Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Jolani and van Buuren, J Biomet Biostat 2014, 5:3
DOI: 10.472/2155-6180.1000194

Research Article Open Access

Doubly Robust Imputation of Incomplete Binary Longitudinal Data

Shahab Jolani[1] and Stef van Buuren[1,2]*

[1]*Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands*
[2]*Department of Statistics, TNO, Leiden, The Netherlands*

Abstract

Estimation in binary longitudinal data by using generalized estimating equation (GEE) becomes complicated in the presence of missing data because standard GEEs are only valid under the restrictive missing completely at random assumption. Weighted GEE has therefore been proposed to allow the validity of GEE's under the weaker missing at random assumption. Multiple imputation offers an attractive alternative, by which the incomplete data are pre-processed, and afterwards the standard GEE can be applied to the imputed data. Nevertheless, the imputation methodology requires correct specification of the imputation model. Dual imputation method provides a new way to increase the robustness of imputations with respect to model misspecification. The method involves integrating the so-called doubly robust ideas into the imputation model. Focusing on incomplete binary longitudinal data, we combine DIM and GEE (DIM-GEE) and study the relative performance of the new method in a case study of obesity among children, as well as a simulation study.

Keywords: Double protection; Incomplete data; Ignorable missingness; Multiple imputation

Introduction

In a longitudinal study, each subject or unit is measured at several time points, thereby allowing the direct study of change over time. In many biomedical applications, the longitudinal response is binary, or in general non-Gaussian, for instance, the presence or absence of illness in an intervention study for a period of seven weeks. The generalized linear mixed model is a widely used approach with binary longitudinal responses [1,2]. In many practical settings with moderate to large length, however, these models imply complex and hard to manipulate likelihoods, for example, in the presence of missing data. An alternative modeling approach is generalized estimating equations [3]. This method essentially allows confining attention to the mean structure provided that one is willing to adopt working' assumptions about the association structure.

When data are incomplete, GEE suffers from its frequentist nature and is only valid under the restrictive missing completely at random (MCAR) assumption, where the missingness is independent of both unobserved and observed data [4]. For this reason, Robins et al. [5] have developed a class of GEE methods, the so-called weighted GEE (WGEE), that allows for the weaker missing at random (MAR) assumption, where the missingness is independent of the unobserved data given the observed data [4,6]. WGEE methods use the inverse of the subject's probability of being observed as a weight contributed in the estimating equation to reduce possible bias in the regression parameter estimates.

More recently, WGEE methods have been extended to the so-called doubly robust (DR) estimating equations, where the weighting idea is integrated with the use of a predictive model for the missing data given the observed data. The DR methods provide consistent estimates of the parameters given correct specification of either the weights or the predictive model, but not necessarily both. Excellent reviews can be found in Bang and Robins [7] and Rotnitzky [8].

The idea of doubly protection (or doubly robustness) is advantageous because it provides the analyst two routes to valid inferences, rather than just one. Nevertheless, the DR methods can be unstable in practice when both models are misspecified [9], or they can be disastrous when the propensity scores (i.e., the probabilities of being observed) are close to zero [9,10]. Moreover, these methods lack generalization to intermittent missing data, where the subjects return to the study after skipping one or more visits.

A viable alternative approach is multiple imputation [11,12]. Standard MI requires MAR to hold, even though extensions exist. Missing values are imputed several times, and then the resulting completed data sets are analyzed using a standard method like GEE. Beunckens et al. [13], among others, combined MI and GEE such that the missing data are multiply imputed, and then inferences are obtained by GEE, and combined into a single summary using Rubin's pooling rules (MI-GEE). However, this method, like the other imputation approaches, needs correct specification of the imputation model.

Jolani et al. [14] combined DR ideas with MI and constructed an imputation model with a doubly protected property, the so-called dual imputation method (DIM). This method makes use of the weighting idea within the imputation model. More specifically, a function of the propensity scores (e.g., the inverse of the propensity score) is included into the imputation model with the aim of increasing robustness of imputations against misspecification of the imputation model. Also, DIM can handle the problem of intermittent as well as monotone (or dropout) incomplete longitudinal data.

Until now, DIM has only been tried for continuous data. In this paper, we extend the methodology to binary data. Our focus is thus on the combination of DIM and GEE (DIM-GEE) for incomplete longitudinal binary data when the pattern of missing data is general. This involves multiply imputing binary responses by means of DIM and then applying GEE to the completed data sets. DIM-GEE is a new imputation method that makes it possible to model incomplete longitudinal binary data under the MAR assumption.

***Corresponding author:** Stef van Buuren, Department of Methodology and Statistics, Utrecht University, Utrecht, The Netherlands, Tel: 31 30 253 5194; E-mail: s.vanbuuren@uu.nl

Received March 17 2014; **Accepted** May 06, 2014; **Published** May 12, 2014

Citation: Jolani S, van Buuren S (2014) Doubly Robust Imputation of Incomplete Binary Longitudinal Data. J Biomet Biostat 5: 194. doi:10.4172/2155-6180.1000194

This paper is organized as follows. In GEE for binary longitudinal data, an overview of GEE for analyzing longitudinal binary data is given. MI is briefly outlined in multiple imputation. The new imputation method is presented in dual imputation based GEE. In case study, DIM-GEE is used to analyze a case study of obesity among children and to compare with MI-GEE. A simulation study comparing DIM-GEE with MI-GEE was conducted and results are presented in simulation study.

GEE for Binary Longitudinal Data

Suppose the random variable Y_{ij} denotes a sequence of binary measurements at time j, $j = 1,\ldots,N$ for subject i, $i = 1,\ldots,n$. The observed value y_{ij} is a realization of the binary response variable Y_{ij}, and we assume independence across subjects. The focus of this study is on the marginal models that describe the binary outcome vector, given a set of predictor variables. The association structure (correlation among the components) is captured by an assumed model. Let π_{ij} denote the marginal probability of observing a 'success' for subject i at time j, i.e., $\pi_{ij} = \mathrm{E}(Y_{ij}) = \mathrm{P}(Y_{ij} = 1)$,

$$\varepsilon_{ij} = \frac{Y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}$$

be standardized deviation between the data and the model predictor for subject i at time j, and $\rho_{ij_1j_2} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}), \rho_{ij_1j_2j_3} = E(\varepsilon_{ij_1}\varepsilon_{ij_2}\varepsilon_{ij_3}),\ldots$ be associations among responses. Following Bahadur [15], the model can be represented by

$$f(\mathbf{y}_i) = c(\mathbf{y}_i)\prod_{j=1}^{N}\pi_{ij}^{y_{ij}}(1-\pi_{ij})^{1-y_{ij}}, \quad \textbf{(1)}$$

where $\mathbf{y}_i = (y_{i1},\ldots,y_{iN})$ is a vector of measurements for subject i, and

$$c(\mathbf{y}_i) = 1 + \sum_{j_1<j_2}\rho_{ij_1j_2}e_{ij_1}e_{ij_2} + \sum_{j_1<j_2<j_3}\rho_{ij_1j_2j_3}e_{ij_1}e_{ij_2}e_{ij_3} + \ldots + \rho_{i1\ldots N}e_{i1}\ldots e_{iN},$$

with

$$e_{ij} = \frac{y_{ij} - \pi_{ij}}{\sqrt{\pi_{ij}(1-\pi_{ij})}}.$$

The joint probability mass function is thus the product of individual mass functions and the correlation factor $c(\mathbf{y}_i)$. The later can be viewed as a model for overdispersion.

The use of full likelihood-based methods for the above marginal model can be unattractive due to prohibitive computational requirements. Therefore, alternative methods such as GEE have been proposed. GEE is very useful in marginal models because by adopting working assumptions about the association structure, one only needs correctly specifying the univariate marginal distributions.

For a binary response Y_{ij}, suppose $\mathbf{x}_{ij}$ is a p-dimensional vector of complete covariates. Assuming the logit link function, the mean structure of the binary model can be expressed as

$$\mathrm{logit}\left\{P(Y_{ij} = 1 \mid \mathrm{x}_{ij}, \beta)\right\} = \mathrm{x}_{ij}^{'}\beta,$$

where β is the vector of model parameters. The classical GEE can thus take the form

$$U(\beta) = \sum_{i=1}^{N}\frac{\partial\pi_i}{\partial\beta^{'}}V_i^{-1}(\mathbf{y}_i - \boldsymbol{\pi}_i) = 0,$$

where $\boldsymbol{\pi}_i = E(\mathbf{Y}_i), V_i = A_i^{1/2}C_iA_i^{1/2}$ is the covariance matrix of $\mathbf{Y}_i$, A_i is a diagonal matrix with the marginal variances, and C_i is the marginal correlation matrix for the repeated measures. The correlation matrix C_i is typically expressed in terms of a vector of nuisance parameters that needs to be replaced by a consistent estimate, e.g., a moment-based estimator [3]. Given a correct specified marginal mean π_i, it can be shown, under mild regularity conditions, the estimate of $\hat{\beta}$ is asymptotically normal with mean vector β and covariance matrix $\mathrm{Var}(\hat{\beta}) = I_0^{-1}I_1I_0^{-1}$ where

$$I_0 = \sum_{i=1}^{N}\frac{\partial\pi_i^{'}}{\partial\beta}V_i^{-1}\frac{\partial\pi_i}{\partial\beta^{'}} \quad \text{and} \quad I_1 = \sum_{i=1}^{N}\frac{\partial\pi_i^{'}}{\partial\beta}V_i^{-1}\mathrm{Var}(\mathrm{y}_i)V_i^{-1}\frac{\partial\pi_i}{\partial\beta^{'}}.$$

When the working correlation structure is misspecified there is no price to pay in terms of consistency of the asymptotic normality of $\hat{\beta}$. However, this misspecification may result in loss of efficiency. Because GEE is not a likelihood based approach, it suffers from its frequentist nature in the presence of missing data. Therefore, GEE is only valid under MCAR.

Multiple Imputation

The idea of multiple imputation is to replace each missing value with a set of M plausible values drawn from the conditional distribution of the missing values given the observed data. M imputed data sets are then analyzed using standard methods. The final step is to combine the results into a single summary using Rubin's rule [11].

A popular approach to create imputed datasets is multiple imputation by chained equations [16-18]. The basic idea is to specify a set of imputation models, one model for each variable with missing values, and then impute data on a variable-by-variable basis. We briefly outline the MICE algorithm for the case of binary longitudinal responses. Suppose, for each subject i, the vector of measurements $\mathbf{Y}_i=(Y_{i1},\ldots,Y_{iN})$ has missing values in an arbitrary pattern. We drop i for notation convenience. All missing values are initially filled in at random. The first incomplete measurement, say Y_1, is regressed (here a logistic regression) on the other measurements $Y_2,\ldots,Y_N$ and possibly covariates x restricted to subjects with observed Y_1. Missing values in Y_1 are then imputed using the posterior predictive distribution of Y_1 given $Y_2,\ldots,Y_N$ and $\mathbf{x}$. The missing values in the second incomplete measurement, say Y_2, are then imputed by measurements Y_1, $Y_3,\ldots,Y_N$ and $\mathbf{x}$. The process is repeated for all other measurements with missing values in turn. The cycle is repeated for several times (say 10 or 20) to produce a single imputed dataset. The whole procedure is repeated M times from different seeds, thus producing M completed data sets. The resulting completed data sets are finally used to estimate β using standard methods.

Dual Imputation Based GEE

In this section we show how to impute the missing measurements in binary longitudinal data using DIM methodology, when the missingness mechanism is MAR. The key idea is to incorporate a function of the propensity scores into the imputation model [14]. The aim of including this function (the inverse of the propensity scores) into the model is to reduce the effect of a possible misspecified imputation model.

Suppose Y_{ij} can be observed or missing. Let R_{ij} denote a binary response indictor for subject i at time j; that is, R_{ij}=1 if Y_{ij} is observed and R_{ij}=0 otherwise. For each subject i, r_{ij} is a realisation of R_{ij}. Suppose the probability of being observed (i.e., the propensity score) for subject i at time j follows the logistic model

$$\text{logit}\left\{P(R_{ij}=1\,|\,\mathbf{w}_{ij},\boldsymbol{\alpha}_j)\right\}=\mathbf{w}_{ij}'\boldsymbol{\alpha}_j,\ j=1,\ldots,\mathrm{N}, \quad (2)$$

where $\mathbf{w}_{ij}$ is a q-dimensional vector of covariates associated with the unknown parameters $\boldsymbol{\alpha}_j$, typically including the other outcomes y_{is} ($s \neq j$) and covariates $\mathbf{x}_{ij}$. We further allow that the missingness can happen in an arbitrary pattern (e.g., an intermittent pattern).

We first consider an unrealistic but pedagogical case where all propensity scores $\tau_{ij} = \mathrm{P}(R_{ij} = 1|\mathbf{w}_{ij},\boldsymbol{\alpha}_j)$ are assumed to be known. Then, inclusion of τ_{ij}^{-1} in the imputation model is a sufficient condition to obtain a DR estimator of β [19]. In what follows it is convenient to define $\mathbf{v}_{ij} = (\mathbf{x}_{ij},\mathbf{y}_{i(-j)})'$, where $\mathbf{y}_{i(-j)}$ includes all outcome variables excluding y_{ij}.

Therefore, for each incomplete variable $\mathbf{Y}_{ij}$, the dual imputation model fits the following model restricted to its observed part

$$\text{logit}\left\{P(Y_{ij}=1\,|\,\mathbf{v}_{ij},\tau_{ij}^{-1},\gamma_j,\delta_j)\right\}=\mathbf{v}_{ij}'\gamma_j+\tau_{ij}^{-1}\delta_j \quad (3)$$

where γ_j is a vector of parameters in the imputation model corresponding to $\mathbf{v}_{ij}$, and δ_j is a regression coefficient for the new predictor τ_{ij}^{-1}. A random draw $\left(\gamma_j^*,\delta_j^*\right)$ is generated from its posterior distribution, and then the missing values of the j^{th} incomplete variable are imputed using the drawn values of the parameters. After all incomplete variables are imputed in turn, and the cycle is repeated for an adequate number, a completed data set will be produced. Each completed data set then is analyzed using the conventional GEE, and the results are pooled by Rubin's rule into a single inference.

The propensity scores often are unknown so need to be estimated. Estimation, however, is not straightforward when the pattern of missing data is intermittent. Because estimation of the propensity scores in a particular time depends on the other time points that might be incomplete. For the continuous case, Jolani et al. [14] have developed an extension of MICE algorithm that successively estimates the propensity scores and imputes the missing values for each incomplete variable.

Here we outline the algorithm in detail. Initially, all missing values are filled in at random. Suppressing i form the notation, for each incomplete variable $\mathbf{Y}_j$, j=1,…,N, the propensity score model 2 is used to draw a random value of $\boldsymbol{\alpha}_j$ and to estimate the propensity score $\boldsymbol{\tau}_j^{-1}$ based on the drawn value. The imputation model 3 then generates imputations for the missing part of $\mathbf{Y}_j$. Cycling through all the models, posterior draws of the parameters are made given current values of the other variables. More specifically, steps of the DIM are:

1. Impute initially missing data by taking a random draw from the observed data.

2. Repeatedly, for j = 1,…,N

(a) Estimate $\boldsymbol{\alpha}_j$ in the propensity score model 2, and draw a random value $\dot{\boldsymbol{\alpha}}_j$ from its posterior distribution.

(b) Calculate the propensity score $\hat{\boldsymbol{\tau}}_j$ given the drawn value $\dot{\boldsymbol{\alpha}}_j$.

(c) Add $\hat{\boldsymbol{\tau}}_j^{-1}$ into the imputation model 3 as an additional predictor.

(d) Estimate the parameters $(\boldsymbol{\gamma}_j,\delta_j)$ in the imputation model 3 only from its observed part.

(e) Draw a random value $\left(\dot{\gamma}_j,\dot{\delta}_j\right)$ from their posterior distributions.

(f) Impute impute the missing values in the j^{th} incomplete variable using the drawn values in the previous step

3. Return to step 2 to repeat the algorithm a small number of times, say 10 or 20.

The algorithm is a possibly incompatible Gibbs sampler. Although there is no guarantee for the existence of the joint distribution from which the values are drawn, experience has shown that it often leads to valid statistical inferences in a variety of cases Van Buuren et al. [16]; Gelman and Raghunathan [20]; Van Buuren [21]; Lee and Carlin [22]; White et al. [23].

Case Study

The data used in this paper were obtained from the Muscatine Coronary Risk Factor study [24], a longitudinal survey of school- age children in Muscatine, Iowa. The aim of the study was to examine the development and persistence of risk factors for coronary disease in children. In total, 4856 children (boys and girls) were followed biennially from 1977 to 1981, resulting in 3 measurements per child. The outcome of interest was the status of obesity, coded as 0 (non-obese) or 1 (obese), which was obtained on the basis of a comparison of their weight to age-gender specific norms. One objective was whether the risk of obesity increases with age and whether patterns of change in obesity are the same for boys and girls.

Due to many reasons, the child's obesity status could not be measured on all scheduled time points. Fewer than 40% of the children provided complete data at all three measurements. The patterns of missingness were displayed in Table 1 along with their corresponding frequency and percent of missing data for boys and girls separately. We see that the occurrence of missingness is similar in both groups.

The rate of children classified as obese at each of the three measurement occasions is also depicted in Figure 1. These percentages were calculated based on the complete case analysis at each occasion for both boys and girls. The graph indicates that the rates of obesity were increased for boys over time. For girls, the rates of obesity were increased first, but declined thereafter. The graph also shows that the rates of obesity were higher for girls at all occasions.

The marginal probability of obesity is modeled as a logistic function with time, sex and their interaction as covariates:

$$\text{logit}\left\{P(Y_{ij}=1\right\}=\beta_0+\beta_1\text{time}_{ij}+\beta_2\text{sex}_i+\beta_3\text{time}_{ij}\times\text{sex}_i,$$

where Y_{ij}=1 if the i^{th} child at the j^{th} occasion is classified as obese, and Y_{ij}=0 otherwise; sex_i=1 if the i^{th} child is girl, and sex_i=0 if the i^{th} child is boy; time_{ij}=j, j=1, 2, 3, represents time at each occasion; β=(β_0, β_1, β_2, β_3)' is a vector of parameters 'in which we are interested.

Missing values pose a problem in this study, and, unfortunately, an intermittent pattern of missingness makes the analysis of the data even more complicated. Standard GEE may produce biased results because

Year			Boys		Girls	
1977	1979	1981	Frequency	Percent	Frequency	Percent
O	O	O	897	36%	873	37%
O	O	M	318	13%	313	13%
O	M	O	88	4%	96	4%
O	M	M	389	16%	367	15%
M	O	O	317	12%	328	14%
M	O	M	196	8%	174	7%
M	M	O	281	11%	219	10%

Note: `O' denotes observed measurement, `M' denotes missing measurement.

Table 1: Muscatine Coronary Risk Factor data: Frequency and percent of children per missing data pattern for boys and girls separately.

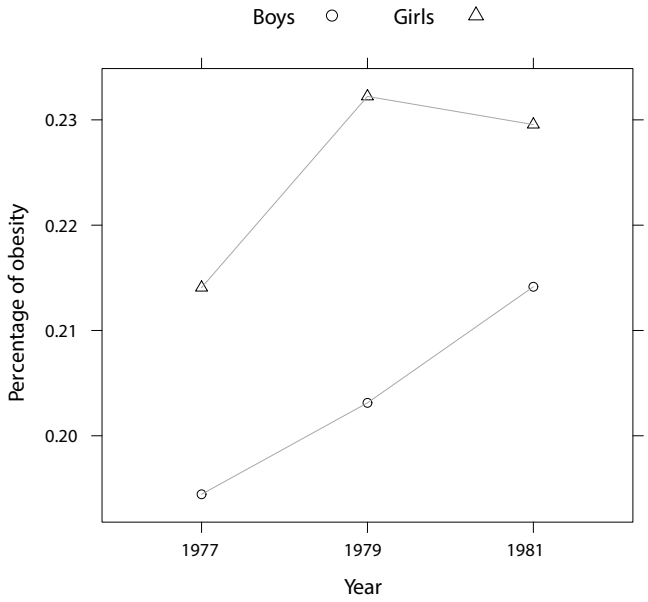

Figure 1: The rates of obesity among children in Muscatine Coronary Risk Factor study during 1977-1981 for boys and girls separately.

Method	Intercept		TIME		SEX		TIME X SEX	
	Est.	Std.err	Est.	Std.err	Est.	Std.err	Est.	Std.err
GEE	-1.551	0.083	0.097	0.034	0.153	0.114	-0.002	0.047
MI-GEE	-1.822	0.15	0.042	0.012	0.164	0.209	-0.008	0.016
DIM-GEE	-1.835	0.138	0.039	0.011	0.127	0.197	-0.003	0.015

Note: GEE is standard generalised estimating equation, MI-GEE is standard multiple imputation based GEE, and DIM-GEE is dual imputation based GEE. Est. is the parameter estimate, and Std.err is the standard error.

Table 2: Parameter estimates and standard errors from the Muscatine Coronary Risk Factor study for GEE, MI-GEE, and DIM-GEE methods.

it is very hard to verify an MCAR assumption. Moreover, complete case analysis is wasteful due to a large fraction of missing data. WGEE cannot also be per- formed because of an intermittent pattern of missingness. Performing an imputation strategy is therefore a reasonable solution to estimate the parameters of interest.

We applied our proposed method (DIM-GEE) as follows. First, a propensity score (i.e., the probability of being observed) was estimated from model 2 for every child at each occasion. Background variables sex and age (mid-point of age group) were considered as covariates in the propensity score model. Second, we included sex, time and their interaction into the imputation model plus the inverse of the propensity scores. The latter variable aims at correcting for possible biases in the imputation model. We created 20 multiply imputed data sets. Each imputed data set was then analyzed using standard GEE. The final results were pooled to obtain a single inference.

Table 2 shows results based on three approaches: Standard GEE, MI- GEE, and DIM-GEE. Standard GEE uses the complete case only. MI-GEE multiply imputes the missing values using the standard MI procedure and then performs GEE for each imputed data set. For this the number of imputations was also 20.

In line with research questions, the effect of time was significant in all methods indicating that the risk of obesity was increased with age. This implies that the older the age of the children was, the higher the rate of obesity was. Nevertheless, the estimated time effect by standard GEE where complete cases were used only was more than twice as large as those of the imputation methods (0.097 versus 0.042 and 0.039) showing a possible overestimation of this effect in standard GEE.

An interaction between time and sex was not significant. This implies that although the rate of obesity was increased with age, this increment did not statistically differ among boys and girls.

Apart from standard GEE, the parameter estimates were very similar based on MI-GEE and DIM-GEE. However, standard errors in DIM-GEE were marginally lower than MI-GEE. Thus, while the substantive conclusion did not differ, imputation using DIM-GEE provided the most efficient estimates. To further investigate the relative merits of both imputation methods, we conducted a limited simulation study aiming at a comparison between DIM-GEE and MI-GEE in the next section.

Simulation Study

This section reports the results of a simulation study comparing DIM-GEE and MI-GEE. We consider a situation where the imputation model for both methods is correctly specified. For DIM-GEE, we also consider a correct propensity score model. Simulation studies comparing misspecified models (in a slightly different setting with continuous outcomes) were reported in Jolani et al. [14]. Since complete case analysis is known to be biased, we concentrated on the comparison between DIM-GEE and MI-GEE.

For the simulation study, we generated data by mimicking the obesity case study. A total of n = 4856 subjects was initially divided into two groups of equal size representing their sex. Then, the binary outcome at three time points was generated based on the Bahadur model formulation 1 with

$$\text{logit}\left\{P(Y_{ij}=1)\right\}=\beta_0+\beta_1\text{time}_{ij}+\beta_2\text{sex}_i+\beta_3\text{time}_{ij}\times\text{sex}_i,$$

where (β_0, β_1, β_2, β_3)=(-1.4, 0.2, -0.2, 0.4), and two-and three-way correlation coefficients equal to $\rho_{ij_1j_2}=0.45$ and $\rho_{ij_1j_2j_3}=0$, respectively. The latter defines an exchangeable correlation structure. The outcome Y_{ij} represents the obesity status (1 coded as obese) for subject i, i=1,…,n, at time j, j = 1,2,3; time$_{ij}$=j represents time at each occasion; and sex$_i$ = 1 if the subject is a girl and zero otherwise.

We assumed the missing data process is MAR and adopted the general methodology proposed by Van Buuren et al. [18] for creating intermittent missing data under MAR. Similar to the obesity data, we specified six missing data patterns (Table 1). Then, the missing data were created such that they formed an approximation of the missing data percentages presented in Table 1. We created missing values in each pattern conditional on the observed data. For instance, the missing data in pattern {OOM} were conditioned on Y_1,Y_2. A full description of this procedure can be found in Van Buuren et al. [18].

The incomplete data sets afterwards were multiply imputed and analyzed by DIM-GEE and MI-GEE methods respectively. For MI-GEE, the imputation model included sex and obesity status at other time points as covariates. For DIM-GEE, we first obtained the propensity scores by fitting the logistic model 2, and then included the inverse of the propensities as an additional covariate into the imputation model. The number of imputations was set to 10 with 1,000 Monte Carlo simulations. All calculations were done in R 3.0.2 using MICE [25].

Several measures were computed to investigate the performance of both methods. First, we defined the relative bias (RB)

Parameter	MI-GEE				DIM-GEE			
	RB%	RMSE	CIW	95%COV	RB%	RMSE	CIW	95%COV
β_0	-1.363	0.093	0.288	88	-2.137	0.093	0.282	88
β_1	-1.915	0.036	0.119	88	-3.598	0.034	0.111	92
β_2	3.363	0.115	0.422	87	3.223	0.112	0.411	92
β_3	-0.578	0.041	0.175	95	-0.025	0.043	0.167	96

Note: MI-GEE is standard multiple imputation based GEE, and DIM-GEE is dual imputation based GEE. RB% is the relative bias percent; RMSE is the root mean squared error; COV is the 95% confidence interval coverage; CIW is the 95% confidence interval width.

Table 3: Simulation results for incomplete binary longitudinal data under MAR based on DIM-GEE and MI-GEE methods.

$$RB = \frac{\bar{\hat{\beta}} - \beta}{\beta},$$

where β is the true parameter value and $\bar{\hat{\beta}}$ is its estimate averaged over all simulations. Further, we calculated the root mean squared error (RMSE) of the parameter estimate

$$RMSE = \left[\left(\bar{\hat{\beta}} - \beta\right)^2 + \mathrm{Var}\left(\bar{\hat{\beta}}\right)\right]^{1/2},$$

where $\mathrm{Var}\left(\bar{\hat{\beta}}\right) = \sum_{s=1}^{S} (\hat{\beta}_s - \bar{\hat{\beta}})^2 / (S-1)$, and S is the number of simulations.

Moreover, a 95% confidence interval width (CIW), as well as the coverage of a 95% confidence interval (COV) were computed. The results for the parameter estimates based on these methods are presented in Table 3.

The relative bias was negligible for both methods showing asymptotically unbiased parameter estimates. However, the RMSE based on DIM-GEE was marginally smaller than that of MI-GEE, pointing a greater efficiency of the estimators by the former method. In addition, the confidence interval width was always shorter for DIM-GEE, and the empirical coverage rates were very close to the nominal level. In contrast, the 95% coverage rates of MI-GEE were lower. In sum, although both methods were performed equally well in terms of bias, the newly developed method provided more efficient parameter estimates.

Concluding Remarks

We have presented a version of generalized estimating equations for in-complete binary longitudinal data under MAR. This extension is based on the principals of multiple imputation, inverse probability weighting and its doubly robustness counterpart, and GEE. Our particular attention was on the extension of dual imputation method [14] to incomplete binary measurements. The proposed method facilitates computational intricacy ofWGEEs in complex patterns of missing data, and is easy to implement in existing software.

In view of previous work on the comparison between WGEE and MI, Clayton et al. [26] and Beunckens et al. [13], among others, provided evidence on preference of MI over WGEE in longitudinal binary data. The simulation studies by Beunckens et al. [13] provided insight about the efficiency of MI based GEE, the so-called MI-GEE, over WGEE particularly in small samples. Nevertheless, misspecification of imputation model can- not be disregarded in practice, and biased results can be expected when the imputation model is incorrect [23,27].

For incomplete binary longitudinal data, the new imputation method (DIM-GEE) was particularly designed to increase the robustness of imputations. By adopting the doubly robust property into the imputation model, one might expect improvement under doubly protected imputation methods.

In this paper, we have compared versions of generalized estimating equations in a real life example with missing data, as well as a simulation study. The results revealed that DIM-GEE produced parameter estimates with smaller estimated variances than the other methods.

Acknowledgements

The authors are thankful to the referees for comments and suggestions that enhanced the manuscript.

References

1. Fitzmaurice GM, Laird NM, Ware JH (2011) Applied Longitudinal Analysis. (2ndedn), New York.
2. Molenberghs G, Verbeke G (2005) Models for Discrete Longitudinal Data. Springer, New York.
3. Liang KY, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73: 13-22.
4. Rubin DB (1976) Inference and missing data. Biometrika 63: 581-592.
5. Robins JM, Rotnitzky A, Zhao LP (1995) Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90: 106-129.
6. Bang H, Robins JM (2005) Doubly robust estimation in missing data and causal inference models. Biometrics 61: 962-973.
7. Little RJ, Rubin DB (2002) Statistical analysis with missing data. Wiley, New York.
8. Rotnitzky A (2009) Longitudinal data analysis. chapter Inverse probability weighted methods. Boca Raton, Chapman and Hall, Florida, 453-476.
9. Tsiatis AA, Davidian M (2007) Comment: Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Stat Sci 22: 569-573.
10. Vansteelandt S, Carpenter J, Kenward MG (2010) Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology 6: 37-48.
11. Mehrotra DV, Li X, Liu J, Lu K (2012) Analysis of longitudinal clinical trials with missing data using multiple imputation in conjunction with robust regression. Biometrics 68: 1250-1259.
12. Rubin DB (1996) Multiple imputation after 18+ years. Journal of the American Statistical Association 91: 473-489.
13. Beunckens C, Sotto C, Molenberghs G (2008) A simulation study comparing weighted estimating equations with multiple imputation based estimating equations for longitudinal data. Computational Statistics and Data Analysis 52: 1533-1548.
14. Jolani S, Frank LE, van Buuren S (2014) Dual imputation model for incomplete longitudinal data. Br J Math Stat Psychol 67: 197-212.
15. Bahadur RR (1961) Studies in item analysis and prediction, Stanford mathematical studies in the social sciences vi. chapter A representation of the joint distribution of responses to n dichotomous items. Stanford University Press, Stanford, USA.
16. van Buuren S, Boshuizen HC, Knook DL (1999) Multiple imputation of missing blood pressure covariates in survival analysis. Stat Med 18: 681-694.
17. Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P (2001) A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27: 85-95.
18. Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB (2006) Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76: 1048-1064.
19. Scharfstein DO, Rotnitzky A, Robins JM (1999) Adjusting for nonignorable drop-out using semi-parametric nonresponse models (with comments). Journal of the American Statistical Association 94: 1096-1146.
20. Gelman A, Raghunathan TE (2001) Using conditional distributions for missing

data imputation. Discussion of 'Conditionally specified distributions' by Arnorld et al. Statistical Science 16: 268-269.

21. van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 16: 219-242.

22. Lee KJ, Carlin JB (2010) Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. Am J Epidemiol 171: 624-632.

23. White IR, Royston P, Wood AM (2011) Multiple imputation using chained equations: Issues and guidance for practice. Stat Med 30: 377-399.

24. Woolson R, Clarke W (1984) Analysis of categorical incomplete longitudinal data. J R Statist Soc A 147: 87-99.

25. Van Buuren S, Groothuis-Oudshoorn CGM (2011) MICE: Multivariate Imputation by Chained Equations in R. Journal of Statatistical Software 45.

26. Clayton D, Spiegelhalter D, Dunn G, Pickles A (1998) Analysis of longitudinal binary data from multiphase sampling. Journal of the Royal Statistical Society, Series B 60: 71-87.

27. Jolani S, Van Buuren S, Frank LE (2013) Combining the complete data and nonresponse models for drawing imputations under MAR. Journal of Statistical Computation and Simulation 83: 866-877.

Citation: Jolani S, van Buuren S (2014) Doubly Robust Imputation of Incomplete Binary Longitudinal Data. J Biomet Biostat 5: 194. doi:10.4172/2155-6180.1000194

AppendixA. R-syntax to run the DIM-GEE for binary longitudinal measurements

This Appendix provides a sample code for running the DIM-GEE method. We assume an incomplete binary longitudinal data set with three measurements (y_1, y_2, y_3), as well as one baseline covariate (x) is available.

```
# Functions needed to run DIM-GEE

# Propensity score model
psm <- function(cr,v1,v2,v3){
  1/glm(cr ~ v1 + v2 + v3, family = binomial(link =
   logit))$fitted.values
}

# Multiple imputation by DIM method
# This code is only works for a binary longitudinal data
with 3 measurements and 1 baseline covariate
dualimpute <- function(data, ...){
  require(mice)
  data <- as.data.frame(data)
  colnames(data) <- c("y1", "y2", "y3", "x")
  data[,"y1"] <- as.factor(data[,"y1"])
  data[,"y2"] <- as.factor(data[,"y2"])
  data[,"y3"] <- as.factor(data[,"y3"])
  datadim <- data.frame(data, dr1 = NA, dr2 = NA,
   dr3 = NA, r1 = NA, r2 = NA, r3 = NA)
  datadim[,"r1"] <- !is.na(data[,"y1"])
  datadim[,"r2"] <- !is.na(data[,"y2"])
  datadim[,"r3"] <- !is.na(data[,"y3"])
  inidim <- mice(datadim, max=0, print=FALSE)
  meth <- inidim$meth
  meth["dr1"] <- "~psm(r1, y2, y3, x)"
  meth["dr2"] <- "~psm(r2, y1, y3, x)"
  meth["dr3"] <- "~psm(r3, y1, y2, x)"
  pred <- inidim$pred
  pred["y1", "dr1"] <- 1
  pred["y2", "dr2"] <- 1
  pred["y3", "dr3"] <- 1
```

```
  pred[,c("r1", "r2", "r3")] <- 0
  impdim <- mice(datadim, pred = pred, meth=meth,
   print=FALSE, maxit = 20, ...)
  return(impdim)
}

# pooling multiple imputions
poolres <- function(fit){
  m <- length(fit)
  dim <- length(fit[[1]]$coefficients)
  q <- matrix(NA, m, dim)
  u <- matrix(0,  dim, dim)
  for (i in 1:m){
    q[i,] <- fit[[i]]$coefficients
    u <- u + fit[[i]]$geese$vbeta
  }
  Qbar <- apply(q, 2, mean)
  Ubar <- u/m
  Bvar <- cov(q)
  Tvar <- Ubar + Bvar*(1 + 1/m)
  res <- t(rbind(Qbar, sqrt(diag(Tvar))))
  rownames(res) <- rownames(summary(fit[[1]])$
   coefficients)
  colnames(res) <- c("Estimate", "Std. Error")
  return(res)
}

# data must be in wide format
# m is the number of imputations
# seed is an arbitrary number
impdim <- dualimpute(data, m = 5, seed = 12345)
require(geepack)
fitdim <- list()
for (i in 1:impdim$m){
  temp <- complete(impdim,i)
  temp <- reshape(temp, varying = c("y1", "y2", "y3"),
   direction = "long", v.names = "y")
  temp[,"y"]  <- as.numeric(as.character(temp[,"y"]))
```

```
  temp <- temp[order(temp$id),]
  fitdim[[i]] <- geeglm(y ~ x + time + time*x, id = id,
   family = binomial, data = temp, corstr = "exchangeable")
  }
dimgee <- poolres(fitdim)
dimgee
```