DOI: http://doi.org/10.4038/sljastats.v22i3.8053

Sri Lankan Journal of Applied Statistics, Vol (22-3)

Multivariate Modelling of Binary Responses with Normal and Non–Normal Random Effects

H.A.C.S. Hapuarachchi[1*], S. Samita[2], N. Withanage[3]

[1]Department of Sports Sciences and Physical Education, Sabaragamuwa University of Sri Lanka, Sri Lanka.
[2]Department of Crop Science, University of Peradeniya, Sri Lanka.
[3]Department of Statistics, University of Sri Jayewardenepura, Sri Lanka.

[*]Corresponding Author: sampathhac@appsc.sab.ac.lk

Received: 15th October 2021 / Revised: 2nd December 2021 / Published: 31st December 2021

ABSTRACT

In different fields of study, multivariate binary data is often found, especially when several different qualitative characteristics or attributes are measured in the same unit or from the same person. These bivariate or multivariate responses observed from the same individual or a unit are likely to be correlated. This study aimed to evaluate the influence on the regression estimates of the parameters when binary responses are modeled jointly. The correlation between binary outcomes was captured by incorporating random effects. Normal and bridge distributions were assumed for the random effects. A simulation study was performed to illustrate the impact on the marginal parameter estimates of the joint response model when using the bridge and normal distributions for the random effects. The simulation study revealed that the joint model with either normal or bridge random effects provides a better gain in efficiency in the parameter estimates compared to the individual models which assume responses are independent. Furthermore, the parameter estimates of the joint model are more or less the same under the normal distribution and bridge distribution of the random effects when outcomes are correlated. However, slight differences are noted in the standard errors of the parameter estimates. In addition, when two outcomes are not correlated there is no gain in the fitting joint model over separate univariate models. Finally, these methods were applied to the Bangladesh Demographic and Health Survey 2011 (BDHS 2011) data.

Keywords: bridge distribution, correlated binary responses, GLMMs, joint modeling, random intercept logistic regression

1. Introduction

In different fields of study, multivariate binary data is often found, especially when several different characteristics or attributes are measured in the same unit or from the same individual. It is expected that bivariate or multivariate responses observed from the same individual or a unit tend to be correlated. Ignoring such correlation, and assuming outcomes are independent may lead to invalid inferences and hence lead to misleading conclusions (Glynn and Rosner, 1992). Methods that account for these correlations are thus needed and preferred. Several approaches have been proposed to analyze correlated binary responses (Laird and Ware, 1982; Liang and Zeger, 1986; Zeger and Liang, 1986; Connolly and Liang, 1988; le Cessie and Van Houwelingen, 1994; Wang and Louis, 2003; Ghebremichael, 2015).

Generalized linear mixed models (GLMMs) (Breslow and Clayton 1993; Laird and Ware 1982) are widely used to model correlated binary responses in the recent past and are also suitable for the analysis of non-normal data with multivariate outcomes. Furthermore, GLMMs accommodate correlated and overdispersed data by adding random effects to the linear predictors as an extension of the generalized linear models (Feddag and Mesbah, 2006; Torabi, 2015). In addition, several authors studied the joint modeling of clustered binary outcomes using GLMMs incorporating random effects to the linear predictor to capture the correlation between the outcomes. Ghebremichael (2015) jointly modeled the two binary variables of HSV-2 and HIV-1 infections that incorporate a shared random effect to take into account the interdependence between the two binary variables. Fang et al., (2018) extend Ghebremichael, (2015) approach for clustered data.

The GLMMs for binary data typically adopt logit or probit link functions with random effects on the linear predictor for clustered or correlated binary data. The random intercept logistic regression model is basically used to analyze clustered or correlated binary responses. The random intercept is allowed to capture the variability between the correlated binary data or the clustered binary data (Wang and Louis, 2003). In general, the marginal functional form of the conventional generalized linear mixed model of binary outcomes with logit link and normal random effect is no longer a logistic form (Ghebremichael, 2015). Wang and Louis (2003) assumed a bridge distribution for the random effect in the binary logistic regression, while the marginal functional shape is still logistic and likelihood inference can be obtained for either marginal or conditional regression inference within a single model framework.

This study focuses on studying the joint modeling of correlated binary responses under the GLMMs framework with different distributional assumptions for random effects. The shared random intercept is used to capture the correlation between binary responses. Two distributional assumptions for the shared random intercept were assumed, namely: normal distribution and the bridge distribution (Wang and Louis

2003). The efficiency of the parameter estimates was discussed using relative efficiency. A comprehensive simulations study was conducted to see how the different distributions of random effects influence parameter estimates and standard errors. Finally, the methodology was applied to the data from the Bangladesh Demographic and Health Survey 2011 (BDHS 2011).

The rest of the paper is organized as follows. In Section 2, we give a brief description of the generalized linear models (GLMs) and GLMMs, estimation, and bridge distribution. Section 3 and Section 4 are devoted to simulation study and application to real data, respectively. The final section presents the discussion and conclusions of the research with important findings.

2. Methodology

2.1 Generalized Linear Mixed Models

The general linear model is extensively used to model the continuous responses that follow the normal distribution. Generalized linear models (GLMs) are used when a response variable has a distribution of the exponential family. A random component, a systematic component, and a link function are the three components of the GLMs (McCullagh and Nelder, 1989; Agresti, 2018). The random component of a GLM identifies the distribution of the response variable Y and the typical GLMs treat $y_1, y_2, \dots, y_N$ as an independent.

The systematic component of a GLM specifies the linear function of the covariates, and this linear combination of covariates is called the linear predictor. The linear predictor for covariates $\mathbf{X}_i^T = (x_{i1}, x_{i2}, \dots, x_{ip})$ of Y_i is $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ $(i = 1,2, \dots, n)$ where x_{ik} is the value of the k^{th} predictor $(k = 1,2, \dots p)$ for subject i. The third component of a GLM, the link function, which relates $\mu_i = E(Y_i)$ to the linear predictor η_i through the function $g(\mu_i)$ such that $\eta_i = g(\mu_i)$. That is the link function, which connects the random and systematic components.

Generalized linear mixed models have been used to model correlated binary responses and are appropriate for the analysis of non-normal data with multivariate form. Moreover, GLMMs accommodate correlated and overdispersed data by adding random effects to the linear predictor in the GLMs (Breslow and Clayton 1993; Feddag and Mesbah 2006; Laird and Ware 1982; Torabi 2015). The prime concern of the GLMM is to investigate the fixed effects on response. Whereas the random effects are used to characterize the correlation between observations within the same cluster or correlation between responses when multiple outcomes have been observed on the same subject. Usually, the random effects are assumed to follow a normal (or a multivariate normal) distribution with zero mean and fixed covariance matrix. The

parameters in the random-effects model are conditional effects while effects in the marginal models (integrated random effects) are population averages.

Let Y_{ij} denote the jth response measured for subject i, $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, J$. $\boldsymbol{y}_i$ is the J dimensional vector of all measurements available for subject i. It is assumed that conditionally on q dimensional random effects $\mathbf{b}_i$, the J outcomes are independent, with density functions given $\mathbf{b}_i$ belonging to the exponential family

$$f(Y_{ij}|\theta_{ij}, \varphi, \mathbf{b}_i) = \exp\left[\varphi^{-1}\{Y_{ij}\theta_{ij} - \Psi(\theta_{ij})\} + c(Y_{ij}, \varphi)\right] \tag{2.1}$$

where φ is a scale parameter, $c(.)$ is a function only depending on y_{ij} and φ, and $\Psi(.)$ is a function satisfying $E(\mathrm{Y_{ij}}|\mathbf{b_j}) = \Psi'(\theta_{ij})$ and $\mathrm{Var}(\mathrm{Y_{ij}}|\mathbf{b_j}) = \Psi''(\theta_{ij})$. Further, $\mathrm{E}(\mathrm{Y_{ij}}|\mathbf{b_j}) = \upsilon(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + \mathbf{z}_{ij}^{\mathrm{T}}\mathbf{b_j})$, where $\upsilon(.)$ denotes a known inverse link function, $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ are vectors of covariates, and $\boldsymbol{\beta}$ is a vector of unknown fixed regression coefficients (Abad et al. 2010). In general, it is assumed that $\mathbf{b}_i \sim \mathrm{N_q}(\mathbf{0}, \mathbf{D})$.

Fitting the model requires maximizing the marginal likelihood, which is obtained by integrating over the random effect. Let the contribution of subject i to the marginal likelihood is given by

$$f(\mathbf{y_i}|\beta, \mathrm{D}, \varphi) = \int \prod_{j=1}^{J} f(\mathrm{Y_{ij}}; \theta_{ij}, \varphi|\mathbf{b}_i)\, f(\mathbf{b}_i; \mathbf{D})\mathrm{d}\mathbf{b}_i \tag{2 .2}$$

and the marginal likelihood for all the subjects is

$$f(\mathbf{y}; \beta, \mathrm{D}, \varphi) = \prod_{i=1}^{n} \int \prod_{j=1}^{J} f(\mathrm{Y_{ij}}; \theta_{ij}, \varphi|\mathbf{b}_i)\, f(\mathbf{b}_i; \mathbf{D})\mathrm{d}\mathbf{b}_i. \tag{2.3}$$

Due to possibly high-dimensional intractable integrals involved in this likelihood function, the inferences in GLMMs are made cumbersome and hence, numerical approximations are needed to solve the integrals in this likelihood. In general, the numerical approximations can be classified into those based on approximations of the integrand, those based on approximations of the data, and those based on approximations of the integral itself (Molenberghs and Verbeke, 2006).

2.2 Joint Model for Two Correlated Binary Responses

Let Y_{ij} denote the jth binary response ($j = 1, 2$) variable of the ith subject ($i = 1, 2, \dots, n$) and $\boldsymbol{Y}_i^T = [Y_{i1}, Y_{i2}]$ is the binary outcome vector of the ith subject. Let $\boldsymbol{x}_{i1}$ and $\boldsymbol{x}_{i2}$ be the vector of covariates associated with outcomes Y_{i1} and Y_{i2}, respectively, and $\boldsymbol{\beta}_j$ ($j = 1, 2$) is the corresponding vector of regression coefficients. We assume a set of latent, random effects b_i are shared by two responses of the same individual. This shared random intercept captures the unobserved factors specific to each

individual which may influence the responses and tends to account for correlation between the two outcomes of the same individual. The random intercepts are assumed to vary independently from one individual to another. Given the random effects, b_i, with logit transformation, the GLMM is given by,

$$\text{logit}\{\text{E}(\text{Y}_{\text{ij}}|b_i, \boldsymbol{x}_{ij})\} = \text{logit}\{\Pr(\text{Y}_{\text{ij}} = 1|b_i, \boldsymbol{x}_{ij})\} = b_i + \boldsymbol{\beta}_j^T \boldsymbol{x}_{ij} \tag{2.4}$$

Suppose the distribution function of the random effects b_i is $G(b_j)$. Under the GLMM framework, it is assumed that Y_{i1}, Y_{i2} are independent given the random effects b_i. Then the likelihood function of the n individuals, $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_n$, is given by

$$\text{L}(.) = \prod_{\text{i}=1}^{\text{n}} \left\{ \int \prod_{j=1}^{2} \left(\frac{e^{b_i + \boldsymbol{\beta}_j^T \boldsymbol{x}_{ij}}}{1 + e^{b_i + \boldsymbol{\beta}_j^T \boldsymbol{x}_{ij}}} \right)^{y_{ij}} \left(\frac{1}{1 + e^{b_i + \boldsymbol{\beta}_j^T \boldsymbol{x}_{ij}}} \right)^{1 - y_{ij}} dG(b_j) \right\} \tag{2.5}$$

2.3.1 Joint model assuming a normal distribution for the random effect

Under the GLMMs framework, it is common practice to assume that random effects follow multivariate (or univariate) normal distribution with mean **0** and variance-covariance matrix $\boldsymbol{D}$ (Breslow and Clayton 1993; Laird and Ware 1982). In our model framework, first, we assumed that $b_i \sim N(0, \sigma_b^2)$ and with this assumption, Equation (2.5) does not have a closed-form expression. This leads to intractable expressions for marginal and conditional distributions. Numerous methods are available in the literature to solve the above integral in Equation (2.5). For example, as an integral approximation, Ghebremichael, (2015) used the Gaussian Adaptive Quadrature (GAQ) approximation for parameter estimation. The GAQ gives a better approximation of the integrals (Cagnone & Monari, 2013), which is more important when used for ordinal data with small sample sizes (Rabe-Hesketh, Skrondal and Pickles, 2005; Cagnone and Monari, 2013). Furthermore, maximum likelihood estimates based on adaptive quadrature have more desirable properties like these estimators are consistent and asymptotically normally distributed (Liu and Pierce, 1994; Bianconcini, 2014; Ghebremichael, 2015).

Unlike, general linear mixed models with normal random effects, the above model does not preserve the same logit link function for marginal and conditional means. For a discussion on this topic, we refer to Ghebremichael (2015). This implies that parameters in equation (2.4) should be interpreted conditional on the shared random effect. Hence, these parameter estimates are not directly comparable with the parameter estimates of the binary logistic regression models. However, following Ghebremichael (2015) the logit of the marginal mean can be approximated by

$$\text{logit}\left[\Pr\left(Y_{ij}=1\right)\right] \approx \frac{1}{\sqrt{1+c^2\sigma_b^2}}\boldsymbol{\beta}_j^T\boldsymbol{x}_{ij}$$

where $c=\frac{16\sqrt{3}}{15\pi}$the marginal parameters which correspond to the conditional parameters ($\boldsymbol{\beta_j}$) is given by

$$\tilde{\boldsymbol{\beta}}_{\boldsymbol{j}}^{\ T} \approx \frac{1}{\sqrt{1+c^2\sigma_b^2}}\boldsymbol{\beta_j}^T .$$

2.3.3 Joint model assuming bridge distribution for the random effect

Wang & Louis (2003) proposed bridge distribution for b_i which preserves the same link function for the marginal and conditional means. The advantage of this approach is likelihood inference can be obtained for either marginal or conditional regression parameters within a single model framework. The bridge density function for the logit link is,

$$f(b_i)=\frac{1}{2\pi}\frac{\sin(\phi\pi)}{\cosh(\phi b_i)\ +\cos(\phi\pi)} \tag{2.6}$$

with $\phi=\left(1+\frac{3}{\pi^2}d\right)^{-\frac{1}{2}}$ and d the random-intercept variance (Wang and Louis, 2003). With the above bridge distribution for the random effects, the marginal distribution can be modeled directly by,

$$\text{logit}\left[\Pr\left(\ Y_{ij}=1\right)\right]=\boldsymbol{\beta}_{mj}^T\boldsymbol{x}_{ij} \tag{2.7}$$

where $\boldsymbol{\beta}_{mj}$ measures a marginal regression effect associated with the covariate $\boldsymbol{x}_{ij}$ for the j^{th} response variable.

The relationship of regression parameters between conditional and marginal regression models is given by

$$\boldsymbol{\beta}_{mj}=\frac{1}{\sqrt{(1+3d/\pi^2)}}\ \boldsymbol{\beta}_j.$$

2.4 The *phi*-coefficient

The *phi*-coefficient (also known as Pearson's *phi*-coefficient) is used to measure the strength of association between two binary variables. The *phi*-coefficient (r_{Phi}) is derived from Pearson's Chi-Square statistic of tabular association. The modifications restrict the resulting statistic to a range of −1.0 to 1.0, analogously to (although not

the same as) Pearson's Correlation Coefficient. If the variables are not associated, then the r_{Phi} value should be 0; a perfect positive (negative) association yields a r_{Phi} of 1 (−1) (Guilford, 1941; De Cáceres, Font and Oliva, 2008; Ekström, 2011).

Let P_{11}, P_{10}, P_{01}, and P_{00} denote the joint probabilities of the two binary random variables (Y_1, Y_2) and $P_1 = P(Y_1 = 1)$, $P_2 = P(Y_2 = 1)$ are the marginal probabilities of Y_1 and Y_2 (Table 2.1).

Table 2.1: Joint and marginal probabilities of random variables Y_1 and Y_2.

		Y_2		
		1	0	
Y_1	1	P_{11}	P_{10}	P_1
	0	P_{01}	P_{00}	$1 - P_1$
		P_2	$1 - P_2$	

The *phi*-coefficient (r_{Phi}) is the linear correlation between postulated underlying discrete univariate distributions of Y_1 and Y_2. Following Ekström (2011), r_{Phi} is given by

$$r_{\text{Phi}} = \frac{P_{11} - P_1 P_2}{\sqrt{P_1 P_2 (1 - P_1)(1 - P_2)}}$$

$$P_{11} = P_1 P_2 + r_{\text{Phi}} \sqrt{P_1 P_2 (1 - P_1)(1 - P_2)} \tag{2.8}$$

3. Simulation Study

Simulation studies were carried out to study the finite sampling behavior of the estimates of the joint model assuming (1) normal distribution for the random effects, $b_i \sim N(0, \sigma_b^2)$, and (2) bridge distribution, for the random effects, Further, it was compared the joint model estimates with marginal model estimates which ignore the correlation between two binary outcomes. Two sets of simulation studies were conducted with varying sample sizes $n = 20$, 50, 100, 200, 300, and 500 and with varying levels of correlations between two binary outcomes; $\rho = 0.0, 0.3, 0.5$, and 0.7.

R software version 4.0.2 was used to simulate data and "PROC NLMIXED" in SAS University Edition® was used to estimate the parameters in the joint model.

For two correlated binary outcomes, (Y_{i1}, Y_{i2}) following joint marginal model with a continuous covariate, X_{i1}, and a binary covariate, X_{i2}, was used for the simulation study.

$$\log\left[\frac{P_{i1}}{1-P_{i1}}\right] = \beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2}$$

$$\log\left[\frac{P_{i2}}{1-P_{i2}}\right] = \beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2} \tag{2.9}$$

where,

$$P(Y_{i1} = 1) = \mathrm{P}_{i1} = \frac{exp(\beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2})}{1 + exp(\beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2})}$$

and

$$P(Y_{i2} = 1) = \mathrm{P}_{i2} = \frac{exp(\beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2})}{1 + exp(\beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2})}.$$

The covariates X_{i1} and X_{i2} for the ith individual were generated from U (0, 2) and Bernoulli (0.5) distributions, respectively. Following Lipsitz et al (1990), correlated binary outcomes (Y_{i1}, Y_{i2}) were generated as follows. First, by Equation (2.8) and for a given correlation coefficient (ρ) between the underlying distributions of Y_{i1} and Y_{i2}, the joint probabilities P_{11}, P_{10}, P_{01} and P_{00} were computed. Next, given the joint probabilities, P_{11}, P_{10}, P_{01} and P_{00}, the outcomes of the ith individuals (Y_{i1}, Y_{i2}) was generated using the multinomial distribution.

The parameter configurations for both simulation studies were $\beta_{10} = 1.25$, $\beta_{11} = 0.50$, $\beta_{12} = 1.00$, $\beta_{20} = 1.50$, $\beta_{21} = 0.5$, and $\beta_{22} = 0.75$. In the first simulation study, a total of 1000 repeated samples of size $n = 20, 50, 100, 200, 300$, and 500 were generated with $\rho = 0.50$.

In the second simulation study, the performance of the parameter estimates of two joint models was compared under varying levels of correlations $\rho = 0.0, 0.3, 0.5$ and 0.7 and with a fixed sample size $n = 100$. For each combination of ρ and parameter value, a total of 1000 simulations were performed. The performance of the estimates was evaluated based on the standard errors of the estimates, relative efficiency, and the coverage probability (CP). The relative efficiency of an estimate is computed by dividing the average of standard errors by the empirical standard deviation of the estimate, where the empirical standard deviation is the standard deviation of the estimated obtained by 1000 repeated samples (Withanage et al. 2015), and the CP was computed as the number of times 95% confidence intervals contained the true parameter.

The results are presented in Tables 3.1, 3.2 A, and 3.2 B. Results suggest that parameter estimates of the two joint models perform well in finite samples. Parameter estimates are generally approached the true parameter with increasing sample size. More importantly, the joint model provides smaller standard errors compared to the standard errors of the marginal model estimates when two outcomes are correlated. Besides, the coverage probability is considerably small for the fewer sample sizes. Both the joint model GLMMs with normal random effect and Bridge random effect perform equally. Amini et al., (2018); Heagerty & Kurland (2001; Ghebremichael (2015); Heagerty & Kurland (2001) and Wang & Louis (2003) have also shown somewhat similar results to ours. In addition, lower standard errors were reported when fitting the univariate models when the two outcomes are not correlated to each other compared to the joint models.

4. Application to Bangladesh Demographic and Health Survey

4.1 Study population and measures

Data from the 2011 Bangladesh Demographic and Health Survey was used to study the use of the joint modeling of the correlated bivariate binary responses with the real-life data. It is the sixth national-level demographic and health survey aimed to provide information on demographic and maternal and child health in Bangladesh. The survey data used for this study consisted of 17,842 ever-married women, ages 12 to 49 (BDHS 2011, 2013).

Two binary questions related to knowledge of HIV prevention are considered: (1) Y_1 = "*Thinks having only one sex partner reduces AIDS risk (yes/no)*", (2) Y_2= "*Thinks always using a condom reduces AIDS risk (yes/no)*". It is reasonable to assume that Y_1 and Y_2 are correlated outcomes since the two outcomes are observed from the same individual and are likely to be interrelated.

Table 3.1 Summary results based on 1000 simulations under ρ values 0, 0.3, 0.5, and 0.7 with n = 100

Parameter	True	ρ	Bridge				Normal				Univariate			
			Ave Est	Ave SE	RE	CP	Ave Est	Ave SE	RE	CP	Ave Est	Ave SE	RE	CP
β_{10}	1.25	0.0	1.3488	0.7067	0.9116	0.963	1.3501	0.7067	0.9127	0.963	1.3437	0.7049	0.9378	0.959
		0.3	1.3414	0.6957	0.9554	0.959	1.3522	0.6859	0.9280	0.953	1.3406	0.7027	0.9401	0.959
		0.5	1.3557	0.6798	0.8583	0.923	1.3356	0.6668	0.8786	0.917	1.2940	0.6966	0.9607	0.957
		0.7	1.5149	0.6290	0.7967	0.838	1.4488	0.6461	0.7355	0.801	1.0306	0.7714	1.0437	0.959
β_{11}	0.50	0.0	0.5252	0.6766	0.9372	0.964	0.5307	0.6822	0.9236	0.966	0.5272	0.6753	0.9423	0.959
		0.3	0.5606	0.6822	0.9127	0.966	0.5475	0.6671	0.8791	0.962	0.5456	0.6871	0.9168	0.968
		0.5	0.5513	0.6852	0.7899	0.916	0.5247	0.6544	0.7994	0.933	0.5451	0.6941	0.8690	0.952
		0.7	0.5636	0.6793	0.7608	0.916	0.5312	0.6257	0.7529	0.862	0.5269	0.6726	0.9392	0.966
β_{12}	1.00	0.0	1.0081	0.7792	1.0344	0.952	1.0072	0.7791	1.0349	0.950	1.0075	0.7793	1.0354	0.969
		0.3	0.9916	0.7755	0.9790	0.944	0.9743	0.7478	0.9595	0.945	0.9893	0.7757	0.9830	0.959
		0.5	1.0058	0.7617	1.0132	0.923	0.9588	0.7155	0.9824	0.936	1.0325	0.7755	1.0375	0.974
		0.7	0.9914	0.7216	0.9266	0.859	0.9699	0.7049	0.9183	0.841	1.0306	0.7714	1.0437	0.973
β_{20}	1.50	0.0	1.5925	0.7519	0.9672	0.963	1.5993	0.7549	0.9512	0.963	1.5932	0.7521	0.9673	0.964
		0.3	1.5911	0.7489	0.9689	0.966	1.6113	0.7371	0.9464	0.965	1.5828	0.7528	0.9770	0.968
		0.5	1.6760	0.7428	0.8568	0.939	1.6512	0.7239	0.8676	0.943	1.5818	0.7569	0.9335	0.960
		0.7	1.7878	0.7119	0.7718	0.909	1.7130	0.6862	0.8272	0.877	1.5393	0.7409	0.9981	0.969
β_{21}	0.50	0.0	0.5297	0.6377	0.9127	0.971	0.5359	0.6405	0.8684	0.970	0.5254	0.6361	0.9204	0.966
		0.3	0.5359	0.6281	0.9368	0.959	0.5285	0.6153	0.8967	0.954	0.5318	0.6325	0.9486	0.964
		0.5	0.5712	0.6192	0.8215	0.925	0.5604	0.5993	0.8194	0.931	0.5434	0.6343	0.9036	0.956
		0.7	0.4526	0.5882	0.8168	0.842	0.5247	0.5898	0.7123	0.853	0.5535	0.6305	0.9417	0.954
β_{22}	0.75	0.0	0.7513	0.7959	1.0656	0.982	0.7509	0.7960	1.0636	0.981	0.7511	0.7964	1.0644	0.974
		0.3	0.7114	0.7948	1.0439	0.981	0.6877	0.7646	1.0310	0.968	0.7125	0.7954	1.0499	0.972
		0.5	0.7224	0.7944	1.0203	0.952	0.6948	0.7432	1.0234	0.941	0.7345	0.8031	1.0352	0.973
		0.7	0.7342	0.7529	1.0431	0.889	0.6485	0.7058	1.0191	0.856	0.7683	0.7951	1.0709	0.978

Ave Est = *Average estimates;* ***Ave SE*** = *Average standard errors; and* ***RE*** = *Relative efficiency*

Table 3.2A Summary results based on 1000 simulations under sample size 20, 50, 100, 200, 300, and 500 when $\rho = 0.5$ [Response variable Y_1]

Parameter	True	n	Bridge				Normal				Univariate			
			Ave Est	Ave SE	RE	CP	Ave Est	Ave SE	RE	CP	Ave Est	Ave SE	RE	CP
β_{10}	1.25	20	0.8840	1.4080	1.0302	0.606	0.9176	1.5155	1.0845	0.649	0.8165	1.6367	1.2047	0.763
		50	1.2211	0.9949	0.9436	0.848	1.2266	0.9923	0.9718	0.882	1.2229	1.0375	1.0246	0.936
		100	1.3557	0.6798	0.8583	0.923	1.3356	0.6668	0.8786	0.917	1.2940	0.6966	0.9607	0.957
		200	1.3158	0.4722	0.9659	0.951	1.3267	0.4494	0.9702	0.911	1.2696	0.4733	0.9843	0.963
		300	1.2818	0.3776	1.0059	0.958	1.3350	0.3612	0.9453	0.909	1.2558	0.3817	0.9810	0.950
		500	1.2569	0.2965	1.0142	0.952	1.3481	0.2782	0.9257	0.893	1.2429	0.2928	0.9984	0.845
β_{11}	0.50	20	0.4117	1.5059	1.0807	0.603	0.3737	1.5388	1.1161	0.620	0.3847	1.6464	1.2073	0.788
		50	0.5647	1.0276	0.9655	0.867	0.5607	1.0280	0.9646	0.900	0.5378	1.0548	0.9882	0.957
		100	0.5513	0.6852	0.7899	0.916	0.5247	0.6544	0.7994	0.933	0.5451	0.6941	0.8690	0.952
		200	0.5420	0.4553	0.9082	0.929	0.5043	0.4158	0.8783	0.919	0.5425	0.4592	0.9274	0.954
		300	0.5129	0.3594	1.0352	0.961	0.4710	0.3240	0.9643	0.908	0.5356	0.3673	0.9876	0.954
		500	0.5101	0.2818	0.9826	0.941	0.4439	0.2443	0.8904	0.881	0.5235	0.2808	0.9568	0.943
β_{12}	1.00	20	0.7417	1.3143	1.0632	0.375	0.6539	1.3589	1.1377	0.363	0.5103	1.4794	1.4990	0.431
		50	0.8662	1.0410	1.0998	0.732	0.8007	1.0002	1.1323	0.731	0.8330	1.0682	1.2127	0.782
		100	1.0058	0.7617	1.0132	0.923	0.9588	0.7155	0.9824	0.936	1.0325	0.7755	1.0375	0.974
		200	1.0787	0.5428	0.8887	0.947	0.9741	0.4907	0.8297	0.886	1.0677	0.5370	0.9306	0.965
		300	1.0440	0.4188	0.9355	0.957	0.9092	0.3878	0.8135	0.854	1.0515	0.4284	0.9209	0.956
		500	1.0248	0.3216	1.0073	0.959	0.8459	0.2954	0.8553	0847	1.0366	0.3252	0.9819	0.960

Ave Est = *Average estimates;* ***Ave SE*** = *Average standard errors; and* ***RE*** = *Relative efficiency*

Table 3.2B Summary results based on 1000 simulations under sample size 20, 50, 100, 200, 300, and 500 when $\rho = 0.5$ [Response variable Y_2]

Parameter	True	n	Bridge				Normal				Univariate			
			Ave Est	Ave SE	RE	CP	Ave Est	Ave SE	RE	CP	Ave Est	Ave SE	RE	CP
β_{20}	1.50	20	1.0651	1.4570	1.0413	0.532	1.0800	1.5423	1.1114	0.550	1.0447	1.6688	1.2311	0.675
		50	1.5570	1.0589	1.0146	0.823	1.5285	1.0566	1.0098	0.856	1.5067	1.1157	1.0645	0.902
		100	1.6760	0.7428	0.8568	0.939	1.6512	0.7239	0.8676	0.943	1.5818	0.7569	0.9335	0.960
		200	1.5989	0.5109	0.9490	0.944	1.5875	0.4756	0.9456	0.921	1.5445	0.5103	0.9575	0.956
		300	1.5593	0.4037	0.9986	0.956	1.5780	0.3759	0.9670	0.902	1.5185	0.4093	0.9724	0.949
		500	1.5330	0.3141	0.9773	0.947	1.5887	0.2868	0.9249	0.904	1.5104	0.3139	0.9673	0.953
β_{21}	0.50	20	0.5164	1.4465	1.0595	0.649	0.4365	1.5001	1.0940	0.687	0.5224	1.6010	1.1931	0.834
		50	0.6377	0.9463	0.8876	0.876	0.6230	0.9451	0.9217	0.889	05743	0.9764	0.9729	0.948
		100	0.5712	0.6192	0.8215	0.925	0.5604	0.5993	0.8194	0.931	0.5434	0.6343	0.9036	0.956
		200	0.5126	0.4241	0.9691	0.937	0.4939	0.3914	0.9315	0.919	0.5343	0.4279	0.9469	0.948
		300	0.5083	0.3387	0.9852	0.951	0.4672	0.3088	0.9124	0.908	0.5284	0.3442	0.9478	0.954
		500	0.5162	0.2617	0.9925	0.957	0.4470	0.2337	0.9193	0.888	0.5231	0.2633	0.9844	0.956
β_{22}	0.75	20	0.4660	1.3231	1.0494	0.343	0.4058	1.3198	1.0522	0.320	0.2965	1.4899	1.4851	0.395
		50	0.5879	1.0565	1.1739	0.698	0.5089	1.0158	1.2459	0.687	0.5464	1.1033	1.2945	0.759
		100	0.7224	0.7944	1.0203	0.952	0.6948	0.7432	1.0234	0.941	0.7345	0.8031	1.0352	0.973
		200	0.7460	0.5561	0.9305	0.969	0.6877	0.5012	0.8714	0.925	0.7737	0.5529	0.9326	0.972
		300	0.7536	0.4309	0.9703	0.957	0.6436	0.3915	0.8573	0.883	0.7702	0.4404	0.9535	0.964
		500	0.7521	0.3303	1.0130	0.962	0.6000	0.2961	0.8828	0.848	0.7594	0.3347	0.9666	0.953

Ave Est *= Average estimates;* ***Ave SE*** *= Average standard errors; and* ***RE*** *= Relative efficiency*

The objective of this application is to identify the economic and demographic factors associated with the above two outcome variables.

The covariates related to demographic factors included current age [age], total children ever born [children], highest educational level [education], urban-rural status [urban], religion, region [region] and place of residence [place]. The economic factors considered in the analysis are the living condition of the household, whether the household had a radio [radio], whether the household had a television [television], and the family wealth index [wealth index]. In addition to the above variable some sociological variables have been considered: whether she is reading a newspaper or magazine [readpaper] and can she be able to read [read]. The Educational level has four categories namely no education, primary school, secondary school, and higher education and above. The wealth index is an ordinal family level covariate having five categories, namely poorest, poorer, middle, richer, and richest. The objective of this application is to identify the socio-economic and demographic factors associated with the above two outcome variables.

4.2 Results

The Pearson chi-square test of the interdependence of Y_1 and Y_2 is highly significant indicating that two binary outcomes are statistically associated (Chi-Square = 11027.0; $p < 0.0001$). The *phi*-coefficient between Y_1 and Y_2 is positive implying that one who believes Y_1 is true is more likely to believe that Y_2 is also true (r_{Phi} = 0.3395).

The fitted joint models of Y_1 and Y_2 are

$$\begin{aligned}\text{logit}\{\Pr(Y_{i1} = 1|b_i)\} &= \beta_{10} + \beta_{11}[\text{age}]_i + \beta_{12}[\text{read}]_i + \beta_{13}[\text{region}]_i + \beta_{14}[\text{urban}]_i \\ &+ \beta_{15}[\text{readpaper}]_i + \beta_{16}[\text{education\_0}]_i + \beta_{17}[\text{education\_1}]_i \\ &+ \beta_{18}[\text{education\_2}]_i + b_i\end{aligned}$$

$$\begin{aligned}\text{logit}\{\Pr(Y_{i2} = 1|b_i)\} &= \beta_{20} + \beta_{21}[\text{age}]_i + \beta_{22}[\text{read}]_i + \beta_{23}[\text{region}]_i + \beta_{24}[\text{urban}]_i \\ &+ \beta_{25}[\text{readpaper}]_i + \beta_{26}[\text{education\_0}]_i + \beta_{27}[\text{education\_1}]_i \\ &+ \beta_{28}[\text{education\_2}]_i + b_i\end{aligned}$$

where b_i is the shared random effect assuming normal or bridge distribution.

Table 4.1: Marginal regression coefficients for joint model (bridge and normal random effect) and independence model for outcome variables Y_1 and Y_2

Parameter	Bridge			Normal			Univariate		
	Estimate	**SE**	**p-value**	**Estimate**	**SE**	**p-value**	**Estimate**	**SE**	**p-value**
Outcome variable Y_1									
Intercept	1.2446	0.1839	<0.0001	1.2812	0.1813	<0.0001	1.2149	0.18600	<0.0001
age	0.01014	0.00333	0.0023	0.00989	0.00328	0.0026	0.00987	0.00338	0.0034
read	0.1965	0.1077	0.0683	0.1952	0.1081	0.0711	0.1970	0.10910	0.0708
region	-0.04156	0.01471	0.0047	-0.0392	0.01461	0.0073	-0.0379	0.01490	0.0110
urban	0.1359	0.05724	0.0176	0.1322	0.05642	0.0191	0.1295	0.05800	0.0256
readpaper	0.119	0.07626	0.1187	0.1281	0.07459	0.0860	0.1352	0.07760	0.0814
education_0	-0.1165	0.1579	0.4604	-0.09481	0.1564	0.5443	-0.0856	0.15970	0.5923
education_1	-0.2124	0.1118	0.0574	-0.19080	0.1091	0.0804	-0.1828	0.11310	0.1060
education_2	-0.08939	0.09639	0.3538	-0.07342	0.09338	0.4317	-0.0710	0.09740	0.4660
Outcome variable Y_2									
Intercept	1.256	0.1749	<0.0001	1.2773	0.1752	<0.0001	1.2293	0.1764	<0.0001
age	0.00794	0.00313	0.0113	0.00788	0.00315	0.0123	0.00775	0.00316	0.0143
read	0.3082	0.09935	0.0019	0.3142	0.1024	0.0022	0.3083	0.1002	0.0021
region	-0.04162	0.01387	0.0027	-0.0401	0.01404	0.0043	-0.0390	0.0140	0.0055
urban	0.2005	0.05416	0.0002	0.1993	0.05435	0.0002	0.1947	0.0547	0.0004
readpaper	0.03795	0.07234	0.5998	0.0492	0.07195	0.4941	0.0514	0.0734	0.4837
education_0	-0.3947	0.1500	0.0085	-0.3638	0.1509	0.0159	-0.3671	0.1515	0.0154
education_1	-0.5531	0.1094	<0.0001	-0.5265	0.1074	<0.0001	-0.5261	0.1105	<0.0001
education_2	-0.3303	0.09586	0.0006	-0.3066	0.09296	0.0010	-0.3121	0.0968	0.0013
σ_b	2.0498	0.06759	<0.0001	1.9654	0.06328	<0.0001			
ϕ	0.6627	0.01225	<0.0001						
AIC	16686			16690					
-2 Log L	16648			16652					

The results are reported in Table 4.1. Results show that the estimates based on two joint models (bridge and normal) are almost the same, but the standard errors of the parameter estimate obtained by assuming the bridge distribution for random effect are slightly smaller than the standard errors of the parameter estimates estimated by assuming the normal distribution for the random effect. Further, these standard errors of the joint model are lower than the standard errors of the independent model. Although the estimates based on two joint models are almost alike, the AIC values are slightly different, and the lowest AIC is observed under the assumption of a bridge distribution for the random intercept. This implies that the joint model with bridge random effect is slightly better than the normal random effect model. The significant variations of the random intercept for the joint model show a high level of heterogeneity between subjects ($p < 0.0001$). The rescaling parameter ϕ for the model with the bridge distribution was estimated as 0.6627.

However, when considering the model parameter estimates, some insignificant parameters were kept in the model since the one of main purposes of this study was, not only to identify the best-fitted model but also wanted to compare how the parameter estimates and their corresponding standard errors vary with the different distributional assumptions of random effects. In the joint model, the covariate of "ability to read newspapers" is not statistically significant under the normal random effect model of Y_1 but it is significant under the bridge random effect model of Y_1 at a 10% level of significance.

The covariates of the respondent's age, education level, the region where she lived, urban-rural status, and whether she is able to read are significant factors related to the knowledge of HIV at a 10% level of significance after adjusting for other factors. However, reading newspapers, watching television, the total number of children she has, and religion are not associated with knowledge of HIV of ever-married women in Bangladesh.

5. Conclusion

In this study, we jointly modeled two binary outcomes under the framework of GLMMs with a shared random effect. Two different distributional assumptions for random effect were assumed, namely, bridge distribution and normal distribution. Parameter estimates and standard errors of the joint model were compared with the independent model which assumes two outcomes are independent using two simulation studies. In the simulation study, correlated binary responses were generated without incorporating the random effect into the linear predictors of the model. Therein we wanted to know, under GLMMs with different distributional assumptions for random effects, whether GLMMs can equally perform regardless of

the distribution of the random effects. Finally, the results were compared with Bangladesh health survey data.

Simulations study reveals that jointly modeling binary data under GLMMs with a bridge or a normal shared random effect performs more or less the same. The results of the simulation studies suggest that joint modeling of the correlated binary responses provides a better gain in efficiency in the parameter estimates. In addition, when two outcomes are not correlated there is no gain in the fitting joint model over separate univariate models.

As for modeling the knowledge of HIV prevention by GLMMs, it was seen that the estimates of the marginal parameters that were obtained from the joint models have smaller values in standard errors compared to the independent models. These findings upheld the need for a joint model of two correlated binary responses.

Acknowledgment

The authors thank the Demographic and Health Surveys (DHS) program for providing free access to DHS data sets of Bangladesh (https://www.idhsdata.org/idhs/).

References

1. Abad, A.A., Litière, S. and Molenberghs, G. (2010) 'Testing for misspecification in generalized linear mixed models', *Biostatistics*, 11(4), pp. 771–786.

2. Agresti, A. (2018) *An introduction to categorical data analysis*. Third Edition. John Wiley & Sons.

3. Amini, P. *et al.* (2018) 'Longitudinal joint modelling of binary and continuous outcomes: A comparison of bridge and normal distributions', *Epidemiology, Biostatistics and Public Health*, 15(1).

4. *Bangladesh Demographic and Health Survey 2011* (2013). Dhaka, Bangladesh: NIPORT, Dhaka, Bangladesh.

5. Bianconcini, S. (2014) 'Asymptotic properties of adaptive maximum likelihood estimators in latent variable models', *Bernoulli*, 20(3), pp. 1507–1531.

6. Breslow, N.E. and Clayton, D. (1993) 'Approximate inference in generalized linear mixed models', *Journal of the American Statistical Association*, 88(421), pp. 9–25.

7. Cagnone, S. and Monari, P. (2013) 'Latent variable models for ordinal data by using the adaptive quadrature approximation', *Computational Statistics*, 28(2), pp. 597–619.

8. le Cessie, S. and Van Houwelingen, J.C. (1994) 'Logistic regression for correlated binary data', *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1), pp. 95–108.

9. Connolly, M.A. and Liang, K.-Y. (1988) 'Conditional logistic regression models for correlated binary data', *Biometrika*, 75(3), pp. 501–506. doi:10.1093/biomet/75.3.501.

10. De Cáceres, M., Font, X. and Oliva, F. (2008) 'Assessing species diagnostic value in large data sets: A comparison between phi-coefficient and Ochiai index', *Journal of Vegetation science*, 19(6), pp. 779–788.

11. Ekström, J. (2011) 'The phi-coefficient, the tetrachoric correlation coefficient, and the Pearson-Yule Debate', *Department of Statistics, UCLA*. [Preprint]. Available at: https://escholarship.org/uc/item/7qp4604r.

12. Fang, D., Sun, R. and Wilson, J.R. (2018) 'Joint modeling of correlated binary outcomes: The case of contraceptive use and HIV knowledge in Bangladesh', *PloS one*, 13(1), p. e0190917.

13. Feddag, M.L. and Mesbah, M. (2006) 'Approximate estimation in generalized linear mixed models with applications to the Rasch model', *Computers & Mathematics with Applications*, 51(2), pp. 269–278.

14. Ghebremichael, M. (2015) 'Joint modeling of correlated binary outcomes: HIV-1 and HSV-2 co-infection', *Journal of Applied Statistics*, 42(10), pp. 2180–2191.

15. Glynn, R.J. and Rosner, B. (1992) 'Accounting for the correlation between fellow eyes in regression analysis', *Archives of ophthalmology*, 110(3), pp. 381–387.

16. Guilford, J.P. (1941) 'The phi coefficient and chi square as indices of item validity', *Psychometrika*, 6(1), pp. 11–19.

17. Heagerty, P.J. and Kurland, B.F. (2001) 'Misspecified maximum likelihood estimates and generalised linear mixed models', *Biometrika*, 88(4), pp. 973–985.

18. Laird, N.M. and Ware, J.H. (1982) 'Random-Effects Models for Longitudinal Data', *Biometrics*, 38(4), pp. 963–974. doi:10.2307/2529876.

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

20. Lipsitz, S.R., Laird, N.M. and Harrington, D.P. (1990) 'Maximum likelihood regression methods for paired binary data', *Statistics in Medicine*, 9(12), pp. 1517–1525.

21. Liu, Q. and Pierce, D.A. (1994) 'A note on Gauss—Hermite quadrature', *Biometrika*, 81(3), pp. 624–629.

22. McCullagh, P. and Nelder, J.A. (1989) *Generalized linear models*. Second Edition. Chapman and hall/CRC.

23. Molenberghs, G. and Verbeke, G. (2006) *Models for discrete longitudinal data*. Springer Science & Business Media.

24. Rabe-Hesketh, S., Skrondal, A. and Pickles, A. (2005) 'Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects', *Journal of Econometrics*, 128(2), pp. 301–323.

25. Torabi, M. (2015) 'Likelihood inference for spatial generalized linear mixed models', *Communications in Statistics-Simulation and Computation*, 44(7), pp. 1692–1701.

26. Wang, Z. and Louis, T.A. (2003) 'Matching conditional and marginal shapes in binary random intercept models using a bridge distribution function', *Biometrika*, 90(4), pp. 765–775.

27. Withanage, N., de Leon, A.R. and Rudnisky, C.J. (2015) 'Joint estimation of multiple disease-specific sensitivities and specificities via crossed random effects models for correlated reader-based diagnostic data: application of data cloning', *Statistics in medicine*, 34(29), pp. 3916–3928.

28. Zeger, S.L. and Liang, K.-Y. (1986) 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, pp. 121–130.