Int. Journal of Math. Analysis, Vol. 5, 2011, no. 45, 2217 - 2227

Analysis of Imputation Methods for Missing Data in AR(1) Longitudinal Dataset

Michikazu Nakai

Innovation Center for Medical Redox Navigation,
Kyushu University,
Fukuoka, 812-8582, Japan
mnakai@redoxnavi.med.kyushu-u.ac.jp

Abstract

When missing data occur in longitudinal analysis, multiple imputation (MI) tends to be considered as the most efficient method for imputation. In simulation study with AR(1) longitudinal dataset, three basic imputation methods (Complete Case, Last Observation Carried Forward and Mean Imputation) are compared with MI to comprehend and classify each performance with empirical mean and mean square error. The final section provides summary.

Mathematical Subject Classification: 62-07, 62H15, 62Q05

Keywords: Longitudinal analysis, Imputation, Missing data

1 Introduction

Missing data are inevitable problem in longitudinal analysis. The most

common way of dealing with missing data is to delete all cases, that is called Complete Case (CC) method. However, this method describes as an ineffective technique to handle missing data in most reviews because the reduced sample used for the analysis leads to loss of statistical power [1,4,14]. Some claims that Multiple Imputation (MI) method is the most sophisticated technique to deal with missing data [5,10]. Yet, its method still doesn't become widespread because of computational difficulty. Moreover, there is little guidance of how to treat missing data with an appropriate method.

Research and comparison of each imputation for analyzing with missing data have been conducting continuously. For example, Blanker et al. [3] concluded that CC method and Last Observation Carried Forward (LOCF) method didn't perform well in their experimental setting. Moreover, Donders et al. [4] stated CC method and Mean Imputation method produced biased result compared with other methods. There was a comparison of CC method, Mean Imputation method, and MI method; CC method found less precise [1]. Furthermore, when White and Carlin [17] performed a research of efficiency between CC method and MI method, even though MI method was effective, there was no difference between CC method and MI method when dataset was binary. Heijenet al. [7] concluded that single imputation methods performed as well as MI method when there were low overall numbers of missing data. In other words, the accuracy of MI method has not been confirmed with strong evidence. Therefore, the appropriate choice of the method is a crucial aspect to deal with the missing data. Besides, it helps researchers to have board-range information about each imputation method including advantages/disadvantages with a variety of different circumstances.

This paper focuses on comparison among selected imputation methods and provides characteristics of imputation methods using simulation study in AR(1) longitudinal dataset. The first section states an overview of missing mechanism and imputation methods. The subsequent section presents a comparison of four popular imputation methods and explains advantages and disadvantages of each imputation. The final section concludes with some remarks.

2. Method

2.1 Missing Mechanism

There are three different missing mechanisms [10]. The first mechanism is called as Missing at Random (MAR). Under MAR, the probability of missingness depends on the observed data, but not the unobserved data. The special case of MAR is the second mechanism, called Missing Completely at Random (MCAR). Under MCAR, the probability of missingness is independent of both observed data and unobserved data. The last mechanism is called Not Missing at Random (NMAR). Under NMAR, the probability of missingness depends on both observed data and unobserved data. Since the probability of missing data is related to at least some unobserved data, NMAR is often referred as non-ignorable, which implies to the fact that missing data mechanism cannot be ignored. More detailed explanations for each mechanism are referred to [4,13].

2.2 Imputation

This section summarizes advantages/disadvantages of four imputation methods which are used in simulation study.

As stated, Complete Case (CC) method is simply to delete all case with missing data at any measurement occasion. The advantage of its method is that it is easy to conduct and its method is a default method in most statistical packages to treat missing data. Researches have already shown that CC method requires MCAR for unbiased estimation [3,5,12].

Last Observation Carried Forward (LOCF) method is regularly used in epidemiological research, especially in clinical trial [18].This method is for every missing data to be replaced by the last observed value from the same subject. Although the original assumption of missing data is MAR, a recent research has shown that its method doesn't give a valid analysis if the missing mechanism is anything other than MCAR [9]. In addition, LOCF method creates bias even in

strong MCAR assumption [12].

Mean Imputation method is to replace missing data with the mean of the variable. That is, mean of the non-missing data is used in place of missing data. Even though this strategy is simple to impute, it can severely distort the distribution for its variable and attenuate variance estimate [3,15]. This method also assumes missing mechanism to be MCAR for unbiased estimation. Donders et al. explained that its imputation always results in biased estimation even in MCAR [4]. Median can be replaced for imputation instead of mean.

The primary imputation method of handling missing data is Multiple Imputation (MI) method in which replaces each missing item with two or more acceptable values representing a distribution of possibilities. The advantage of its method is that once the imputed dataset has been generated, the analysis can be carried out using procedures in virtually any statistical packages. However, using MI method, the uncertainty inherent in missing data is ignored because the analysis doesn't distinguish between the observed and imputed values [6].Also, missing data individuals are allowed to have varying probability in MI method, which means that individual variation is ignored [2].

3. Simulation Study

For a dataset, suppose repeated measurements $Y_{it}(i = 1, \cdots, 100; t = 1, \cdots, 5)$ are generated from a multivariate normal distribution with mean response $E(Y_{it}) = \beta_0 + \beta_1 t$ where $\beta_0 =$ intercept, $\beta_1 =$ slope and $\rho^{|s-t|}$ =correlation for $\rho \geq 0$, then simulated N=200 different random longitudinal datasets in SAS®. The variance at each occasion is assumed to be constant over time, while the correlations have a first-order autoregressive (AR(1)) pattern with positive coefficient [5]. Assuming that the first occasion was fully observed, simple random sampling without replacement was used to make MCAR datasets and to test following cases:

Case I: 5% missingness at each time point;

Case II: 0%, 5%, 10%, 15% and 20% at time points 1, 2, 3, 4, 5, respectively.

The experiment itself consists of mean of the 200 empirical means (1000 for MI) and Mean Square Error (MSE) from a fitted mean $E(Y_{it})$. In addition, normality Shapiro-Wilk test is performed to each imputation method at each time point and Analysis of Variance (ANOVA) test is conducted to verify the significance whether means are different between original dataset and imputed dataset with $\alpha = 0.05$ level. Multiple comparisons with Turkey procedure are used for mean comparisons. If normality test fails, its imputation excludes from comparison. At last, two different slope values (0.1 and 2) are tested to investigate the effectiveness for imputations. The default numbers for each parameters are following: $\rho = 0.7$, $\sigma^2 = 1$, and $\beta_0 = 10$.

The computation was mainly carried out using the computer facilities at Research Institute for Information Technology, Kyushu University.

4. Result

At first, Table1 yields that all time points have equally 5% missingness. The purpose of this case is to consider any influence of AR(1) structure with missing data, that is, to judge whether imputation methods have any relationship as correlation decreases. Now, all cases in CC method are not rejected with $\alpha = 0.05$ level and normality tests are clear. The sample frequency for CC method in Case I ranges from 81 to 83 out of 100 subjects. Even though there is 5% missingness at each time point, since the missingness at different time points is independent, the estimated percentage of missing data would increase more. Mean Imputation method doesn't seem to appear any disadvantage of distorting the distribution for its variable in Table1. In Case I, both CC method and Mean Imputation method seem to appear unbiased at all-time points. LOCF method already estimates poorly in $\text{slope} = 2$ when missingness is 5% whereas its method is accurate in $\text{slope} = 0.1$. This explains that it is not important for LOCF method how many missing percentage it is in the dataset, but the accuracy of LOCF method is

affected by a slope value at each time point. Furthermore, MI method fails normality test in 5% missingness. MI method is generally to model

Table 1: Case I for mean of 200 empirical means with different slopes and missing data with MSE in parentheses

Slope	Method	Missing percentage				
		0%	5%	5%	5%	5%
0.1	Fitted	10.1	10.2	10.3	10.4	10.5
	Original	10.0886 (0.01086)	10.1978 (0.01048)	10.2954 (0.00919)	10.3975 (0.00982)	10.4969 (0.00963)
	Complete	10.0858 (0.01429)	10.1951 (0.01372)	10.2939 (0.01173)	10.3971 (0.01243)	10.4968 (0.01156)
	Mean	10.0886 (0.01086)	10.1985 (0.01142)	10.2959 (0.00934)	10.3958 (0.01041)	10.4997 (0.00982)
	LOCF	10.0886 (0.01086)	10.1926 (0.01037)	10.2917 (0.00920)	10.3929 (0.00960)	10.4940 (0.00962)
	MI	10.0864 (0.01086)	10.1972† (0.01074)	10.2953† (0.00939)	10.3913† (0.01003)	10.4914† (0.01009)
2.0	Fitted	12	14	16	18	20
	Original	11.9886 (0.01086)	13.9978 (0.01048)	15.9954 (0.00919)	17.9975 (0.00982)	19.9969 (0.00963)
	Complete	11.9857 (0.01275)	13.9919 (0.01313)	15.9897 (0.01251)	17.9966 (0.01326)	19.9957 (0.01146)
	Mean	11.9886 (0.01086)	13.9969 (0.01095)	15.9926 (0.00966)	17.9959 (0.01066)	19.9975 (0.01016)
	LOCF	11.9886 (0.01086)	13.8958* (0.02128)	15.8936* (0.02039)	17.8924* (0.02149)	19.8911* (0.02100)
	MI	11.9886 (0.01086)	13.9957† (0.01095)	15.9944† (0.00936)	17.9894† (0.01022)	19.9883† (0.00987)

*p-value<0.05 comparing with original dataset †Fails normality test

misspecification when the amounts of missing data are not large [2]. MI method doesn't seem to be suitable method when missingness is small.

Table2 inquires the increment of 5% missingness at time points. At first, both CC method and Mean Imputation method remain unchanged with Table1. That is, both methods keep the efficiency up to 20% missingness. This result agrees with

literature that CC method is satisfied imputation up to 25% [11]. However, MSE of CC method is almost double of Original dataset which indicates some bias. In contrast, MSE of Mean Imputation is as close as Original dataset. This describes imputed means are not so different from Original dataset. Moreover, LOCF

Table 2: Case II for mean of 200 empirical means with different slopes and missing data with MSE in parentheses

Slope	Method	Missing percentage				
		0%	5%	10%	15%	20%
0.1	Fitted	10.1	10.2	10.3	10.4	10.5
	Original	10.0886 (0.01086)	10.1978 (0.01048)	10.2954 (0.00919)	10.3975 (0.00982)	10.4969 (0.00963)
	Complete	10.0852 (0.02060)	10.1989 (0.02113)	10.2897 (0.01741)	10.3886 (0.01689)	10.4876 (0.01843)
	Mean	10.0886 (0.01086)	10.1967 (0.01049)	10.2943 (0.01086)	10.3971 (0.01135)	10.4941 (0.01277)
	LOCF	10.0886 (0.01086)	10.1931 (0.01017)	10.2843 (0.00901)	10.3820† (0.00967)	10.4732 (0.01082)
	MI	10.0886 (0.01086)	10.2040† (0.01058)	10.2992† (0.00966)	10.3866† (0.01121)	10.5344† (0.01434)
2.0	Fitted	12	14	16	18	20
	Original	11.9886 (0.01086)	13.9978 (0.01048)	15.9954 (0.00919)	17.9975 (0.00982)	19.9969 (0.00963)
	Complete	11.9832 (0.01949)	13.9935 (0.02076)	15.9868 (0.01801)	17.9876 (0.01705)	19.9915 (0.01615)
	Mean	11.9886 (0.01086)	13.9956 (0.01118)	15.9985 (0.01058)	17.9947 (0.01087)	19.9889 (0.01115)
	LOCF	11.9886 (0.01086)	13.8980* (0.02075)	15.7900* (0.05290)	17.6693* (0.11881)	19.5340* (0.22777)
	MI	11.9886 (0.01086)	14.0048† (0.01087)	16.0004† (0.00976)	17.9840 (0.01097)	20.0321† (0.01249)

*p-value<0.05 comparing with original dataset †Fails normality test

method rejects at all-time points in $\text{slope} = 2$. Since 5% missingness for LOCF method indicates significant in Case I, it is natural prediction to observe that higher missingness is also significant. However, normality test in 15%

missingness with slope $= 0.1$ indicates to reject while 20% missingness is accepted. Even though it may have happened by chance, LOCF method may have less accuracy in 15% to 20% missingness even for a small slope. Moreover, MI method fails normality test in most of missing percentages. White and Carlin mention that more missingness enhances the efficiency advantage of MI method. Therefore, MI method should be used when missingness is large enough or when the number m of imputed datasets increases [17].

5. Discussion

In this paper, we have reviewed missing mechanism briefly, examined the consequences of missing data in longitudinal study and compared different imputation methods under different circumstance. To summarize the simulation results, CC method and Mean Imputation method yield reasonable mean values to accept null hypothesis of its analysis. Accuracy of CC method doesn't influence covariance effect (Case I) and slope difference (β). However, as stated, CC method has a disadvantage of losing sample size and statistical power. In addition, MSE in CC method shows higher values than other methods, which indicates appearance of some bias. Mean Imputation method forces the variance decrease. In this simulation study, such a disadvantage doesn't appear. However, we need to use its imputation with consideration for potential change of the distribution especially when missingness is large. LOCF method often fails normality test when missingness increases and slope is large. Literature stated that any imputation methods should be satisfactory when missingness was less than 5% [8,11]. However, LOCF method in this simulation study doesn't even satisfy in 5% missingness when slope is large. Therefore, the slope must be small for LOCF method to be fairly accurate. Finally, according to this simulation, MI method has poor performance for normality test. Furthermore, MI method doesn't seem to exert its efficiency when missingness is small. In theory, MI method functions to

determine imputed values by maximum likelihood. That is, the imputed values influence overall dataset. Therefore, the accuracy in MI method probably stabilizes when the dataset at each time point has standard normal distribution N(0,1).

In conclusion, CC method approximates a good imputation method with condition of small missing percentage and large sample size to fix its disadvantages. Mean Imputation method should be used when samples are close to each other. As long as MCAR assumption satisfies, its method imputes fairly well. When values at each time point are small, then LOCF method estimates appropriate values. However, its method is not recommended when values at each time point are large and dataset is other than MCAR assumption. MI method imputes poor approximation in normality for small missingness. To consider using its method with small missingness, the number m of imputed dataset should be increased for accuracy. Its method is more adaptable for large missing data.

In further research, accuracy of imputation in categorical dataset is expected to investigate. Since MI doesn't procedure categorical data for imputation [16], single imputation such as LOCF method or mean/median method must feature for missing data in categorical analysis. Furthermore, CC method must nominate as one of imputation methods for handling missing data in categorical analysis.

References

[1] A.M.G. Ali., S-J Dawson., F.M. Blows., E. Provenzano., I.O. Ellis., L. Baglietto., D. Huntsman.,C. Caldas and P.D. Pharoah, Comparison of methods for handling missing data on immunohistochemical markers in survival analysis of breast cancer, British Journal of Cancer, **104** (2011), 693-699.

[2] F. Barzi and M. Woodward, Imputation of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies, American Journal of

Epidemiology, **160** (2004), 34-45.

[3] M. Blankers., M.W.J Koeter., and G.M. Schippers, Missing Data Approaches in eHealth Research: Simulation Study and a Tutorial for Nonmathematically Inclined Researchers, Journal of Medical Internet Research, **12** (2010), e54

[4] R.T. Donders., J.M.G. Van der Heijden., T. Stijnen., and K.G.M. Moons, Review: A gentle introduction to imputation of missing values, Journal of Clinical Epidemiology, **59** (2006), 1087-1091

[5] G.M. Fitzmaurice., N.M. Laird., and J.H. Ware, Applied Longitudinal Analysis, Wiley, New York, 2004

[6] D. Hedeker., R.J. Mermelstein., and H. Demirtas, Analysis of binary outcomes with missing data: missing=smoking, last observation carried forward, and a little multiple imputation, Addiction, **102** (2007), 1564-1573

[7] J.M.G. Van der Heijden., R.T. Donders., T. Stijnen., and K.G.M. Moons, Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostics research: A clinical example, Journal of Clinical Epidemiology, **59** (2006), 1102-1109

[8] D.F. Heitjan, Annotation: What Can Be Done about Missing Data? Approaches to Imputation, American Journal of Public Health, **87** (1997), 548-549

[9] P. Lane, Handling drop-out in longitudinal clinical trial: a comparison of the LOCF and MMRM approaches, Pharmaceutical Statistics, **7** (2008), 93-106

[10] R.J.A Little and D.B Rubin, Statistical Analysis with Missing Data, second edition, Wiley, New York, 2002

[11] A. Marshall., D.G. Altman., and R.L. Holder, Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study, BMC Medical Research Methodology, **10** (2010), 112

[12] G. Molenberghs., H. Thijs., I. Jansen., and C. Beunckens, Analyzing incomplete longitudinal clinical trial data, Biostatistics, **5** (2004), 445-464

[13] M. Nakai and W. Ke, Review of the Methods for Handling Missing Data in Longitudinal Data Analysis, International Journal of Mathematical Analysis, **5**(1) (2011), 1-13

[14] P.L. Roth, Missing data: A conceptual review for applied psychologists, Personnel Psychology, **47**(3) (1994), 537-560

[15] J. K. Strike., K.E. Emam., and N. Madhavji, Software Cost Estimation with Incomplete Data, IEEE Transactions on Software Engineering, **27**(2001), 890-908

[16] L. Tang., T.R. Belin., and J. Song, A Comparison of imputation methods for missing data in a multi-center randomized clinical trial: The impact study, Joint Statistical Meetings-section on Health Policy Statistics (2002), 3430-3435

[17] I.R. White and J.B. Carlin, Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values, Statistics in Medicine, **29** (2010), 2020-2931

[18] A. M. Wood., I.R. White., and S.G. Thompson, Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journal, Clinical Trials, **1** (2004), 368-376

Received: June, 2011