Generalized Bernoulli Model for Correlated Binary Responses with an Application to Child Nutrition Data in Bangladesh

Mohammad Junayed Bhuyan*
Institute of Statistical Research and Training, University of Dhaka, Dhaka, Bangladesh, jbhuyan@isrt.ac.bd

M. Shafiqur Rahman
Institute of Statistical Research and Training, University of Dhaka, Dhaka, Bangladesh, shafiq@isrt.ac.bd

M. Ataharul Islam
Department of Applied Statistics, East West University, Dhaka, Bangladesh, mataharul@yahoo.com

Abstract

In longitudinal studies, response from the same subjects are usually correlated and hence standard statistical models based on independence assumption may draw misleading inference. To analyze such data one needs special statistical models allowing for such dependence. Most of the studies has made attempts to address this problem using either marginal models or conditional models. However, using the marginal models or conditional models alone, it is difficult to specify the measures of dependence in outcomes precisely. Islam et al (2013) proposed a joint modeling approach for bivariate binary response using both the conditional and marginal models where the dependence in outcome variables can be measured and tested using a link function of the models. The authors investigated, via simulation studies, only the dependence of the outcome variables. However no investigation on the performance of the regression coefficient and the models as a whole was performed. In this paper, the properties of the estimates of regression coefficients, such as bias and coverage probability, of the joint model are investigated by using an extensive simulation study. The results showed that the models performed well in all simulation scenarios. An application of the model is provided using child nutrition data from Bangladesh demographic and health survey.

Keywords: marginal model; conditional model; joint model; link function.

1. Introduction

Longitudinal data sets are comprised of repeated observations of an outcome variable and a set of covariates for each of many subjects. These repeated outcomes are usually correlated. In analyzing longitudinal response one must take into account the correlation between repeated observation from the same subject. Ignoring this correlation may lead misleading inference. There is overwhelming use of correlated binary data since work on repeated measures data by Zeger and Liang (1986) that was proposed on the generalized estimating equation (GEE). We observe that if the marginal variates are assumed independent, then the analysis can be based on a standard generalized linear model. However, we need to deal with repeated binary outcomes which are correlated with a longitudinal analysis. The GEE models are proposed based on probability of the event and correlations or the first and the second moments by Zeger and Liang (1986). Azzalini (1994) proposed a marginal model based on the binary Markov Chain for a single stationary process. It is worth noting that Le Cessie and van Houwelingen (1994) have proposed measures of dependence for correlated binary data using logistic regression. It has been observed that the marginal measures may fail to provide the measure of dependence of binary outcomes due to lack of proper specification of the underlying model. As compared to the marginal models, a relatively small number of studies have been conducted on the conditional approaches. However, without a joint model of the correlated outcome variables, marginal or conditional models alone do not resolve the problems associated with dependence in outcomes. At this backdrop, a joint model has been proposed by Islam *et al.* (2013) which takes account of both marginal and conditional probabilities of correlated binary events such that the joint function can be specified fully by

unifying marginal and conditional probabilities. A test based on one of the link functions of the joint models for association for binary bivariate data has also proposed and investigated via simulation studies. However, the authors did not assess the properties of the estimate of regression coefficients of the models and the models as a whole, except for testing the dependence between the response. In this paper we evaluate the properties of the estimate of regression coefficient such as bias, coverage probability by using an extensive simulation study and application is provided using child nutrition data from Bangladesh Demographic and Health Survey(BDHS) 2011.

2. Methodology

Islam *et al.* (2013) propose the following model based on the marginal-conditional approach to obtain joint models.

The bivariate Bernoulli distribution for outcomes Y_1 and Y_2 can be expressed as

$$P(Y_1 = y_1, Y_2 = y_2) = p_{00}^{(1-y_1)(1-y_2)} p_{01}^{(1-y_1)y_2} p_{10}^{y_1(1-y_2)} p_{00}^{y_1 y_2}. \tag{1}$$

The joint probability can be shown in a 2×2 table as follows:

	y_2		
y_1	0	1	Total
0	p_{00}	p_{01}	p_{0+}
1	p_{10}	p_{11}	p_{1+}
	p_{+0}	p_{+1}	1

The joint probability can be obtained from the conditional and marginal probability as

$$P(Y_1 = y_1, Y_2 = y_2) = P(Y_2 = y_2 | Y_1 = y_1) P(Y_1 = y_1). \tag{2}$$

The bivariate probabilities as a function of covariates X are as follows:

$$P(Y_1 = y_1, Y_2 = y_2 | x) = P(Y_2 = y_2 | Y_1 = y_1; x) P(Y_1 = y_1 | x). \tag{3}$$

The joint probability mass function in Equation (1) can be demonstrated in terms of the exponential family for the generalized linear models as

$$P(Y_1 = y_1, Y_2 = y_2) = \exp\left\{ y_1 \log\left(\frac{p_{10}}{p_{00}}\right) + y_2 \log\left(\frac{p_{01}}{p_{00}}\right) + y_1 y_2 \log\left(\frac{p_{00} p_{11}}{p_{01} p_{10}}\right) + \log p_{00} \right\},$$

where $(y_1, y_2) = (0,0), (0,1), (1,0), (1,1)$ and $\sum_{i,j} p_{ij} = 1$.

Let us consider a sample of size n then the log likelihood function in this case is given by

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left\{ y_{1i} \log\left(\frac{p_{10i}}{p_{00i}}\right) + y_{2i} \log\left(\frac{p_{01i}}{p_{00i}}\right) + y_{1i} y_{2i} \log\left(\frac{p_{00i} p_{11i}}{p_{01i} p_{10}i}\right) + \log p_{00i} \right\}.$$

Then the components of the link function can be denoted as follows:

$$\eta_0 = \log p_{00}, \quad \eta_1 = \log\left(\frac{p_{01}}{p_{00}}\right), \quad \eta_2 = \log\left(\frac{p_{10}}{p_{00}}\right) \text{ and } \eta_3 = \log\left(\frac{p_{00} p_{11}}{p_{01} p_{10}}\right),$$

where η_0 is the baseline link function, η_2 is the link function for Y_1, η_1 is the link function for Y_2 and η_3 is the link function for dependence between Y_1 and Y_2. The link functions for Y_1 and Y_2 are expressed this way for convenience in the expression of the conditional models shown later.

We have demonstrated the probabilities without function of covariates in the previous expressions. Now let us consider $\mathbf{X} = (1, X_1, X_2, \ldots, X_p)$ and $\mathbf{x} = (1, x_1, x_2, \ldots, x_p)$ where $\mathbf{X}$ and $\mathbf{x}$ are the vector of covariates and their corresponding covariates' values, respectively. Then we can express the conditional probabilities in terms of the logit link functions as follows:

$$P(Y_2 = 1 | Y_1 = 0, \mathbf{x}) = \frac{e^{\mathbf{x}\boldsymbol{\beta_{01}}}}{1 + e^{\mathbf{x}\boldsymbol{\beta_{01}}}} = \pi_{01}(\mathbf{x}) \quad \text{and} \quad P(Y_2 = 0 | Y_1 = 0, \mathbf{x}) = \frac{1}{1 + e^{\mathbf{x}\boldsymbol{\beta_{01}}}} = \pi_{00}(\mathbf{x}),$$

$$P(Y_2=1|Y_1=1,\mathbf{x})=\frac{e^{\mathbf{x}\boldsymbol{\beta_{11}}}}{1+e^{\mathbf{x}\boldsymbol{\beta_{11}}}}=\pi_{11}(\mathbf{x}) \quad \text{and} \quad P(Y_2=0|Y_1=1,\mathbf{x})=\frac{1}{1+e^{\mathbf{x}\boldsymbol{\beta_{11}}}}=\pi_{10}(\mathbf{x}),$$

where

$$\boldsymbol{\beta_{01}}=(\beta_{010},\beta_{011},\beta_{012},\ldots,\beta_{01p})' \qquad \text{and} \qquad \boldsymbol{\beta_{11}}=(\beta_{110},\beta_{111},\beta_{112},\ldots,\beta_{11p})'.$$

The marginal probabilities are as follows:

$$P(Y_1=1|\mathbf{X}=\mathbf{x})=\frac{e^{\mathbf{x}\boldsymbol{\beta_1}}}{1+e^{\mathbf{x}\boldsymbol{\beta_1}}}=\pi_1(\mathbf{x}) \text{ and } P(Y_1=0|\mathbf{X}=\mathbf{x})=\frac{1}{1+e^{\mathbf{x}\boldsymbol{\beta_1}}}=1-\pi_1(\mathbf{x}),$$

where

$$\boldsymbol{\beta_1}=(\beta_{10},\beta_{11},\beta_{12},\ldots,\beta_{1p})'.$$

Also, we can write

$$p_{01}(\mathbf{x})=P(Y_2=1|Y_1=0,\mathbf{X}=\mathbf{x}).P(Y_1=0|\mathbf{X}=\mathbf{x})=\frac{e^{\mathbf{x}\boldsymbol{\beta_{01}}}}{1+e^{\mathbf{x}\boldsymbol{\beta_{01}}}}.\frac{1}{1+e^{\mathbf{x}\boldsymbol{\beta_1}}}, \tag{4}$$

$$p_{00}(\mathbf{x})=P(Y_2=0|Y_1=0,\mathbf{X}=\mathbf{x}).P(Y_1=0|\mathbf{X}=\mathbf{x})=\frac{1}{1+e^{\mathbf{x}\boldsymbol{\beta_{01}}}}.\frac{1}{1+e^{\mathbf{x}\boldsymbol{\beta_1}}}, \tag{5}$$

$$p_{11}(\mathbf{x})=P(Y_2=1|Y_1=1,\mathbf{X}=\mathbf{x}).P(Y_1=1|\mathbf{X}=\mathbf{x})=\frac{e^{\mathbf{x}\boldsymbol{\beta_{11}}}}{1+e^{\mathbf{x}\boldsymbol{\beta_{11}}}}.\frac{e^{\mathbf{x}\boldsymbol{\beta_1}}}{1+e^{\mathbf{x}\boldsymbol{\beta_1}}}, \tag{6}$$

$$p_{10}(\mathbf{x})=P(Y_2=0|Y_1=1,\mathbf{X}=\mathbf{x}).P(Y_1=1|\mathbf{X}=\mathbf{x})=\frac{1}{1+e^{\mathbf{x}\boldsymbol{\beta_{11}}}}.\frac{e^{\mathbf{x}\boldsymbol{\beta_1}}}{1+e^{\mathbf{x}\boldsymbol{\beta_1}}}. \tag{7}$$

The log-likelihood function can be rewrite as

$$l_i=y_{1i}\eta_{2i}+y_{2i}\eta_{1i}+y_{1i}y_{2i}\eta_{3i}+\eta_{0i}, \tag{8}$$

where

$$\eta_{0i}=-\ln(1+e^{x_i\beta_{01i}})-\ln(1+e^{x_i\beta_{1i}})\,; \qquad \eta_{1i}=e^{x_i\beta_{01i}};$$
$$\eta_{2i}=e^{x_i\beta_{1i}}+\ln(1+e^{x_i\beta_{01i}})-\ln(1+e^{x_i\beta_{11i}})\,; \qquad \eta_{3i}=x_i(\beta_{11i}-\beta_{01i}).$$

which indicates that if there is no association between Y_1 and Y_2 then $\eta_3=0$ and this is true for $\boldsymbol{\beta_{01}}=\boldsymbol{\beta_{11}}$. The use of the proposed conditional and marginal models provide the necessary background to obtain the test for dependence in the repeated outcome variables based on the above link functions.
The estimating equations for $j=0,1,\ldots,p$ are as follows

$$\frac{\partial l}{\partial\beta_{01j}}=\sum_{i=1}^{n}\sum_{s=0}^{3}\frac{\partial l_i}{\partial\eta_s}\frac{\partial\eta_s}{\partial\beta_{01j}}, \quad \frac{\partial l}{\partial\beta_{11j}}=\sum_{i=1}^{n}\sum_{s=0}^{3}\frac{\partial l_i}{\partial\eta_s}\frac{\partial\eta_s}{\partial\beta_{11j}} \text{ and } \frac{\partial l}{\partial\beta_{1j}}=\sum_{i=1}^{n}\sum_{s=0}^{3}\frac{\partial l_i}{\partial\eta_s}\frac{\partial\eta_s}{\partial\beta_{1j}}.$$

Hence the score equations are:

$$\left[\frac{\partial l}{\partial\beta_j}\right]=\begin{bmatrix}\frac{\partial l}{\partial\beta_{01j}}\\ \frac{\partial l}{\partial\beta_{11j}}\\ \frac{\partial l}{\partial\beta_{1j}}\end{bmatrix}=\begin{bmatrix}-\sum_{i=1}^{n}x_{ij}(1-y_{1i})(\pi_{01}(x_i)-y_{2i})\\ -\sum_{i=1}^{n}x_{ij}y_{1i}(\pi_{11}(x_i)-y_{2i})\\ -\sum_{i=1}^{n}x_{ij}(\pi_1(x_i)-y_{1i})\end{bmatrix} \qquad \text{where } j=0,1,\ldots,p.$$

And the second derivatives are shown below as in the following:

$$\left[\frac{\partial^2 l}{\partial\beta_j\partial\beta_{j'}}\right]=\begin{bmatrix}\frac{\partial^2 l}{\partial\beta_{01j}\partial\beta_{01j'}}\\ \frac{\partial^2 l}{\partial\beta_{11j}\partial\beta_{11j'}}\\ \frac{\partial^2 l}{\partial\beta_{1j}\partial\beta_{1j'}}\end{bmatrix}=\begin{bmatrix}-\sum_{i=1}^{n}x_{ij}x_{ij'}(1-y_{1i})\pi_{01}(x_i)(1-\pi_{01}(x_i))\\ -\sum_{i=1}^{n}x_{ij}x_{ij'}y_{1i}\pi_{11}(x_i)(1-\pi_{11}(x_i)\\ -\sum_{i=1}^{n}x_{ij}x_{ij'}\pi_1(x_i)(1-\pi_1(x_i)\end{bmatrix}$$

where $j,j'=0,1,\ldots,p$. We can test for the overall significance of a model using the likelihood ratio test and the dependence can be examined on the basis of η_3 . In case of no dependence, it is expected that $\eta_3=0$

which is evident if, alternatively, $\boldsymbol{\beta_{01}} = \boldsymbol{\beta_{11}}$. We can test the equality of two sets of regression parameters, $\boldsymbol{\beta_{01}}$ and $\boldsymbol{\beta_{11}}$ using the following test statistic:

$$\chi^2 = (\hat{\boldsymbol{\beta}}_{01} - \hat{\boldsymbol{\beta}}_{11})' \Big[\hat{Var}(\hat{\boldsymbol{\beta}}_{01} - \hat{\boldsymbol{\beta}}_{11})\Big]^{-1} (\hat{\boldsymbol{\beta}}_{01} - \hat{\boldsymbol{\beta}}_{11}) \tag{9}$$

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

3. Simulation

To generate correlated binary data for simulations, we used the technique proposed by Leisch *et al.* (1998) employing the bindata package for R. We first simulated one explanatory variable (X) from standard normal distribution. Then we calculated conditional and marginal probabilities. We calculated joint probabilities using the true regression coefficients of the models specified in equations(4)-(7). We use Marshall-Olkin correlation formula

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

to calculate pairwise correlation between the outcome variables, which is equivalent to odd ratio of the outcome variables i.e. log odd ratio η_3. Finally using correlation and two marginal probabilities, we simulate two correlated outcome variables (Y_1, Y_2). Different simulation scenarios were considered varying the degree of dependence between the outcome variables and sample size. Three different structure of dependence were considered: no dependence, mild and high dependence. Under each dependence structure, three different sample size such as 100, 250, and 500 were considered. For each scenarios, the estimates of the regression coefficients were reported as the average of the 500 simulated datasets and bias and coverage of 95% nominal confidence interval (CI) of the estimates ware investigated. Bias was calculated as the difference between the estimated value and true value of the regression coefficients. Coverage was calculated as the proportion of normal based CIs that include the true value. Results showed that the models performed well in all simulation scenarios.

Table 1: Results based on 500 simulations with true odds ratio 1.000 :

sample size	$\hat{OR}$	coefficient	true coef.	estimate	sd	bias	coverage
500	.993	β_{010}	-0.5	-0.509	0.117	0.009	0.948
		β_{011}	0.5	0.499	0.131	0.001	0.960
		β_{110}	-0.7	-0.720	0.189	0.020	0.948
		β_{111}	0.5	0.515	0.163	0.015	0.950
		β_{10}	-1.0	-1.003	0.113	0.003	0.954
		β_{11}	1.0	1.008	0.126	0.008	0.962
250	1.049	β_{010}	-0.5	-0.510	0.166	0.010	0.942
		β_{011}	0.5	0.525	0.188	0.025	0.936
		β_{110}	-0.7	-0.735	0.315	0.035	0.970
		β_{111}	0.5	0.573	0.296	0.073	0.954
		β_{10}	-1.0	-1.023	0.161	0.023	0.948
		β_{11}	1.0	1.021	0.181	0.021	0.950
100	1.051	β_{010}	-0.5	-0.526	0.267	0.026	0.936
		β_{011}	0.5	0.532	0.308	0.032	0.950
		β_{110}	-0.7	-0.784	0.532	0.084	0.958
		β_{111}	0.5	0.582	0.506	0.082	0.960
		β_{10}	-1.0	-1.038	0.259	0.038	0.958
		β_{11}	1.0	1.036	0.292	0.036	0.958

Here in Table 1 we assume no correlation, in Table 2 we assume mild correlation and in Table 3 we assume strong correlation.

Table 2: Results based on 500 simulations with true odds ratio 2.014 :

sample size	$\hat{OR}$	coefficient	true coef.	estimate	sd	bias	coverage
500	2.073	β_{010}	-0.5	0.506	0.117	0.006	0.946
		β_{011}	0.5	0.504	0.131	0.004	0.956
		β_{110}	-0.7	-0.724	0.237	0.024	0.948
		β_{111}	1.2	1.233	0.253	0.033	0.954
		β_{10}	-1.0	-1.012	0.113	0.012	0.942
		β_{11}	1.0	1.008	0.126	0.008	0.954
250	2.109	β_{010}	-0.5	-0.509	0.166	0.009	0.960
		β_{011}	0.5	0.524	0.189	0.024	0.946
		β_{110}	-0.7	-0.759	0.343	0.059	0.956
		β_{111}	1.2	1.270	0.366	0.070	0.952
		β_{10}	-1.0	-1.011	0.161	0.011	0.950
		β_{11}	1.0	1.030	0.181	0.030	0.916
100	2.319	β_{010}	-0.5	-0.503	0.267	0.003	0.958
		β_{011}	0.5	0.538	0.305	0.038	0.966
		β_{110}	-0.7	-0.776	0.577	0.076	0.976
		β_{111}	1.2	1.379	0.644	0.179	0.966
		β_{10}	-1.0	-1.029	0.259	0.029	0.956
		β_{11}	1.0	1.048	0.294	0.048	0.956

Table 3: Results based on 500 simulations with true odds ratio 4.482 :

sample size	$\hat{OR}$	coefficient	true coef.	estimate	sd	bias	coverage
500	4.721	β_{010}	-0.5	-0.507	0.120	0.007	0.936
		β_{011}	-0.5	-0.517	0.128	0.017	0.942
		β_{110}	-0.7	-0.729	0.231	0.029	0.958
		β_{111}	1.0	1.035	0.234	0.035	0.936
		β_{10}	-1.0	-1.007	0.113	0.007	0.954
		β_{11}	1.0	1.007	0.126	0.007	0.950
250	4.646	β_{010}	-0.5	-0.494	0.170	0.006	0.964
		β_{011}	-0.5	-0.503	0.182	0.003	0.968
		β_{110}	-0.7	-0.716	0.326	0.016	0.950
		β_{111}	1.0	1.033	0.334	0.033	0.972
		β_{10}	-1.0	-0.990	0.159	0.010	0.954
		β_{11}	1.0	1.006	0.179	0.006	0.954
100	6.013	β_{010}	-0.5	-0.518	0.274	0.018	0.952
		β_{011}	-0.5	-0.528	0.297	0.028	0.962
		β_{110}	-0.7	-0.894	0.585	0.194	0.966
		β_{111}	1.0	1.266	0.629	0.266	0.978
		β_{10}	-1.0	-1.016	0.257	0.016	0.970
		β_{11}	1.0	1.036	0.291	0.036	0.962

4. Application

An application of the model is provided to child nutrition data from Bangladesh demographic and health survey.

	underweight	
stunted	no	yes
no	77.93%	22.07%
yes	22.02%	77.98%

The responses are Height-for-age (stunted or not) and Weight-for-age (underweight or not) of a child, which are correlated. A set covariates such as mothers education, household socioeconomic status, gender, birth order, preceding birth interval etc is consider. But results are not shown here.

5. Conclusions

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 very strict assumptions. The models based on GEE with various correlation structures have been employed in most of the cases. Another widely used technique is the regressive logistic regression model. However, both these approaches provide either inadequate or, in some instances, misleading results due to use of only marginal or conditional approaches, instead of joint models. We need to specify the bivariate or multivariate outcomes specifying the underlying correlations for a more detailed and more meaningful models. This paper shows the model for bivariate binary data using the conditional and marginal models to specify the joint bivariate probability functions by which the dependence between the outcome variables can be estimated and tested. An application of the joint models is provided to real-life data, which shows that the estimated of the regression coefficient of the models have meaningful interpretation. Simulation studies shows that the model performed well in all simulation scenarios. Therefore, we suggest to use joint modeling approach rather than using only marginal or only condition to analyze correlated binary response.

References

Antelman, G. R. (1972). Interrelated bernoulli processes. Journal of the American Statistical Association, 67(340):831-841.

Azzalini, A. (1994). Logistic regression for autocorrelated data with application to repeated measures. Biometrika, 81(4):767-775.

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(5):1064-1075.

Le Cessie, S. and Van Houwelingen, J. (1994). Logistic regression for correlated binary data. Applied Statistics, pages 95-108.

Leisch, F., Weingessel, A. and Hornik, K. (1998). On the generation off correlated artificial binary data, Working Paper Series, Working Paper No. 13, Vienna University of Economics and Business Administration, Austria.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika, 73(1):13-22.

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(390):332-338.

McCullagh, P., & Nelder, J. A. (1983). Generalized linear models. London, England, Chapman and Hall.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. Biometrics, pages 121-130.