Journal of Data Science **9**(2011), 171-180

A Selection Model for Longitudinal Data with Non-Ignorable Non-Monotone Missing Values

Ahmed M. Gad
Cairo University

Abstract: Missing values are not uncommon in longitudinal data studies. Missingness could be due to withdrawal from the study (dropout) or intermittent. The missing data mechanism is termed non-ignorable if the probability of missingness depends on the unobserved (missing) observations. This paper presents a model for continuous longitudinal data with non-ignorable non-monotone missing values. Two separate models, for the response and missingness, are assumed. The response is modeled as multivariate normal whereas the binomial model for missingness process. Parameters in the adopted model are estimated using the stochastic EM algorithm. The proposed model (approach) is then applied to an example from the International Breast Cancer Study Group.

Key words: Intermittent missing, informative missing, selection models, the stochastic EM algorithm.

1. Introduction

Longitudinal data consist of time sequence of measurements on several subjects. Longitudinal data frequently involve some missing values. Subjects may withdraw from the study prematurely resulting in a dropout pattern or they may missed some occasions; the intermittent pattern. Little and Rubin (1987) introduce different mechanisms for missing values. The missing values is termed non-ignorable if the probability of being missing depends on the unobserved measurements. In this case a model is needed for both the observed and missing data for unbiased inference.

Many models have been proposed that link the response and missingness in some way. Shared parameter models that relate the response with the probability of missingness introduced by Wu and Carroll (1988). The selection models (Heckman, 1979) relate the probit model for missigness and a normal error regression model for the response. In selection framework Diggle and Kenward

(1994) propose a model that combine logistic model for the dropout process and normal linear model for the response. They formulate the log-likelihood and the parameter estimates are obtained by using Simplex method. This model is in dropout pattern. Albert and Follmann (2003) propose a shared parameter model for longitudinal binary data with informative missingness. The missing values are due to dropout or intermittent missing. Gad and Ahmed (2006) develop the stochastic EM algorithm to handle longitudinal data with intermittent missing data. Also, Gad and Ahmed (2007) apply the same procedure in sensitivity analysis context.

The aim of this paper is to propose a selection model for longitudinal data with dropout and intermittent missing values. So far as we are aware, this is the first paper to propose a model for continuous longitudinal data with non-ignorable intermittent and dropout missing values. This paper deals with longitudinal data with intermittent and dropout missing data, whereas the previous papers (Gad and Ahmed, 2006; Gad and Ahmed, 2007) concern with only intermittent missing values. In section 2 the proposed model is presented. The estimation procedure is described in Section 3. Section 4 contains analysis of a real data, and a discussion is given in Section 5.

2. Model and Notation

Let $Y_i = (Y_{i1}, \cdots, Y_{in_i})'$ represent the sequence of intended measurements on the ith subject and $Y = (Y_1', \cdots, Y_m')'$ the entire set of measurements on the m subjects. We assume that Y_i follows a multivariate normal distribution,

$$Y_i \sim \mathrm{MVN}(\mathrm{X_i}\beta, \mathrm{V_i}(\alpha)), \tag{1}$$

where X_i is an $n_i \times p$ matrix of explanatory variables (or design matrix) and β is a $p-$vector of unknown parameters (mean parameters). The variance-covariance matrix $V_i(\alpha)$, of order $n_i \times n_i$, is a function of a vector of unknown parameters α of length q. The explicit parameterization of the variance-covariance matrix allow us to assume different models for the covariance structure, ranging from a very simple model to the unstructured model.

Due to the missing data not all the intended measurements are available. Assume that the observed measurements for the ith subject are stacked in a vector Y_{iobs} of length n_{iobs}. Also, the missing values are stacked in a vector Y_{imis}. Assume that this vector is of dimension $1 \times (r+1)$, where the first r elements include those who are intermittent $(Y_{i,mis1}, \cdots, Y_{i,misr})$ and the last element is the dropout Y_{i,d_i}. Assume that H_{ij} represents the part of Y_{iobs} preceding a missing value (the history). Note that H_{ij} is the same as Y_{iobs} when the jth observation

is dropout. Let R_{ij} be a missing value indicator that takes three values as:

$$R_{ij} = \begin{cases} 0, & \text{if } Y_{ij} \text{ observed,} \\ 1, & \text{if } Y_{ij} \text{ intermittent missing,} \\ 2, & \text{if } Y_{ij} \text{ dropout.} \end{cases} \tag{2}$$

The missing data mechanism is assumed to be, conditionally, depends on the history of measurements up to and including the jth observation , i.e.,

$$\mathrm{P}(R_{ij} = r|H_{ij}) = P_j(H_{ij}, Y_{ij}; \psi), \tag{3}$$

where ψ is a vector of unknown parameters.

Suppress the dependence on i for simplicity, let

$$\eta_{j1} = \psi_{01} + \psi_{11} y + \sum_{k=2}^{j} \psi_{k1} Y_{j+1-k}$$

and

$$\eta_{j2} = \psi_{02} + \psi_{12} y + \sum_{k=2}^{j} \psi_{k2} Y_{j+1-k}.$$

The missing data mechanism is modeled as a multinomial regression with three states (Albert and Follmann, 2003) as

$$\mathrm{P}(R_{ij} = r|H_{ij}, R_{ij-1} \neq 2; \psi) = \begin{cases} \frac{1}{1+\sum_{r=1}^{2} \exp(\eta_{jr})}, & r = 0, \\ \\ \frac{\exp(\eta_{jr})}{1+\sum_{r=1}^{2} \exp(\eta_{jr})}, & r = 1, 2. \end{cases} \tag{4}$$

The parameters ψ_{11} and ψ_{12} that relate the intermittent missing and dropout, respectively, with the response process. The missing data mechanism is non-ignorable when these two parameters take non-zero values. Also, it is assumed that $R_{i0} = 0$. Note that R_{i2} is an absorbing state. Thus,

$$P(R_{ij} = 2|R_{ij-1} = 2) = 0.$$

The joint distribution of Y_i can be written as

$$f(Y_{iobs}) \left[\prod_{j=1}^{n_i} \{1 - P_{1ij} - P_{2ij}\}^{I(R_{ij}=0)} P_{1ij}^{I(R_{ij}=1)} P_{2ij}^{I(R_{ij}=2)} \right],$$

where n_i is the last observation or the observation prior to dropout and $I(R_{ij} = r)$ are indicator functions which have the value 1 when the condition is met.

The log-likelihood for the observed data, $\log f(Y_{iobs})$ can be written as

$$\log f(Y_{iobs}) = -\frac{n_{iobs}}{2}\log(2\pi) - \frac{1}{2}\log|V_{iobs}| - \frac{1}{2}(Y_{iobs} - \mu_{iobs})' V_{iobs}^{-1}(Y_{iobs} - \mu_{iobs}),$$

where μ_{iobs} represent the relevant elements of $X_i\beta$.

The log-likelihood function of the i subject is given as:

$$\ell_i(\mu, \alpha, \psi) = \log f(Y_{iobs}) + \sum_{j=1}^{n_i} \log\{1 - P_{1ij} - P_{2ij}\}^{I(R_{ij}=0)} P_{1ij}^{I(R_{ij}=1)} P_{2ij}^{I(R_{ij}=2)}$$

Accordingly, the log-likelihood for the m subjects is

$$\begin{aligned}\ell(\mu, \alpha, \psi) &= \sum_{i=1}^{m} \ell_i(\mu, \alpha, \psi) \\ &= \sum_{i=1}^{m} \left[\log f(Y_{iobs}) + \sum_{j=1}^{n_i} \log\{1 - P_{1ij} - P_{2ij}\}^{I(R_{ij}=0)} P_{1ij}^{I(R_{ij}=1)} P_{2ij}^{I(R_{ij}=2)} \right]. \quad (5)\end{aligned}$$

3. Estimation

Clearly, the function in equation (5) involves a very high-dimensional integration and does not have a closed form in general. Therefore, maximizing the observed data likelihood directly is not at all feasible. When some components of Y are nonignorably missing, the estimation problem based on the observed data likelihood becomes much more complicated. Thus, to make the estimation problem feasible, we develop the stochastic EM algorithm (Diebolt and Ip, 1996, chap. 15) that facilitates estimation of the parameters.

The stochastic EM algorithm (SEM) was proposed by Diebolt and Ip (1996, chap. 15) as a stochastic version of the EM algorithm. The SEM algorithm consists of iterating two steps: the S-step and the M-step. In the S-step, the missing values 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 data is maximized using any standard maximization procedure. These two steps are iterated for a sufficient number of iterations.

The developed approach includes two steps: the S-step and the M-step. In the S-step, a single draw is obtained from the conditional distribution of the missing data, Y_{imis}, given the observed data, (Y_{iobs}, R_i). The Gibbs sampling algorithm, see for example Gelfand (2000), is adopted in this paper to carry out the simulation step. At the $(t+1)$th iteration $Y_{i,mis}^{(t+1)} = (Y_{i,mis1}^{(t+1)}, \cdots, Y_{i,misr}^{(t+1)}, Y_{i,d_i}^{(t+1)})$ is simulated from the full conditional distributions. This iteration is executed in

$(r+1)$ sub-steps. First, $Y_{i,mis1}^{(t+1)}$ is simulated from the conditional distribution $f(Y_{i.mis1}|Y_{i,mis2}^{(t)},\cdots,Y_{i,misr}^{(t)},Y_{i,d_i}^{(t)},Y_{i,obs},R_i,\theta^{(t)})$. Then, in the second sub-step, $Y_{i,mis2}^{(t+1)}$ is simulated from the conditional distribution $f(Y_{i.mis2}|Y_{i,mis1}^{(t+1)},\cdots,Y_{i,misr}^{(t)}, Y_{i,d_i}^{(t)},Y_{i,obs},R_i,\theta^{(t)})$. In the third sub-step, $Y_{i,mis3}^{(t+1)}$ is simulated from the distribution $f(Y_{i.mis3}|Y_{i,mis1}^{(t+1)},Y_{i,mis2}^{(t+1)},\cdots,Y_{i,misr}^{(t)},Y_{i,d_i}^{(t)},Y_{i,obs},R_i,\theta^{(t)})$. In the last sub-step, the last missing value $Y_{i,d_i}^{(t+1)}$ is simulated from the conditional distribution $f(Y_{i,d_i}\mid Y_{i,mis1}^{(t+1)},\cdots,Y_{i,misr}^{(t+1)},Y_{i,obs},R_i,\theta^{(t)})$.

This simulation is not possible because the full conditional distribution has no standard distribution. To overcome this problem we suggest an accept-reject procedure as follow.

1. Generate a candidate value y^* from the conditional distribution $f(Y_{i,misj}|$ $Y_{i,obs},Y_{i,mis1}^{(t+1)},\cdots,Y_{i,misj-1}^{(t+1)},Y_{i,misj+1}^{(t)},\cdots,Y_{i,misr}^{(t)},Y_{i,d_i}^{(t)},\theta^{(t)})$ for $j=1,2,\cdots,$ $r+1$. This distribution is normal, hence the direct simulation is possible using any available software.

2. Compute the probability of missingness for the candidate value, y^*, according to the assumed model in Eq. (4) assuming that ψ is fixed at $\psi^{(t)}$. Assume that this probability is P_i.

3. Simulate a single value from the uniform distribution $[0,1]$, U, and take $Y_{i,misj}=y^*$ if $U\leq P_i$; otherwise go to step 1.

The M-step consists of two sub-steps: the multinomial step and the normal step. In the multinomial step, the missingness parameters are obtained using any iterative procedure, see for example McCullagh and Nelder (1989). In the normal step, the EM scoring algorithm (Jennrich and Schluchter, 1986) is used to obtain the model parameters.

4. Application: Breast Cancer Data

The proposed approaches are applied to the breast cancer data. This data concerning quality of life among breast cancer patients in a clinical trial taken by the International Breast Cancer Study Group (**IBCSG**). In the **IBCSG** trial $\mathbb{VII}$ (Hürny *et al.*, 1992), premenopausal women with breast cancer are followed for traditional outcomes such as relapse, death and also focused on quality of life. Patients were randomized to four different chemotherapy regimes denoted by A, B, C and D. It is intended to compare quality of life among the four different treatments. The patients were asked to complete quality of life questionnaires at baseline (before starting treatment) and at three months intervals for fifteen months. Hence, each questionnaire should be filled out six times. Essentially,

these six time points cover the time during the administration of chemotherapy across all the four treatments.

One of the relevant determinants of quality of life was the Perceived Adjustment to Chronic Illness Scale (**PACIS**). This is one-item scale comprising a global patient rating of the amount of effort costs to cope with illness. In this trial the **PACIS** assessments for patients who remained alive during the 15 months of the study are analyzed. Ten patients who die within the study period are excluded from the analysis, so the missing responses is not due to death. The total number of patients survive the study period is 446 patients. The patients with missing response at the first assessment (64 cases) are excluded from the analysis, leaving 382 patients. Compliance was not compulsory and patients did refuse, on occasion, to complete the assessment. Even when they refused, the patients were asked to complete an assessment at their next scheduled follow-up visit. Thus, the structure of this trial result in hybrid pattern of missing data (intermittent pattern and dropout pattern). A patient may not appear to fill the questionnaire if her mood is poor, and therefore the missing data mechanism is nonrandom.

The **PACIS** values were missing for 77% of the patients for at least one occasion, so the study completers are 89 (23%) patient. Out of the patients with missing data there are 184 (63%) patient as dropout pattern and 109 (37%) as intermittent pattern. The **PACIS** measured on a continuous scale from 0 to 100 where a larger score indicates that a greater amount of effort is required for the patient to cope with her illness. Following Hürny *et al.* (1992), we use a square-root transformation to normalize the data. The averages of the assessments using all available transformed data are 6.1, 5.7, 5.6, 5.1, 4.7, 5.1 respectively and the standard deviations are 2.50, 2.46, 2.49, 2.51, 2.51, 2.51.

A preliminary version of this data, the responses for the first 9 months of the study, were analyzed by Hürny *et al.* (1992). Only patients with complete responses are included in the analysis (complete cases analysis). Another preliminary analysis of this data has been conducted by Troxel *et al.* (1998). The missing data have been taken into account in this analysis and the analysis is based on the responses for the first 6 months of the study. Ibrahim *et al.* (2001) have analyzed the patient's self of her mood variable for the 18 months of the study. They used random effects model with AR(1) model for the covariance structure. The missing data mechanism is modeled using the logistic model that includes the previous and the current responses.

In this article the PACIS response variable is of main interest. We adopt a mean model that allow each treatment to have its own effect. That is:

$$\mu_j = \mu_{0j} + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 \qquad \text{for} \quad j = 1, \cdots, 6,$$

where μ_{0j} is a constant shift at each assessment time and

$$(x_1, x_2, x_3) = \begin{cases} (1,0,0), & \text{for treatment A,} \\ (0,1,0), & \text{for treatment B,} \\ (0,0,1), & \text{for treatment C,} \\ (0,0,0), & \text{for treatment D.} \end{cases}$$

Previous analyses of these data suggested using the first order auto-regressive AR(1) model rather than any other covariance structures. In this model, the (i,j)th element of the covariance matrix, σ_{ij} equal to $\sigma^2\rho^{|i-j|}$ for $i,j = 1,\cdots,6$. In this paper the AR(1) covariance structure and unstructured covariance matrix are used. Different covariance structure are possible and have been tried but the results for the above two structures only are presented. For the missing data mechanism, we use the model in Eq. (4). To keep the model simple only the previous and the current outcomes are included, that is:

$$\eta_{j1} = \psi_{01} + \psi_{11}y + \psi_{21}Y_{j-1}$$

and

$$\eta_{j2} = \psi_{02} + \psi_{12}y + \psi_{22}Y_{j-1},$$

for $j = 1, 2, \cdots, 6$. The developed SEM algorithm is applied to obtain parameter estimates of this model. The iterations number n is set at 5000 iterations with a burn-in period of 2000 iterations. We assume that $\psi_{01} = \psi_{01} = \psi_0$ and $\psi_{21} = \psi_{22} = \psi_2$ for parsimony. Hence, there are 4 parameters to be estimated for the missingness process. The SEM estimates of the mean, covariance and missingness parameters are displayed in Table 1. The standard errors (SE) of the SEM estimates have been obtained using the proposed simulation method. Results are, also, shown in in Table 1.

Another model of the covariance structure, the unstructured model, has been used. In this structure there are 21 variance/covariance parameters for the six time points. The SEM estimates of mean and missingness parameters for this model are also given in Table . The SEM estimate of the covariance matrix, $\tilde{V}$, is as:

$$\tilde{V} = \begin{pmatrix} 6.18 & 1.52 & 1.81 & 1.78 & 1.54 & 0.87 \\ & 4.65 & 1.63 & 1.24 & 1.42 & 0.43 \\ & & 4.32 & 1.95 & 1.48 & 1.23 \\ & & & 3.78 & 1.67 & 1.24 \\ & & & & 3.59 & 0.94 \\ & & & & & 2.86 \end{pmatrix}.$$

The -2log-likelihood value for the unstructured covariance model is 5399 with 34 parameters and for the first order auto-regressive model is 5899 with 15 parameters. The -2log-likelihood difference between the two models is 500 on 19 degrees of freedom. Hence there is an evidence for AR(1) model.

Table 1: The SEM estimates and Standard errors (SE) for the PACIS response

AR(1) Covariance Model				Unstructured Covariance Model			
Parameter	Est. (SE)	Parameter	Est. (SE)	Parameter	Est.	Parameter	Est.
μ_{01}	6.17 (0.16)	α_1	-0.22 (0.15)	μ_{01}	6.18	α_1	0.18
μ_{02}	5.88 (0.14)	α_2	0.05 (0.15)	μ_{02}	5.60	α_2	0.13
μ_{03}	5.87 (0.16)	α_3	- 0.62 (0.16)	μ_{03}	5.92	α_3	-0.96
μ_{04}	6.15 (0.16)	ψ_0	1.12 (0.07)	μ_{04}	5.16	ψ_0	3.19
μ_{05}	5.23 (0.14)	ψ_{11}	1.59 (0.06)	μ_{05}	5.55	ψ_{11}	1.94
μ_{06}	5.16 (0.14)	ψ_{12}	0.84 (0.08)	μ_{06}	5.12	ψ_{12}	1.04
ρ	0.52 (0.03)	ψ_2	1.02 (0.11)			ψ_2	1.55
σ^2	4.19 (0.12)						

The positive values for the parameters ψ_{11} and ψ_{12} imply that high values of the PACIS are more likely to be missing. This is natural because high values of PACIS indicate that more difficulty in coping with the disease. Hence we would expect that a woman costs great amount of effort to cope with her illness is more likely to refuse to complete quality of life questionnaire. The ψ_{11} and ψ_{12} is significantly different from 0, supporting that the missing data mechanism is nonrandom. Also ψ_2 is significantly different from 0. This indicates the importance of the response at the previous time point. The Z-values for testing both null hypotheses are significant at any reasonable degree of confidence.

The covariates treatment C is significant at any reasonable significance level.

5. Conclusion

In this paper we proposed a selection model (Diggle and Kenward, 1994) for longitudinal data with non-ignorable missing values. The proposed model cover the case of the intermittent and 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.

In the context of the proposed model, direct simulation is not possible because there is no formula for the density function of the missing data given the observed data. Hence, a reject-accept sampling procedure is proposed and incorporated in

the simulation step of the Stochastic EM algorithm.

The proposed algorithm is applied to a data set from breast cancer field. The approach can be easily implemented in many fields where the missingness process is suspected to be non-ignorable.

Acknowledgements

The author is grateful to the Quality of Life Committee of the IBCSG for permission to use the quality of life data. Also, the author is grateful to the Editor and an anonymous referee for their comments on the manuscript.

References

Albert, P. S. and Follmann, D. A. (2003). A random effects transition model for longitudinal binary data with informative missingness. *Statistica Neerlandica* **57**, 100–111.

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.

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

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**, 2701–2714.

Gad, A. M. and Ahmed, A. S. (2007). Sensitivity analysis of longitudinal data with intermittent missing values. *Statistical Methodology* **4**, 217–226.

Gelfand, A. E. (2000). Gibbs sampling. *Journal of the American Statistical Association* **50**, 1300–1304.

Hurny, C., Bernard, J. G., Gelber, R. D., Coates, A., Gastiglione, M., Isley, M., Dreher, D., Peterson, H., Goldhirsch, A., and Senn, H. J. (1992). Quality of life measures for patients receiving adjutant therapy for breast cancer: an international trial. *European Journal of Cancer* **28**, 118-124.

Heckman, J. J. (1979). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economics and Social Measurement* **5**, 475-492.

Ibrahim, J. G., Chen, M. and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika* **88**, 551-564.

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

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2 edn. Chapman and Hall, London.

Troxel, A. B., Harrington, D. P. and Lipsitz, S. R. (1998). Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics* **47**, 425–438.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modelling the censoring process. *Biometrics* **44**, 175-188.

Received August 24, 2009; accepted January 18, 2010.

Ahmed M. Gad
Statistics Department
Faculty of Economics and Political Science
Cairo University, Cairo, Egypt
hagas10@hotmail.com