Bangladesh J. Sci. Res. **27**(2): 133-142, 2014 (December)

EM ALGORITHM FOR LONGITUDINAL DATA WITH NON-IGNORABLE MISSING VALUES: AN APPLICATION TO HEALTH DATA

Radia Taisir* and M. Ataharul Islam[1]

Department of Statistics, Biostatistics & Informatics, University of Dhaka, Dhaka-1000, Bangladesh

Abstract

Longitudinal studies involves repeated observations over time on the same experimental units and missingness may occur in non-ignorable fashion. For such longitudinal missing data, a Markov model may be used to model the binary response along with a suitable non-response model for the missing portion of the data. It is of the primary interest to estimate the effects of covariates on the binary response. Similar model for such incomplete longitudinal data exists where estimation of the regression parameters are obtained using likelihood method by summing over all possible values of the missing responses. In this paper, we propose an expectation-maximization (EM) algorithm technique for the estimation of the regression parameters which is computationally simple and produces similar efficient estimates as compared to the existing complex method of estimation. A comparison of the existing and the proposed estimation methods has been made by analyzing the Health and Retirement Survey (HRS) data of United States.

Key words: Incomplete data, Informative missingness, logistic regression, repeated measurement, EM algorithm.

Introduction

Longitudinal studies are designed to collect data on every individual at each time of follow-up and it is very common that all responses are not observed at all occasions. This incomplete or missing data leads standard analysis more difficult or inappropriate to implement, consequently the parameter estimates may become inefficient and/or biased. When missingness occurs depending on the response of that time point that is, the probability of being a non-respondent depends on the unobserved response, the data are said to be affected by non-ignorable missingness. If the missingness is non-ignorable, the resulting estimates are seriously biased.

Several researchers have worked over the last decade in variety of ways in analyzing longitudinal missing data. For non-ignorable missing data, a class of log-linear models were introduced by Fay (1986) and Baker and Laird (1988). The maximum likelihood estimates were obtained by using EM algorithm. The log-linear modeling approach for contingency tables was extended by Park and Brown (1994) and Green and Park (2003) under a Bayesian framework.

*Corresponding Author: e-mail: <rtysr86@gmail.com>. 1Department of Applied Statistics, East West University, Dhaka-1219, Bangladesh.

For longitudinal data, Bonetti *et. al.* (1999) proposed a method-of-moments estimation. This estimation technique is useful in some situation where likelihood maximization is problematic. Fitzmaurice *et. al.* (2001) described how bias can arise in generalized estimating equations (GEE) estimators where the missingness is informative. For longitudinal binary data with non-ignorable drop-out, Ten Have *et. al.* (1998) proposed mixed effects logistic regression models and these models were extended to ordinal response data with multiple causes of informative drop-out by Ten Have *et. al.* (2000) in a later paper. Accommodating intermittent missingness in addition to monotone missingness for second order dependency, a Markov chain model was proposed by Huang and Brown (1999). For Longitudinal continuous data with non-ignorable non-monotone missingness, Troxel *et. al.* (1998) proposed a full likelihood method involving a Markov assumption regarding the correlation structure of the longitudinal outcomes. A class of semi-parametric marginal regression models were developed by Rotnizky *et. al.* (1998) for handling non-ignorable missing mechanism. Fairclough (2002) described multiple imputation techniques

Cole *et. al.* (2005) developed a multistate Markov chain model for the analysis of longitudinal, categorical outcomes derived from QOL measures with the advantage over existing methods by allowing two or more QOL states, while accommodating both intermittent, informative missingness and covariate effects for first order dependency. For the purpose of inference, estimation of the regression parameters was carried out by a maximum likelihood method, summing over all possible values of the missing observations, which involves huge number of parameters to be estimated. Because of this and computational complexity, this inference procedure becomes complex and computationally intensive. Also for a data set containing very small number of missing observations, this approach can not produce efficient estimates of all regression parameters associated with the non-response model.

Considering the importance of the role of non-ignorable missingness in estimation, we focus on estimating the model parameters with informative missing values by using EM algorithm. The model and the inference procedure are outlined in the next sections. An application of the proposed estimation approach to the Health and Retirement Survey (HRS) binary data is discussed later.

The model for longitudinal data with non-ignorable missing values

Let, $x_{it} = (x_{it1}, x_{it2}, \ldots, x_{itp})'$ be the time-varying p-dimensional covariate vector for t^{th} individual at the t^{th} time point. For binary response y_{it}, the transition probabilities can be modelled by using logistic regression as

$$p_{il}(x_i(i, t-1)) = Pr[(Y_{it} = 1 \mid Y_i(i, t-1) = l, x_i(i, t-1))] = (\exp(\beta_{il}' x_i(i, t-1)))/(1+\exp(\beta_{il}' x_i(i, t-1))) \; ; l = 0, 1, \quad (1)$$

where $\beta_l = (\beta_{l0}, \beta_{l1}, \cdots, \beta_{lp})'$ is the set of regression parameter associated with the transition model from l to 1. It follows that, $p_{l0}(x_{i,t-1}) = 1 - p_{l1}(x_{i,t-1})$.

Let R_{it}'s are the observation indicator for i^{th} individual at t-th time such that $R_{it} = 1$, if Y_{it} is observed; 0 otherwise. Under non-ignorable missing mechanism, R_{it} depends on the observed responses. Accordingly a common logistic regression model is assumed for the non-response model. That is, the conditional probability that Y_{it} is observed given that $Y_{it} = j$ is defined by

$$q_{lj}(z_{it}) = \Pr(R_{it} = l \mid Y_{it} = j, z_{it}) = \exp(\eta_{lj}' z_{it}) / (1 + \exp(\eta_{lj}' z_{it})); \; l, j = 0, 1. \quad (2)$$

Following Cole *et. al.* (2005), for $l, j = 0, 1$, the non-ignorable incomplete binary data model may be written as

$$\Pr(Y_{it} = j, R_{it} = r_{it} \mid Y_{i,t-1} = l, x_{it}, z_{it}) = p_{lj}(x_{it}) q_j(z_{it})^{r_{it}} \{1 - q_j(z_{it})\}^{1-r_{it}}. \quad (3)$$

In (3), it is assumed that the likelihood for the initial state $\Pr(Y_{i1} = j)$ does not depend on any of the parameters associated with the transition probabilities and the initial state is always observed and also the covariate vectors are always observed.

Therefore, using (1) in (3) for $l, j = 0, 1$, one obtains the Markov model for longitudinal binary data subject to non-ignorable missingness

$$\Pr(Y_{it} = j, R_{it} = r_{it} \mid Y_{i,t-1} = l, x_{it}, z_{it}) = p_{lj}(x_{it}) q_j(z_{it})^{r_{it}} \{1 - q_j(z_{it})\}^{1-r_{it}}. \quad (4)$$

In the next section, we outline the proposed estimation method for estimating $\beta = (\beta_0', \beta_1')$, the two sets of parameter vectors for transition from 0 and 1, respectively.

Estimation technique by EM algorithm

Let y_i^{Obs} and y_i^{Miss} denote the observed and missing components of y_i, respectively and all the chains of the data is represented by y. Let, $\theta = (\beta, \eta)'$ be the vector of parameters associated with incomplete data model (4). Cole *et. al.* (2005) proposed ML estimation for the parameter $\theta = (\beta, \eta)'$ by maximizing the likelihood function

$$L^c(\theta; y_i^{Obs}) = \sum_{y_i^{Miss}} \left[\prod_{t=2}^{T_i} p_{y_{i(t-1)} y_{it}}(x_{i,t-1}) q_{y_{it}}(z_{it})^{r_{it}} \left\{ 1 - q_{y_{it}}(z_{it}) \right\}^{1-r_{it}} \right]. \quad (5)$$

It is clear from equation (5) that as the number of missing value increases, this likelihood estimation becomes complicated and computationally intensive. As an alternative, we propose EM algorithm approach for the estimation of the regression parameters $\theta = \beta$ of (4). Assuming the data is complete, the conditional likelihood for the sample of chains is expressed as

$$L(\theta; y_i) = \prod_{i=1}^{n} \Pr(Y_{i1} = y_{i1}) \prod_{t=2}^{T_i} p_{y_{i(t-1)} y_{it}}(x_{i,t-1}) \,. \qquad (6)$$

Under the assumption that the parameters of these components are distinct, for the estimation of the parameters for the state transitions, the initial-state likelihood can be ignored and (6) takes the following form

$$L(\theta; y_i) = \prod_i (i=1)^n \prod_t (t=2)^{T_i} p(y_i(t) = j \mid y_i(i,t-1) = l, x_i(i,t-1)) = \prod_i (i=1)^n \prod_t (t=2)^{T_i} p_{lj}(x_i(i,t-1)). \quad (7)$$

The E step of the EM algorithm sets the complete data-sufficient statistic

$$E(y_t = 1 | y_{t-1} = l, x_{t-1}) = p_{i1}^{(x_{t-1})} = \frac{\exp(\beta_l' x_{t-1})}{1 + \exp(\beta_l' x_{t-1})}; \quad l, j = 0, 1.$$

From the incomplete data, we calculate $E(y_t = 1 | y_{t-1} = l, x_{t-1}, \beta_l) = p_{i1}^{(x_{t-1})}$ and if $p_{i1}^{(x_{t-1})} \geq 0.5$ then in missing values we consider $y = 1$. But if $p_{i1}^{(x_{t-1})} < 0.5$, then in missing values we consider $y = 0$. Note that, we estimate the initial β parameters assuming the data as complete ignoring the missing values.

Once we impute the missing values in the E-step, we then maximize the likelihood (7) in the M-step. The score functions and the elements of the information matrix are give in equation (8) and (9) respectively.

$$\frac{\delta l(\beta; y_i)}{\delta \beta_u} = \sum_{i=1}^{n} \sum_{t=2}^{T_i} x_{i,t-1,lu} \left[1 - \frac{\exp(\beta_l' x_{i,t-1,l})}{1 + \exp(\beta_l' x_{i,t-1,l})} \right] \qquad (8)$$

$$\frac{\delta^2 l(\beta; y_i)}{\delta \beta_u \delta \beta_v} = -\sum_{i=1}^{n} \sum_{t=2}^{T_i} x_{i,t-1,lu} \times x_{i,t-1,lv} \left[1 - \frac{\exp(\beta_l' x_{i,t-1,l})}{\{1 + \exp(\beta_l' x_{i,t-1,l})\}^2} \right] \qquad (9)$$

Finally using the score vector and information matrix we get the estimates of the regression parameters by applying Newton-Raphson algorithm.

Analysis of HRS data

To compare the two estimation methods discussed in previous section, we fit the Markov model (4) to the Mental Health Index Data taken from Health and Retirement Survey (HRS) Data by both approaches. The HRS is a longitudinal household survey data set for the study of retirement and health among the elderly in the United States that surveys more than 22,000 Americans over the age of 50 on subjects like health care, housing, assets, pensions, employment and disability in every two years at the University of Michigan in Ann Arbor. Respondents in the initial HRS cohort were those who born during 1931 to 1941. This cohort was first interviewed in 1992 and subsequently every two years and the last interview was held in 2006. Detailed on the dataset can be found at the the HRS website (http://hrsonline.isr.umich.edu) and in Islam *et al.* (2009).

For this study, we have considered only last two waves (follow-ups) of the study and selected only those individuals whose response at the first wave are complete and covariate information on both waves are available. In this subset of the data, there are 16504 individuals in the 1st wave and 372 individuals responses were missing at the 2nd wave.

Our objective is to estimate the effect of gender (x_{it1}) and age (x_{it2}) on the dependent variable mental health index (y_{it}) by two estimation methods. This mental health index was derived using a score on the Center for Epidemiologic Studies Depression (CESD) scale. The CESD score (ranges 0 to 8) is the sum of the eight indicators such as 'felt sad', 'felt alone'. Considering the CESD score equal to 0 as 'no depression' and the CESD score greater than 0 as 'depression' we categorized the dependent variable. Then numerical scores 0 and 1 are assigned to the categories 'no depression' and 'depression' respectively. The distribution of the selected individals is reported in Table 1.

Table 1. Frequency distribution of Depression status by the selected covariates.

		Depression Status		
		No Depression (%)	Depression (%)	Total
Gender*	Male	3358 (51.9)	3114 (48.1)	6472
	Female	4185 (41.7)	5847 (58.3)	10032
Age*	<40	37 (39.8)	56 (60.2)	93
	40-50	378 (42.2)	517 (57.8)	895
	50-60	2028 (45.9)	2387 (54.1)	4415
	60-70	2770 (48.8)	2910 (51.2)	5680
	>70	2330 (43)	3091 (57)	5421

*p-value<0.01

From the Table 1, we obtain that the proportion of depression is higher for females as compared to males. It is also clear that, the proportion of depression is quite large in <40, 40-50 and >70 age intervals. Both of the covariates have significant association with depression status.

Estimation of the regression parameters obtained by the EM algorithm technique are reported in Table 2. ML estimates proposed by Cole *et. al.* (2005) are also reported in the same table.

Table 2 shows that both covariates gender and age have significant effect on transition from the state 'no depression' to 'depression'. The covariate 'gender' has negative impact but the covariate 'age' has positive impact to change the status from 'no depression' to 'depression'. For transition type 'depression' to 'depression', gender and age also have significant effects, gender has negative and age has positive impact to stay at the 'depression' state.

Table 2. Estimates of the regression parameters by likelihood method and EM algorithm approach for the HRS incomplete data.

		Likelihood Method		EM Method	
Parameter	Variable	Estimate	SE	Estimate	SE
Transitions from 'no depression'					
β_{00}	Intercept	-1.455	0.162	-1.105	0.161
β_{01}	Gender	-0.282*	0.051	-0.266*	0.051
β_{02}	Age	0.012*	0.002	0.006*	0.002
Transitions from 'depression'					
β_{10}	Intercept	0.723	0.125	0.787	0.148
β_{11}	Gender	-0.223*	0.043	-0.190*	0.052
β_{12}	Age	0.008*	0.002	0.007*	0.002
Logits of observation probabilities for 'no depression'					
η_{00}	Intercept	2.321	1.189	-	-
η_{01}	Gender	-1.968*	0.470	-	-
η_{02}	Age	0.054*	0.019	-	-
Logits of observation probabilities for 'depression'					
η_{10}	Intercept	10.225	0.539	-	-
η_{11}	Gender	0.123	0.142	-	-
η_{12}	Age	-0.093*	0.007	-	-

*$p-value$<0.01. Female is used as the reference for gender

Table 3. Parameter estimates and standard errors under likelihood and EM algorithm approaches for different hypothetical samples with different missing proportions, γ.

		Likelihood Method		EM Method	
Parameter	Variable	Estimate	SE	Estimate	SE
γ = 5% (n = 7440)					
β_{00}	Intercept	-1.916	0.246	-1.173	0.240
β_{01}	Gender	-0.380*	0.077	-0.341*	0.077
β_{02}	Age	0.020*	0.004	0.007**	0.004
β_{10}	Intercept	0.268	0.186	0.397	0.220
β_{11}	Gender	-0.257*	0.065	-0.179**	0.077
β_{12}	Age	0.015*	0.003	0.013*	0.003
η_{00}	Intercept	2.214	1.146	-	-
η_{01}	Gender	-1.977*	0.468	-	-
η_{02}	Age	0.042**	0.018	-	-
η_{10}	Intercept	9.307	0.559	-	-
η_{11}	Gender	0.148	0.152	-	-
η_{12}	Age	-0.091*	0.007	-	-
γ = 15% (n = 2480)					
β_{00}	Intercept	-3.058	0.444	-1.090	0.411
β_{01}	Gender	-0.100	0.135	-0.035	0.134
β_{02}	Age	0.038*	0.007	0.003	0.006
β_{10}	Intercept	0.055	0.341	0.451	0.395
β_{11}	Gender	-0.423*	0.117	-0.232***	0.140
β_{12}	Age	0.021*	0.005	0.016*	0.006
η_{00}	Intercept	-0.109	1.364	-	-
η_{01}	Gender	-2.035*	0.504	-	-
η_{02}	Age	0.064*	0.022	-	-
η_{10}	Intercept	7.420	0.559	-	-
η_{11}	Gender	0.069	0.158	-	-
η_{12}	Age	-0.083*	0.007	-	-

Female is used as the reference for gender. *p-value <0.01, **P-value<0.05 and ***p-value<0.1

This finding makes sense, because, as age increases, individuals are more likely to transit from 'no depression' to 'depression' state (therefore positive effect for transition from 0 to 1) and as they reach 'depression' state, they remain depressed (hence effect for 1 to 1 transition model). On the other hand, males are psychologically stronger than females. Thus they are less likely to get depressed when they are not depressed (hence negative effect for transition type 0 → 1), and once they are depressed, they are less likely to remain depressed (hence negative effect for transition type 1 → 1).

For the observation probabilities for 'no depression', we observe that both of the covariates have significant effects on the responses to be observed. On the other hand,the result from non-response model indicates that the chance of missing response increases as age increases.

From the table it is clear that the parameter estimates obtained by the proposed EM technique are almost equally efficient as compared to that of likelihood approach, the standard error of the estimates produced by two approaches are almost identical.

Estimation under small and large proportion of missing data

Here to compare the performance of estimation technique under different proportion of missing cases, we draw some hypothetical samples. To do so, we fix 372 missing responses and select random sample of size n^* from the remaining (16504 - 372 = 16132) individuals such that there are γ% missing responses in the sample of size *n* (= n^* + 372). Note that this is not a random sample.

Table 3 summarizes the estimation performance for γ = 5% and 15%. Irrespective of the missing proportion, the standard errors under both approaches are almost identical for 0→1 transition model. On the other hand, the performance of likelihood method is slightly better than EM algorithm approach for 1→1 transition model, but this efficiency gain is not too much.

Conclusion

We have used an alternative EM algorithm approach of estimation of the regression parameters of the Markov model for longitudinal informative missing data. In a position to pick one out of two alternative inference methods that are equally efficient, the simple answer is to pick the one that is simple in theory, easy to apply and computationally less intensive. In all of these respects our proposed EM approach outperforms the likelihood approach proposed by Cole *et. al.* (2005). Therefore, one can avoid doing complex algebra and complicated programming algorithm by using our proposed EM algorithm technique accommodating both longitudinal nature of the data and non-ignorable missingness and get efficient estimates.

Further note that, in EM algorithm approach we do not need to estimate huge number of parameters. As we have seen, the likelihood approach requires 12 parameters including the

parameters for the non-response model. On the other hand, we can achieve similar efficient regression effect by estimating only 6 parameters. That is why the estimation procedure becomes more simple, takes less time for computation. But in likelihood estimation approach, this huge number of parameters make the whole procedure computationally inconvenient. However, imposing appropriate restrictions, this large parameter set can be reduced.

Acknowledgement

The authors acknowledge gratefully to the HRS (Health and Retirement Study) sponsored by the National Institute of Aging and conducted by the University of Michigan for giving permission to use the dataset for this research.

References

Baker, S. G. and N. M. Laird. 1988. Regression analysis for categorical variables with outcomes subject to non-ignorable nonresponse. *Journal of the American Statistical Association*. **83**(401) : 62-69.

Bonetti, M., B. F. Cole and R. D. Gelber. 1999. A method-of-moments estimmation procedure for categorical quality-of-life data with non-ignorable missingness. *Journal of the American Statistical Association.* **94**(448) :1025-1034.

Cole, B. F., M. Bonetti, A. M. Zalavasky and R. D. Gelber. 2005. A multistate markov chain model for longitudinal, categorical quality-of-life data subject to non-ignorable missingness. *Statistics in Medicine.* **24**(15) :2317-2334.

Fairclough, D. L. 2002. Multiple imputation for non-random missing data in longitudinal studies of health related quality of life. *In* Statistical Methods for Quality of Life Studies; Design, Measurements and Analysis. Mesbah M, Cole BF, Lee M-LT (eds.). Springer, US. pp. 323-337.

Fay, R. E. 1986. Causal Models for Patterns of Nonresponse. *Journal of the American Statistical Association.* **81**(394):354-365.

Fitzmaurice, G. M., S. R. Lipsitz, G. Molenberghs and J. G. Ibrahim. 2001. Bias in estimating association parameters for longitudinal binary responses with drop-outs. *Biometrics.* **57**(1): 15-21.

Green, P. E. and T. Park. 2003. A bayesian hierarchical model for categorical data with non-ignorable nonresponse. *Biometrics.* **59**(4) :886-896.

Health and Retirement Study (HRS). 2015. Public release data files. The University of Michigan. Retrieved from http://hrsonline.isr.umich.edu

Huang, S. and M. B. Brown. 1999. A markov chain model for longitudinal categorical data when there may be non-ignorable non-response. *Journal of Applied Statistics.* **26**(1) :5-18.

Islam, M.A., R.I. Chowdhury, and S. Huda. 2009. *Markov models with covariate dependence for repeated measures*. Nova Science, New York.

Park, T. and M. B. Brown. 1994. Models for categorical data with non-ignorable nonresponse. *Journal of the American Statistical Association.* **89** (425):44-52.

Rotnitzky, A., J. M. Robins and D. O. Scharfstein. 1998. Semiparametric regression for repeated outcomes with non-ignorable nonresponse. *Journal of the American Statistical Association.* **93**(444):1321-1339.

Ten Have, T.R., A. R. Kunselman, E. P. Pulksteins, J. R. Landis (1998). Mixed effects logistic regression models for longitudinal binary response data with informative drop-out. *Biometrics.* **54**(1):367-383.

Ten Have, T.R., M. E. Miller, B.A. Reboussin and M.K. James.2000. Mixed effects logistic regressions models for longitudinal ordinal functional response data with multiple-cause drop-out from the longitudinal study of aging. *Biometrics.* **56**(1):279-287.

Troxel, A. B., D. P. Harrington and S. R. Lipsitz. 1998. Analysis of longitudinal data with non-ignorable non-monotone missing values. *Applied Statistics.* **47**(3):425-438.

(Manuscript received on 26 May, 2014; revised on 12 November, 2014)