American Journal of Applied Mathematics and Statistics, 2016, Vol. 4, No. 3, 59-66
Available online at http://pubs.sciepub.com/ajams/4/3/1

DOI:10.12691/ajams-4-3-1

Maximum Likelihood Approach for Longitudinal Models with Nonignorable Missing Data Mechanism Using Fractional Imputation

Abdallah S. A. Yaseen[1], Ahmed M. Gad[2,*], Abeer S. Ahmed[1]

[1]The National Centre for Social and Criminological Research, Cairo, Egypt
[2]Statistics Department, Faculty of Economics and Political Science, Cairo University, Egypt
*Corresponding author: dr_ahmedgad@yahoo.co.uk

Abstract In longitudinal studies data are collected for the same set of units for two or more occasions. This is in contrast to cross-sectional studies where a single outcome is measured for each individual. Some intended measurements might not be available for some units resulting in a missing data setting. When the probability of missing depends on the missing values, missing mechanism is termed nonrandom. One common type of the missing patterns is the dropout where the missing values never followed by an observed value. In nonrandom dropout, missing data mechanism must be included in the analysis to get unbiased estimates. The parametric fractional imputation method is proposed to handle the missingness problem in longitudinal studies and to get unbiased estimates in the presence of nonrandom dropout mechanism. Also, in this setting the jackknife replication method is used to find the standard errors for the fractionally imputed estimates. Finally, the proposed method is applied to a real data (mastitis data) in addition to a simulation study.

Keywords: *longitudinal data, mastitis data, missing data, nonrandom dropout, parametric fractional imputation, repeated measures, standard errors*

Cite This Article: Abdallah S. A. Yaseen, Ahmed M. Gad, and Abeer S. Ahmed, "Maximum Likelihood Approach for Longitudinal Models with Nonignorable Missing Data Mechanism Using Fractional Imputation." *American Journal of Applied Mathematics and Statistics*, vol. 4, no. 3 (2016): 59-66. doi: 10.12691/ajams-4-3-1.

1. Introduction

The defining characteristic of longitudinal studies is that sample units are measured repeatedly over time. That is, data are collected for the same set of units for two or more occasions. Missing values are not uncommon with longitudinal data.

Missing data mechanisms can be classified according to the process causing missingness, as defined by Little and Rubin [17]. These include; missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) mechanism. Missing not at random mechanism is always termed nonignorable missing data mechanism. In this case the missing data mechanism must be included in the analysis, so as to get unbiased estimates.

Another important classification is the missingness pattern: the dropout and intermittent pattern. In dropout pattern a subject who leaves the study at some time point does not appear again; a missing value never followed by an observed value, whereas in intermittent pattern a missing value may be followed by an observed value.

Handling missing data requires jointly modeling the longitudinal outcome and the missing data process. There are many approaches for parametric modeling of the longitudinal outcome and the missing data process. The first is the selection models [6]. The selection models are better choice if the interest is on the inference about the marginal distribution of the response. This why we choose such models in this article. The second is the pattern mixture models [19]. The third is the shared parameter models [8]. For more details, refer to Molenberghs and Fitzmaurice [22].

The stochastic EM algorithm (SEM), suggested by Celeux and Diebolt [2], has been developed to facilitate the E-step of the EM algorithm. The stochastic EM algorithm has been extended to the longitudinal studies by Gad and Ahmed [9]. Other alternatives include the stochastic approximation EM (SAEM) algorithm [5] and the Monte Carlo EM (MCEM) algorithm [25]. Booth and Hobert [1] used an automated Monte Carlo EM algorithm to compute the E-step of the EM algorithm. A disadvantage of the MCEM algorithm is that the generated values are updated at each iteration which requires heavy computations and as a result this affects the speed of the convergence. In addition, the convergence is not guaranteed for a fixed Monte Carlo sample size [26].

Thus, the MCEM is developed using the parametric fractional imputation to facilitate the expectation step. Also, this can speed the convergence and to guarantee the existence of convergence [14,15,16,27].

Kim and Kim [16] applied the parametric fractional imputation in the context of cross-sectional studies to deal with the missingness problem in the case of nonignroable missing mechanism. Yang et al. [27] generalized the

approach to deal with the nonignorable missing mechanism in longitudinal studies using the shared parameter model.

The aim of this article is to develop the parametric fractional imputation to handle the nonignorable dropout in the context of longitudinal studies using the selection model of Diggle and Kenward [6]. In addition, the Jackknife replication method is used to obtain the standard errors of the fractionally imputed estimates. The performance of the proposed method is evaluated using a simulation study. Also, the proposed methods are applied to a real data (mastitis data). The rest of the article is organized as follows. In Section 2, the basic notations are introduced. In Section 3 the selection model for longitudinal data is introduced. The developed parametric fractional imputation method is described in Section 4. Section 5 is devoted to the proposed Jackknife method to evaluate the standard errors of the estimates. A simulation study is presented in Section 6 to evaluate the performance of the proposed methods. In Section 7 the proposed techniques are applied to the mastitis data Finally, Section 8 is devoted to conclusion.

2. Notations

Let y_{ij} be the sequence of the response outcomes and x_{ij} be the p-vector of fully observed covariate for the j^{th} measurement from the ith subject, made at time t_{ij}, $j = 1,..,n_i$, $i = 1,..,m$. Let t_{ij} denote the time at which the measurements are taken. It is supposed that t_{ij} are common for all subjects. The set of responses for the subject i are gathered into a n_i-vector $Y_i = (y_{i1},..y_{in_i})$. The variable Y_i is assumed to be normally distributed with mean μ_i and variance-covariance matrix V_i, i.e.

$$Y_i \sim \text{MVN}(\mu_i, V_i),$$

where $\mu_i = X_i\beta$, X_i is $n_i \times p$ matrix representing the covariates, β is $p \times 1$ vector of unknown parameters, V_i is the covariance matrix of dimension $n_i \times n_i$ such that its jk-element, σ_{jk}, represents the covariance between y_{ij} and y_{ik}.

The response variable Y_i can be modeled using the general linear model

$$Y_i = X_i\beta + \epsilon_i, \quad (1)$$

where ϵ_i is assumed to follow multivariate normal distribution. That is,

$$\epsilon_i \sim \text{MVN}(0, V_i).$$

The responses for all subjects are collected in an $n_i \times m$-vector, $Y = (Y_1,..,Y_m)$. The covariance matrix of Y is V. Because it is assumed that the measurements from each subject are correlated, but uncorrelated with other measurements from other subjects, the matrix V is a block diagonal matrix with non-zero blocks V_i. The matrix V_i may be unstructured containing $n_i(n_i+1)/2$ parameters or it may have a parametric structure then it is function of a vector of unknown parameters, see Diggle et al. [7].

In the missing data context the response Y_i which represents the intended observations can be classified into two sub-vectors $\{Y_{iobs}, Y_{imis}\}$, where Y_{iobs} denotes the observed measurements of the ith subject and Y_{imis} denotes the missing observations. A binary variable R_{ij} is assumed to represent the missing data process parameterized by ϕ. The R_{ij} equals 1 if y_{ij} is observed and equals 0 if y_{ij} is missing. The R_{ij} is assumed to follow Bernoulli distribution with a probability $\pi(Y_i, \phi)$. The R_{ij}'s of the subject i can be arranged in a vector R_i. The complete data of subject i can be considered as (Y_i, R_i). Let C_i be an indicator of completeness that C_i equals one if the individual i has the complete measurements and zero otherwise. Let l represents the log-likelihood function of a specific parameter.

3. Selection Model for Incomplete Longitudinal Data

Under the selection model (Diggle and Kenward, 1994), the joint distribution function of the response variable Y_i and the indicator R_i can be written as:

$$f(Y_i, R_i|X_i, \theta, \emptyset) = f(Y_i|X_i, \theta)P(R_i|Y_i; \emptyset),$$

where θ is a vector of parameters describing the response variable Y_i, X_i is a fully observed matrix of covariates (design matrix), and $\emptyset$ is a vector of parameters describing the response indicator R_i.

As defined by Little and Rubin [17], missingness is defined to be missing completely at random (MCAR) if R_i is independent of Y_{imis} and Y_{iobs}, i.e.,

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

The missing data mechanism is missing at random (MAR) if R_i is independent of Y_{imis} conditionally on Y_{iobs}, i.e.,

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

Otherwise, the missing data mechanism is missing not at random (MNAR).

Following Diggle and Kenward [6], the probability of dropout is modeled as a logistic model depending on the measurement at the time of dropout d_i; y_{d_i}, the previous measurements; H_{d_i} and the unknown parameter $\emptyset$; that is,

$$P(D_i = d_i|\, history) = P_{d_i}(H_{d_i}, y_{d_i}, \emptyset),$$

and the logistic model for the dropout process can be expressed as

$$\text{logit}\{P_{d_i}(H_{d_i}, y_{d_i}, \emptyset)\} = \emptyset_0 + \sum_{j=1}^{d_i} \emptyset_j y_{d_i-j+1}.$$

4. Maximum Likelihood Estimation for Longitudinal Data with Missing Values Using Parametric Fractional Imputation

The log-likelihood function of θ and $\emptyset$, $l(\theta, \emptyset|Y, R)$, can be any function proportional to $f(Y, R|\theta, \emptyset)$

$$l(\theta, \emptyset|Y, R) \propto f(Y, R|\theta, \emptyset).$$

If there are missing values, the observed density function can be written as

$$f(Y_{obs}, R|\theta, \varnothing) = \int f(Y|\theta)\, P(R|Y; \varnothing) dY_{mis}$$
$$= \int f(Y_{obs}, Y_{mis}|\theta)\, P(R|Y_{obs}, Y_{mis}; \varnothing) dY_{mis}$$

and the observed log-likelihood function of θ and ∅ , $l(\theta, \emptyset|Y_{obs}, R)$, will be any function proportional to $f(Y_{obs}, R|\theta, \emptyset)$, i.e.

$$l(\theta, \emptyset|Y_{obs}, R) \propto f(Y_{obs}, R|\theta, \emptyset).$$

Instead of maximizing $l(\theta, \emptyset|Y_{obs}, R)$ to get the maximum likelihood estimator of θ and ∅, Louis [20] tried to obtain the MLE by maximizing

$$Q(\theta, \varnothing) = E\left(l(\theta, \varnothing|Y, R)|Y_{obs}, R\right). \quad (2)$$

This is due to the fact that maximizing the maximizing the observed log likelihood function requires getting an explicit form for the distribution of the observed data which is hard to calculate as we have to integrate over the distribution of the observed data. So a suitable solution is to use the distribution of the complete data and maximize the expectation of the complete data given the observed data as a proxy to avoid the calculations of the observed log-likelihood function

The EM algorithm can be applied in the t^{th} iteration by calculating $Q(\theta_t, \emptyset_t)$ in the E-step. In the M-step, $\theta_{t+1}$and $\emptyset_{t+1}$are chosen to maximize the Q-function, i.e.

$$Q(\theta_{t+1}, \emptyset_{t+1}) \geq Q(\theta_t, \emptyset_t).$$

However, calculating the conditional expectation in (2) is cumbersome and time consuming. Thus, numerical approximation is needed. The MCEM approximates the Q-function in the E-step but the generated values are changed in each iteration and the convergence is not guaranteed.

The Parametric fractional imputation (PFI) develops the MCEM using the idea of the fractional weights where the generated values do not change in each iteration. Only the fractional weights are updated iteratively which guarantees the convergence and accelerates its rate. Kim and Kim [16] applied the parametric fractional imputation in cross-sectional studies. We will try to develop the parametric fractional imputation to longitudinal studies context with nonrandom dropout using the selection model of Diggle and Kenward [6] and the general linear model in (1).

The Parametric fractional imputation (PFI) algorithm can be conducted in the following steps.

(1) Generate M imputed values for the missing data Y_{imis}. Kim and Kim [16] and Yang et al [27] recommended generating the imputed values from an initial density $q(Y|X)$ with the same support as the density of the response variable. We recommend generating the imputed values from the conditional distribution of the missing data given the observed data, the response indicator and initial parameter estimates, $f(Y_{imis}|Y_{iobs}, R, X_i)$, which has the same support as the density of the outcome variable and take into account the dropout process. Unfortunately, the distribution doesn't have a standard form and it is not possible to simulate from it. Hence, an accept-reject procedure can be used to overcome this problem and to mimics the dropout process. The imputed values are generated from $f(Y_{imis}|Y_{iobs}, X_i)$ instead of $f(Y_{imis}|Y_{iobs}, R, X_i)$ and then, using an accept-reject procedure, the value can be accepted or rejected. Assuming normality, the conditional distribution, $f(Y_{imis}|Y_{iobs}, X_i)$, is also normal distribution with mean $u_{im.o}$ and covariance matrix $V_{im.o}$, where

$$u_{im.o} = u_{im} + V_{imo}V_{ioo}^{-1}(Y_{iobs} - u_{io})$$

and

$$V_{im.o} = V_{imm} - V_{imo}V_{ioo}^{-1}V_{iom}$$

where $u_{io}, u_{im}, V_{ioo}, V_{iom}$ and V_{imm} are suitable partitions of the mean vector u_i and the covariance matrix V_i.

(2) Given the M imputed values, $Y_{imis}^{*(1)}$, ..., $Y_{imis}^{*(M)}$ for the vector of missing for individual i, $Y_{mis,i}$, and the current parameter estimates θ_t, the joint density of the imputed values in the K^{th} replicate gathered in the vector $Y_{mis,i}^{*(K)}$, for $K = 1, .., M$, will be

$$f^*\left(Y_{mis,i}^{*(K)}|\theta_t\right) = \prod_{j=d_i}^{n} f\left(y_{ij}^{*(k)}|\theta_t\right)$$

where $y_{ij}^{*(k)}$ is the K^{th} imputed value for the i^{th} individual in the j^{th} time point. The K^{th} replicated data for the i^{th} individual are denoted by $Y_i^{*(K)} = (Y_{obs,i}, Y_{mis,i}^{*(K)})$. Given $Y_{mis,i}^{*(K)}$ and the current estimates θ_t and $\emptyset_t$, a fractional weight $W_{i(t)}^{*(K)}$ is assigned in the t^{th} iteration for each $Y_i^{*(K)}$ and can be calculated by

$$W_{i(t)}^{*(K)} = \frac{f^*(Y_{mis,i}^{*(K)}|\theta_t) * \prod_{j=d_i}^{n} \left\{1 - \pi\left(y_{ij}^{*(k)}, \varnothing_t\right)\right\}}{f(Y_{mis,i}^{*(K)}|Y_{obs,i})},$$

where $\pi\left(y_{ij}^{*(k)}, \emptyset_t\right)$ is the probability of missing for the value $y_{ij}^{*(k)}$ given the current estimate $\emptyset_t$.

(3) Using the K^{th} fractional weight, $W_i^{*(K)}$, and the K^{th} imputed vector, $Y_i^{*(K)}$, the Monte Carlo approximation of (2) is given by

$$Q^*(\theta_t, \emptyset_t) = [Q^*(\theta_t), Q^*(\emptyset_t)],$$

and

$$Q^*\left(\theta_t\right) = \sum_{i=1}^{m}[C_i l(\theta_t|Y_i) + (1 - C_i)\sum_{K=1}^{M} W_{i(t)}^{*(K)} l(\theta_t|Y_i^{*(K)})]$$

$$Q^*\left(\varnothing_t\right) = \sum_{i=1}^{m}\begin{bmatrix} C_i l(\varnothing_t|R_i, Y_i) \\ +(1 - C_i)\sum_{K=1}^{M} W_{i(t)}^{*(K)} l(\varnothing_t|R_i, Y_i^{*(K)}) \end{bmatrix},$$

It is worth noting that this step corresponds to the E-step in the EM algorithm.

(4) Update the parameter estimates in two sub steps; the normal step and the logistic step. In the normal step, the maximum likelihood estimate for θ is obtained using an appropriate optimization procedure, for example the Jennrich- Schulchter algorithm [13]. In the logistic step, the maximum likelihood estimates for the logistic model

$$logit\left\{P_{d_i}\left(H_{d_i}, y_{d_i}, \varnothing\right)\right\} = \varnothing_0 + \sum_{j=1}^{d_i} \varnothing_j y_{d_i - j + 1},$$

are obtained using the iterative reweighted least squares [3,21].

(5) Repeat the previous three sub-steps until convergence.

It is of great interest to mention that the imputed values are not necessarily regenerated at each iteration. Only the fractional weights are updated in each iteration. Thus, the rate of convergence is fast and the convergence is guaranteed. For sufficiently large M, the final estimates are asymptotically equivalent to the ML estimates [16].

There are many techniques depends on the idea of weights such as importance resampling [24]. However, the way of calculating the weights is different in the proposed method. Also, the aim of sampling importance resampling is to sampling from difficult distributions. This is not the case in the proposed method.

5. Standard Error Estimation

The standard errors of the estimated parameters, for fractionally imputed estimator, can be obtained using a replication method such as Jackknife or bootstrap. Kim and Kim [16] used Jackknife method to estimate the standard errors of the estimates. The Jackknife method can be conducted as follows:

(1) Generate n independent samples, $S_1, \ldots, S_n$ of size $n-1$ from the original sample by deleting one individual observation in each sample systematically, i.e. S_1 will contain $\{y_2, \ldots, y_n\}$ while S_2 will contain $\{y_1, y_3, \ldots, y_n\}$…etc.

(2) In each of the generated samples, calculate the fractionally imputed estimators $\hat{\theta}^*_{(i)}$, for $i = 1, \ldots, n$.

(3) Estimate the standard error of $\hat{\theta}^*$ using the formula

$$\widehat{S.E}\left(\hat{\theta}^*\right) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}^*_{(i)} - \hat{\theta}^*_{\cdot})^2},$$

where

$$\hat{\theta}^*_{\cdot} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}^*_{(i)}.$$

The Jackknife method is used to obtain the estimated standard errors for the estimated parameters. The Jackknife method will be applied as in Kim and Kim [16]. The only difference is that, in generating the sample S_i, the vector of the observed data of the ith individual, $Y_{i,obs}$, is omitted instead of deleting the observation y_i.

6. The Simulation Study

The aim of this simulation study is to judge the performance of the proposed method. A number of replications B = 5000 Monte Carlo samples were generated at different sample sizes. Sample sizes are chosen as 30, 50, and 100 with five time points for each individual. This choice covers small, moderate and large sample size.

The response variable Y_i was simulated from $\mathrm{MVN}(u_i, V_i)$ where $u_i = X_i\beta$. The matrix X_i is a design matrix and the vector β is of length 3; $\beta = (\beta_0, \beta_1, \beta_2)$. The logistic regression model used for the mastitis data is adopted here.

The covariance matrix is left unstructured. This means that there are 15 covariance parameters. Different covariance structures are also tried, such as compound symmetric, exponential structure. For definition of these structures see Diggle et al [7]. Data were generated to meet the assumptions of the multivariate normal, the assumed covariance model and the missingness model. The missing data is generated by a similar technique to that used by Kim and Kim [16], that the binary random variable is generated from the Bernoulli distribution with parameter equal to the probability of missingness for the specified value. The value is omitted if its associated binary variable is one. The logistics model is used to describe the dropout process is

$$logit\left\{P_{d_i}\left(H_{id_i}, y_{id_i}, \varnothing\right)\right\} = \varnothing_0 + \varnothing_1 y_{id_i-1} + \varnothing_2 y_{id_i},$$

$$d_i = 2, \ldots, 5.$$

Under this setup, the vector of parameters is $\theta = (\beta, \sigma, \varphi)$, where $\beta = (\beta_0, \beta_1, \beta_2)$, and $\emptyset = (\emptyset_0, \emptyset_1, \emptyset_2)$ and

$$\sigma = \begin{pmatrix} \sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2, \sigma_{12}, \sigma_{13}, \sigma_{14}, \\ \sigma_{15}, \sigma_{23}, \sigma_{24}, \sigma_{25}, \sigma_{34}, \sigma_{35}, \sigma_{45} \end{pmatrix}.$$

These parameters are fixed at the values $\beta = (0.5,1,6)$, $\sigma = (10,10,9,10,10,6,7,6,9,5,6,6,9,7,6)$ and $\emptyset = (-5.6, 0.03, 0.04)$ for low missingness rate and $\emptyset = (-5.6, 0.07, 0.04)$ for high missingness rate respectively. The simulation is conducted under two missingness rates (percentage of individuals with missing data over the 5000 replications). The low missingness rate with missingness percentage ranges from 13% to 17%; and high missingness rate with missingness percentage ranges from 40% to 50%.

The parameter estimates have been obtained using the following methods;

(1) The multiple imputation with $M = 10$ where M is the number of the imputed values (MI).

(2) The fractional regression nearest neighbor imputation with $M = 10$ (FRNNI).

(3) The parametric fractional imputation with M=10 (PFI).

The choice of ten replicates for the PFI is to test if the proposed method can compete with other techniques at modest number of replicates, also to simplify the calculations. As proved by Yang et al [27], the more replicates are used, the better estimates are obtained. Therefore, it is expected to obtain more precise estimates by increasing the number of replicates. In the multiple imputations, we use predictive mean matching method described in Grannell and Murphy [11]. For the fractional regression nearest neighbor imputation, we apply the method described in Paik and Larsen [23].

The results are shown in Table 1 - Table 6. The *multiple imputation* (MI) estimates of the mean parameters have small relative bias but large relative bias for the covariance parameters in the case of low missingness rate. For high missingness rate, the estimates are seriously biased comparable to the other methods.

Table 1. The relative bias percentage for the simulation study at n=30, high missingness rate; PFI=parametric fractional imputation,, FRNNI= fractional nearest neighbor imputation, MI=multiple imputation

parameter	PFI	FRNNI	MI
β_0	0.0	22.0	24.0
β_1	0.0	-1.0	0.0
β_2	0.0	-22.5	-26.8
σ_1^2	-1.1	110.5	201.5
σ_2^2	4.2	92.1	209.3
σ_3^2	-0.9	118.0	240.1
σ_4^2	0.4	103.2	249.5
σ_5^2	-4.3	155.8	369.9
σ_{12}	-12.7	83.2	154.6
σ_{13}	-7.4	71.3	141.8
σ_{14}	-11.9	89.5	172.4
σ_{15}	-2.6	59.1	133.8
σ_{23}	-13.6	85.5	144.2
σ_{24}	-7.9	69.0	146.8
σ_{25}	-10.9	95.6	160.0
σ_{34}	-1.2	15.5	94.8
σ_{35}	-6.6	71.5	153.7
σ_{45}	-11.9	86.8	149.7
$\emptyset_0$	-4.3	---	---
$\emptyset_1$	-8.9	---	---
$\emptyset_2$	-5.3	---	---

Table 2. The relative bias percentage for the simulation study at n=50, high missingness rate ; PFI=parametric fractional imputation, FRNNI= fractional nearest neighbor imputation, MI=multiple imputation

parameter	PFI	FRNNI	MI
β_0	0.0	26.0	32.0
β_1	0.0	1.0	1.0
β_2	0.0	-18.0	-24.2
σ_1^2	0.6	16.9	102.1
σ_2^2	5.3	15.1	115.0
σ_3^2	0.8	35.4	125.0
σ_4^2	2.2	29.4	128.5
σ_5^2	-1.7	60.4	214.0
σ_{12}	-10.5	-9.7	55.4
σ_{13}	-6.6	-11.9	51.0
σ_{14}	-11.4	-5.8	61.8
σ_{15}	-1.8	-14.2	41.9
σ_{23}	-12.5	-8.8	55.8
σ_{24}	-7.5	-15.8	47.9
σ_{25}	-9.9	-1.7	72.3
σ_{34}	0.3	-37.0	33.7
σ_{35}	-6.1	-9.9	66.4
σ_{45}	-12.1	-6.3	44.2
$\emptyset_0$	2.6	---	---
$\emptyset_1$	-4.6	---	---
$\emptyset_2$	3.8	---	---

The *fractional regression nearest neighbor* (FRNNI) imputation leads to reasonable estimates for low missingness rate. It leads to relatively biased estimates, especially for the covariance model, in the case of high missingness rate. This can be noticed clearly for small and moderate sample sizes.

The *parametric fractional imputation* (PFI) estimates are relatively unbiased for most parameters regardless of the missingness rate and the sample size. The covariance estimates have small bias that has a negative relation with the response rate and the sample size. In general, the bias ranges from small to moderate and decreases with larger sample sizes. Using the proper (right) weights produce estimates with small bias. In general, the PFI estimates have lower bias rates comparable to the other two methods. In fact the parametric fractional imputation (PFI) method approximates the maximum likelihood estimates in the case of very large replications.

Table 3. The relative bias percentage for the simulation study at n=100, high missingness rate; PFI=parametric fractional imputation, FRNNI= fractional nearest neighbor imputation, MI=multiple imputation

parameter	PFI	FRNNI	MI
β_0	0.0	24.0	-26.0
β_1	0.0	3.0	4.0
β_2	0.0	-13.5	-14.2
σ_1^2	-0.1	-30.9	-18.8
σ_2^2	4.3	-23.9	51.9
σ_3^2	-0.5	-10.8	44.1
σ_4^2	1.0	-11.9	82.1
σ_5^2	-2.3	4.4	134.1
σ_{12}	-12.1	-51.2	-56.3
σ_{13}	-7.6	51.3	-53.2
σ_{14}	-12.6	-50.5	-56.6
σ_{15}	-2.6	-50.5	-54.8
σ_{23}	-14.2	-51.0	-49.8
σ_{24}	-8.8	-53.1	-34.5
σ_{25}	-11.4	-46.5	-34.2
σ_{34}	-1.0	-60.9	3.8
σ_{35}	-7.4	-50.0	-22.6
σ_{45}	-11.9	-50.2	-59.4
$\emptyset_0$	1.2	---	---
$\emptyset_1$	-2.5	---	---
$\emptyset_2$	1.4	---	---

Table 4. The relative bias percentage for the simulation study at n=30, low missingness rate; PFI=parametric fractional imputation, FRNNI= fractional regression nearest neighbor imputation, MI=multiple imputation

parameter	PFI	FRNNI	MI
β_0	0.0	6.0	4.0
β_1	0.0	-1.0	-1.0
β_2	0.0	-5.7	-5.0
σ_1^2	0.4	8.5	8.1
σ_2^2	0.6	11.6	34.2
σ_3^2	0.0	18.7	37.1
σ_4^2	0.3	18.8	55.7
σ_5^2	-0.8	23.8	78.0
σ_{12}	-1.3	8.3	4.9
σ_{13}	-1.4	2.7	0.2
σ_{14}	-2.4	7.5	3.4
σ_{15}	-0.3	-3.3	-6.4
σ_{23}	-2.4	12.8	5.9
σ_{24}	-1.9	7.1	7.9
σ_{25}	-1.4	7.5	6.8
σ_{34}	-0.1	-5.0	16.8
σ_{35}	-1.6	1.6	8.1
σ_{45}	-2.6	7.0	-4.1
$\emptyset_0$	0.7	---	---
$\emptyset_1$	-5.2	---	---
$\emptyset_2$	2.8	---	---

Table 5. The relative bias percentage for the simulation study at n=50, low missingness rate ; PFI=parametric fractional imputation, FRNNI= fractional regression nearest neighbor imputation, MI=multiple hot deck imputation

parameter	PFI	FRNNI	MI
β_0	0.0	4.0	4.0
β_1	0.0	0.0	0.0
β_2	0.0	-4.8	-4.3
σ_1^2	0.4	-5.1	-5.5
σ_2^2	-1.0	-2.1	19.6
σ_3^2	-0.9	3.8	21.3
σ_4^2	-0.9	3.9	39.5
σ_5^2	-1.3	7.7	55.0
σ_{12}	-2.1	-6.0	-8.5
σ_{13}	-1.7	-9.2	-10.9
σ_{14}	-3.1	-6.7	-9.8
σ_{15}	-0.6	-13.2	-15.6
σ_{23}	-3.2	-3.6	-7.5
σ_{24}	-2.9	-7.1	-3.8
σ_{25}	-2.6	-7.0	-5.1
σ_{34}	-1.3	-14.1	8.8
σ_{35}	-2.2	-10.3	-1.7
σ_{45}	-3.7	-7.4	-12.4
$\emptyset_0$	0.9	---	---
$\emptyset_1$	-5.1	---	---
$\emptyset_2$	3.5	---	---

Table 6. The relative bias percentage for the simulation study at n=100, low missingness rate; PFI=parametric fractional imputation, FRNNI= fractional regression nearest neighbor imputation, MI=multiple imputation

parameter	PFI	FRNNI	MI
β_0	0.0	3.7	2.0
β_1	0.0	-0.5	-1.0
β_2	0.0	-4.1	-3.5
σ_1^2	0.2	-12.6	-14.2
σ_2^2	-0.1	-10.4	7.2
σ_3^2	-0.1	-5.9	9.3
σ_4^2	0.0	-5.5	25.7
σ_5^2	-0.9	-2.2	36.1
σ_{12}	-2.2	-13.5	-16.6
σ_{13}	-1.4	-15.5	-17.9
σ_{14}	-2.6	-14.0	-17.8
σ_{15}	-0.5	-18.0	-21.0
σ_{23}	-2.8	-12.2	-14.7
σ_{24}	-2.2	-14.5	-11.3
σ_{25}	-2.4	-14.2	-11.6
σ_{34}	-0.3	-18.8	2.1
σ_{35}	-1.4	-16.1	-7.5
σ_{45}	-2.7	-14.4	-15.0
$\emptyset_0$	2.2	---	---
$\emptyset_1$	-3.2	---	---
$\emptyset_2$	4.8	---	---

Hence, within the simulation context and depending on its results, we can conclude that the parametric fractional imputation method (PFI) provides reasonable estimates for the parameters in the case of nonrandom dropout even with small sample size.

7. Application (Mastitis Data)

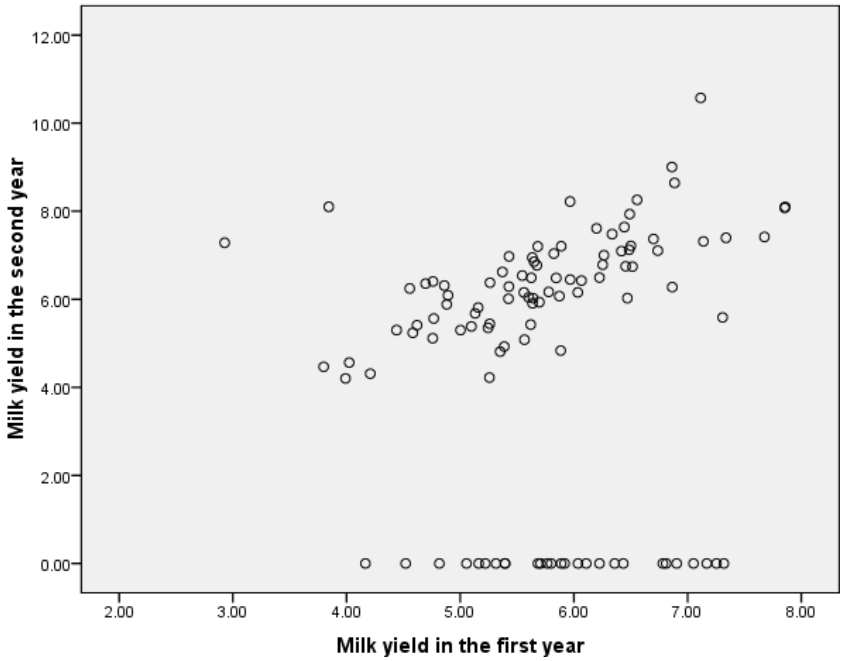

Figure 1. The first year yield vs. the second year yield of mastitis data

Mastitis is an infection of the udder causes reduction in the milk yield of the infected animals. The data set contains the total milk yield, in thousands litres, for 107 cows from a single herd for two successive years. 27 cows became infected in the second year and as a result their observations, although recorded, are considered missing in the second year. It is intended to compare the average of the milk yield in the two years. The data set has been analyzed by Diggle and Kenward [6] resulting that the type of missing is NMAR. They suggest using Nelder-Mead simplex algorithm to find parameters estimates. Also, the data has been analyzed by Gad and Kenward [10] where they used the stochastic EM algorithm to obtain parameters estimates. Figure 1 shows a scatter plot of the first year yield against the second year milk.

It is clearly from the graph that there is a strong positive correlation between the two yields and there are two values with high milk yield in the second year and low milk yield in the first year.

Figure 2 shows the profile lines of the completers; cows with data available for the two years. The graph shows a positive increase of the measurements from the first year to the second year.

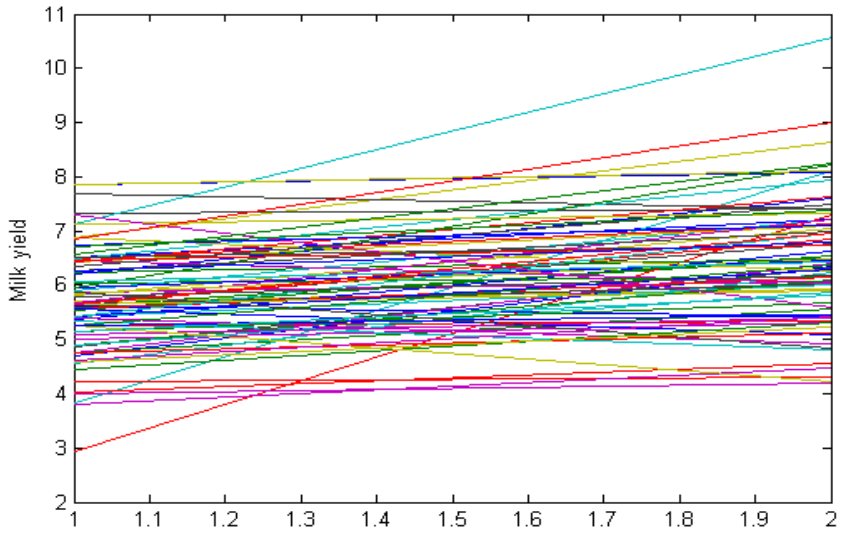

Figure 2. Milk yields for the completers (80 cows)

The data is analyzed using the linear model

$$Y_i = \mu + \epsilon_i \text{ for } i = 1, \ldots, 107$$

where Y_i is a vector containing the two observations for the i^{th} animal, and

$$\mu = (\mu_1, \mu_2),$$

where μ_1 and μ_2 is the average of the response in the first and second year respectively. The compound symmetric covariance structure is chosen for the covariance matrix. In the compound symmetric structure, the covariance matrix, V, takes the form

$$V = \sigma_1^2 I + \sigma_2^2 J$$

where I is the identity matrix of order 2×2 and J is 2×2 matrix all of whose elements are ones. Hence, there are two covariance parameters σ_1^2 and σ_2^2.

The unstructured covariance structure is also used for the covariance matrix. In this case three covariance parameters need to be estimated; $\sigma_{11}, \sigma_{12}, \sigma_{22}$. The estimates of the parameters are almost identical with the compound symmetric structure. So, the results of the unstructured covariance structure are only shown.

Without loss of generality, we assume that the dropout process depends on the measurement at the dropout time, the previous measurement and the unknown parameters $\emptyset$. The dropout process is modeled as

$$logit\left\{P_{d_i}\left(y_{i2}, H_2\right)\right\} = \varnothing_0 + \varnothing_1 y_{i1} + \varnothing_2 y_{i2}.$$

The PFI is applied to the data with M=10 and the standard errors are calculated using Jackknife replication method. The results are shown in Table 7. MI and FRNNI are not applicable in this kind of data where all the subjects share the same independent variable, thus we apply the MI in this data set using the propensity score method described in Grannell and Murphy [12]. For the sake of comparison between the proposed method and the previous analyses we also include in Table 7 the results of Diggle and Kenward [6] and Gad and Kenward [10].

Table 7. the PFI estimates and their standard errors for mastitis data. Also, the MI estimates, the Diggle-Kenward estimates and Gad-Kenward estimates

Parameter	PFI Estimates	S.E	D-K Estimates	G-K Estimates	MI Estimates
μ_1	5.76	0.09	5.77	5.76	5.76
μ_2	6.32	0.12	6.09	6.09	6.48
σ_1^2	0.87	0.12	0.83	0.87	.87
σ_2^2	1.37	0.28	1.67	1.63	1.17
σ_{12}	0.56	0.18	0.56	0.56	.66
$\emptyset_0$	2.01	2.09	0.15	0.40	-
$\emptyset_1$	2.22	0.58	2.38	2.38	-
$\emptyset_2$	-2.79	0.58	- 2.63	- 2.70	-

The results in Table 7 show that the average of the second year yield is larger than the average of the first year yield. The two estimates are statistically significant. A closer look at the dropout parameters shows that the probability of missingness has a negative relation with the second observation. This is natural because the infection with mastitis reduces the milk yield and this also supports the assumption of MNAR. The estimates of $\emptyset_2$ is slightly bigger in absolute value than $\emptyset_1$. Both of the estimates are statistically significant. The Z-values for testing both the null hypotheses are significant at 95% confidence interval. For the MI estimates, the estimates of the mean parameters are reasonable but it seems the estimate of σ_{22} may be underestimated. Our results are similar to the results of Diggle and Kenward [6] and Gad and Kenward

[10] for the model estimates and the covariance estimates. There are slightly differences in the estimates of the missingness. This is may be due to using different approach and different dropout models.

8. Conclusion

The parametric fractional imputation is proposed as an innovative tool for parameters estimation in the presence of the missing values. If the parametric fractional imputation is used to construct the score function, the solution to the imputed score equation is approximately the maximum likelihood estimator. The PFI is superior to the MCEM or SEM in the sense that the imputed values are not regenerated at each iteration which guarantees the convergence and accelerate its rate. Variance estimation can be obtained using a replication method such as Jackknife or bootstrap. The simulation results show that the proposed techniques provide reasonable estimates. However, one limitation of this technique is that its accuracy, like other variants of the EM algorithm, depends heavily on the assumptions of the selection model which may not be totally correct. Kim and Yu (2009) proposed a semi parametric approach for applying the fractional imputation method when the assumptions of the assumed model are suspected. A further research is recommended in this topic especially for the repeated measures data but this is out of the scope of this article.

References

[1] Booth, J. G. and Hobert, J. P. (1999) Maximizing generalized linear models with an automated Monte Carlo EM algorithm, *Journal of the Royal Statistical Society* B, 61, 625-85.

[2] Celeux, G. and Diebolt, J. (1985) The SEM algorithm: A probabilistic Teacher algorithm derived from the EM Algorithm for the mixture problem, *Computational Statistics Quarterly*, 2, 73-82.

[3] Collett, D. (1991) *Modelling Binary Data*, Chapman and Hall, London.

[4] Dempster, A. P, Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Society B*, 39, 1-38.

[5] Delyon, B, Laird, N. M. and Rubin, D. B. (1999) Convergence of a stochastic approximation version of the EM algorithm, *The Annals of Statistics*, 27(1), 94-128.

[6] Diggle, P. J. and Kenward, M. G. (1994) Informative dropout in longitudinal data analysis, *Journal of the Royal Statistical Society B*, 43, 49-93.

[7] Diggle, P.J. Liang, K. Y. and Zeger, S. L. (1994) *Analysis of Longitudinal Data*, Oxford: Oxford Science, UK.

[8] Follmann, D. and Wu, M. (1995) An approximate generalized linear model with random effects for informative missing data, *Biometrics*, 51, 151-168.

[9] Gad, A. M. and Ahmed, A. S. (2006) Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm, *Computational Statistics & Data Analysis*, 50, 2702-2714.

[10] Gad, A. M. and Kenward, M. G. (2001) The Stochastic EM algorithm and sensitivity analysis for nonrandom dropout models, *Proceedings of the 12th Conference for Statistics and Computer Modelling in Human and Social Sciences*, Faculty of Economics and Political Science, Cairo University, Cairo, Egypt.

[11] Gad, A. M. and Youssif, N. A. (2006) Linear mixed models for longitudinal data with nonrandom dropouts, *Journal of Data Science*, 4(4), 447-460.

[12] Grannell, A. and Murphey, H. (2011) Using multiple imputation to adjust for survey non-response, *Proceedings of the sixth ASC international conference*, University of Bristol, UK.

[13] Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated-measures models with structured covariance matrices, *Biometrics*, 42, 805-820.

[14] Kim, J. K. (2011) Parametric fractional imputation for missing data analysis, *Biometrika*, 98, 119-132.

[15] Kim, J. K. and Fuller, W. (2008) Parametric fractional imputation for missing data analysis. Proceeding of the section on survey research method, *Joint Statistical Meeting*, pp. 158-169.

[16] Kim, J. Y. and Kim, J. K. (2012) Parametric fractional imputation for nonignorable missing data, *Journal of the Korean Statistical Society*, 41, 291-303.

[17] Little, R. J. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*, John Wiley & Sons, New York.

[18] Kim, J. K. and Yu, C. L. (2009) A semi-parametric approach to fractional imputation for nonignorable missing data, *Survey Research Methods Proceedings*: 2603-2610.

[19] Little, R. J. (1993) Pattern mixture models for multivariate incomplete data, *Journal of American Statistical Association*, 88, 125-134.

[20] Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm, *Journal of the Royal Statistical Society B*, 44, 226-233.

[21] McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, Chapman and Hall, London.

[22] Molenberghs, G. and Fitzmaurice, G. (2009) *Longitudinal Data*, Fitzmaurrice, G. Davidian, M. Verbeke, M. and Molenberghs, G. Editors, Chapman & Hall/CRC Taylor & Francis Group, USA, ch.17.

[23] Paik, M. and Larsen, M. D. (2006) Fractional regression nearest neighbor imputation, *Proceedings of the joint statistical meeting, American Statistical Association*, Alexandria, VA.

[24] Rubin, D. B. (1987) A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the SIR algorithm. Discussion of Tanner and Wong, Journal of the American Statistical Association, 82, 543-546.

[25] Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm, *Journal of the American Statistical Association*, 85, 699-704.

[26] Yang, X. Li, J. and Shoptaw, S. (2008) Imputation –based strategies for clinical trial longitudinal data with nonignorable missing values, *Statistics in Medicine*, 27, 2826-2849.

[27] Yang, S., Kim, J. K. and Zhu, Z. (2012) Parametric fractional imputation using adjusted profile likelihood for linear mixed models with nonignorable missing data, Proceeding of the section on survey research method, *Joint Statistical Meeting*, pp. 4366-4376.