Journal of Applied Statistics
Vol. 00, No. 00, August 2009, 1–14

RESEARCH ARTICLE

Likelihood-based Approach for Analysis of Longitudinal Nominal Data using Marginalized Random Effects Models

Keunbaik Lee[a*] Sanggil Kang[b] Xuefeng Liu[c] Daekwan Seo[d]

[a]*Assistant Professor, Biostatistics Program, Louisiana State University-Health Sciences Center, New Orleans, LA 70122, USA*; [b]*Associate Professor, Department of Data Information, Sangji University, Wonju, Korea*; [c]*Assistant Professor, Department of Biostatistics and Epidemiology, College of Public Health, East Tennessee State University, Johnson City, TN 37614, USA* [d]*Visiting Fellow, Laboratory of Experimental Carcinogenesis, NCI/NIH Bethesda, MD 20892, USA*

(*Received 00 Month 200x; in final form 00 Month 200x*)

Likelihood-based marginalized models using random effects have become popular for analyzing longitudinal categorical data. These models permit direct interpretation of marginal mean parameters and characterize the serial dependence of longitudinal outcomes using random effects (Heagerty [1] and Lee and Daniels [2]). In this paper, we propose a model that expand the use of previous models to accommodate longitudinal nominal data. Random effects using a new covariance matrix with a Kronecker product composition are used to explain serial and categorical dependence. The Quasi-Newton algorithm is developed for estimation. These proposed methods are illustrated with a real dataset and compared with other standard methods.

Keywords: likelihood-based model; random effects; marginal model; Quasi-Newton; Kronecker product

*Corresponding author. Email: klee4@lsuhsc.edu

ISSN: 0233-1934 print/ISSN 1360-0532 online

DOI: 10.1080/0266476YYxxxxxxxx
http://www.informaworld.com

1. Introduction

Longitudinal data arise from repeated measurements on the same study subject periodically over time. This is different compared to the data from cross-sectional studies where each measurement is taken only once at a single point in time for each subject. The repeated observations for longitudinal data are typically correlated with one another across time [3]. Although the population-averaged effects of the covariates are usually of primary interest in the analysis of longitudinal data, the effect of serial correlation must also be taken into account for proper inference of the population-averaged effects of covariates in marginal models [4, 5].

Likelihoood-based marginalized models have recently been developed for the analysis of longitudinal categorical data [1, 2, 6–11]. These models permit direct interpretation of marginal mean parameters and characterize the serial dependence of longitudinal outcomes by using one of the following approaches: 1) random effects (marginalized random effects model (MREM)) [1, 2, 6], 2) a Markov structure (marginalized transition model (MTM)) [7–9], or 3) both random effects and Markov dependence [10, 11]. In this paper, we focus on the MREM and propose a model that extends the MREM to accommodate longitudinal nominal data.

Different models have been developed or proposed for the analysis of nominal data. Multinomial logit models were developed for the purpose of analyses of longitudinal or clustered nominal data [13–17]. A Bayesian two-level generalized logit model was proposed by Daniels and Gatsonis [18]. Revelt and Train [19] proposed discrete choice models with random coefficients for the explanatory variables which can vary according to the response category. Similar models were proposed in Hedeker [20] to describe clustered or longitudinal nominal response data. Hartzel et al. [21] surveyed mixed-effects models for both clustered ordinal and nominal responses. Chen et al. [22] proposed likelihood-based joint marginal and conditional models for longitudinal nominal data using a Markovian structure. Lee and Mercante [9] proposed nominal marginalized transition models using a Markovian structure and presented consistency and robustness of estimates of marginal mean parameters to misspecification of dependence models.

Heagerty [1] and Lee and Daniels [2] used the covariance matrix with AR(1) structure (Σ_1) to account for serial correlation of binary or ordinal responses, respectively. However this structure cannot be used for longitudinal nominal data directly because correlations among categories cannot be explained. In longitudinal nominal data analysis, two types of correlations among responses must be taken into account: correlations among categories at the same time and correlations among responses on the same subject over time. To account for serial and categorical dependence, we introduce normally distributed random effects with a covariance matrix with a Kronecker product composition. The covariance matrix (Σ) is specified as the Kronecker product of the covariance matrix accounting for serial dependence (Σ_1) and the correlation among categories at same time (Σ_2), that is, $\Sigma = \Sigma_1 \otimes \Sigma_2$ [6, 23, 24]. A key advantage of this specification lies in the ease of interpretation in terms of the independent contribution of two types of correlations (categorical and serial) to the overall within subject covariance matrix.

In this paper, we propose a marginalized random effects model [1, 2] to be used for longitudinal nominal data. The marginal model for marginal mean parameters is same to that in Lee and Mercante [9] and the dependence of responses is modeled 'separately' via random effects. We will also consider a Kronecker product covariance structure for the covariance matrix of responses to analyze longitudinal nominal data.

The paper is organized as follows. We describe the marginalized random effects models for longitudinal nominal data in Section 2. In Section 3, we conduct a simulation study to examine bias and efficiency of estimation of marginal mean parameters. Methods are illustrated by using health service data from the McKinney Homeless Research Project (MHRP) study in Section 4. Finally, a conclusion is provided in Section 5.

2. Marginalized Random Effects Models for Longitudinal Nominal Data

In this section, we describe our proposed model and the maximum likelihood method for it.

2.1 *Proposed Model*

If the designed completion time is denoted by T, we will have $n_i \leq T$ measures for each unit. Let $Y_i^{*T} = (Y_{i1}^*, \cdots, Y_{in_i}^*)$ be the vector of multinominal responses with K categories on subject $i = 1, \cdots, N$ and let $Y_{it} = (Y_{it1}, \cdots, Y_{itK})$ be indicator vectors for Y_{it}^* at times $t = 1, \cdots, n_i$ where $Y_{itk} = 1$ if $Y_{it}^* = k$ and $Y_{itk} = 0$ otherwise.

Let $x_{it}^T = (x_{it1}, \cdots, x_{itp})$ indicate covariates corresponding to $y_{it\cdot}$. In longitudinal nominal data, two types of response correlations are considered: 1) correlations among categories at the same time and 2) correlations due to serial dependence. We propose a Kronecker product of structured covariance matrices, that is, $\Sigma = \Sigma_1 \otimes \Sigma_2$, where Σ_1 and Σ_2 are, respectively, the covariance matrix for serial dependence and the covariance matrix among categories at same time [23, 24]. The marginalized random effects model for nominal data (NMREM) is given by

$$\log \frac{P(Y_{itk} = 1|x_{it})}{P(Y_{itK} = 1|x_{it})} = x_{it}^T \beta_k, \tag{1}$$

$$\log \frac{P(Y_{itk} = 1|b_{it}, x_{it})}{P(Y_{itK} = 1|b_{it}, x_{it})} = \triangle_{itk} + b_{itk}, \tag{2}$$

$$b_i \sim N(0, \Sigma), \tag{3}$$

where $b_i^T = (b_{i1}^T, \cdots, b_{in_i}^T) = (b_{i11}, \cdots, b_{i1K-1}, \cdots, b_{in_i1}, \cdots, b_{in_iK-1})$ and β_k is the $p \times 1$ vector of regression coefficients for $i = 1, \cdots, N$; $t = 1, \cdots, n_i$; $k = 1, \cdots, K-1$. Here, $\beta^T = (\beta_1^T, \cdots, \beta_{K-1}^T)$ is the vector of marginal mean parameters used in making inferences about the covariates effects. The parameters $\triangle_{itk}$ in (2) are the subject-time-category-specific intercept and function of both the marginal mean parameters, β and the dependence parameter, Σ.

From (1)-(3), we have the following relationship, for all i, t, and k,

$$P_{itk}^M = \int P_{itk}^c(b_{it})\phi(b_{it})db_{it}, \tag{4}$$

where $P_{itk}^M = P(Y_{it}^* = k|x_{it})$, $P_{itk}^c(b_{it}) = P(Y_{it}^* = k|b_{it}, x_{it})$ and $\phi(\cdot)$ is a multivariate normal distribution with mean 0 and variance matrix Σ_2. Given β and Σ_2, we calculate $\triangle_{it}$ from the relationship (4) using a Newton-Raphson algorithm. More detailed calculations are given in the Appendix.

The model as described above accounts for the longitudinal association of responses using random effects. The fixed effects parameters have both the subject-specific and marginal interpretations.

Heagerty [1] and Lee and Daniels [2] proposed an autoregressive covariance matrix to provide measures of random variation both across individuals and over time for longitudinal binary and ordinal data, respectively. We assume the variance-covariance matrix $\Sigma = \Sigma_1 \otimes \Sigma_2$ of b_i where Σ_1 is an autoregressive covariance structure

$$\Sigma_1 = \begin{pmatrix} 1 & e^{-\alpha} & e^{-2\alpha} & \cdots & e^{-(n_i-1)\alpha} \\ e^{-\alpha} & 1 & e^{-\alpha} & \cdots & e^{-(n_i-2)\alpha} \\ e^{-2\alpha} & e^{-\alpha} & 1 & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ e^{-(n_i-1)\alpha} & e^{-(n_i-2)\alpha} & e^{-(n_i-3)\alpha} & \cdots & 1 \end{pmatrix}, \tag{5}$$

and Σ_2 is a unstructured covariance matrix to explain correlation of categories as follows

$$\Sigma_2 = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1,K-1} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} & \cdots & \sigma_{2,K-1} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} & \cdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,K-1} & \sigma_{2,K-1} & \sigma_{3,K-1} & \cdots & \sigma_{K-1,K-1} \end{pmatrix}. \tag{6}$$

Galecki [23] noted that a lack of identifiability can result with Kronecker product covariance structure. The indeterminacy stems from the fact that Σ_1 and Σ_2 are not unique since for $c \neq 0$, $c\Sigma_1 \otimes (1/c)\Sigma_2 = \Sigma_1 \otimes \Sigma_2$. However, this nonidentifiability can be resolved using Σ_1 as a correlation matrix. This same covariance structure was used in the models for bivariate longitudinal binary data in Lee et al. [6].

The regression structure is characterized using a generalized linear model for the marginal probabilities that incorporates random effects to account for serial dependence and the correlation among categories. Also, the conditional model accounts for the longitudinal association from α in Σ_1. The marginal parameters have both subject-specific and marginal interpretations, as is typical in linear mixed models [3]. The advantage in the MREM approach is the ability to use conditional mean models for association while preserving the ability to structure the marginal mean directly, using the regression model. So, the interpretation of regression coefficients, β, does not depend on the specification of the dependence model. For longitudinal data analysis with random effects b_i, the marginal probability captures the systematic variation that is due to x_{it}, whereas parameters in $cov(b_i)$ provide a measure of random variation of categories at the same time (Σ_2) and over time ($\Sigma_2 e^{\alpha|t-t'|}$). Heagerty and Kurland [25] investigated the robustness of regression coefficients estimates to incorrect assumptions regarding the random effects in generalized linear mixed models and MREMs. They found that the MREMs are much less susceptible than generalized linear mixed models to bias resulting from random effects model misspecification. Lee and Daniels [6, 8] also examined robustness of marginal parameter estimates to misspecification of the dependence model.

To deal with the random effects from a computational perspective, we transform the random effects

$$b_i = \left\{ \Sigma_1^{\frac{1}{2}} \otimes \Sigma_2{}^{\frac{1}{2}} \right\} a_i,$$

where a_i is a $n_i(K-1) \times 1$ vector of independent standard normals, and $\Sigma_1{}^{\frac{1}{2}}$

and $\Sigma_2^{\frac{1}{2}}$ are lower triangular matrices with positive diagonal elements and the Cholesky factor of the $n_i \times n_i$ matrix Σ_1 and that of the $(K-1) \times (K-1)$ matrix Σ_2, respectively. Computationally, it is convenient to standardize the random effects and estimate the Cholesky elements of Σ (Hedeker and Gibbons, 1994). The reparameterized conditional model is then given by

$$\log \frac{P(Y_{itk} = 1|a_i, x_{it})}{P(Y_{itK} = 1|a_i, x_{it})} = \triangle_{itk} + s_k^{(t)} a_i,$$
$$a_i \sim N(0, I),$$

where $s_k^{(t)}$ is the $((t-1)(K-1)+k)$th row vector of $\Sigma^{\frac{1}{2}}$ and I is the identity matrix of order $n_i(K-1)$. This transformation allows us to estimate the Cholesky factor $\Sigma^{\frac{1}{2}}$, which is a lower-triangular matrix, instead of the covariance matrix Σ. As the Cholesky factor is the square root of the covariance matrix, this then allows more stable estimation of near-zero variance terms [26].

2.2 *Maximum Likelihood Estimation*

Now the maximum likelihood method for these models is described. Since the random effects are latent or unobserved, to obtain the likelihood function we construct the usual product of multinominals that would apply if they were known and then integrate out the random effects. This integral does not have closed form and this necessitates the use of some approximation for the likelihood function (e.g., marginalized likelihood function). We can then maximize the marginalized likelihood using a variety of standard methods. There are several algorithmic approaches to integrate the random effects. Gauss-Hermite quadrature is popular for simple models such as random intercept models [26]. To increase efficiency of Gauss-Hermite quadrature, an adaptive version of Gauss-Hermite quadrature [27, 28] is used. It centers the nodes with respect to the mode of the function being integrated and scales them according to the estimated curvature at the mode. Monte Carlo (MC) methods are used for models with higher-dimensional integrals. The MC methods use the randomly sampled nodes to approximate integrals. However, locating points at random does not guarantee an optimal distribution of the points. That is, the points may not be distributed exactly uniformly because of the sampling error.

An alternative to both the Gauss-Hermite and Monte Carlo methods is the 'Quasi-Monte Carlo (QMC) method' [29–31]. QMC works like MC but instead of using a uniformly and randomly distributed set of points it utilizes uniformly distributed deterministic sequences, called low discrepancy sequences [29]. The strength of the QMC method is that the distribution of points is optimal [32]. In particular, for high dimensional correlated data, the QMC method apparently outperforms both the Gauss-Hermite and Monte Carlo methods [32]. In our model, we use the QMC method to evaluate the integral in (7) due to high dimensionality of a_i. The NMREM marginal likelihood function is given by,

$$L(\theta; y) = \prod_{i=1}^{N} \int \prod_{t=1}^{n_i} \prod_{k=1}^{K} (P_{itk}^c(a_i))^{y_{itk}} \phi(a_i) da_i. \tag{7}$$

where $\theta = (\beta, \gamma, \alpha)$, γ is nonzero elements of $\Sigma_2^{1/2}$, and $y_{itk} = 1$ if $Y_{it}^* = k$ and 0 otherwise. We propose a Quasi-Newton algorithm to find the MLE of the parameters of interest in the MREM.

For estimation of the p covariate coefficients β_j for $j = 1, \cdots, K-1$ and (γ, α), we need to evaluate $\log L(\theta; y)$. We obtain

$$\begin{aligned}\log L(\theta; y_i) &= \log \int \prod_{t=1}^{n_i} \prod_{k=1}^{K} \{P_{itk}^c(a_i)\}^{y_{itk}} \phi(a_i) da_i \\ &= \log \int \exp\left[\sum_{t=1}^{n_i} \left\{ \sum_{k=1}^{K-1} y_{itk}(\triangle_{itk} + s_k^{(t)} a_i) + \log P_{itK}^c(a_i) \right\}\right] \phi(a_i) da_i.\end{aligned}$$

The marginal likelihood for a sample of N independent subjects is given by $L(\theta; y) = \prod_{i=1}^{N} L(\theta; y_i)$. Maximizing the log-likelihood, $\log L$, with respect to θ yields the likelihood equation

$$\frac{\partial \log L(\theta; y)}{\partial \theta} = \sum_{i=1}^{N} L^{-1}(\theta; y_i) \int \frac{\partial L(\theta, a_i; y_i)}{\partial \theta} \phi(a_i) da_i = 0$$

where

$$\begin{aligned}L(\theta, a_i; y_i) &= \prod_{t=1}^{n_i} \prod_{k=1}^{K} \{P_{itk}^c(a_i)\}^{y_{itk}} \qquad (8)\\ &= \exp\left[\sum_{t=1}^{n_i} \left\{ \sum_{k=1}^{K-1} y_{itk}(\triangle_{itk} + s_k^{(t)} a_i) + \log P_{itK}^c(a_i) \right\}\right].\end{aligned}$$

The $(K-1)p + (K-1)K/2 + 1$-dimensional likelihood equations are given in the Appendix.

The matrix of second derivatives of the observed data log-likelihood has a very complicated form. One alternative is to base an approximation on the observed information. The sample empirical covariance matrix of the individual scores in any correctly specified model is a consistent estimator of the information matrix and involves only the first derivatives. The Quasi-Newton method can be used to solve the likelihood equations, using

$$\theta^{(n+1)} = \theta^{(n)} + \left[I_e\left(\theta^{(n)}; y\right)\right]^{-1} \frac{\partial \log L}{\partial \theta^{(n)}},$$

where $I_e(\theta)$, an empirical and consistent estimator of the information matrix at step n, is given by

$$I_e(\theta; y) = \sum_{i=1}^{N} \frac{\partial L(\theta; y_i)}{\partial \theta} \frac{\partial L(\theta; y_i)}{\partial \theta^T}.$$

At convergence, the large-sample variance-covariance matrix of the parameter estimates is then obtained as the inverse of the information matrix.

For the explicit form of the Quasi-Newton algorithm and the detailed derivatives calculations (using (4); see the Appendix).

3. Simulations

We conducted several simulation studies to examine the bias in estimating the marginal mean parameters in the setting of misspecification of the dependence model under no missing data and under MAR missingness. We simulated longitudinal nominal data under an NMREM. Covariates were time and group (2 levels). The marginal probabilities for NMREM were specified as

$$\log\left(\frac{P(Y_{it}^* = k)}{P(Y_{it}^* = 3)}\right) = \beta_{k0} + \beta_{k1}\text{time}_{it} + \beta_{k2}\text{group}_i + \beta_{k3}\cdot\text{time}_{it} * \text{group}_i,$$
$$\beta_1 = (\beta_{10}, \beta_{11}, \beta_{12}, \beta_{13}) = (0.1, 0.2, 0.5, 0.7),$$
$$\beta_2 = (\beta_{20}, \beta_{21}, \beta_{22}, \beta_{23}) = (0.2, 0.1, 0.4, 0.5),$$

where $\text{time}_{it} = (t-1)/8$ for $t = 1, \cdots, 8$ and $\text{group}_i = 0$ or 1 such that the sample size per group was approximately equal for all times t. The conditional probabilities were based on (2) and (3) with $s = (s_{11}, s_{21}, s_{22}) = (1.5, 0.1, 1.4)$ and $\alpha = 0.3$ (correlation= $\exp(-0.3) = 0.74$). Note that s are nonzero elements of $\Sigma_2^{1/2}$.

We generated 200 simulated data sets with sample sizes of 200 and 300. We fit 2 models. Model 1 is the true model; Model 2 is special case of Model 1 with $b_{it} = b_{i0} \sim N(0, \Sigma_2)$ (ignoring the temporal dependence).

For missingness, we specified the following MAR dropout model,

$$\text{logit}P(\text{dropout} = t|\text{dropout} \geq t) = \begin{cases} -2.0 + 0.1Y_{it-12} + 0.2Y_{it-13}, \text{ group}_i = 1; \\ -1.5 + 0.2Y_{it-11} + 0.1Y_{it-12}, \text{ group}_i = 0. \end{cases}$$

Based on this specification, the observed dropout rates were approximately 40 percent.

Table 1 presents mean estimates, the root mean squared error (RMSE), and the 95% Monte Carlo error intervals of the parameters when there was no missing data and the two sample sizes of 200 and 300 for each scenario. The estimates were essentially unbiased with very similar RMSE's for Models 1 and 2. The 95% Monte Carlo error intervals contained the true values for the parameters. Thus, the estimates of marginal mean parameters in our proposed model were robust to misspecification of the dependence model.

In the presence of MAR dropout, we saw considerable bias when using the sample size of 200, with percentage relative bias as large as 32% in the Models 1 and 2 for the coefficients of interaction of group and time (β_{23}) (see Table 2). Monte Carlo error intervals for these estimated coefficients did not contain the true values. This bias occurred because of small sample size. With the sample size of 300, the estimates were essentially unbiased and the 95% Monte Carlo error intervals contained the true values.

Overall, the simulation results indicate that marginal mean parameter estimates were robust when the dependence model was incorrectly specified for complete data but not for incomplete data. However, as sample size increased, the marginal mean parameter estimates were also robust to the misspecification of dependence model.

Table 1. Bias of maximum likelihood estimators under complete data sets. Displayed are the average regression coefficient estimates and the root mean squared error (RMSE=$\sqrt{\sum(\hat{\beta}-\beta)^2/M}$) where M is the number of simulated data sets. 95% Monte Carlo error intervals ($\bar{\beta} \pm 1.96\sqrt{var(\bar{\beta})/M}$) in parentheses.

		$N = 200$				$N = 300$			
		Model 1		Model 2		model 1		model 2	
para.	truth	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
β_{10}	1.0	0.11	0.19	0.10	0.19	0.11	0.16	0.10	0.17
		(0.08,0.13)		(0.08,0.13)		(0.09,0.13)		(0.08,0.13)	
β_{11}	0.2	0.18	0.41	0.19	0.41	0.20	0.37	0.20	0.38
		(0.13,0.24)		(0.13,0.24)		(0.15,0.25)		(0.15,0.25)	
β_{12}	0.5	0.48	0.27	0.49	0.27	0.50	0.23	0.50	0.24
		(0.45,0.52)		(0.45,0.53)		(0.46,0.53)		(0.47,0.53)	
β_{13}	0.7	0.71	0.61	0.72	0.62	0.72	0.57	0.72	0.57
		(0.63,0.80)		(0.63,0.81)		(0.64,0.80)		(0.64,0.80)	
β_{20}	0.2	0.22	0.19	0.22	0.19	0.21	0.14	0.21	0.14
		(0.20,0.25)		(0.20,0.25)		(0.19,0.23)		(0.19,0.22)	
β_{21}	0.1	0.06	0.44	0.06	0.44	0.08	0.33	0.09	0.33
		(-0.00,0.12)		(0.00,0.12)		(0.03,0.12)		(0.04,0.13)	
β_{22}	0.4	0.38	0.28	0.38	0.27	0.40	0.22	0.40	0.22
		(0.34,0.42)		(0.34,0.41)		(0.36,0.43)		(0.37,0.43)	
β_{23}	0.5	0.50	0.63	0.52	0.62	0.51	0.53	0.51	0.53
		(0.42,0.59)		(0.44,0.61)		(0.44,0.59)		(0.44,0.59)	

Table 2. Bias of maximum likelihood estimators under MAR missingness. Displayed are the average regression coefficient estimates and the root mean squared error (RMSE=$\sqrt{\sum(\hat{\beta}-\beta)^2/M}$) where M is the number of simulated data sets. 95% Monte Carlo error intervals ($\bar{\beta} \pm 1.96\sqrt{var(\bar{\beta})/M}$) in parentheses.

		$N = 200$				$N = 300$			
		Model 1		Model 2		model 1		model 2	
para.	truth	Mean	RMSE	Mean	RMSE	Mean	RMSE	Mean	RMSE
β_{10}	1.0	0.13	0.20	0.13	0.20	0.11	0.17	0.10	0.17
		(0.10,0.16)		(0.10,0.15)		(0.08,0.13)		(0.08,0.13)	
β_{11}	0.2	0.19	0.66	0.18	0.66	0.19	0.52	0.19	0.52
		(0.10,0.28)		(0.09,0.28)		(0.12,0.26)		(0.12,0.26)	
β_{12}	0.5	0.46	0.31	0.46	0.31	0.51	0.25	0.51	0.25
		(0.42,0.50)		(0.42,0.51)		(0.47,0.54)		(0.47,0.54)	
β_{13}	0.7	0.70	0.95	0.71	0.95	0.73	0.71	0.73	0.72
		(0.57,0.83)		(0.58,0.84)		(0.63,0.83)		(0.63,0.83)	
β_{20}	0.2	0.20	0.18	0.19	0.19	0.21	0.18	0.21	0.18
		(0.17,0.22)		(0.17,0.22)		(0.19,0.24)		(0.18,0.23)	
β_{21}	0.1	0.14	0.62	0.14	0.64	0.04	0.55	0.05	0.55
		(0.05,0.22)		(0.06,0.23)		(-0.03,0.12)		(-0.02,0.13)	
β_{22}	0.4	0.42	0.27	0.42	0.27	0.39	0.25	0.40	0.25
		(0.38,0.46)		(0.38,0.46)		(0.36,0.43)		(0.36,0.43)	
β_{23}	0.5	0.34	0.87	0.34	0.89	0.58	0.75	0.59	0.75
		(0.22,0.46)		(0.22,0.46)		(0.48,0.69)		(0.49,0.70)	

4. Example

4.1 *Description of Data*

Data from the McKinney Homeless Research Project (MHRP), a longitudinal study employing a randomized factorial design [33, 34], were analyzed to assess whether the use of Section 8 housing certificates effectively provided housing options for homeless individuals with severe mental illness. The 361 clients in this data set were randomly assigned to either comprehensive or traditional supportive case management, and to either of two levels of access to independent housing using Section 8 certificates. Hedecker [20] examined the effect of acess to Section 8 certificates on repeated housing status outcomes across time using mixed-effects multinomial regression models.

Housing status outcomes were nominal-level and were determined on each individual at four different times: baseline, and at 6, 12, and 24 months of follow-up. There were three different outcomes: 1) streets/shelters, 2) community housing, and 3) independent housing. As analyzed in Hedecker [20], we focus on assessing the effect of whether there is access to Section 8 certificates (yes or no) on housing status outcomes over time in a homeless mentally ill population using NMREM. For our analyses, the housing status outcome of street/shelter was chosen as the reference category.

About 25 percent of the subjects dropped out of the study during the follow-up period resulting in some missing housing outcome status data. Since estimation of model parameters is based on a full-likelihood approach, the missing data are assumed to be ‘ignorable’ conditional on both the model covariates and the observed nominal responses [3]. To focus on describing our proposed model, we assume the missing data are ignorable in our analysis.

4.2 *Computations*

Implementing the Quasi-Newton algorithm in this setting is computationally intensive because estimates and derivatives of Δ_{itk} require the Newton-Raphson algorithm for all subjects at all times within each Fisher-scoring step. We used simultaneously R 2.8.1 software (http://www.r-project.org) and FORTRAN 77. The R software was used for the Quasi-Newton iteration and FORTRAN 77 was used to make subroutines (.dll files) to implement the calculation of Δ using the Newton-Raphson and of derivatives of Δ. The R code and the Fortran .dll files are available upon request. Each Quasi-Newton step for the NMREM required approximately 40 seconds. To minimize the number of iterations until convergence, we used good initial values obtained by fitting an independent nominal logit model. For example, in our analysis below, we obtained convergence in about 50 iterations using a fairly strict convergence criterion,

$$\sqrt{\left(\hat{\theta}^{old} - \hat{\theta}^{new}\right)^T \left(\hat{\theta}^{old} - \hat{\theta}^{new}\right)} \leq 10^{-4}$$

where $\hat{\theta}^{new}$ and $\hat{\theta}^{old}$ are current and previous fitted values of parameters, respectively.

4.3 *Model Fits*

We fit four models under an assumption of ignorable dropout. Model 1 was NMREM with Σ_1 and Σ_2 as in (5) and (6) and Model 2 was a special case of Model 1 with only Σ_2. Model 3 was an independent multinomial model. Model 4 was a mixed-effects multinomial logistic regression model that was proposed in Hedeker (2003).

Maximized loglikelihood and AIC for Models 1-3 are given in Table 3 and those for Model 4 are -1109.365 and 2254.730, respectively. The likelihood ratio test for the comparison of Model 3 and Model 2 indicated that Model 2 fits much better than Model 3 ($\triangle D_{23} = 2 \times (1176.523 - 1090.551) = 171.944$, $p-$value$= 0.000$ on 3 d.f.). To compare the fit of the two models (Models 1 and 2), we computed the likelihood ratio test. Comparison of deviances for Models 1 and 2 which were nested yielded $\triangle D_{12} = 2 \times (1090.551 - 1078.677) = 23.748$, $p-$value< 0.001 on 1 d.f. This comparison indicated that Model 1 provided a significantly better fit than Model 2. Using a penalized model selection criterion (AIC) indicated Model 1 provided a better fit than Model 4 (2197.354 and 2254.730 for Model 1 and 4, respectively). These comparisons indicated that the Model 1 fit best among the four models. Compared with the best marginalized transition model (MTM) in Lee and Mercante (2010), the AIC for the MTM was 2194.690 which is very similar to that for Model 1 (2197.354) in this paper. This means that the MREM is a good alternative to the MTM for analysis of this data set.

Since our main interest was marginal relationship between marginal mean and covariates, we compared estimates of marginal mean parameters from Models 1, 2, and 3. Table 3 presents maximum likelihood estimates of marginal mean parameters. Results are organized into two parts based on which nominal housing status outcome responses are being compared: either 1) street/shelters compared to community or 2) street/shelters compared to independent housing.

Table 3. Maximum likelihood estimates for NMREMs (Models 1 and 2) and independent multinomial logit model (Model 3). Standard errors are in parentheses.

	Model 1	Model 2	Model 3
Community vs Street/Shelter			
Intercept	−0.52* (0.17)	−0.50* (0.16)	−0.49* (0.16)
6 month vs. baseline	0.43 (0.23)	0.43 (0.23)	0.43 (0.23)
12 month vs. baseline	1.59* (0.21)	1.62* (0.23)	1.63* (0.27)
24 month vs. baseline	2.35* (0.33)	2.32* (0.34)	2.37* (0.34)
Section 8 (yes=1,no=0)	1.70* (0.31)	1.71* (0.29)	1.79* (0.31)
Section 8 by 6 month	−0.32 (0.35)	−0.36 (0.39)	−0.46 (0.43)
Section 8 by 12 month	−2.10* (0.51)	−2.06* (0.49)	−2.12* (0.49)
Section 8 by 24 month	−0.92 (0.46)	−0.94 (0.43)	−1.09* (0.45)
Independent vs Street/Shelter			
Intercept	−1.52* (0.24)	−1.64* (0.25)	−1.66* (0.25)
6 month vs. baseline	0.40 (0.32)	0.52 (0.33)	0.54 (0.34)
12 month vs. baseline	1.73* (0.25)	1.84* (0.29)	1.90* (0.35)
24 month vs. baseline	2.75* (0.37)	2.88* (0.38)	2.97* (0.40)
Section 8 (yes=1,no=0)	2.60* (0.34)	2.75* (0.32)	2.88* (0.37)
Section 8 by 6 month	1.32* (0.38)	1.24* (0.43)	1.13* (0.50)
Section 8 by 12 month	−0.09 (0.48)	−0.02 (0.48)	−0.04 (0.52)
Section 8 by 24 month	0.22 (0.47)	0.05 (0.43)	−0.07 (0.50)
s_{11}	4.97 (2.85)	1.65* (0.43)	
s_{21}	6.87 (3.94)	1.76* (0.20)	
s_{22}	4.30* (2.08)	1.66* (0.18)	
α	0.41* (0.09)		
Max. loglik.	−1078.677	−1090.551	−1176.523
AIC	2197.354	2219.102	2385.046

* indicates significance at 5% level.

Point estimates and standard errors for marginal mean parameters for Models 1, 2, and 3 were similar. Now we focus on Model 1 which was the best fit among four models. The first part of the Table 3 presents estimates of coefficients and corresponding standard errors for comparing the two nominal response categories of community and street/shelter (reference category).

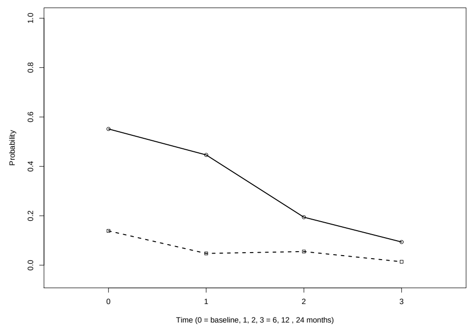
$$\log \frac{P(\text{Community})}{P(\text{Street/Shelter})} = -0.52 + 0.43t_1 + 1.59t_2 + 2.35t_3 + 1.70\text{Section 8}$$
$$-0.32t_1\text{Section 8} - 2.10t_2\text{Section 8} - 0.92t_3\text{Section 8},$$

where t_1, t_2, and t_3 are indicators for 6, 12, and 24 months, respectively. The regression coefficients of the indicator variables for the association between Section 8 certificate status and 12 and 24 months of follow-up were statistically significant. The regression coefficient of the interaction of 12 month and Section 8 certificate was also significant. Combining the logit estimates for the main effect of Section 8 certificate (1.70) and the interaction of 12 month and Section 8 certificate (-2.10) yielded an estimated odds ratio (OR) of $e^{-0.40} = 0.670$, suggesting individuals with access to Section 8 certificates were less likely to be in community housing as opposed to street/shelter housing at 12 month.

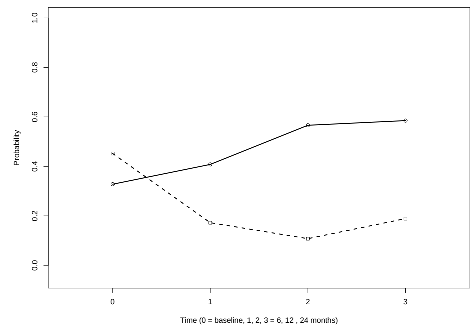
In the lower part of Table 3, independent housing is compared to the reference category, street/shelter.

$$\log \frac{P(\text{Independent})}{P(\text{Street/Shelter})} = -1.52 + 0.40t_1 + 1.73t_2 + 2.75t_3 + 2.60\text{Section 8}$$
$$1.32t_1\text{Section 8} - 0.09t_2\text{Section 8} + 0.22t_3\text{Section 8},$$

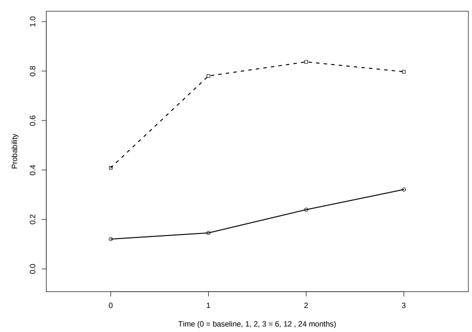
The regression coefficients of the indicator variables for 12, 24 months, and Section 8 certificate status were significant and the regression coefficients of the interaction of 6 month and Section 8 certificate was also significant. Combining the logit estimates for Section 8 (2.60) and the interaction of 6 month and Section 8 certificate (1.32) yielded an estimated odds ratio of $e^{3.92} = 50.400$ suggesting individuals with Section 8 certificates were much more likely to be in independent housing in



(a) Street/shelter housing



(b) Community housing



(c) Independent housing

Figure 1. Maximum likelihood estimates of marginal probabilities for street/shelter, community, and independent housings, respectively. Solid line is for control group; Dashed line is for Section 8 group.

6 month than in street/shelter housing.

The estimates of correlation parameter (α) was significant and corresponded to an estimated correlation of $\hat{\rho} = \exp(-\alpha) = \exp(-0.41) = 0.66$. The estimated values for s_{11}, s_{21}, and s_{22} in Model 1 were relatively large compared with those in Model 2 because the number of replications was small ($T = 4$). MLEs of the marginal probabilities of three housing types are given by Figures 1(a)-1(c). Each figure compares the MLEs of the marginal probabilities for the Section 8 vs. control groups. In the community housing, the two estimated marginal probabilities were different as month increased. Whereas, the difference of the two estimated marginal probabilities in the street/shelter housing decreased as month increased. However, we know that there were large differences between the Section 8 and control groups in independent housing. The estimated marginal probabilities for community with section 8 group was higher than those with control group. However, the estimated marginal probabilities with control group was higher than those with section 8 in the street/shelter and community housings.

5. Conclusion

In this paper the use of marginalized random effects models to analyze longitudinal nominal data were proposed. Marginal probabilities as a function of covariates were used for modeling the average effects of the covariates while dependence probabilities accounted for the longitudinal and categorical correlations via random effects. To explain these correlations, we used a covariance matrix with a Kronecker product composition which accounted for serial dependence and correlation among categories at same time.

The proposed models were implemented using a likelihood approach for parameter estimations. Quasi-Monte Carlo methods were used in likelihood estimations and in calculation of Δ to numerically integrate over the distribution of random effects. Simulation studies indicated that marginal mean parameter estimates were robust when the dependence model was incorrectly specified for complete data and for incomplete data with large sample size.

In the NMREM, the number of variance parameters in Σ_2 increases exponentially with the number of categories, K. Therefore, we can apply the NMREM to the data with moderate or large sample size when K is moderately large. Alternatively, we consider a simple structure of Σ_2 with equal elements of covariances.

Computing time is major potential problem with the use of NMREM. Due to using Newton-Raphson for the calculation of Δ, it takes considerable time for the calculations with large sample sizes. However, with our example data, our algorithm took a more reasonable time for the calculations.

Acknowledgements

The authors would like to thank Dr. James L Anderson from the Department of Biostatistics and Epidemiology at East Tennessee State University for his suggestions and inputs. We also gratefully acknowledge three referees for their helpful comments on this paper.

References

[1] Heagerty PJ. Marginally specified logistic-normal models for longitudinal binary data. *Biometrics.* 1999; **55**: 688–698.
[2] Lee K, Daniels M. Marginalized Models for Longitudinal Ordinal Data with Application to Quality of Life Studies. *Statistics in Medicine.* 2008; **27**: 4359–4380.
[3] Diggle PJ, Heagerty P, Liang KY, Zeger S. *Analysis of Longitudinal Data.* 2002; Oxford University Press.
[4] Liang KY, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika.* 1986; **73**: 13–22.
[5] Fitzmaurice GM, Laird NM. A likelihood-based method for analysing longitudinal binary responses. *Biometrika.* 1993; **80**: 141–151.
[6] Lee K, Joo Y, Yoo JK Lee J. Marginalized random effects models for multivariate longitudinal binary data. *Statistics in Medicine.* 2009; **28**: 1284–1300.
[7] Heagerty PJ. Marginalized Transition Models and Likelihood Inference for Longitudinal Categorical Data. *Biometrics.* 2002; **58**: 342–351.
[8] Lee K Daniels M. A class of markov models for longitudinal ordinal data. *Biometrics.* 2007; **63**: 1060–1067.
[9] Lee K, Mercante D. Longitudinal nominal data analysis using marginalized models. 2010; **54**: 208–218.
[10] Miglioretti D, Heagerty PJ. Marginal Modeling of Multilevel Binary Data with Time-Varying Covariates. *Biostatistics.* 2004; **5**: 381–398.
[11] Schildcrout J, Heagerty PJ. Marginalized models for moderate to long series of longitudinal binary response data. *Biometrics.* 2007; **63**: 322-331.
[12] Heagerty PJ, Zeger SL. Marginalized Multilevel models and likelihood inference (with discussion). *Statistical Science.* 2000; **15**: 1–26.
[13] Bock RD. Estimating multnomial response relations. *Contributions to Statistics and Probability.* 1970; ed. R.C. Bose. Chapel Hill, NC: University of North Carolina Press.
[14] Haberman SJ. *The Analysis of Frequency Data.* 1974; Chicago: University of Chicago Press.
[15] Nerlove M, Press S.J. Univariate and multivariate log-linear and logistic models. *Technical Report.* 1973; R-1306-EDA/NIH, Rand Corporation, Santa Monica, CA.
[16] Theil H. A multinomial extension of the linear logit model. *International economic review.* 1969; **10**: 251–259.
[17] Theil H. On the estimation of relationships involving qualitative variables. *American Journal of Sociology.* 1970; **76**: 103–154.
[18] Daniels M, Gatsonis C. Hierarchical polytomous regression models with applications to health services research. *Statistics in Medicine.* 1997; **16**: 2311–2325.
[19] Revelt D, Train K. Mixed logit with repeated choices : household choices of appliance efficiency level. *Review of Economics and Statistics.* 1998: **80**: 647–657.
[20] Hedeker D. A mixed-effects multinomial logistic regression model. *Statistics in medicine.* 2003: **23**: 1433–1446.
[21] Hartzel J, Agresti A, Caffo B. Multinomial logit random effects models. *Statistical modelling.* 2001; **1**: 81–102.

[22] Chen B, Yi GY, Cook RJ. Likelihood analysis of joint marginal and conditional models for longitudinal categorical data. *Canadian Journal of Statistics.* 2009; **37**, 182–205.

[23] Galecki, A. T. (1994). General class of covariance structures for two or more repeated factors in longitudinal data analysis. *Communication in Statistics-Theory and Method.*, 23, 3105-3119.

[24] Chaganty N.R. and Naik D. N. (2002) Analysis of multivariate longitudinal data using quasi-least squares. *Journal of Statistical Planning and Inference*, 103, 421–436.

[25] Heagerty PJ, Kurland BF. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika.* 2001; **88**: 973–985.

[26] Hedeker D, Gibbons RD. A random-effects ordinal regression model for multilevel analysis. *Biometrics.* 1994; **50**: 933–944.

[27] Liu Q, Pierce DA. A note on Gauss-Hermite quadrature. *Biometrika.* 1994; **81**: 624–629.

[28] Pinheiro JC, Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics.* 1995; **4**: 12–35.

[29] Niederreiter H. Random Number Generation and Quasi-Monte Carlo Methods. *CBMS-NSF Regional Conference Series in Applied Mathematics.* 1992; No. 63.

[30] Judd K. *Numerical Methods in Economics.* 1998; MIT Press, Boston.

[31] Caflisch R. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica.* 1998; **7**: 1–49.

[32] Gonzalez J, Tuerlinckx F, Boeck PD, Cools R. Numerical integration in logistic-normal models. *Computational Statistics and Data Analysis.* 2006; **51**: 1535–1548.

[33] Hurlburt MS, Wood PA, Hough RL. Providing independent housing for the homeless mentally ill: a novel approach to evaluating long-term longitudinal housing patterns. *Journal of Community Psychology.* 1996; **24**: 291–310.

[34] Hough RL, Harmon S, Tarke H, Yamashiro S, Quinlivan R, Landau-Cox P, Hurlburt MS. Supported independent housing: Implementation issues and solutions in the Sandiego Mckinney homeless demonstration research project. In *Mentally Ill and Homeless: Special Programs for Special Needs.* 1997; Breakey WR, Thompson JW (eds). 95–117

APPENDIX

Detailed Calculation of Quasi-Newton for MREM

The forms of the derivatives for Quasi-Newton algorithm follow

$$\frac{\partial \log L}{\partial \beta_j} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta; y_i, a_i) \left[\sum_{t=1}^{n_i} \left\{ \sum_{k=1}^{K} (y_{itk} - P^c_{itk}(a_i)) \frac{\partial \triangle_{itk}}{\partial \beta_j} \right\} \right] da_i,$$

$$\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta; y_i, a_i) \left[\sum_{t=1}^{n_i} \left\{ \sum_{k=1}^{K} (y_{itk} - P^c_{itk}(a_i)) \frac{\partial s^{(t)}}{\partial \alpha} a_i^{(t)} \right\} \right] da_i,$$

$$\frac{\partial \log L}{\partial \gamma_j} = \sum_{i=1}^{N} L(\theta; y_i)^{-1} \int L(\theta; y_i, a_i) \left[\sum_{t=1}^{n_i} \left\{ \sum_{k=1}^{K} (y_{itk} - P^c_{itk}(a_i)) \left(\frac{\partial \triangle_{itk}}{\partial \gamma_j} + \frac{\partial s^{(t)}}{\partial \gamma_j} a_i^{(t)} \right) \right\} \right] da_i,$$

where $\gamma_j = (\gamma_{j1}, \cdots, \gamma_{jj})^T$ is nonzero elements of $\Sigma_2^{1/2}$.

To compute the score vector and information matrix, we also need derivatives of $\triangle_{it}$ with respect to c and β. They can be obtained as the solution to the following system of linear equations, To compute the score vector and information matrix, we also need derivatives of $\triangle_{it}$ with respect to β, α and γ. They can be obtained as the solution to the following system of linear equations,

$$\frac{\partial P^M_{itk}}{\partial \beta_j} = \frac{\partial \triangle_{it1}}{\partial \beta_j} \int \frac{\partial P^c_{itk}(b_{it})}{\partial \triangle_{it1}} \phi(b_{it}) db_{it} + \cdots + \frac{\partial \triangle_{itK-1}}{\partial \beta_j} \int \frac{\partial P^c_{itk}(b_{it})}{\partial \triangle_{itK-1}} \phi(b_{it}) db_{it},$$

$$- \int \sum_{l=1}^{K-1} \frac{\partial P^c_{itk}(b_{it})}{\partial s_l^{(t)}} \frac{\partial s_l^{(t)}}{\partial \gamma_j} \phi(b_{it}) db_{it}$$

$$= \frac{\partial \triangle_{it1}}{\partial \gamma_j} \int \frac{\partial P^c_{itk}(b_{it})}{\partial \triangle_{it1}} \phi(b_{it}) db_{it} + \cdots + \frac{\partial \triangle_{itK-1}}{\partial \gamma_j} \int \frac{\partial P^c_{itk}(b_{it})}{\partial \triangle_{itK-1}} \phi(b_{it}) db_{it},$$

where

$$\frac{\partial P^c_{itk}(b_{it})}{\partial \triangle_{itj}} = \begin{cases} P^c_{itk}(b_{it})(1 - P^c_{itk}(b_{it})), & \text{if } j = k; \\ -P^c_{itk}(b_{it}) P^c_{itj}(b_{it}), & \text{if } j \neq k, \end{cases}$$

for $k, j = 1, \cdots, K-1$. We solve for $\frac{\partial \triangle_{itg}}{\partial \beta_j^{(n)}}$ and $\frac{\partial \triangle_{itg}}{\partial c_j}$ using these $K-1$ equations.

Calculation of $\triangle_{it}$

From (4), we know that $\triangle_{itk}$ are a function of β and σ. Estimates of $\triangle_{it} = (\triangle_{it1}, \cdots, \triangle_{it,K-1})$ can be obtained using Newton-Raphson as follows. Let $f(\triangle_{it}) = (f_1(\triangle_{it}), \cdots, f_{K-1}(\triangle_{it}))$ where $f_k(\triangle_{it}) = \int P^c_{itk}(b_{it}) \phi(b_{it}) db_{it} - P^M_{itk}$. We obtain

$$\triangle_{it}^{(n+1)} = \triangle_{it}^{(n)} - H^{-1}(\triangle_{it}^{(n)}) f(\triangle_{itk}^{(n)}),$$

where

$$H(\triangle_{it}) = \begin{pmatrix} \frac{\partial f_1(\triangle_{it})}{\partial \triangle_{it1}} & \cdots & \frac{\partial f_1(\triangle_{it})}{\partial \triangle_{it,K-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_{K-1}(\triangle_{it})}{\partial \triangle_{it1}} & \cdots & \frac{\partial f_{K-1}(\triangle_{it})}{\partial \triangle_{it,K-1}} \end{pmatrix},$$

$$\frac{\partial f_k(\triangle_{it})}{\partial \triangle_{itj}} = \begin{cases} \int P^c_{itk}(b_{it})(1 - P^c_{itk}(b_{it}))\phi(b_{it})db_{it}, & \text{if } j = k; \\ -\int P^c_{itk}(b_{it})P^c_{itj}(b_{it})\phi(b_{it})db_{it}, & \text{if } j \neq k. \end{cases}$$