Int. Journal of Math. Analysis, Vol. 5, 2011, no. 1, 1 - 13

Review of the Methods for Handling Missing Data in Longitudinal Data Analysis

Michikazu Nakai and Weiming Ke

Department of Mathematics and Statistics
South Dakota State University
Brookings, SD 57007, USA
Weiming.Ke@sdstate.edu

Abstract

Even in well-controlled situations, missing data always occur in longitudinal data analysis. Missing data may degrade the performance of confidence intervals, reduce statistical power and bias parameter estimate. In this paper, we review and discuss general approaches for handling miss data in longitudinal studies. We first illustrate the mechanism of missing data. Then we focus on the methods for handling missing values in longitudinal data analysis. In the end, we summarize and discuss the characteristics of each method.

Mathematical Subject Classification: 62J10, 62J12, 62-07

Keywords: Missing value, longitudinal analysis, complete case, imputation, selection model, pattern-mixture model

1 Introduction

Longitudinal data analysis is defined as the study of the data resulting from the observations of subjects which are repeatedly measured over a series of time-points [6].

Missing data are commonly occurred in longitudinal studies and are defined as that no data values are stored for the variable in the current observation. Missing data include the problem of attrition or “drop-out”, that is, some individuals “drop out” of the longitudinal study or withdraw from the study before its intended completion. Hence, data record for the subject terminates prematurely.

Missing data have three important implications for longitudinal data analysis [10]. First, when longitudinal data are missing, the data set is necessarily unbalanced over time since not all individuals have the same number of repeated measurements at a common set of occasions. This imbalance data let the methods of analysis differ from the one of balanced data. Secondly, when there are missing data, there must necessarily be some loss of information. The missingness spread sporadically over many subjects and how highly correlated the missing data are with the observed data will affect loss of precision. Finally, under certain circumstances, missing data can introduce bias and thereby lead to misleading inferences about changes in the mean response. The higher attrition is likely to have bias and the potential for serious bias makes the longitudinal analysis more complicated. Some important references in the field of missing data can be found in [1, 5, 7, and 11].

This article is organized as follows. The next section provides some background information on the mechanisms of missing data. Then, we present an overview of longitudinal data models and implications of missing values on these models. The subsequent section presents some of the most frequently used methods for handling missing values in longitudinal data analysis and discuss their advantages/disadvantages to yield valid analysis. It is concluded with some remarks.

2 Missing Values

Rubin first developed a very useful taxonomy for describing the assumptions concerning the dependence of the missingness process [13]. Generally there are three types of missing data mechanisms.

2.1 Missing At Random

Data are said to be Missing At Random (MAR) when the probability that responses are missing depends on the set of observed responses, but is unrelated to the specific missing values that, in principle, should have been obtained.

In notation, when Y= complete data matrix, Y^O= observed part of Y, Y^M= missing part of Y and R is missing data indicator matrix where R_{ij}=1 for missing, 0 for observed, then $P(R|Y, \varphi) = P(R|Y^O, \varphi)$ for all Y^M, φ.
where φ denotes unknown parameter.

When dropout is MAR, it means that the probability of dropout at each occasion is conditionally independent of current and future responses, given the history of the observed responses prior to that occasion.

2.2 Missing Completely At Random

Data are said to be Missing Completely At Random (MCAR) when the probability that response are missing is unrelated to either the specific values that in principle, should have been obtained or the set of observed responses. MCAR is a special case of MAR, and occurs when the distribution doesn't depend on observed data, either.

In notation, $P(R|Y, \varphi) = P(R| \varphi)$ for all Y, φ.

The distinction between MCAR and MAR is that missingness cannot depend on observed values of the dependent variable Y^O in MCAR, but can in MAR. Therefore, the test of MCAR is based on analysis involving Y^O.

2.3 Not Missing At Random

Data are said to be Not Missing At Random (NMAR) when the probability that responses are missing depends on both the set of observed responses and the specific missing values that, in principle, should have been obtained. Sometimes it is referred as Missing At Not Random (MANR) or Missing Not At Random (MNAR).

Since the probability of missing data is related to at least some elements of Y^M, NMAR is often referred as non-ignorable missingness. The term non-ignorable refers to the fact that missing data mechanism cannot be ignored. When missingness is non-ignorable, it means that we cannot predict future unobserved responses, conditional on past observed responses; instead, we need to incorporate a model for the missingness mechanism.

2.4 Ignorable Mechanism

In contrast, MCAR and MAR are often referred to as ignorable mechanisms. The ignorable mechanisms have two conditions to satisfy. One is that the data are MAR. Secondly, the parameters that govern the missing data process are unrelated to the parameters to be estimated. Some important references in the field of missing mechanism can be found in [2, 4, 9, 11, and 14].

3 Longitudinal Data Analysis

Now, we discuss the statistical models for longitudinal data analysis.

3.1 ANOVA and MANOVA

The most popular methods in longitudinal analysis are univariate and multivariate analysis of variance, called ANOVA and MANOVA respectively. These methods are well-understood and most developed. Both models assume interval measurement and normally distributed errors that are homogeneous across groups. The weak aspect of these methods is that they only estimate and compare the group means but are not informative about individual growth. As an assumption, ANOVA requires compound symmetry which has little validity for longitudinal data. Also, MANOVA assumes a general form for the correlation of repeated measurements over time. The main difference between ANOVA and MANOVA is that MANOVA approach must discard all missing data because MANOVA treats the repeated measures as one vector and the entire data vector must be complete for the subject to be included in the analysis.

3.2 Mixed-Effect Regression Model

Next method is Mixed-effect Regression Model (MRM). This method is quite widely used for analysis of longitudinal data and can be extended to ordinal outcomes or nominal or count outcomes that have a Poisson distribution as well. MRM is more flexible in term of repeated measures and does not require restrictive assumptions concerning missing data across time and the variance–covariance structure of the repeated measures. Hence, MRM can be used for unbalanced/incomplete longitudinal data. The disadvantage of MRM is that full-likelihood methods are more computationally complex than quasi-likelihood methods.

3.3 Generalized Estimating Equation

Last method is called Generalized Estimating Equation (GEE) introduced by Liang and Zeger [9]. They are extension of generalize linear model (GLM) to longitudinal analysis using quasi-likelihood. GEE treats covariance structure as a nuisance and GEE is not concerned about variance of each data. Besides, GEE is often used as a general and computationally convenient method. The disadvantage is that missing data are only ignorable if the missing data are explained by covariates in the model. That is, GEE doesn't perform well unless the missing data are MCAR. This is a more stringent assumption than MRM since MRM analysis is acceptable as long as the mean and variance-covariance structure are correctly modeled. Therefore GEE models have somewhat limited applicability to incomplete longitudinal data.

4 Missing Values in Longitudinal Data Analysis

In this section, we describe some of the most commonly used methods and discuss the characteristics of the method to yield valid analysis. Some important references in the field can be found in [1, 3, 4, 7, 8, and 11].

4.1 Complete Case Analysis

One approach to handling missing values is to simply omit all cases with missing values at any measurement occasion. This is called a Complete-Case Analysis (CCA). The advantage of this method is that it can be used for any kind of statistical analysis and no special computational methods are required. However, it will yield unbiased estimate of mean response trends only when the missingness is MCAR. When the missing data are not MCAR, the results from CCA may be biased because the complete case can be unrepresentative of the full population. Also, it can result in a very substantial loss of information by deleting all case with missing value, and this gives an impact on reduced statistical precision and power. If the missing-data problem can be resolved by discarding only a small part of the sample, then the method can be quite effective. But, CCA is very problematic and is rarely an acceptable approach to the analysis. This method can be done using PROC REG or PROC FACTOR in SAS®.

Suppose θ is a scalar parameter of interest. Denote $\hat{\theta}_{CC}$, $\hat{\theta}_{NM}$ and $\hat{\theta}_{EFF}$ the complete case estimator, estimator based on the hypothetical complete data and the efficient estimator on the observed data, respectively. From

Var($\hat{\theta}_{CC}$) = Var($\hat{\theta}_{NM}$)(1 +Δ^{*} CC),

$\mathrm{Var}(\hat{\theta}_{CC}) = \mathrm{Var}(\hat{\theta}_{EFF})(1 + \Delta_{CC})$,
where Δ^{*}_{CC} and Δ_{CC} can be used to evaluate the relative efficiency of $\hat{\theta}_{CC}$, that is, the proportional increase in variance from the loss of information.

4.2 Available Case Analysis

Another approach to handling missing values is Available Case Analysis. This is a general term for a variety of different methods that use the available information to estimate means and covariance. It can readily incorporate vectors of repeated measures of unequal length in the analysis. The popular method in available case analysis is pair-wise deletion method. In this method, a covariance (or correlation) matrix is computed where each element is based on the full number of cases with complete data for each pair of variables. The attempt is to maximize sample size by not requiring complete data on all variables in the model. In general, Available Case Analysis is more efficient than CCA because it incorporate the partial information obtained from those who are missing. The disadvantage of this method is that the sample base changes from variable to variable according to the pattern of missing data. This variability in the sample base creates practical problems such as the determination of sample size and degree of freedom. Also, it yields biased estimates of treatment comparisons unless missing data are MCAR. This method can be done using PROC CORR in SAS®.

4.3 Single Imputation

Third approach to handling missing values is Single Imputation. This is a method that involves replacing an incomplete observation with complete information based on an estimate of the true value of the unobserved variable. It is widely used in practice because the analysis is straightforward once imputation is done. The obvious disadvantage of single imputation is that imputing a single value treats that value as known, and thus without special adjustments, single imputation cannot reflect sampling variability under one model for non-response or uncertainty about the correct model for non-response.

One of the most widely used imputation methods in longitudinal analysis is Last Observation Carried Forward (LOCF). This method is for every missing value to be replaced by the last observed value from the same subject. Whenever a value is missing, the last observed value is substituted. LOCF is routinely used in the pharmaceutical industry, and elsewhere, in the analysis of randomized parallel group trials for which a primary objective is to test the null hypothesis of no difference between treatment groups. Although simple, this method makes the strong assumption that the value of the outcome remains unchanged after missing, which seems likely to be unrealistic in many setting. The one of a few settings where this assumption might conceivably be appropriate is when missing data is due to recovery or cure.

Let Yi= (Y_{i1},…Y_{ik}) be a (M×1) complete data vector of outcomes for subject *i*, possibly incompletely observed. Suppose R_i denote a missing data indicator, with $R_i = 0$ for complete cases and $R_i = k$ if a subject drops out between the (*k*-1) th and *k* th observation time.

For cases *i* with $M_i = k$, $Y_{it} = Y_{i,k-1}$ where t = k,…,M.
That is, missing values are imputed by the last recorded value of a respondent.

Another single imputation method to handling missing value is mean imputation. This method is to fill in any missing values with mean of the non-missing values. It therefore assumes that mean of the variable is the best estimate for any observation that has missing value on the variable. Even though it is simple to impute, this strategy can severely distort the distribution for the variable, leading to complication with summary measures including underestimates of the standard deviation. Also, the missing values require being MCAR as an assumption. Therefore, mean imputation is getting an unaccepted method nowadays. This method can be done using PROC STANDARD in SAS®.

Next single imputation method to handling missing value is hot-deck imputation. This method replaces missing values with values from similar responding units in the sample. The imputed values do not distort the distribution of the sampled values. Hot-deck imputation is common in survey practice and can involve very elaborate schemes for selecting units that are similar for imputation. The disadvantage is that it is difficult to find such similar responding units in the sample. Also, the distortion of correlations and covariance can be the serious drawback with this method.

Suppose $y_1 \cdots y_r, y_{r+1}, \cdots y_n \sim \text{iid}(\mu, \sigma^2)$, the first *r* units are recorded. Then,

$$\bar{Y}_{HD} = \frac{\{r\bar{Y}_R + (n-r)\bar{Y}^*_{NE}\}}{n}$$

where $\bar{Y}^*_{NE} = \sum_{i=1}^{r} \frac{H_i Y_i}{n-r}$ and H_i is the number of times Y_i is used as a substitute for a missing value of Y. The properties of $\bar{Y}_{HD}$ depend on the procedure used to generate the number $\{H_1, H_2 \cdots H_r\}$.

Last single imputation method to handling missing value is Expectation Maximization

(EM) algorithm. EM algorithm is an iterative algorithm that finds the parameters which maximize the log likelihood when there are missing values. It capitalizes on the relationship between missing data and the unknown parameters of a data model [12]. A disadvantage of EM algorithm is that its rate of convergence can be painfully slow when there is a large fraction of missing values. Each iteration of EM consists of an E step (expectation step) and M step (maximization step). Given a set of parameter estimates, E-step calculates the conditional expectation of the complete data log likelihood given the observed data and the parameter estimates. Suppose θ^t is the current estimate of the parameter θ. Then,

$Q(\theta| \theta^t) =\int g(\theta| Y) f(Y^M|Y^O, \theta=\theta^t)\, dY^M$

Where $g(\theta| Y)$ is the complete data log likelihood. Given a complete data log likelihood, the M step finds the parameter estimates to maximize the complete data log likelihood from E step.

$Q(\theta^{(t+1)}| \theta^t) \geq Q(\theta| \theta^t)$ for all θ

And, these two steps are iterated until the iteration converges.

4.4 Multiple Imputation

The most popular imputation method to handling missing value is multiple imputation (MI). The MI is to replace each missing item with two or more acceptable values, representing a distribution of possibilities. The advantage of the method is that once the imputed data set have been generated, the analysis can be carried out using procedures in virtually any statistical package, which makes the analysis simple. Also, the inferences such as standard error or p-value etc obtained from MI are generally valid because they incorporate uncertainty due to missing values. The MI can be highly efficient even if the number of imputation is relatively small, especially when between-imputation variance is not too large. However, there are some disadvantages in MI. First, since we impute some values into missing values, missing value individuals are allowed to have varying probability. Thus, individual variation is being ignored. Secondly, the uncertainty inherent in missing values is ignored because the analysis doesn't distinguish between the observed and imputed values. At last, the only disadvantage of MI over single imputation is that it takes more work to create the imputations and analyze the results. However, from SAS® version8.2, they have developed PROC MI and PROC MIANALYZE, which improve the computing environment and save time to analyze and space to store data.

The multiple imputation inference involves three distinct phases: (a) The missing data are filled in *m* times to generate *m* complete data sets (b) The *m* complete data sets are analyzed by using standard procedures (c) The result from the *m* complete data sets are combined for the inference. PROC MI creates imputed data sets for incomplete multivariate data. Its uses methods that incorporate appropriate variability across the *m* imputations. SAS multiple imputation procedures assume that the missing data are ignorable. Once the *m* complete data sets are analyzed by using standard procedures such as PROC REG, PROC GLM or PROC MIXED, then PROC MIANALYZE can be used to generate valid statistical inference about these parameters by combining results from *m* complete data sets.

There are three imputation mechanisms in PROC MI. The method of choice depends on the type of missing data pattern. For monotone missing data patterns, either a regression method or propensity score method can be used. For an arbitrary missing data pattern, a Markov chain Monte Carlo (MCMC) method can be used. Without the detail of theatrical methods, Regression method is fitted for each variable with missing values with previous variables as covariates. Propensity Score method is that observations are grouped based on propensity scores, and an approximate Bayesian bootstrap imputation is applied to each group. At last, MCMC constructs a Markov chain long enough for distribution of the elements to stabilize to a common distribution [15].

4.5 Selection Model and Pattern-Mixture Model

Finally, when longitudinal data are NMAR, the data are called non-ignorable. In the case, standard statistical models are not valid and can yield badly biased results. The models to be used for handling missing values in such a situation are selection model and pattern-mixture model.

The selection model specifies the model for both longitudinal and missing process simultaneously [4]. Both models depend on random subject effects, most or all of which are shared by both models. That is why selection model is sometimes called shared parameter model. In the model, one uses a complete data model for the longitudinal outcomes, and then the probability of missingness is modeled conditional on the possibly unobserved outcomes. The joint distribution of M and Y of the selection model is given by

$$f\,(\mathrm{M,Y}|\ \theta,\psi) = f\,(\mathrm{Y}|\theta)\, f\,(\mathrm{M}|\mathrm{Y},\psi)$$

where the first factor characterizes the distribution of Y in the population, the second factor models the incidence of missing value as a function of Y, and θ and ψ are unknown vector parameters and both are distinct [11].

For identification, the set of outcomes is usually restricted in some way. That is, with selection model, identification comes from unverifiable models for the dependence of the dropout probabilities on the unobserved outcomes. However, it tends to be very difficult to determine the identifying restrictions that must be placed on the model. At last, even though the selection model is easy to formulate hypotheses about dropout process, it is computationally intractable because it is difficult to infer how assumptions on dropout process translate into assumptions about distribution of unobserved response.

In contrast, pattern-mixture model use missing value pattern as between-subject variable in the longitudinal model. The idea behind pattern-mixture model is to explicitly model the missing data distribution by first identifying different patterns of missing data and then including parameters in the outcomes model that capture this effect [6]. Pattern-mixture model write as follows:

$$f\,(\mathrm{M,Y}|\ \xi,\ \omega) = f\,(\mathrm{Y|M},\ \xi)\, f\,(\mathrm{M}|\ \omega)$$

where the first distribution characterizes the distribution of Y in the strata defined by different patterns of missing data M, the second distribution models the incidence of the different patterns. ξ and ω are unknown vector parameter and both are distinct [11].

The pattern-mixture method is computationally simple, however it makes explicit assumptions about distribution of unobserved response since missingness process is not immediately transparent. To decide on an appropriate grouping of the missing-data pattern, three things need to be considered. At first, the grouping has the sparseness of the patterns. If a pattern has very few observations, it may not make sense to treat it a separate group in the analysis. Also, we need to be careful with the potential influence of the missing-data pattern on the response variable. In longitudinal study, it is often reasonable to assume that the intermittent missing observations are randomly missing. Thus, how we consider groupings may cause difference in the analysis. At last, it is important to realize that, by knowing what one is interested, some observations may provide no information for missing-data pattern.

5 Discussion

Missing values are crucial knowledge to know when dealing in longitudinal analysis. In this paper, we discussed general types of missing data mechanisms and introduced basic methods for handling missing values. Complete-case method, pair-wise deletion method and mean imputation require the missing data to be MCAR for unbiased estimate. Last Observation Carried Forward is a simple method to use, but it makes the strong assumption that the value of the outcome remains unchanged after missing, which seems likely to be unrealistic in many setting. Hot-deck imputation replaces missing values with values from similar responding units in the sample. However, the distortion of correlations and covariance can be the serious drawback with this method. Both EM algorithm and MI consist of Bayesian technique. EM algorithm is an iterative algorithm that finds the parameters which maximize the log likelihood when there are missing values. It can be very slow to converge with large fractions of missing data. MI also has similar disadvantage as EM algorithm. We may solve such a problem using PROC MI and PROC MIANALYZE in SAS®, which improve the computing environment and save time to analyze and space to store data. When data are NMAR, then selection model and pattern-mixture model can be used. The selection model specifies the model of both longitudinal and missing process simultaneously. The pattern-mixture model explicitly produces the missing data distribution by first identifying different patterns of missing data and then including parameters in the outcomes model that capture this effect. However, in reality, it is very difficult to distinguish among MCAR, MAR, and NMAR. The assumptions about missing data are almost impossible to assess with the observed data. Hence, knowing several methods to determine the appropriate analytic approach and testing to compare the results are essential ability for accurate statistical analysis.

References

[1] P. Allison, Missing Data, Sage Publications Inc, California, 2001

[2] H. Demirtas, Assessment of Relative Improvement Due to Weights Within Generalized Estimating Equations Framework for Incomplete Clinical Trials Data, Journal of Biopharmaceutical Statistics, **14**, 1085-1098,2004

[3] H. Demirtas & J. L. Schafer, On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out, Statistics in Medicine, **22**, 2003

[4] G.M. Fitzmaurice, Methods for handling dropouts in longitudinal clinical trials, Statistica Neerlandica, **57**, 75-99, 2003

[5] G.M. Fizmaurice, N.M, Laird, and J.H. Ware, Applied Longitudinal Analysis, Wiley-Interscience, New Jersey, 2004

[6] D. Hedeker & R. D. Gibbons, Application of Random-Effects pattern-Mixture models for Missing Data in Longitudinal Studies, Psychological Methods, **2**, 64-78, 1997

[7] D. Hedeker & R.D. Gibbons, Longitudinal Data Analysis, Wiley-Interscience, New Jersey, 2006

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

[9] K.-Y. Liang and S. L. Zeger, Longitudinal Data Analysis Using Generalized Linear Models, Biometrika, **73**, 13-22, 1986

[10] R.J.A Little, Modeling the drop-out mechanism in repeated-measures studies, Journal of the American Statistical Association, **90**, 1112-1121,1995

[11] R. J. A. Little & D. B. Rubin, Statistical analysis with missing data-second edition, Wiley-Interscience, New Jersey, 2002

[12] F. V. Nelwamondo, S. Mohamed & T. Marwala, Missing data: A comparison of neural network and expectation maximization techniques, Current Science, **93**, 2007

[13] D. B. Rubin, Inference and Missing Data, Biometrika, **63**, 581-592, 1976

[14] J. L. Schafer & J. W. Graham, Missing Data: Our View of the State of the Art, Psychological Methods, **7**, 147-177, 2002

[15] Y. C. Yuan, Multiple Imputation for Missing Data: Concepts and New Development, Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, 267-25, SAS Institute, Cary NC, 2000

Received: March, 2009