Effectiveness 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

Based on previous simulation dataset in AR(1) longitudinal studies, three basic imputation methods (Complete Case, Last Observation Carried Forward and Mean Imputation) are compared to examine the effectiveness of its methods. Furthermore, these methods comprehend and classify each performance with empirical mean and mean square error. Then, specific conditions for each method have defined to perform in real dataset. The final section provides summary.

Mathematical Subject Classification: 62-07, 62H15

Keywords: Longitudinal analysis, Imputation, Missing data

1. Introduction

Missing data are one of unavoidable problems in longitudinal analysis. Due to the technological development, several imputation methods are available to fill in missing data. But, there are not many references to check the compatibility of each method. In this paper, we focus on three basic imputation methods; Complete Case method, Mean imputation method, and Last Observation Carried Forward (LOCF) with various conditions. More detailed characteristics for each method are referred to [2,4,6].
2. Simulation and data structure

This paper is an extended simulation research by Nakai [5,6]. Multiple Imputation has been omitted from this research because it did not satisfy Normality test in most situations based on previous studies. For a dataset, suppose repeated measurements \( Y_{it} (i=1,\ldots,100; t=1,\ldots,5) \) are generated from a multivariate normal distribution with mean response \( \text{E}(Y_{it}) = \beta_0 + \beta_1 \cdot t \) where \( \beta_0 \) = intercept, \( \beta_1 \) = slope and \( \rho = \text{correlation for } \rho \geq 0 \). In this paper, simulated sample size is increased from \( N=200 \) to \( N=300 \) different random longitudinal datasets in SAS®. Since samples have 100 increments for each simulation, investigation has not just changed 100 samples but \( 100^2 \) more factors to evaluate, which reflects a better result and leads to have more justification in statistical analysis. Also, 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 [4].

Assuming that the first occasion is fully observed, simple random sampling without replacement is used to make Missing Complete at Random (MCAR) [3] dataset 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;
- Case III: 0%, 10%, 20%, 30% and 50% at time point 1, 2, 3, 4, 5, respectively.

The experiment itself consists of mean of the 300 empirical means and mean square error (MSE) from a fitted mean \( E(Y_{it}) \). Moreover, 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 \( \beta_0 = 10. \) In addition, we test by changing \( \rho = 0.2, \rho = 0.9, \) and \( \sigma^2 = 10. \)

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

3. Simulation Results

At first, the main improvement for this simulation is that a value of MSE has been decreased. A largest MSE among samples was 0.001 and most values were around \( 1 \times 10^{-7}. \) This indicates the accuracy of its simulation has been improved by increasing sample size and the results are more concrete than previous studies. In simulation result, the lack of stability clearly appears for LOCF method at slope=2 in all cases and all conditions. Also, MSE values are larger compared to other methods. Since this pattern does not occur at slope=0.1, the primary effect to
Influence for accuracy of LOCF method is slope, that is, the range of each interval. For slope=0.1, 30% missing percentage is the borderline for LOCF method. However, when \( \sigma^2 = 10 \), LOCF method works fine. Therefore, for a large slope, LOCF method is not beneficial method to consider for any missing percentages. For a small slope, LOCF method is limited to less than 30% missing percentage unless \( \sigma^2 \) is large.

For a case I and \( \rho = 0.9 \), normality check does not satisfy in most situations. There seems to be some relationship among a small slope, AR(1) covariate structure, and \( \rho \). Besides, for \( \rho = 0.2 \), normality of Complete Case method is not stable from 20% or larger missing percentage for a large slope. Therefore, a different correlation coefficient may relate to the accuracy for Complete Case method. On the other hand, mean imputation is a stable and unbiased method out of three methods. One reason is that parameters in this experiment are total mean and MSE. And, its MSE has decreased extremely. Thus, imputed values for mean imputation have a good estimation for a true total mean. However, we need to keep in mind that there are several disadvantages for mean imputation as well [1,7].

4. Summary

In the conclusion, LOCF method does not impute the appropriate values for a large slope not matter what other conditions are. We find out that Complete Case method shows inconsistent result depended on correlation coefficient. Mean imputation is the most reasonable method of all three imputations. In the future research, we need to change its parameters to explore more different aspect as well as look into more detail to fit Missing at Random (MAR) by expanding the limited range of imputation methods.

References


Received: September, 2011