Bull. Malays. Math. Sci. Soc. (2022) 45 (Suppl 1):S251–S270
https://doi.org/10.1007/s40840-022-01290-4

BULLETIN *of the*
MALAYSIAN MATHEMATICAL
SCIENCES SOCIETY
http://www.springer.com/mathematics/journal/40840

ORIGINAL ARTICLE

Modelling Correlated Bivariate Binary Data: A Comparative View

Jahida Gulshan[1] · Azmeri Khan[1] · M Ataharul Islam[1]

Received: 11 January 2022 / Revised: 7 April 2022 / Accepted: 8 April 2022 /
Published online: 13 May 2022

Abstract

This study focused on comparing selected commonly used marginal models with marginal-conditional models for analyzing correlated longitudinal binary data. A simulation study shows that for explaining the relationship among the covariates and the repeated outcomes, each of the proposed models show competitive results in terms of bias and coverage probability as compared to the marginal models. If the repeated outcomes are associated or if the distribution of outcome variables are not identical at different follow-ups, the marginal-conditional models give better results in terms of bias and coverage probability of the estimates. For keeping the number of parameters to be estimated as small as possible, the regressive model is suggested for data with more than three follow-ups. The methods are illustrated with an example using Health and Retirement Study data.

Keywords Bivariate binary outcomes · Marginal model · Conditional model · Marginal conditional model

Mathematics Subject Classification 62-08 · 62H99

Communicated by Rafiqul I. Chowdhury.

✉ Jahida Gulshan
gulshan@isrt.ac.bd

Azmeri Khan
azmeri@isrt.ac.bd

M Ataharul Islam
mataharul@yahoo.com

[1] Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh

1 Introduction

Analysis of repeated measures categorical data has drawn the interest of many researchers in the last few decades and has become an important and active area of research [7, 9, 14]. Most of the previous works on correlated outcome variables were based on the marginal response probabilities. Generalized estimating equations (GEE) is a popular and one of the most widely used methods for analyzing longitudinal data which is a quasi-likelihood approach that uses a population averaged model [20, 32]. GEE do not require to meet the classical assumptions of independence and normality, which are too restrictive for many problems [26]. Carey et al. [5] introduced alternating logistic regression (ALR) models based on marginal odds ratios instead of correlations between pairs of binary responses combining the first-order GEE for regression coefficients with a new logistic regression equation for estimating the correlation parameter. Due to lack of proper specification of the underlying model, marginal models such as GEE or ALR may fail to provide the measure of dependence of binary outcomes. The induced correlations considered in these methods and anomalies caused by the induced correlation between repeated outcomes is beyond any explanation. Working correlation structure of GEE has been a concern of many studies mainly focusing on examining the existing selection criteria and/or proposing new selection criteria for correlations structures [10, 12, 23, 24, 27, 30]. Many studies, see, for example, Darlington and Farewell, Guerra et al. [6, 11], tried to address this problem by modifying the approaches based on marginal models using Markov-based transition probabilities.

A good number of researchers, for example, Muenz and Rubinstein [22], Zeger et al. [33] and Azzalini [1], explored Markov models for binary longitudinal data. Islam and Chowdhury [18], Islam et al. [13, 16, 17] carried out a series of research works using Markov-based conditional models and joint models based on marginal conditional approaches for repeated binary data. The conditional regressive logistic models of Bonney [3, 4] were generalized by Islam et al. [13] to include both binary outcomes in previous times and the covariates in the conditional models.

A longitudinal data offers the advantage of visualizing the change in the individual responses with respect to time. GEE- or GEE-based models, being constructed to describe the population averaged or marginal distribution of repeated measurements, may sometimes be appropriate for descriptive observational studies but should be used carefully in causal experiments [21]. Moreover, GEE or other marginal models may not provide the measure of dependence of binary outcomes due to lack of proper specification of the underlying model. The conditional models alone are, also, not adequate to model the longitudinal data. In this study, we proposed two joint models using marginal-conditional approaches for longitudinal data. These models can be used as alternatives to GEE-based models for longitudinal data where marginal models are not appropriate. Starting with an extension of the Markov-based model proposed by Darlington and Farewell [6], we proposed, consequently, a more generalized form that takes into account the correlation structure in an appropriate manner. Finally, we proposed the use of a regressive model-based joint model in case of more than three repeated outcomes in a longitudinal data. The proposed joint models and their inference procedures are simple. Nevertheless, the proposed models take care of the covariate

dependence of the conditional probabilities (of occurrence of events) in second or later follow-ups given the earlier responses of the same subject. One can use of the proposed models for any number of follow-ups, equal or unequal, without making the underlying model complex. Furthermore, the estimation and test procedures for both the specific parameters of interest and the overall model is easy and simple for practical uses on any longitudinal data. Through a simulation study, we compared the proposed two joint models (based on a marginal conditional approach) with GEE and ALR based on marginal models. Finally, we illustrated the selected methods using Health and Retirement Study data [29].

2 Models for Analyzing Repeated Binary Data

Let Y_{ij} be a Bernoulli outcome variable for subject i at jth occasion, $i = 1, 2, ..., N$ and $j = 1, 2, ..., n_i$. Then the outcome vector for subject i can be defined as $\boldsymbol{Y}_i = \left(Y_{i1}\ Y_{i2}\ ...\ Y_{in_i}\right)'$ with mean vector $\boldsymbol{\mu}_i = E(\boldsymbol{Y}_i) = \left(\mu_{i1}\ \mu_{i2}\ ...\ \mu_{in_i}\right)' = \left(p_{i1}\ p_{i2}\ ...\ p_{in_i}\right)'$. Also let $\boldsymbol{X}_{ij}$ be the $1 \times (p+1)$ vector of covariates for subject i at jth occasion.

Let us consider the simplest case of two repeated outcomes on each individual. The vector of responses can be defined as $\boldsymbol{Y}_i = (Y_{i1}\ Y_{i2})$. For binary outcome variables Y_{i1} and Y_{i2} of ith individual, the marginal probability of Y_{ij} observing an event can be expressed as

$$\begin{aligned} p_{ij} &= \Pr(Y_{ij} = 1|\boldsymbol{x}_{ij}) \\ &= \frac{e^{\boldsymbol{x}_{ij}\boldsymbol{\beta}_j}}{1 + e^{\boldsymbol{x}_{ij}\boldsymbol{\beta}_j}}\ i = 1, 2, \ldots, N;\ j = 1, 2, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_j = (\beta_{j0}, ..., \beta_{jp})'$ is a $(p+1) \times 1$ vector of parameters of the marginal model of Y_{ij}. Consequently, the marginal probability of not observing an event can be expressed as $1 - p_{ij} = 1 - \Pr(Y_{ij} = 1|\boldsymbol{x}_{ij}) = \frac{1}{1+e^{\boldsymbol{x}_{ij}\boldsymbol{\beta}_j}}$.

If Y_{i2} depends on Y_{i1}, then for each possible values of y_{i1}, we get one conditional model for Y_{i2}. As we assumed Y_{ij} to be binary random variables, Y_{i1} can take values 0 and 1. When $Y_{i1} = 0$, the conditional probability of $Y_{i2} = 1$ can be defined as

$$\begin{aligned} p^*_{i2} &= \Pr(Y_{i2} = 1|y_{i1} = 0, \boldsymbol{x}_{i2}) \\ &= \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{01}}}{1 + e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{01}}};\ i = 1, 2, \ldots, N, \end{aligned} \tag{2}$$

here $\boldsymbol{\beta}_{01}$ is the vector of parameters of the conditional model of $P(Y_{i2} = 1|Y_{i1} = 0, X_{i2} = x_{i2})$; $i = 1, 2, \ldots N$. Here, the suffix (01) of $\boldsymbol{\beta}$ is used to show the transition from $Y_{i1} = 0$ to $Y_{i2} = 1$.

Similarly, when $Y_{i1} = 1$, the conditional probability of $Y_{i2} = 1$ can be defined as

$$p^*_{i2} = \Pr(Y_{i2} = 1|y_{i1} = 1, \boldsymbol{x}_{i2})$$

$$= \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{11}}}{1+e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{11}}};\ i = 1, 2, \ldots, N, \tag{3}$$

where $\boldsymbol{\beta}_{11}$ is the vector of parameters of the conditional model of $P(Y_{i2} = 1|Y_{i1} = 1, X_{i2} = x_{i2}); i = 1, 2, \ldots N$. The suffix (11) of $\boldsymbol{\beta}$ is used to show the transition from $Y_{i1} = 1$ to $Y_{i2} = 1$.

For $i = 1, 2, \ldots, N$, the joint probabilities can be expressed as the product of marginal and conditional probabilities,

$$P(Y_{i1}, Y_{i2}) = P(Y_{i2} = 1|Y_{i1} = y_{i1}, \boldsymbol{x}_{i2})P(Y_{i1} = y_{i1}|\boldsymbol{x}_{i1}). \tag{4}$$

The repeated measures data are naturally correlated and the major challenge of the methods for analyzing repeated measures categorical data is to model the probable correlations among the repeated observations on the same subject.

2.1 Marginal Models

Following the quasi-likelihood approach [31], with a mean model, μ_{ij}, and variance structure, V_{ij}, the GEE [20, 32] for $\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ denotes the parameters of the marginal model, can be expressed as:

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{N} U_i(\boldsymbol{\beta}) = \sum_{i=1}^{N} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i) = 0, \tag{5}$$

where $\boldsymbol{D}_i = \frac{\delta \boldsymbol{\mu}_i}{\delta \boldsymbol{\beta}}$ and $\boldsymbol{V}_i$ is a working or approximate covariance matrix of $\boldsymbol{Y}_i$ that allows the time dependence to be specified in different ways. The GEE approach uses an induced correlation matrix to define the correlation among the repeated responses. The commonly used correlation structures are independence, autoregressive, exchangeable or unstructured correlation.

The alternating logistic regression (ALR) procedure proposed by Carey et al. [5] combines the first-order GEE for $\boldsymbol{\beta}$ with new logistic regression equations for estimating correlation parameter. ALR regress the response on explanatory variables and model the association among responses in terms of pairwise odds ratio simultaneously. The ALR estimate of $(\alpha, \boldsymbol{\beta})$, where α is the pairwise log odds ratio and $\boldsymbol{\beta}$ is the regression coefficient, is the simultaneous solution of the following unbiased estimating equations:

$$U_{\boldsymbol{\beta}} = \textstyle\sum_{i=1}^{N} \left(\frac{\delta \boldsymbol{\mu}_i}{\delta \boldsymbol{\beta}}\right)' \boldsymbol{V}_i^{-1}(\boldsymbol{Y_i} - \boldsymbol{\mu}_i) = 0, \tag{6}$$

$$U_{\alpha} = \textstyle\sum_{i=1}^{N} \left(\frac{\delta \boldsymbol{\zeta}_i}{\delta \alpha}\right)' \boldsymbol{S}_i^{-1}(\boldsymbol{Y_i} - \boldsymbol{\zeta}_i) = 0, \tag{7}$$

where $\boldsymbol{\zeta}_{ijk} = E(Y_{ij}|Y_{ik} = y_{ik})$ and S_i is the ${}^{n_i}C_2 \times {}^{n_i}C_2$ diagonal matrix with elements $\boldsymbol{\zeta}_{ijk}(1 - \boldsymbol{\zeta}_{ijk})$. Equations (6) and (7) are solved simultaneously for $\boldsymbol{\beta}$ and α.

2.2 Dependence in Bivariate Binary Outcomes

Consider binary outcomes Y_{i1} and Y_{i2} for ith individual. If Y_{i1} and Y_{i2} are not independent, then the conditional probability of Y_{i2} given Y_{i1} can be expressed as [6, 25]

$$\begin{aligned} &P(Y_{i2} = 1|Y_{i1}, \boldsymbol{X}_{i2} = \boldsymbol{x}_{i2}) \\ &= P(Y_{i2} = 1|\boldsymbol{X}_{i2} = \boldsymbol{x}_{i2}) + \rho_i \left(Y_{i1} - P(Y_{i1}|\boldsymbol{X}_{i1} = \boldsymbol{x}_{i1})\right) \end{aligned} \tag{8}$$

where ρ is the correlation between Y_{i1} and Y_{i2}. For $Y_{i1} = 0$, Eq. (8) can be expressed as,

$$\begin{aligned} &P(Y_{i2} = 1|Y_{i1} = 0, \boldsymbol{X}_{i2} = \boldsymbol{x}_{i2}) \\ &= P(Y_{i2} = 1|\boldsymbol{x}_{i2}) + \rho_i \left(0 - P(Y_{i1}|\boldsymbol{x}_{i1})\right) \\ &\text{or,} \\ &\frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{01}}}{1 + e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{01}}} = \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{2}}}{1 + e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{2}}} - \rho_i . \frac{e^{\boldsymbol{x}_{i1}\boldsymbol{\beta}_{1}}}{1 + e^{\boldsymbol{x}_{i1}\boldsymbol{\beta}_{1}}} \end{aligned} \tag{9}$$

and for $Y_{i1} = 1$, Eq. (8) can be expressed as:

$$\begin{aligned} &P(Y_{i2} = 1|Y_{i1} = 1, \boldsymbol{x}_{i2}) \\ &= P(Y_{i2} = 1|\boldsymbol{x}_{i2}) + \rho_i \left(1 - P(Y_{i1}|\boldsymbol{x}_{i1})\right) \\ &or, \\ &\frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{11}}}{1 + e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{11}}} = \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{2}}}{1 + e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{2}}} + \rho_i \left(1 - \frac{e^{\boldsymbol{x}_{i1}\boldsymbol{\beta}_{1}}}{1 + e^{\boldsymbol{x}_{i1}\boldsymbol{\beta}_{1}}}\right) \end{aligned} \tag{10}$$

Clearly ρ_i is a function of $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_{2.1}$ where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are the parameters of the marginal models (Eq. 1), $j = 1, 2$ and $\boldsymbol{\beta}_{2.1} = \boldsymbol{\beta}_{01}$ or $\boldsymbol{\beta}_{11}$, are the vectors of parameters of the two conditional models (Eq. 2) for $j = 2$. When Y_{i1} and Y_{i2} are not correlated, then $\rho_i = 0$ and

$$\begin{aligned} &P(Y_{i2} = 1|Y_{i1}, \boldsymbol{X}_{i2} = \boldsymbol{x}_{i2}) \\ &= P(Y_{i2} = 1|\boldsymbol{X}_{i2} = \boldsymbol{x}_{i2}) \\ &= \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{2}}}{1 + e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{2}}}. \end{aligned} \tag{11}$$

Theoretically, the observed correlation between two repeated outcome variables, Y_{i1} and Y_{i2}, can be shown as:

$$\rho_i = \frac{\text{cov}(Y_{i1}, Y_{i2})}{\sqrt{V(Y_{i1})}\sqrt{V(Y_{i2})}} = \frac{E(Y_{i1}Y_{i2}) - E(Y_{i1})E(Y_{i2})}{\sqrt{\mu_{i1}(1 - \mu_{i1})}\sqrt{\mu_{i2}(1 - \mu_{i2})}} \tag{12}$$

where

$$\begin{aligned} E(Y_{i1}Y_{i2}) &= \sum_{y_{i1},y_{i2}=0}^{1} y_{i1}y_{i2}P(Y_{i1}=y_{i1},Y_{i2}=y_{i2}) \\ &= P(Y_{i1}=1,Y_{i2}=1) \\ &= P(Y_{i2}=1|Y_{i1}=1)P(Y_{i1}=1) \\ &= \frac{e^{x_{i2}\boldsymbol{\beta}_{11}}}{1+e^{x_{i2}\boldsymbol{\beta}_{11}}}.\frac{e^{x_{i1}\boldsymbol{\beta}_{1}}}{1+e^{x_{i1}\boldsymbol{\beta}_{1}}} \end{aligned} \tag{13}$$

If $\boldsymbol{X}_{ij}$ is time invariant then the correlation between Y_{i1} and Y_{i2}, can be shown as:

$$\rho_i = e^{\frac{1}{2}\boldsymbol{x}_i(\boldsymbol{\beta}_1-\boldsymbol{\beta}_2)}\frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}_{11}}-e^{\boldsymbol{x}_i\boldsymbol{\beta}_2}}{(1+e^{\boldsymbol{x}_i\boldsymbol{\beta}_{11}})}. \tag{14}$$

Equation (14) shows that correlation between Y_{i1} and Y_{i2} is equal to zero when $\boldsymbol{\beta}_{11}=\boldsymbol{\beta}_2$. However, this condition does not completely define no association between Y_{i1} and Y_{i2}. Equations (9) and (10) show that for the independence of Y_{i1} and Y_{i2}, it is necessary that both $\boldsymbol{\beta}_{01}$ and $\boldsymbol{\beta_{11}}$ are equal and equal to $\boldsymbol{\beta}_2$. If $\boldsymbol{\beta}_{01}\neq\boldsymbol{\beta}_{11}$ then Y_{i1} and Y_{i2} are associated. Islam et al. [17] showed that the dependence in bivariate Bernoulli outcome variables can be tested by testing the equality of the conditional models. Y_{i1} and Y_{i2} are independent if $P(Y_{i2}=1|Y_{i1}=s_1,\boldsymbol{X}_{i2}=\boldsymbol{x}_{i2})=P(Y_{i2}=y_{i2}|Y_{i1}=0,\boldsymbol{X}_{i2}=\boldsymbol{x}_{i2})=P(Y_{i2}=y_{i2}|Y_{i1}=1,\boldsymbol{X}_{i2}=\boldsymbol{x}_{i2})=P(Y_{i2}=y_{i2}|\boldsymbol{X}_{i2}=\boldsymbol{x}_{i2})$, i.e. $\boldsymbol{\beta}_{2.s_1}=\boldsymbol{\beta}_{01}=\boldsymbol{\beta}_{11}=\boldsymbol{\beta}_2$. It should also be noted that even if the distribution of Y_{i1}, Y_{i2}, ..., Y_{ij} are independent, i.e., $\boldsymbol{\beta}_{j.12\ldots j-1}=\boldsymbol{\beta}_j$, this does not necessarily mean that the distribution of Y_{ij}'s are identical. Distribution of Y_{i1}, Y_{i2}, ..., Y_{ij} are identical only if $\boldsymbol{\beta}_1=\boldsymbol{\beta}_2=\ldots=\boldsymbol{\beta}_j=\boldsymbol{\beta}$.

GEE is a method for marginal or population averaged model and it considers $\boldsymbol{\beta}_1=\boldsymbol{\beta}_{2.1}=\ldots=\boldsymbol{\beta}_{n_i.12\ldots n_i-1}=\boldsymbol{\beta}$, although inducing a (nuisance) correlation structure. ALR is also a marginal model-based approach and hence the association of repeated responses cannot be addressed in a true sense in ALR. Clearly, while analyzing longitudinal data with correlated response variables or response variables from independent but non-identical populations at different time points, fitting marginal-conditional models for Y_{ij}'s is a more appropriate choice as marginal models fail to utilize the major advantage of longitudinal data of observing the change in the outcome variable with respect to time because a marginal model is not able to apprehend the scenario.

It might be noted here that Darlington and Farewell [6] proposed a transition probability model based on the transition probability $P(Y_{i2}=1|Y_{i1}=1,\boldsymbol{x_i})=\frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}_{11}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}_{11}}}$ and the marginal probability $P(Y_{i2}=1|\boldsymbol{x_i})=\frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i\boldsymbol{\beta}}}$, where β is the vector of parameters of the marginal model $P(Y_{ij}=1|\boldsymbol{x}_i)$. Essentially Darlington and Farewell [6] addressed the correlation partially, as they have not considered the transition probability $P(Y_{i2}=1|Y_{i1}=0,\boldsymbol{x_i})$ in their model.

3 Proposed Models

In this study, we propose two joint models based on marginal conditional approach for repeated binary outcomes. We start from the model considered by Darlington and Farewell [6] with the working likelihood function:

$$L(\boldsymbol{\beta}, \boldsymbol{\beta}_{11}) = \prod_{i=1}^{N} p_i^{\,y_{i1}}(1-p_i)^{1-y_{i1}} \prod_{j=2}^{n_i} p_{ij}^{*\,y_{ij}}(1-p_{ij}^{*})^{1-y_{ij}} \tag{15}$$

where $p_i = \Pr(Y_{ij} = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}'\boldsymbol{x}_i}}$ and $p_{ij}^{*} = \Pr(Y_{ij} = 1|Y_{ij-1}, \boldsymbol{X}_i = \boldsymbol{x}_i) = E(Y_{ij}|Y_{ij-1}, \boldsymbol{X}_i = \boldsymbol{x}_i) = p_i + \rho_i(Y_{ij-1} - p_i)$and $\rho_i = \frac{e^{\boldsymbol{\beta}_{11}{}'x_i} - e^{\boldsymbol{\beta}^{*\prime}x_i}}{1+e^{\beta_{11}{}'\boldsymbol{x}_i}}$, $\max(-\frac{p_i}{1-p_i}, -\frac{1-p_i}{p_i}) < \rho_i < 1$ because the likelihood must be maximized at $0 < p_i < 1$ and $0 < p_{ij} < 1$. The limitations of this model is that it does not consider the transition probability from $Y_{ij-1} = 0$ to $Y_{ij} = 1$ and considered $p_{ij}^{*} = \Pr(Y_{ij} = 1|Y_{ij-1} = 1, \boldsymbol{X}_i)$. Although, the transition probability $P(Y_{ij} = 1|Y_{ij-1} = 0)$ was not considered in determining the correlation, while defining the range of ρ_i, the transition from $Y_{ij-1} = 0$ was considered which contradicts with the definition of ρ_i. A straight forward and simple way to improve the model discussed by Darlington and Farewell [6] by including both the transition probabilities, $P(Y_{ij} = 1|Y_{ij-1} = 0)$ and $P(Y_{ij} = 1|Y_{ij-1} = 1)$, in the working likelihood function is discussed in the following subsection.

3.1 Proposed Model 1

For any order of Markov chain with covariate dependence, a model based on marginal and conditional models can be used. Consider the simplest case of two repeated measures on each individuals. If Y_{i2} depends on Y_{i1}, then for each possible values of y_{i1}, we get one conditional model for Y_{i2}. As we assumed Y_{ij} to be binary random variables, Y_{i1} can take values 0 and 1. When $Y_{i1} = 0$, the conditional probability of $Y_{i2} = 1$ can be defined as

$$\begin{aligned} p_{i2}^{*} &= \Pr(Y_{i2} = 1|y_{i1} = 0, \boldsymbol{x}_{i2}) \\ &= \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{01}}}{1+e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{01}}};\ i = 1, 2, \ldots, N, \end{aligned} \tag{16}$$

where $\boldsymbol{\beta}_{01}$ is the vector of parameters of the conditional model of $P(Y_{i2} = 1|Y_{i1} = 0, X_{i2} = x_{i2})$; $i = 1, 2, \ldots N$. Here the suffix (01) of $\boldsymbol{\beta}$ is used to show the transition from $Y_{i1} = 0$ and $Y_{i2} = 1$.

Similarly, when $Y_{i1} = 1$, the conditional probability of $Y_{i2} = 1$ can be defined as

$$\begin{aligned} p_{i2}^{*} &= \Pr(Y_{i2} = 1|y_{i1} = 1, \boldsymbol{x}_{i2}) \\ &= \frac{e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{11}}}{1+e^{\boldsymbol{x}_{i2}\boldsymbol{\beta}_{11}}};\ i = 1, 2, \ldots, N, \end{aligned} \tag{17}$$

where $\boldsymbol{\beta}_{11}$ is the vector of parameters of the conditional model of $P(Y_{i2} = 1|Y_{i1} = 1, X_{i2} = x_{i2})$; $i = 1, 2, \cdots N$. The suffix (11) of $\boldsymbol{\beta}$ is used to show the transition from $Y_{i1} = 1$ and $Y_{i2} = 1$.

The joint probabilities can be expressed as, for $i = 1, 2, \ldots, N$,

$$P(Y_{i1}, Y_{i2}) = P(Y_{i2} = 1|Y_{i1} = y_{i1}, \boldsymbol{x}_{i2})P(Y_{i1} = y_{i1}|\boldsymbol{x}_{i1}). \tag{18}$$

In general, the joint mass function for n_i outcome variables, $Y_{i1}, Y_{i2}, ..., Y_{in_i}$, for subject i at follow-ups $1, 2..., n_i$, respectively, in the presence of covariates $\boldsymbol{X}_{ij} = (1, X_{ij1}, X_{ij2}, ..., X_{ijp})$, can be expressed as product of the conditional and marginal probability mass functions for given values of covariates as follows:

$$\begin{aligned} &\Pr(Y_{i1} = y_{i1}, ..., Y_{in_i} = y_{in_i}|\boldsymbol{X_i} = \boldsymbol{x_i}) \\ &= \Pr(Y_{i1}) \prod_{j=2}^{n_i} \Pr(Y_{ij}|Y_{i1} = y_{i1}, ..., Y_{ij-1} = y_{ij-1}, \boldsymbol{X_i} = \boldsymbol{x_i}). \end{aligned} \tag{19}$$

In general, let us consider n_i possibly correlated outcome variables $(Y_{i1}, Y_{i2}, ..., Y_{in_i})$ on each of N individuals. Let $\boldsymbol{\theta} = (\theta_1, \theta_{2.1}, ..., \theta_{n_i.1,2,...,n_i-1})$ be the vector of unknown parameters where $\theta_j = g(\mu_{ij}) = X_{ij}\boldsymbol{\beta}_j$, $\theta_{j.12...j-1} = g(\mu_{ij.12...j-1}) = \boldsymbol{X}_{ij}\boldsymbol{\beta}_{j.12...j-1}$ and g is an appropriate link function. The joint probability mass function of $Y_{i1}, ..., Y_{in_i}$ can be expressed as:

$$\begin{aligned} &P(Y_{i1} = y_{i1}, \cdots, Y_{in_i} = y_{in_i}) \\ &= P(Y_{i1} = y_{i1}|\boldsymbol{x}_{i1}).P(Y_{i2} = y_{i2}|\boldsymbol{x}_{i2}, y_{i1}) \\ &\cdots P(Y_{in_i} = y_{in_i}|\boldsymbol{x}_{in_i}, y_{i1}, \cdots, y_{in_i-1}). \end{aligned} \tag{20}$$

The likelihood function can be expressed as:

$$\begin{aligned} L(\boldsymbol{\beta}) = \prod_{i=1}^{N} &f(y_{i1}|\boldsymbol{x}_{i1}, \boldsymbol{\beta}_1) f(y_{i2.1}|\boldsymbol{x}_{i2}, \boldsymbol{\beta}_{2.1}) \\ &\cdots f(y_{n_i.1,2,...,n_i-1}|\boldsymbol{x}_{in_i}, \boldsymbol{\beta}_{n_i.1,2,...,n_i-1}), \end{aligned} \tag{21}$$

where $f(y_{i1}|\boldsymbol{x}_{i1}, \boldsymbol{\beta}_1)$ is the marginal distribution of y_{i1} for given $\boldsymbol{X}_{i1} = \boldsymbol{x}_{i1}$ and the conditional probabilities of y_{ij}, given $Y_{i1} = y_{i1}, ..., Y_{ij-1} = y_{ij-1}$, and $\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}$ are $f(y_{ij.12...j-1}) = f(y_{ij}|\boldsymbol{x}_{ij}, y_{i1}, ..., y_{ij-1}, \beta_{j.1,2,...,j-1})$, $j = 2, 3, ..., n_i$. Let l_{ij} be the contribution of ijth term to the log likelihood function. Differentiating the log-likelihood, $l = \sum_{i=1}^{N} \sum_{j=1}^{n_i} l_{ij}$, with respect to corresponding parameters, and equating to zero, the estimating equations are:

$$\frac{\partial l}{\delta \beta_k} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{\partial l_{ij}}{\partial \theta_j} \frac{\partial \theta_j}{\partial \mu_{ij}} . \frac{\partial \mu_{ij}}{\partial \beta_k} = 0. \tag{22}$$

The estimates of $\boldsymbol{\beta}$ can be obtained by maximum likelihood method.

The variance of the estimates, $V(\hat{\boldsymbol{\beta}})$, is obtained from the inverse of the information matrix I, where I is a $(2^{n_i}-1)(p+1)\times(2^{n_i}-1)(p+1)$ matrix with kk'th elements $-\frac{\partial^2 l}{\partial\beta_k\partial\beta_k'}; k, k' = 0, 1, ..., p$.

For example, consider possibly correlated Bernoulli outcome variables $Y_{i1}, ..., Y_{in_i}$, with probability of success $p_{i1}, p^*_{i2}, ..., p^*_{in_i}$, where p^*_{ij} denotes the conditional probability, $P(Y_{ij}=1|y_{i1}, ..., y_{ij-1}, \boldsymbol{x}_{ij}), j = 2, ..., n_i$. Then $f(y_{i1}|\boldsymbol{x}_{i1}, \boldsymbol{\beta}_1) = p_{i1}^{y_{i1}}(1-p_{i1})^{1-y_{i1}}$ and $f(y_{ij.12...j-1}|\boldsymbol{x}_{ij}, \boldsymbol{\beta}_{j.12...j-1}) = {p^*}_{ij}^{y_{ij}}(1-{p^*}_{ij})^{1-y_{ij}}$, $j = 2, ..., n_i$. The likelihood function can be expressed as

$$L = \prod_{i=1}^{N} e^{y_{i1}ln\frac{p_{i1}}{1-p_{i1}}+y_{i2}ln\frac{p^*_{i2}}{1-p^*_{i2}}+...+y_{in_i}ln\frac{p^*_{in_i}}{1-p^*_{in_i}}}. \tag{23}$$

Similar representation was shown previously by Islam and Chowdhury [14]. Differentiating the log-likelihood with respect to the respective parameters and equating to zero, the score equations for $\boldsymbol{\beta}$ are obtained as:

$$\begin{aligned}\frac{\partial l}{\partial\beta_k} &= \sum_{i=1}^{N} X_{i1k}(y_{i1}-p_{i1}) \\ &\quad + \sum_{i=1}^{N}\sum_{j=2}^{n_i} X_{ijk}(y_{ij}-p^*_{ij}), k = 0, 1, ..., p.\end{aligned} \tag{24}$$

The information matrix I_β is a $(2^{n_i}-1)(p+1)\times(2^{n_i}-1)(p+1)$ matrix with elements $-\frac{\partial^2 l}{\partial\beta_k\partial\beta_k'}, k, k' = 0, 1, ..., p$.

Test of Hypothesis

To test the significance of the overall model, the null and alternative hypothesis can be expressed as: $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta_0}$ vs $H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta_0}$ where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_{2.1}, ..., \boldsymbol{\beta}_{n_i.1,2,...n_i-1})$ and $\boldsymbol{\beta}_0$ is the value of $\boldsymbol{\beta}$ under null hypothesis of no covariate effect. The test statistic $\Lambda = -2[\ln L(\boldsymbol{\beta}_0) - \ln L(\boldsymbol{\beta})]$ has a chi-square distribution under H_0 with $(2^{n_i}-1)p$ d.f. Here $\ln L(\boldsymbol{\beta})$ is the log likelihood of the full model and $\ln L(\boldsymbol{\beta_0})$ is the log likelihood of the reduced model for no covariate effects, i.e. the value of $\ln L(\boldsymbol{\beta})$ under H_0. For testing $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$ the test statistic is $z = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$ which follows $N(0, 1)$ under H_0. The major limitation of the proposed joint model based on Markov transition probability in Eq. (21) is the rapid increase in the number of parameters for increasing number of follow-ups. With n_i follow-ups, the number of parameters to be estimated is as big as $(2^{n_i}-1)(p+1)$ where p is the number of covariates. To overcome the limitations of the proposed model 1, in the following section, we propose a second set of joint models as an alternative.

3.2 Proposed Model 2

If there are more than three follow-ups in a longitudinal data, the number of parameters of the joint model becomes as big as $(2^4 - 1)(p + 1)$, for 4 follow-ups where p is the number of covariates.. In this section, an alternative to GEE approach is developed based on the regressive models [3] in order to analyze repeated measures data. The generalized form of the regressive model was proposed by Islam et al. [16]. Following the notations of Islam et al. [16], let us define $\boldsymbol{\lambda}'_{j-1} = (\boldsymbol{\beta}', \boldsymbol{\gamma}'_{j-1}, \boldsymbol{\rho}'_{j-1}, \boldsymbol{\eta}'_{j-1})$ and $\boldsymbol{W}'_{j-1} = (\boldsymbol{X}'_{ij}, \boldsymbol{Y}'_{j-1}, \boldsymbol{\nu}'_{j-1}, \boldsymbol{Z}'_{j-1})$ where $\boldsymbol{X}_{ij} = (1, X_{ij1}, X_{ij2}, ..., X_{ijp})$ and $\boldsymbol{Y}_{j-1} = Y_{i1}, ..., Y_{ij-1}$, $\boldsymbol{\nu}_{j-1} = (\nu_{12}, \nu_{123}, ..., \nu_{12...j-1})' = (y_{i1}y_{i2}, y_{i1}y_{i2}y_{i3}, ..., y_{i1}y_{i2}...y_{ij-1})'$ are the interaction terms among Y_{ij}s, $j = 1, ..., n_i$ and $\boldsymbol{Z}_{j-1} = (z_{11}, ..., z_{1p}, ..., z_{j-1p})'$ are the interaction terms among $\boldsymbol{X}_{ij}$ and $\boldsymbol{Y}_i = (x_{i1}y_{i1}, ..., x_{ip}y_{i1}, ..., x_{i1}y_{ij-1}, ..., x_{ip}y_{ij-1})'$. $\boldsymbol{\beta}' = (\beta_0, \beta_1, ..., \beta_p)$ are the coefficients of $\boldsymbol{X}_{ij}$; $\boldsymbol{\gamma}'_j = (\gamma_1, ..., \gamma_{j-1})$, the parameters corresponding to $Y_{i1}, ..., Y_{ij-1}$; $\boldsymbol{\rho}'_{j-1} = (\rho_{12}, \rho_{123}, ..., \rho_{12...j-1})$, the coefficients of the interaction terms among Y_{ij}'s, and $\boldsymbol{\eta}_{-1} = (\eta_{11}, ..., \eta_{j-1p})$ be the parameters corresponding to $\boldsymbol{Z}_{j-1}$. The regressive model for the jth follow-up is defined as:

$$P(Y_{ij} = s|w_{j-1}) = \frac{e^{\boldsymbol{\lambda}'_j \boldsymbol{w}_{j-1} s}}{1 + e^{\boldsymbol{\lambda}'_j \boldsymbol{w}_{j-1}}}, s = 0, 1, \ \ j = 2, ..., n_i. \tag{25}$$

The likelihood function can be expressed as:

$$L = \prod_{i=1}^{N} f(y_{i1}|\boldsymbol{x}_1, \boldsymbol{\lambda}_1) f(y_{i2}|\boldsymbol{x}_i, \boldsymbol{\lambda}_2)...f(y_{n_i}|\boldsymbol{x}_i, \boldsymbol{\lambda}_{n_i}). \tag{26}$$

The score equations can be obtained by differentiating the log likelihood, $l = \log L$, with respect to the respective parameters. The information can be obtained as $-\frac{\partial^2 l}{\partial \boldsymbol{\lambda} \partial \boldsymbol{\lambda}'}$.

Test of Hypothesis

To test for the dependence of the jth outcome on earlier outcomes and other related terms, the null hypothesis can be shown as: $H_0 : \boldsymbol{\lambda}^*_{j-1} = 0$ against $H_1 : \boldsymbol{\lambda}^*_{j-1} \neq 0$ where $\boldsymbol{\lambda}^*_{j-1} = (\boldsymbol{\gamma}_{j-1}, \boldsymbol{\rho}_{j-1}, \boldsymbol{\eta}_{j-1})'$. The total number of parameters need to be tested is $(2^{j-1} - 1)$ for $\boldsymbol{\gamma}_{j-1}$ and $\boldsymbol{\rho}_{j-1}$ and $(j - 1) \times p$ parameters for $\boldsymbol{\eta}_{j-1}$. The test statistic is a likelihood ratio and follows chi-square distribution with $(2^{j-1} - 1) + (j - 1) \times p$ degrees of freedom [13]. Under independence, the model in Eq. (25) can be defined as

$$P(Y_{ij} = s|\boldsymbol{x}_{ij}, \boldsymbol{Y}_{ij-1}, \boldsymbol{\nu}_{j-1}\boldsymbol{Z}_{j-1}) = \frac{e^{\boldsymbol{\beta}'\boldsymbol{x}_{ij}s}}{1 + e^{\boldsymbol{\beta}'\boldsymbol{x}_{ij}}}, s = 0, 1. \tag{27}$$

If the outcomes are independent, one can simply fit the reduced model using a maximum likelihood method. If the outcomes are associated, the full model as given in Eq. (25) is suggested.

4 Simulation Study

A simulation study was carried out to compare the properties of estimates of regression coefficients of the models discussed in the earlier sections. The repeated measures can be associated in a variety of ways and in this study, the cases considered are:(i) Y_{ij}'s are identically and independently distributed, (ii) Y_{ij}'s are identically distributed and associated (iii) Y_{ij}'s are not identical and their distributions are independent.

4.1 Simulation Design

For simplicity of the study, we restrict the simulation study for the conditional marginal model to two follow-ups, Y_{i1} and Y_{i2} on ith subject and only one explanatory variable, X_{i1} for each of the N individuals where X_{i1} is fixed and time invariant. We assumed that Y_{i1} and Y_{i2} are two binary random variables with $Y_{i1} \sim B(1, p_{i1})$ and $Y_{i2} \sim B(1, p_{i2})$. The corresponding generalized linear models are $g(\mu_{i1}) = \frac{e^{\boldsymbol{\beta}_1' \boldsymbol{X}_i}}{1+e^{\boldsymbol{\beta}_1' \boldsymbol{X}_i}}$ and $g(\mu_{i2}) = \frac{e^{\boldsymbol{\beta}_2' \boldsymbol{X}_i}}{1+e^{\boldsymbol{\beta}_2' \boldsymbol{X}_i}}$.

The simulation followed the following steps: a time invariant explanatory variable X_i was generated first from Bernoulli distribution with probability of success 0.5. Then p_{i1}, the probability of success of Y_{i1} was calculated using the equation $P(Y_{i1} = 1|\boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}_1' \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}_1' \boldsymbol{x}_i}}$ for selected values of $\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11})'$ where $\boldsymbol{X}_i = (1, X_i)$, β_{10} is the intercept term and β_{11} is the coefficient of X_i. N values, a_i, were generated from uniform distribution within range (0, 1) and then Y_{i1} was generated such that $Y_{i1} = 1$ if $a_i < P(Y_{i1} = 1|x_i)$ and 0 otherwise. To generate data on Y_{i2}, first, the probability of success at time point 2, p_{i2}, was calculated as $P(Y_{i2} = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{e^{\boldsymbol{\beta}_2' \boldsymbol{x}_i}}{1+e^{\boldsymbol{\beta}_2' \boldsymbol{x}_i}}$. Here $\boldsymbol{\beta}_2 = (\beta_{20} + \gamma_1 y_{i1}, \beta_{21})$ where $\beta_{20} + \gamma_1 y_{i1}$ is the intercept term, β_{21} is the coefficient of X_i and γ_1 is the coefficient of Y_{i1}. Similar as Y_{i1}, N values, b_i, were generated from uniform distribution within range (0, 1) and then Y_{i2} was generated such that $Y_{i2} = 1$ if $b_i < P(Y_{i2} = 1|X_i = x_i)$ and 0 otherwise.

For illustration of the regressive model, Y_{i1}, Y_{i2}, Y_{i3} and Y_{i4} were generated in a similar way as Y_{i1} with $\boldsymbol{\beta}_1 = (\beta_0, \beta_1)'$, Y_{i2} with $\boldsymbol{\beta}_2 = (\beta_0, \beta_1, \gamma_1)'$, Y_{i3} with $\boldsymbol{\beta}_3 = (\beta_0, \beta_1, \gamma_1, \gamma_2)'$ and Y_{i4} with $\boldsymbol{\beta}_4 = (\beta_0, \beta_1, \gamma_1, \gamma_2, \gamma_3)'$, respectively.

Y_{ij}'s are independently and identically distributed when $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and $\gamma_1 = 0$; distribution of Y_{ij}'s are identical ($\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$) but they are associated ($\gamma_1 \neq 0$); Y_{ij}'s are independent ($\gamma_1 = 0$) but the distribution of Y_{ij}'s are not identical ($\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$). GEE under different correlation structures (independent, exchangeable and autoregressive), ALR under exchangeable correlation and joint models were fitted. The bias, the standard error of the estimates and coverage probability of the 95 % confidence interval

were constructed over a range of scenarios for large samples and varying association among the repeated responses.

4.2 Simulation Results

The findings of the simulation study (estimates, bias, standard error and coverage probability) are summarized in Tables 1, 2 and 3. In all the following tables, GEE(In), GEE(Ex), GEE(AR) stand for GEE models under independent, exchangeable and autoregressive correlations, respectively. ALR(Ex)denotes the ALR model under an exchangeable correlation. The parameters of the joint model are β_{10}, β_{11}, β_{010}, β_{011}, β_{110} and β_{111}. Here, β_{10} and β_{11}, respectively, denote the intercept and the regression coefficients of the marginal model $P(Y_{i1} = 1|\boldsymbol{X}_i = \boldsymbol{x}_i)$; β_{010} and β_{011}, respectively, denote the intercept and the regression coefficient of the conditional model $P(Y_{i2} = 1|Y_{i1} = 0, \boldsymbol{X}_i = \boldsymbol{x}_i)$; and β_{110} and β_{111}, respectively, denote the intercept and regression coefficient of the conditional model $P(Y_{i2} = 1|Y_{i1} = 1, \boldsymbol{X}_i = \boldsymbol{x}_i)$. GEE or ALR, being approaches based on marginal models, estimate the parameters of such models as an average of the parameters of two populations from where Y_{i1} and Y_{i2} were generated. To distinguish the parameters of GEE and ALR from joint model, we used the notation $\boldsymbol{\beta}^* = (\beta_0^*, \beta_1^*)'$ to denote the parameters of GEE and ALR in the following tables.

In Table 1, $P(Y_{i1} = 1|\boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{e^{\beta_{10}+\beta_{11}x_{i11}}}{1+e^{\beta_{10}+\beta_{11}x_{i11}}} = \frac{e^{0.5+0.2x_{i11}}}{1+e^{0.5+0.2x_{i11}}}$ and $P(Y_{i2} = 1|Y_{i1} = y_{i1}, \boldsymbol{X}_i = \boldsymbol{x}_i) = \frac{e^{\beta_{20}+\beta_{21}x_{i21}+\gamma_1 y_{i1}}}{1+e^{\beta_{20}+\beta_{21}x_{i21}+\gamma_1 y_{i1}}}$, where, β_{20} and β_{21}, respectively, denote the intercept and regression coefficients of the marginal model $P(Y_{i2} = 1|\boldsymbol{X}_i = \boldsymbol{x}_i)$; So if $\gamma_1 = 0$, the true values of the parameters to be estimated for the joint model are $\beta_{10} = 0.5$, $\beta_{11} = 0.2$, $\beta_{010} = \beta_{20} + \gamma_1 \times (y_{i1} = 0) = 0.5 + 0 \times 0 = 0.5 = \beta_{20}$, $\beta_{011} = \beta_{21} = 0.2$, $\beta_{110} = \beta_{20} + \gamma_1 \times (y_{i1} = 1) = 0.5 + 0 \times 1 = 0.5$ and $\beta_{111} = \beta_{21} = 0.2$. When $\gamma_1 = 1$, the true values of the parameters to be estimated for the joint model are $\beta_{10} = 0.5$, $\beta_{11} = 0.2$, $\beta_{010} = \beta_{20} + \gamma_1 \times (y_{i1} = 0) = 0.5 + 1 \times 0 = 0.5$, $\beta_{011} = \beta_{21} = 0.2$, $\beta_{110} = \beta_{20} + \gamma_1 \times (y_{i1} = 1) = 0.5 + 1 \times 1 = 1.5$ and $\beta_{111} = \beta_{21} = 0.2$.

Table 1 shows that bias and the standard error of estimates of the proposed Model 1 (extension of Darlington and Farewell [6]), Proposed Model 2, GEE and ALR are competitive for longitudinal data when the repeated measures are independent ($\gamma_1 = 0.0$). Inadequacy of GEE or ALR to portray the relationship between X and Y are visible with the presence of dependence relationship between Y_{i1} and Y_{i2} as shown in Table 1 where the data are generated from two associated populations ($\gamma_1 = 1.0$). The marginal parameters in the Model 1 proposed as an extension of Darlington and Farewell [6] does not make much improvement in the performance of the parameters in terms of bias and standard error.

The proposed joint model (Model 2) gives better estimates in this case. The inadequacy of GEE or ALR to portray the relationship between X and Y are also observed in Table 2 where the data are generated from two independent but nonidentical populations. The estimates of parameters of GEE do not portray the actual relationship between the covariates and the response variable because of the variation in the rela-

Table 1 Parameters (Par), estimates(Est.), bias, standard error(SE) and coverage probability (CP) of estimates for independent ($\gamma_1 = 0.0$) and correlated outcomes ($\gamma_1 = 1.0$) with identical distributions of Y_{i1} and Y_{i2}, $(\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}) = (0.5, 0.2), \boldsymbol{\beta}_2 = (\beta_{20}, \beta_{21}) = (0.5, 0.2))$

	$\gamma_1 = 0$					$\gamma_1 = 1$				
	Par	Est	Bias	SE	CP	Par	Est	Bias	SE	CP
Model 1	$\beta_{10} = 0.5$	0.5127	− 0.0127	0.2066	0.9580	$\beta_{10} = 0.5$	0.5127	− 0.0127	0.2066	0.9580
	$\beta_{11} = 0.2$	0.2030	− 0.0030	0.2985	0.9550	$\beta_{11} = 0.2$	0.2030	− 0.0030	0.2985	0.9550
	$\beta_{20} = 0.5$	0.4997	0.0003	0.2064	0.9510	$\beta_{010} = 0.5$	0.5013	− 0.0013	0.3420	0.9660
	$\beta_{21} = 0.2$	0.2109	− 0.0109	0.2982	0.9460	$\beta_{011} = 0.2$	0.2046	− 0.0046	0.5138	0.9500
						$\beta_{110} = 1.5$	1.5365	− 0.0365	0.3357	0.9640
						$\beta_{111} = 0.2$	0.2228	− 0.0228	0.4896	0.9630
GEE(In)	$\beta_0^* = 0.5$	0.5037	− 0.0037	0.1453	0.9470	$\beta_0^* = 0.5$	0.7756	− 0.2756	0.1659	0.6400
	$\beta_1^* = 0.2$	0.2054	− 0.0054	0.2101	0.9510	$\beta_1^* = 0.2$	0.2181	− 0.0181	0.2409	0.9520
GEE(Ex)	$\beta_0^* = 0.5$	0.5037	− 0.0037	0.1453	0.9470	$\beta_0^* = 0.5$	0.7756	− 0.2756	0.1659	0.6400
	$\beta_1^* = 0.2$	0.2054	− 0.0054	0.2101	0.9510	$\beta_1^* = 0.2$	0.2181	− 0.0181	0.2409	0.9520
ALR(Ex)	$\beta_0^* = 0.5$	0.5037	− 0.0037	0.1453	0.9470	$\beta_0^* = 0.5$	0.7756	− 0.2756	0.1659	0.6400
	$\beta_1^* = 0.2$	0.2054	− 0.0054	0.2101	0.9510	$\beta_1^* = 0.2$	0.2181	− 0.0181	0.2409	0.9520

Table 2 Estimates(Est.), bias, standard error(SE) and coverage probability (CP) of estimates for independent outcomes with non-identical distributions, ($\boldsymbol{\beta}_1 = (\beta_{10}, \beta_{11}) = (0.5, 0.2)$, $\boldsymbol{\beta}_2 = (\beta_{20} + \gamma_1 y_{i1}, \beta_{21}) = (0.2, 0.7)$, $\gamma_1 = 0.0$)

Methods	Par	Est	Bias from $\boldsymbol{\beta}_1$	Bias from $\boldsymbol{\beta}_2$	SE	CP for $\boldsymbol{\beta}_1$	CP for $\boldsymbol{\beta}_2$
Model 1	β_{10}	0.5127	− 0.0127	–	0.2066	0.9580	–
	β_{11}	0.2030	− 0.0030	–	0.2985	0.9550	–
	β_{20}	0.1992	–	0.0008	0.2010	–	0.9520
	β_{21}	0.7157	–	− 0.0157	0.3010	–	0.9520
GEE(In)	β_0^*	0.3522	0.1478	− 0.1522	0.1426	0.818	0.828
	β_1^*	0.4566	− 0.2566	0.2434	0.2101	0.763	0.780
GEE(Ex)	β_0^*	0.3522	0.1478	− 0.1522	0.1426	0.818	0.828
	β_1^*	0.4566	− 0.2566	0.2434	0.2101	0.763	0.780
ALR(Ex)	β_0^*	0.3522	0.1478	− 0.1522	0.1426	0.818	0.828
	β_1^*	0.4566	− 0.2566	0.2434	0.2101	0.763	0.780

tionship at different time points. And the actual bias from population 1 (from where Y_{i1} were generated) and population 2 (from where Y_{i2} were generated) are shown in two columns of Table 2. Clearly, even if the repeated measures are not associated, while data come from two different populations, the GEE or ALR are not adequate to capture the relationship between the covariates and the response variable. Proposed model 2 is suggested in such cases.

While there are more than three repeated measurements on the same subject, the covariate-dependent Markov Chain-based joint models need to estimate too many parameters and a general form of the regressive model approach [13] is suggested as an alternative of GEE-based approaches. The results of the simulation study (Table 3) show that when the outcomes are independent and identically distributed, the estimates of the parameters of a regressive model produce similar results as GEE or ALR in terms of bias and coverage probability. The regressive model performs better than GEE or ALR while the repeated responses are associated.

Indubitably, GEE and ALR performed well only when repeated measures come from identical population and are not associated. The simulation study also finds that basically there is no difference in the estimates of GEE under different correlation structures (Tables 1, 2, 3). Also, ALR does not show any noticeable difference from GEE estimates in most cases. The proposed model 2 (for 3 or fewer repeated outcomes) and proposed model 3 (for more than 3 repeated outcomes) produce better estimates in terms of bias and coverage probability than GEE or ALR in the cases when responses are associated or the responses at different time points have different distributions.

5 Application to HRS Data

The first three waves of the longitudinal data from the Health and Retirement Study (HRS) conducted by the University of Michigan [29] were used for comparison of

Table 3 Parameters(Par), Estimates(Est), Bias, standard error (SE) and coverage probability (CP) of estimates of different models for independent and associated distribution ($\beta_{10} = \beta_{20} = \beta_{30} = \beta_{40} = \beta_0^* = 0.2$, $\beta_{11} = \beta_{21} = \beta_{31} = \beta_{41} = \beta_1^* = 0.7$, $(\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0)$ and $(1, 1, 1)$)

Methods	$(\gamma_1, \gamma_2, \gamma_3) = (0, 0, 0)$					$(\gamma_1, \gamma_2, \gamma_3) = (1, 1, 1)$				
	Par	Est	Bias	SE	CP	Par	Est	Bias	SE	CP
Model 2	$\beta_0^* = 0.2$	0.208	− 0.008	0.226	0.946	$\beta_0^* = 0.2$	0.252	− 0.052	0.316	0.952
	$\beta_1^* = 0.7$	0.698	0.002	0.199	0.953	$\beta_1^* = 0.7$	0.748	− 0.048	0.397	0.944
	$\gamma_1 = 0.0$	− 0.012	0.012	0.197	0.940	$\gamma_1 = 1.0$	1.010	− 0.010	0.370	0.942
	$\gamma_2 = 0.0$	− 0.001	0.001	0.197	0.942	$\gamma_2 = 1.0$	0.987	0.013	0.352	0.949
	$\gamma_3 = 0.0$	0.003	− 0.003	0.197	0.945	$\gamma_3 = 1.0$	1.001	− 0.001	0.366	0.950
GEE	$\beta_0^* = 0.2$	0.201	− 0.001	0.062	0.950	$\beta_0^* = 0.2$	0.927	− 0.727	0.085	0.000
(In)	$\beta_1^* = 0.7$	0.695	0.005	0.095	0.950	$\beta_1^* = 0.7$	0.799	− 0.099	0.136	0.889
GEE	$\beta_0^* = 0.2$	0.201	− 0.001	0.062	0.950	$\beta_0^* = 0.2$	0.927	− 0.727	0.085	0.000
(Ex)	$\beta_1^* = 0.7$	0.695	0.005	0.095	0.950	$\beta_1^* = 0.7$	0.799	− 0.099	0.136	0.889
GEE	$\beta_0^* = 0.2$	0.201	− 0.001	0.062	0.948	$\beta_0^* = 0.2$	0.921	− 0.721	0.085	0.000
(AR)	$\beta_1^* = 0.7$	0.695	0.005	0.095	0.951	$\beta_1^* = 0.7$	0.788	− 0.088	0.136	0.902
ALR	$\beta_0^* = 0.2$	0.201	− 0.001	0.062	0.950	$\beta_0^* = 0.2$	0.927	− 0.727	0.085	0.000
(Ex)	$\beta_1^* = 0.7$	0.695	0.005	0.095	0.950	$\beta_1^* = 0.7$	0.799	− 0.099	0.136	0.889

the selected methods. The study started in 1992 on American individuals over the age of 50 years and their spouses and the subjects are observed every two years. In wave 1, the sample size was 9760 and the sample size was reduced to 9750 due to the dropping of 10 cases with missing values of outcome variable at first round. Finally, the number of individuals were 8657 who reported that they were not hospitalized at wave 1. The panel data from the waves for 1992, 1994 and 1996 have been used in this study. An Elderly population may suffer from repeated spells of depression which may change over time [8, 15] and result in other health problems and chronic illness [19]. The literature on depression among elderly helped filling many gaps in our understanding of the factors associated with depression and also the outcome of depression [2]. But understanding depression and its associated factors more explicitly is important. In many studies on clinical and non-clinical populations, CESD (Center for Epidemiologic Studies Depression) scale is employed to measure depressive symptoms [28]. The dependent variable for this study is Depression status (no depression (CESD score $= 0$), depression (CESD score > 0)). The independent variables are gender (male=1), marital status (married/partnered=1), education, ethnicity: Black (Black $= 1$), ethnicity: White (White $= 1$), drinking habit (drink=1) and the number of health conditions. In Tables 4 and 5, Mstat stands for marital status, White stands for white ethnicity, Black stands for Black ethnicity, Drink means drinking habit and No. of Cond. is the number of health conditions. In GEE models, we observe that marital status, education year, ethnicity: White and number of health conditions were significantly associated with depression. The GEE model under the assumption of independence and exchangeable correlation produces the same results and finds that marital status, education, White ethnicity and number of health conditions had significant influence on depression among the study population. ALR under an exchangeable correlation, in addition, finds drinking habit as a significant factor for depression. GEE model under the assumption of autoregressive correlation shows that marital status, education, white ethnicity and number of health conditions were significantly associated with the depression status but gender was not significant in GEE-based models.

The joint model shows that the effects of the covariates were different on the depression status at different follow-ups. At the baseline, marital status, education, white ethnicity and the number of conditions had a significant effect on depression. Married people were less depressed as compared to their single counterparts, education lessened the risk of depression, white people were less depressed, number of physical conditions increased the risk of depression.

In the first follow-up, covariate effects were different on depression status depending on what the CESD score was in the baseline (Y_1). If the respondent was not depressed in the baseline, gender, marital status, education year and being white had a significant influence on the dependent variable. Male, married, educated persons and people from White ethnicity are at less risk of being depressed. Gender had no significant effect on those at the first follow-up. Being married and being educated lessens the risk of being depressed for those who were depressed at the baseline.

In the second follow-up, the effects of the covariates were notably different depending on the depression status of the respondent in the previous follow-ups. Depression status of patients (who were not depressed in the baseline or the first follow-up) was

Table 4 Estimates of parameters of GEE and ALR on HRS data

	GEE(In)			GEE(Ex)		
	Est	SE	*p* value	Est	SE	*p* value
Intercept	2.023	0.206	0.000	2.023	0.206	0.000
Gender	− 0.059	0.059	0.321	− 0.059	0.059	0.321
Mstat	− 0.621	0.065	0.000	− 0.621	0.065	0.000
Education	− 0.153	0.010	0.000	− 0.153	0.010	0.000
White	− 0.363	0.166	0.029	− 0.363	0.166	0.029
Black	− 0.085	0.177	0.629	− 0.085	0.177	0.629
Drink	− 0.091	0.055	0.097	− 0.091	0.055	0.097
No. of Cond.	0.389	0.024	0.000	0.389	0.024	0.000
	GEE(AR)			ALR(Ex)		
	Est	SE	*p* value	Est	SE	*p* value
Intercept	1.944	0.206	0.000	2.019	0.192	0.000
Gender	− 0.056	0.060	0.351	− 0.059	0.057	0.153
Mstat	− 0.613	0.065	0.000	− 0.624	0.063	0.000
Education	− 0.151	0.010	0.000	− 0.153	0.010	0.000
White	− 0.338	0.166	0.042	− 0.366	0.153	0.008
Black	− 0.067	0.177	0.706	− 0.085	0.163	0.301
Drink	− 0.082	0.055	0.135	− 0.091	0.053	0.045
No. of Cond.	0.391	0.024	0.000	0.391	0.023	0.000

significantly associated with marital status, education and drinking habit. Depression status of patients (who were not depressed in the baseline but were depressed in the first follow-up) was significantly associated with education. Education had a significant effect on depression status of patients in second follow-up for those who were depressed in the baseline but not depressed in the first follow-up. Respondents' depression status was significantly associated with marital status and education for those who were depressed in both the first and the second follow-ups. These findings confirm our assertion that the extensive use of GEE-based models may result in failure to specify the covariate effects adequately for longitudinal data. The results demonstrate that a joint model based on marginal conditional approach explains the covariate effects more meaningfully.

Table 5 Estimates of parameters of the proposed Model 1 for HRS data

	β_1			β_{01}			β_{11}		
	Est	SE	p value	Est	SE	p value	Est	SE	p value
Intercept	1.230	0.179	0.000	1.837	0.254	0.000	2.856	0.319	0.000
Gender	− 0.012	0.054	0.826	− 0.273	0.067	0.000	− 0.047	0.093	0.614
Mstat	− 0.525	0.060	0.000	− 0.334	0.081	0.000	− 0.455	0.103	0.000
Education	− 0.111	0.009	0.000	− 0.140	0.012	0.000	− 0.140	0.016	0.000
White	− 0.454	0.144	0.002	− 0.595	0.199	0.003	− 0.288	0.250	0.249
Black	− 0.094	0.154	0.544	− 0.312	0.214	0.146	− 0.143	0.266	0.591
Drink	−0.079	0.054	0.147	− 0.076	0.069	0.272	− 0.127	0.095	0.182
No. of Cond.	0.354	0.024	0.000	0.285	0.034	0.000	0.282	0.041	0.000

	β_{001}			β_{011}		
	Est	SE	p value	Est	SE	p value
Intercept	− 0.038	0.378	0.920	1.187	0.363	0.001
Gender	− 0.139	0.087	0.110	− 0.039	0.109	0.722
Mstat	− 0.213	0.111	0.056	− 0.121	0.126	0.339
Education	− 0.083	0.016	0.000	− 0.092	0.018	0.000
White	− 0.056	0.307	0.855	− 0.092	0.284	0.746
Black	0.155	0.327	0.636	0.060	0.308	0.847
Drink	0.256	0.094	0.007	0.002	0.110	0.989
No. of Cond.	0.175	0.048	0.000	0.320	0.052	0.000

	β_{101}			β_{111}		
	Est	SE	p value	Est	SE	p value
Intercept	1.937	0.581	0.001	1.973	0.350	0.000
Gender	0.032	0.156	0.835	− 0.117	0.121	0.334
Mstat.	− 0.335	0.181	0.064	− 0.346	0.130	0.008
Education	− 0.115	0.028	0.000	− 0.076	0.019	0.000
White	− 0.373	0.435	0.392	0.228	0.280	0.414
Black	− 0.029	0.467	0.950	0.136	0.296	0.647
Drink	− 0.188	0.161	0.245	− 0.192	0.123	0.119
No. of Cond.	0.052	0.070	0.453	0.317	0.053	0.000

6 Conclusion

Majority of the longitudinal models, for example, the GEE and ALR, are based on marginal approaches with an induced correlation among repeated outcomes on one subject and lack in proper specification of the dependence in binary or multivariate repeated outcomes. Naturally, these models may fail to provide an efficient estimation of parameters of the model considered. At this backdrop, this study proposed the usage of two joint models based on marginal conditional approaches as alternatives to GEE or related models based on marginal approaches.

The joint models, (proposed models 1 and 2), take care of the correlation among the repeated measures in a built-in nature and can be extended for any order of dependence without complicating the theory. First of all, the proposed model 1 is an extension of Darlington and Farewell [6] showing the likelihood for models based on the Markovian assumption of first order more explicitly. The second model (proposed model 2) is a further generalization of proposed model 1 based on marginal and conditional models for any order of a Markov chain with covariate dependence. Although the estimates of parameters of the proposed model 1 have less bias and greater coverage probability as compared to the same of GEE or ALR, the proposed model 1 has restricted use due to an overwhelming increase in the number of models and parameters to be estimated when there are more than three observations on a single subject. To overcome these limitations, we suggested the regressive model (Proposed model 2), when a subject is observed more than three times. It might be noted that the biggest advantage of the proposed model 2 is its minimum number of parameters for any order of the underlying Markov Chain. Furthermore, in terms of bias and coverage probability, the proposed model 2 appears to be as good as other alternatives, say, proposed models 1. Hence, for practical reasons, the proposed model 2 can be used to analyze longitudinal data effectively and conveniently for more than three follow-ups. In addition to the simulation study, the applications of the selected models to HRS data [29] show that the proposed model 2 is a more specified model in a simpler setup, as compared to GEE, ALR, the Darlington and Farewell's [6] method or proposed model 1. Nevertheless, in case of more than 3 repeated outcomes, the proposed model 2 is not only the most convenient model but also it performs better than GEE or ALR. Indubitably, both the theoretical and practical users will find the results more useful using the proposed models.

References

1. Azzalini, A.: Logistic regression for autocorrelated data with application to repeated measures. Biometrika **81**(4), 767–775 (1994)
2. Blazer, D.G.: Depression in late life: review and commentary. J. Gerontol. Ser. A **58**(3), 249–265 (2003)
3. Bonney, G.E.: Regressive logistic models for familial disease and other binary traits. Biometrics **42**(3), 611–625 (1986)
4. George, E.B.: Logistic regression for dependent binary observations. Biometrics **43**(4), 951–973 (1987)
5. Carey, V., Zeger, S.L., Diggle, P.: Modelling multivariate binary data with alternating logistic regressions. Biometrika **80**(3), 517–526 (1993)
6. Darlington, G.A., Farewell, V.T.: Binary longitudinal data analysis with correlation a function of explanatory variables. Biom. J. **34**(8), 899–910 (1992)

7. Diggle, P., Heagerty, P., Liang, K.Y., Zeger, S.: Analysis of Longitudinal Data. Oxford University Press, New York (2002)
8. Evans, M., Mottram, P.: Diagnosis of depression in elderly patients. Adv. Psychiatr. Treat. **6**(1), 49–56 (2000)
9. Fitzmaurice, G.M., Laird, N.M., Ware, J.H.: Applied Longitudinal Analysis. Wiley, New Jersey (2012)
10. Fu, L., Hao, Y., Wang, Y.G.: Working correlation structure selection in generalized estimating equations. Comput. Stat. **33**(2), 983–996 (2018)
11. Guerra, M.W., Shults, J., Amsterdam, J., Have, T.T.: The analysis of binary longitudinal data with time-dependent covariates. Stat. Med. **31**(10), 931–948 (2012)
12. Imori, S.: Consistent selection of working correlation structure in GEE analysis based on stein'loss function. Hiroshima Math. J. **45**(1), 91–107 (2015)
13. Islam, M.A., Alzaid, A.A., Chowdhury, R.I., Sultan, K.S.: A generalized bivariate Bernoulli model with covariate dependence. J. Appl. Stat. **40**(5), 1064–1075 (2013)
14. Islam, M.A., Chowdhury, R.I.: Analysis of Repeated Measures Data. Springer, Singapore (2017)
15. Islam, M.A., Chowdhury, R.I.: Prediction of disease status: a regressive model approach for repeated measures. Stat. Methodol. **7**(5), 520–540 (2010)
16. Islam, M.A., Chowdhury, R.I., Alzaid, A.A.: Tests for dependence in binary repeated measures data. J. Stat. Res. **46**(2), 203–217 (2012)
17. Islam, M.A., Chowdhury, R.I., Briollais, L.: A bivariate binary model for testing dependence in outcomes. Bull. Malays. Math. Sci. Soc. **35**(4), 845–858 (2012)
18. Islam, M.A., Chowdhury, R.I.: A higher order Markov model for analyzing covariate dependence. Appl. Math. Model. **30**(6), 477–488 (2006)
19. Karakus, M.C., Patton, L.C.: Depression and the onset of chronic illness in older adults: a 12-year prospective study. J. Behav. Health Serv. Res. **38**(3), 373–382 (2011)
20. Liang, K.Y., Zeger, S.L.: Longitudinal data analysis using generalized linear models. Biometrika **73**(1), 13–22 (1986)
21. Lindsey, J.K., Lambert, P.: On the appropriateness of marginal models for repeated measurements in clinical trials. Stat. Med. **17**(4), 447–469 (1998)
22. Muenz, L.R., Rubinstein, L.V.: Markov models for covariate dependence of binary sequences. Biometrics **41**(1), 91–101 (1985)
23. Nikoloulopoulos, A.K.: Correlation structure and variable selection in generalized estimating equations via composite likelihood information criteria. Stat. Med. **35**(14), 2377–2390 (2016)
24. Carmen Pardo, M., Alonso, R.: Working correlation structure selection in GEE analysis. Stat. Pap. **60**(5), 1447–1467 (2017)
25. Pitt, M.K., Chatfield, C., Walker, S.G.: Constructing first order stationary autoregressive models via latent processes. Scand. J. Stat. **29**(4), 657–663 (2002)
26. Prentice, R.L.: Correlated binary regression with covariates specific to each binary observation. Biometrics **44**(4), 1033–1048 (1988)
27. Shults, J., Sun, W., Xin, T., Kim, H., Amsterdam, J., Hilbe, J.M., TenHave, T.: A comparison of several approaches for choosing between working correlation structures in generalized estimating equation analysis of longitudinal binary data. Stat. Med. **28**(18), 2338–2355 (2009)
28. Steffick, D.E.: Documentation of affective functioning measures in the Health and Retirement Study. An online report accessed on January 18, 2018 from http://hrsonline.isr.umich.edu/sitedocs/userg/dr-005.pdf (2000)
29. University of Michigan. Health and Retirement Study Data. Accessed on January 20, 2018, from http://hrsonline.isr.umich.edu/data/index.html (2014)
30. Gan Wang, Y., Liya, F.: Selection of working correlation structure in generalized estimating equations. Stat. Med. **36**(14), 2206–2219 (2017)
31. Wedderburn, R.W.M.: Quasi-likelihood functions, generalized linear models, and the Gauss–Newton method. Biometrika **61**(3), 439–447 (1974)
32. Zeger, S.L., Liang, K.Y.: Longitudinal data analysis for discrete and continuous outcomes. Biometrics **42**(1), 121–130 (1986)
33. Zeger, S.L., Liang, K.Y., Self, S.G.: The analysis of binary longitudinal data with time-independent covariates. Biometrika **72**(1), 31–38 (1985)