Bangladesh J. Sci. Res. 27(2): 121-131, 2014 (December)

CHEMICAL SUITABILITY OF GROUND WATER FOR IRRIGATION IN TRIMOHONI AND SAGARDARI UNION, KESHABPUR UPAZILA, JESSORE, BANGLADESH

S. K.Saha[*], B. M. Rabby Hossain[1], Md. Anwar Jahid[2]

Department of Geology, University of Dhaka, Dhaka-1000, Bangladesh

Abstract

The paper intends to provide guidance to evaluate and identify a standard ground water chemistry data for irrigation in south western region of Bangladesh. During the course of hydrogeological studies in the study area, twenty water samples were collected from twenty different villages and chemically analyzed. The analytical results revealed that the water was slightly acidic to slightly alkaline (pH 6.68 - 7.32) and TDS values range from 565 to 1073 mg/l. The other parameters like sodium adsorption ratio, (SAR) (0.10 - 0.27), sodium percentage (3.22 - 7.13), residual sodium carbonate, (RSC) (3.2 - 5.33) and potential soil salinity, (PS) (less than 30) were below the desired limit suggesting the suitability for irrigation purpose. Considering SAR, permeability index and salinity hazard, all waters could be applied safely for irrigation without any hazard to crops.

Key words: Ground water, irrigation, water chemistry, south-west Bangladesh

Introduction

Irrigation is an age-old art. Historically, civilization has followed the development of irrigation. The duration of civilized people is probably dependent on many factors, of which a permanent profitable agriculture is significantly important (Hansen *et al.* 1979). Around 80% people in Bangladesh are living in rural areas (BBS 2001) and are belonging to agrarian structure. Food security of that inhabitant is mainly agro based. The agriculture sector plays a pivotal role in the economy of the country accounting for 31.6 per cent of total GDP in 2000-2001. The agricultural sector comprises of crops, forests, fisheries and livestock. Of the agricultural GDP, the crops sub-sector contributes to 71 per cent itself. Keshabpur Upazila (Fig. 1) belongs to mature deltaic plain (Choudhury 2001). More especially this area is a part of Gopalganj-Khulna peat basin. Differences in elevation between river banks and basin centre usually are about one meter (Brammer 1997). During the time that the delta was being built, the activity of the river Ganges helped to build it up latter as a plain (Abedin *et al.* 1990). Kopothaksha is the main river

[*] Corresponding author: <sahask03@gmail.com; sks@univdhaka.edu>. [1]Department of Disaster Science and Management, University of Dhaka, Bangladesh. [2]Institute of Water Resources and Flood Management, Bangladesh University of Engineering and Technology, Dhaka-1000.

channel in the region. But due to Farrakka Barrage at the upstream, the natural flow of the river is being shut down day by day. This accelerates the siltation at the river. As a result of acute water shortage in the dry season, awful drainage congestion and overflowing in the monsoon are the common feature in the study area.

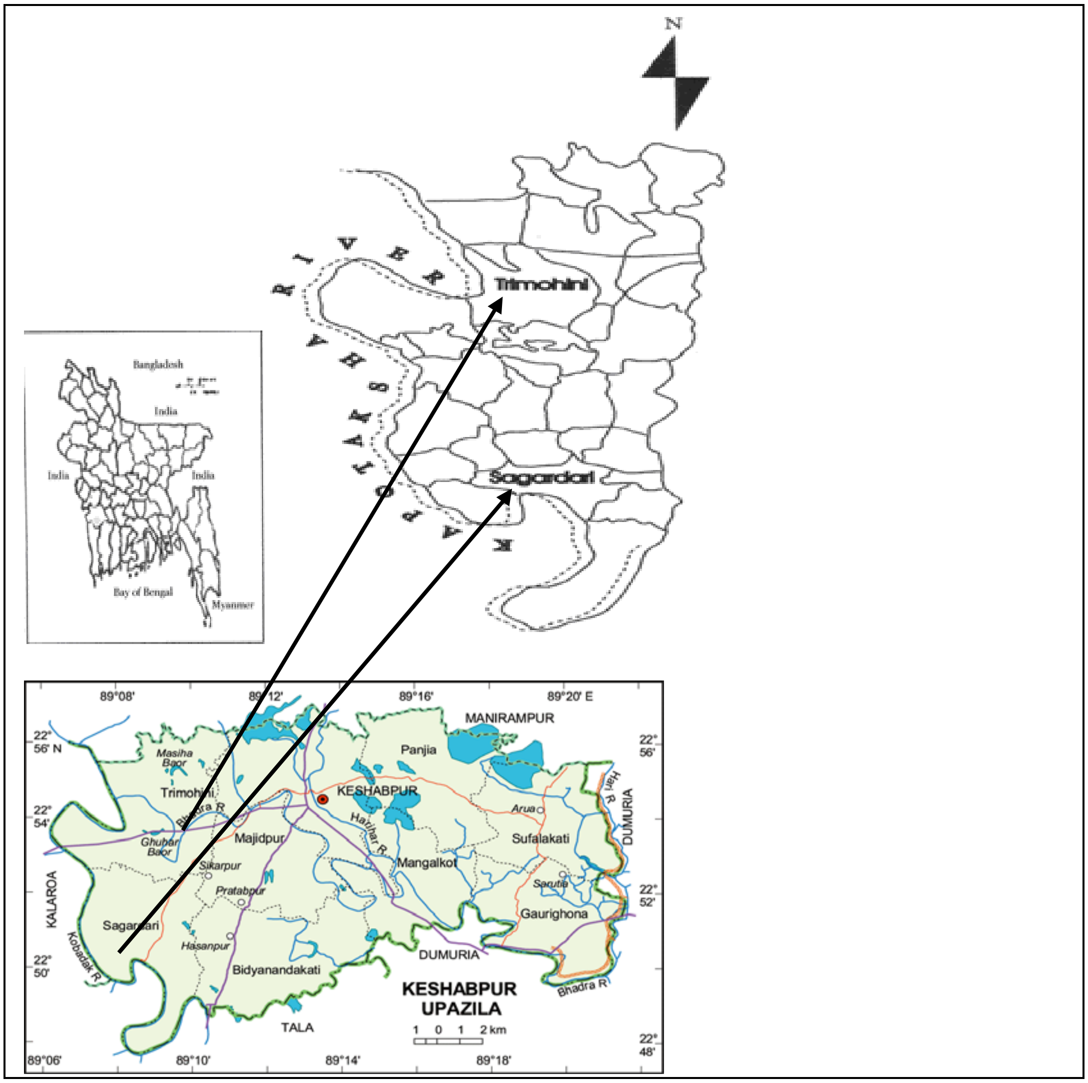


Fig.1. Location map of the study area (Source: Banglapedia).

In the Keshabpur Upazila (Fig.1), total irrigable land area are 194 sq. km, among it 163 sq. km are at present in irrigated condition. The irrigation system of this area was totally based on ground water. For the irrigation purpose total five thousand and five hundred shallow (80 m) tube-wells were established in the Keshabpur Upazila. The study area, Trimohini union consist four hundred ninety five and Sagardari union consisted of six hundred thirty six shallow tube wells. The irrigation system was maintained by the earthen channel system. In Bangladesh, irrigation water quality standard is recommended for surface water designating of pH, BOD, DO and total coliform whereas no standard prevail for irrigation water quality on ground water resources. So, it is an urgent need to highlight and evaluate irrigation water quality standards of ground water in Bangladesh.

Materials and Methods

The study area Keshabpur Upazila, south-western part of Bangladesh is situated under the district of Jessore. Twenty villages were studied for sample collection from two Unions. The data on existing agricultural scenario has been collected through ‘Structured Interview’ during November to December, 2003. The water samples for laboratory analysis (Table 1.) were collected from irrigation pump of the study area. Clean and dried plastic bottles without any contamination were used for sampling. The bottle was rinsed with sampled water during the collection of samples and then sealed with proper labeling and preserved for chemical analysis.

Table 1. Analytical methods of parameters used in the present study.

Parameters	Method	Procedure
pH	pH Meter (p^{Hep}, H $_1$98107 HANNA)	Field
EC	Conductivity/TDS meter (H1-9635)	,,
TDS	Conductivity/TDS meter (H1-9635)	,,
Na^+	Flame photometer (APHA 1995)	Laboratory
K^+	Flame photometer (APHA 1995)	,,
Ca^{++}	Titrimetric method (Ramesh and Anbu 1996)	,,
Mg^{++}	Titrimetric method (Ramesh and Anbu 1996)	,,
Cl^-	Titrimetric method (Ramesh and Anbu 1996)	,,
SO_4^{2-}	Turbidimetry method, Helis UV- visible Spectrophotometer (APHA1995)	,,
CO_3^{2-}	Titrimetric method (Ramesh and Anbu 1996)	,,
HCO_3^-	Titrimetric method (Ramesh and Anbu 1996)	,,

Results and Discussion

Major ion chemistry of irrigation water: Chemical composition of ground water is the combined result of the composition of water that enters the ground water reservoir and reactions with minerals present in the rock that may modify the water composition. The major components of ground water were Na^+, K^+, Mg^{2+}, Ca^{2+}, HCO_3^-, SO_4^{2-}, Cl^- and PO_4^{3-} (Appelo and Postma 1994). In laboratory analysis, the electro neutrality of up to 2% are inevitable in almost all laboratories and the differences between 2 - 5% are acceptable

(Appelo and Postma 1994). The electro neutrality (E.N. %) value of all samples were within the standard range.

The results of the chemical analysis for all water samples collected from different sites have been reported in Table 2.

Table 2. Major ionic constituents of ground water used for irrigation.

Sampling site	E.NVal-ue	pH	EC µS /cm	TDS mg/l	Ca^{2+} mEq/l	Mg^{2+} mEq/l	Na^{+} mEq/l	K^{+} mEq/l	Cl^{-} mEq/l	SO_4^{2-} mEq/l	CO_3^{2-} mEq/ l	HCO_3^{-} mEq/l
1.Bhalukghar	− 1.73	7.01	669	447	1.99	5.92	0.26	0.05	0.50	0.58	3.8	7.4
2.Chalitabaria	+0.23	6.69	928	650	2.19	5.89	0.46	0.05	1.00	0.54	5.0	7.0
3. Janpur	− 3.28	7.32	571	400	1.89	5.36	0.22	0.04	0.50	0.50	5.0	7.0
4. Srirampur	− 1.75	7.01	785	549	1.79	5.46	0.26	0.05	0.50	0.41	4.6	6.9
5. Barandali	− 1.68	6.8	803	562	2.09	5.46	0.26	0.06	0.50	0.43	3.8	7.2
6. Chandra	+0.25	6.74	1073	749	2.19	5.23	0.52	0.05	0.50	0.43	4.0	7.0
7. Mrizanagar	− 3.75	6.68	884	620	1.79	5.35	0.27	0.03	0.50	0.50	4.0	7.0
8. Begampur	+1.58	6.85	850	592	1.89	6.81	0.22	0.07	0.25	0.41	6.0	8.03
9. Kariakhali	− 0.66	6.74	980	684	1.99	5.92	0.23	0.05	0.75	0.54	4.2	7.0
10. Satbaria	− 3.10	6.79	765	530	2.19	4.93	0.30	0.06	0.25	0.70	4.0	7.0
11. Barihati	− 0.93	6.75	744	517	2.19	5.89	0.32	0.08	0.25	1.16	4.8	7.2
12. Gopsona	− 4.45	6.73	735	513	2.09	6.31	0.32	0.07	0.25	0.95	3.6	8.39
13. Kasta	− 1.91	6.8	634	443	2.09	5.49	0.32	0.06	0.25	1.00	4.0	7.0
14. Meherpur	− 1.11	6.68	967	677	2.39	5.49	0.50	0.03	0.50	0.87	4.8	7.2
15. Fatehpur	+0.32	6.78	807	562	2.19	4.93	0.46	0.04	0.25	0.70	4.6	6.6
16. Jhikra	+1.88	6.7	928	649	2.09	5.46	0.50	0.05	0.75	0.43	4.6	6.6
17. Sagardari	+1.00	6.75	917	640	2.29	5.26	0.40	0.06	0.25	0.58	4.0	7.0
18. Chingra	− 2.00	7.0	585	414	2.09	5.46	0.23	0.06	0.25	0.50	5.0	7.4
19.Gobindapur	− 1.02	7.04	565	393	2.09	6.31	0.28	0.03	0.25	0.62	3.6	8.0
20. Sekhpura	− 2.38	7.03	709	497	1.99	5.26	0.27	0.04	0.25	0.66	4.2	7.0

The pH value of the study areas ranged from 6.68 to 7.32 indicating slightly acidic to slightly alkaline in nature. In the study area, irrigation water was suitable for crop production as per the pH context. As because, the acceptable range of pH in irrigation water is from 6.5 to 7.5 (Hansen *et al.* 1979).

EC values of ground water in all sampling sites ranged from 565 to 1073 µS/cm and with the mean value of 795 µS/cm. EC value below 750 µS/cm is more suitable for irrigation and EC of 750 to 2250 µS/cm is moderately safe for irrigation (Wilcox 1963). As a result, the overall EC values of ground water in the study area was safe for irrigation purpose. The total dissolve solids (TDS) content in ground water of sampling sites vary from 393 to 749 mg/l with the average value of 554.4 mg/l. TDS below 525 mg/l is good for irrigation and 525 to1400 mg/l is in permissible level (Wilcox 1963). So, the ground water samples of the study area were suitable for irrigation. The concentration of Sodium in ground water varied from 0.22 - 0.50 mEq/l. Potassium concentration in ground water ranged from 0.03 - 0.08 mEq/l with the average concentration of 0.05 mEq/l. Calcium concentration in ground water varied from 1.79 - 2.39 mEq/l. Magnesium content in ground water was found to vary from 4.93 - 6.31 mEq/l with the average concentration of 5.25 mEq/l. Bicarbonate concentration in ground water samples varied from 402.6 to 512 mg/l. with the average concentration is 439 mg/l. On the other hand the carbonate concentration in ground water of the study areas varies from 108 to 180 mg/L and the average concentration is 131.4 mg/l. Chloride concentration in ground water was within 0.25 - 1.00 mEq/l. Sulfate concentration in ground water varied from 0.41 - 1.41 mEq/l as reported in Table 2. The sources of major ions in water can be defined by plotting the samples according to the variation of weight ratio of Na^+/ (Na^+ + Ca^{++}) as a function of the TDS (Gibbs 1970). It is observed from the Gibbs diagram that the major mechanism

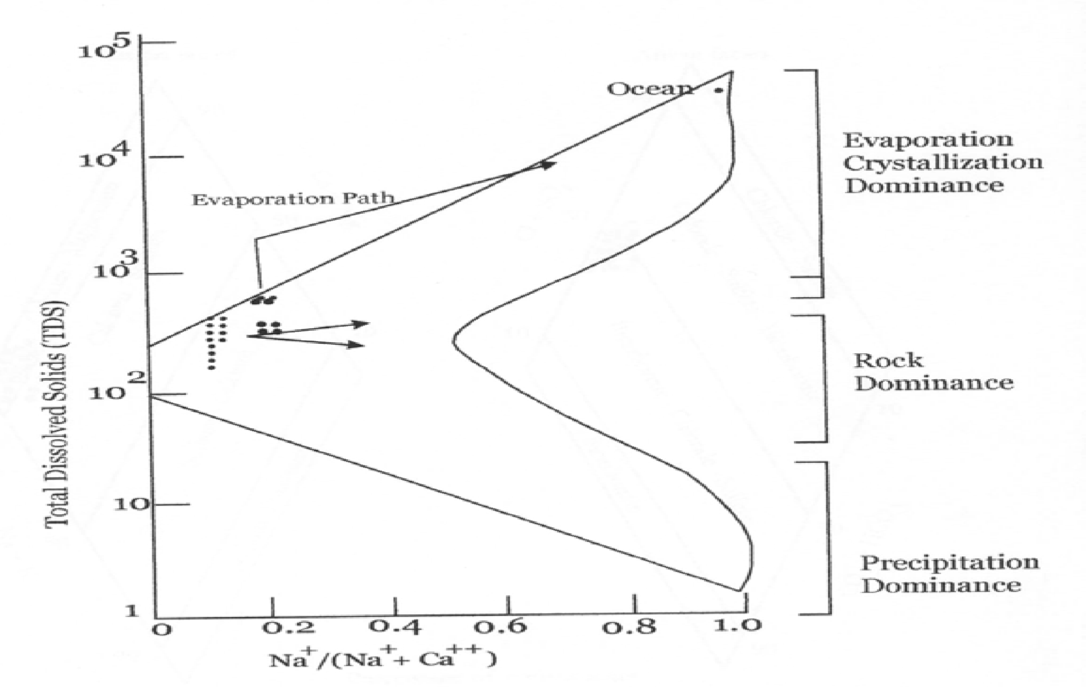


Fig. 2. Gibbs diagram to find our sources of groundwater chemistry in the study area (Gibbs 1970).

controlling the water chemistry was rock dominated and a slight influence from sea or ocean by crystallization and evaporation process.

Status of irrigation water chemistry: The chemical quality of water is an essential factor to be considered in evaluating its suitability for irrigation use. In this study, it is noted that total dissolved solids (TDS), electrical conductivity (EC), sodium percentage (% Na) and sodium adsorption ratio (SAR) are considered for assessing the suitability of ground water for irrigation purpose.

Sodium hazard: Sodium concentration is very important in classifying irrigation water because sodium by the process of base exchange replaces calcium in the soil thereby reducing the permeability of soil which has greater effect on plant growth. Sodium content in chemical analysis is reported as percent sodium, which is determined as:

Na percentage = $\{(Na^{+}+ K^{+})/ (Ca^{2+}+ Mg^{2+}+ Na^{+}+ K^{+})\}$ x 100, where concentrations of cations are expressed in mEq/l (Richards 1954). The relative activity of sodium ion in exchange reaction with soil is expressed in terms of sodium absorption ratio (SAR). The ability of water to expel calcium and magnesium by sodium can be estimated with the aid of the sodium absorption ratio (Richards 1954) as follows: $SAR=Na^{+}/\{(Ca^{2+}+Mg^{2+})/2\}^{1/2}$, where concentrations of cations are expressed as mEq/l. From the analysis it is found that Na% in the study area was within the range of 3.22 - 7.13 and SAR values were in the range of 0.10 to 0.27. All waters were excellent for irrigation purposes (Table 2) because the obtained SAR value didn't exceed the specified limit (SAR < 3) as per Richards (1954) which were excellent for irrigation purpose.

*Bicarbonate hazard:*Bicarbonate concentration of water has been suggested as an additional criterion for irrigation purpose. The convenient way of expressing values of the water in terms of residual sodium carbonate (RSC) is as follows:

RSC = $(CO_3^{2-} + HCO_3^{-}) – (Ca^{2+} + Mg^{2+})$, where all the concentrations of ions are expressed in me/l.

The recorded RSC values ranged from 3.2 to 5.33 indicating bicarbonate hazards (Table 2) where water samples having > 2.50 RSC is considered as bad in quality as per Mohato (1994).

Doneen classification: Doneen (1962) proposed a classification of water based on the salinity, permeability and toxicity of irrigation water. Potential soil salinity (PS) is defined as the concentration of chloride and half of the sulphates ions.

PS = $Cl^{-} + 1/2\ SO_4^{2-}$, where concentrations of all ions are expressed in mEq/l.

The PS values of all water samples were within the limit of 0.455 to 1.27 mEq/l revealing that all samples were in excellent to good in category for irrigation purpose.

The PS value versus total ionic concentration was plotted on Doneen's curve (Fig. 3). The Doneen classification diagram revealed that the collected water samples were in class I category. It shows that all ground water samples of the studied area were good for irrigation purpose.

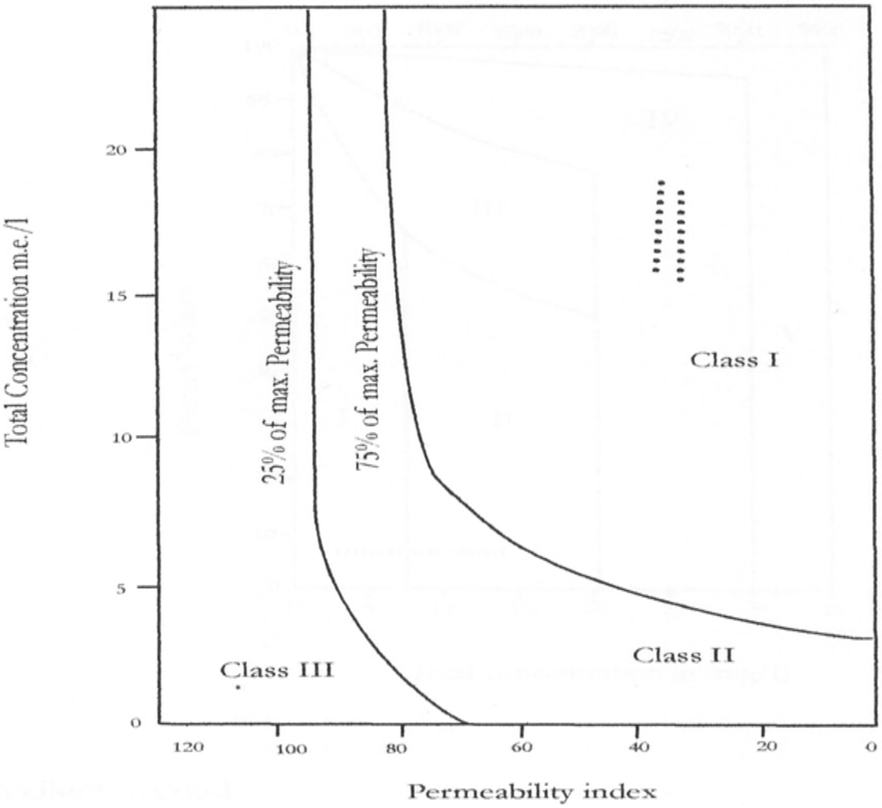
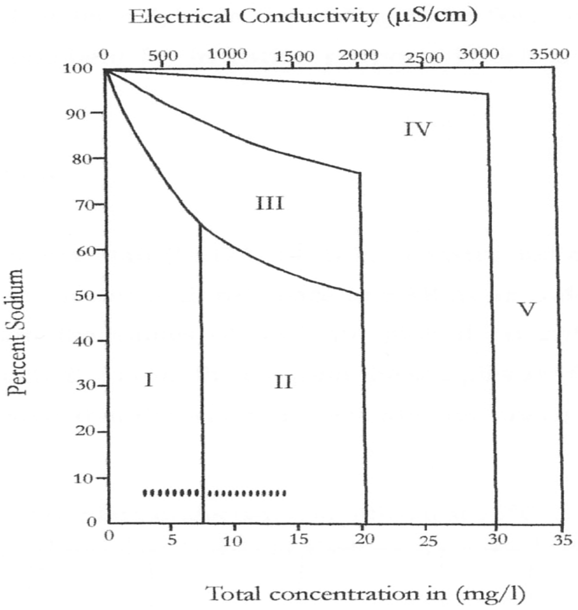


Fig. 3. Quality rating of waters for irrigation in the study area (Doneen 1962).

Wilcox classification: Since the quality requirement of irrigation water vary between types and drainage ability of soils and climate, thus the Universal Standards for irrigation water cannot be formulated. However, Wilcox (1963) proposed the following figure (Fig. 4) depending upon the percent sodium and electrical conductivity. The samples are within category I and II reveals that suitable for irrigation purpose.



Index:

I. Excellent to good

II. Good to permissible

III. Permissible to doubtful

IV. Doubtful to unsuitable

V. Unsuitable

Fig. 4. Quality classification of waters for irrigation (after Wilcox 1963).

USDA classification: Richards (1954) has constructed a diagram (Fig. 5) for classification of irrigation water with reference to SAR as an index of salinity hazard. In this diagram the values of SAR were plotted on arithmetic scale against electrical conductivity

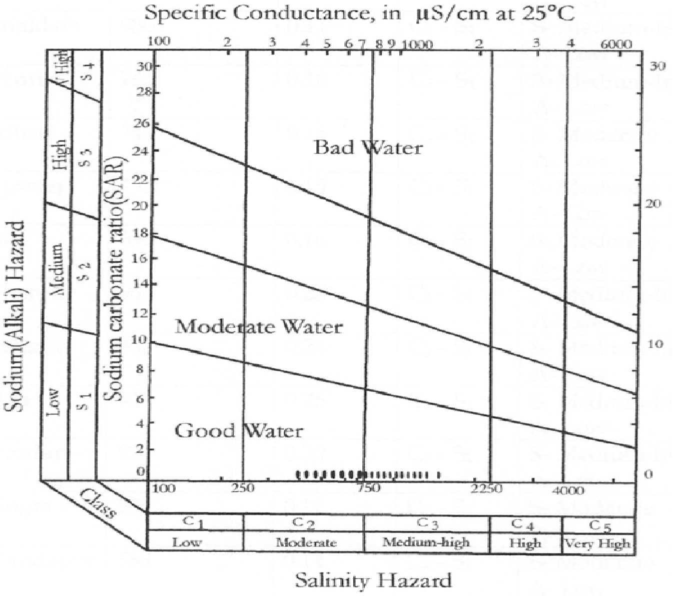


Fig. 5. Suitability of water for irrigation in the study areas (Richards 1954).

Table 3. The quality classification of groundwater samples for irrigation in the study area (USDA 1954).

Sample area	EC in µs/cm	SAR	Class	Hazards
1. Bhalukghar	669	0.13	$C_2 – S_1$	S- Moderate A- Low
2. Chalitabaria	928	0.23	$C_3– S_1$	S- Medium-high A- Low
3. Janpur	571	0.11	$C_2 – S_1$	S- Moderate A- Low
4. Srirampur	785	0.14	$C_3 – S_1$	S- Medium-high A- Low
5. Barandali	803	0.13	$C_3 – S_1$	S- Medium-high A- Low
6. Chandra	1073	0.27	$C_3 – S_1$	S- Medium-high A- Low
7. Mrizanagar	884	0.14	$C_3 – S_1$	S- Medium-high A- Low
8. Begampur	850	0.10	$C_3 – S_1$	S- Medium-high A- Low
9. Kariakhali	980	0.11	$C_3 – S_1$	S- Medium-high A- Low
10. Satbaria	765	0.16	$C_3 – S_1$	S- Medium-high A- Low
11. Barihati	744	0.16	$C_2 – S_1$	S- Moderate A- Low
12. Gopsona	735	0.15	$C_2 – S_1$	S- Moderate A- Low
13. Kasta	634	0.16	$C_2 – S_1$	S- Moderate A- Low
14. Meherpur	967	0.25	$C_3 – S_1$	S- Medium-high A- Low
15. Fatehpur	807	0.24	$C_3 – S_1$	S- Medium-high A- Low
16. Jhikra	928	0.26	$C_3 – S_1$	S- Medium-high A- Low
17. Sagardari	917	0.20	$C_3 – S_1$	S- Medium-high A- Low
18. Chingra	585	0.12	$C_2 – S_1$	S- Moderate A- Low
19. Gobindapur	565	0.14	$C_2 – S_1$	S- Moderate A- Low
20. Sekhpura	709	0.14	$C_2 – S_1$	S- Moderate A- Low

Note: A = Alkalinity, S = Salinity.

(EC). From the diagram all water samples fall into $C_2 – S_1$ and $C_3 – S_{1,}$ which showed that the water of the study area was suitable for irrigation.

Conclusion

A huge amount of ground water is used for irrigation purposes and therefore need irrigation water management. From the above research it was found that a slight seawater intrusion is remarkable in the study area, which in long run would be disaster for irrigation. So it is the high time to decline the pressure on ground water for irrigation purpose.

The ground water samples were showing SAR values less than one, which indicates that they were excellent for irrigation. As per sodium percentage below twenty (< 20) indicated excellent for irrigation. On the basis of Wilcox diagram the forty percent samples falling in excellent to good quality and other sixty percent falling in good to permissible class. On the basis of U.S. salinity diagram, forty percent samples were falling in C_2-S_1 and other sixty percent samples fell in C_3 - S_1 categories, which were suitable for irrigation purposes. The potential soil salinity (PS) value was less than three, which fell in excellent to good category. And the Doneen classification diagram on the basis of permeability index (PI) revealed that the all water samples in the study area fell in class 1. It showed that water is good for irrigation purpose in the study area.

References

Abedin, M.Z.; Hai, C.K. and Ali, M.O., (eds.). 1990. Homestead Plantation and Agro-forestry in Bangladesh, Proceeding of a national Workshop held on July 17-19, 1988 in Joypur, India. p 170.

APHA. 1995. Standard methods for the examination of water and wastewater, 19th eds, Washington DC, 200005, USA.

Appelo, C. A. J. and Postma, D. 1994. Geochemistry, groundwater and pollution, A.A. Balkema Publisher, Rotterdam, Netherlands. p 14-19.

Bangladesh Bureau of Statistics (BBS). 2001. Statistical Pocket Book of Bangladesh: 2000. Bangladesh Bureau of Statistics, Government of the People’s Republic of Bangladesh (GOB), Dhaka.

Banglapedia, http://www.banglapedia.org/

Brammer, H. 1997. Agricultural Development Possibilities in Bangladesh, University Press Limited, Dhaka, Bangladesh. p. 357.

Choudhury, Q. I. 2001. State of Biodiversity, Forum of Environmental Journalists of Bangladesh (FEJB). pp. 35-49.

Doneen, L. D. 1962. The influence of crop and soil on percolating waters. Proc. 1961 Biennial Conference on groundwater recharge: 156-163.

Gibbs, R. J. 1970. Mechanisms Controlling World Water Chemistry. *Science* **170**: 1088 - 1090.

Hansen, V. E.; Israelsen, O. W. and Stringham, G. E. 1979. Irrigation Principles and Practices, John Wiley and Sons, New York, U. S. A. pp. 366.

Mahato, D. 1994. Ground Water Quality in Weathered Singhbhum Granite Around Karanjia, Southern Bihar, India, Jour. Geol. Soc. India **43**: 685-689.

Ramesh, R. and Anbu, M., 1996. Chemical Methods for Environmental Analysis of Water and Sediment, Rajiv Beri for Macmillian India Limited, Madras 600002, pp. 56.

Richards, L. A., 1954. Diagnosis and improvement of saline and alkali soils., USDA Agricultural Handbook No. 60, US Department of Agriculture, Washington DC. pp. 160.

Wilcox, L. V., 1963. Factors for calculating the Sodium Adsorption Ratio (SAR), US Salinity Laboratory Mimeo Report.

(Manuscript received on 29 June 2014; revised on 09 September, 2014)

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

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

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.

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 *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 *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 for non-ignorable missing longitudinal quality-of-life (QOL) data.

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 missing values

Let, x_{it} = $(x_{it1}, x_{it2}, x_{it3},..., x_{itp})'$ be the time-varying p-dimensional covariate vector for i^{th} individual at the t^{th} time point. For binary response y_{it}, the transition probabilities can be modelled by using logistic regression as

$$p_{l1}\left(x_{i,t-1}\right) = \Pr\left(Y_{it} = 1 \mid Y_{i,t-1} = l, x_{i,t-1}\right) = \frac{\exp(\beta_l' x_{i,t-1})}{1+\exp(\beta_l' x_{i,t-1})}; \quad l = 0, 1. \qquad (1)$$

Where $\beta_l = (\beta_{l0}, \beta_{l2},...., \beta_{lp})'$ is the set of regression parameter associated with the transition model from l to 1. It follows that, $p_{l0}\left(x_{i,t-1}\right) = 1 - p_{l1}\left(x_{i,t-1}\right)$.

Let, R_{it}'s are the observation indicator for i^{th} individual at the 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

$$q_j\left(z_{it}\right) = \Pr(R_{it} = l \mid Y_{it} = j, z_{it}) = \frac{\exp(\eta_j' z_t)}{1+\exp(\eta_j' z_t)}; \quad l, \ j = 0, 1. \qquad (2)$$

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

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

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

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

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

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

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

$$L(\theta, y_i) = \prod_{i=1}^{n} \prod_{t=2}^{T_i} p(y_{it} = j \mid y_{i,t-1} = l, x_{i,t-1}) = \prod_{i=1}^{n} \prod_{t=2}^{T_i} p_{lj}(x_{i,t-1}). \qquad (7)$$

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

$$E(y_t = 1 \mid y_{t-1} = l, x_{t-1}) = \hat{p}_{l1}(x_{t-1}) = \frac{\exp(\beta_l' x_{t-1})}{1 + \exp(\beta_l' x_{t-1})}; \quad l, j = 0, 1.$$

From the incomplete data, we calculate $E(y_t = 1 \mid y_{t-1} = l, x_{t-1}, \beta_l) = \hat{p}_{l1}(x_{t-1})$ and if $\hat{p}_{l1}(x_{t-1}) \geq 0.5$ then in missing values we consider $y = 1$. But if $\hat{p}_{l1}(x_{t-1}) < 0.5$, 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 E-step, we get the complete data likelihood $L(\beta; y_i) = \prod_{i=1}^{n} \prod_{t=2}^{T_i} p_{lj}\left(x_{i,t-1}\right)$. Then we maximize the loglikelihood function in the M-step. The score functions and the elements of the information matrix are given in equation (8) and (9) respectively.

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

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

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 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 individuals 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.

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).

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.

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 *y%* missing responses in the sample of size n (= $n*$ + 372). Note that this is not a random sample.

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.

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. *J. 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. *J. 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. *J. 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. *J. 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 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. *J. 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)