Journal of Statistical Research *ISSN 0256 - 422 X*
2018, Vol. 52, No. 1, pp. 61-73

A GENERALIZED LINEAR MODEL FOR MULTIVARIATE CORRELATED BINARY RESPONSE DATA ON MOBILITY INDEX

MD NAZIR UDDIN
Ball State University, Muncie, Indiana 47306, USA
Email: muddin@bsu.edu

MUNNI BEGUM*
Ball State University, Muncie, Indiana 47306, USA
Email: mbegum@bsu.edu

SUMMARY

Dependence in multivariate binary outcomes in longitudinal data is a challenging and an important issue to address. Numerous studies have been performed to test the dependence in binary responses either using conditional or marginal probability models. Since the conditional and marginal approach provide inadequate or misleading results, the joint models based on both are implemented for bivariate correlated binary responses. In the current paper, we consider a joint modeling approach and a generalized linear model (GLM) for tri-variate correlated binary responses. The link function of the GLM is used to test the dependence of response variables. The mobility index with two categories, no difficulty and difficulty, over the duration of three waves of Health and Retirement Survey (HRS) is chosen as the binary response variable. Initial analysis with Marshall-Olkin correlation coefficients and logistic regression coefficients provide moderate correlation in mobility indices implying dependence in the response variables. We also found statistically significant dependence among the response variables using the joint modeling approach. The mobility at current wave not only depends on the previous mobility status, but also depends on covariates such as age, gender, and race.

Keywords and phrases: Multivariate Correlated Binary Responses, Dependency, Generalized Linear Models, Mobility Index.

1 Introduction

Dependence in multivariate response variables collected over time in longitudinal studies is both a challenging and an important issue to address while analyzing the relationship among the correlated response variables and a set of explanatory variables. In modeling repeated measures outcome variables, researchers need to incorporate both the dependence among the repeated outcome measures and the dependence among these measures and a set of explanatory variables. The literature in modeling correlated quantitative response variables that follow multivariate normal distribution is quite

* Corresponding author

rich. However, methodologies for modeling discrete repeated measures outcomes are scattered. Repeated or correlated binary outcome measures are a special case of discrete repeated response variables that arise commonly in many fields including biomedical sciences, public health, epidemiology, agriculture, business, economics, and social sciences.

The generalized estimating equations (GEE) are the most commonly used methodologies (Liang and Zeger, 1986; Hardin and Hilbe, 2002; Diggle et al., 2002) for modeling repeated binary outcome measures and a set of explanatory variables. GEE methods implement marginal models and a working correlation structure to address the dependence between the outcome measures and a set of explanatory variables and the dependence among the outcome measures respectively. Marginal models implemented in GEE are basically population-averaged fixed-effects models where the dependence among the outcome measures are incorporated by assuming some prespecified correlation structures on them. A second approach that explicitly models the dependence among the repeated outcome measures, is the generalized linear mixed effects model (GLMM) (Stroup, 2016; Fitzmaurice et al., 2012) where the random variation among the group of repeated measures is added explicitly to the systematic part of the model.

In recent literature on repeated measures analysis (Islam et al., 2012, 2013; Islam and Chowdhury, 2017) both marginal and conditional models for outcome measures are considered to analyze a number of non-normal repeated response variables collected over time. In the current paper, we consider such approaches for repeated binary outcome measures collected from a longitudinal study. In particular, we consider marginal and conditional models for tri-variate binary outcome measures on mobility index of elderly people in the USA from a longitudinal household survey on Health and Retirement Study (HRS). The mobility indices collected over three waves (years) of elderly people are considered as response variables.

The term mobility is defined as the condition of moving independently. In general, the ability of moving from one place to another, decreases as people become older. From the public health point of view it is important to study the dependence structure of mobility over time for elderly people as well as the dependence of mobility on other associated factors. Mobility issue in elderly people gained great attention in recent time. Numerous studies have been conducted to identify the factors that influence mobility. Truong et al. (2011) studied mobility of older people and identified that age, gender, income, living status (alone or accompanied), and neighborhood characteristics were the most influential factors that might affect mobility. Schwanen and Páez (2010) identified that factors such as gender and ethnicity might vary across geographical space in relation to mobility.

The primary objectives of the current paper are: (i) to develop a generalized linear model for multivariate correlated binary response variables (i.e. mobility indices, in this case), (ii) to establish a testing procedure to test the strength of dependence among the response variables, and (iii) to identify the statistically significant factors that affect the mobility index.

2 Background

Dependence in categorical variables particularly in binary outcome variables is studied for a long time in different contexts. There had been many substantial contributions to statistical theory for

determining the dependence across binary response variables.

For longitudinal data analysis, Liang and Zeger (1986) and Prentice (1988) proposed the method of moments to analyze binary response variables. The generalized estimating equations (GEE) technique based on a marginal model was used to estimate the parameters associated with the binary response variables and the correlation among the repeated outcomes was addressed with a working correlation structure. Lipsitz et al. (1991) then modified the GEE proposed by Prentice (1988) to estimate the odds ratio providing more efficient measures of association between binary response variables. Liang et al. (1992) and Carey et al. (1993) also used the odds ratio to measure the association between binary responses. Different association measures such as odds ratio and tetrachoric correlation were used by Le Cessie and Van Houwelingen (1994). Introduction to generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) set the stage for analyzing multivariate binary responses in many applications. To analyze the repeated measures data with correlated binary outcomes, Darlington and Farewell (1992) showed that the relationship between response and predictors can be described by the dependence in response as well as the dependence on the predictors. They focused on marginal probabilities and dependence on predictors. Islam et al. (2012) extended their model by considering both marginal and conditional probabilities. In their proposed model, both marginal and conditional probabilities were expressed as a function of predictors and they suggested a testing procedure to check the dependence of response variables. One drawback of the model is that without a joint model for the correlated outcomes, models based on marginal or conditional probabilities alone can not solve the problem of dependence in the outcomes entirely.

To overcome this problem, Islam et al. (2013) and Islam and Chowdhury (2017) proposed a joint model which takes both marginal and conditional probabilities of correlated binary response variables. The application was carried out on a bi-variate binary response. We implemented their proposed model to tri-variate correlated binary responses in longitudinal data. This approach can be carried out in a similar manner to multivariate binary response data with time dependent predictors.

3 Methodology

3.1 Terminology and Notations

We begin with the notations and terminologies associated with the methodology for analyzing correlated tri-variate binary responses. Let Y_1 be the binary response variable at time point 1. For $Y_1 = m$, $m = 0, 1$, let Y_{m2} and Y_{m3} denote two additional binary response variables at time points 2 and 3 respectively given the mth status of the response variable at time point 1.

The joint probability of Y_{m2} and Y_{m3} can be illustrated using a 2×2 table (Table 1), where P_{m+0}, P_{m+1}, P_{m0+} and P_{m1+} represent the marginal probabilities and P_{m00}, P_{m01}, P_{m10} and P_{m11} are the joint probabilities given the mth status of the response at time point 1. In particular $P_{mij}, (i, j) = \{(0, 0), (0, 1), (1, 1), (1, 0)\}$ denote the joint probabilities of the response variables at time points 2 and 3 given that the response variable at time point 1 takes value $m(m = 0, 1)$. P_{mi+} and P_{m+j} denote the marginal probabilities of the response variables at time points 2 and 3

Table 1: Joint probability distribution

		Y_{m3}		
		0	1	Total
Y_{m2}	0	P_{m00}	P_{m01}	P_{m0+}
	1	P_{m10}	P_{m11}	P_{m1+}
Total		P_{m+0}	P_{m+1}	P_{m++}

respectively given that the response variable at time point 1 is $m(m = 0, 1)$.

Under this setup we can test the dependence between the response variables at time points 2 and 3 using Marshall-Olkin correlation coefficient (Marshall and Olkin, 1985) which is defined as,

$$\rho_M = \frac{P_{m00}P_{m11} - P_{m10}P_{m01}}{\sqrt{P_{m1+}P_{m0+}P_{m+1}P_{m+0}}}. \tag{3.1}$$

If the estimated correlation coefficient $\hat{\rho}_M$ is close to zero, then it concludes that there is no correlation between the response variables 2 and 3 given the response at time point 1.

Another approach to test the dependence between two correlated binary response variables is to consider a logistic regression model that can be expressed as follows:

$$p\big(Y_{m3}|y_{m2}, x\big) = \frac{e^{(X\boldsymbol{\beta}+\alpha_{m2}y_{m2})}}{1 + e^{(X\boldsymbol{\beta}+\alpha_{m2}y_{m2})}} \quad ; m = 0, 1 \tag{3.2}$$

This representation of a logistic model is a Markov type model since the response variable at time point 2 (Y_{m2}) is considered as an explanatory variable to model the response variable at time point 3 (Y_{m3}) along with other explanatory variables X. If the regression coefficient α_{m2} becomes zero then the response variables Y_{m2} and Y_{m3} are said to be independent irrespective of Y_1. On the other hand, if

$$\begin{cases} \alpha_{m2} = 0, & \text{for } m = 0 \\ \alpha_{m2} \neq 0, & \text{for } m = 1 \end{cases}$$

then the responses Y_{m2} and Y_{m3} are said to be conditionally independent and it is true for reciprocal condition on α_{m2}.

3.2 Data and Variable Selection

We consider secondary data from the longitudinal household survey on Health and Retirement Study (HRS). The survey was sponsored by the National Institute on Aging (NIA) and Social Security Administration (SSA). The HRS was administered by the Institute for Social Research (ISR), University of Michigan. The survey was conducted on individuals aged 50 years and higher and their spouses. Data were collected on demographics such as health, financial and housing wealth, income, social security, pension, health insurance, family structure, retirement plan, and employment

history from both the respondent and his/her spouse, if any. The main goal is to provide panel data that enable research and analysis in support of policies on retirement, health insurance, saving, and economic well-being.

The data set contains 12 waves (years) denoted as wave 1 (W1:1992), wave 2 (W2:1994) and so on up to wave 12 (W12:2014). In the current study, we consider last three waves, W10-W12 and rephrased W10, W11, and W12 as waves 1-3. The mobility index among elderly people is considered as the response variable. The mobility index was calculated based on the five tasks such as walking several blocks, walking one block, walking across the room, climbing several flights of stairs and climbing one flight of stairs. The sum of the mobility index ranges from 0 to 5 and the difficulty level is defined as no difficulty with sum zero and difficulty with sum in one to five. For the k^{th} $(k = 1, 2, 3)$ wave, the response variable is defined as,

$$Y_k = \begin{cases} 0, & \text{if no difficulty} \\ 1, & \text{if difficulty} \end{cases}$$

The mobility index may depend on many factors including age, sex, race and body mass index. For our analysis we considered time invariant covariates such as age (X_1), gender (X_2) and race (X_3). The predictors are defined as:

$$X_1 = \begin{cases} 0, & \text{if age is 50-60 years} \\ 1, & \text{if age } >60 \end{cases}, X_2 = \begin{cases} 0, & \text{if Female} \\ 1, & \text{if Male} \end{cases}, X_3 = \begin{cases} 0, & \text{if White/Caucasian} \\ 1, & \text{if Others} \end{cases}$$

The complete data set is separated into two groups. The first group is constructed by taking all respondents' information who are at no difficulty status of movement and the other group contains information who are at difficulty status of movement at wave-1. A sample of size 20295 is selected for the analysis. The data set is further splitted according to the mobility index status at wave 1. With no difficulty status at wave 1, the sample has 10416 individuals and with difficulty status at wave 1, the sample has 10509 individuals. All the missing information in both data sets are assumed to be missing completely at random and is excluded from further analysis.

3.3 A Generalized Linear Model Approach

Regression techniques based on Markov type models as in (3.2) may fail to identify the correct relationship between responses and predictors especially when there are more than one response variable. In order to investigate the correlations among the response variables at multiple time points and a set of explanatory variables a generalized linear model (GLM) is considered based on conditional and marginal probability distributions of the response variables in two time points given the response at a third time point. Thus we are able to address the correlatedness among tri-variate binary responses and a set of explanatory variables.

For $Y_1 = m$, $m = 0, 1$ the bi-variate Bernoulli distribution for outcome variables at time points 2 and 3, $Y_{m3} = y_{m3}$ and $Y_{m2} = y_{m2}$ can be written as

$$P\big(Y_{m2} = y_{m2}, Y_{m3} = y_{m3}\big) = P_{m00}^{(1-y_{m2})(1-y_{m3})} P_{m01}^{(1-y_{m2})y_{m3}} P_{m10}^{y_{m2}(1-y_{m3})} P_{m11}^{y_{m2}y_{m3}}. \qquad (3.3)$$

The joint probabilities can be expressed in terms of conditional and marginal probabilities as follows:

$$P(Y_{m2} = y_{m2}, Y_{m3} = y_{m3}) = P(Y_{m3} = y_{m3}|Y_{m2} = y_{m2})P(Y_{m2} = y_{m2}).$$

In the presence of covariate information, the bi-variate probabilities can be written as a function of covariates X's as follows:

$$P\big(Y_{m2} = y_{m2}, Y_{m3} = y_{m3}|X = x\big) = P\big(Y_{m3} = y_{m3}|Y_{m2} = y_{m2}; x\big)P\big(Y_{m2} = y_{m2}|X = x\big).$$

The exponential family representation of the joint probability mass function of the response variables at time points 2 and 3 in equation (3.3) can be written as follows:

$$\begin{aligned} P(Y_{m2} = y_{m2}, Y_{m3} = y_{m3}) = \exp\Bigg[& y_{m2} \log\Big(\frac{P_{m10}}{P_{m00}}\Big) + y_{m3} \log\Big(\frac{P_{m01}}{P_{m00}}\Big) \\ & + y_{m2}y_{m3} \log\Big(\frac{P_{m00}P_{m11}}{P_{m01}P_{m10}}\Big) + \log P_{m00}\Bigg], \end{aligned} \tag{3.4}$$

where $(Y_{m2}, Y_{m3}) = \{(0,0), (0,1), (1,0), (1,1)\}$, $\sum_{i,j} P_{mij} = 1$, $m = 0, 1$. The log likelihood function is then expressed as follows:

$$\begin{aligned} l = \sum_{i=1}^{n} \Big[& y_{2mi} \log\Big(\frac{P_{m10i}}{P_{m00i}}\Big) + y_{3mi} \log\Big(\frac{P_{m01i}}{P_{m00i}}\Big) \\ & + y_{2mi}y_{3mi} \log\Big(\frac{P_{m00i}P_{m11i}}{P_{m01i}P_{m10i}}\Big) + \log P_{m00i}\Big]. \end{aligned} \tag{3.5}$$

The components of the link function for the generalized model are composed from the exponential representation of the joint mass function in equation (3.4) as follows:

$$\eta_{m0} = \log\big(P_{m00}\big),\ \eta_{m1} = \log\Big(\frac{P_{m01}}{P_{m00}}\Big),\ \eta_{m2} = \log\Big(\frac{P_{m10}}{P_{m00}}\Big),\ \eta_{m3} = \log\Big(\frac{P_{m00}P_{m11}}{P_{m01}P_{m10}}\Big),$$

where η_{m0} is the baseline link function, η_{m2} is the link function for Y_{m2}, η_{m1} is the link function for Y_{m3} and η_{m3} is the link function for the dependence between Y_{m2} and Y_{m3}. To write the systematic components for generalized linear model via three link functions we require to write the joint probabilities in terms of conditional and marginal probabilities.

Let $X = (X_1, \ldots, X_p)'$ be the covariate vector with x representing the realized covariate vector. Then conditional probabilities for response Y_{m3} at time point 3 given the response Y_{m2} at time point 2 can be expressed as function of covariates as follows:

$$\begin{aligned} P(Y_{m3} = 1|Y_{m2} = 0; x) &= \frac{e^{x\beta_{m01}}}{1 + e^{x\beta_{m01}}} = \pi_{m01}(x), \\ P(Y_{m3} = 1|Y_{m2} = 1; x) &= \frac{e^{x\beta_{m11}}}{1 + e^{x\beta_{m11}}} = \pi_{m11}(x), \\ P(Y_{m3} = 0|Y_{m2} = 0; x) &= \frac{1}{1 + e^{x\beta_{m01}}} = \pi_{m00}(x), \\ P(Y_{m3} = 0|Y_{m2} = 1; x) &= \frac{1}{1 + e^{x\beta_{m11}}} = \pi_{m10}(x), \end{aligned} \tag{3.6}$$

where $\beta_{m01} = \left(\beta_{m010}, \beta_{m011}, \beta_{m012}, \ldots, \beta_{m01p}\right)'$ and $\beta_{m11} = \left(\beta_{m110}, \beta_{m111}, \beta_{m112}, \ldots, \beta_{m11p}\right)'$.

Next we need the marginal probabilities for the response Y_{m2} at time point 2:

$$P(Y_{2m} = 1|X = x) = \pi_{m2}(x) = \frac{e^{x\beta_{m1}}}{1 + e^{x\beta_{m1}}}$$

and

$$P(Y_{2m} = 0|X = x) = 1 - \pi_{m2}(x) = \frac{1}{1 + e^{x\beta_{m1}}},$$

where $\beta_{m1} = \left(\beta_{m10}, \beta_{m11}, \beta_{m12}, \ldots, \beta_{m1p}\right)'$. Now combining the conditional and marginal probabilities, the joint probabilities can be written as:

$$\begin{aligned}
P_{m01}(x) &= P(Y_{m3} = 1|Y_{m2} = 0, X = x)P(Y_{m2} = 0|X = x) = \frac{e^{x\beta_{m01}}}{1 + e^{x\beta_{m01}}} \cdot \frac{1}{1 + e^{x\beta_{m1}}} \\
P_{m00}(x) &= P(Y_{m3} = 0|Y_{m2} = 0, X = x)P(Y_{m2} = 0|X = x) = \frac{1}{1 + e^{x\beta_{m01}}} \cdot \frac{1}{1 + e^{x\beta_{m1}}} \\
P_{m11}(x) &= P(Y_{m3} = 1|Y_{m2} = 1, X = x)P(Y_{m2} = 0|X = x) = \frac{e^{x\beta_{m11}}}{1 + e^{x\beta_{m11}}} \cdot \frac{e^{x\beta_{m1}}}{1 + e^{x\beta_{m1}}} \\
P_{m10}(x) &= P(Y_{m3} = 0|Y_{m2} = 1, X = x)P(Y_{m2} = 1|X = x) = \frac{1}{1 + e^{x\beta_{m11}}} \cdot \frac{e^{x\beta_{m1}}}{1 + e^{x\beta_{m1}}}.
\end{aligned} \tag{3.7}$$

Using the joint probabilities in (3.7), the components of the link function are expressed as follows:

$$\begin{aligned}
\eta_{m0} &= \log P_{m00}(x) = -\log\left[1 + e^{x\beta_{m01}}\right] - \log\left[1 + e^{x\beta_{m1}}\right], \\
\eta_{m1} &= \log\left(\frac{P_{m01}}{P_{m00}}\right) = x\beta_{m01}, \\
\eta_{m2} &= \log\left(\frac{P_{m10}}{P_{m00}}\right) = x\beta_{m1} + \log\left[1 + e^{x\beta_{m01}}\right] - \log\left[1 + e^{x\beta_{m11}}\right], \\
\eta_{m3} &= \log\left(\frac{P_{m00}P_{m11}}{P_{m01}P_{m10}}\right) = x(\beta_{m11} - \beta_{m01}).
\end{aligned} \tag{3.8}$$

If the value of η_{m3} equals to 0 for $m = 0, 1$, then the responses at time points 2 and 3 are independent. This implies that, the test of uncorrelatedness between Y_{m2} and Y_{m3} can be carried out by setting $\beta_{m11} = \beta_{m01}$.

The method of maximum likelihood estimation is used to estimate the parameters of the model presented in (3.8). Due to lack of analytical solutions, numerical methods based on quasi-Newton method (BFGS algorithm) is used to find numerical approximations to maximum likelihood estimates of the parameters. The *optim()* function of R software (R Core Team, 2018) is applied to perform the computations.

In order to test the dependence structure among the response variables across two time points given the status of the response at the third time point, we proceed as follows: the response variables at time points 2 and 3 are said to be independent if $\eta_{m3} = 0$ for $m = 0, 1$. Equivalently, we can test whether $H_0 : \beta_{m01} = \beta_{m11}$. The hypothesis can be tested using the following test statistic:

$$\chi^2 = (\hat{\beta}_{m01} - \hat{\beta}_{m11})'[var(\hat{\beta}_{m01} - \hat{\beta}_{m11})]^{-1}(\hat{\beta}_{m01} - \hat{\beta}_{m11}), \tag{3.9}$$

which is distributed as chi-squared with $(p+1)$ degrees of freedom asymptotically.

4 Data Analysis and Results

4.1 Exploratory Data Analysis

As discussed in *Data and Variable Selection* section, we consider the last three waves, W10-W12 from the Health and Retirement Study (HRS) in the current paper. The mobility indices (redefined as 0: no difficulty, and 1: difficulty) of elderly people at waves 10, 11, and 12 are considered as the response variables at time points 1, 2, and 3 denoted by $Y_1 - Y_3$ respectively. Table 2 shows the frequency distribution of the responses at last two waves (Y_2, Y_3) given the mobility status of the subjects at wave 1. About eighty percent of the respondents had no difficulty whereas about twenty percent had difficulty with mobility at wave 2 when the mobility status at wave 1 was no difficulty. For the same mobility status at wave 1, the percentage of respondents with no difficulty decreased to about seventy three percent and the percentage of respondents with difficulty status increased to twenty seven percent at wave 3.

Table 2: Frequency distributions of responses at wave 2 and 3

	Y_{02}		Y_{03}		Y_{12}		Y_{13}	
	Freq	%	Freq	%	Freq	%	Freq	%
no difficulty	7060	80.4	6424	73.1	1443	18.6	1261	16.27
difficulty	1721	19.5	2357	26.8	6307	81.3	6489	83.7

From Table 2 we see that when the mobility status of the respondents at wave 1 was difficulty, the above scenario changes drastically. Only little over eighteen percent of the respondents are at no difficulty status at wave 2 and reduces to about sixteen percent at wave 3. Thus about eighty two and eighty four percent of the respondents are at difficulty status at waves 2 and 3 respectively.

Table 3: Transition counts and probabilities from Y_2 and Y_3 given Y_1; probabilities are in parenthesis

		Y_{03}					Y_{13}		
		0	1	Total			0	1	Total
Y_{02}	0	5796 (0.821)	1264 (0.179)	7060 (1.0)	Y_{12}	0	718 (0.498)	725 (0.502)	1443 (1.0)
	1	628 (0.365)	1093 (0.635)	1721 (1.0)		1	543 (0.086)	5764 (0.914)	6307 (1.0)

Table 3 shows the transition counts and probabilities of mobility status from wave 2 to wave 3 for people who had difficulty and no difficulty at wave 1. The probability that people with no difficulty at waves 1 and 2 tend to be at the same state at wave 3 is 82% while the probability that people with difficulty at wave 1 will be at the same state at waves 2 and 3 is about 64%. However, the probability

that the people having difficulty with mobility at wave 1 will have no difficulty at waves 2 and 3 is about 50% while the probability that they will have difficulty at waves 2 and 3 is higher than 91%. In what follows, we explore the interdependence of the response variables (mobility indices) at the three time points (W10-W12).

4.2 Test of Dependence

One of the main goals of the study is to test the dependence of binary response variables at waves 2 and 3 given the values of the response variable at wave 1. Since both response variables are binary, Marshall-Olkin correlation coefficient is used to test the dependence in mobility at waves 2 and 3. The Marshall-Olkin correlation coefficients of the dependence between Y_2 and Y_3 is 0.406 when $Y_1 = 0$ and is 0.4356 for $Y_1 = 1$. The correlation coefficients indicate that the association between the mobility indices for both models are moderately positive.

The Markov model analysis with logistic regression where the mobility index at wave 2 is considered as an explanatory variable in addition to all explanatory variables also provides the evidence of dependence between the mobility indices at waves 2 and 3 given the mobility status at wave 1. The predicted logistic regression model for $Y_1 = 0$ is written as:

$$p\big(Y_{03} = 1|y_{02}, x, y_1 = 0\big) = \frac{e^{-2.0259+0.55x_1+0.24x_2+0.26x_3+2.01y_{02}}}{1 + e^{-2.0259+0.55x_1+0.24x_2+0.26x_3+2.01y_{02}}}. \tag{4.1}$$

The equation in (4.1) gives the probability of having difficulty in movement at wave 3 for the given value of predictors, when a respondent does not face any difficulty in movement at waves 1 and 2. The logistic regression for $Y_1 = 1$ is expressed as,

$$p\big(Y_{13} = 1|y_{12}, x, y_1 = 1\big) = \frac{e^{-0.40+0.41x_1+0.19x_2+0.14x_3+2.32y_{12}}}{1 + e^{-0.40+0.41x_1+0.19x_2+0.14x_3+2.32y_{12}}}. \tag{4.2}$$

The equation in (4.2) is the probability of having difficulty in movement at wave 3 for the given values of predictors when a respondent does have difficulty in movement at waves 1 and 2. The regression coefficients of y_{02} and y_{12} indicate that there is large amount of evidence for dependence between the mobility indices at waves 2 and 3 for given the mobility status at wave 1.

4.3 Results from the Generalized Linear Model

The parameter estimates from the generalized linear models based on the conditional and marginal probability models given that the respondents had no difficulties at wave 1 (i.e. $Y_1 = 0$) are presented in Table 4.

The p-value of the dependence test when the respondents had no difficulties in movement at wave 1 (i.e. $Y_1 = 0$) indicates that their mobility states at waves 2 and 3 are dependent. In other words, there is statistically significant dependence between the mobility indices at waves 2 and 3 when the respondents had no difficulty at wave 1. We also see from Table 4 that all the covariates are significant in both conditional and marginal models when mobility states were no difficulty and difficulty at waves 2 and 3 given no difficulty at wave 1. When mobility is at difficulty at both

Table 4: Parameter estimation of the proposed model for $Y_1 = 0$ ($\chi^2_0 = 292.131$, p-value=0.000)

	conditional model β_{001}			conditional model β_{011}			marginal model β_{01}		
$Predictors$	Estimate	SE	p-value	Estimate	SE	p-value	Estimate	SE	p-value
Const.	-2.077	0.0664	0.000	0.135	0.112	0.222	-2.038	0.0596	0.000
Age	0.615	0.0653	0.000	0.391	0.1071	0.000	0.569	0.0571	0.000
Gender	0.262	0.0627	0.000	0.217	0.1020	0.033	0.417	0.0548	0.000
Race	0.298	0.0723	0.000	0.167	0.1169	0.152	0.274	0.062	0.000

waves 2 and 3 given no difficulty wave 1 then the covariate race becomes insignificant. That is, the odds of having difficulty at wave 3 is not affected by race given that a subject had difficulty in the previous year. However, in the absence of difficulty in the prior year, race may be a significant factor of mobility status. Thus when the respondents had no difficulty in movement at wave 1, age and gender were the two factors that affected movement status consistently for the respondents at wave 2 to wave 3. These results are consistent with respect to the risk factors in the progression of difficulty in movement for the elderly people.

Table 5 presents the regression parameter estimates from the generalized linear models based on conditional and marginal probability models given that the respondents had difficulties at wave 1 (i.e. $Y_1 = 1$).

Table 5: Parameter estimation of the proposed model for $Y_1 = 1$ ($\chi^2_1 = 92$, p-value=0.000)

	conditional model β_{101}			conditional model β_{111}			marginal model β_{11}		
$Predictors$	Estimate	SE	p-Value	Estimate	SE	p-value	Estimate	SE	p-value
Const.	-0.395	0.120	0.001	1.9131	0.105	0.000	0.968	0.067	0.000
Age	0.3828	0.112	0.000	0.4369	0.096	0.000	0.373	0.062	0.000
Gender	0.1885	0.108	0.081	0.1993	0.094	0.034	0.354	0.0603	0.000
Race	0.1923	0.120	0.109	0.1042	0.102	0.307	0.118	0.066	0.076

Similar to the case presented in Table 4, the p-value of the dependence test with respondents having difficulty in movement at wave 1 (i.e. $Y_1 = 1$) indicates that the responses at waves 2 and 3 are dependent. Thus we conclude that regardless of the mobility status at wave 1 there is statistically significant dependence in mobility indices at waves 2 and 3. From Table 5 we see that only age is significant in both conditional and marginal models. Gender is not significant in the conditional model and race is not significant in conditional and marginal models. Thus when the respondents had difficulty in movement at wave 1, only age affected consistently the difficulty in movement status for the respondents at waves 2 and 3.

5 Conclusion

Dependency in multiple correlated categorical response variables is challenging to explore. A number of studies are conducted to test the dependence in multiple categorical response variables using various forms of correlation coefficients and regression analysis. In recent years, generalized estimating equations and logistic regression models are considered for testing dependency among binary response variables at multiple time points. Other methodological development includes generalized linear models based on marginal or conditional probability models. However, these techniques do not provide true nature of dependency since they use a single probability model. In this paper, we consider a generalized linear model technique based on both conditional and marginal probabilities. The model is applied to practical data which provides adequate information for testing the dependence for tri-variate binary outcomes at three time points.

Our initial approach for testing dependency among the response variables was to implement Marshall- Olkin correlation coefficient. Next we considered the logistic regression technique based on Markov model. Although both of these approaches indicate dependence in response variables, the results are not adequate to describe the true relation since these approaches demonstrate only conditional dependence. In order to study the dependence in tri-variate binary responses we adopted a generalized linear model based on joint models of responses at the last two waves conditional on the response at the first wave.

The response variable at wave 1, Y_1 is categorized into two categories with labels of no difficulty ($Y_1 = 0$) and difficulty ($Y_1 = 1$). Conditional to $Y_1 = m, m = 0, 1$ we tested dependency among the response variables at waves 2 and 3. The results from the generalized linear model demonstrate that the mobility status in two consecutive years depends on the mobility status in the previous year. In addition, the test of dependence indicates strong correlation among the status of mobility at waves 2 and 3 conditional to the mobility status at wave 1. Our results are comparable to those presented by Islam and Chowdhury (2017).

In brief, the proposed analysis is performed under two conditions, testing the dependency of mobility indices at last two waves when the movement of elderly people is at (i) not difficult and (ii) difficult at the first wave. Under both scenarios, the mobility of elderly people depend primarily on respondent's age and gender in addition to the prior mobility status. Although prior studies found race to affect mobility status significantly, results from the current study do not support this evidence. One explanation could be a large number of missing values for this covariate.

Due to the longitudinal nature of the responses, conditional and joint probability models for the mobility status at waves 2 and 3 conditioning on wave 1 are considered. The mobility status of a respondent at wave 3 can only depend on mobility at waves 2 and 1. As such, the following cases are not considered in our analysis: i) conditional and joint probability models for the mobility status at waves 1 and 2 conditioning on wave 3 and ii) conditional and joint probability models for the mobility status at waves 1 and 3 conditioning on wave 2.

One of the drawbacks of the research is that we considered only time invariant covariates due to lack of sufficient data. Time variant covariates may provide further insight to the dependency nature of bi-, tri-, and in general multi-variate responses as suggested in the literature. Nontheless, a general algorithm can be proposed to test the dependence in multi-variate binary outcomes of

longitudinal data by simple implementation of this approach to responses collected at more than three time points. As in any bio-medical or public health research, missing values are common in both response and covariate variables. Time variant covariates and incorporation of missing values are left as future research.

References

Carey, V., Zeger, S. L., and Diggle, P. (1993), "Modelling multivariate binary data with alternating logistic regressions," *Biometrika*, 80, 517–526.

Darlington, G. and Farewell, V. (1992), "Binary longitudinal data analysis with correlation a function of explanatory variables," *Biometrical Journal*, 34, 899–910.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K.-Y., Zeger, S., et al. (2002), *Analysis of longitudinal data*, Oxford University Press.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012), *Applied longitudinal analysis*, vol. 998, John Wiley & Sons.

Hardin, J. W. and Hilbe, J. M. (2002), *Generalized estimating equations*, Chapman and Hall/CRC.

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

Islam, M. A., Chowdhury, R. I., and Briollais, L. (2012), "A bivariate binary model for testing dependence in outcomes," *Bull. Malays. Math. Sci. Soc.(2)*, 35, 845–858.

Le Cessie, S. and Van Houwelingen, J. (1994), "Logistic regression for correlated binary data," *Journal of the Royal Statistical Society, series C*, 95–108.

Liang, K.-Y. and Zeger, S. L. (1986), "Longitudinal data analysis using generalized linear models," *Biometrika*, 73, 13–22.

Liang, K.-Y., Zeger, S. L., and Qaqish, B. (1992), "Multivariate regression analyses for categorical data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–40.

Lipsitz, S. R., Laird, N. M., and Harrington, D. P. (1991), "Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association," *Biometrika*, 78, 153–160.

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.

McCullagh, P. and Nelder, J. A. (1989), *Generalized linear models*, vol. 37, CRC press.

Nelder, J. and Wedderburn, R. (1972), "General linearized models," *Journal of the Royal Statistical Society, Series A*, 135, 370–384.

Prentice, R. L. (1988), "Correlated binary regression with covariates specific to each binary observation," *Biometrics*, 1033–1048.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Schwanen, T. and Páez, A. (2010), "The mobility of older people: an introduction," *Journal of Transport Geography*, 18, 591–595.

Stroup, W. W. (2016), *Generalized linear mixed models: modern concepts, methods and applications*, CRC press.

Truong, L. T., Somenahalli, S., et al. (2011), "Exploring mobility of older people: a case study of Adelaide," Australasian Transport Research Forum.

Received: August 28, 2017

Accepted: August 28, 2018