the multiple imputation approach. Estimators from multiple imputation approach at this point can fit well to the MCAR as well.

The purpose of this research is to strengthen the output of the improved algorithm to the repeated measurements. This is due to the limited simulation studies which have been issued by the previous researchers while only limited licensed software can run this simulation study. Although there are constraints and limitation to the proposed study, it is useful for the future study references. For instance in this study, it is proved that the results are reasonable and the concepts are logically observed.

ACKNOWLEDGMENTS

The authors would like to thank the Malaysian Ministry of Higher Education and Universiti Teknologi Malaysia. This paper will contribute as part of the PhD thesis.

REFERENCES

1. N. M. Laird and J. H. Ware, *Biometrics* **38**, 963-974 (1982).
2. H. Goldstein, *Biometrika* **73**, 43-56 (1986).
3. W. Y. Chin, Z. M. Khalid and M. K. Ho, *Menemui Matematik* **34**, 7-15 (2012).
4. P. J. Diggle, *Biometrics* **44**, 959-971 (1988).
5. P. J. Diggle and M. G. Kenward, *J. R. Stat. Soc.: Ser. C (Appl. Statist.)* **43**, 49-93 (1994).
6. R. L. Presti E. Barca and G. Passarella, *Environ. Monit. Assess.* **160**, 1-22 (2010).
7. I. Mohamad, "Data Analysis in the Presence of Missing Data", Ph.D. Thesis, Lanchester University, 2003.
8. M. K. Ho, F. Yusof and I. Mohamad, *Jurnal Teknologi* **57**, 57-67 (2012).
9. A. N. Baraldi and C. K. Enders, *J. School Psychol.* **48**, 5-37 (2010).
10. Y. Yuan, *J. Stat. Softw.* **45**, 1-25 (2011).
11. P. D. Allison, "Missing Data" in *The SAGE Handbook of Quantitative Methods in Psychology*, edited by R. E Millsap and A. Maydeu-Olivares, London: SAGE Publications Ltd, 2009, pp. 72-89.