JIRSS (2010)
Vol. 9, No. 2, pp 93-113

A Bayesian Approach to Estimate Parameters of a Random Coefficient Transition Binary Logistic Model with Non-monotone Missing Pattern and some Sensitivity Analyses

S. Eftekhari Mahabadi, M. Ganjali

Department of Statistics, Faculty of Mathematical Sciences, Shahid Beheshti University, Iran. (s-eftekhari@sbu.ac.ir,m-ganjali@sbu.ac.ir)

Abstract. A transition binary logistic model with random coefficients is proposed to model the unemployment statues of household members in two seasons of spring and summer. Data correspond to the labor force survey performed by Statistical Center of Iran in 2006. This model is introduced to take into account two kinds of correlation in the data; one due to the longitudinal nature of the study, that will be considered using a transition model, and the other due to the assumed correlation between responses of members of the same household which is taken into account by introducing random coefficients into the model. Due to the use of special sampling method in this survey (rotation sampling), some kinds of non-monotone missing pattern occur that are considered in the proposed model using the breakdown of the joint distribution of the response variables. A Bayesian approach is used to estimate model parameters via the Gibbs sampling method and data augmentation. Results of using

Key words and phrases: Clustered longitudinal data, Gibbs sampling, non-monotone missing pattern, transitional binary logistic model, unemployment Statues.

this model are compared with those of three other transitional models. The most applicable model which gains more interpretability and precision due to consideration of all aspects of the collected data is found. Also some sensitivity analysis are performed to assess asymmetric departures from the logistic link function and robustness of the posterior estimation of the transition parameter to the perturbations of the prior parameters.

1 Introduction

In a panel or longitudinal study, each subject is measured at several occasions. Sometimes there exist clusters of subjects in each period due to some kinds of initial relationship between subjects at each time, for example subjects that are members of the same family may indicate a cluster. In such clustered longitudinal studies, Y_{ijt} indicates the response variable for the jth member ($j = 1, \ldots, n_i$) of the ith ($i = 1, \ldots, K$) cluster at time t ($t = 1, \ldots, T$). We use n_i for the number of elements in the ith cluster to allow for variation of cluster sizes in the study as it seems reasonable if we assume households as clusters.

In these kinds of studies there exist two aspects of correlation that must be taken into account in the model; one is the correlation between responses of a subject in different occasions, and the other is due to clustering that results in correlated responses for subjects in the same cluster at each time period. We use random coefficients to allow for the clustering correlation in the model. Different models can be used to take into account the correlation between responses raised due to the longitudinal nature of the study. One possibility is marginal modeling, which can be used to make inferences about parameters averaged over the whole population (Stiratelli et al., 1984; Liang et al., 1992) A second possibility is random effects modeling, which makes inferences about variability between respondents (Harvile and Mee, 1984; Berridge and Dos Santos, 1996; Verbeke and Molenberghs, 1997; Tutz, 2005) The third approach would be to use Markov (transition) models (see Anderson and Goodman, 1957; Kaciroti et al., 2006; Sengul et al., 2007). For reviews of transition and other models for longitudinal responses, see McCullagh (1980), Diggle et al. (2002) and Agresti (2002).

In such longitudinal studies mentioned above, often some of the subjects do not respond in some occasions which cause for missing

responses. There are often two kinds of missing pattern considered for longitudinal studies; one is the monotone pattern that is most common in medical surveys and is caused by subject's dropout before the study is completed, and the other is the non-monotone or general pattern that occurs when some of the subjects withdraw from the study at some occasion and return to the study again at some other occasion.

Rubin (1976), Little and Rubin (2002) and Diggle and Kenward (1994), made important distinctions between the various types of missing mechanisms for each of the above mentioned patterns. For illustration of different mechanisms, a dummy variable M that indicates the missing situation of the variable Y may be defined as $M = 0$ if Y is observed and $M = 1$ otherwise. A missing mechanism is missing completely at random (MCAR) if the conditional distribution of M given Y and the model covariates is independent of the variable Y, Missing mechanism is missing at random (MAR), when this conditional distribution depends only on the observed part of the variable Y given the model covariates, finally missing mechanism is not missing at random (NMAR) when the missingness is dependent on both parts of Y, missed and observed. From a likelihood point of view MCAR and MAR are ignorable given disjoint parameter spaces of the response and the missing indicator models, but NMAR is non-ignorable.

In this article a Markov transition model with random coefficients for binary response variables is proposed. Moreover, we consider a general non-monotone missing data pattern with a MAR mechanism. Also we will introduce low information prior distributions for all model parameters in order to take advantage of using a Bayesian approach.

To analyze the dependence of binary response data on explanatory variables, the logistic transformation is probably the one most commonly used. However, it is tentative and therefore some consideration of adequacy is needed. If some non-logistic model gives a better fit it is important to discover this. An appealing and informative way to examine the accuracy of the logistic link is to construct extended models which include the logistic model as a special case. We use the family of the asymmetric transformations introduced by Aranda-Ordaz (1981) to access departures from the logistic link.

As we use a Bayesian approach toward estimating, another important issue is the robustness of the posterior distribution. We can

assess how robust the posterior distribution is to the selection of the prior distribution via sensitivity analysis, in which we assess changes in the posterior distribution over different prior distributions. When prior information is available, sensitivity analysis focuses on the structure of the prior distribution; when noninformative or low informative priors are used, it focuses on how different choices of prior parameters may influence the posterior inference. In our model we do a sensitivity analysis for the most interesting parameter in our model which is the transition parameter.

In the present article, we apply the data belonging to the labor force survey performed by Statistical Center of Iran 2006 in two seasons of spring and summer. The target population in this survey consists of all people who are, according to the definition, a member of ordinary households in urban or rural areas of the country. The sampling method in this survey is two stage clustered with classification. The first stage sampling units are clusters and for the second stage the sampling unit contains a group of three households those are usually adjacent. The respondent unit is the private settled living in the place considered in the sample. A panel survey is performed in this plan to allow study of variations between seasons in addition to within season's variation. Our main interest in analyzing these data is based on the potential clustering of the household members and the special missing pattern of this survey which is a kind of non-monotone missing data pattern.

In the next Section, the attributes of the data belonging to the labor force survey are discussed. In Section 3, the random coefficient transitional binary logistic model is presented. In this section the likelihood function and the posterior distribution of the model parameters are also given. In Section 4, the appropriate model for labor force data and its computational approach using Gibbs sampling method are discussed. Also the parameter estimates for the proposed model and a comparison of this model with three other models with different missing and random effect considerations are given. Some sensitivity analysis are also performed in Section 5 to access appropriateness of the logistic model and the robustness of the model estimates to different prior selections. Finally a brief conclusion along with possible further works are given in Section 6.

2 Labor force survey data

The data used in this paper are related to the unemployment statues of household members in two seasons of spring and summer of 2006. The information about 224691 people who were present in the study for at least one season is available. Among the sampling units, 9237 people refuse to fill the questionnaire and we omit them from the study because they only form four percent of the whole sample. Therefore there are 215454 subjects remained in the study. In researches related to unemployment issues, only the economical active part of the population is used toward calculations of unemployment indexes and rates. The economical active population is defined as all people who are at least ten years old and have either taken part in production and services (employed) in the week before the sampling week (reference week), or they had the employment opportunity in the reference week but they are unemployed. Because of this reason we first remove inactive part of the spring sample from the study that results in 130344 economically active people in the study. Again, among these we omit the ones who were active in spring but became inactive in summer that leads to 83144 subjects in the whole sample. At last, we remove 13 people from the resulting sample due to incomplete covariates. This finally leads us to have a sample of 83131 economically active people that includes 51791 households.

Due to the special rotating method of sampling in this survey, the sample is divided into three parts. There are 31153 people in the sample who are present only in the first season, 29890 people who are included in the survey only in the second season and 22088 numbers of them were present in both seasons. The binary response variable of interest in this study is the unemployment status of each member of household. Also the list of covariates which will be used through the analysis process is given in Table 1. Some descriptive statistics are also given for each variable in different seasons (considering the present part of sample in each season).

Figure (1) shows how the probability of being employed in summer depends on the spring employment status for different age groups based on part of sample which includes individuals who are present in both seasons. According to this Figure, the probability of employment in summer is higher for people who were employed in spring and that this probability increases as the age increases.

Table (1): List of variables in the labor force survey data with some descriptive statistics

Variable	Values	Percentage in Spring	Percentage in Summer
Unemployment Status	1: employed	90.10	91.10
	2: unemployed	9.90	9.00
Gender	1: male	79.60	80.10
	2: female	20.40	19.90
Highest educational qualification	1: illiterate	15.30	15.00
	2: under diploma	48.30	47.70
	3: diploma or upper diploma	24.00	24.70
	4: bachelor degree	7.50	7.60
	5: master of science or PhD	1.10	1.10
	6: others	1.90	4.00
Number of people living in the household	1: one member	0.60	0.70
	2: two members	7.00	6.80
	3: three members	16.60	16.50
	4: at least four members	75.80	76.00
Marital status	1: married	69.50	68.40
	2: widowed	1.20	1.30
	3: divorced	0.50	0.40
	4: single	28.80	29.90
Living area	1: urban	41.20	41.80
	2: rural	58.80	58.20
Age	continuous	mean: 35.75	mean: 35.72

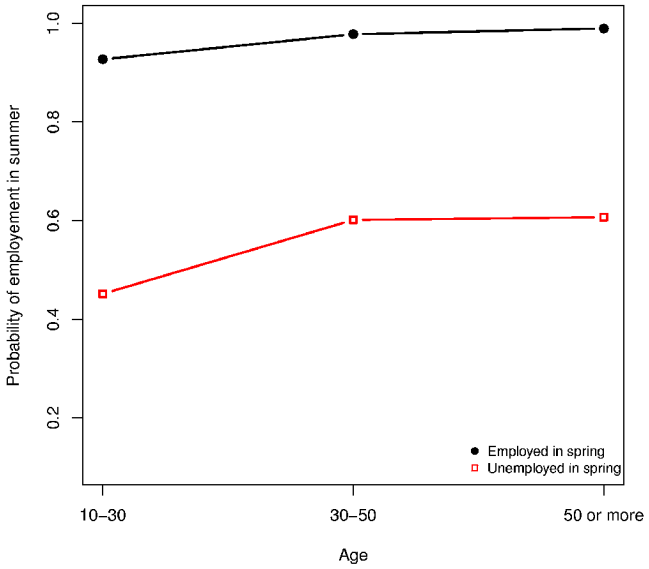

Figure (1): Probability of employment in summer conditional on the spring employment status and age.

3 Random coefficient transitional binary logistic model

Let Y_{ijt} denotes the binary response variable for the j-th individual of the i-th cluster at time t, in a longitudinal study in T periods. Also let $\{X_{ijt}\}$ be the set of all covariates in the study those which may be regarded as time varying. We assume a continuous latent variable U_{ijt} related to the binary response of Y_{ijt} with the following mechanism:

$$\begin{gathered} Y_{ijt} = 0 \Leftrightarrow U_{ijt} \leq \alpha_{0t} \\ i = 1, ..., K; \quad j = 1, ..., n_i; \quad t = 1, ..., T \end{gathered} \tag{1}$$

where α_{0t}, for $t = 1, \ldots, T$, are cut point parameters. Also We consider following transition model with random coefficients for the latent variable:

$$\begin{gathered} U_{ijt} = -X'_{ijt}\beta_t - \gamma_t y_{ij,t-1} - b_{it} + \varepsilon_{ijt} \\ \varepsilon_{ijt} \overset{iid}{\sim} F; b_i = (b_{i1}, b_{i2}, ..., b_{iT}) \overset{iid}{\sim} MVN(0, \Sigma_b); b_{it} \perp \varepsilon_{ijt} \\ i = 1, ..., K; \quad j = 1, ..., n_i; \quad t = 1, ..., T. \end{gathered} \tag{2}$$

In which $y_{ij0} = 0$ for all i and j, and we assume Σ_b as a diagonal matrix with diagonal elements $(\tau_1, ..., \tau_T)$. Using the above model U_{ijt} is independent of $U_{ij,t-1}$ given $Y_{ij,t-1}$ as a consequence of transitional nature of the model. Also random effects b_{it} are introduced to account for the correlation due to clustering. We can consider the binary response of Y_{ijt} as a Bernoulli distributed random variable with success probability $\pi_{ijt} = \Pr(Y_{ijt} = 1 | X_{ijt}, y_{ij,t-1}, b_{it})$ which according to the relation between Y_{ijt} and U_{ijt}, given in equation (1) and the proposed model in (2) can be regarded as:

$$\pi_{ijt} = F(\alpha_{0t} + X'_{ijt}\beta_t + \gamma_t y_{ij,t-1} + b_{it}) \tag{3}$$

Applying Different distribution functions, F, lead to different models. In our application, We especially use the logistic distribution that leads into a binary logistic model with success probability of

$$\pi_{ijt} = \frac{e^{\alpha_{0t} + X'_{ijt}\beta_t + \gamma y_{ij,t-1} + b_{it}}}{1 + e^{\alpha_{0t} + X'_{ijt}\beta_t + \gamma y_{ij,t-1} + b_{it}}}. \tag{4}$$

3.1 likelihood and posterior function

In order to obtain the likelihood function for the set of model parameters $\Theta = \{\alpha_{0t}, \beta_t, \gamma_t, \tau_t; t = 1, ..., T\}$ using the model proposed in the previous section, we have:

$$\begin{aligned} L(\Theta|Y_1, ..., Y_T) &= \prod_{i=1}^{K} \int_{b_i} \prod_{t=1}^{T} \prod_{j=1}^{n_i} f(Y_{ijt}|Y_{ij,t-1}, X_{ijt}, \alpha_{0t}, \beta_t, \gamma_t, b_{it}, \tau_t) \\ &\times \phi_{1,...,T}(b_i) db_i. \end{aligned} \quad (5)$$

In which $f(Y_{ijt}|Y_{ij,t-1}, X_{ijt}, \alpha_{0t}, \beta_t, \gamma_t, b_{it}, \tau_t)$ is the Bernoulli density with corresponding success probability π_{ijt} and also $\phi_{1,...,T}(.)$ is the multivariate normal density with mean 0 and covariance matrix Σ_b. The first product on the right hand side of the equation (5) is based on the assumption of having no correlations between different clusters and the second one results from the transitional property of the introduced model and the third product is obtained by the conditional independence assumption of responses in the same cluster given random effects. To complete the model specification, a diffuse prior distribution for Θ, $P(\Theta)$, is assumed where for $t = 1, \ldots, T$, $(\alpha_{0t}, \beta_t, \gamma_t)$ have independent diffuse normal priors with mean 0 and some large variance and τ_t follows a diffuse inverse gamma distribution.

The primary inferential quantity of interest is $(\alpha_{0t}, \beta_t, \gamma_t; t = 1, .., T)$. Obviously other parameters $(\tau_t, b_{it}; t = 1, ..., T; i = 1, ..., K)$ are also of interest. Given the complexity of the model, inference based on the complete data needs to be based on simulation techniques. For example Gibbs sampling or Markov Chain Monte Carlo (MCMC) methods can be used to construct inferences based on values drawn from the joint posterior density,

$$P(\Theta, b|Y_1, \ldots, Y_T, X) \propto L(\Theta|Y_1, \ldots, Y_T, X)P(\Theta)$$

When there are missing values in $Y = (Y_1, \ldots, Y_T)$, and missing data mechanism is ignorable (MAR), the Gibbs sampling for the complete data model can be easily modified. We include missing values in Y in the Gibbs sampling steps simply by drawing values from its conditional predictive distribution, given the observed values and the current draw of parameters that will be discussed more in the next sections.

4 Model and results for labor force survey data

In the labor force survey data we assume the unemployment statues of household members as the binary response variable of interest. The extracted data only contains two seasons, hence we have $T = 2$. Also we consider X as the matrix of all covariates introduced in Table 1. We consider π_{ijt} as the employment probability conditional on model covariates, previous response and random effects as defined in equation (3). Similar to model in (4), we assume a transition binary logistic model with random effects distributed as mentioned in (2) for the response variable. For the complete data we have the likelihood as (5) but considering the missing data problem with a non-monotone pattern which occurred in the labor force survey data, we have to compute the appropriate likelihood function. This pattern has some respondents in the first period who are removed from the study in the second period and some individuals who are not in the study at the first period but are included in the study, as new individuals, for the second period. The missing responses in these data are based on the sampling design so that, response variables have a MAR mechanism. Hence the likelihood function for these data should be viewed in three parts. First part consists of households who were present in both seasons:

$$L^{(1)}(\Theta|Y_{1\text{obs}}, Y_{2\text{obs}}, X) = \prod_{i\in I_1} \int_{b_{i2}} \int_{b_{i1}} \prod_{j=1}^{n_i} [\pi_{ij1}^{y_{ij1}}(1-\pi_{ij1})^{1-y_{ij1}} \times \pi_{ij2}^{y_{ij2}}(1-\pi_{ij2})^{1-y_{ij2}}]\phi_{1,2}(b_{i1}, b_{i2})db_{i1}db_{i2}.$$

where I_1 is the set of all households who were present in both seasons and Θ is the set of all parameters.

For the second part of the likelihood, there are some households that were only present in the first season with the corresponding set I_2:

$$L^{(2)}(\Theta_1|Y_{1\text{obs}}, X_1) = \prod_{i\in I_2} \int_{b_{i1}} \prod_{j=1}^{n_i} [\pi_{ij1}^{y_{ij1}}(1-\pi_{ij1})^{1-y_{ij1}}] \times \varphi_1(b_{i1})db_{i1}.$$

Where $\Theta_1 = (\alpha_{01}, \beta_1, \tau_1)$. Here it is obvious that there is no need to consider the remaining vector of parameters $\{\Theta - \Theta_1\}$ due to the

transitional nature of the model and that the missing mechanism is MAR.

The third part consists of households who only their summer response is observed with all individuals belonging to I_3:

$$L^{(3)}(\Theta|Y_{2\text{obs}}, X) = \prod_{i\in I_3}\prod_{j=1}^{n_i}\sum_{y_{ij1}=0}^{1}\{\int_{b_{i2}}\int_{b_{i1}}[\pi_{ij1}^{y_{ij1}}(1-\pi_{ij1})^{1-y_{ij1}} \times\pi_{ij2}^{y_{ij2}}(1-\pi_{ij2})^{1-y_{ij2}}]\phi_{1,2}(b_{i1},b_{i2})db_{i1}db_{i2}\}.$$

We have used the marginal distribution of the second response by summing over two possible outcomes of the first season to be able to find conditional probabilities of the second response.

From a frequentist point of view, product of all individuals' likelihood might be used to obtain parameter estimates using optimizing functions (for maximizing the overall likelihood or minimizing minus logarithm of the likelihood) available in softwares S-Plus or R (for example function 'optim' in R). However, considering independent diffuse normal priors for $(\alpha_{01}, \alpha_{02}, \beta_1', \beta_2', \gamma)$ with mean 0 and variance 1.0×10^6 and independent diffuse gamma priors for $(1/\tau_1, 1/\tau_2)$ with shape parameter 0.001 and scale parameter 0.001, we can base inference on the resulting posterior distribution. We will use Gibbs sampling along with the above decomposition to account for the missing problem as will be explained in the next section.

4.1 Computations based on Gibbs sampling

Posterior distribution of Θ is interactable; hence the inferential statistics on the parameters of interest can be constructed based on values drawn from the joint posterior distribution, $P(\Theta, b|Y_1, Y_2, X)$ obtained using Gibbs sampling. WinBUGS software (Ntzoufras, 2009; Spiegelhalter et al. 2003; Gilks et al. 1996) will be used to obtain the draws and to derive inferences on parameters of interest.

For the complete data inferences, Gibbs sampling (Gelfand and Smith, 1990) involves iteratively drawing from the known conditional distributions. Draws from $P(\beta, \gamma, b, \tau|Y, X)$ where $\beta = (\beta_1, \beta_2)$, $b = (b_1, b_2)$ and $\tau = (\tau_1, \tau_2)$, are generated by Gibss sampling based on the following conditioned distributions:

$$\begin{array}{ll}(i) & [\beta_1, b_1, \tau_1|X, Y_1] \\ & (i.1)\quad [\beta_1|b_1, \tau_1, X, Y_1]\end{array}$$

$$
\begin{array}{ll}
 & (i.2)\quad [b_1|\beta_1,\tau_1,X,Y_1] \\
 & (i.3)\quad [\tau_1|b_1,\beta_1,X,Y_1] \\
(ii) & [\beta_2,\gamma,b_2,\tau_2|Y_1,Y_2,X] \\
 & (ii.1)\quad [\beta_2|\gamma,b_2,\tau_2,Y_1,Y_2,X] \\
 & (ii.1)\quad [\gamma|\beta_2,b_2,\tau_2,Y_1,Y_2,X] \\
 & (ii.2)\quad [b_2|\beta_2,\gamma,\tau_2,Y_1,Y_2,X] \\
 & (ii.3)\quad [\tau_2|b_2,\beta_2,\gamma,Y_1,Y_2,X]
\end{array}
$$

In practical problems, however, not all of the conditional distributions are known or have closed form. In such cases, rejection sampling (Ripley, 1987), adaptive rejection sampling (Gilks and Wild, 1992), the Metropolis algorithm (Metropolis et al., 1953), or the Metropolis-Hastings algorithm (Hastings, 1970) are commonly used for drawing values from the distributions (Chib and Greenberg, 1995).

For inferences under ignorable missing data mechanism for the response variables in both seasons (non-monotone pattern) with disjoint parameter spaces in spring and summer, inferences can be made by drawing values from $P(\beta,b,\tau|Y_{obs},X)$ which is equivalent to drawing from $P(\beta,b,\tau,Y_{miss}|Y_{obs},X)$. These draws are obtained using Gibbs sampling based on the data augmentation algorithm (Tanner and Wong, 1987) implemented in the following conditional distributions:

$$
\begin{array}{ll}
(i) & [Y_{1,miss}|\beta_1,b_1,\tau_1,X] \\
(ii) & [\beta_1,b_1,\tau_1|Y_1,X] \\
 & (i.1)\quad [\beta_1|b_1,\tau_1,Y_1,X] \\
 & (i.2)\quad [b_1|\beta_1,\tau_1,Y_1,X] \\
 & (i.3)\quad [\tau_1|b_1,\beta_1,Y_1,X] \\
(i') & [Y_{2,miss}|\beta_2,\gamma,b_2,\tau_2,Y_1,X] \\
(ii') & [\beta_2,\gamma,b_2,\tau_2|Y_1,Y_2,X] \\
 & (ii'.1)\quad [\beta_2|\gamma,b_2,\tau_2,Y_1,Y_2,X] \\
 & (ii'.1)\quad [\gamma|\beta_2,b_2,\tau_2,Y_1,Y_2,X] \\
 & (ii'.2)\quad [b_2|\beta_2,\gamma,\tau_2,Y_1,Y_2,X] \\
 & (ii'.3)\quad [\tau_2|b_2,\beta_2,\gamma,Y_1,Y_2,X]
\end{array}
$$

The four blocks (i), (ii), (i') and (ii') represent an outer Gibbs sampling, from which draws from $P(\beta,b,\tau,Y_{miss}|Y_{obs},X)$ are obtained. Because drawing directly from blocks (ii) and (ii') is not feasible, an inner Gibbs sampling, as described for complete cases, was implemented within each block. Blocks (i) and (i') are consisted of the

I-step in data augmentation, with missing data generated to create a complete data set.

4.2 Results for labor force survey data

In this section, four different transitional models will be compared. The first two models only consider complete cases (22088 individuals who were present in both seasons) whereas the other two models work with all available cases. Model (I) does not assume any random effect parameters but the second model [Model (II)] includes random effects, due to household clustering. Both models are of the form (3) in Section 3, but excluding b_{it} for Model (I). Model (III) and Model (IV) are assumed to be transitional models for the available cases with non-monotone missing pattern as discussed in the previous section. Model (III) does not include random effect parameters but Model (IV) takes into account the existing correlation within household members by including random coefficients (similar to the model proposed in Section 3).

All the above Models are fitted using a Bayesian approach toward parameter estimations or data augmentation [if needed as in Model (III) and Model (IV)]. Using Bayesian approach, we have performed the iterative Gibbs sampling procedure in 10000 iterations, ignoring the first 5000 iterations as burn-in, we obtain inferences about the model parameters using 5000 remained iterations. We use the posterior mean of each parameter as its estimate and the sample standard deviation as the estimated standard deviation of the parameter of interest. Results for the proposed models are presented in Table (2) and Table (3).

According to the results of using complete cases given in Table (2), in both seasons, employment odds of the active part of population increases as the age grows by both Models (I) and (II) with an stronger effect in Model (II), fixing all other model covariates. In Model (I), the odds ratios of employment for men v.s. women are 1.002 and 1.176 respectively in spring and summer, and these values respectively increase to 1.039 and 1.249 by Model (II). The odds of employment for married people in Model (I) is 2.775 (2.519) times than that of singles in spring (summer). Model (II) indicates a higher employment odds ratio of 3.414 and 3.296 for married people v.s. single ones respectively in spring and summer. Both models show that people with master degree or PhD have more odds of employment

than illiterate ones in spring and summer. Based on Model (I), in spring (summer) the employment odds of households having 1 member is 1.846 (1.531) times than that of households having at least 4 members and these odds are respectively 2.003 and 1.84 by Model (II).

Table (2): Results for complete data; (parameter estimations highlighted in **Bold** are significant at 0.05 level and DIC stands for Deviance Information criterion)

		Model (I)		Model (II)	
Par.		Spring	Summer	Spring	Summer
		Est. (S. E.)	Est. (S. E.)	Est. (S. E.)	Est. (S. E.)
Intercept		**1.400** (0.232)	0.722 (0.367)	**1.766** (0.304)	0.583 (0.457)
Age		**0.041** (0.003)	**0.023** (0.004)	**0.051** (0.004)	**0.032** (0.005)
Gender	Female (ref.)	-	-	-	-
	Male	0.002 (0.066)	**0.162** (0.079)	0.039 (0.084)	**0.223** (0.104)
Marital statues	Single (ref.)	-	-	-	-
	Married	**1.021** (0.068)	**0.924** (0.084)	**1.228** (0.094)	**1.193** (0.114)
	Widowed	0.615 (0.369)	**1.710** (0.577)	0.684 (0.419)	**2.327** (0.712)
	Divorced	-0.215 (0.282)	0.261 (0.357)	-0.179 (0.357)	0.393 (0.466)
Educational degree	Other (ref.)	-	-	-	-
	illiterate	-0.252 (0.244)	**-1.010** (0.388)	-0.388 (0.281)	**-1.163** (0.478)
	Under diploma	-0.295 (0.222)	**-1.241** (0.361)	-0.293 (0.261)	**-1.340** (0.451)
	Diploma or upper	**-0.699** (0.224)	**-1.574** (0.363)	**-0.709** (0.263)	**-1.700** (0.460)
	Bachelor	**-0.874** (0.232)	**-1.573** (0.369)	**-0.948** (0.276)	**-1.677** (0.468)
	Master or PhD	0.918 (0.488)	-0.795 (0.519)	**1.168** (0.549)	-0.744 (0.654)
Family size	At least 4(ref.)	-	-	-	-
	1 member	0.613 (0.414)	0.426 (0.468)	0.695 (0.502)	0.612 (0.627)
	2 members	0.053 (0.114)	-0.104 (0.131)	0.116 (0.143)	-0.086 (0.172)
	3 members	0.143 (0.075)	-0.031 (0.086)	0.188 (0.094)	-0.040 (0.111)
Living Area	Rural (ref.)	-	-	-	-
	Urban	**-0.739** (0.059)	**-0.623** (0.069)	**-0.901** (0.083)	**-0.810** (0.097)
Effect of spring response		-	**2.740** (0.062)	-	**3.756** (0.153)
Variance of random effect		-	-	**2.203** (0.079)	**2.994** (0.053)
DIC		11402.900	8570.300	10670.100	7752.730

It is noteworthy that employment odds of people belonging to rural households is 2.096 (1.864) times than that of people in urban households in spring (summer) by Model (I) and these odds increase in Model (II). Also the summer employment odds of people who were employed in spring is 15.487 (42.777) times than that of others that were unemployed in spring based on Model (I) (Model (II)). The

overall comparison of Model (I) and Model (II) indicates that the parameter effects are stronger when the random coefficients are included in the model and that the variance of both random effects in summer and spring are highly significant in Model (II). Hence, for complete data, consideration of the existing correlation between household members (taking into account by random effects) has made the effects more obvious.

Table (3): Results for the available data (assuming MAR); (parameter estimations highlighted in **Bold** are significant at 0.05 level and DIC stands for Deviance Information criterion)

Par.		Model (III)		Model (IV)	
		Spring	Summer	Spring	Summer
		Est.(S. E)	Est.(S. E)	Est.(S. E)	Est.(S. E)
Intercept		**1.135** (0.018)	**0.851** (0.224)	**1.340** (0.161)	**1.029** (0.258)
Age		**0.038** (0.0003)	**0.026** (0.002)	**0.052** (0.003)	**0.036** (0.003)
Gender	Female (ref.)	-	-	-	-
	Male	**0.197** (0.013)	**0.427** (0.038)	**0.284** (0.052)	**0.604** (0.056)
Marital statues	Single (ref.)	-	-	-	-
	Married	**0.954** (0.015)	**1.086** (0.048)	**1.217** (0.057)	**1.355** (0.067)
	Widowed	**0.832** (0.208)	**0.684** (0.217)	**0.995** (0.279)	**0.909** (0.267)
	Divorced	**-0.466** (0.132)	0.169 (0.206)	-0.338 (0.229)	0.271 (0.269)
Educational degree	Other (ref.)	-	-	-	-
	illiterate	0.056 (0.041)	-0.166 (0.223)	-0.057 (0.157)	-0.289 (0.240)
	Under diploma	**-0.172** (0.009)	**-0.556** (0.210)	-0.149 (0.140)	**-0.604** (0.227)
	Diploma or upper	**-0.627** (0.014)	**-1.017** (0.210)	**-0.616** (0.142)	**-1.154** (0.227)
	Bachelor	**-0.702** (0.047)	**-1.055** (0.214)	**-0.686** (0.149)	**-1.184** (0.231)
	Master or PhD	**0.765** (0.126)	-0.054 (0.296)	**0.906** (0.305)	0.062 (0.348)
Family size	At least 4 (ref.)	-	-	-	-
	1 member	0.062 (0.101)	**0.885** (0.266)	0.224 (0.269)	**1.172** (0.346)
	2 members	**0.160** (0.046)	-0.048 (0.074)	0.079 (0.089)	0.008 (0.099)
	3 members	**0.168** (0.018)	0.003 (0.049)	**0.154** (0.059)	0.034 (0.066)
Living Area	Rural (ref.)	-	-	-	-
	Urban	**-0.767** (0.028)	**-0.757** (0.041)	**-0.914** (0.053)	**-0.938** (0.059)
Effect of spring response		-	**1.167** (0.044)	-	**1.550** (0.071)
Variance of random effect		-	-	**2.544** (0.024)	**2.878** (0.061)
DIC		29938.200	25920.900	27659.100	23640.000

Table (3) summarizes the results for models fitted to all available cases with the assumption of MAR mechanism. The interpretation of parameter estimates are nearly the same as Model (I) and Model (II) but these two models gain more precision due to consideration of

available observations and the use of data augmentation procedure to consider the special missing pattern of the data. In comparison with the complete data models, these two models lead to significant effects for all covariates except for some of the levels in multilevel covariates. It is also apparent that the significance of the effect of all covariates except for some of their levels are obtained in comparison with models for complete data. Also the effect of spring outcome on the employment statues in summer has been decreased in comparison with Model (I) and Model (II) due to data augmentation. In Model (IV) the parameter effects are stronger than Model (III).

According to the deviance information criterion (DIC; Spiegelhalter et al., 2002) calculated for all 4 models (see Table (2) and (3)), consideration of random effect parameters has reduced DIC in both models for complete and available data. Also, comparing Table (2) results with its corresponding results in Table (3) illustrates that using data augmentation method toward available data analysis and including random effects as in Model (IV) have reduced variance of parameter estimates and increased their significant level due to consideration of the potential correlation between household members, hence yielding more precision for analyzing these data. Hence, we would suggest Model (IV) as the best for the employment probability prediction and inference in these data.

5 Sensitivity Analysis

Statistical conclusions can be viewed as the end result of synthesis of the relevant information provided by the observed data and the prior information provided by the model which is usually a plausible, but necessarily imprecise, description of the actual process that generated the data. The discussions presented by Cook (1986) are based on the informal notion that important conclusions should not depend critically on the hypothesized model or unusual aspects of the data. If our conclusions do depend critically on the model or the data, there is surely cause for concern and knowledge of such dependence must become a part of conclusions. Otherwise our ignorance of the precise process that generated the data should do no harm. An obvious way to see if perturbations of the model influence key results of the analysis is to compare the results derived from the original and perturbed models using an influence graph.

In this section we access the sensitivity of the best model selected

in the pervious section (Model IV) around two key aspects of the assumed model. One the use of logit link function for both spring and summer response variables which for our data set leads to logistic models. Next the robustness of the posterior distribution to the selection of the prior distribution via sensitivity analysis.

5.1 Link Function

To analyze the dependence of binary response data on explanatory variables, the logistic transformation is probably the one most commonly used. However, it is tentative and therefore some consideration of adequacy is needed. If some non-logistic model gives a better fit it is important to discover this. An appealing and informative way to examine the accuracy of the logistic link is to construct extended models which include the logistic model as a special case. We use the following family of the asymmetric transformations introduced by Aranda-Ordaz (1981),

$$W(\theta) = \{(1-\theta)^{-\lambda} - 1\}/\lambda$$

where θ is the success probability for the binary response of interest. Now we assume that

$$log(W(\theta)) = X'\beta \tag{6}$$

where $X'\beta$ is some linear predictor for the model.

For $\lambda = 1$, (6) reduces to the logistic model, while when $\lambda \to 0$ the complementary log log model is obtained. Hence, these two models may be compared by means of a single parameter, λ.

The inverse of (6) takes the form

$$\begin{array}{rcll} \theta(X'\beta) & = & 1-(1+\lambda e^{X'\beta})^{-1/\lambda}, & if \lambda e^{X'\beta} > -1 \\ & & 1 \qquad\qquad\qquad\quad , & otherwise. \end{array}$$

Now the model (IV) for the labor force survey data changes to,

$$\begin{array}{rcl} log(\{(1-\pi_{ijt})^{-\lambda_t} - 1\}/\lambda_t) & = & \alpha_{0t} + X'_{ij}\beta_t + \gamma y_{ij,t-1} + b_{it}, \\ & & i = 1,\ldots,K; j = 1,\ldots,n_i; t = 1,2 \end{array}$$

The results of using a Bayesian approach with a uniform(0,1) prior distribution for λ_t and the same diffuse priors introduced before for all other parameters, show that the posterior means are $\hat{\lambda}_1 = 0.967$ $(S.E. = 0.031)$ and $\hat{\lambda}_2 = 0.963$ $(S.E. = 0.032)$. Hence, as $\hat{\lambda}_1 \approx \hat{\lambda}_2 \approx 1$ the logistic link is the appropriate one for these data.

5.2 Prior Distribution

Another important issue in the context of the approach we use is the robustness of the posterior distribution. We can assess how robust the posterior distribution is to the selection of the prior distribution via sensitivity analysis, in which we assess changes in the posterior distribution over different prior distributions. When prior information is available, sensitivity analysis focuses on the structure of the prior distribution; when noninformative priors are used, it focuses on how different choices of prior parameters may influence the posterior inference. When prior distributions are used, it is common practice to proceed to the implementation of sensitivity analysis with different values of the prior mean or variance.

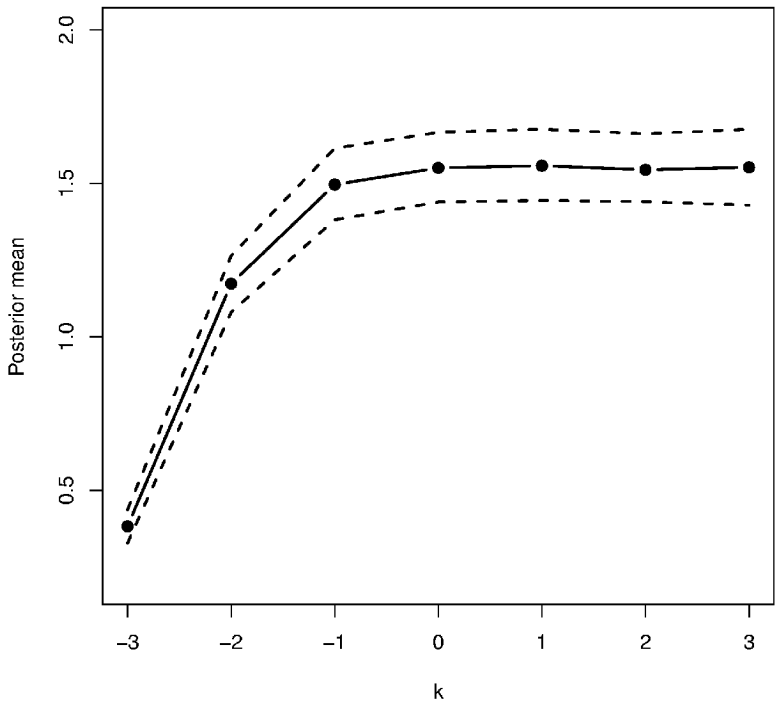

Figure (2): Sensitivity plot of posterior mean of γ for different values of k with prior mean equal to 1.55; dotted lines represent the 2.5% and 97.5% quantiles of the posterior distribution.

The most interesting parameter for us in the Model (IV) is the transition effect (γ) which shows the dependence of two response variables along two seasons. Hence, we examine the sensitivity of its posterior distribution to the perturbation of the mean and variance of the normal prior used in the model.

The sensitivity of the posterior mean of γ over different values of the prior variance σ^2 when mean equals to 1.55 (the estimation from Model (IV)), is depicted in Figure 2. For σ^2 we consider $\sigma^2 = 10^k$ for values of $k = -3, -2, -1, 0, 1, 2, 3$. From the graph we clearly see that the posterior mean is quite robust with values ranging from 1.545 to 1.556. Even for relatively low values of σ^2 (e.g., for $\sigma^2 = 1$), the estimate is equal to 1.552, which is quite close to the value of 1.55 resulted using large values of σ^2. Even if we consider as prior mean the extreme value of zero (which is far away from realistic values), the posterior mean of γ is still quite robust as is shown in Figure 3. Actually, for values of $\sigma^2 > 10$ the posterior mean ranges from 1.49 to 1.55, see Figure 3.

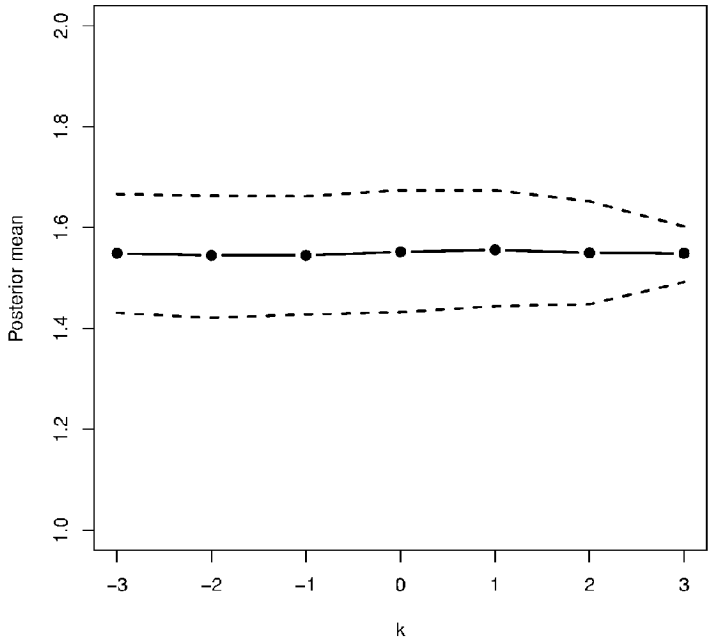

Figure (3): Sensitivity plot of posterior mean of γ for different values of k with prior mean equal to zero; dotted lines represent the 2.5% and 97.5% quantiles of the posterior distribution.

6 Conclusion

We use a transition logistic model with random coefficients for longitudinal binary response with MAR non-monotone missing pattern. The model is so flexible to be used with different distributions for the measurement error of the latent variable model. This model provides the ability to model not only monotone missing (dropout) but also non-monotone missing data for two period longitudinal studies. Gibbs sampling is used to obtain Bayesian parameter estimates. For labor force data (Statistical Center of Iran, 2006) we find a random effect transitional model as the best available. To access asymmetric departures from the logistic model, a general class of power transformations for the binary response variable was used which shows that the logistic link is appropriate. Also the sensitivity analysis of the posterior estimation of the transition parameter to the prior assumptions was performed which shows that our best model is nearly robust to the prior selection. Based on the results, we obtain that the response variable in summer is dependent on the spring response and that the correlation due to being a member of the same household is strongly significant. For further work the model can be extended to be used for longitudinal ordinal responses with non-ignorable missing mechanism.

Acknowledgements

This work was guided by the statistical research group of "Bayesian Inferential Statistics".

References

Agresti, A. (2002), Analysis of Categorical Data. New York: Wiley.

Anderson, T. W. and Goodman, LA. (1957), Statistical Inference about Markov Chains. Ann. Math. Stat, **28**, 89-110.

Aranda-Ordaz, F. J. (1981), On two families of transformations to additivity for binary response data. Biometrika, **68(2)**, 357-63.

Berridge, D. M. and Dos Santos, D. M. (1996), Fitting a random effects model to ordinal recurrent events using existing software.

J. Statist. Comput. Simul., **55**, 73-86.

Chib, S. and Greenberg, E. (1995), Understanding the Metropolis-Hastings Algorithm. The American Statistician, **49**, 327-335.

Cook, R. D. (1986), Assessment of local influence. Royal Statistical Society, **48**, 133-169.

Diggle, P. J. and Kenward, M. G. (1994), Informative Drop-out in longitudinal data analysis. Journal of Applied Statistics, **43**, 49-93.

Diggle, P. J., Heagerty, P., Liang, K. Y., and Zeger, S. L. (2002), Analysis of Longitudinal Data. Oxford: University Press.

Gelfand, A. E., and Smith, A. F. M. (1990), Sampling-Based Approaches to Calculate Marginal Densities. Journal of the American Statistical Association, **85**, 398-409.

Gilks, W. R., Richardson, D. B., and Spiegelhalter, D. J. (1996), Markov Chain Monte Carlo in Practice. London: Chapman & Hall.

Gilks, W. R. and Wild, P. (1992), Adaptive Rejection Sampling for Gibbs Sampling. Applied Statistics, **41**, 337-348.

Harvile, D. A., Mee, R. W. (1984), A mixed model procedure for analyzing ordered categorical data. Biometrics, **40**, 393-408.

Hastings, W. K. (1970), Monte Carlo Sampling Method Using Markov Chains and Their Applications. Biometrika, **57**, 97-109.

Kaciroti, N. A., Raghunathan, T. E., Schork, M. A., Clark, N. M., and Gong, M. (2006), A Bayesian Approach for Clustered Longitudinal Ordinal Outcome With Nonignorable Missing Data: Evaluation of an Asthma Education Program. Journal of the American Statistical Association, **474**, 435-446.

Liang, K. Y., Zeger, S. L., and Qaqish, B. F. (1992), Multivariate regression analyzes for categorical data (with discussion). J. Roy. Statist. Soc. B, **54**, 3-40.

Little, R. J. A. and Rubin, D. B. (2002), Statistical Analysis with Missing Data. New York: John Wily.

McCullagh, P. (1980), Regression models for ordinal data (with discussion). J. Roy. Statist. Soc. B, **42**, 109-142.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, **21**, 1087-1091.

Ntzoufras, I. (2009), Bayesian Modeling Using WinBUGS. John Wiley & Sons.

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

Ripley, B. (1987), Stochastic Simulation. New York: Wiley.

Sengul, T. K., Stoffer, D. S., and Day N. L. (2007), A residuals-based transition model for longitudinal analysis with estimation in the presence of missing data. Stat. Med., **26**, 3330-3341.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. (2002). Bayesian measures of model complexity and fit (with discussion). Journal of Royal Statistical Society B, **64**, 583-639.

Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003). Win- BUGS User Manual, Version 1.4, MRC Biostatistics Unit, available at www.mrc-bsu.cam.ac.uk/bugs.

Stiratelli, R., Laird, N., Ware, J. H. (1984), Random-effect models for serial observations with binary response. Biometrics, **40**, 961-971.

Tanner, M. and Wong, W. H. (1987), The Calculation of Posterior Distribution by Data Augmentation (with discussion). Journal of the American Statistical Association, **82**, 528-550.

Tutz, G. (2005), Modeling of repeated ordered measurements by isotonic sequential regression. Statist. Mod, **5**, 269-287.

Verbeke, G., Molenberghs, G. (1997), Linear mixed models in practice: A SAS Oriented Approach. Springer.