Journal of Statistical Research *https://doi.org/10.47302/jsr.2020540207*
2020, Vol. 54, No. 2, pp. 207-221 *ISSN 0256 - 422 X*

GENERALIZED LINEAR MODELING OF A PANEL MOBILITY INDEX WITH CORRELATED MULTIVARIATE CATEGORICAL RESPONSES

DREW M. LAZAR*
Department of Mathematical Sciences, Ball State University, Muncie, IN 47304
Email: dmlazar@bsu.edu

MUNNI BEGUM
Department of Mathematical Sciences, Ball State University, Muncie, IN 47304
Email: mbegum@bsu.edu

SUMMARY

Data with multivariate, longitudual categorical responses often occur in applications. It can be difficult to analyze and model such data while simultaneously taking into account explanatory variables and correlations between the responses over time. We take a generalized linear model approach to this problem in analyzing panel data from the Health and Retirement Survey (HRS) that includes older Americans' mobility over several years as a response. We provide a general formula for the likelihood of such data and apply it to the case when there are three binary responses. This approach can be taken, with computational limits, for data with multivariate, categorical responses with any number of categories. We consider, simultaneously, interpretations of coefficients, dependence of responses and goodness-of-fit in reduced models for parsimony while taking into account explanatory data. The gradient of the objective function is provided for use in gradient descent and the coded optimization algorithm is tested with a Monte Carlo simulation. Dependence of responses in mobility is shown before taking explanatory variables into account, and dependence is shown in a Markov logistic regression model and in the generalized linear model taking into account race, age, gender and interactions between them.

Keywords and phrases: Generalized Linear Modeling, Correlation, Categorical Data Analysis, Computational Statistics

AMS Classification: 62-07

1 Background and Introduction

Modeling repeated categorical responses on a group of subjects is a challenging problem, in part, because such responses are typically not independent across time periods. The difficulty in creating such models is noted in Sutradhar (2014), particularly when each of the responses is multinomial and not just binary. A number of approaches have been formulated to address this problem including Generalized Estimating Equations (GEE) (Lipsitz et al., 1994; Hardin, 2005; Islam and Chowdhury, 2017) and Generalized Linear Mixed Models (GLMM) (Fitzmaurice et al., 2007; Stroup, 2012).

* Corresponding author

GEE is a method that assumes a particular covariance structure but does not fully specify a parametric distribution. Using iterative weighted least squares (Miller et al., 1993), with the covariance providing the weights, the generalized estimating equations are fit which provide estimates of population average effects. As a semi-parametric method, GEE provides robustness, however, it does not estimate individual subject probabilities beyond population averages. Islam and Chowdhury (2017) include a chapter providing a background, further explanation, estimation procedures and examples of GEE.

GLMM is an extension of generalized linear models (GLM) that utilizes different distributions of link functions in order to simultaneously model random variation among and within the group of repeated measures. The allows the inclusion of both fixed and random effects which adds complexity to the model in comparison to ordinary GLM models. Stroup (2012) provides an in-depth text which explains, explores and applies the GLMM method.

In this paper, we take the approach of Miller et al. (1993), Sun and Sutradhar (2015) and Sutradhar (2014). In this approach both marginal and conditional probabilities are simultaneously fit in a joint distribution in a GLM model in order to model and analyze multivariate, categorical responses. In particular, we extend and generalize the results of Uddin and Begum (2018) in analyzing longitudinal data from the 2012, 2013 and 2014 waves of the University of Michigan Health and Retirement Study (HRS) (University of Michigan, 2012-2014).

Mobility is the ability to move in one's environment with ease and without restriction (Farlex, 2020) and is of crucial concern for seniors. Mobility impairments have significant quality of life and economic implications and the associated lack of physical activity often has detrimental health implications such as depression, weight gain, bone fractures, and further loss of mobility (Kabiri et al., 2018; Rosenbloom, 2005). Mobility generally declines with age (Satariano et al., 2012) and can be influenced by race (Allman et al., 2004) and gender (Legato and Bilezikian, 2004). It is important to have flexible statistical models of mobility changes that can take into account these and other factors over time. We generalize and significantly approve upon the approach of Uddin and Begum (2018) in creating such a model which, of course, can be used in analyzing other longitudinal or panel data while taking account important covariates.

In Uddin and Begum (2018), three longitudinal binary responses were considered over three periods (or "waves"), but the data was partitioned by the binary response in the first period and general linear models built on each of the partitions. To extend the results of Uddin and Begum (2018), in this paper, we (1) build the full model on responses over the three periods, (2) include a general formula for the likelihood and with this formula the method employed can be applied to any number of responses with any number of categories, (3) conduct a Monte Carlo simulation to demonstrate convergence of the gradient descent Matlab algorithm used for parameter estimation and optimization, (4) allow for interaction between explanatory variables, include coefficient interpretation and demonstrate model selection for parsimony and interpretation, and (5) code all optimization, model selection, parameter estimation and simulation for the case of binary responses over each of three periods in Matlab. All code and data is available at `https://github.com/DrewLazar/GLM_Long_MobilityInd/tree/master/MatlabCode`. The general formula is also applied and optimization is also coded in the case of three categorical responses over each of two periods.

The general formula we provide for the likelihood was developed independently, but a number of similar formulas and approaches, also using inverse logits for the marginal and conditional probabilities, are given in Sutradhar (2014), Sun and Sutradhar (2015), Islam and Chowdhury (2017) and Islam et al. (2013). Sutradhar (2014) provides general formulas for the derivatives of the likelihood function for maximimum likelihood optimization similar to what we do in this paper. They also employ the generalized quasilikelihood approach (GQL) for estimation in which the correlation structure is assumed and estimated by a method of moments and then used in a quadratic distance function which is minimized for optimization. Sutradhar (2014) also considers interaction effects, as we do in this paper, and allows for different numbers of categories over different periods whereas we assume that the number of categories stays the same. In Sun and Sutradhar (2015), the GQL approach is also used to model bivariate, multinomial data from left and right eye retinopathy status.

General linear regression models are examined in Islam and Chowdhury (2010) where the HRS data set, with disease status as the response, is also analyzed. The focus of Islam and Chowdhury (2010) is prediction and assessment of predictive power in modeling disease status, while this paper examines parameter estimation and interpretation in regards to the mobility index in the HRS. Further discussion of the development and background of GLM models in these settings can be found in Uddin and Begum (2018), the paper which is the basis for this work.

2 Methodology

2.1 Terminology and Notation

Let $Y_{k,x}$ be the categorical response variable at time point $k = 1, \ldots, m$ with explanatory covariate $X = x$. We assume that the sample space of $Y_{k,x}$ is $\{0, 1, \ldots, d-1\}$ for $k = 1, \ldots, m$ and $X = x$, i.e., each $Y_{k,x}$ can be in one of d different categories. We denote by $Y_{k,i}$ the categorical response at time point $k = 1, \ldots, m$ for individual $i = 1, \ldots, n$ with observed covariate $x = x_i$.

When not taking into account the effect of explanatory variables, we let Y_k be our response for time periods $k = 1, \ldots, m$. We can nonparametrically conduct tests of independence between our responses in several ways. For the case where $m = 3$ and with each variable in $d = 2$ categories, using chi-squared tests and the Marshall-Olkin correlation coefficient (Marshall and Olkin, 1985) as used by Uddin and Begum (2018), we carry out such tests as in Section 6.1.

Throughout this paper we let IL be the inverse logit function, that is,

$$\mathrm{IL}(t) = \frac{\exp(t)}{1 + \exp(t)} \text{ for } t \in \mathbb{R}.$$

We can also test the dependence between our response variables, taking into account explanatory covariates, by adding response variables to our explanatory covariates in logistic regression models. We carry out such tests in Section 6.2.

We can test the conditional dependence of $Y_{2,x}$ on $Y_{1,x}$ by considering the model

$$p(Y_{2,x}|y_1, x) = \frac{\exp(x'\boldsymbol{\beta} + \alpha_1 y_1)}{1 + \exp(x'\boldsymbol{\beta} + \alpha_1 y_1)} = \mathrm{IL}(x'\boldsymbol{\beta} + \alpha_1 y_1) \tag{2.1}$$

where y_1 and x are observed values of response variable $Y_{1,x}$ (as an explanatory variable) and explanatory variable X, respectively, $\boldsymbol{\beta}$ a vector of population parameters and α_1 a population parameter. If α_1 is significantly non-zero based on sample $y_{2,i}, y_{1,i}, i = 1, \ldots, n$, this suggests $Y_{2,x}$ is dependent on $Y_{1,x}$. We can test the conditional dependence of $Y_{3,x}$ on $Y_{2,x}$ and $Y_{1,x}$ by considering the model

$$\begin{aligned} p(Y_{3,x}|y_1, y_2, x) &= \frac{\exp(x'\boldsymbol{\beta} + \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_1 y_2)}{1 + \exp(x'\boldsymbol{\beta} + \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_1 y_2)} \\ &= \mathrm{IL}(x'\boldsymbol{\beta} + \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_1 y_2) \end{aligned} \tag{2.2}$$

where y_1, y_2 and x are observed values of response variables $Y_{1,x}, Y_{2,x}$ (as explanatory variables) and explanatory variable X, respectively, $\boldsymbol{\beta}$ is a vector of population parameters, and α_1, α_2 and α_3 are population parameters. We include the term $y_1 y_2$ to estimate dependence conditionally and to test for interaction. That is, if $y_1 = 0$, a change in status of y_2 from 0 to 1 changes the log-odds of difficulty in wave 3 by α_2 and if $y_1 = 1$ a change in status of y_2 from 0 to 1 changes the log-odds of difficulty in wave 3 by $\alpha_2 + \alpha_3$. Thus, we can test whether status in wave 1 affects how difficulty in wave 2 affects difficulty in wave 3 by testing whether $\alpha_3 \neq 0$.

3 Data and Variable Selection

As in Uddin and Begum (2018), we consider secondary data from the longitudinal household survey from the Health and Retirement Study (HRS) (University of Michigan, 2012-2014). This is a yearly survey of approximately 20,000 Americans over the age 50. This in-depth survey considers a large host of demographic and external factors such as race, income, health insurance coverage, living arrangements, etc., to provide a rich database to researchers and practitioners concerned with important issues of aging in the United States. The data we consider is for the three years 2012, 2013 and 2014 to give us our categorical responses and these are considered as periods $k = 1, 2, 3$, respectively. The ability to move to different places and locations in your environment is a crucial issue for seniors. A mobility index is given based on responses about the ability to walk several blocks, walking one block, walking across the room, climbing several flights of stairs, etc. The mobility index is scored from 0 up to 5 with 0 indicating little difficulty and 5 indicating the most difficulty. With $M_{k,i}$ the mobility index in period k for individual $i = 1, \ldots, n$, $M_{k,i} = 0$ indicates no difficulty and $M_{k,i} = 1, \ldots, 5$ indicates some difficulty. We recoded the mobility index for the three periods $k = 1, 2, 3$ as

$$Y_{k,i} = \begin{cases} 0, \text{ if } M_{k,i} = 0 \\ 1, \text{ if } M_{k,i} = 1, \ldots, 5 \end{cases}$$

so that $Y_{k,i} = 0$ indicates no difficulty and $Y_{k,i} = 1$ indicates difficulty. For a further example in the code, with $d = 3$ categories, we considered the $M_{k,i} = 0$ to be no difficulty, $M_{k,i} = 1, 2, 3$ to be moderate difficulty and $M_{k,i} = 4, 5$ to be difficulty in $m = 2$ (2013, 2014) periods. The explanatory

variables, including interaction terms, in this data set that we considered were

$$X_1 \equiv 1(\text{main effect}),\ X_2 = \begin{cases} 0, \text{ if female} \\ 1, \text{ if male} \end{cases},\ X_3 = \begin{cases} 0, \text{ if White} \\ 1, \text{ if other race} \end{cases},\ X_4 = \text{Age in Years/5},$$

$$X_5 = X_3 * X_4 \text{ (RacebyAge)}, X_6 = X_2 * X_3 \text{ (RacebyGender)}, X_7 = X_2 * X_4 \text{ (GenderbyAge)}.$$

Our data set consists of 17,350 observations. All missing information is assumed to be missing completely at random and is excluded from further analysis. Age is divided by 5 for optimization purposes and for interpretability of coefficients.

4 A Generalized Linear Model Approach

We consider a full generalized linear model that simultaneously allows for relationships and interactions between explanatory variables X, between response variables and explanatory variables and between the response variables themselves over m time periods. We first formulate the likelikood function in the general case when we have m responses $Y_{1,x}, \ldots, Y_{m,x}$ with d categories for each response. Let

$$P_{y_1,\ldots,y_m} = P(Y_{1,x} = y_1, \ldots, Y_{m,x} = y_m; X = x) \text{ with } y_i = 0, 1, \ldots, d-1 \text{ for } i = 1, \ldots, m,$$

where $P_{y_1,\ldots,y_m}$ depends on x. With $A \subset \mathbb{R}^{d-1}$ let

$$A = \{(0,0,\ldots,0),(1,0,\ldots,0),(0,1,\ldots,0),\ldots,(0,0,\ldots,1)\} \text{ and}$$
$$h : A \to \mathbb{R},\ h(0,0\ldots,0) = 0, h(1,0,\ldots,0) = 1, h(0,1,\ldots,0) = 2, \ldots, h(0,0,\ldots,1) = d-1.$$

With $a = m(d-1) - d + 2, b = m(d-1)$ and $z_1, \ldots, z_b$ such that

$$h(z_1, \ldots, z_{d-1}) = y_1, h(z_d, \ldots, z_{2(d-1)}) = y_2, \ldots, h(z_a, \ldots, z_b) = y_m,$$

the m-variate joint distribution for outcome variables $Y_{1,x}, \ldots, Y_{m,x}$ at time periods $1, \ldots, m$ can then be written as

$$P(Y_{1,x} = y_1, Y_{2,x} = y_2, \ldots, Y_{m,x} = y_m) = \prod_{i_1=0}^{d-1} \cdots \prod_{i_m=0}^{d-1} P_{i_1,\ldots,i_m}^{\prod_{k=1}^{b}\left(1-w_k^{(i_1,\ldots,i_m)}+(-1)^{1-w_k^{(i_1,\ldots,i_m)}}z_k\right)},$$

where $w_1^{(i_1,\ldots,i_m)}, \ldots, w_b^{(i_1,\ldots,i_m)}$ is such that

$$h\big(w_1^{(i_1,\ldots,i_m)}, \ldots, w_{d-1}^{(i_1,\ldots,i_m)}\big) = i_1, \ldots, h\big(w_a^{(i_1,\ldots,i_m)}, \ldots, w_b^{(i_1,\ldots,i_m)}\big) = i_m.$$

For example, for $m = 2, d = 3$ we have

$$\begin{aligned} P(Y_{1,x} = y_1, Y_{2,x} = y_2) = {} & P_{00}^{(1-z_1)(1-z_2)(1-z_3)(1-z_4)} P_{01}^{(1-z_1)(1-z_2)z_3(1-z4)} \\ & \times P_{02}^{(1-z_1)(1-z_2)(1-z_3)z_4} P_{10}^{z_1(1-z_2)(1-z_3)(1-z4)} P_{11}^{z_1(1-z_2)z_3(1-z4)} \\ & \times P_{12}^{z_1(1-z_2)(1-z_3)z_4} P_{20}^{(1-z_1)z_2(1-z_3)(1-z4)} P_{21}^{(1-z_1)z_2z_3(1-z_4)} P_{22}^{(1-z_1)z_2(1-z_3)z_4}, \end{aligned}$$

where $h(z_1, z_2) = y_1$ and $h(z_3, z_4) = y_2$. This example is coded for parameter estimates and maximum likelihood estimation.

In this paper, we extend and generalize the results of Uddin and Begum (2018) and cover the case where $m = 3, d = 2$. With $m = 3, d = 2$ we have

$$\begin{aligned} P(Y_{1,x} = y_1, Y_{2,x} = y_2, Y_{3,x} = y_3) = P_{000}^{(1-z_1)(1-z_2)(1-z_3)} P_{010}^{(1-z_1)z_2(1-z_3)} \\ P_{001}^{(1-z_1)(1-z_2)z_3} P_{011}^{(1-z_1)z_2z_3} P_{100}^{z_1(1-z_2)(1-z_3)} P_{110}^{z_1z_2(1-z_3)} P_{101}^{z_1(1-z_2)z_3} P_{111}^{z_1z_2z_3}, \end{aligned} \tag{4.1}$$

where h is the identity, $h(z_1) = y_1, h(z_2) = y_2, h(z_3) = y_3$, so that in this case, $z_1 = y_1, z_2 = y_2$ and $z_3 = y_3$.

The exponential family representation of the joint probability mass function of the response variables at time points 1, 2 and 3 of (4.1) is then given by

$$\begin{aligned} &P(Y_{1,x} = y_1, Y_{2,x} = y_2, Y_{3,x} = y_3) \\ &\qquad = \exp\left[\ln(P(Y_{1,x} = y_1, Y_{2,x} = y_2, Y_{3,x} = y_3)\right] \\ &\quad = \exp\Bigg[\log(P_{000}) + z_1 \log\left(\frac{P_{100}}{P_{000}}\right) + z_2 \log\left(\frac{P_{010}}{P_{000}}\right) + z_1z_2 \log\left(\frac{P_{000}P_{110}}{P_{010}P_{100}}\right) \\ &\qquad + z_3 \log\left(\frac{P_{001}}{P_{000}}\right) + z_1z_3 \log\left(\frac{P_{000}P_{101}}{P_{001}P_{100}}\right) + z_2z_3 \log\left(\frac{P_{000}P_{011}}{P_{001}P_{010}}\right) \\ &\qquad + z_1z_2z_3 \log\left(\frac{P_{001}P_{010}P_{100}P_{111}}{P_{000}P_{011}P_{101}P_{110}}\right)\Bigg]. \end{aligned} \tag{4.2}$$

For a sample of individuals $i = 1, \ldots, n$, the log-likelihood function is then expressed as

$$\begin{aligned} \ell = \sum_i^n \Bigg[&\log(P_{000i}) + z_{1i} \log\left(\frac{P_{100i}}{P_{000i}}\right) + z_{2i} \log\left(\frac{P_{010i}}{P_{000i}}\right) + z_{1i}z_{2i} \log\left(\frac{P_{000i}P_{110i}}{P_{010i}P_{100i}}\right) \\ &+ z_{3i} \log\left(\frac{P_{001i}}{P_{000i}}\right) + z_{1i}z_{3i} \log\left(\frac{P_{000i}P_{101}}{P_{001i}P_{100i}}\right) + z_{2i}z_{3i} \log\left(\frac{P_{000i}P_{011i}}{P_{001i}P_{010i}}\right) \\ &+ z_{1i}z_{2i}z_{3i} \log\left(\frac{P_{001i}P_{010i}P_{100i}P_{111i}}{P_{000i}P_{011i}P_{101i}P_{110i}}\right)\Bigg], \end{aligned} \tag{4.3}$$

where $P_{i_1i_2i_3i} = P(Y_{1,x_i} = i_1, Y_{2,x_i} = i_2, Y_{3,x_i} = i_3)$ for $(i_1 = 0, 1; i_2 = 0, 1; i_3 = 0, 1)$ and x_i is the observed value of x from subject $i = 1, \ldots, n$ from the sample.

The components of the links functions for the generalized model from the joint probability mass function in (4.2) are now

$$\begin{aligned} &\eta_{0,x} = \log(P_{000}),\ \eta_{1,x} = \log\left(\frac{P_{100}}{P_{000}}\right),\ \eta_{2,x} = \log\left(\frac{P_{010}}{P_{000}}\right),\ \eta_{3,x} = \log\left(\frac{P_{001}}{P_{000}}\right), \\ &\eta_{4,x} = \log\left(\frac{P_{000}P_{110}}{P_{010}P_{100}}\right),\ \eta_{5,x} = \log\left(\frac{P_{000}P_{101}}{P_{001}P_{100}}\right), \\ &\eta_{6,x} = \log\left(\frac{P_{000}P_{011}}{P_{001}P_{010}}\right),\ \eta_{7,x} = \log\left(\frac{P_{001}P_{010}P_{100}P_{111}}{P_{000}P_{011}P_{101}P_{110}}\right), \end{aligned} \tag{4.4}$$

where $\eta_{0,x}$ is the baseline link function.

We now introduce parametric models for the joint probabilities in terms of conditional and marginal probabilities. Let $X = (X_1, X_2, \ldots, X_p)'$ be the covariate vector with $X_1 \equiv 1$ to admit main effects. For $j = 0, 1$ and $k = 0, 1$ let

$$\boldsymbol{\beta}_1 = (\beta_{1,1}, \ldots, \beta_{1,p})', \boldsymbol{\beta}_{j1} = (\beta_{j1,1}, \ldots, \beta_{j1,p})', \boldsymbol{\beta}_{jk1} = (\beta_{jk1,1}, \ldots, \beta_{jk1,p})'. \tag{4.5}$$

Then for $j, k = 0, 1$ let

$$\begin{aligned}
&P(Y_{1,x} = 1; X = x) = \frac{e^{x'\boldsymbol{\beta}_1}}{1 + e^{x'\boldsymbol{\beta}_1}} \implies P(Y_{1,x} = 0; x) = \frac{1}{1 + e^{x'\boldsymbol{\beta}_1}}.\\
&P(Y_{2,x} = 1|Y_{1,x} = j; X = x) = \frac{e^{x'\boldsymbol{\beta}_{j1}}}{1 + e^{x'\boldsymbol{\beta}_{j1}}} \implies P(Y_{2,x} = 0|Y_{1,x} = j; X = x) = \frac{1}{1 + e^{x'\boldsymbol{\beta}_{j1}}}.\\
&P(Y_{3,x} = 1|Y_{1,x} = j, Y_{2,x} = k; X = x) = \frac{e^{x'\boldsymbol{\beta}_{jk1}}}{1 + e^{x'\boldsymbol{\beta}_{jk1}}} \implies\\
&P(Y_{3,x} = 0|Y_{1,x} = j, Y_{2,x} = k; X = x) = \frac{1}{1 + e^{x'\boldsymbol{\beta}_{jk1}}}.
\end{aligned} \tag{4.6}$$

Then for $j = 0, 1, k = 0, 1$ and $m = 0, 1$, expressing joint probabilities

$$\begin{aligned}
P_{jkm} = P(Y_{1,x} = j; X = x)&P(Y_{2,x} = k|Y_{1,x} = j; X = x)\\
&\times P(Y_{3,x} = m|Y_{1,x} = j, Y_{2,x} = k; X = x)
\end{aligned} \tag{4.7}$$

and substituting parameterized, conditional probabilities in the link functions we have

$$\begin{aligned}
&\eta_{0,x} = -\log(1 + e^{x'\boldsymbol{\beta}_{001}}) - \log(1 + e^{x'\boldsymbol{\beta}_{01}}) - \log(1 + e^{x'\boldsymbol{\beta}_1}),\\
&\eta_{1,x} = x'\boldsymbol{\beta}_1 + \log(1 + e^{x'\boldsymbol{\beta}_{01}}) + \log(1 + e^{x'\boldsymbol{\beta}_{001}}) - \log(1 + e^{x'\boldsymbol{\beta}_{11}}) - \log(1 + e^{x'\boldsymbol{\beta}_{101}})\\
&\eta_{2,x} = x'\boldsymbol{\beta}_{01} + \log(1 + e^{x'\boldsymbol{\beta}_{001}}) - \log(1 + e^{x'\boldsymbol{\beta}_{011}}),\ \eta_{3,x} = x'\boldsymbol{\beta}_{001},\\
&\eta_{4,x} = x'(\boldsymbol{\beta}_{11} - \boldsymbol{\beta}_{01}) - \log(1 + e^{x'\boldsymbol{\beta}_{001}}) - \log(1 + e^{x'\boldsymbol{\beta}_{111}}) + \log(1 + e^{x'\boldsymbol{\beta}_{011}}) + \log(1 + e^{x'\boldsymbol{\beta}_{101}}),\\
&\eta_{5,x} = x'(\boldsymbol{\beta}_{101} - \boldsymbol{\beta}_{001}),\ \eta_{6,x} = x'(\boldsymbol{\beta}_{011} - \boldsymbol{\beta}_{001}),\ \eta_{7,x} = x'(\boldsymbol{\beta}_{001} + \beta_{111} - \boldsymbol{\beta}_{011} - \boldsymbol{\beta}_{101}).
\end{aligned} \tag{4.8}$$

By considering the conditional probabilities in (4.6) and the link functions in (4.8) we can examine the dependence structure among the response variables across the three time periods. For example,

$$\begin{aligned}
&Y_{3,x} \text{ is conditionally independent of } Y_{1,x} \text{ and } Y_{2,x} \text{ for all } x\\
&\iff \eta_{5,x} = \eta_{6,x} = \eta_{7,x} = 0 \text{ for all } x\\
&\iff \boldsymbol{\beta}_{001} = \boldsymbol{\beta}_{111} = \boldsymbol{\beta}_{011} = \boldsymbol{\beta}_{101}.
\end{aligned} \tag{4.9}$$

Further,

$$Y_{2,x} \text{ is conditionally independent of } Y_{1,x} \text{ for all } x \iff \boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{01} \tag{4.10}$$

and we have (4.9) and (4.10) if and only $\eta_{4,x} = \eta_{5,x} = \eta_{6,x} = \eta_{7,x} = 0$ for all x. Thus, we can test the dependence structure among the response variables across the three time points with hypothesis tests:

$$H_{0,1} : \boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{01},\ H_{0,2} : \boldsymbol{\beta}_{001} = \boldsymbol{\beta}_{111} = \boldsymbol{\beta}_{011} = \boldsymbol{\beta}_{101}. \tag{4.11}$$

These tests can be computed individually or simultaneously with a multiple comparison adjustment.

4.1 Maximum Likelihood Estimation

Observing covariates x_i and responses y_{1i}, y_{2i}, y_{3i} from sample of subjects $i = 1, \ldots, n$, to estimate parameters introduced in (4.6), we substitute the link functions in (4.8) into the log-likelihood in (4.3) according to (4.4) and maximize with respect to the parameters. To do so we use the MATLAB routine fminunc with a user-supplied gradient given in (4.14). The code for optimization and data for this paper can be found in `https://github.com/DrewLazar/GLM_Long_MobilityInd/tree/master/MatlabCode`. For given $x \in \mathbb{R}^p$ consider the function

$$f(a) = \log[1 + \exp(x'a)] \text{ for } a \in \mathbb{R}^p.$$

The directional derivative of f at a in the direction of $c \in \mathbb{R}^p$ is

$$d_a f(c) = \frac{d}{dt} f(a + tc)|_{t=0} = \frac{(x'c)\exp(x'a)}{1 + \exp(x'a)}.$$

Thus, the gradient of f at a (as a row vector) is given by

$$\nabla_a f = \left[\frac{\exp(x'a)}{1 + \exp(x'a)}\right] x' = \text{IL}(x'a)x' \tag{4.12}$$

where $\text{IL}(x'a)$ is the inverse logit function evaluated at $x'a$. Let B and $\mathbf{0}$ be the vectors

$$B = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{11}', \boldsymbol{\beta}_{001}', \boldsymbol{\beta}_{101}', \boldsymbol{\beta}_{011}', \boldsymbol{\beta}_{111}')' \in \mathbb{R}^{7p \times 1} \text{ and } \mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^{1 \times p}.$$

Using (4.12) and (4.8) we have

$$\begin{aligned}
&\nabla_B \eta_{0,x} = -(\text{IL}(x'\boldsymbol{\beta}_1)x', \text{IL}(x'\boldsymbol{\beta}_{01})x', \mathbf{0}, \text{IL}(x'\boldsymbol{\beta}_{001})x', \mathbf{0}, \mathbf{0}, \mathbf{0}) \\
&\nabla_B \eta_{1,x} = (x', \text{IL}(x'\boldsymbol{\beta}_{01})x', -\text{IL}(x'\boldsymbol{\beta}_{11})x', \text{IL}(x'\boldsymbol{\beta}_{001})x', -\text{IL}(x'\boldsymbol{\beta}_{101})x', \mathbf{0}, \mathbf{0}) \\
&\nabla_B \eta_{2,x} = (\mathbf{0}, x', \mathbf{0}, \text{IL}(x'\boldsymbol{\beta}_{001})x', \mathbf{0}, -\text{IL}(x'\boldsymbol{\beta}011)x', \mathbf{0}) \\
&\nabla_B \eta_{4,x} = (\mathbf{0}, -x', x', -\text{IL}(x'\boldsymbol{\beta}_{001})x', \text{IL}(x'\boldsymbol{\beta}_{101})x', \text{IL}(x'\boldsymbol{\beta}_{011})x', -\text{IL}(x'\boldsymbol{\beta}_{111})x') \\
&\nabla_B \eta_{3,x} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, x', \mathbf{0}, \mathbf{0}, \mathbf{0}),\ \nabla_B \eta_{5,x} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, -x', x', \mathbf{0}, \mathbf{0}) \\
&\nabla_B \eta_{6,x} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, -x', \mathbf{0}, x', \mathbf{0}),\ \nabla_B \eta_{7,x} = (\mathbf{0}, \mathbf{0}, \mathbf{0}, x', -x', -x', x').
\end{aligned} \tag{4.13}$$

This gives the gradient of the log-likelihood in (4.3) as

$$\begin{aligned}
\nabla_B \ell = \sum_i^n \Big(&\nabla_B \eta_{0,i} + z_{1i}\nabla_B \eta_{1,i} + z_{2i}\nabla_B \eta_{2,i} + z_{1i}z_{2i}\nabla_B \eta_{4,i} + z_{3i}\nabla_B \eta_{3,i} \\
&+ z_{1i}z_{2i}z_{3i}\nabla_B \eta_{7,i} + z_{1i}z_{3i}\nabla_B \eta_{5,i} + z_{2i}z_{3i}\nabla_B \eta_{6,i} \Big)
\end{aligned} \tag{4.14}$$

with $\eta_{k,i} = \eta_{k,x_i}$ where x_i is the observed value of x from subject $i = 1, \ldots, n$ from the sample.

4.2 Monte Carlo Simulation

After substituting our parameterized link functions in (4.8) in (4.3) according to (4.4) we have 49 parameters to estimate with 7 intercept terms and 42 slopes for our covariates. Accordingly, in two simulations, we generate $n = 17,350$ observations with coefficients $\boldsymbol{\beta} \in \mathbb{R}^{49}$ and explanatory variables X as

Simulation 1: $X \sim (1, W)$ where $W \sim \mathcal{N}(\mathbf{0}, \mathrm{I}_6)$ with I_6 the 6×6 identity matrix and
Simulation 2: $X \sim (1, X_2, \ldots, X_7)$ with $X_2 \sim b(1, 0.594), X_3 \sim b(1, 0.719)$,
$X_4 \sim \mathcal{N}(12.906, 5.013)$ and with interaction terms X_5, X_6 and X_7 as in (3).

We then generate responses according to the distribution in (4.1) by a Monte Carlo simulation. Specifically, we code functions for the marginal and conditional probabilities in (4.6). Then, for the population $\boldsymbol{\beta}$ and for each generated x_i, we

1. compute P_{jkm} in (4.7) for $j = 0, 1, k = 0, 1$ and $m = 0, 1$,
2. partition [0,1] into eight intervals of lengths given by the P_{jkm} for $j = 0, 1, k = 0, 1$ and $m = 0, 1$,
3. generate an observation, y, on [0,1] by the uniform distribution,
4. set $(y_{1,i}, y_{2,i}, y_{3,i}) = (j, k, m)$ according to the interval in which y falls.

In agreement with simulation 2, in our data in section 3 we have 59.4% females, 71.9% white and Age/5 with mean 12.906, variance 5.013 and a bell-shaped histogram. We maximize our log-likelihood using our Matlab gradient descent algorithm with the user-supplied gradient in (4.14) to estimate $\boldsymbol{\beta}$. We observe convergence, with the sample mean of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ tending to the vector $\mathbf{0}$ over an increasing number of runs. This is demonstrated in Table 1 where we compute the Euclidean norm of the sample mean,

$$\left\| \overline{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}} \right\|$$

over an increasing number of runs.

Table 1: $\left\| \overline{\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}} \right\|$ over increasing numbers of runs for Simulations 1 and 2.

	Number of runs							
Similations	10	50	100	150	200	300	400	500
1	0.0917	0.0353	0.0297	0.0267	0.0204	0.0164	0.0123	0.0078
2	0.6266	0.2931	0.1657	0.1375	0.0985	0.0785	0.0685	0.0485

5 Data Analysis and Results

5.1 Exploratory Data Analysis

As in Uddin and Begum (2018), we consider the last waves from 2012, 2013 and 2014 from the Health Retirement Study (HRS). The mobility indices of elderly people with 0 as no difficulty and 1 as difficulty are considered as response variables at time points 1, 2 and 3 and are denoted as Y_1, Y_2 and Y_3. The full distribution of the $n = 17,350$ responses is given in Table 2 with marginal distributions broken down for observed values of Y_2 given $Y_1 = y_1$ and observed values of Y_3 given $Y_1 = y_1, Y_2 = y_2$.

Table 2: Distribution of Sample Responses

	No Difficulty				**Difficulty**			
y_1	9341 (0.538)				8009 (0.462)			
$y_2 \vert y_1$	7548 (0.808)		1793 (0.192)		1516 (0.189)		6493 (0.811)	
$y_3 \vert (y_2, y_1)$	6220 (0.821)	1328 (0.176)	655 (0.365)	1138 (0.635)	766 (0.505)	750 (0.495)	577 (0.089)	5916 (0.911)

Note that there a fewer observations here than in Uddin and Begum (2018) due to a coding mistake in Uddin and Begum (2018). The percentage of subjects with no difficulty in wave 1 (53.8%) is roughly the same as the percentage of subjects with difficulty (46.2%). In the second wave, subjects are nearly equally likely to stay at difficulty (80.8%) or stay at no difficulty (81.1%) from the first wave. In the third wave, subjects are more likely to stay at difficulty or no difficulty from the previous wave if they are at the same state in first wave suggesting interaction between y_1 and y_2 in explaining y_3 (before accounting for covariates). For example, 82.1% of those that had difficulty in the first wave and the second wave had difficulty in the third wave, but 50.5% of those that had no difficulty in the first wave but difficulty in the second wave had difficulty in the third wave. In addition, the proportion of those that had difficulty increases from wave 1 (46.2%) to wave 2 (47.8%) to wave 3 (52.63%) suggesting a positive effect of age.

6 Tests of Dependence

6.1 Nonparametric tests

We consider the tests of dependence of our three response variables Y_1, Y_2, Y_3 as outlined in Section 2.1. Nonparametrically, and as suggested by Table 2, using chi-squared tests we find dependence of: 1) Y_2 and Y_1, 2) Y_1, Y_2 and Y_3, and 3) Y_3 on Y_2 and Y_1. These tests all return large chi-squared values with p-values ≈ 0. We also compute the Marshall-Olkin correlation coefficients (Marshall and Olkin, 1985) and as used by Uddin and Begum (2018) for: 1) Y_2 and Y_1, 2) Y_3 and Y_2, and 3)

Y_3 and Y_1. These coefficients are, respectively,

$$p_{1,2} = 0.6176,\ p_{2,3} = 0.6233,\ \text{and}\ p_{1,3} = 0.5674.$$

These all indicate strong correlation. That there is lower correlation between Y_1 and Y_3 agrees with the fact that there is an intervening year between these responses.

6.2 Markov Model Analysis

We now conduct a Markov model analysis with logistic regression where in (2.1) the mobility index in wave 1, $Y_{1,x}$, is taken as an explanatory variable and in (2.2) the mobility indices in waves 1 and 2, $Y_{1,x}$ and $Y_{2,x}$, are taken as explanatory variables. Including the explanatory variables, X, in the model takes into account confounding from the explanatory variables of age, race and sex and their interactions in assessing independence. By maximum likelihood estimation we find

$$\hat{p}(y_2|y_1, x) = \text{IL}(x'\hat{\beta} + \hat{\alpha}_1 y_1), \tag{6.1}$$

where

$$(\hat{\beta}, \hat{\alpha_1}) = (-3.888, 0.460, 0.625, 0.170, -0.028, -0.031, -0.005, 2.778)'.$$

This means that accounting for x, the odds of increased difficulty in wave 2 increases by a factor of approximately $\exp(2.778) = 15.959$ when going from no difficulty to difficulty in wave 1. To conduct a likelihood ratio test we also fit $\hat{p}(y_2|x)$ as the reduced model. We compute $-2(LR - LF)$ which has an approximate $\mathcal{X}^2(1)$ distribution under $\alpha_2 = 0$ and find a p-value ≈ 0.

By maximum likelihood estimation we also find

$$\hat{p}(y_3|y_1, y_2, x) = \text{IL}(x'\hat{\beta} + \hat{\alpha}_1 y_1 + \hat{\alpha}_2 y_2 + \hat{\alpha}_3 y_1 y_2), \tag{6.2}$$

where

$$(\hat{\beta}, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3) = (-4.233, 0.677, 1.073, 0.197, -0.066, -0.002, -0.034, 1.451, 2.009, 0.271)'.$$

Likelihood ratio tests to test $\alpha_1 = 0, \alpha_2 = 0, \alpha_3 = 0$ returned p-values ≈ 0 for the first two tests, but the test for $\alpha_3 = 0$ had a p-value $= 0.072$. Assuming significance of interaction, the odds of difficulty in wave 3, in going from no difficulty to difficulty in wave 2, increase by approximately a factor of $\exp(2.009 + 0.271) = \exp(2.280) = 9.777$ or 877.7% if you have difficulty in wave 1 and increase by approximately a factor of $\exp(2.009) = 7.456$ or 645.6% if you don't have difficulty in wave 1. These results are similar with increases in empirical odds observed in Table 2 before considering covariates. In Table 2, odds of difficulty in wave 3 increase by a factor of (5916/577)/(750/766)=10.4718 in going from no difficulty to difficulty in wave 2 if you have difficulty in wave 1 and increase by a factor of (1138/655)/(1328/6220) = 8.175 in going from no difficulty to difficulty in wave 2 if you don't have difficulty in wave 1.

All together, both logistic regression models suggest dependence of $Y_{1,x}, Y_{2,x}$ and $Y_{3,x}$ in agreement with our non-parametric measures which don't account for x.

Table 3: Full Model Parameter Estimates

	Intercept	Gen	Race	Age	Race*Age	Race*Gen	Gen*Age	Wave	y_1, y_2
$\hat{\boldsymbol{\beta}}_1$	-3.243	0.377	0.811	0.202	**-0.038**	0.128	**0.017**	1	-,-
$\hat{\boldsymbol{\beta}}_{01}$	-4.169	0.559	0.795	0.188	**-0.033**	-0.135	**-0.008**	2	0,-
$\hat{\boldsymbol{\beta}}_{11}$	-0.739	0.279	0.413	0.146	**-0.023**	0.112	**0.004**	2	1,-
$\hat{\boldsymbol{\beta}}_{001}$	-5.094	0.773	1.781	0.260	-0.117	0.068	**-0.039**	3	0,0
$\hat{\boldsymbol{\beta}}_{011}$	-2.536	0.790	-0.074	0.182	**0.017**	0.188	-0.049	3	0,1
$\hat{\boldsymbol{\beta}}_{101}$	0.108	-1.222	0.567	0.023	-0.036	0.064	0.108	3	1,0
$\hat{\boldsymbol{\beta}}_{111}$	-0.737	1.693	0.537	0.216	**-0.018**	-0.283	-0.108	3	1,1

Gen (0=female, 1=male), Race (0=white,1=non-white), Age is divided by 5, parameters in bold font were not significant

7 Results from the Generalized Linear Model

MLE estimates of parameters in (4.5) are obtained by maximizing (4.3) after substituting link functions in (4.8) in (4.3). Maximization is done by supplying the gradient given in (4.14) to Matlab routine fminunc. The estimates for the parameters are given in Table 3.

We get an overall chi-squared for the full model by fitting a reduced model with only intercepts, i.e., we set each entry in $\boldsymbol{\beta}_1, \boldsymbol{\beta}_{j1}, \boldsymbol{\beta}_{jk1}$ in (4.5) to zero except the first in each and fit a model. The seven approximated intercepts should then be the log-odds of the main effects in Table 2. For example, in the reduced model, $\hat{\boldsymbol{\beta}}_{11} = \log(0.462/0.538) = -0.1523$ which agrees with our fitted result of the reduced model in Matlab . We then find $-2(LR - LF) = 119.3144 > \mathcal{X}^2_{0.05,42} = 58.124$ which suggests overall goodness-of-fit.

Considering (4.6), and taking into account that we included interaction terms in (3), we can use entries in each row in Table 3 to estimate the change, per unit change in a covariate, in the log-odds of difficulty in that wave given the difficulty status given in the last column and given the values of other covariates. For example, inspecting $\hat{\boldsymbol{\beta}}_{001}$, the estimated change, for every additional five years of age of a white female, in the log-odds of difficulty in wave 3 given no difficulty in the previous two waves is 0.260. This translates to a change in odds by a factor of $\exp(0.260) = 1.297$, i.e., the odds increase by 29.7%. We also see the estimated change, for every five years of age of a non-white male, in the log-odds of difficulty in wave 3 given no difficulty in the previous two waves is 0.260-0.117=0.143 which translates to a change in odds by a factor of $\exp(0.143) = 1.154$.

We do simultaneous Wald tests to test the significance of each coefficient, β, using the test statistic

$$\chi^2 \sim \frac{(\hat{\beta} - \beta_0)^2}{\text{var}(\hat{\beta})}.$$

The standard errors vary from 0.0004 to 0.0469 with the age and intercept coefficients generally with the largest standard errors. Testing the significance each coefficient simultaneously and using

Table 4: Reduced Model Estimates: Race*Age = Gen*Age = 0, Age effect fixed

	Intercept	Gen	Race	Age	Race*Gen	Wave	y_1, y_2
$\hat{\boldsymbol{\beta}}_1$	-2.957	0.625	0.333	0.179	0.100	1	x,x
$\hat{\boldsymbol{\beta}}_{01}$	-4.062	0.466	0.383	0.179	-0.113	2	0,x
$\hat{\boldsymbol{\beta}}_{11}$	-1.108	0.227	0.047	0.179	0.296	2	1
$\hat{\boldsymbol{\beta}}_{001}$	-3.993	0.204	0.196	0.179	0.246	3	0,0
$\hat{\boldsymbol{\beta}}_{011}$	-2.398	**-0.017**	-0.071	0.179	0.596	3	0,1
$\hat{\boldsymbol{\beta}}_{101}$	-1.841	0.073	**0.036**	0.179	0.413	3	1,0
$\hat{\boldsymbol{\beta}}_{111}$	-0.119	0.039	**-0.013**	0.179	0.343	3	1,1

Gen (0=female, 1=male), Race (0=white,1=non-white), Age is divided by 5, parameters in bold font were not significant

a Bonferroni correction at an overall level of 0.05, we find all parameters are significant except the coefficients highlighted in bold font in Table 3.

Using the full model we test (4.10). Fixing $\beta_{11} = \beta_{01}$ and finding a reduced model, we find $-2(LR - LF) = 468.8130 > \mathcal{X}^2_{0.05,13} = 22.3620$ which suggests conditional dependence of $Y_{2,x}$ on $Y_{1,x}$. Similarly, testing (4.9) we find $-2(LR - LF) = 536.5492 > \mathcal{X}^2_{0.05,21} = 32.6706$ which suggests conditional dependence of $Y_{3,x}$ on $Y_{2,x}$ and $Y_{1,x}$. Testing (4.11) we find $-2(LR - LF) = 996.6268 > \mathcal{X}^2_{0.05,40} = 55.7585$ which suggests conditional dependence of $Y_{2,x}$ on $Y_{1,x}$ and conditional dependence of $Y_{3,x}$ on $Y_{2,x}$ and $Y_{1,x}$.

We also fit reduced models and consider the AIC criteria, AIC $= 2p - 2L$, where L is the log-likelihood of the model. A smaller AIC is considered an improvement in the model by taking into account a larger likelihood while penalizing additional parameters. In line with Table 3, we find a reduced model without the interactions Race*Age and Gen*Age and with the effect of Age set to a constant. We find $\text{AIC}_R = 3387.4 < \text{AIC}_F = 3421.3$. For the reduced model, we have $-2(LR - LR^*) = 56.6175 > \mathcal{X}^2_{0.05,22} = 33.9244$, where in this case we signify the main effects model by R^*. Note that the overall effect of age, a 0.1789 change in the log-odds of difficulty for every five years of age or a $\exp(0.17891) = 19.59\%$ increase in probability of difficulty, is very similar to the overall estimated conditional effects found in fitted logistic regression models (6.1) and (6.2). Other reduced models can be fit in this manner depending on research interest, parameter selection criteria and results in the full model. We might look for the effect of age, for example, given no difficulty in the previous wave, and set $\hat{\beta}_{01,4} = \hat{\beta}_{001,4} = \hat{\beta}_{101,4}$ and let other age parameters vary.

8 Discussion

In this paper, in generalizing and extending the results of Uddin and Begum (2018), we considered a full generalized linear model for m-variate categorical responses each in d possible categories. This technique naturally lends itself to the analysis of longitudinal data as marginal and conditional

distributions are considered simultaneously to fit the model. Using maximum likelihood estimation, we applied this technique to panel mobility data over $m = 3$ periods with $d = 2$ binary responses.

In the full model, we fit effects of covariates on odds over all three time periods and generally found positive effects for gender (from female to male), race (from white to other race) and age (for every five years) on status of difficulty in mobility. We also examined interactions over the three time periods and generally found positive interactions of race by gender, and negative interactions of race by age and gender by age. We also considered a reduced model without race by age and gender by age interactions and with an overall effect of age. The overall effect of age was very similar to what we found when fitting the logistic models in 2.1 and 2.2.

This method can be applied generally with m time periods and d categories. However, the number of parameters in the full model grows linearly in the number of covariates and exponentially in the number of categories and time periods. For example, with $d = 2$ and p covariates the number of number of parameters to estimate is $p(2^m - 1)$ which quickly becomes untenable with increasing m. Thus, this approach is best for a small number of d categories and m time periods, perhaps starting with a reduced model depending on researcher interest. Other statistical techniques can be incorporated including imputation of missing data, dimension reduction in the presence of a large number of covariates or collinearity and the inclusion of time-varying covariates.

Also, by simulation, we observe that estimation becomes unstable when there is a very large skew in any of the observed conditional distributions of the responses. In the mobility data as shown in Table 2, the largest skew is 91.1% to 8.9% and we observe stable convergence and consistent results in the full and reduced models. Additional computational considerations and optimization algorithms for this type of model can be addressed in future research. In addition, as the size of the data set from the HRS study that we analyzed has 17,350 observations, there is an opportunity for training and validation to assess predictive power. Accordingly, we coded simple validation for our data set with 90% of our data partitioned for training and 10% partitioned for validation. Future research, along the lines of Islam and Chowdhury (2010) can focus on predictive power of the general linear models given here on the HRS data set and large data sets with multivariate, longitudual categorical responses, in general.

References

Allman, R. M., Baker, P. S., Maisiak, R. M., Sims, R. V., and Roseman, J. M. (2004), "Racial similarities and differences in predictors of mobility change over eighteen months," *Journal of General Internal Medicine*, 19, 1118–1126.

Farlex, F. M. D. (2020), "The Free Medical Dictionary," `https://medical-dictionary.thefreedictionary.com/mobility` Last accessed on 2020-06-15.

Fitzmaurice, G. M., Lipsitz, S. R., and Ibrahim, J. G. (2007), "A note on permutation tests for variance components in multilevel generalized linear mixed models," *Biometrics*, 63, 942–946.

Hardin, J. W. (2005), "Generalized estimating equations (GEE)," in *Encyclopedia of Statistics in Behavioral Science*, Wiley Online Library, pp. 721–728.

Islam, M. A., Alzaid, A. A., Chowdhury, R. I., and Sultan, K. S. (2013), "A generalized bivariate Bernoulli model with covariate dependence," *Journal of Applied Statistics*, 40, 1064–1075.

Islam, M. A. and Chowdhury, R. I. (2010), "Prediction of disease status: A regressive model approach for repeated measures," *Statistical Methodology*, 7, 520–540.

— (2017), *Analysis of repeated measures data*, Springer.

Kabiri, M., Brauer, M., Shafrin, J., Sullivan, J., Gill, T. M., and Goldman, D. P. (2018), "Long-term health and economic value of improved mobility among older adults in the United States," *Value in Health*, 21, 792–798.

Legato, M. J. and Bilezikian, J. P. (2004), *Principles of gender-specific medicine*, vol. 2, Gulf Professional Publishing.

Lipsitz, S. R., Kim, K., and Zhao, L. (1994), "Analysis of repeated categorical data using generalized estimating equations," *Statistics in Medicine*, 13, 1149–1163.

Marshall, A. W. and Olkin, I. (1985), "A family of bivariate distributions generated by the bivariate Bernoulli distribution," *Journal of the American Statistical Association*, 80, 332–338.

Miller, M. E., Davis, C. S., and Landis, J. R. (1993), "The analysis of longitudinal polytomous data: generalized estimating equations and connections with weighted least squares," *Biometrics*, 1033–1044.

Rosenbloom, S. (2005), *The mobility needs of older Americans: Implications for transportation reauthorization*, Brookings.

Satariano, W. A., Guralnik, J. M., Jackson, R. J., Marottoli, R. A., Phelan, E. A., and Prohaska, T. R. (2012), "Mobility and aging: new directions for public health action," *American Journal of Public Health*, 102, 1508–1515.

Stroup, W. W. (2012), *Generalized linear mixed models: modern concepts, methods and applications*, CRC press.

Sun, B. and Sutradhar, B. (2015), "Bivariate categorical data analysis using normal linear conditional multinomial probability model," *Statistics in Medicine*, 34, 469–486.

Sutradhar, B. C. (2014), *Longitudinal categorical data analysis*, Springer.

Uddin, M. N. and Begum, M. (2018), "A generalized linear model for multivariate correlated binary response data on mobility index," *Journal of Statistical Research*, 52, 61–73.

University of Michigan (2012-2014), "Health and Retirement Study, (Public Survey Data) public use dataset," Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740). Ann Arbor, MI, (2012-2014).

Received: July 22, 2020

Accepted: February 27, 2021