BULLETIN *of the*
MALAYSIAN MATHEMATICAL
SCIENCES SOCIETY
http://math.usm.my/bulletin

Bull. Malays. Math. Sci. Soc. (2) **35***(4) (2012), 845–858*

A Bivariate Binary Model for Testing Dependence in Outcomes

[1]M. ATAHARUL ISLAM, [2]RAFIQUL I CHOWDHURY AND [3]LAURENT BRIOLLAIS
[1]Department of Statistics and OR, King Saud University, PO Box 2455, Riyadh 11451, Saudi Arabia
[2]Department of Epidemiology and Biostatistics, University of Western Ontario, Canada
[3]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada
[1]mataharul@yahoo.com, [2]mchowd23@uwo.ca, [3]laurent@lunenfeld.ca

Abstract. The problem of dependence in the outcome variables has become an increasingly important issue of concern during the past two decades attributable mainly to the increase in the demand for techniques in analyzing repeated measures data. In the past, most of the longitudinal models developed are based on marginal approaches and relatively few are based on conditional models. The joint models are examined mainly to focus on the characterization problems but not much has been employed to focus the covariate dependent models with dependence in the outcomes. This paper develops a new simple procedure to take account of the bivariate binary model with covariate dependence. The model is based on the integration of conditional and marginal models. Test procedures are suggested for testing the dependence in binary outcomes. Simulations are employed to demonstrate the utility of the proposed test procedures in different dependence settings. Finally, an application to the depression data has been shown. All the results confirm that the proposed model for testing the dependence in outcomes can be applied very successfully for a wide variety of situations.

2010 Mathematics Subject Classification: 62H20, 62F03, 62F10, 62N03

Keywords and phrases: Conditional model, marginal model, joint model, correlated outcomes, test for dependence, bivariate Bernoulli.

1. Introduction

The Bernoulli distribution has a very important role and is connected with univariate distributions such as binomial, geometric, negative binomial, Poisson, gamma, hypergeometric, exponential, normal etc. either as a limit or as a sum or other functions. In the univariate case, there is a family of interrelated distributions. Marshall and Olkin [27] demonstrated that some distributions arise naturally from bivariate Bernoulli distribution as well. At an earlier time, Antelman [1] also suggested some interrelated Bernoulli processes.

The importance of the bivariate, or more specifically, multivariate Bernoulli has increased since the introduction of the generalized linear models [28], and more so, after the extension for repeated measures since the work of Zeger and Liang [40] was published on the generalized estimating equation (GEE). In the simplest case, we may assume that the marginal variables are also independent for each subject. Then the analysis reduces to

Communicated by Anton Abdulbasah Kamil.
Received: November 26, 2011; *Revised:* April 13, 2012.

a standard generalized linear model [28, 29]. In longitudinal studies, we need to deal with repeated binary outcomes which are correlated. Liang and Zeger [21] and Prentice [33] proposed the GEE models based on probability of the event and correlations or the first and the second moments. On the other hand, Lipsitz *et al.* [24], Liang *et al.* [22] and Carey *et al.* [8] employed the marginal odds ratios, instead of correlations between pairs of binary responses [14]. Cessie and Houwelingen [9] proposed use of different measures of dependence in modeling for logistic regression for correlated binary data. A marginal model of multivariate categorical data was proposed by Molenberghs and Lesaffre [30]. These marginal models may fail to provide efficient estimation of parameters due to lack of proper specification of the dependence of binary outcomes in the model. Azzalini [2] proposed a marginal model based on binary Markov Chain for a single stationary process $(Y_1, \ldots, Y_T)$ where Y's take values of 0 or 1 at subsequent times.

The quadratic exponential form model has been proposed on the basis of the Bahadur representation [3]. Cox [10] showed that the multivariate binary probability can be explained by a quadratic exponential form by employing the logistic regression model. This was further studied by Zhao and Prentice [41], Cox and Wermuth [11] and Lee and Jun [25]. A generalized multivariate logistic model was proposed by Glonek and McCullagh [15] and Glonek [16]. A marginal modeling of correlated data was proposed by Molenberghs and Lesaffre [31] using a multivariate Plackett distribution. In their model, they considered three link functions for marginal and association parameters. This approach has been studied in the setting of log-linear models [4, 5, 34]. Wakefield [38] summarizes the limitation of the marginal models with specific reference to the well known Simpson's paradox [35]. Some of the recent expositions of the bivariate and multivariate Bernoulli approaches include Juan and Vidal [19], Lin and Clayton [23], Lovison [26], Sun *et al.* [36], Yee and Dronbock [39] and Lee and Jun [25]. The conditional approach was shown by Bonney [6, 7], Muenz and Rubinstein [32], Islam and Chowdhury [17], and Islam *et al.* [18]. Bonney's regressive model approach takes previous outcome into account in addition to covariates. This model also fails to address the dependence in the binary outcomes unconditionally. Darlington and Farewell [13] proposed two approaches for analyzing longitudinal data with correlation as a function of explanatory variables. They pointed out that the relationship between outcome and explanatory variables may also depend on the dependence in outcomes and explanatory variables. According to Darlington and Farewell, the models are designed to focus on the marginal probability along with the dependence of correlation structure on explanatory variables. At this backdrop, we propose a new model based on both the marginal and conditional probabilities of the correlated binary events such that the joint function can be specified fully by unifying the marginal and conditional probabilities. In the proposed model, both the marginal and conditional probabilities are expressed as a function of explanatory variables and a test for dependence in outcomes is proposed.

2. Bivariate Bernoulli

Let us consider outcomes Y_{j-1} and Y_j at time points t_{j-1} and t_j respectively. If we consider, $j = 2$, then the bivariate probabilities are

	Y_2		
Y_1	*0*	1	Total
0	$P_{00} = P(Y_2 = 0, Y_1 = 0)$	$P_{01} = P(Y_2 = 1, Y_1 = 0)$	$P(Y_1 = 0)$
1	$P_{10} = P(Y_2 = 0, Y_1 = 1)$	$P_{11} = P(Y_2 = 1, Y_1 = 1)$	$P(Y_1 = 1)$
			1

The bivariate probability mass function for Y_1 and Y_2 can be shown as:

$$P(y_1, y_2) = P_{00}^{(1-y_1)(1-y_2)} P_{01}^{(1-y_1)y_2} P_{10}^{y_1(1-y_2)} P_{11}^{y_1 y_2} = \prod_{j=0}^{1} \prod_{k=0}^{1} P_{jk}^{y_{jk}} \tag{2.1}$$

where

$$\begin{aligned} y_{00} &= (1-y_1)(1-y_2), \quad j = 0, k = 0 \\ y_{01} &= (1-y_1)y_2, \quad j = 0, k = 1 \\ y_{10} &= y_1(1-y_2), \quad j = 1, k = 0 \\ y_{11} &= y_1 y_2, \quad j = 1, k = 1. \end{aligned}$$

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

$$\begin{aligned} P(Y_1 = 0, Y_2 = 1) &= P(Y_2 = 1 \,|Y_1 = 0\,) P(Y_1 = 0), \\ P(Y_1 = 0, Y_2 = 0) &= P(Y_2 = 0 \,|Y_1 = 0\,) P(Y_1 = 0), \\ P(Y_1 = 1, Y_2 = 1) &= P(Y_2 = 1 \,|Y_1 = 1\,) P(Y_1 = 1), \\ P(Y_1 = 1, Y_2 = 0) &= P(Y_2 = 0 \,|Y_1 = 1\,) P(Y_1 = 1). \end{aligned}$$

Using these relationships in the joint probability function, we obtain

$$P(y_1, y_2) = \prod_{j=0}^{1} \prod_{k=0}^{1} [P(Y_2 = k \,|Y_1 = j\,) P(Y_1 = j)]^{y_{jk}} \tag{2.2}$$

Now, the conditional probabilities can be shown as follows:

	Y_2		
Y_1	*0*	1	Total
0	π_{00}	π_{01}	1
1	π_{10}	π_{11}	1

The bivariate probability mass function can be obtained from conditional and marginal probability functions as displayed below:

$$P(y_1, y_2) = \prod_{j=0}^{1} \prod_{k=0}^{1} [\pi_{jk} P(Y_1 = j)]^{y_{jk}} \tag{2.3}$$

Let

$$\begin{aligned} \gamma_{01} &= [\gamma_{010}, \gamma_{011}, \ldots, \gamma_{01p}], \\ \gamma_{11} &= [\gamma_{110}, \gamma_{111}, \ldots, \gamma_{11p}], \\ \gamma_{1+} &= [\gamma_{1+0}, \gamma_{1+1}, \ldots, \gamma_{1+p}], \end{aligned}$$

$$\gamma_{+1} = [\gamma_{+10}, \gamma_{+11}, \ldots, \gamma_{+1p}],$$
$$\mathbf{X}'_{\mathrm{i}} = [1, X_{1i}, \ldots, X_{pi}].$$

The first order transition model can be expressed as function of covariates as shown below:

$$\pi_{01i}(\mathbf{X}_{\mathrm{i}}) = P(Y_{2i} = 1 \,|Y_{1i} = 0, \mathbf{X}_{\mathrm{i}}) = \frac{e^{\gamma_{01}\mathbf{X}_{\mathrm{i}}}}{1 + e^{\gamma_{01}\mathbf{X}_{\mathrm{i}}}} \tag{2.4}$$

and

$$\pi_{11i}(\mathbf{X}_{\mathrm{i}}) = P(Y_{2i} = 1 \,|Y_{1i} = 1, \mathbf{X}_{\mathrm{i}}) = \frac{e^{\gamma_{11}\mathbf{X}_{\mathrm{i}}}}{1 + e^{\gamma_{11}\mathbf{X}_{\mathrm{i}}}}. \tag{2.5}$$

These can be expressed as logit functions as follows:

$$\mathrm{logit}[\pi_{01i}(\mathbf{X}_{\mathrm{i}})] = \gamma_{01}\mathbf{X}_{\mathrm{i}}, \quad \text{and} \quad \mathrm{logit}[\pi_{11i}(\mathbf{X}_{\mathrm{i}})] = \gamma_{11}\mathbf{X}_{\mathrm{i}}.$$

These are new models for revealing the nature of dependence in the outcome variables $Y_1$and Y_2in the presence of covariates. It can be shown that under independence of $Y_1$and Y_2, the conditional models in the presence of covariates, (2.4) and (2.5), are equal (i.e. $\gamma_{01} = \gamma_{11}$). Here, it is noteworthy that γ_{01} and γ_{11} are the parameters of the conditional logit models for given covariates $\mathbf{X}$ and $Y_1 = 0$ and $Y_1 = 1$, respectively. The covariates are assumed to be time independent for this model.

The marginal probabilities for Y_1 and Y_2 are:

$$\pi_{1+i} = P(Y_{1i} = 1 \,|\mathbf{X}_{\mathrm{i}}) = \frac{e^{\gamma_{1+}\mathbf{X}_{\mathrm{i}}}}{1 + e^{\gamma_{1+}\mathbf{X}_{\mathrm{i}}}} = P_{1+i}(\mathbf{X}_{\mathrm{i}}), \tag{2.6}$$

$$\pi_{+1i}(\mathbf{X}_{\mathrm{i}}) = P(Y_{2i} = 1 \,|\mathbf{X}_{\mathrm{i}}) = \frac{e^{\gamma_{+1}\mathbf{X}_{\mathrm{i}}}}{1 + e^{\gamma_{+1}\mathbf{X}_{\mathrm{i}}}} = P_{+1i}(\mathbf{X}_{\mathrm{i}}). \tag{2.7}$$

It is evident that under independence of Y_1 and Y_2 , the conditional probabilities (2.4) and (2.5) can be shown as equal to the marginal probability of Y_2 in equation (2.7). This provides the basis for a new model for two dependent binary variables in the presence of covariates where independence is a special case for $\gamma_{01} = \gamma_{11}$.

It is noteworthy that Darlington and Farewell [13] proposed a transition probability model based on the following logit functions with marginal specification:

$$\pi_{11i}(\mathbf{X}_{\mathrm{i}}) = P(Y_{2i} = 1 \,|Y_{1i} = 1, \mathbf{X}_{\mathrm{i}}) = \frac{e^{\gamma_{11}\mathbf{X}_{\mathrm{i}}}}{1 + e^{\gamma_{11}\mathbf{X}_{\mathrm{i}}}}$$

and

$$\pi_{+1i}(\mathbf{X}_{\mathrm{i}}) = P(Y_{2i} = 1 \,|\mathbf{X}_{\mathrm{i}}) = \frac{e^{\gamma_{+1}\mathbf{X}_{\mathrm{i}}}}{1 + e^{\gamma_{+1}\mathbf{X}_{\mathrm{i}}}} = P_{+1i}(\mathbf{X}_{\mathrm{i}}).$$

They have not considered transition probability $\pi_{01i}(\mathbf{X}_{\mathrm{i}})$ in their model. Darlington and Farewell observed that there is asymmetry in this section and may not be suitable for all applications. Thus the method proposed by Darlington and Farewell can be shown as a special case of the new model where equality of conditional probability (2.5) and marginal probability (2.7) can be employed for testing for independence. In that case, $\gamma_{11} = \gamma_{+1}$, in other words, the conditional probability of Y_2 for the given Y_1 and $\mathbf{X}$ and the marginal probability of Y_2 for the given $\mathbf{X}$ are equal if Y_1 and Y_2 are independent. However, equality of models (2.4) and (2.5) reveals this more explicitly due to underlying conditional dependence on $\mathbf{X}$ for conditional models of Y_2 for the given values of Y_1 .

Then the likelihood function is

$$
\begin{aligned}
L &= \prod_{i=0}^{n}\prod_{j=0}^{1}\prod_{k=0}^{1}\left[\pi_{jki}(\mathbf{X}_i)P(Y_{1i}=j\,|\mathbf{X}_i)\right]^{y_{jkl}} \\
&= \prod_{i=1}^{n}\left[\left\{\frac{e^{\gamma_{01}\mathbf{X}_i}}{1+e^{\gamma_{01}\mathbf{X}_i}}\right\}^{y_{01i}}\left\{\frac{1}{1+e^{\gamma_{01}\mathbf{X}_i}}\right\}^{y_{00i}}\left\{\frac{e^{\gamma_{11}\mathbf{X}_i}}{1+e^{\gamma_{11}\mathbf{X}_i}}\right\}^{y_{11i}}\left\{\frac{1}{1+e^{\gamma_{11}\mathbf{X}_i}}\right\}^{y_{10i}}\right. \\
&\quad \left.\times\left\{\frac{e^{\gamma_{1+}\mathbf{X}_i}}{1+e^{\gamma_{1+}\mathbf{X}_i}}\right\}^{y_{1+i}}\left\{\frac{1}{1+e^{\gamma_{1+}\mathbf{X}_i}}\right\}^{y_{0+i}}\right]
\end{aligned}
\tag{2.8}
$$

where

$$
\sum_k Y_{jki}=Y_{j+i},\quad \sum_j\sum_k Y_{jki}=1,\quad \sum_i\sum_j\sum_k Y_{jki}=n,\quad \sum_i\sum_k Y_{jki}=n_j,\quad \sum_i Y_{jki}=n_{jk};
$$

$$
j=0,1;\quad k=0,1;\quad i=1,2,\ldots,n.
$$

Hence the log likelihood function can be obtained as follows:

$$
\begin{aligned}
\ln L = \sum_i &\left[\left\{y_{01i}\gamma_{01}\mathbf{X}_i - y_{00i}\ln(1+e^{\gamma_{01}\mathbf{X}_i})\right\}+\left\{y_{11i}\gamma_{11}\mathbf{X}_i - y_{10i}\ln(1+e^{\gamma_{11}\mathbf{X}_i})\right\}\right] \\
&+\sum_i\left[\left\{y_{1+i}\gamma_{1+}\mathbf{X}_i-(y_{1+i}+y_{0+i})\ln(1+e^{\gamma_{1+}\mathbf{X}_i})\right\}\right].
\end{aligned}
\tag{2.9}
$$

Differentiating (2.9) with respect to parameters, we obtain the following score functions:

$$
\frac{\partial \ln L}{\partial \gamma_{j1l}}=0,\quad j=0,1,\quad l=0,1,2,\ldots,p
$$

and

$$
\frac{\partial \ln L}{\partial \gamma_{1+l}}=0,\quad l=0,1,2,\ldots,p
$$

and we obtain the estimates $\hat{\gamma}_{j1l}$ and $\hat{\gamma}_{1+l}$, $l=0,1,\ldots,p$, by solving the above equations iteratively. The elements of variance-covariance matrix can be obtained from the observed information matrix as

$$
-\frac{\partial^2 \ln L}{\partial \gamma_{j1l}\partial \gamma_{j1l'}}\quad j=0,1;\quad l,l'=0,1,\ldots,p
$$

and

$$
-\frac{\partial^2 \ln L}{\partial \gamma_{1+l}\partial \gamma_{1+l'}},\quad l,l'=0,1,\ldots,p.
$$

3. Measure of dependence

For bivariate Bernoulli variates, $\operatorname{cov}(Y_1,Y_2)=\sigma_{12}=P_{11}P_{00}-P_{10}P_{01}$, hence, the correlation is

$$
\rho=\frac{P_{11}P_{00}-P_{10}P_{01}}{\sqrt{P_{0+}P_{1+}P_{+0}P_{+1}}}
\tag{3.1}
$$

as shown by Marshall and Olkin [27] and the empirical estimator is:

$$
\hat{\rho}=\frac{\hat{P}_{11}\hat{P}_{00}-\hat{P}_{10}\hat{P}_{01}}{\sqrt{\hat{P}_{0+}\hat{P}_{1+}\hat{P}_{+0}\hat{P}_{+1}}}
$$

where $P(Y_1 = j, Y_2 = k) = P_{jk},\ j = 0,1; k = 0,1$ and P_{j+}or P_{+k} are the marginal probabilities, $\hat{P}_{jk}, \hat{P}_{j+}$ and $\hat{P}_{+k}$ are the corresponding estimators. The correlation coefficient, $\rho = 0$ (denoted as ρ_{MO} in the tables) will indicate no association between Y_1 and Y_2. In other words, $P_{11}P_{00} - P_{01}P_{10} = 0$ can also be examined from the odds ratio, $\psi = (P_{11}P_{00}/P_{01}P_{10}) = 1$. If we define

$$E(Y_1) = \mu_1 = P_{1+}, \quad E(Y_2) = \mu_2 = P_{+1}, \quad E(Y_1Y_2) = \sigma_{12} + P_{1+}P_{+1},$$

then it is evident that $\sigma_{12} = 0$ indicates independence of the two binary outcomes as demonstrated by Teugels [37] and obtained a measure of correlation coefficient similar to (3.1).

Following Dale [12], the joint probability P_{11} for correlated binary variables can be expressed as [9]:

$$P_{11} = \begin{cases} 1/2(\psi-1)^{-1}\{1 + (P_{1+} + P_{+1})(\psi-1) - S(P_{1+}, P_{+1}, \psi)\}, & \text{if } \psi \neq 1 \\ P_{1+}P_{+1}, & \text{if } \psi = 1 \end{cases}$$

where

$$S(P_{1+}, P_{+1}, \psi) = \sqrt{[\{1 + (P_{1+} + P_{+1})(\psi-1)\}^2 + 4\psi(1-\psi)P_{1+}P_{+1}]}.$$

Darlington and Farewell [13] proposed the following measure for correlation to examine the dependence in outcome variables:

$$\rho_i = \text{corr}(Y_{1i}, Y_{2i}|\mathbf{X}_i) = \frac{e^{\gamma_{11}\mathbf{X}_i} - e^{\gamma_{+1}\mathbf{X}_i}}{1 + e^{\gamma_{11}\mathbf{X}_i}}.$$

If $\gamma_{11} = \gamma_{+1}$in the above relationship, then it confirms complete independence. The motivation behind this measure of correlation is very straightforward, i.e., if we need to study the relationship between binary outcomes and a set of explanatory variables then the dependence in outcomes are also dependent on the explanatory variables.

4. Test for the model

We need to test the following null hypothesis for the overall fit of the models comprising of the conditional and marginal models as functions of explanatory variables:

$$H_0 : \bar{\gamma}_{\mathbf{H}} = [\bar{\gamma}_{01}, \bar{\gamma}_{11}, \bar{\gamma}_{1+}] = 0$$
$$H_1 : \bar{\gamma}_{\mathbf{H}} \neq 0$$

where

$$\bar{\gamma}_{\mathbf{01}} = (\gamma_{011}, \gamma_{012}, ..., \gamma_{01p})$$
$$\bar{\gamma}_{\mathbf{11}} = (\gamma_{111}, \gamma_{112}, ..., \gamma_{11p})$$
$$\bar{\gamma}_{\mathbf{1+}} = (\gamma_{1+1}, \gamma_{1+2}, ..., \gamma_{1+p}).$$

Then

$$-2\left[\ln L(\hat{\gamma}_{010}, \hat{\gamma}_{110}, \hat{\gamma}_{1+0}) - \ln L\left(\hat{\gamma}_{01}, \hat{\gamma}_{11}, \hat{\gamma}_{1+}\right)\right] \tag{4.1}$$

which is asymptotically distributed as χ^2_{3p}. For testing

$$H_0 : \gamma_{j1l} = 0$$
$$H_1 : \gamma_{j1l} \neq 0$$

we can use the following Wald test statistic:

$$W = \frac{\hat{\gamma}_{j1l}}{s\hat{e}\left(\hat{\gamma}_{j1l}\right)}. \tag{4.2}$$

Similarly, for testing

$$H_0 : \gamma_{1+l} = 0$$
$$H_1 : \gamma_{1+l} \neq 0$$

We can use the Wald test statistic

$$W = \frac{\hat{\gamma}_{1+l}}{s\hat{e}\left(\hat{\gamma}_{1+l}\right)}. \tag{4.3}$$

5. Test for dependence

A simple test procedure can be developed for the bivariate Bernoulli model proposed in section 2. Using (2.4) and (2.5), we can obtain the odds ratio as follows:

$$\psi_i = \frac{\pi_{11i}(\mathbf{X}_i)/[1-\pi_{11i}(\mathbf{X}_i)]}{\pi_{01i}(\mathbf{X}_i)/[1-\pi_{01i}(\mathbf{X}_i)]} = \frac{e^{\gamma_{11}\mathbf{X}_i}}{e^{\gamma_{01}\mathbf{X}_i}} = e^{(\gamma_{11}-\gamma_{01})\mathbf{X}_i} \tag{5.1}$$

and

$$\ln \psi_i = (\gamma_{11}-\gamma_{01})\mathbf{X}_i.$$

If $\gamma_{01} = \gamma_{11}$ then $\psi = 1$ and $\ln \psi = 0$ indicate independence of the two binary outcomes in the presence of covariates. Any departure from $\psi = 1$ will indicate the extent of dependence. We can measure the dependence between Y_1 and Y_2, in terms of odds ratio, as $e^{(\gamma_{11}-\gamma_{01})\mathbf{1}}$ where $\mathbf{1}$ is the column vector of 1's for the unit difference in the values of covariates. The null hypothesis $H_0 : \gamma_{01} = \gamma_{11}$ can be tested for independence in the presence of covariates between the binary outcome variables Y_1 and Y_2 using the following test statistic:

$$\chi^2 = (\hat{\gamma}_{01}-\hat{\gamma}_{11})' \left[\widehat{Var}(\hat{\gamma}_{01}-\hat{\gamma}_{11})\right]^{-1} (\hat{\gamma}_{01}-\hat{\gamma}_{11}) \tag{5.2}$$

which is distributed asymptotically as chi-square with p degrees of freedom. Here the estimators, $\hat{\gamma}$'s, are the maximum likelihood estimators based on the equations shown in section 2 by differentiating the log likelihood function (2.9) with respect to the parameters of interest. We have employed this test statistic for testing dependence between Y_1 and Y_2.

Another alternative test can be obtained from the relationship between the conditional and marginal probabilities for the outcome variable, Y_2, as displayed in equations (2.4), (2.5) and (2.7). It may be noted here that under independence of Y_1 and Y_2, in the presence of covariates, the conditional probabilities (2.4) and (2.5) are equal and can be expressed in terms of the marginal probability (2.7). In other words, the null hypothesis is: $H_0 : \gamma_{01} = \gamma_{11} = \gamma_{+1}$. This can be tested employing the following asymptotic chi-squares for hypotheses: $H_{01} : \gamma_{01} = \gamma_{+1}$ and $H_{02} : \gamma_{11} = \gamma_{+1}$, respectively:

$$\chi^2 = \left(\hat{\gamma}_{01}-\hat{\gamma}_{+1}\right)' \left[\widehat{Var}(\hat{\gamma}_{01}-\hat{\gamma}_{+1})\right]^{-1} \left(\hat{\gamma}_{01}-\hat{\gamma}_{+1}\right) \tag{5.3}$$

$$\chi^2 = \left(\hat{\gamma}_{11}-\hat{\gamma}_{+1}\right)' \left[\widehat{Var}(\hat{\gamma}_{11}-\hat{\gamma}_{+1})\right]^{-1} \left(\hat{\gamma}_{11}-\hat{\gamma}_{+1}\right). \tag{5.4}$$

It is noteworthy that the measure of correlation proposed by Darlington and Farewell [13], ie.,

$$\rho_i = \text{corr}(Y_{1i}, Y_{2i}|\mathbf{X}_i) = \frac{e^{\gamma_{11}\mathbf{X}_i} - e^{\gamma_{+1}\mathbf{X}_i}}{1 + e^{\gamma_{11}\mathbf{X}_i}}$$

can be tested by (5.4). However, it is necessary for the independence that both (5.3) and (5.4) should support the null hypotheses $H_{01} : \boldsymbol{\gamma_{01}} = \boldsymbol{\gamma_{+1}}$ and $H_{02} : \boldsymbol{\gamma_{11}} = \boldsymbol{\gamma_{+1}}$, respectively. Both are asymptotically chi-squares with p degrees of freedom. If one or both of (5.3) and (5.4) show significant results then it is likely that there is dependence between Y_1 and Y_2. The extent of dependence can be estimated as $e^{(\hat{\gamma}_{11} - \hat{\gamma}_{01})\mathbf{1}}$.

6. Simulation

To generate correlated binary data for simulations, we have used technique proposed by Leisch *et al.* [20] known as bindata package for R. Based on their method, data are first generated from multivariate normal random variates and they are transformed into binary data. We simulated three variables, two are dependent outcomes Y_1 and Y_2 and one is covariate, X. We have considered all the three variables binary for clear exposition of the proposed tests. We have considered different combination of the correlation between the two outcome variables and their relationship with the covariate, X. Each simulation was performed 500 times with samples of size 250 and 500.

Table 1 displays results averaged from 500 simulations with samples of size 250 for different correlation structures between the three variables denoted as Y_1, Y_2 and X. The various dependence patterns between these variables are employed here to obtain 12 different models for samples of size 250. Models 1, 2 and 3 show that there are no evidences of association between Y_1 and Y_2 and the odds ratios are close to 1. However, the conditional odds ratios for given X seem to deviate from 1 indicating substantial association between dependent and independent variables. Models 4-12 display different types of associations between Y_1 and Y_2 as well as between dependent variables Y_1 and Y_2 and X. The estimated correlation coefficients for Y_1 and Y_2 based on Marshall-Olkin, and odds ratio, ψ show no association for models 1-3 for observed data and models using logit link function. In addition, the estimated correlations employing Marshall-Olkin indicate values close to zero for models 8 and 9. The Marshall-Olkin correlation coefficients show that the association between Y_1 and Y_2 are also close to zero which other measures failed to recognize. The test for models indicate that all the models are significant due to association between the independent and dependent variables as shown in the conditional odds ratios.

Now if we examine the pattern of dependence based on the proposed test then models 1,2,3 and 8 (less than 5%) clearly fail to show any dependence. Model 9 shows independence in 90% of the simulations. This is supported by their corresponding measure of dependence by Marshall-Olkin correlation coefficient. The alternative tests based on hypotheses: $H_{01} : \gamma_{01} = \gamma_{+1}$ and $H_{02} : \gamma_{11} = \gamma_{+1}$, respectively, also reveal that the models 1,2,3,8,9 and 12 clearly fail to show any dependence.

We observe almost similar findings for samples of size 500 as displayed in Table 2 with some minor differences although the number of increased significant cases might be attributed to the increased sample size from 250 to 500.

Table 1. Sample Size of 250 with 500 Simulations for Obtaining the Estimates of Measures of Associations Based on Observed Data and Logistic Regression Models

Simulation No	1	2	3	4	5	6	7	8	9	10	11	12
00	100	99	100	113	112	82	87	38	38	50	113	13
01	100	100	100	87	87	118	37	62	63	74	87	37
10	25	25	25	13	13	43	38	63	62	75	12	37
11	25	25	25	38	38	7	88	87	88	50	38	163
$\psi_{(Y1-Y2)}$	1.041	1.056	1.085	4.244	4.238	0.123	5.668	0.877	0.873	0.461	4.315	1.571
$\hat{\rho}_{MO}$	-0.002	0.000	0.006	0.251	0.250	-0.356	0.397	-0.039	-0.041	-0.198	0.253	0.067
Model χ^2	20.33	36.48	59.06	4.13	43.93	36.77	65.11	245.65	3.86	106.49	105.91	4.75
# Sig. p-values	416	496	500	17	500	499	500	500	8	500	500	13
Test for Dependencies												
$\gamma_{01} vs. \gamma_{11}$	2.04	1.91	2.03	14.59	17.86	19.36	20.21	1.36	2.49	14.46	4.82	3.16
# Sig. p-values	24	19	19	483	500	494	490	2	51	476	161	79
$\gamma_{01} vs. \gamma_{+1}$	0.28	0.26	0.31	1.94	2.95	2.94	6.63	0.61	1.10	5.53	1.16	1.98
# Sig. p-values	0	0	0	0	6	7	284	0	2	203	4	21
$\gamma_{11} vs. \gamma_{+1}$	1.31	1.22	1.28	10.03	12.40	13.99	7.77	0.40	0.63	5.47	2.98	0.43
# Sig. p-values	5	3	6	439	486	492	350	0	0	196	44	0

Table 2. Sample Size of 500 with 500 Simulations for Obtaining the Estimates of Measures of Associations Based on Observed Data and Logistic Regression Models

Simulation No	1	2	3	4	5	6	7	8	9	10	11	12
00	200	200	201	201	225	165	176	75	75	100	226	25
01	201	200	199	199	175	235	75	125	125	150	175	75
10	50	50	50	50	25	85	74	125	125	151	25	75
11	49	50	50	50	75	15	176	175	175	100	74	325
$\Psi_{(Y1-Y2)}$	1.004	1.014	1.044	1.044	4.105	0.125	5.722	0.867	0.861	0.446	4.064	1.500
$\hat{\rho}_{MO}$	-0.005	-0.002	0.003	0.003	0.252	-0.350	0.405	-0.038	-0.040	-0.202	0.250	0.064
Model χ^2	38.069	72.832	122.565	122.565	86.116	72.932	132.621	495.042	4.877	209.969	210.372	6.732
# Sig. p-values	496	500	500	500	500	500	500	500	31	500	500	45
Test for Dependencies												
$\gamma_{01} vs. \gamma_{11}$	2.216	2.212	2.410	2.410	36.751	38.952	40.139	1.791	2.982	29.871	7.883	3.895
# Sig. p-values	30	32	41	41	500	500	500	9	66	500	305	101
$\gamma_{01} vs. \gamma_{+1}$	0.284	0.297	0.342	0.342	5.828	5.374	12.155	0.824	1.319	11.178	1.457	2.437
# Sig. p-values	0	0	0	0	218	166	489	0	4	475	8	37
$\gamma_{11} vs. \gamma_{+1}$	1.441	1.408	1.533	19.746	25.355	28.149	16.179	0.501	0.737	11.150	5.403	0.499
# Sig. p-values	8	7	9	499	500	500	500	0	0	472	191	0

7. Application

For this study, an application is displayed in this section from the Health and Retirement Study (HRS) data. The HRS is sponsored by the National Institute of Aging (grant number NIA U01AG09740) and conducted by the University of Michigan. This study is conducted nationwide for individuals over age 50 and their spouses. We have used the panel data from the two rounds of the study conducted on individuals over age 50 years in 1992 (Wave I) and 1994 (Wave II) and documented by RAND. We have used the panel data on depression for the period, 1992–1994. The depression index is based on the score on the basis of the scale proposed by the Center for Epidemiologic Studies Depression (CESD). As indicated in the documentation of the RAND, the CESD score is computed on the basis of eight indicators attributing depression problem. The indicators of depression problem are based on six negative (all or most of the time: depressed, everything is an effort, sleep is restless, felt alone, felt sad, and could not get going) and two positive indicators (felt happy, enjoyed life). These indicators are yes/no responses of the respondent's feelings much of the time over the week prior to the interview. The CESD score is the sum of six negative indicators minus two positive indicators. Hence, severity of the emotional health can be measured from the CESD score. From the panels of data, we have used 9761 respondents for analyzing depression among the elderly in the USA during 1992–2002.

We considered the following dependent and explanatory variables: depression status (no depression, if CESD score $= 0$ then depression status $= 0$, depression, if CESD score > 0 then depression status = 1), gender (male $= 1$, female $= 0$), marital status (married/partnered $= 1$, single/widowed/divorced $= 0$), ethnic group (white $= 1$, else 0; black $= 1$, else 0; others $=$ reference category). Table 3 displays the transition counts and transition probabilities during 1992–94 period in Waves I and II. It is evident from the transition probabilities that the probability of outcome status remains depression free during the period is 0.650 and outcome status is changed from depression free to depression is 0.350. However, the probability of remaining in the state of depression during the period is 0.715. The estimated odds ratio for depression in Waves I and II is 4.67 and the conditional odds ratios for given $X = 0$ and $X = 1$ are 4.48 and 4.85 respectively. The Marshall-Olkin correlation coefficient between depression status in Waves I (Y_1) and II (Y_2) is 0.354. This indicates a positive correlation between the depression status in consecutive time points. The conditional and marginal models are significant indicating association between the outcome variables and selected covariates (using the test statistic demonstrated in equation (4.1)) and we observe negative association between depression among elderly and gender, marital status and white race as compared to other races. The proposed test statistic for testing equality of parameters of the conditional models, using the test statistic (5.2), indicates dependence in the depression status in Waves I and II (p-value<0.01). The alternative tests based on equality of conditional and marginal model parameters, based on the tests (5.3) and (5.4), support this finding of dependence (p-value<0.01).

Table 3. Transition Count and Probability Based on Consecutive Follow-ups I and II

WAVE I	WAVE II					
	Transition Count			Transition Probability		
	0	1	Total	0	1	Total
0	3296	1773	5069	0.650	0.350	1.000
1	868	2179	3047	0.285	0.715	1.000

Table 4. Application Using WAVE I and WAVE II from HRS Data (Dependent Variables=CESD, 0,1+)

	Conditional Models						Marginal Models					
Covariates	$\hat{\gamma}_{01j}$			$\hat{\gamma}_{11j}$			$\hat{\gamma}_{1+j}$			$\hat{\gamma}_{+1j}$		
	$\hat{\beta}_{01j}$	s.e	p-value	$\hat{\beta}_{11j}$	s.e	p-value	$\hat{\beta}_{1j}$	s.e	p-value	$\hat{\beta}_{2j}$	s.e	p-value
Constant	0.249	0.183	0.158	1.861	0.225	0.000	0.525	0.128	0.000	1.104	0.134	0.000
Gender	-0.279	0.061	0.000	-0.179	0.083	0.038	-0.080	0.048	0.098	-0.243	0.046	0.000
Marital Status	-0.299	0.075	0.000	-0.525	0.093	0.000	-0.606	0.054	0.000	-0.547	0.054	0.000
White	-0.591	0.177	0.002	-0.587	0.217	0.010	-0.671	0.124	0.000	-0.752	0.129	0.000
Black	-0.125	0.191	0.322	-0.300	0.232	0.173	-0.144	0.134	0.222	-0.225	0.139	0.108
Model χ^2 (p-value)	451.36 (0.000)											

Note: $\psi_{(Y1-Y2)} = 4.67$; $\hat{\rho}_{MO} = 0.354$; χ^2 for testing γ_{01} vs. $\gamma_{11} = 838.504$(p-value <0.001);
χ^2 for testing γ_{01} vs. $\gamma_{+1} = 210.668$(p-value < 0.001); χ^2 for testing γ_{11} vs. $\gamma_{+1} = 391.919$($p-$value < 0.001).

8. Conclusion

The problem of dependence in the repeated measures outcomes is one of the formidable challenges to the researchers. In the past, the problem had been resolved on the basis of marginal models with a varied range of assumptions. The models based on GEE with various correlation structures are some examples of arbitrariness contained in the procedures. Some attempts have been made to address this problem employing conditional models too. However, we need to specify the bivariate or multivariate outcomes specifying the underlying correlations for a more detailed and more meaningful models. This paper shows a new model for bivariate binary data using the conditional and marginal probabilities to specify the joint bivariate probability functions and applies the proposed estimation procedures to real life data and simulations. Some test procedures are suggested for testing the dependence of the bivariate outcomes in the presence of covariates. The numerical examples clearly show the utility of the proposed procedures for testing dependence in the binary outcome variables.

Acknowledgement. The authors acknowledge gratefully to the HRS (Health and Retirement Study) which is sponsored by the National Institute of Aging (grant number NIA U01AG09740) and conducted by the University of Michigan. This project was supported by King Saud University, Deanship of Scientific Research, College of Science Research Center. The authors are grateful to the anonymous reviewers for their useful comments that contributed to improvement in the exposition of the paper to a great extent.

References

[1] G. R. Antelman, Interrelated Bernoulli processes, *J. Amer. Statist. Assoc.* **67** (1972), no. 340, 831–841.
[2] A. Azzalini, Logistic regression for autocorrelated data with application to repeated measures, *Biometrika* **81** (1994), 767–775.
[3] R. R. Bahadur, A representation of the joint distribution of responses to *n* dichotomous items, in *Studies in Item Analysis and Prediction*, 158–168, Stanford Univ. Press, Stanford, CA.
[4] W. P. Bergsma and T. Rudas, Modeling conditional and marginal association in contingency tables, *Ann. Fac. Sci. Toulouse Math. (6)* **11** (2002), no. 4, 455–468.
[5] W. P. Bergsma and T. Rudas, Marginal models for categorical data, *Ann. Statist.* **30** (2002), no. 1, 140–159.
[6] G. E. Bonney, Regressive logistic models for familial disease and other binary trials, *Biometrics* **42** (1986), 611–625.
[7] G. E. Bonney, Logistic regression for dependent binary observations, *Biometrics* **43** (1987), 951–973.
[8] V. Carey, S. L. Zeger and P. J. Diggle, Modelling multivariate binary data with alternating logistic regressions, *Biometrika* **80** (1993), 517–526.
[9] S. Cessie and J. C. Houwelingen, Logistic regression for correlated binary data, *J. Roy. Statist. Soc. Ser. C* **43** (1994), 95–108.
[10] D. R. Cox, The analysis of multivariate binary data, *J. Roy. Statist. Soc. Ser. C* **21** (1972), 113–120.
[11] D. R. Cox and N. Wermuth, A note on the quadratic exponential binary distribution, *Biometrika* **81** (1994), no. 2, 403–408.
[12] J. R. Dale, Global cross-ratio models for bivariate, discrete, ordered responses, *Biometrics* **42** (1986), 909–917.
[13] G. A. Darlington and V. T. Farewell, Binary longitudinal data analysis with correlation a function of explanatory variables, *Biometrical J.* **34** (1992), 899–910.
[14] G. M. Fitzmaurice and S. R. Lipsitz, A model for binary time series data with serial odds ratio patterns, *Appl. Statist.* **44** (1995), 51–61.
[15] G. F. V. Glonek and P. McCullagh, Multivariate logistic models, *J. Roy. Statist. Soc. Ser. B* **57** (1995), 533–546.

[16] G. F. V. Glonek, A class of regression models for multivariate categorical responses, *Biometrika* **83** (1996), no. 1, 15–28.
[17] M. A. Islam and R. I. Chowdhury, A higher order Markov model for analyzing covariate dependence, *Appl. Math. Model.* **30** (2006), 477–488.
[18] M. A. Islam, R. I. Chowdhury and S. Huda, *Markov Models with Covariate Dependence for Repeated Measures*, Nova Sci. Publ., New York, 2009.
[19] A. Juan and E. Vidal, On the use of Bernoulli mixture models for text classification, *Pattern Recogn.* **35** (2002), 2705–2710.
[20] F. Leisch, A. Weingessel and K. Hornik, *On the generation off correlated artificial binary data,* Working Paper Series. Working Paper No. 13, Vienna University of Economics and Business Administration, August 2–6, 1090 Wien, Austria, 1998.
[21] K.-Y. Liang and S. L. Zeger, Longitudinal data analysis using generalized linear models, *Biometrika* **73** (1986), no. 1, 13–22.
[22] K.-Y. Liang, S. L. Zeger and B. Qaqish, Multivariate regression analyses for categorical data, *J. Roy. Statist. Soc. Ser. B* **54** (1992), no. 1, 3–40.
[23] P.-S. Lin and M. K. Clayton, Analysis of binary spatial data by quasi-likelihood estimating equations, *Ann. Statist.* **33** (2005), no. 2, 542–555.
[24] S. R. Lipsitz, N. M. Laird and D. P. Harrington, Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association, *Biometrika* **78** (1991), no. 1, 153–160.
[25] S. H. Lee and C. H. Jun, Discriminant analysis of binary data following multivariate Bernoulli distribution, *Expert Syst. Appl.* **38** (2011), 7795–7802.
[26] G. Lovison, A matrix-valued Bernoulli distribution, *J. Multivariate Anal.* **97** (2006), no. 7, 1573–1585.
[27] A. W. Marshall and I. Olkin, A family of bivariate distributions generated by the bivariate Bernoulli distribution, *J. Amer. Statist. Assoc.* **80** (1985), no. 390, 332–338.
[28] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Second Edition, Monographs on Statistics and Applied Probability, Chapman & Hall, London, 1989.
[29] B. W. McDonald, Estimating logistic regression parameters for bivariate binary data, *J. Roy. Statist. Soc. Ser. B* **55** (1993), no. 2, 391–397.
[30] G. Molenberghs and E. Lesaffre, Marginal modelling of multivariate categorical data, *Statist. Medicine* **18** (1999), 2237–2255.
[31] G. Molenberghs and E. Lesaffre, Marginal modelling of correlated ordinal data using a multivariate Plackett distribution, *J. Amer. Statist. Assoc.* **89** (1996), 633–644.
[32] L. R. Muenz and L. V. Rubinstein, Markov models for covariate dependence of binary sequences, *Biometrics* **41** (1985), no. 1, 91–101.
[33] R. L. Prentice, Correlated binary regression with covariates specific to each binary observation, *Biometrics* **44** (1988), no. 4, 1033–1048.
[34] T. Rudas and W. P. Bergsma, On applications of marginal models for categorical data, *Metron* **62** (2004), no. 1, 15–37.
[35] E. H. Simpson, The interpretation of interaction in contingency tables, *J. Roy. Statist. Soc. Ser. B.* **13** (1951), 238–241.
[36] Z. Sun, O. Rosen and A. R. Sampson, Multivariate Bernoulli mixture models with application to postmortem tissue studies in schizophrenia, *Biometrics* **63** (2007), no. 3, 901–909.
[37] J. L. Teugels, Some representations of the multivariate Bernoulli and binomial distributions, *J. Multivariate Anal.* **32** (1990), no. 2, 256–268.
[38] J. Wakefield, Ecological inference for 2×2 tables, *J. Roy. Statist. Soc. Ser. A* **167** (2004), no. 3, 385–445.
[39] T. W. Yee and T. Dirnbock, Models for analyzing species' presence/absence data at two time points, *J. Theor. Biol.* **259** (2009), 684–694.
[40] S. L. Zeger and K. Y. Liang, Longitudinal data analysis for discrete and continuous outcomes, *Biometrics* **42** (1986), 121–130.
[41] L. P. Zhao and R. L. Prentice, Correlated binary regression using a quadratic exponential model, *Biometrika* **77** (1990), no. 3, 642–648.