International Journal of Statistics and Applications 2013, 3(4): 106-112
DOI: 10.5923/j.statistics.20130304.03

A New Imputation Strategy for Incomplete Longitudinal Data

Ahmed M. Gad*, Iman Z. I. Abdelraheem

Statistics Department, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt

Abstract Longitudinal studies are very common in public health and medical sciences. Missing values are not uncommon with longitudinal studies. Ignoring the missing values in the analysis of longitudinal data leads to biased estimates. Valid inference about longitudinal data must incorporate the missing data model into the analysis. Several approaches have been proposed to obtain valid inference in the presence of missing values. One of these approaches is the imputation techniques. Imputation techniques range from single imputation (the missing value is imputed by a single observation) to multiple imputation, where the missing value is imputed by a fixed number of observations. In this article we propose a new imputation strategy to handle missing values in longitudinal data. The new strategy depends on imputing the missing values with donors from the observed values. The donors represent quantiles of the observed data. This imputation strategy is applicable if the missing data mechanism is missing not at random. The proposed technique is applied to a real data set of antidepressant clinical trial. Also, a simulation study is conducted to evaluate the proposed strategy.

Keywords Antidepressant, Hot Deck Imputation, Jennrich and Schluchter Algorithm, Longitudinal Data, Missing Data, Multiple Imputations

1. Introduction

Longitudinal studies are often used in public health and medical sciences. Such studies are often designed to investigate changes in a specific variable, which is measured repeatedly over time for every participant. A typical example of longitudinal studies is to measure blood pressure for patients with hypertension several times during a specific period of time. In fact longitudinal studies are powerful design since they are measuring the changes of the variable of interest over time.

Longitudinal studies are rarely complete due to patient attrition, mistimed visits, premature study termination, death, and other factors. Missing data in longitudinal studies can be a difficult problem to overcome. Although investigators may devote substantial efforts to minimize the number of missing values, some amount of missing data is inevitable in the practice. Missing data could be in the response variable, the covariates, or both. Our interest is on the missingness in the response variables.

A vast amount of work has been done in the area of missing values, see for example Fitzmuarice et al.[1] and references therein. This has led, on the one hand, to a rich taxonomy of missing data concepts, issues, and methods, and, on the other hand, to a variety of data-analytic tools. If missing data are treated inadequately, the statistical power of detecting treatment effects may be reduced, the variability might be underestimated and bias may affect the estimation of the treatment effect, the comparability of the treatment groups, and the generalizability of the results. So, it might be necessary to accommodate missing values in the modelling process to avoid either biased parameter estimates or invalid inferences.

The missing data pattern can be categorized into two patterns: *intermittent* missing values (i.e., missing values due to occasionally withdrawal) with observed values afterwards, intermittent missingness is also termed *non-monotone* missing pattern, and *dropout* (i.e., missing values due to earlier withdrawal), with no observed values. Dropout missing pattern is *monotone* pattern ([2];[3]).

Let y_{ij} be the response for the jth measurement from the ith subject, made at time t_{ij} j=1,…,n_i, i =1,…,m. The times t_{ij} are assumes to be common for all subjects. Let $Y_i = (y_{i1}, y_{i2}, \dots, y_{in})$ be the vector contains the response of the ith subject at all times. Assume that the vector of responses can be partitioned to the observed part and the missing part, i.e. $Y_i = (Y_{iobs}, Y_{imis})$. The impact of missing data and the ways to handle incomplete data depend much upon the mechanism of incompleteness. A set of definitions for missing data mechanisms has been proposed by Rubin[4], also in[5]. These are *missing completely at random* **(MCAR)**, *missing at random* **(MAR)**, and *missing not at random* **(MNAR)**. They introduce an indicator variable $R_i =$

* Corresponding author:
dr_ahmedgad@yahoo.co.uk (Ahmed M. Gad)
Published online at http://journal.sapub.org/statistics

$(R_{i1}, \ldots, R_{in_i})$, where

$$R_{ij} = \begin{cases} 1 & \text{if } Y_{ij} \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

The missing data are missing completely at random (MCAR), if the random variable R_i is independent of both Y_{iobs} and Y_{imis}, in notation:

$$P(R_i|Y_{iobs}, Y_{imis}; \psi) = P(R_i|\psi)$$

i.e., If the missingness is independent of both unobserved and observed data. If the missingness in not related to the data being missing, conditional on the observed measurements, this called missing at random (MAR). i.e., if the random variable R_i is independent of Y_{imis}

$$P(R_i|Y_{iobs}, Y_{imis}; \psi) = P(R_i|Y_{iobs}; \psi).$$

If the missingness also depends on the unobserved data Y_{imis}, it is considered to be missing not at random (MNAR) also called (informative). Informative missingness refers to the fact that the probability of missingness depends on the underlying individual characteristics. i.e., the random variable R_i is conditionally dependent on Y_{imis} given Y_{iobs}

$$P(R_i|Y_{iobs}, Y_{imis}; \psi)$$

When data are MNAR valid analysis requires explicit incorporation of the missing data mechanism, which in most situations will be unknown. That is because under informative missingness, the estimation methods ignoring missing data process can lead to severely biased estimators ([6];[7]).

Imputation techniques have become an important and influential approach in the statistical analysis of incomplete data. The imputation techniques can be classified according to the number of imputed values as single imputation (SI) and multiple imputations (MI). In single imputation techniques each missing observation is replaced by a single value. In multiple imputation techniques each missing value is imputed by a group of observations[8,9]. Standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. These analysis results are then combined to produce one overall analysis.

The key advantages of MI are flexibility and simplicity. It applies to a wide range of missing data situation and is simple enough to be used by non-statisticians. Theoretically this approach is superior to other models because it often produces the most robust effects[7]. The second advantage is bias correction. When the missing data mechanism is MAR, as opposed to MCAR, the method corrects for biases in complete case analysis and other ad hoc analyses[6];[10]. Moreover, auxiliary variables that are not part of the analysis procedure can be incorporated into the imputation procedure to increase efficiency and reduce bias[11].

Many variants of imputation techniques have emerged in literatures. In cross-sectional studies, Yulei[7] summarizes some of the key steps involved in a typical MI project for practitioners. Yulei[7] introduced the basic concepts and general methodology of multiple imputationand provided some guidance for application in regression analysis. Spratt et. al.[10] investigated the effect of including auxiliary variables. Donders, et al.[12] compared MCMC and chained equations approaches in the context of estimating coefficients in a linear regression model. Allison[13] examined two approaches of MI for missing data: one that combines a propensity score with the approximate Bayesian bootstrap (ABB), and regression based method (data augmentation MCMC) method. Siddique, et. al.[11] presented a framework for generating multiple imputations for continuous data when the missing data mechanism is unknown. Munoz and Rueda[14] proposed a novel single imputation method based on quantiles obtained from available data. They applied this method on a cross-sectional data. They concluded that this method can provide desirable estimates of parameters. Deng[15] proposed a multiple imputation method based on conditional linear mixed effects model.

Kenward and Carpenter[6] outlined how MI proceeds in practice. They concluded that MI incorporates information from subjects with incomplete sets of observations, and a good advantage of MI is bias correction. When the missing data mechanism is MAR, their method corrects for bias in completers-only analysis and other ad hoc analyses. Yang et. al.[16] introduced alternative imputation-based strategies for implementing longitudinal models with full-likelihood function in dealing with intermittent missing values and dropouts that are potentially non-ignorable. They have demonstrated the application of *multiple partial imputation* (MPI) and *two-stage* MPI. Tang et. al.[17] compared a multiple imputation method that handles all variables at once in a multivariate normal model (MVNMI) with an approach that combines hot-deck (HD) multiple imputations. Olsen et. al.[18] described and implemented both linear mixed effects models and an inclusive MI strategy, which generates multiple imputations data sets under a multivariate normal model via the Bayesian simulation technique (MCMC or data augmentation). In general the performance of imputation techniques depends on the missing data mechanism, the trajectory of repeated measures, and the distribution of the variables[19].

The aim of this paper is to introduce an imputation strategy to handle missing data in longitudinal studies. The Munoz and Rueda[14] method depends on imputing the missing values using quantiles of the observed values. This method is used with cross-sectional studies. The proposed strategy is a development of Munozand Rueda method to the longitudinal data setting. Once, the pseudo complete data have been obtained we suggest using selection model to model the data. The parameter estimates and their standard errors have been obtained.

2. Modelling Longitudinal Data

Let y_{ij} and x_{ij} be respectively the response and p-vector of explanatory variables for the jth measurement from the ith subject, made at time t_{ij} j=1,…,n_i, i =1,…,m. the times t_{ij} are assumed to be common for all subjects. The mean and variance of y_{ij} are represented by $E(y_{ij}) = \mu_{ij}$ and

Var(y_{ij}) = σ_{jj} . All measurements of the subject i are collected into an $1 \times n_i$ vector, $Y_i = (y_{i1}, \dots, y_{in_i})$, with mean $E(Y_i) = \mu_i$ and covariance matrix $\mathrm{Var}(Y_i) = V_i$ of order $n_i \times n_i$, where the jk th element of V_i is the covariance between y_{ij} and y_{ik} , denoted by $\mathrm{Cov}(y_{ij}, y_{ik}) = \sigma_{jk}$. The responses for all subjects are denoted as $Y = (Y_1, \dots, Y_m)'$, which is an N-vector with *N*= $\sum_{i=1}^{m} n_i$. Let n be the maximum value of n_i and $\mathrm{Cov}(Y) = V$. The matrix V_i may be unstructured, i.e. containing $n_i(n_i + 1)/2$ covariance parameters, or it may have a specific structure, i.e. its elements are functions of smaller number of parameters α, in this case it is written as $V_i(\alpha)$.

The response for the subject i is modelled as a linear regression model

$$Y_i = X_i\beta + \epsilon_i,$$

where $\beta = (\beta_1, \dots, \beta_p)$ is a p-vector of unknown regression coefficients, X_i is a known $n_i \times p$ matrix of explanatory variables with x_{ij} in the jth row and $\epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{in_i})$ is a zero-mean random variable representing the deviation of the response from the model prediction, $X_i\beta$. It is assumed that $\epsilon_i \sim N(0, V_i)$.

In the selection model[20], the joint distribution of Y_i and R_i is factored as a product of the marginal distribution of Y_i and the conditional distribution of R_i given Y_i, i.e.

$$f(Y_i, R_i|\theta, \psi) = f(Y_i|\theta)\mathrm{P}(R_i|Y_i, \psi),$$

where $f(Y_i|\theta)$ represents the complete data model for Y_i, $\mathrm{P}(R_i|Y_i, \psi)$ represents a model for missing data mechanism, and (θ, ψ) are unknown parameters. Diggle and Kenward [20] suggested modelling missing data mechanism using logistic model as

$$\mathrm{logit}\{\mathrm{P}(R_{ij} = 0)\} = \psi_0 + \psi_1 y_{ij-1} + \psi_2 y_{ij}.$$

Also, Diggle and Kenward[20] formulate the log-likelihood function. This log-likelihood function can maximised to obtain the parameter estimates using a suitable optimization method.

3. The Proposed Imputation Strategy

3.1. Imputation and Estimation

The proposed strategy is a multiple imputation technique based on quantiles, to handle analysis of longitudinal data with missing values. This approach is able to deal with different assumptions of missingness mechanisms. Generally multiple imputations techniques consist of two distinct steps: the first is the imputation; the second is the analysis of the imputed data sets, where the uncertainty introduced in the imputation part is included in the estimates. Thus, two steps can be handled separately, even using different models. Therefore, the proposed strategy accordingly divided into two parts; the methodology of the proposed multiple imputations method, and the analysis of the longitudinal data based on selection model. The proposed method is a development of the Muniz and Rueds method[14].

The proposed method has very useful features. First, it is robust against model violation.Second, it is more robust in the presence of outliers than methods based on means, since quantiles are known to be less affected than the mean in the presence of outliers.Finally, the proposed method can be applicable in the case of discrete data.

Let y_{ij} be the *j*th measurement on the *i*th respondent (patient) and let n_{iobs} and n_{imis} be the number of observed and missing measurements, respectively. The proposed method consists of using *s* donors (quantiles) obtained from the sample of the *i*th subject. Following Munoz and Rueda [14], the proposed imputed values are given by

$$y_k{}^* = \hat{Q}_{iobs}(\alpha_k) = \inf\{t: \hat{F}_{iobs} \geq \alpha_k\}\ k = 1, \dots, s,$$

$\alpha_k \in [0,1]$ and $\hat{F}_{iobs}(t) = r^{-1}\sum_{k \in iobs} \Delta(t - y_k)$ is the customary distribution function estimator based on the sample n_{iobs}.

The value α_k is chosen such that the data set after imputation leads to a better overall estimate for the parameters of interest. It seems reasonable to assume that the value α_k could be taken at regular intervals from the cumulative distribution function of variable Y_{iobs} observed on the sample n_{iobs}. Taking this consideration into account, we proposed two choices: the first is

$$\alpha_k = \frac{k}{s+1}\ \ k = 1,2, \dots, s. \qquad (1)$$

As an example, the proposed imputation method when $s = 3$, with values α_k given by Eq. (1), use the quartilesfor imputation.

The second choice, according to Munonz and Rueda[14], is

$$\alpha_k = \frac{k-1}{s-1}\ \ k = 1,2, \dots, s. \qquad (2)$$

In the case of *s*=3; Eq. (2) use the minimum, the median and the maximum values y_i, with $i \in n_{iobs}$.

Note that the value of s can not exceed the number of observed values for each subject, since this method can be categorized as a hot deck method and, the donors assigned for missing values are taken from respondents in the current sample. Also, only the first missing observation in each subject is imputed. Once, we imputed this value the remaining missing values can be considered as missing completely at random.

After applying the proposed imputation method, we have s datasets. Apply the chosen statistical analysis on each of these imputed datasets, for ignorable mechanism. For each pseudo complete data we fit the selection model and obtain the parameter estimates. Fitting selection model mean estimating the parameters in two steps. The first step, the maximum-likelihood estimates of the model parameters are obtained using an appropriate optimization approach. The EM scoring algorithm[21] is used in this paper. In the second step, the maximum-likelihood estimates of the missing data mechanism parameters (the logistic model) are obtained. An iterative maximum-likelihood estimation approach of binary data models, see for example[22], can be used.

Once the analyses have been completed for each imputed data set, we combine these analyses to produce one overall set of estimates. Combining the estimates from the imputed data sets is done using rules established by Rubin[8]. These

rules allow the analyst to produce one overall set of estimates like the produced from a non-imputation analysis.

Suppose $\hat{\theta}$ is the point estimate for a scalar parameter, θ. Rubin's rules specify that combining the estimates of the parameter of interest is accomplished simply by averaging the individual estimates produced by the analysis of each imputed data set. In mathematical terms, this is written generally as:

$$\bar{\theta} = \frac{1}{s}\sum_{l=1}^{s}\hat{\theta}_l, \quad (3)$$

for s imputed data sets.

In the case of ignorable missing data the Jennrich and Schucher method[21] can be used to obtain maximum likelihood estimates. This is an iterative technique and adopted in this paper.

3.2. Standard Errors of Estimates

We suggest a bootstrap method to obtain the standard errors of the estimates. The bootstrap method was introduced by Efron[23], also in[24], as a computer based method to estimate the standard deviation of the parameters estimates $\hat{\theta}$. This approach utilizes resampling from the original data. When the sample size is large, the bootstrapping estimates will converge to the true parameters as the number of repetitions increases.

The basic idea of the bootstrap involves repeated random samples with replacements from the original data. This produces random samples of the same size of the original sample, each of which is known as a bootstrap sample and each provides an estimate of the parameter of interest.A great advantage of bootstrap is simplicity. It is a straightforward way to derive estimates of standard errors and confident-intervals for complex estimators of complex parameters of the distributions, such as percentiles, proportions, and odds ratio, since it is completely automatic, and requires no theoretical calculations. Moreover, it is an appropriate way to control and check the stability of the results.

Generally, bootstrapping follows the same basic steps:

1. Select B independent bootstrap samples $X_1^*,\dots,X_B^*$, each consists of n data values drawing with replacement from X. This create bootstrap data sets of the same size as the original.

2. Apply the estimate procedure to the bootstrap samples by estimating the desired statistic, which forms the sampling distribution of $\hat{\theta}$. Such that:

$$\hat{\theta}^*(b) = S(X_b^*) \quad b = 1,\dots,B$$

3. Estimate the standard error $SE(\hat{\theta})$ by the sample standard error of the B replicates

$$\widehat{SE}(\theta) = \left[\frac{1}{B-1}\sum_{b=1}^{B}\left\{\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)\right\}^2\right],$$

with

$$\hat{\theta}^*(\cdot) = B^{-1}\sum_{b=1}^{B}\hat{\theta}^*(b)$$

Let θ_i be the estimated parameter in sample i of the B samples, and let S_i be its estimated standard error. The mean of θ_i is $\bar{\theta}$ and it's estimated standard error is given by Rubin[8]:

$$\sqrt{\frac{1}{B}\sum_{i=1}^{B}S_i^2 + \left(1+\frac{1}{B}\right)\left(\frac{1}{B-1}\right)\sum_{i=1}^{B}(\theta_i - \bar{\theta})^2}$$

In other words, this is the square root of the average of the sampling variances plus the variance of estimates multiplied by a correction factor $\left(1+\frac{1}{B}\right)$ (Allison[13]).

4. Application (Antidepressant Data)

The data represent a multi-centre study of depression. The number of individuals enrolled in the trial is 367 depressed patients from six centres. In each centre, participants were randomly assigned to 3 different treatments. About 20 participants assigned to each treatment in each centre. Each participant was rated on the Hamilton depression score (HAMD). A response scale from 0 to 50, was produced from 16 test items. The HAMD score were expected to take on each participant over 5-weekly visits. The first visit made before treatment, the remaining four during treatment.

Table 1 shows the distribution of unit response pattern over 5 weeks. There were 243 (66.2%) of the participants completed all follow-up assessments. At the termination of the study, up to 124 (33.8%) of the participants had withdrawn from the study.

Table 1. Unit response patterns*

Weeks 1	2	3	4	5	Overall	%
x	*x*	*X*	*x*	*x*	243	66.2
x	*x*	*X*	*x*	*m*	44	12.0
x	*x*	*X*	*m*	*m*	27	7.4
x	*x*	*M*	*m*	*m*	53	14.4

* *x*: respondent; *m*: missing

Table 2. Mean Profile of the treatments

Treat	Value	Statistic		SE
1	Mean		17.44	0.430
	95% CI for Mean	L. Bound	16.59	
		U. Bound	18.30	
	5% Trimmed Mean		17.42	
2	Mean		15.72	0.480
	95% CI for Mean	L. Bound	14.77	
		U. Bound	16.68	
	5% Trimmed Mean		15.58	
3	Mean		17.51	0.479
	95% CI for Mean	L. Bound	16.56	
		U. Bound	18.46	
	5% Trimmed Mean		17.47	

Table 2 shows the profile means of each treatment which based on the observed data. The second treatment has the lowest mean profile of a value 15.72. Nevertheless; these means are considered to be biased.

Figure 1 shows the set of measurements for completers at each centre. It can be noticed from the figure the following: first, there is a typical decrease over time of HAMD scores in all centres. Second, the highest rate of variation of the measurements between participants is at week 5. Third, there is a participant in centre two taking treatment 3 who starts and remains at high value.

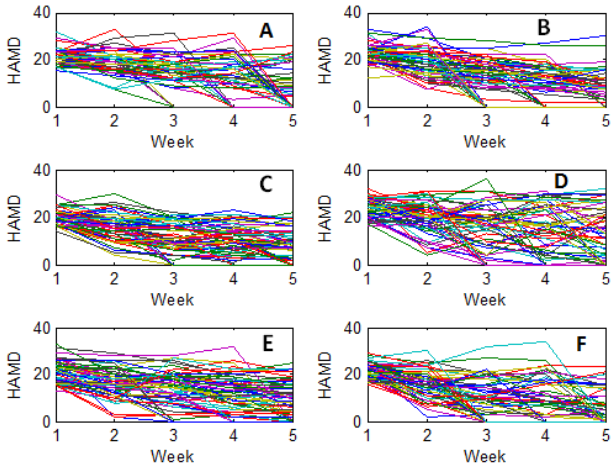

Figure 1. Observed measurements of antidepressant data and centers: : (A) center 1; (B) center 2; (c) center 3; (D) center4; (E) center 5; (F) center 6

Table 3 shows number of participants who dropouts at each treatment in each centre. Centre 6 has the highest rate of dropout, regarding the treatment. The highest rate of dropouts occurs at treatment 1. Dropouts occur consistently from week 3 to week five, except in the second treatment in centre 6, all dropouts occur in week 3. Generally, week 3 has the highest number of dropouts.

Table 3. Number of dropouts at centre*treatment combination

Centre	Treat 1	Treat 2	Treat 3	Total
1	8	8	8	24
2	10	5	5	20
3	5	8	2	15
4	5	7	8	20
5	5	6	5	16
6	12	8	8	28
Total	45	42	36	123

Non-random missing data is of particular concern in depression trials because it is very likely that the reason for a participant missing an assessment or dropping out of the study is related to their underlying depression status[11]. For example, a depressed participant may feel that the treatment is not of benefit for him/her and may be unwilling to sit through an interview or any kind of assessment. Conversely, a high-functioning, non-depressed participant may feel that he/she no longer needs to remain in the trial or may not be available for an assessment because he/she is busy with life. Failure to take into account the missing data mechanism may result in inferences that make a treatment appears more or less effective.

We considered two choices of multiple imputation based on quantiles to handle the missing data. Each is based on assuming that the dropout rate is non-ignorable. As mentioned earlier that the dropout occur from week three, and we have five measurements to each participant, i.e. the smallest observed measurements for each subject is two observations. The proposed MI method is considered to be a hot deck procedure; hence we set the number of multiple imputations to be three. We impute the first missing value as we argue that the remaining missing values are missing completely at random. We apply Eq. (1) and Eq. (2) to impute the missing value. In the special cases were the number of observed measurements is two (found in 53 cases (14.4%)), we imputed the missing value with the average of the two observed measurements.

Table 4. The estimates and their standard errors for antidepressant data

Par	Est.	SE	Pa.	Est.	SE
α	20.51	0.0500	σ_{13}	4.42	1.3900
β_1	-0.04	0.0800	σ_{14}	4.16	1.2700
β_2	1.40	0.0900	σ_{15}	3.33	1.1900
β_3	-1.89	0.0800	σ_{22}	44.82	1.1400
β_4	2.63	0.1100	σ_{23}	41.29	1.1200
β_5	-0.49	0.0900	σ_{24}	42.72	1.1700
γ_1	0.48	0.0400	σ_{25}	44.62	1.3500
γ_2	-0.45	0.0500	σ_{33}	64.67	1.8400
ψ_0	3.89	0.0100	σ_{34}	60.43	1.8100
ψ_1	0.10	0.0001	σ_{35}	63.31	1.1300
ψ_2	-0.19	0.0001	σ_{44}	81.96	1.0700
σ_{11}	13.19	2.0300	σ_{45}	79.21	1.3500
σ_{12}	7.32	1.6000	σ_{55}	102.03	2.5300

The mean profiles are modelled as

$$E(Y_{ct}) = \alpha + \beta_c + \gamma_t, \quad (4)$$

for centre c, treatment t, and where $c = 1, \ldots, 5, t = 1,2$, giving 8 parameters for the model. The sixth centre and the third treatment are the base categories. The covariance matrix is left unstructured, so there are 15 covariance parameters for the five time points. Other covariance structures can be used. The dropout process is modelled using the logistic model:

$$\text{logit}\{P(y_k, y_{k-1}, \psi)\} = \psi_0 + \psi_1 y_{k-1} + \psi_2 y_k,$$

where $k = 3, \ldots 5$. In this case the underlying dropout rate is set to 0 for the first and second weeks, when there are no dropouts, and constant for weeks 3 - 5.

The proposed strategy is used to obtain the parameter estimates. The results are presented in Table 4. Also, the bootstrap standard errors are presented.

From the results we can see that all the β's parameters are significant except β_1. This means that the first centre effect

is not significant and the remaining centres effect is significant. Also, all treatments have significant effects because γ's estimates are significant. The parameters ψ's are significant. The parameter ψ_2 is of main interest. This parameter represents the missing not at random process. The significance of this parameter means that the missing data process is missing not at random.

5. Simulation Study

The purpose of this simulation study is to evaluate the performance of the estimates of proposed multiple imputation method. The simulation setup is as follows. The sample size, m, is chosen to range from small to large. We consider the sample sizes m=20, m= 50 and m=100 to represent small, moderate and, large sample sizes respectively. The number of time points (repeated measures) is fixed at 5. Each subject is allocated randomly to one of two groups; treatment or control group with 50% probability. We use the linear regression model $E(Y_i) = \beta_1 + \beta_2 x_i$, where $x_i = (1, Trt_i), Trt_i = 1$ or 0 indicating treatment or control group. The covariance matrix is left unstructured.

For the missingness model we used the model; $\text{logit}\{P(y_k, y_{k-1}, \psi)\} = \psi_0 + \psi_1 y_{k-1} + \psi_2 y_k$, where k=3,4,5.

Data are simulated from the multivariate normal model, the unstructured covariance matrix and the missingness model. The parameters values are $\beta_1 = 10.7, \beta_2 = 0.4$, $\sigma_{11} = 13.2, \sigma_{12} = 7.3, \sigma_{13} = 4.5, \sigma_{14} = 4.2, \sigma_{15} = 3.5, \sigma_{22} = 44.8, \sigma_{23} = 41.4, \sigma_{24} = 42.8, \sigma_{25} = 44.8, \sigma_{33} = 63.7, \sigma_{34} = 59.7, \sigma_{35} = 62.6, \sigma_{44} = 81.3, \sigma_{45} = 78.6, \sigma_{55} = 100.8,$ $\psi_0 = 7.5, \psi_1 = 0.3, \psi_2 = 11.7.$

The proposed approach has been applied and then the parameter estimates are obtained. For each simulation we used 10000 replications. The relative bias (RB%) is obtained for each parameter. The simulation results are summarized in Table 5.

From the simulation results we notice that the relative bias of all parameters is below 20% which is reasonable value. This indicates that the proposed technique gives reasonable results for different sample sizes. The relative bias of β_2 is around 10% for all sample sizes. This parameter is of main interest because it represents the difference in slope between the two groups.

Other models for the covariance structure have been tried. Also different values formissingness model parameters have been used. The qualitative results are the same as theabove results, so they are not reported.

6. Conclusions

Modelling and analysing the missing values in the longitudinal data context have gained popularity in recent years. Imputation methods are emerged as an aid to such analysis. Multiple imputations is a branch of imputation techniques in which the missing value is replaced by a group of values, resulting in a group of pseudo complete data sets. These data sets are analysed separately and then the overall analysis is obtained. In this paper we present a new multiple imputations strategy based on quantiles. The proposed method is a development of Munoz and Ruedo method to the longitudinal data setting. This method has many advantages; at least it is less biased comparable to other methods. The proposed method is applied in the case of dropout pattern. The method can be easily extended to the intermittent pattern. The selection model[6] is used to fit the data after imputation step. The proposed technique is applied to a real data set.

Table 5. The relative bias of parameter estimates

Para.	n =20		n =50		n=100	
	Est.	RB%	Est.	RB%	Est.	RB%
β_1	11.01	2.9	11.15	4.2	11.15	4.2
β_2	0.35	12.5	0.36	10.0	0.36	10.0
σ_{11}	12.49	5.4	12.73	3.6	12.98	1.7
σ_{12}	6.66	8.8	6.83	6.4	7.09	2.9
σ_{13}	4.34	3.5	4.94	9.7	4.54	0.8
σ_{14}	4.15	1.2	4.70	11.9	14.31	2.6
σ_{15}	3.61	3.1	4.13	18.0	3.57	2.0
σ_{22}	44.40	0.8	44.29	1.1	44.51	0.6
σ_{23}	40.04	3.3	38.39	7.2	38.11	7.8
σ_{24}	41.40	3.3	39.83	7.9	39.65	7.3
σ_{25}	43.13	3.7	41.49	7.4	41.12	8.2
σ_{33}	59.76	6.2	56.51	11.3	55.54	12.8
σ_{34}	55.85	6.4	52.65	11.8	51.73	13.3
σ_{35}	58.11	7.2	54.63	12.7	53.53	14.5
σ_{44}	76.80	5.5	73.84	9.2	72.82	10.5
σ_{45}	73.47	6.5	70.05	10.9	68.81	12.5
σ_{55}	94.81	5.9	91.06	9.6	89.68	19.9
ψ_0	8.84	17.8	9.01	20.1	8.99	19.9
ψ_1	0.33	10.0	0.33	10.0	0.33	10.0
ψ_2	12.39	5.9	12.26	4.8	12.23	4.5

REFERENCES

[1] Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G., 2009, Longitudinal data analysis, Chapman and Hall, London.

[2] Albert, S. P., 1999, Tutorial in Biostatistics Longitudinal Data Analysis (Repeated Measures) in Clinical Trials, Statistics in Medicine, 18, 1707-1732.

[3] Gad, M. A. and Ahmed, S. A., 2006, Analysis of Longitudinal Data with Intermittent Missing Values Using the Stochastic EM Algorithm, Computational Statistics and Data Analysis, 50, 2702-2714.

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

[5] Little, R. J. A. and Rubin, D. B., 1987, Statistical Analysis with Missing Data, Wiley, New York.

[6] Kenward, G. M. and Carpenter, J., 2007, Multiple Imputation Current Perspectives, Statistical Methods in medical Research, 16, 199-218.

[7] Yulei, H., 2010, Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter, Circulation Cardiovascular Quality and Outcomes, 3, 98-105.

[8] Rubin, D. B., 1987, *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New York.

[9] Rubin, D. B., 1996, Multiple Imputation after 18+ years, Journal of American Statistical Association, 91, 473-489.

[10] Spratt, M., Carpenter, J., Sterne, C. J., Carlin, B. J., and Heron, J., 2010, Strategies for Multiple Imputation in Longitudinal Studies, American Journal of Epidemiology, 172, 478-487.

[11] Siddique, J. Harel, O., and Crespi, M. C., 2012, Addressing Missing Data mechanism Uncertainty Using Multiple-Model Multiple Imputation: Application to a Longitudinal Clinical Trial, The Annals of Applied Statistics, 6, 1814-1837.

[12] Donders, T. A., Heijden, G., Stijnen, T., and Moons, K., 2006, Review: A gentle Introduction to Imputation of Missing Values, Journal of Clinical Epidemiology, 59, 1087-1091.

[13] Allison, D. P., 2000, Multiple Imputation for Missing Data: A Cautionary Tale, Sociological Methods and Research, 28, 301-309.

[14] Munoz, F, J. and Rueda, M., 2009, New Imputation Methods for Missing Data Using Quantiles, Journal of Computational and Applied Mathematics, 232, 295-304.

[15] Deng, W., 2005, Multiple Imputation for Marginal and Mixed Models in Longitudinal Data with Informative missingness, Ph.D. thesis, School of Public Health, Ohio State University, USA.

[16] Yang, X., Li, J., and Shoptaw, S., 2008, Imputation Based Strategies for Clinical Trial Longitudinal Data with Non-ignorable Missing Values, Statistics in Medicine, 27, 2826-2849.

[17] Tang, L., Song, J., Belin, R. T., and Unutzer, J., 2005, A comparison of Imputation Methods in Longitudinal Randomized Clinical Trial, Statistics in Medicine, 24, 2111-2128.

[18] Olsen, K. M., Stechuchak, M. K., Edinger, D. J., Ulmer, S. C., and Woolson, F. R., 2012, Move Over LOCF: Principled Methods for Handling Missing Data in Sleep Disorder Trials, Sleep Medicine, 13, 123-132.

[19] Houck, R. P., Mazumdar, S., Koro-Sengul, T., Tang, G., Mulsant, H. B., Pollock, G. B., and Reynolds F. C., 2004, Estimating Treatment Effects From Longitudinal Clinical Trial Data with Missing values: Comparative Analyses Using Different Methods, Psychiatry Research, 129, 209-215.

[20] Diggle, P. J. and Kenward, M. G.,1994, Informative dropout in longitudinal data analysis, Journal of Royal Statistical Society B, 43, 49 – 93.

[21] Jennrich, R. I. and Schluchter, M. D., 1986, Unbalanced repeated measures models with structured covariance matrices, Biometrika, 42, 805-820.

[22] Collett, D., 1991, Modelling binary Data, Chapman and Hall, London.

[23] Efron, B., 1979, Bootstrap methods: another look at the Jacknife, Annals of Statistics, 7, 1-29.

[24] Efron, B. and Tibshirani, R. J., 1993, An Introduction to the Bootstrap, Chapman and Hall, London.