Longitudinal Data with Intermittent Missing Values: A Sensitivity Analysis Approach

Ahmed M. Gad*

Statistics Department, Faculty of Economics & Political Science, Cairo University, Egypt.

and

Abeer S. Ahmed

The National Centre for Social and Criminological Research, Cairo, Egypt.

Abstract

Intermittent missing data are not uncommon in longitudinal data studies. In selection models, the probability of being missing for any observation is modeled as a function of the current observation and the previous observations. The parameter that relates the probability of missingness and the current observation has special interpretation. The degree of informativeness of the missing data process depends on this parameter's value. We conduct sensitivity analysis to evaluate the effect of this parameter value (the sensitivity parameter) on study results. In the proposed approach, the sensitivity parameter is assumed to be fixed at a set of plausible values. This allows us to examine several degrees of informativeness of the missing data process. The stochastic EM algorithm is used to obtain parameter estimates. The proposed method is evaluated via a simulation study and then applied to a real data set. Sensitivity analysis shows that the conclusion depends on the degree of informativeness. Hence, when estimating the sensitivity parameter the results should be interpreted cautiously.

Keywords: Diggle-Kenward model, Informative missing, Intermittent missing, Selection models,The stochastic EM algorithm.

*Correspondence Address: *Statistics Department, Faculty of Economics & Political Science, Cairo University, Egypt. Email: dr_ahmedgad@yahoo.co.uk*

1 Introduction

In longitudinal studies each subject is measured on several occasions. Missing observations are not uncommon in longitudinal studies. In dropout pattern (monotone missing) a missing observation is never followed by an observed value, whereas in intermittent pattern (non-monotone missing) a missing value may followed by an observed value. Less attention has been paid to intermittent missing values in literature. The focus of this article is on intermittent missing values, where responses are available for a subject even after a missing response.

Little and Rubin (1987, Chapter 6) and Little (1993) have introduced the terminology of the missing data process. A missing data process is said to be missing completely at random (MCAR) if the probability of missingness is independent of both observed and unobserved data and missing at random (MAR) if, conditional on the observed data, the probability of missingness is independent of the unobserved data. A process that is neither MCAR nor MAR is said to be missing not at random or informative (MNAR).

Many approaches that deal with missing data are formulated as selection models (Heckman, 1976). In selection models the joint density function of the response and a missing data indicator is factorized into a distribution of the response conditional on the missing data indicator and a marginal distribution of the missing data indicator. Diggle and Kenward (1994) have proposed a selection model for longitudinal data with informative dropout. The probability of dropout is assumed to depend on the unobserved measurement and the measurement history. They use the normal model for the responses and the logistic regression model for the dropout process. This approach has been generalized to the intermittent setting by Troxel *et al.* (1998) and Gad and Ahmed (2006).

It has been noted by many discussants to Diggle and Kenward (1994) that study conclusions of this model relies on assumptions which cannot be verified from the observed data. So, sensitivity analysis of study conclusions to such assumptions is needed. Sensitivity analysis is a set of tools showing the influence of the model assumptions on the study conclusion. Several sensitivity analysis tools have been proposed in dropout setting; see, for example, Daniels and Hogan (2000), Molenberghs *et al.* (2003), Minini and Chavance (2004) and Verbeke *et al.* (2001). However, in the intermittent setting few work concerning sensitivity analysis are available and many research need to be done.

In selection models context, the missingness model is a key assumption for conducting sensitivity analysis. Ibrahim *et al.* (2001) propose the Monte Carlo EM algorithm for estimating parameters in the generalized lin-

ear mixed model with non-random non-monotone missing data. Several sensitivity analyses were conducted by fitting several models for the missing data mechanism based on various covariates. They concluded that parameter estimates are quite robust with respect to changes in the missingness model. Minini and Chavance (2004) suggest using a shared parameter (sensitivity parameter) that relates the response variable and the dropout process. A range of different values of this parameter is considered, which allow us to assess the sensitivity of study conclusions to the dropout mechanism.

Celeux and Diebolt (1985) have introduced the stochastic EM algorithm as an alternative to the EM algorithm. The stochastic EM algorithm can be used when the E-step of the EM algorithm is intractable. The stochastic EM algorithm involves iterating two steps. 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 likelihood function of the pseudo complete data is maximized using any conventional procedure. For more details on the stochastic EM algorithm; see, for example, Diebolt and Ip (1996).

The purpose of this paper is to conduct sensitivity analysis of study conclusion, in the intermittent setting, using different assumptions of the missingness process. The proposed approach is an extension to the Minini-Chavance's approach (Minini and Chavance, 2004) to the intermittent setting. A shared (sensitivity) parameter relating the response and the missingness process is introduced, so different degrees of informativeness can be considered. The rest of the paper is outlined as follows. In Section 2 the basic notation are described. In Section 3 the proposed method is described. In Section 4 the proposed approach is applied to a data set concerning quality of life among breast cancer patients in a clinical trial undertaken by the International Breast Cancer Study Group. The final section is devoted to concluding remarks.

2 Notation

Assume that m subjects are participating in the study. Assume that for the ith subject, $i = 1, ..., m$, a sequence of responses Y_{ij} is planned to be measured at times $j = 1, ..., n$. The responses of the ith subject are gathered into a vector Y_i, $Y_i = (Y_{i1}, ..., Y_{in})'$. The vector Y_i is split into two sub-vectors Y_i^{mis} and Y_i^{obs}, where Y_i^{mis} contains the missing components and Y_i^{obs} contains the observed components. Also, assume that R_{ij} is a missing value indicator that takes the value of one if Y_{ij} is observed and the value of zero if Y_{ij} is missing. Let R_{ij} are grouped into a vector $R_i = (R_{i1}, ..., R_{in})'$. Assume that

Y_i satisfies the linear regression model

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

where X_i is a known $n \times p$ matrix of explanatory variables, β is a $p \times 1$ vector of fixed effect parameters, and V_i is an $n \times n$ positive definite covariance matrix. The matrix V_i can be unstructured with $n(n+1)/2$ parameters. Also, the covariance matrix V_i can be structured, i.e. its elements are functions of vector of parameters α, and can be written as $V_i(\alpha)$. The main reason for modeling the covariance matrix as a function of parameters α is to examine different covariance structures, and for parsimony. The parameters β and α are grouped in a vector of parameters $\theta = (\beta', \alpha')'$.

In selection models the density of the complete data, Y_i and R_i, is factorized into two components as

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

where the parameter vectors θ and ψ describes the measurement and missingness processes respectively. Following Diggle-Kenwards' model (Diggle and Kenward, 1994) the missingness process is modeled as a function of the current response and the measurement history, i.e. $P(R_{ij} = r_{ij}|\text{history}) = \text{P}_{\text{ij}}(\text{Y}_{\text{ij}}, \text{H}_{\text{ij}}; \psi)$. The marginal distribution of the response $f(Y_i|\theta)$ is assumed to be the normal distribution. The logistic model is used to model the missingness process, assuming that the dropout time is d_i, as

$$\text{logit}\left\{P_{di}(H_{d_i}, Y_{di}; \psi)\right\} = \psi_0 + \sum_{j=1}^{d_i} \psi_{di-j+1} Y_{di-j+1}. \quad (1)$$

For simplicity we assume that the probability of missingness depends only on the current and the previous responses as

$$\text{logit}\{\text{P}_{\text{ij}}(\text{Y}_{\text{ij}}, \text{H}_{\text{ij}}; \psi)\} = \psi_0 + \psi_1 \text{Y}_{\text{i,j-1}} + \psi_2 \text{Y}_{\text{ij}}. \quad (2)$$

In this model, if the $\psi_2 = 0$ the missingness process is MAR whereas the situation is MNAR if ψ_2 is different from 0. In the later case inference concerning the measurement process cannot be performed independently from inference concerning the missingness process. The sensitivity parameter ψ_2 is of main interest in the proposed approach. Finally, let Ω denote the parameter vector contains all parameters except ψ_2, i.e. $\Omega = (\theta', \psi_0, \psi_1)'$.

3 The proposed approach

The parameter ψ_2 is fixed at a range of plausible values, ψ_2^f. For each value ψ_2^f, the stochastic EM algorithm is used to obtain parameter estimates. Gad and Ahmed (2006) develop the stochastic EM algorithm to

handle longitudinal data with informative intermittent missing values. At the S-step, the missing data Y_i^{mis} are simulated from the conditional distribution $f(Y_i^{mis}|Y_i^{obs}, R_i)$ at the current parameter estimate $\Omega^{(t)}$ and ψ_2^f. Direct simulation is not possible from this distribution. So, the accept-reject procedure proposed in Gad and Ahmed (2006) can be used. This imputation provides us with a plausible pseudo complete data set. At the M-step, any maximization procedure for the complete data can be used to update parameter estimates. In this paper we use the Jennrich-Schluchter algorithm (Jennrich and Schluchter, 1986).

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 SEM estimate, $\tilde{\Omega}$, is the mean of points generated by the stochastic EM algorithm ignoring the early first points as a burn-in period.

The stochastic EM algorithm does not provide the estimate standard errors. 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. Gad and Ahmed (2006) have developed the Monte Carlo method for longitudinal data with informative intermittent missing. This method is used in this paper to obtain the estimates standard errors.

4 Simulation study

A simulation study is conducted to evaluate the proposed method. It is based on a data set with m subjects and n observations for each subject. We adopt the simple model $E(Y_{ij}) = \mu_j$, where $i = 1, ..., m$ and $j = 1, ..., n$. A stationary first-order auto-regressive, AR(1), process is used to generate the residual component of the repeated measurement, i.e. $V(\epsilon_{ij}) = \sigma^2$ and $Cov(\epsilon_{ij}, \epsilon_{ik}) = \sigma^2\rho^{|j-k|}$. The missingness model is assumed as in Eq. (2). The data are simulated to satisfy the multivariate normal distribution, the AR(1) covariance structure and the missingness model (2), with number of subjects $m = 100$ and time points $n = 5$. According to this setting the parameters vector is $(\mu_1, \mu_2, \mu_3, \mu_4, \mu_5, \sigma^2, \rho, \psi_0, \psi_1, \psi_2)'$. These parameters are set to the values $\mu_1 = 6$, $\mu_2 = 6$, $\mu_3 = 6$, $\mu_4 = 5$, $\mu_5 = 5$, $\sigma^2 = 4$, $\rho = 0.5$,

Table 1: Simulation results: the relative bias (RB) of parameter estimates and 95% confidence coverage (CP)

	ψ_2^f									
	-2		-1		0		1		2	
Par.	RB	CP	RB	CP	RB	CP	RB	CP	RB	CP
μ_1	-0.2	95	-0.2	95	-0.2	95	-0.2	95	-0.2	95
μ_2	-75.3	12	-68.3	20	-14.4	88	-13.4	94	-13.9	94
μ_3	-70.2	15	-61.0	26	-12.2	91	-11.5	94	-11.9	94
μ_4	-52.3	33	-44.7	48	-10.1	93	-9.2	94	-9.6	96
μ_5	-77.7	14	-67.9	26	-15.3	89	-13.5	91	-14.1	90
σ^2	235.6	4	180.4	11	-7.0	91	-10.1	91	-9.8	92
ρ	-96.6	26	-97.7	27	-19.5	92	-18.1	94	-19.1	93

$\psi_0 = 0$, $\psi_1 = -1$, $\psi_2 = 1$. We used 5000 replication (samples) according to this setting.

The choice of the fixed values of the sensitivity parameter ψ_2 is a crucial step in the proposed approach. Assume that these values are labeled as ψ_2^f which are used for estimation process. We assume that ψ_2^f is fixed at the values $\{-2, -1, 0, 1, 2\}$. This is a reasonable range and allow the true value (ψ_2) to be underestimated, accurately estimated or overestimated. The stochastic EM algorithm, as described in Section 3, is used to find parameter estimates, for each replication. Also, the coverage percentage of 95% confidence interval are obtained. The estimates mean are shown in Table 1.

There is no missing values at the first time point by design, so the estimates of μ_1 are very close to the true value. Also, the percentage coverage of 95% confidence interval is very close to the nominal level. The mean parameters are generally underestimated. The relative bias for negative values of ψ_2^f are greater than those for positive values of ψ_2^f. This means that the bias is smaller when ψ_2 is accurately estimated or overestimated. The smallest bias is at the true value of ψ_2, $\psi_2^f = 1$. The coverage percentage is closer to its nominal level when ψ_2 is accurately estimated whereas we have poor percentage coverage for negative values of ψ_2^f. The relative bias is positive for σ^2 for negative values of ψ_2^f with poor percentage coverage. The parameter ρ has a negative bias with smallest value when ψ_2 is accurately estimated.

Other covariance structures have been tried. Also different values for missingness model parameters have been used. The qualitative results are the same as the above results, so they are not reported.

5 Application: breast cancer data

The proposed method is 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 VI (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. Each patient is randomly allocated to one of four chemotherapy treatments: A, B, C and D. It was planned to collect six measurements from each patient during the treatment period, one every three months. The study objective were to compare the quality of life for patients among the four treatment regimes. Each patient was asked to complete quality of life questionnaire.

The Perceived Adjustment to Chronic Illness Scale (PACIS) was an intended response. This is one-item scale comprising a global patient rating of the amount of effort costs to cope with her illness. Some patients refused in some visits to complete the questionnaire, resulting in intermittent missing values. A patient may not appear to fill the questionnaire if her mood is poor, and therefore the missing data mechanism is nonrandom (informative). The total number of patients survive the study period is 446 patients where 10 patients died during the study. Those patients are excluded from the analysis, so the missing values are not due to death. There are 64 patients with missing response at the first assessment and those are also excluded from the analysis. So, the number of subjects included in the analysis is 382 patients. Only 89 (23%) patient with no missing values whereas 293 (77%) patient with at least one missing measurement. For consecutive visits, starting from the second visit, the percentage of missing values are 29%, 36%, 47%, 54% and 62% respectively. 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.

These data have been analyzed by Hürny *et al.* (1992) ignoring the missing data (complete cases analysis) for responses of the first four measurements. Troxel *et al.* (1998) have analyzed the responses for the first 6 months of the study, including the missing values. Ibrahim *et al.* (2001) have analyzed the PACIS variable for the patients remain on the study long enough to have all assessments. Gad and Ahmed (2006) have analyzed the the PACIS variable for all assessments considering the missing data.

We adopt a mean model that allow each treatment to have its own effect.

Table 2: Application results: parameter estimates at different values of ψ_2

Par.	ψ_2^j						
	-5	-2	-1	0	1	2	5
μ_{01}	6.45	6.47	6.49	6.21	6.18	6.19	6.20
μ_{02}	3.80	3.78	3.89	6.01	6.03	6.02	6.01
μ_{03}	2.98	2.90	2.92	5.87	5.92	5.91	5.88
μ_{04}	1.45	1.32	1.24	5.40	5.53	5.47	5.43
μ_{05}	0.49	0.29	0.08	4.99	5.32	5.17	5.09
μ_{06}	-0.13	-0.40	-0.71	5.25	5.60	5.49	5.44
α_1	-0.15	-0.16	-0.17	-0.13	-0.09	-0.10	-0.12
α_2	-0.53	-0.56	-0.60	0.03	0.07	0.04	0.04
α_3	-0.95	-0.96	-1.01	-0.52	-0.48	-0.50	-0.52
σ^2	19.91	22.24	22.64	6.16	5.74	5.80	5.85
ρ	0.46	0.47	0.50	0.52	0.51	0.52	0.52

That is:

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

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

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

The first order auto-regressive AR(1) model is adopted for the covariance structure. In this model, the (i,j)th element of the covariance matrix, σ_{ij} equal to $\sigma^2\rho^{|i-j|}$ for $i,j = 1, ..., 6$. For the missing data mechanism, we use the logistic regression model as in Eq. (2.1) including only the previous and the current responses to keep the model simple. That is:

$$\text{logit}(r_{ij} = 1 \mid \Psi) = \Psi_0 + \Psi_1 Y_{ij-1} + \Psi_2 Y_{ij},$$

for i= 1,..., 382 and j= 1,2, ..., 6.

Gad and Ahmed (2006) estimate the parameter estimates of the same model for these data. The parameter estimates were $\mu_{01} = 6.27, \mu_{02} = 5.38, \mu_{03} = 5.77, \mu_{04} = 6.35, \mu_{05} = 5.43, \alpha_1 = -0.20, \alpha_2 = 0.04, \alpha_3 = -0.72, \sigma^2 = 4.49, \rho = 0.42, \psi_0 = 1.22, \psi_1 = 1.61$ and $\psi_2 = 1.06$.

The proposed approach is applied to these data. Hence, the values of ψ_2 need to be fixed at a plausible range of values. Minini and Chavance

(2004) suggest obtaining the plausible range of ψ_2 according to the drop-out probabilities at a given visit. As indicated by Gad and Ahmed (2006) that subjects with higher responses tend to be missing. This means that the ψ_2 could be a positive value. Also, some subjects may forget to fill-in the questionnaire. So, both positive and negative values of ψ_2 should be considered. A plausible range for ψ_2 could be between -5 and 5. The results are shown in Table 2.

For negative values of ψ_2^f the mean estimates decrease when ψ_2 moves towards zero for all treatments. Also, the mean estimates for negative values of ψ_2 are smaller than those at $\psi_2 = 0$ (MAR process). This reasonable because the negative values of ψ_2 mean that smaller responses tend to be missing. The mean estimates for positive values of ψ_2 are higher than those at $\psi_2 = 0$. This may be because, with positive values of ψ_2, subjects with higher responses are more likely to be missing. The variance parameters σ^2 for negative values of ψ_2 are higher than those for positive values of ψ_2. However, the estimates of ρ for negative values of ψ_2 are smaller than the estimates at positive values of ψ_2.

In this study, the conclusion are different for different values of ψ_2, i.e. the conclusion depend on the missing data mechanism. Hence, the missing data should be considered a serious source of concern.

6 Concluding Remarks

Modeling the missing values in the longitudinal data context have gained popularity in recent years. Selection model is one of these modeling approaches. However, as noted by many discussants to Diggle and Kenward (1994), such model depends on un-testable assumptions. The results are sensitive to the assumptions have been made. Many articles have been focused on the sensitivity analysis from different point of views.

The MNAR parameter ψ_2 is of major concern. Estimating this parameter confused with other parameter estimates. In this paper we study the sensitivity of results, in the presence of intermittent missing values, by fixing ψ_2 rather than estimating it. If the chosen range for ψ_2 is wide enough, one can expect that the true value of ψ_2 can be within this range. Also, different degrees of informativeness can be considered. This approach have been used by Minini and Chavance (2004). The main difference is that here we are interested in intermittent missing values not the special case, the dropout setting, as in Minini and Chavance (2004). Also, here we depend on the approach proposed by Gad and Ahmed (2006) which is a generalization of The Diggle-Kenward's model to the intermittent setting.

Acknowledgements:
The authors are grateful to the Quality of Life Committee of the IBCSG for permission to use the quality of life data.

References

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.

Daniels, M. J. and Hogan, J. W. (2000), Reparameterizing the pattern mixture model for sensitivity analyses under informative dropout, *Biometrics*, **56**, 1241–1248.

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

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

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

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.

Heckman, J. J. (1976), The common structure of statistical models of truncation, sample selection and limited dependent variables and simple estimator for such models, *Annals of Economic and Social Measurement*, **5**, 475–492.

Hürny, C., Bernhard, J., 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.

Ibrahim, J. G., Chen, M. H.and S. R. Lipsitz (2001), Missing responses in generalized 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, *Biometrika*, **42**, 805–820.

Little, R. J. A. (1993), Pattern mixture models for multivariate incomplete data, *Journal of the American Statistical Association*, 88, 125–134.

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

Louis, T. A. (1982), Finding the observed information matrix when using the EM algorithm, *Journal of Royal Statistical Society*, **B44**, 226–232.

Minini, P. and Chavance, M. (2004), Sensitivity analysis of longitudinal data with drop-outs, *Statistics in Medicine*, **23**, 1039–1054.

Molenberghs, G., Thijs, H., Kenward, M. G. and Verbeke, G. (2003), Sensitivity analysis for continues incomplete longitudinal outcomes, *Statistica Neerlandica*, **57**, 112–135.

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

Verbeke, G., Molenberghs, G., Thijs, H., Lesaffre, E. and Kenward, M. G. (2001), Sensitivity analysis for nonrandom dropout: A local influence approach, *Biometrics*, **57**, 7–14.