

Incremental Methods of Imputation for Longitudinal Data with Informative Missingness

Naum M. Khutoryansky, Novo Nordisk Pharmaceuticals, Inc., Princeton, NJ

ABSTRACT

In longitudinal clinical trials, missing data are mostly related to dropouts. Some dropouts appear completely at random. The source for other dropouts is withdrawal from trials due to lack of efficacy. One of the approaches to comply with the intent-to-treat principle is the imputation of incomplete data. When the dropout process is related to the outcome process (informative dropouts), it creates tremendous challenges in analyzing such data. No commercial software currently considers the dropout mechanisms in dealing with informative or non-random dropout. Consequently, the results can be biased and misleading. This paper deals with the incremental methods of imputation applied to incomplete longitudinal data sets with informative dropouts. It is shown by simulation that the incremental methods are more precise than some other imputation methods (including the last observation carried forward method, the multiple imputation method, mixed models). The drop-out mechanism depends on individual previously and currently observed values of a secondary endpoint correlated with the primary endpoint.

INTRODUCTION

Comparison of drug treatments involves longitudinal measurements of efficacy variables. When individual elements of longitudinal data are missing, the data are incomplete. The incompleteness can have different patterns and different causes. Missing values occur due to a drop-out process if the missing-data process induces a monotone pattern of missing values, i.e. a patient has observations up to a time point and none after that; otherwise the missing values are intermittent.

Statistical analyses can be affected by missing data in two ways. First, there can be a loss of efficiency due to reduced sample size if potentially informative subjects are completely excluded from the analysis because one or more variables are missing in the model. Second, and more serious, results can be biased if missingness itself is related, directly or indirectly, to some of the relevant factors. These problems are particularly important in longitudinal studies, because few subjects will have complete data for all visits, and the fact that missingness could be an indicator of an adverse outcome.

In many cases the drop-out processes possess the property that the missing data are *missing at random* (MAR) (i.e. the drop-out process depends on the observed measurements). A special case of MAR drop-outs is the MCAR class when the missing data are *missing completely at random* (i.e. the drop-out and measurement processes are independent). Data are missing not at random (MNAR) if missingness depends on (is explained by) the unobserved outcomes. In this case, many methods of imputation can be biased and misleading due to the "ignorability of missingness" assumption behind these methods.

The response variables in longitudinal trials are often correlated. Many situations arise in which two response variables are observed simultaneously on each patient at each occasion. For example, in diabetes longitudinal studies two glycemic variables are of most importance: Glycosylated Haemoglobin (HbA1c) and Fasting Plasma Glucose (FPG). These variables are correlated at each visit. In some trials, only the FPG level is used for withdrawal criterion: if the FPG response of a patient at a visit exceeds a pre-specified threshold, the patient is withdrawn from the trial after this visit.

Bivariate response (HbA1c, FPG) can often be considered as a Markov process which means that the dependence of the current value of this response on its history of during the trial can be reduced to the dependence on its most recent previous value. The drop-out mechanism described above can also be presented in the framework of Markov processes depending on the current and the most recent previous observations.

This paper is an attempt to compare several methods of means and variances estimation (two of which are presented in SAS procedures) for simulated longitudinal bivariate incomplete data satisfying the Markov process assumptions with the drop-out mechanism described above. The methods under consideration are the last observation carried forward method (LOCF), multiple imputation method [1] (represented by PROCs MI and MIANALYZE), linear mixed models method [2] (represented by PROC MIXED and incremental method [3]. The comparison of these methods for simulated longitudinal univariate data with ignorable dropout mechanisms were presented in [3]. Consideration of bivariate correlated responses with the drop-out mechanism involving only one variable much closer resembles the practice of the diabetes clinical trials than the univariate approach. Such an approach with ignorable (MAR) dropouts was presented in [4]. In the current paper, we consider the bivariate response with non-ignorable dropouts.

INFORMATIVE DROP-OUT MECHANISM

Consider time points t_1, t_2, \dots, t_n . Now suppose that two response variables X and Y are represented by random variables X_i and Y_i at time points t_i ($i=1, \dots, n$). Denote pair $\{X_i, Y_i\}$ by Z_i . In this paper, we assume that Z_i is an inhomogeneous Markov process which means that

$$P(X_i < x, Y_i < y | Z_1, \dots, Z_{i-1}) = P(X_i < x, Y_i < y | Z_{i-1}) \quad (1)$$

Random variables X_i and Y_i are assumed to be correlated with correlation coefficient ρ_i .

The drop-out mechanism implemented is based on a threshold value T for random variable X_i . Let X_{ij} and Y_{ij} be the values of X_i and Y_i for the j th subject. If a linear combination of $X_{k-1,j}$ and X_{kj} (with coefficients a and b where $a+b=1$) is greater than T then the j th subject is a drop-out starting from time point t_k . It means that X_{kj} and Y_{kj} are not observable. Therefore, the following properties hold for a drop-out at time t_k :

$$\begin{aligned} &\text{If} \\ &\quad X_{ij} \leq T_i, i = 1, 2, \dots, k-1, \\ &\quad aX_{k-1,j} + bX_{kj} > T \\ &\text{then} \\ &\quad X_{ij} \text{ and } Y_{ij} \text{ are missing for } i \geq k. \end{aligned} \quad (2)$$

In the next sections, we consider and compare estimation of the means and variances of random variables X_i and Y_i using the four methods mentioned above if the dropout mechanism satisfies (2).

DATA SET SIMULATION

We consider the simulation based on longitudinal data sets resembling that in diabetes clinical trials. Two endpoints comprising the bivariate response are FPG (Fasting Plasma Glucose) and HbA1c (Glycosylated Haemoglobin). The drop-out mechanism of type (2) is based on values of FPG with threshold T .

As an example we consider such a trial with six time points of measurements starting from baseline. The first three time increments are equal to four weeks, the next two are six weeks. We assume that the baseline distributions of FPG and HbA1c can be approximated by normal distributions and specify time behavior of these variables resembling their dependence on time in real diabetes trials for different treatments.

Two longitudinal variables X and Y are correlated at each time point t_i with a correlation coefficient $\rho_i = \rho$. The coefficients a and b had values 0.5. It means that the value of FPG used in comparison with the threshold is the average of the current (possibly unobserved) and previous FPG values. The correlation coefficient was chosen to be 0.5 or 0.8. The FPG threshold for the drop-out process was chosen to render about 50% of the total dropout rate.

The data sets (complete and incomplete) simulated by this approach can be used to compare different methods of parameter estimation for the complete data set using only the incomplete data set. The results obtained by each method under consideration could be assessed by their comparison with the corresponding results for the complete data set.

INCREMENTAL METHODS

Let D_{ij} be a bivariate increment of $Z=(X, Y)$ for subject i from time point j to time point $j+1$. Next, let Z_{ij}^* and Z_{ij}^{\dagger} be observed and missing values of Z , respectively. Similarly, let D_{ij}^* and D_{ij}^{\dagger} be observed and missing bivariate increments, respectively. Denote by $D_{\cdot j}^*$ the bivariate sample mean of observed increments D_{ij}^* at time point j . Suppose that values Z_{i1} are all observed. The incremental mean method [3] creates two matrices: a basic imputed matrix and an uncertainty matrix. To build the basic imputed matrix B the unknown bivariate increments D_{ij}^{\dagger} are prescribed to be equal to $D_{\cdot j}^*$ and, next, the missing values $Z_{i,j+1}^{\dagger}$ are replaced by

$$Z_{i,j+1}^{\dagger} = Z_{ij}^{\dagger} + D_{\cdot j}^* \quad \text{for } j = 1, \dots, k-1 \quad (3)$$

Where bivariate values Z_{ij} are observed or calculated on the previous step. The elements B_{ij} that coincide either with Z_{ij}^* or with Z_{ij}^{\dagger} (calculated step by step as specified above) create the basic imputed matrix B . The approximate bivariate mean values of Z at each time point j will be calculated as the bivariate sample mean $B_{\cdot j}$ using the bivariate j th column of matrix B . The columns of matrix B can be used to calculate the variances of Z at each time point. However, these calculated values V_j^{\dagger} underestimate the real variances due to the lack of uncertainty in (3) and will be named the partial variances. To induce an additional uncertainty it is proposed to build an uncertainty matrix V . Let U_j^* be the sample variance of the set $\{D_{ij}^*\}$ for fixed j . Let R_{ij} be equal to 1 if Z_{ij} is observed and 0 otherwise. Then the elements V_{ij}^{\dagger} of matrix V are calculated as follows:

$$\begin{aligned} V_{ij}^{\dagger} &= 0; \quad V_{ij}^{\dagger} = (V_{i,j-1}^{\dagger} + U_j^*) (1 - R_{ij}) \\ & \quad j = 2, \dots, k. \end{aligned} \quad (4)$$

Therefore, if Z_{ij} is observed then $V_{ij} = 0$ (there is no uncertainty). Otherwise, an additional variance (as an uncertainty measure) is accumulated over time until time point j for each subject i . The pooled additional variance V_{ij} at time point j is defined as the average of V_{ij} over all the subjects. The total variance V_j at time point j is defined as

$$V_j = V_j' + V_j'' \quad (5)$$

RESULTS OF SIMULATION

Comparisons of the methods mentioned above were done for two different sample sizes and two different values of ρ . The sample sizes were 50 or 100 subjects. For each sample size and each value of the correlation coefficient, 100 random samples were simulated, and the missingness mechanism (2) was applied to each sample. The estimators of Mean and SD for the incomplete sample were compared with their values for the corresponding complete sample using each of the abovementioned methods. To make the comparisons, we considered the average bias and mean square error (MSE) of the Mean and SD across all the samples for each estimation method involved in comparison.

The results of calculations are shown in tables 1 to 4.

Table 1

N=50 Missing: 50% $\rho = 0.5$	Mean		SD	
	bias	MSE	bias	MSE
Observed	-0.61	0.45	-0.09	0.27
LOCF	0.42	0.20	0.02	0.02
Proc MIXED: Type=un	0.17	0.08	0.31	0.11
Multiple imputation (Proc MI)	-0.28	0.16	0.24	0.11
Incremental mean method	-0.13	0.06	-0.003	0.01

Table 2

N=50 Missing: 49% $\rho = 0.8$	Mean		SD	
	bias	MSE	bias	MSE
Observed	-0.97	1.03	-0.31	0.138
LOCF	0.40	0.18	0.09	0.022
Proc MIXED: Type=un	-0.30	0.17	0.20	0.094
Multiple imputation (Proc MI)	-0.41	0.28	0.10	0.052
Incremental mean method	-0.18	0.07	-0.05	0.020

Table 3

N=100 Missing: 51% $\rho = 0.5$	Mean		SD	
	bias	MSE	bias	MSE
Observed	-0.52	0.31	-0.05	0.021
LOCF	0.41	0.18	0.02	0.010
Proc MIXED: Type=un	-0.12	0.03	0.34	0.144
Multiple imputation (Proc MI)	-0.22	0.09	0.05	0.014
Incremental mean method	-0.10	0.03	0.02	0.010

Table 4

N=100 Missing: 50% $\rho = 0.8$	Mean		SD	
	bias	MSE	bias	MSE
Observed	-0.91	0.87	-0.25	0.085
LOCF	0.40	0.17	0.09	0.017
Proc MIXED: Type=un	-0.24	0.08	0.23	0.084
Multiple imputation (Proc MI)	-0.35	0.16	-0.04	0.028
Incremental mean method	-0.15	0.04	0.03	0.012

The comparisons based on these results show the following:

1. The incremental mean method rendered the most precise results (in average) for both Mean and SD in comparison with the other methods tested.
2. The observed data approach was much less precise than all other methods tested.
3. The LOCF method was the second (after the incremental method) most precise method for estimation of SD. However, for estimation of Mean, it was less precise than the other methods except for the observed data approach.
4. The multiple imputation method in comparison with the linear mixed models method was less precise for estimation of Mean but more precise for estimation of SD

CONCLUSION

The paper is concerned with comparison of several imputation and estimation techniques applied to incomplete longitudinal data sets with two correlated variables. The data sets are simulated to resemble time behavior of HbA1c and FPG in diabetes clinical trials. The missingness mechanisms employed resemble the process of withdrawal from trials based on a threshold for only one variable.

The imputation techniques being compared include the mixed effects model repeated measures method, the multiple imputation method and the incremental method. The results of simulation presented in the paper show that all the methods for different numbers of patients (50 or 100) and a relatively large percentage of missing values the incremental mean method give, in average, more precise estimations of the means and standard deviations (as measured by MSE) than the other estimation methods under comparison.

REFERENCES

1. Little, R.J.A., and Rubin, D.B. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.
2. Verbeke, G., and Molenberghs, G. *Linear Mixed Models in Practice*. Springer, 1997
3. Khutoryansky, N.M., and Huang, W.C. *Imputation Techniques Using SAS Software for Incomplete Data in Diabetes Clinical Trials*, Pharmaceutical Industry SAS Users Group, 2001; 335-339.
4. Khutoryansky, N. M. *Parameter estimation for incomplete bivariate longitudinal data in clinical trials*, Pharmaceutical Industry SAS Users Group, 2003; 549-553.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Naum Khutoryansky
 Novo Nordisk Pharmaceuticals, Inc.
 100 College Road West
 Princeton NJ 08540
 Work Phone: 609-987-5812
 Email: nakh@novonordisk.com
 Web: www.novonordisk-us.com