Analysis of Longitudinal Data with Missing Values in the Response and Covariates Using the Stochastic EM Algorithm

Ahmed M. Gad[1]
Business Administration Department, Faculty of Business Administration, Economics and Political Science, The British University in Egypt (BUE), Cairo, Egypt

Nesma M. Darwish
Business Administration Department, Faculty of Politics, Economics and Business Administration, May University in Cairo (MUC), Egypt

ABSTRACT

Longitudinal data are not uncommon in many disciplines where repeated measurements on a response variable are collected for each subject. Missing values are unavoidable in longitudinal studies. Missing values could be in the response variable, the covariates or in both. Dropout pattern occurs when some subjects leave the study prematurely. When the probability of missingness depends on the missing value, and may be on the observed values, the missing data mechanism is termed as non-random. Ignoring the missing values in this case leads to biased inferences. In this paper we will handle missing values in covariates using multiple imputations (MI) and the selection model to fit longitudinal data in the presence of non-random dropout. The stochastic EM (Expectation-Maximization) algorithm is developed to obtain the model parameter estimates. Also, parameter estimates of the dropout model have been obtained. Standard errors of estimates have been calculated using the developed Monte Carlo method. The proposed approach performance is evaluated through a simulation study. Also, the proposed approach is applied to a real data set.

Keywords: *Interstitial Cystitis data; missing covariates; dropout missingness; multiple imputation; selection model; the SEM algorithm.*

1. Introduction

In longitudinal studies each subject is measured repeatedly, for the same response variable at different times or under different condition or both. Longitudinal data are very common in biomedical research and clinical trials where some of measurements on a subject develops over time. In these cases, one variable is the underlying characteristic or measurement. The main advantage of longitudinal studies is that it can distinguish changes over time within individuals and enabling direct study of that change.

[1] **Address correspondence to Ahmed M. Gad**: ahmed.gad@feps.edu.eg

Missing data are not uncommon in longitudinal studies. The missing values could be due to many reasons. The missing data in the response occur whenever one or more of measurement sequences are incomplete. Missing data in the response can be categorized into two different patterns: intermittent missing pattern and dropout pattern. In intermittent pattern a missing value could be followed by an observed value. Dropout pattern means a missing value is never followed by an observed value.

It is commonly assumed that the responses are missing at the time of dropout, but all covariates are completely observed. Little (1995) reviews some approaches where the covariates are completely observed. However, the response variable and the associated covariates maybe not observed at the time of dropout. Hence, the assumption of completely observed covariates is often not realistic. In this article we focus on missingness in response and covariates.

In the case of missingness in the response, Erler et al. (2016) evaluate the performance of multiple imputation chained equation using different strategies to include a longitudinal response into the imputation models and compare it with a fully Bayesian approach. Nooraee et al. (2018) investigate a hybrid approach which is a combination of maximum likelihood and multiple imputation, i.e. scales from the imputed data are eliminated if all underlying items were originally missing. Abdelwahab et al. (2019) propose a sensitivity analysis index for shared parameter models in longitudinal studies. Darwish et al. (2020) propose using multiple imputation for missing at random (MAR) cross-sectional covariates. They employ a shared parameter model to fit response variable in the presence of non-random dropout.

Assume that Y_{ij} is the longitudinal response of subject i at time point j and R_{ij} is the missing data mechanism indicator, where R_{ij} equals 1 if Y_{ij} is observed and 0 if Y_{ij} is missing. In the selection model the joint distribution of the response Y_i and R_i are factorized as product of the marginal distribution of Y_i and conditional distribution of R_i given Y_i. Thus

$$f(Y_i, R_i|\theta, \Psi) = f(Y_i|\theta)P(R_i = r_i|Y_i, \Psi), \tag{1}$$

where θ is a vector containing the model parameters, $P(R_i = r_i|Y_i, \Psi)$ is the distribution that characterizes the missing data mechanism, and Ψ is a vector of parameters that govern the missing data mechanism. According to Rubin's taxonomy, the missing data mechanism can be classified to three different mechanisms (Rubin, 1976). The first is missing completely at random (MCAR) if R_i and Y_i are independent, i.e.

$$P\left(R_i = r_i \middle| Y_{i,obs}, Y_{i,mis}\Psi\right) = P(R_i = r_i|\Psi), \tag{2}$$

where $Y_{i,obs}$ and $Y_{i,mis}$ are the observed and missing parts of Y_i, respectively. The second is missing at random (MAR) if the conditional distribution of R_i given Y_i depends only on the observed, $Y_{i,obs}$, i.e.

$$P\left(R_i = r_i \middle| Y_{i,obs}, Y_{i,mis}\Psi\right) = P(R_i = r_i|Y_{i,obs}, \Psi). \tag{3}$$

The third is nonrandom (informative) if it is neither MCAR nor MAR. In dropout pattern, Diggle and Kenward (1994) propose a selection model for longitudinal data with nonrandom dropout. They specified a normal linear model for the response variable, Y_i, and a logistic model for the probability of dropout. They suggest modelling the probability of dropout at time d_i as a function of the measurement at time d_i and the observed measurements (history H_{d_i}) up to time $d_i - 1$; that is,

$$P(D_i = d_i | history) = P_{d_i}(H_{d_i}, y_{d_i}, \Psi). \tag{4}$$

Also, they suggest using the logistic model for the dropout process as

$$\text{logit}\{P_{d_i}\{H_{d_i}, y_{d_i}, \Psi)\} = \psi_0 + \sum_{j=1}^{d_i} \psi_j \, y_{d_{i-j+1}}. \tag{5}$$

The SEM algorithm has been proposed by Celuex and Diebolt (1985) as a stochastic version of the EM algorithm. The SEM algorithm overcomes the main difficulty of the EM algorithm, in some situations, by avoiding explicit calculation of the E-step. The E-step is replaced by the stochastic step (S-step) where the missing data are imputed with a single draw from the conditional distribution of the missing data given the observed data. In the M-step, the log-likelihood function of the pseudo-complete can be maximized using standard maximization procedures. So, the algorithm involves iterating two steps, the S-step and the maximization step (M-step) for sufficient number of iterations.

The estimated parameter values corresponding to each pseudo-complete data form a Markov chain. This Markov chain converges reasonably quickly to its stationary distribution, which is unique (Diebolt and Ip, 1996). The mean of the points, ignoring the early first points as a burn-in period, generated by the SEM algorithm can be considered as an estimate for the parameter β. This mean is called the SEM estimate and denoted by $\tilde{\beta}$ (Diebolt and Ip,1996). Gad and Ahmed (2006) apply the SEM algorithm to longitudinal data with dropout in the response. Different variants of SEM algorithm are also used to escape poor local maxima using the concepts of simulated annealing (Allassonnière and Chevallier, 2021). Yassen and Gad (2020) introduce different variants of the SEM algorithm to deal with mixed continuous and discrete longitudinal data.

The EM algorithm, also the SEM algorithm, does not provide the standard errors of the parameter estimates. Several methods have been proposed in literature to solve this problem. Louis' formula (Louis, 1982) relates the observed information matrix to the conditional expectation of the second derivatives of complete data log-likelihood function and the covariance of the first derivatives of complete data log-likelihood function. Evaluating the integrals in this formula, in the current setting, may not be easy. Efron (1994) suggests using simulation (the Monte Carlo method) to approximate the integrations. The missing values are simulated from their conditional distribution and then each integration is evaluated by its empirical version.

The aim of this article is to suggest a multiple imputation approach for cross-sectional covariates. The selection model is adopted for fitting linear regression model between longitudinal response and cross-sectional covariates where both the response and the covariates have missingness. The rest of the article is organized as follows. In Section 2 we present the two common multiple imputation methods that can be used to handle missingness in covariates. In Section 3 the proposed approach is described in addition to the Monte Carlo method as a way for obtaining the standard error estimates. In section 4, a simulation study is presented to validate the proposed approach. In section 5 the proposed approach is applied to a real data. Finally, Section 6 presents the conclusion and future work.

2. Multiple Imputations (MI) Methods

Grannell and Murphy (2011) discuss the application of four multiple imputations (MI) methods using the SOLAS package. Salfran and Spiess (2015) describe some of the most common imputation methods included in software packages. The most common two multiple imputations methods are described below in more details.

2.1 Regression-based imputation

In the regression method, a regression model is fitted for each variable with missing values using the complete cases. Based on the resulting model we impute the missing values because the data set has a monotone missing data pattern. The process is repeated sequentially for all variables with missing values. That is, for a variable Y_j, with missing values, a model

$$Y_j = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_k X_k \tag{6}$$

is fitted using the observed values and the corresponding covariates $(X_1, \ldots, X_2)$. The fitted model includes the regression parameter estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and the association covariance matrix $\widehat{\sigma^2} V_j$.

The following steps are used to generate the imputed values.

1- New parameter $\beta_* = (\beta_{0*}, \beta_{1*}, \ldots.., \beta_{k*})$and σ^2_{*j} are obtained from the posterior predictive distribution of the parameters. That is, they are simulated from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ and the covariance matrix $\widehat{\sigma^2} V_j$. The variance σ^2_{*j} is obtained as

$$\sigma^2_{*j} = \frac{\sigma^2 (n_j - k - 1)}{g}, \tag{7}$$

where g is a chi-square $\chi^2_{n_j - k - 1}$ random variate and n_j is the number of non-missing observations of Y_j . The regression coefficients are obtained as

$$\beta_* = \hat{\beta} + \sigma^2_{*j} V'_{hj} Z, \tag{8}$$

where V'_{hj} is the upper triangular matrix in the Cholesky decomposition of V_j and Z is a vector of $k+1$ independent standard normal variates with 0 means and a variance of 1.

2- The missing values are then replaced by

$$y_{i*} = \beta_{0*} + \beta_{1*} x_1 + \beta_{2*} x_2 + \ldots.. + \beta_{k*} x_k + z_i \sigma_{*j}, \tag{9}$$

where $(x_1, x_2, \ldots., x_k)$ are the values of the covariates and z_i is a simulated standard normal (random) deviate using standard statistical packages.

The regression method can be extended to deal with longitudinal data with dropout. In the sequential imputation method, assuming that the data are complete at the first time point, the regression imputation-based method described above can be used to impute the missing values at the second time point. The previously imputed values (possibly a subset of them) are used in the imputation model as predictors for future values. This process is repeated at the third time point and sequentially up to the final time point. The process is repeated M times to obtain M completed data sets. In principle, the normal-based regression model can be replaced by any appropriate model, for other types of responses, for example logistic regressions for binary data, or proportional odds models for ordinal data.

2.2 Predictive mean matching method (PMM)

The predictive mean matching method (PMM) can also be used for imputation. It is like the regression method, except that for each missing value, it imputes an observed value which is closest to the predicted value using the simulated regression model (Rubin 1987). The predictive mean matching method ensures that imputed values are plausible, and may be more appropriate than the regression method, if the normality assumption is violated. The steps of the PMM method are the same as the regression method. However, the PMM method needs generating a set of k_0 observations whose corresponding predicted values are closest to y_{i*}. The missing value is then replaced by a value drawn randomly from these k_0 observed values.

3. The Proposed Approach

First, we handle the missingness in covariates through multiple imputation using the regression method or predictive mean matching method. Second the SEM algorithm can be applied using the pseudo complete covariates using the two steps: the S-step and the M-step.

- **Imputing continuous cross-sectional covariates with monotone missingness using regression method.**

Depending on the observed part of the response and the observed cross-sectional covariates, the missing covariates are imputed using the model $x_{i,mis} = \beta_0 + \beta_1 y_{i,obs} + \beta_2 x_{i,obs}$.

The following steps are used to generate the imputed values for each imputation.

1. New parameters $\beta_* = (\beta_{0*}, \beta_{1*}, \beta_{2i*})$and $\sigma_{*_i}^2$ are drawn from the posterior predictive distribution of the parameters. That is, they are simulated from $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)$ and the associated covariance matrix $\hat{\sigma}^2 V_i$. The variance is obtained as $\sigma_{*i}^2 = \frac{\sigma^2(n_i-k-1)}{g}$, where g is a chi-square $\chi_{n_i-k-1}^2$ random variate and n_i is the number of non-missing observations for X_i . The regression coefficients are drawn as $\beta_* = \hat{\beta} + \sigma_{*j}^2 V'_{hj} Z$, where V'_{hj} is the upper triangular matrix in the Cholesky decomposition of V_j and Z is a vector of k+1 independent random standard normal variates.
2. The missing values are then replaced by $x_{i*} = \beta_{0*} + \beta_{1*} y_i + \beta_{2*} x_{i,obs} + z_i \sigma_{*_i}^2$, where y_i is the longitudinal response and $x_{i,obs}$ is the observed covariates and z_i is a simulated standard normal deviate.

- **Imputing continuous cross-sectional covariates with monotone missingness using predictive mean matching method.**

The above steps are used to generate imputed values with extra following two steps.

1- Generate a set of k_0 observations whose corresponding predicted values are the closest to x_{i*}.

2- The missing value is then replaced by a value drawn randomly from these k_0 values.

Here, the SEM algorithm can be applied using the pseudo complete covariates using the two steps: the S-step and the M-step.

The S-Step

In this step, the missing response values are simulated from their conditional distribution, given the observed values and the current parameter estimates, $Y_{i,mis} \sim f\ (\ Y_{i,mis}|Y_{i,obs}, R_i; \theta)$. This distribution does not have a standard form; hence it is not possible to use the direct simulation. To overcome this problem, we adopt an acceptance/rejection Monte Carlo simulation method, to generate the missing values $y_{i,mis}$. This procedure mimics the dropout process assuming that the postulated dropout model is correct. A draw from $f\left(y_{i,mis}\middle|Y_{i,obs}, \theta^{(t)}\right)$ is obtained instead of $f(y_{i,mis}\ |Y_{i,obs}, R_i, \theta^{(t)})$. Then, this value can be accepted or rejected using Metropolis Hasting procedure (Gilks et al, 1996)). The steps of this procedure can be summarized as follows:

1. Generate a candidate value, $\boldsymbol{y}^*$ from the conditional distribution function $\boldsymbol{f\left(y_{i,mis}\middle|Y_{i,obs}, \theta^{(t)}\right)}$ which is normal distribution. Only the first dropped observation is simulated and the remaining dropped values are considered missing at random (Gad and Ahmed, 2006).

2. Calculate the probability of dropout for the candidate value $\boldsymbol{y}^*$, according to the dropout model $\boldsymbol{P(D_i = d_i|H_i d_i) = \Psi_0 + \Psi_1 y_{d_i} + \sum_{j=2}^{d_i} \Psi_j\, y_{i,d_i+1-j}}$, where the parameters $\boldsymbol{\Psi_j}$ are fixed at the current values $\boldsymbol{\Psi_j^{(t)}}$. Let us denote this probability of dropout, $\boldsymbol{P(D_i = d_i|H_i d_i)}$, as $\boldsymbol{P_i}$.

3. Simulate a random variable U from the uniform distribution on the interval [0,1] that is, $\boldsymbol{U \sim U[0, 1]}$, then take $\boldsymbol{y_{i,mis}\ =\ y^*}$ if $\boldsymbol{U \le P_i}$; otherwise repeat Step1.

The M-Step

It consists of two sub-steps: the logistic step (M1-step) and the normal step (M2-step).

- In the logistic step (M1-step), the MLEs for the dropout logistic model

$$\mathbf{logit}\{\boldsymbol{P_i}\} = \boldsymbol{\psi_0 + \psi_1 y_{d_i} + \sum_{j=2}^{d_i} \psi_j\, y_{i,d_i+1-j}} \tag{10}$$

are obtained. The iteratively reweighted least squares method for finding the MLE of binary data models (McCullagh and Nelder,1989) can be used.

- In the normal step (M2 step), the MLE.s for the model parameters can be obtained using an appropriate optimization approach for incomplete data such as Newton- Raphson, Scoring method and Jennrich and Schluchter algorithm (Jennrich and Schluchter, 1986). Newton- Raphson method is used in this article. The obtained estimates are the average of the M imputed data sets, i.e.

$$\hat{\beta} = \tfrac{1}{M}\sum_{i=1}^{M} \hat{\beta}_i. \tag{11}$$

Standard errors

Louis (1982) introduces the following formula to approximate the information matrix:

$$I(\theta) = \mathrm{E}\left(-\frac{\partial^2 l(\theta|Y_{obs}, Y_{mis})}{\partial\theta\partial\theta}\middle|Y_{obs}\right) - \mathrm{cov}\left(\frac{\partial l(\theta|Y_{obs}, Y_{mis})}{\partial\theta}\middle|Y_{obs}\right)$$

$$= -E - C, \tag{12}$$

where θ is fixed at the stochastic EM estimates and $l(\theta|Y_{obs}, Y_{mis})$ is the log-likelihood function. Evaluating the integrals in this formula may not be easy. Efron (1994) suggests using simulation (the Monte Carlo method) to approximate the integrations. The main idea is to simulate *M* identically distributed samples, $q_1, q_2, ..., q_M$ from the conditional distribution of the missing values given the observed values and the parameters estimates, $f\left(Y_{mis}\middle|Y_{obs}, \hat{\theta}\right)$. Then, the formula in Eq. (12) can be approximated by its empirical version, i.e.

$$E \approx \frac{1}{M}\sum_{j=1}^{M}\frac{\partial^2 l\left(\theta\middle|Y_{obs}, q_j\right)}{\partial\theta\partial\theta} \tag{13}$$

and

$$C \approx \mathrm{cov}\left(\frac{\partial l\left(\theta\middle|Y_{obs}, q_j\right)}{\partial\theta}\right). \tag{14}$$

The Monte Carlo method is proposed and developed to obtain the standard errors of the SEM estimates in the current setting. We simulate $q_1, q_2, ..., q_M$ samples from the conditional distribution $f\left(Y_{mis}\middle|Y_{obs}, R; \hat{\theta}\right)$. Then, the information matrix in Eq. (12) can be approximated as

$$\mathrm{E} \approx \frac{1}{M}\sum_{j=1}^{M}\frac{\partial^2 l\left(\theta\middle|Y_{obs}, R, q_j\right)}{\partial\theta\partial\theta} \tag{15}$$

and

$$C \approx \mathrm{cov}\left(\frac{\partial l\left(\theta\middle|Y_{obs}, R, q_j\right)}{\partial\theta}\right), \tag{16}$$

where the parameters $\theta = (\beta, \alpha, \psi)$ is fixed at the SEM estimates; $\hat{\theta} = (\hat{\beta}, \hat{\alpha}, \hat{\psi})$.

Having the *M* pseudo-complete data, the first and second order derivatives of the log-likelihood function are evaluated for each sample. Then it is possible to calculate the quantities *E* and *C* and hence the information matrix. The inverse of the information matrix is the covariance matrix of the stochastic EM estimates. The standard error estimates are the square root of the main diagonal elements of this matrix.

4. Simulation Study

The aim of this simulation is to investigate the performance of the proposed approach. A complete longitudinal outcome Y_{ij}, for the subject *i* at the time point *j*, is generated from the following model $Y_{ij} = \beta_0 + \beta_1 X_{ij} + \varepsilon_{ij}$,, where *i=1,2,...,n* and *j=1,2,...,t.* The continuous cross-sectional covariates X_{ij} are generated from the standard normal distribution. These covariates are independent from the error terms ε_{ij}. The error terms ε_{ij} are assumed to follow a normal distribution with a mean of zero and $\sigma_e^2 = 0.5$. The number of subjects n (the sample size) is fixed at 25 and 50 subjects. The time points are restricted at $t = 5$. The parameters are fixed at $\beta_0 = 5$ and $\beta_1 = 10$. The simulation is replicated 2000 times. The missing values in the cross-sectional covariates are generated according to the logit model:

$$\text{logit}(X_i) = \eta_0 + \eta_1 X_{i-1}. \quad (17)$$

The parameters are fixed at $\eta_0 = -5$ and $\eta_1 = 0.06$. The missing values in the response are generated according to the model:

$$\text{logit}\left(r_{ij} = 1 \middle| \Psi\right) = \Psi_0 + \Psi_1 Y_{ij-1} + \Psi_2 Y_{ij}. \quad (18)$$

The parameters are assumed to be $\Psi = (\Psi_0, \Psi_1, \Psi_2) = (-17, 0.11, 0.13)$. All subjects are assumed to be observed at the first time point $j = 1$. The covariance structure, of the response, is assumed to be autoregressive of order 1, AR(1), with ρ=0.7 and σ =6.0.

We apply the proposed approach where multiple imputation to cross-sectional covariates with number of imputations *M=10*. The final parameter estimates are obtained as the average over the multiply imputed data sets, i.e.

$$\widehat{\boldsymbol{\beta}} = \frac{\mathbf{1}}{\mathbf{10}} \sum_{i=1}^{10} \widehat{\boldsymbol{\beta}}_i. \quad (19)$$

Table 1 and Table 2 present the results assuming the covariates are complete, and the sample size is 25 and 50, respectively. Table 3 and Table 4 present the results assuming that there are missing values in the covariates, and the sample size is 25 and 50, respectively.

Table 1. Parameter estimates (Est) and the relative bias (RB); sample size n= 25, complete covariates.

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.70	6	-17	0.11	0.13
Est.	5.13	9.61	0.64	5.8	-15.02	0.12	0.11
RB	0.03	0.04	0.08	0.03	0.12	0.09	0.15

Table 2. Parameter estimates (Est) and the relative bias (RB); sample size n= 50, complete covariates.

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.7	6	-17	0.11	0.13
Est.	5.34	9.67	0.63	5.81	-16.1	0.12	0.12
RB	0.07	0.03	0.10	0.03	0.05	0.09	0.08

Table 3.Parameter estimates (Est) and the relative bias (RB); sample size n= 25, MAR covariates.

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.7	6	-17	0.11	0.13
Predictive mean matching method							
Est.	4.54	10.09	0.65	5.75	-15.45	0.13	0.10
RB	0.09	0.01	0.07	0.04	0.09	0.18	0.23
Regression method							
Est.	4.80	9.76	0.62	5.85	-14.55	0.10	0.11
RB	0.04	0.02	0.11	0.02	0.14	0.09	0.15

Table 4. Parameter estimates (Est) and the relative bias (RB); sample size n= 50, MAR covariate.s

	β_0	β_1	ρ	σ	Ψ_0	Ψ_1	Ψ_2
True parameter	5	10	0.7	6	-17	0.11	0.13
Predictive mean matching method							
Est.	5.03	10.06	0.67	5.81	-16.01	0.11	0.12
RB	0.01	0.01	0.04	0.03	0.06	0.02	0.08
Regression method							
Est.	4.97	10.01	0.69	5.86	-15.55	0.09	0.10
RB	0.01	0.001	0.014	0.02	0.08	0.17	0.24

From Tables 1 and 2 we can see that the parameter estimates, for both sample sizes, perform well in terms of the relative bias. Tables 3 and 4 show similar results for both sample sizes. As a conclusion depending on this simulation results, the SEM estimates with multiple imputations to MAR in covariates using the regression method has the best performance in terms of relative bias. Hence, the performance of the proposed approach is acceptable even in relatively small sample sizes.

5. Application: Interstitial Cystitis Data Base (ICDB)

The Interstitial Cystitis Data Base (ICDB) have been used by Yang and Kang (2010). The ICDB characteristics are discussed in detail in Propert et al. (2000). The data include 637 patients at the baseline. Patients are followed for symptoms of pain, urgency, and urinary frequency, from January 1993 to November 1997. Yang and Kang (2010) study the joint effect of a group of covariates on the urgency and urinary frequency treating them as continuous and discrete variables, respectively. Each of these variables are measured by asking the patients to rate them in the last week on an ordinal scale ranging from 0; for the lowest severity, to 9 which is the maximum severity. In addition, the patients are required to rate the same variables in three consecutive days. The averages of the study variables over the three days are also recorded. Therefore, Yang and Kang (2010) consider only the data gathered in the first 36 months. The aim of the study is to explore the effect of continuous covariates on the continuous response (urgency).

There are missing values in the response variable (urgency). There are both dropout pattern and intermittent pattern. Because the proposed method deals with dropout pattern, so we omit the intermittently missing values. All continuous covariates are complete, and we generated missing values in the covariates. This is result in a reduced sample of 450 patients. A brief description of the covariates is given in the Table 5.

Table 5. The definition of the continuous covariates used in ICDB data.

Variable	Definition
Age	Patient age
UROD_7	Volume at first sensation
UROD_9	Volume at maximal capacity

Figure 1 presents a histogram of the continuous outcome plotted and compared with normal density function with mean of 4.25 and variance of 2.13 s a pre-analysis step we checked the adequacy of the model to fit the data.

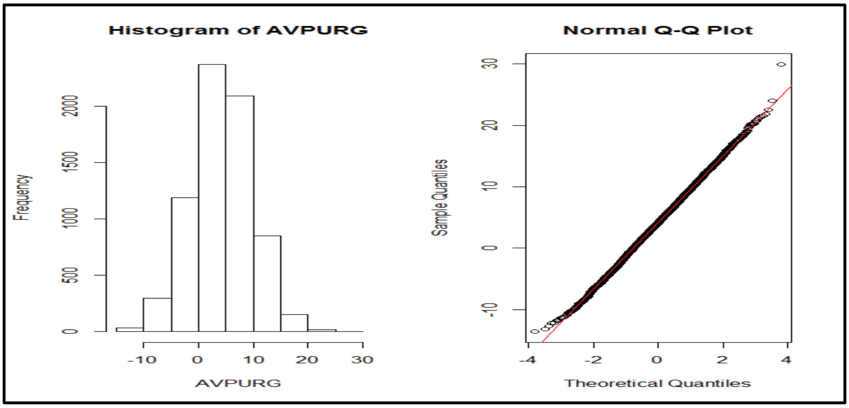

Figure 1. A histogram of the continuous outcome with normal and normal Q-Q plot.

We adopt the following model that allows each covariate to have its own effect, that is

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon_{ij}. \tag{20}$$

The AR(1) covariance structure is used, the elements of the covariance matrix, $\sigma_{ij=\sigma^2\rho^{|i-j|}}$. For the missing data mechanism, we use the linear logistic regression model. To keep the model simple only the previous and the current outcomes are included, that is

$$\operatorname{logit}(r_{ij} = 1 | \Psi) = \psi_0 + \psi_1 Y_{ij-1} + \psi_2 Y_{ij}. \tag{21}$$

The proposed approach is applied to the Interstitial Cystitis Data (ICDB). The results are given in the following Tables 6 and 7.

Table 6. The SEM estimates and standard errors (SE) for the urgency response without imputation to MAR covariates.

	Est.	SE	P-value
Intercept (β_0)	0.3367	0.07	< 0.000
UROD_7(β_1)	-0.001	0.16	< 0.000
UROD_9 (β_2)	-0.003	0.17	< 0.0640
Age (β_3)	0.05	0.081	< 0.004
ρ	0.12	0.079	< 0.000
σ	4.59	0.13	0.030
Ψ_0	-1.22	0.068	< 0.000
Ψ_1	0.13	0.079	< 0.000
Ψ_2	0.18	0.098	< 0.000

Diggle and Kenward (1994) noticed that in nonrandom models, dropout tends to depend on the difference between the current and previous measurements, $Y_{ij-1} - Y_{ij}$. Using this idea the estimated model for the missing data mechanism can be viewed as:

$$\begin{aligned} \mathrm{logit}(P) &= -1.22 + 0.13Y_{ij-1} + 0.18Y_{ij}, \\ &= -1.22 + 0.13(Y_{ij} - Y_{ij-1}) + 0.31Y_{ij}. \end{aligned} \tag{22}$$

From this model, the positive coefficient (0.13) of the difference between $\mathrm{Y_{ij}}$ and $\mathrm{Y_{ij-1}}$ also indicates that the response whose urgency increased are more likely to be missing

Table 7. The SEM estimates and standard errors (SE) for the urgency response with MAR missingness in covariate.

	Est.	SE	P-value
Predictive mean matching method			
Intercept (β_0)	0.445	0.029	< 0.000
UROD_7(β_1)	-0.005	0.01	< 0.000
UROD_9 (β_2)	-0.008	0.127	< 0.001
Age (β_3)	0.03	0.156	< 0.000
ρ	0.68	0.011	< 0.000
Ψ_0	-2.8	0.09	0.000
Ψ_1	0.07	0.07	< 0.000
Ψ_2	0.31	0.09	< 0.000
σ	5.42	0.12	< 0.000
Regression method			
Intercept (β_0)	0.544	0.023	< 0.000
UROD_7(β_1)	-0.041	0.002	< 0.000
UROD_9 (β_2)	-0.001	0.0001	< 0.000
Age (β_3)	0.022	0.001	< 0.000
ρ	0.033	0.002	< 0.000
Ψ_0	-1.023	0.022	0.000
Ψ_1	0.456	0.010	< 0.000
Ψ_2	0. 124	0.003	< 0.000
σ	4.98	0.098	< 0.000

Based in the results in Table 6 and Table 7, the positive values for the parameter Ψ_2 imply that high values of the urgency are more likely to be missing. We can conclude that the null-hypothesis that, $\Psi_2 = 0$ cannot accepted; this may be evidence for non-random dropout. Also Ψ_1 is significantly different from 0. This indicates the importance of the response at the previous time point. Also handling MAR in covariates improves results instead of ignoring them, as ignoring the MAR in covariates at Table 6, it was an insignificant effect of UROD_9 on urgency, after handling missingness with two MI methods the effect of UROD_9 on urgency become significant as in Table 7.

6. Conclusion and Future Work

Most literature, in longitudinal studies with missing values, focus on missingness in the longitudinal response or in the covariates. However, in practice it is possible to have missingness in the longitudinal response and in the covariates at the same time. This article proposes two methods to deal with missingness in both the longitudinal response and covariates at the same time. In this paper we proposed a selection model (Diggle and Kenward, 1994) for longitudinal data with non-ignorable missing values of the response with multiple imputation to missingness on covariates. The proposed model covers the case of the dropout missingness. The obtained likelihood function is intractable and not easy to be maximized. To overcome this difficulty, we suggest using the Stochastic EM algorithm. The proposed approach is applied to a data set from (The Interstitial Cystitis Data Base (ICDB)). The approach can be easily implemented in many fields where the missingness process is suspected to be non-ignorable. The case of intermittent pattern, which has less attention in the literature compared to the dropout, is a very challenging topic for future work.

Acknowledgments

The authors would like to thank Professor Kathleen Joy Propert at University of Pennsylvania, School of Medicine for providing guidelines about how to obtain the ICDB data. The ICDB data reported here were supplied by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) Central Repositories.

Literature Cited

Abdelwahab, H. A., El Kholy, R. B. and Gad, A. M. (2019) Sensitivity analysis index for shared Parameter models in longitudinal studies. *Advances and Applications in Statistics*, 57, 1-20.

Allassonnière, S. and Chevallier, J. (2021) A new class of stochastic EM algorithms. escaping local maxima and handling intractable sampling. *Computational Statistics and Data Analysis*, 159, 107159

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.

Darwish, N. M., Gad, A. M. and Hamid, R. M. (2020) Fitting longitudinal data with missing values in the response and covariates, *Advances and Applications in Statistics,* 64(2), 127-142.

Diebolt, J. and Ip, E. H. S. (1996) Stochastic EM: method and application. In: *Markov chain Monte carlo in practice.* (eds W.R. Gilks, S. Richardson and D. J. Spiegelhalter). Chapman and Hall, London. Chapter 15, pp. 259-273.

Diggle, P. and Kenward, M. G. (1994) Informative drop-out in longitudinal data analysis, *Journal of the Royal Statistical Society C*, 43, 49 – 93.

Efron, B. (1994) Missing data, imputation, and the bootstrap, *Journal of American Statistical Association*, 89, 463–475.

Erler, N. S., Rizopoulos, D., Rasmalen, V., Jaddoe, V. W., Franco, O. H., and Leasffre, E. M. H. (2016) Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and full Bayesian approach, *Statistics in .Medicine,* 35, 2955-2974.

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

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practices*, Chapman and Hall, London.

Grannell, A. and Murphy, H. (2011) Using multiple imputation to adjust for survey nonresponse, *Shifting the Boundaries of Research Proceedings*: 123-135. University of Bristol, UK.

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

Little, R. J. A. (1995) Modelling the dropout mechanism in repeated measures studies, *Journal of American Statistical Association*, 90, 1112-1121

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society,* B 44, 226–232.

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

Nooraee, N., Molenberghs, G., Ormel, J. and Heuve, E. R. V. D. (2018) Strategies for handling missing data in longitudinal studies with questionnaires, *Journal of Statistical Computation and Simulation*, 88 (17), 3415–3436.

Propert, K. J., Schaeffer, A. J. Brensinger, C. M. Kusek, J. W. Nyberg, L. M. and Landis, J. R. (2000) A prospective study of interstitial cystitis: results of longitudinal followup of the interstitial cystitis data base cohort, *The Journal of Urology*, 163, 1434-1439.

Rubin, D. B. (1976) Inference and missing data, *Biometrika* 63, 581-592.

__________(1987) *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons Inc., New York.

Salfran, D. and Spiess, M. (2015) *A Comparison of Multiple Imputation Techniques*. Discussion Paper No. 3, Universtat Hamburg, Germany.

Yang, Y., and Kang, J. (2010) Joint analysis of mixed Poisson and continuous longitudinal data with nonignorable missing values. *Computational Statistics & Data Analysis,* 54 (1), 193–207.

Yassen, A. S. and Gad, A. M. (2020) A stochastic variant of the EM algorithm to fit mixed (discrete and continuous) longitudinal data with nonignorable missingness, *Communication in Statistics - Theory and Methods*, 49(18), 4446-4467.