Journal Homepage: - www.journalijar.com
INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)
ISSN NO. 2320-5407

Article DOI: 10.21474/IJAR01/6101
DOI URL: http://dx.doi.org/10.21474/IJAR01/6101

RESEARCH ARTICLE

GENERATING MULTIVARIATE LONGITUDINAL BINARY RANDOM VARIABLES FOR GEE MODELS USING BRIDGE DISTRIBUTION.

Hissah Alzahrani.
Mathematical science department, college of applied sciences, umm al-qura university, mekkah, 24382, saudi arabia.

Manuscript Info

Manuscript History

Received: 19 October 2017
Final Accepted: 20 November 2017
Published: December 2017

Key words:-
GEE, Bridge distribution, binary , longitudinal simulation

Abstract

Generalized estimating equations (GEE) models are often used to analyze the longitudinal data.It accounts for the within-subject associations through specification of working correlation matrixR. In multivariate longitudinal data, the within-subject correlation is computed by many outcomesare measured over many occasions. Then, the correlation is the main problem in the multivariate longitudinal data. This complicated correlation may affect the parameter estimations precision when it is increased over the outcomes or occasions. Designing a simulation method to investigate the correlation effects on the parameter estimations for the marginal models could be good statistical tool in the longitudinal data analysis. In this paper, we utilize a method to generate correlated binary data for a multivariate longitudinal model with specified R correlation matrix. This specified structure allows the correlation to be induced over the outcomes or occasions. We utilized the methods of Wang and Louis (2003) and Parzen et al. (2011) to use the generalized linear mixed models via a bridge distribution to generate multivariate binary longitudinal data for marginal models. In addition, we conducted a clinical trial simulation study for analyzing multiple and correlated binary outcomes based on control the correlation over the outcomes and occasions, and estimate the effect sample size. This approach could be a good method in simulating the correlated binary data. We include an explanation of some constraints to achieving the best simulation results.

Corresponding Author:- Hissah Alzahrani
Address:- Mathematical Science Department, College Of Applied Sciences, Umm Al-Qura University, Mekkah, 24382, Saudi Arabia.

1 Introduction

In this paper, our interest is generate multivariate binary longitudinal data for marginal models. It is a simulation study for many longitudinal outcomes. The longitudinal data feature is measuring the responses over many occasions. Then, the measurements within each subject are supposed to be correlated. In the multivariate longitudinal data, there are many longitudinal outcomes are obtained in many occasions. The main two factors to build up the within subject correlation in multivariate longitudinal data are outcomes and occasions. Because the multivariate longitudinal data has a complicated correlation structure R, there is not a lot of correlation or covariance patterns are defined for the multivariate longitudinal data. This simulation study will be helpful to build up a correlation pattern for the correlated binary response form the artificial data. Generating the data based on the advantage of controlling the correlation over the outcomes and occasions is the main goal in this simulation study. Then, we can study the changes in the responses means over the time based on controlling the correlation.

Generating correlated binary data under the marginal model requires specification of the marginal means or pairwise correlation in R. Different methods are used based on different structures of R and equal or unequal marginal means. In the case of generate the artificial correlated binary data, Lee (1993) developed a method using Copula to generate correlated binary data, but contains only one parameter for R matrix. Lunn and Davies (1998) and Kang and Jung (2001) improved methods for exchangeable patterns and equal means correlations. Qaqish (2003) introduced the conditional linear family of correlated binary distribution for patterned R under equals and unequal means, or unpatterned R and large sample size. The method of Qaqish (2003) is based on a conditional linear family of multivariate binary distributions. Emrich and Piedmonte (1991) proposed a method based on the multivariate probit model using correlated standard normal variables by solving nonlinear equations. Since our goal is build a desired correlation pattern to adopt the multivariate longitudinal data, the method we use should generate the data for unstructured correlation matrix which means no constraints and the maximum parameters to estimate.

The two most practical and applicable methods for unpatterned R are those described by Emrich and Piedmonte (1991) and Qaqish (2003). Generally, Preisser Jr and Qaqish (2012) compared the two method and showed they have good estimations unless in some patterned structures. In Addition to the two methods of Emrich and Piedmonte (1991) and Qaqish (2003) to generate correlated binary data for unstructured pattern of R, we describe a third method to generate the multivariate binary data using bridge distribution. It is a method considered to be for unstructured R or building up a desired pattern. In section 2, we preview some aspects of the multivariate longitudinal data structure specifically for the binary data. Then, we explained the proposed method and its constraints in sections 3 and section 4. Section 5 is the application to the study design for multivariate binary data. Sections 6 and 7 contain the results and conclusion, respectively.

2 Multivariate Binary Longitudinal Data

The multivariate longitudinal data model is an extension of the univariate longitudinal model, but for more than one outcome. Each individual i has a vector of responses for different outcomes, $k = 1, 2, ..K$. Also, each individual is measured at different times or occasions, $j = 1, 2 \dots J_i$, and has cluster size $n_i = J_i K$. Let us model K vectors of outcomes measured corresponding to a vector of times. Then, the structure is in the following figure:

$$\begin{array}{c} ID \\ \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \\ 2 \\ 2 \\ \vdots \\ \vdots \\ N \\ N \\ \vdots \\ N \end{pmatrix} \end{array} \quad \begin{array}{c} \begin{array}{cccc} Y_{ij1} & Y_{ij2} & \cdots & Y_{ijK} \end{array} \\ \begin{pmatrix} y_{111} & y_{112} & \cdots & y_{11K} \\ y_{121} & y_{122} & \cdots & y_{12K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1J1} & y_{1J2} & \cdots & y_{1JK} \\ y_{211} & y_{212} & \cdots & y_{21K} \\ y_{221} & y_{222} & \cdots & y_{22K} \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ y_{N11} & y_{N12} & \cdots & y_{N1K} \\ y_{N21} & y_{N22} & \cdots & y_{N2K} \\ \vdots & \vdots & \ddots & \vdots \\ y_{NJ1} & y_{NJ2} & \cdots & y_{NJK} \end{pmatrix} \end{array} \quad \begin{array}{c} time \\ \begin{pmatrix} 1 \\ 2 \\ \vdots \\ J \\ 1 \\ 2 \\ \vdots \\ \vdots \\ 1 \\ 2 \\ \vdots \\ J \end{pmatrix} \end{array}$$

Figure 1: Multivariate longitudinal data structure

To simplify these notations, we will refer to J_i as J which is the number of occasions or visit numbers over all the observations, the vector of responses for subject i is :

$$Y_i = [Y_{i11}, Y_{i21}, ...Y_{iJ1}, Y_{i12}, Y_{i22}, Y_{i32}, ..., Y_{iJ2}, ..., Y_{i1K}, Y_{i2K}, ..., Y_{iJK}]^T$$

To illustrate aspects of the multivariate longitudinal data structure, lets assume the simple case where there are two longitudinal outcomes, $k = 1, 2$, are measured over three occasions, $j = 1, 2, 3$.

for observation i, $i = 1, 2, 3...N$. The correlation matrix R consists of the within subject correlation parameters. Then, correlation matrix $R(\gamma)$ is a function of γ, where γ represents a vector of within subject association parameters, $\gamma = [\gamma_1, \gamma_2, ..., \gamma_{15}]^T$.

$$R = \begin{array}{cc} & \begin{array}{cccccc} Y_{11} & Y_{21} & Y_{31} & Y_{12} & Y_{22} & Y_{32} \end{array} \\ \begin{array}{c} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{12} \\ Y_{22} \\ Y_{32} \end{array} & \begin{pmatrix} 1 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 & \gamma_5 \\ - & 1 & \gamma_6 & \gamma_7 & \gamma_8 & \gamma_9 \\ - & - & 1 & \gamma_{10} & \gamma_{11} & \gamma_{12} \\ - & - & - & 1 & \gamma_{13} & \gamma_{14} \\ - & - & - & - & 1 & \gamma_{15} \\ - & - & - & - & - & 1 \end{pmatrix} \end{array}$$

Let γ be a vector of size $\binom{JK}{2}$ of all non-redundant pairwise correlation parameters in R. We will use the idea of modeling the correlation matrix to reduce the length of the vector γ. O'Brien and Fitzmaurice (2004) fit a regression model for marginal pairwise odds ratio to estimate less parameters in the binary multivariate longitudinal data structure correlation for GEE model. We will build a model for pairwise correlation parameters to induce the correlation over the outcomes or the occasions for many scenarios. Consider the correlation matrix R consists of three correlation parameters types:

1- Let $\alpha_{jk,j'k}$ be the inter-outcome correlation parameter which compares the outcome k with the outcome $k^{'}$at time j:

$$\alpha_{jk,j'k} = \frac{P(Y_{jk} = 1, Y_{jk'} = 1) - P(Y_{jk} = 1)P(Y_{jk'} = 1)}{\sqrt{P(Y_{jk} = 1)P(Y_{jk'} = 1)(1 - P(Y_{jk} = 1))(1 - P(Y_{jk'} = 1))}} \tag{1}$$

2- Let $\upsilon_{jk,jk'}$ be the intra-outcome correlation parameter which compares outcome k at time j with the same outcome at time $j^{'}$:

$$\upsilon_{jk,j'k} = \frac{P(Y_{jk} = 1, Y_{j'k} = 1) - P(Y_{jk} = 1)P(Y_{j'k} = 1)}{\sqrt{P(Y_{jk} = 1)P(Y_{j'k} = 1)(1 - P(Y_{jk} = 1))(1 - P(Y_{j'k} = 1))}} \tag{2}$$

3- Let $\tau_{jk,j'k'}$ be the cross correlation parameter which compares the outcome k at time j with outcome $k^{'}$ at time $j^{'}$:

$$\tau_{jk,j'k'} = \frac{P(Y_{jk} = 1, Y_{j'k'} = 1) - P(Y_{jk} = 1)P(Y_{j'k'} = 1)}{\sqrt{P(Y_{jk} = 1)P(Y_{j'k'} = 1)(1 - P(Y_{jk} = 1))(1 - P(Y_{j'k'} = 1))}} \tag{3}$$

$$R = \begin{array}{c|cccccc|} & Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \\ Y_{11} & 1 & \upsilon_1 & \upsilon_2 & \alpha_1 & \tau_1 & \tau_2 \\ Y_{12} & - & 1 & \upsilon_3 & \tau_3 & \alpha_2 & \tau_4 \\ Y_{13} & - & - & 1 & \tau_5 & \tau_6 & \alpha_3 \\ Y_{21} & - & - & - & 1 & \upsilon_4 & \upsilon_5 \\ Y_{22} & - & - & - & - & 1 & \upsilon_6 \\ Y_{23} & - & - & - & - & - & 1 \end{array}$$

Then, correlation matrix R is for the response Y_{jk}, $k = 1, 2$ (outcomes), $J = 1, 2, 3$ (occasions). Here we can specify any pattern or build up a parsimonious model to reduce the number of estimated parameter less than $\binom{JK}{2} = 15$ in R. For example, we could assume the time correlation parameters $\alpha^{'}s$ to be exponential and assume the outcomes correlation parameters $\upsilon^{'}s$ to be compound symmetry . To simplify the parameter estimations size, we conducted the simulation under the assumption of exchangeability for each correlation type υ, α and, τ and $\tau = 0$ for five correlation scenarios. The following table shows the values for each correlation structure and scenario in R matrix:

Table 1: The scenarios of simulation study

	$\alpha = 0.00$	$\alpha = 0.60$	$\alpha = 0.90$
$\upsilon = 0.00$	scenario1	scenario2	scenario3
$\upsilon = 0.60$	scenario4	-	-
$\upsilon = 0.90$	scenario5	-	-

3 The Simulation Method

3.1 Generating correlated binary data using bridge distribution

The goal in this study is to generate correlated binary data for marginal model. We used a regular generalized liner mixed model using bridge distribution for the random effects term. It is known that the parameter estimations under the mixed model have different interpretation than the marginal model because the marginal model integration over the random effect do not keep the logistic form. Using bridge distribution, matched the logistic shape of the conditional and marginal binary response models. The first contribution to use bridge distribution for the random intercept logistic regression model is proposed by Wang and Louis (2003). We will start by the univariate longitudinal data structure. Let Y_{ij} be the binary response that measured at time j, $j = 1, 2, 3, ...J$ form independent observations $i, i = 1, 2, 3...N$. For each individual has a $C \times 1$ vector of covariate X_{ij}. Suppose the marginal distribution of the responses is Bernoulli with mean $E(Y_{ij}) = P(Y_{ij} = 1|X_{ij}, \beta_p) = \mu_{ij}$ through a logit link function, then responses model is:

$$logit(\mu_{ij}) = \beta_p^T X_{ij} \tag{4}$$

where $\beta_p = (\beta_0, \beta_1, \beta_2, ..., \beta_C)^T$ are the regression parameters from the marginal model. Wang and Louis (2003) used a bridge distribution for the random effect in the following mixed effects logistic model:

$$logit(\mu_{ij}|b_i, X_{ij}) = b_i + \phi\ \beta_s^T X_{ij} \tag{5}$$

where ϕ is the a cluster heterogeneity parameter. The relationship between the effects from the marginal regression model β_p and the mixed regression model β_s is related by ϕ such that:

$$\beta_p = \beta_s * \phi$$

Here will give a brief description of the method of Wang and Louis (2003). They introduced a CDF of bridge distribution $G(b)$ for the random effect to gain its advantage of keeping the marginal shape same as the conditional shape such as:

$$\int H(b + \beta_s^T X) dG(b) = H(r + \phi \beta_s^T X) \tag{6}$$

where H is a CDF of bridge distribution and ϕ is rescaling parameter between 0 and 1. The parameters β_s and r are unknown parameters and r is 0 when H is a CDF of symmetric distribution. The parameter β_s is regression effect and X is the covariates. By differentiate both sides of equation (6) respect to $\beta_s^T X$ and taking Fourier transformation F, then after organizing and using the Fourier Inversion theorem, they got the following equation:

$$g_\phi(x) = \frac{1}{2\pi} \int e^{i(\frac{r}{\phi - x})\xi} \frac{Fh(\frac{\xi}{\phi})}{Fh(\xi)} d\xi \tag{7}$$

If the function $H(.) = logit$ link function then $H(\beta_s^T X) = \frac{e^{\beta_s^T X}}{1 + e^{\beta_s^T X}}$. By plugging in Fourier transformation, see Wang and Louis (2003), then they got the pdf of bridge distribution:

$$g_\phi(x) = \frac{1}{2\pi} \frac{sin(\phi\pi)}{cosh(\phi x) + cos(\phi\pi)} \qquad (0 < \phi < 1, -\infty < x < \infty) \tag{8}$$

where $cosh(x) = \frac{e^x + e^{-x}}{2}$. The bridge distribution is symmetric and has slightly heavier tail than the normal distribution and lighter than logistic distribution with mean 0 and variance $\sigma_b^2 = \pi^2(\frac{1}{\phi^2} - 1/3)$. In addition, $\beta_p = \beta_s(1 - \rho_Y)$ where $\rho_Y = corr(Y_{ij}, Y_{ij'})$ is the intracluster correlation in the binary response. ϕ measure the heterogeneity across the clusters between $[0, 1]$. The CDF of the bridge distribution is :

$$G_\phi(x) = 1 - \frac{1}{\pi\phi}\,[\frac{\pi}{2} - arctan\frac{e^{\phi x} + cos(\pi\phi)}{sin(\pi\phi)}] \tag{9}$$

and the inverse of the cumulative density function is:

$$G_\phi^{-1}(x) = \frac{1}{\phi}\, log\frac{sin(\pi\phi x)}{sin(\pi\phi(1 - x))} \tag{10}$$

Using the transformation $\tilde{b} = \Phi^{-1}(G(x))$, where Φ is CDF of standard Gaussian distention, then $\tilde{b} \sim N(0, 1)$. That leads to using the Gaussian-Hermite quadrature method to evaluate the integral over the bridge random effects and estimate MLE parameters. In addition, Parzen et al. (2011) have improved this model of bridge distribution for the random effect in the logistic model. Their contribution has two primary advantages. First, they constructed a model for distinct and correlated random bridge intercepts b_{ij} at each time point. The response Y_{ij} given bridge random intercepts follows the Bernoulli distribution $P(Y_{ij} = 1|b_{ij}, X_{ij}, \beta_s)$ instead of $P(Y_{ij} = 1|b_i, X_{ij}, \beta_s)$. Their method leads to better association modeling for within each subject correlation. Then, they use Copula to model the multivariate bridge random variables. Second, Parzen et al. (2011) recommend using Pearson correlation in terms of Kendall's τ to present the association between the Z's random variables due its advantage of invariance of the monotone transformation.

We exploit the advantage of the flexibility in the association structure in Parzen et al.

(2011)'s method between the bridge random effects. It is a beneficial method to generate multivariate longitudinal data, controlling the within subject correlation over the outcomes and occasions using marginal model. We will generate it from mixed model using multivariate bridge random effects. First, Let Y_{ijk} be the binary response that measured at time j, $j = 1, 2, 3$ for observation i, $i = 1, 2, 3...N$ and for outcome $k = 1, 2$:

$$logit(E(Y_{ijk}|X_{ij}, b_{ijk})) = logit(P(Y_{ijk} = 1|b_{ijk}, X_{ij})) = \beta_{0k} + \beta_{1k}X_{ij} + b_{ijk} \qquad (11)$$

where b_{ijk} is for distinct and correlated random bridge intercepts for each outcome $k = 1, 2$ at each occasion or time $j = 1, 2, 3$. Given the vector of the random effect b_{ijk}, the Y_{ijk} for subject i is assumed to be independent Bernoulli random variables, $Y_{ijk}|b_{ijk} \sim Ber(P(Y_{ijk} = 1))$. The marginal model will be:

$$logit(E(Y_{ijk}|X_{ij})) = 1/\phi_{jk}\ (\beta_{0k} +\ \beta_{1k}X_{ij}) \qquad (12)$$

where the parameter $0 < \phi_{jk} < 1$ is assumed to be toward zero to ensure the maximum heterogeneity of the random effect (clusters) for the response at time j for outcome k. For some reasons will be explained in the next section, we referred to ϕ_{jk} as ϕ which means all the bridge parameter have the same value. The contribution of subject i to the likelihood function is given by:

$$L_i = \int_{b_i} [\prod_{k=1}^{2}\prod_{j=1}^{3} P(Y_{ijk} = y_{ijk}|b_i, X_{ij})\ ] f_b(b_i)db_i, \qquad (13)$$

where $f_b(b_i)$ is the joint density of $(b_{i11}, b_{i21}, b_{i31}, b_{i12}, b_{i22}, b_{i32})$. To simplify the notations, we will refer to the joint bridge random intercepts as $(b_{i1}, b_{i2}, b_{i3}, b_{i4}, b_{i5}, b_{i6})$. The likelihood function will be $\prod_{i=1}^{N} L_i$. The multivariate density of bridge random variables can be modeled using Copula model, a multivariate joint cumulative distribution function used to joint univariate marginal distribution when the inverse cumulative of each variable is uniform distribution on the interval [0,1], Sklar (1959). Here we use the Gaussian Copula to joint bridge random variables.

If $F_1(b_1), F_2(b_2), F_3(b_3), F_4(b_4), F_5(b_5), F_6(b_6)$ are the cumulative distribution functions for the random effect variables $(b_{i1}, b_{i2}, b_{i3}, b_{i4}, b_{i5}, b_{i6})$, then there exist a function C such that the joint CDF is:

$$C(u_1, u_2, u_3, u_3, u_4, u_5, u_6) = P(U_1 \leq u_1, U_2 \leq u_1, U_3 \leq u_3, U_4 \leq u_4, U_5 \leq u_5, U_6 \leq u_6)$$

where $U_1, U_2, ...U_6$ variables are $F_1(b_1), F_2(b_2)...., F_6(b_6)$ has uniformly distributed CDF's and C is the density of Gaussian Copula is given by:

$$C(u_1, u_2, u_3, u_3, u_4, u_5, u_6) = \Phi_{Z_1, Z_2, ...Z_6, \Sigma}(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3), \Phi^{-1}(u_4), \Phi^{-1}(u_5), \Phi^{-1}(u_6 \tag{14}$$

where $\Phi_{Z_1, Z_2, ...Z_6, \Sigma}$ is the CDF of a multivariate normal distribution with mean zero vector and variance covariane matrix is Σ. Then, the bridge variable can be obtained by $b_r = G^{-1}(\Phi(Z_r))$, where $r = 1, 2, ...6$ and $\Phi(.)$ is the CDF of univariate standard normal and $G^{-1}(.)$ is inverse cumulative distribution of marginal bridge distribution. To specify the correlation matrix Σ, we need to specify the Pearson correlation $\rho_{ish} = Corr(Z_{is}, Z_{ih})$ for each pair of Z's random variables. Parzen et al. (2011) recommend using Pearson correlation in terms of Kendall's τ to present the association between the the Z's random variables due its advantage of invariance of the monotone transformation, as discussed in Hougaard (2000).

$$\rho_{ish} = sin(\pi \tau_{ish}/2)$$

Then, inducing the correlation in Copula random variables using Kendall's τ, will be produced in bridge random variables because the bridge random variable are monotone transformation of Z's random variables. The maximum likelihood estimates of the parameter can be obtained by maximizing the likelihood function using Copula method. Because the method does not have a closed form, maximum likelihood estimates can be implemented using numerical approximations.

3.2 Natural constraints

Most of the simulation methods for the binary data have some constraints related to response means or the correlation structure. Using bridge distribution for the random effect is also has some constraints. To explore the limitations, we assumed just two bridge random effects to generate correlated binary data for the GEE model:

$$logit(P(Y_{ik} = 1|b_i, \beta)) = \beta_{0k} + \beta_{1k}\, X_k + b_{ik} \tag{15}$$

where the parameters $\beta_{01} = 1$, $\beta_{02} = 1$, $\beta_{11} = 1$, $\beta_{12} = 1$, and $b_i = (b_{i1}, b_{i2})$ are distinct and correlated random bridge intercepts and $X = 1, 2$. Under the bridge distributional assumption, the rescalling parameter ϕ is the connection between the regression parameters in the marginal and the conditional logistic model such that:

$$\beta_p = \beta_s * \phi$$

where ϕ is the parameter that measures the heterogeneity between the clusters. Also, ϕ is related to the variance of bridge random effect, $\sigma_b^2 = \pi^2(\frac{1}{\phi^2} - 1/3)$. As we see in figure 2, the variability of bridge random effect convergences to zero when ϕ is larger for both the theatrical and empirical relationship. In the context of generating artificial binary responses, we are looking to assume the best ϕ value that leads to better estimations.

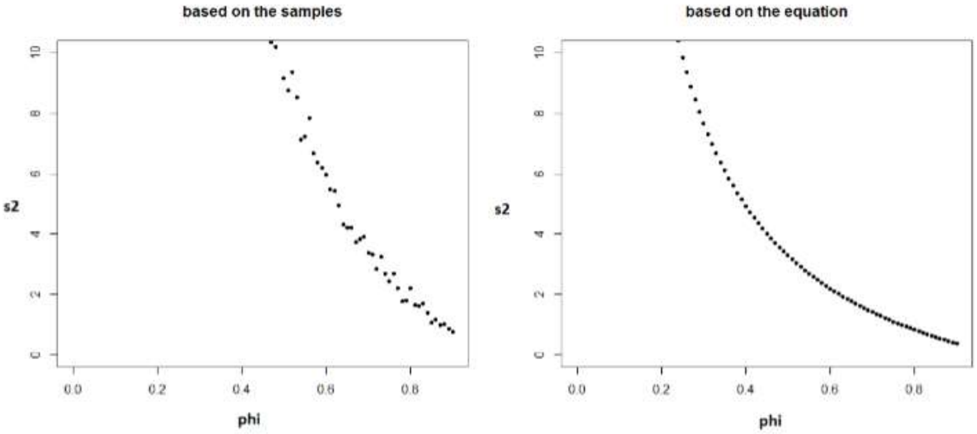

Figure 2: The relationship between the bridge parameter and its variance

In order to reach good estimation, it is important to investigate the connection between the correlation of the bridge random effects and the correlation of the binary responses. The efficiency of this method is based on the good induction of the desired correlation from bridge random effects into the binary responses $corr(Y_{i1}, Y_{i2}) \approx corr(b_{i1}, b_{i2})$. In figure 3, we explore the relationship between the Kendall's tau of the bridge random effect $corr(b_{i1}, b_{i2})$ and the Kendall's tau for the binary responses $corr(Y_{i1}, Y_{i2})$ for the range of correlation [-0.9,0.9] and for sample size=1000. The relationship is implemented for different values of ϕ, $\phi = 0.05, 0.30, 0.60, 0.80$. The best relationship is considered close to 45° degree line between the associations of bridge random effect and the binary responses. Thus, we recommend assuming $\phi = 0.05$ to induce the desired correlation from bridge random effect into the binary responses.

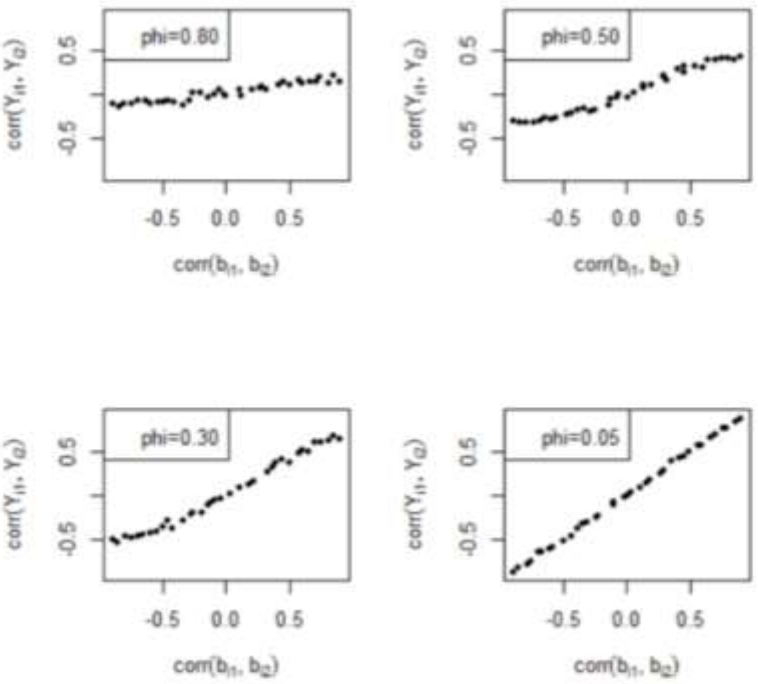

Figure 3: The relationship between the associations of (Y_{i1}, Y_{i2}) and the associations of (b_{i1}, b_{i2}) for different values of bridge parameter

Secondly, the restriction on the estimation parameter β_s is imposed by ϕ since $\beta_p = \beta_s * \phi$. Wang and Louis (2003) said that the marginal parameter shrink toward zero when the heterogeneity is larger. Consequently, it is better to assume smaller values of β when ϕ is smaller. The last constraint is related to using Copula method of estimation. The R correlation matrix should be positive definite to generate Copula random variables.

4 Application to Simulation Design

We conducted a simulation study for the five correlation scenarios in order to explore the properties of using the proposed method. Specifying bridge distribution for the random effects is to generate multivariate longitudinal binary data. One of the goals of designing the simulation study was to determine the efficient sample size for specified model that leads to statistically- significant result in the treatment effect between the outcomes. A larger sample size certainly leads to more accurate parameter estimations, but would raise the research budget. Further, in clinical trials, it would require more human subjects who would be exposed to new treatments that may be harmful. In this section, we conducted a simulation study to estimate the efficient sample size in clinical trails needed to detect statistically significant results for the treatment using the proposed method.

Let $X_i = 0, 1$ is the treatment covariate and $t_j = 1, 2, 3$ is time covariate for three occasions. Let Y_{ijk} be the binary response that measured at time j, $j = 1, 2, 3$ for observation i, $i = 1, 2, 3...N$ and for outcome $k = 1, 2$. Then, the true logistic model be

$$logit(E(Y_{ijk}|X_i, b_i)) = logit(P_{ijk}) = \beta_{0k} + \beta_{1k}X_i + \beta_{2k}t_j + b_{ijk} \tag{16}$$

where $b_i = (b_{i1}, b_{i2}, b_{i3}, b_{i4}, b_{i5}, b_{i6})$ are distinct but correlated random bridge intercept for each outcome at each occasion or time. Given the vector of the random effect b_i, the Y_{ijk} for subject i are assumed to be independent Bernoulli random variables, $Y_{ijk}|b_i \sim Ber(P(Y_{ijk} = 1))$. The marginal model will be:

$$logit(E(Y_{ijk}|X_i)) = 1/\phi\ (\beta_{0k} + \beta_{1k}X_i + \beta_{2k}t_j) \tag{17}$$

where the parameters $\phi = 0.05,\ \beta_{01} = 4,\ \beta_{02} = 2,\ \beta_{11} = 1,\ \beta_{12} = -3,\ \beta_{21} = 1,\ \beta_{22} = -5$.

For the random effect model, we conducted the simulation study based on the GEE model of Shelton et al. (2004) to separate the estimated effects for each outcomes using Kronecker product. The simulation for the five correlation scenarios to explore the properties of the model when the correlation is induced over the outcomes and the occasions. Additional goal of this study was to estimate the effect sample size to reject the null hypothesis of the treatment group for the two outcomes, $H_0 : \beta_{11} = \beta_{12} = 0$.

5 The Results

We ran a simulation method for N=200 samples. The clinical trial is balance for each arm. For 200 samples, we computed the correlation mean for each correlation parameter and the results are in figure 4. The correlation means are calculated for the five scenarios. Starting from scenario 1, it can be seen as a reference scenario since we assume there is no correlation over occasions or outcomes. It seems it has good convergence close to zero of all its parameters. The correlation scenarios is 2 and 3 are supposed to be induced over the outcomes' parameters α and scenario 4 and 5 over the occasions' parameters υ. From figure 4, the estimated correlation matrix in scenarios 4 and 5 has a good results. Also, it is clear that the parameter estimation of the correlation matrix for scenarios 2 and 3 have increased bias over the time, meaning that the correlation between the outcomes in time 1 is better than time 3. Adding the time dependent covariate in the longitudinal response model may effects the correlation for the binary responses.

$$R = \begin{array}{c} \\ Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \\ \end{array} \begin{pmatrix} 1 & \upsilon & \upsilon & \alpha & \tau & \tau \\ - & 1 & \upsilon & \tau & \alpha & \tau \\ - & - & 1 & \tau & \tau & \alpha \\ - & - & - & 1 & \upsilon & \upsilon \\ - & - & - & - & 1 & \upsilon \\ - & - & - & - & - & 1 \end{pmatrix}$$

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.000 & 0.000 & 0.000 & -.002 & -.005 \\ - & 1 & -.003 & 0.001 & 0.006 & -.001 \\ - & - & 1 & -.001 & -.006 & -.005 \\ - & - & - & 1 & 0.000 & 0.002 \\ - & - & - & - & 1 & 0.002 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 1

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.003 & 0.000 & 0.571 & -.001 & 0.006 \\ - & 1 & 0.001 & 0.004 & 0.535 & 0.001 \\ - & - & 1 & 0.002 & -.009 & 0.497 \\ - & - & - & 1 & 0.003 & 0.008 \\ - & - & - & - & 1 & -.001 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 2

$$\begin{array}{c} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{21} \\ Y_{22} \\ Y_{23} \end{array} \begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array} \begin{pmatrix} 1 & 0.000 & -.002 & 0.770 & -.004 & -.002 \\ - & 1 & 0.000 & -.001 & 0.668 & 0.000 \\ - & - & 1 & 0.003 & 0.000 & 0.574 \\ - & - & - & 1 & -.001 & 0.003 \\ - & - & - & - & 1 & 0.001 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 3

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.593 & 0.596 & 0.000 & -.000 & -.002 \\ - & 1 & 0.590 & 0.000 & 0.001 & -.001 \\ - & - & 1 & 0.002 & 0.001 & -.001 \\ - & - & - & 1 & 0.580 & 0.568 \\ - & - & - & - & 1 & 0.583 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 4

$$\begin{array}{cccccc} Y_{11} & Y_{12} & Y_{13} & Y_{21} & Y_{22} & Y_{23} \end{array}$$
$$\begin{pmatrix} 1 & 0.881 & 0.875 & -.000 & 0.000 & 0.000 \\ - & 1 & 0.881 & 0.000 & 0.000 & -.000 \\ - & - & 1 & 0.000 & -.000 & -.001 \\ - & - & - & 1 & 0.849 & 0.769 \\ - & - & - & - & 1 & 0.846 \\ - & - & - & - & - & 1 \end{pmatrix}$$

scenario 5

Figure 4: The estimated correlation structures using the proposed method

The second output is for the parameter estimations. Assuming the log odds the responses changes curvilinerly with the time and X, we got the means of estimated regression coefficients over 200 samples in table 2. Starting from scenario 1, we found the estimated effect of intercept means for the outcomes 1 and 2 respectively are 0.206, 0.089 over 200 samples and each sample size is n=400. The parameter estimations $\beta_{01}, \beta_{11}, \beta_{21}$ are the log odds of $P(Y_{i1} = 1)$ for intercepts, treatment and time covariates respectively and $\beta_{02}, \beta_{12}, \beta_{22}$ are for the second outcome. Generally for all the scenarios, the parameter estimations are approximately close to the true values unless in scenario 4 and 5. The standard deviation std in scenario 4 and 5 for treatment effect is large comparing with the other scenarios. That leads to conclude the strong correlation in the time factor may affects the bias of the regression coefficients especially in scenarios 4 and 5.

Table 2: The covariate parameter estimations

	senario 1		senario2		senario3		senario4		senario5	
True β	mean	std	mean	std	mean	std	mean	std	mean	std
$\beta_{01} = 0.2$	0.206	0.168	0.205	0.166	0.210	0.162	0.213	0.159	0.215	0.148
$\beta_{02} = 0.1$	0.089	0.178	0.120	0.166	0.114	0.156	0.1068	0.083	0.079	0.162
$\beta_{11} = 0.05$	0.059	0.112	0.046	0.117	0.063	0.111	0.051	0.171	0.035	0.221
$\beta_{12} = -0.15$	-0.152	0.117	-0.163	0.130	-0.138	0.114	-0.130	0.173	-0.121	0.214
$\beta_{21} = 0.05$	0.046	0.077	0.048	0.073	0.043	0.070	0.046	0.044	0.051	0.027
$\beta_{22} = -0.25$	-0.245	0.074	-0.259	0.072	-0.258	0.063	-0.249	0.045	-0.249	0.038

One of the goals of this case study was to estimate the best sample size for to detect significant treatment effect in outcome 1 or outcome 2. We applied the proposed method in the clinical trial model for range of sample sizes $n = (400, 800, 1200, 1600, 2000, 2400, 2800, 5000, 8000$ and counted how many times the null hypothesis $H_0 : \beta_{11} = \beta_{12} = 0$ is rejected for each sample size versus at a least one of the parameter estimation is not zero. To get the study power, we estimated the power as a function of the sample size in order to estimate the best sample size leads to get 0.80 power value. In graph 5, we present the effect sample size for each scenario. In scenario 1, the best sample size for two arms is n=2200 that means approximately 1100 for each study arm. It is clear the effect sizes for scenarios 2 and 3, which expressed the correlation between the outcomes in each occasion, are lower than other scenarios. This happens may because this model is designed to separate the effects of the parameter for each outcome, then the correlation over occasions required higher sample size.

The highest required sample sizes are approximately n=8850, 4880 for scenarios 4 and 5, respectively, when the correlation is induced over the occasions for each outcome.

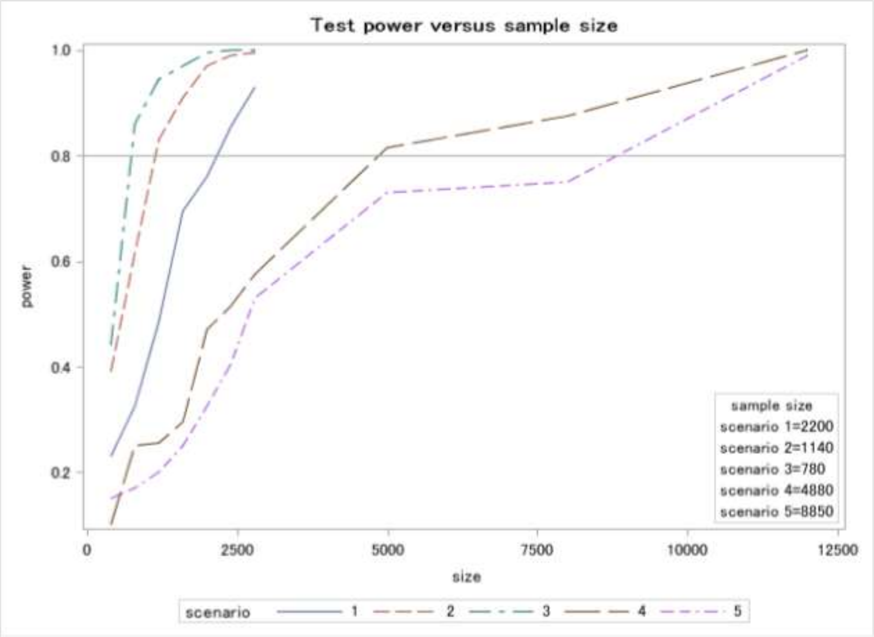

Figure 5: The sample size estimations over the correlation scenarios

In conclusion, the multivariate longitudinal data potentially has complicated correlation. The correlation among the responses comes from the repeated measurements and the outcomes that are measured from the same observation. The estimations of each correlation pattern in multivariate structure for the five scenarios is computed. Scenarios 4 and 5 present the induced correlation over the occasions while scenarios 2 and 3 present the induced correlations between the outcomes. We saw in scenarios 2 and 3 the bias increases gradually over the time and maybe this due to the existence of the dependent coviarate in the model. Also, the parameter estimations over the five scenarios did not changed dramatically when we changed the source and the strength of the correlation over the scenarios unless in the treatment effects. It is clear the strong correlation produce more bias estimates. In fact, The effect sample size for the study model is also effected by the scenario. Based on our model, clinical trials require higher sample sizes for high correlations over the occasions, as we saw in scenarios 4 and 5.

6 Conclusion

Researchers have discussed variety methods to address problems related to generating correlated binary data for marginal models. In this paper, we describe a simple computed method using a linear mixed model via bridge distribution for the random effect. Using bridge distribution, it has the advantage to keep the same logistic shape for the marginal and conditional models. This method could reach good convergence for a desired R correlation matrix. It could be a good future work to study a comparison between the proposed method and Emrich and Piedmonte (1991) and Qaqish (2003) methods. Choosing the appropriate bridge parameter and parameter estimation of the marginal model would effect the results convergence of using bridge distribution. In conclusion, generate the binary responses for the marginal model using bridge distribution for the random effect could be good approach.

References

Emrich, L. J. and Piedmonte, M. R. (1991). A method for generating high-dimensional multivariate binary variates. *The American Statistician*, 45(4):302–304.

Hougaard, P. (2000). *Analysis of multivariate survival data.* Springer Science & Business Media.

Kang, S.-H. and Jung, S.-H. (2001). Generating correlated binary variables with complete specification of the joint distribution. *Biometrical Journal*, 43(3):263–269.

Lee, A. (1993). Generating random binary deviates having fixed marginal distributions and specified degrees of association. *The American Statistician*, 47(3):209–215.

Lunn, A. D. and Davies, S. J. (1998). A note on generating correlated binary variables. *Biometrika*, 85(2):487 490.

O'Brien, L. M. and Fitzmaurice, G. M. (2004). Analysis of longitudinal multiple-source binary data using generalized estimating equations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):177–193.

Parzen, M., Ghosh, S., Lipsitz, S., Sinha, D., Fitzmaurice, G. M., Mallick, B. K., and Ibrahim, J. G. (2011). A generalized linear mixed model for longitudinal binary data with a marginal logit link function. *The annals of applied statistics*, 5(1):449.

Preisser Jr, J. S. and Qaqish, B. F. (2012). A comparison of methods for generating correlated binary variates with specified marginal means and correlations.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455 463.

Shelton, B. J., Gilbert, G. H., Liu, B., and Fisher, M. (2004). A sas macro for the analysis of multivariate longitudinal binary outcomes. *Computer Methods and Programs in Biomedicine*, 76(2):163–175.

Sklar, M. (1959). *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8.

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):765 775.