South East Asian J. of Mathematics and Mathematical Sciences
Vol. 15, No. 1 (2019), pp. 71-88

ISSN (Print): 0972-7752

GENERALISED BERNOULLI MODEL FOR CORRELATED BINARY RESPONSES: APPLICATION TO THE NATIONAL INCOME DYNAMICS STUDY (NIDS) DATASETS

Mawutor Fleku, Ezekiel N.N.Nortey, Anani Lotsi and Kwabena Doku-Amponsah

University of Ghana

E-mail: kdoku-amponsah@ug.edu.gh

(Received: November 22, 2017)

Abstract: The bivariate Bernoulli model was used to estimate covariate parameters for conditional as well as marginal models for the NIDs datasets.The covariate parameters were estimated by first expressing the proposed model in the exponential family form, finding the log-likelihood function and then the corresponding estimating equations. The Nelder Mead method of iteration was used to estimate the covariate parameters. The research revealed that the bivariate Bernoulli model fitted bivariate binary response data significantly better than the conditional logistic and the Generalized Estimating Equation (GEE) logistic marginal model. The result was same for both artificial and real-life data.

Keywords and Phrases: Correlated binary responses,longitudinal study, joint modeling, pre and post testing , likelihood ratio test.

2010 Mathematics Subject Classification: 62J12.

1. Introduction

In longitudinal studies, outcomes are normally taken from same subjects over a period of time. In such situations, there is likely to be correlation between the outcomes. The possibility of correlations between repeated outcomes need to be taken into account when analyzing such data. Using standard statistical models and assuming independence for correlated responses may lead to misleading results particularly, estimation of regression parameters. One therefore, needs a statistical model that takes the dependence in response variables into consideration.

To address this problem, many studies have employed the use of marginal models such as Liang and Zeger (1986) and Molenberghs and Leasaffre (1994). A few however, employed the use of conditional models, notable among them are Bonney (1986) and Bonney (1987). However, in all the above studies, it is not easy to specify the measures of dependence in response variable precisely. Furthermore, marginal and conditional models proposed by various authors to tackle the problem of dependence between outcomes may fail to provide efficient population parameter estimates because they do not specify the dependence of binary outcomes in the model, hence the introduction of joint modeling by Islam *et al.*(2012), for bivariate binary response using both the marginal and conditional models. In this paper; a follow-up on that of Islam *et al.* (2012), involving the usefulness of the proposed model as well as the efficiency of the regression coefficient estimates was investigated by using an extensive simulation study and an application to real life data.

2. The Bivariate Bernoulli model

In this section, we propose the model based on the marginal-conditional approach to obtain joint models. In the univariate case, some distributions such as the binomial, Poisson, negative binomial, hypergeometric, gamma distributions and normal distributions originated from the Bernoulli distribution. They are gotten as sums or limits, thereby forming an interrelated family of distributions. (Marshall and Olkin, 1985)

If (X, Y) has Bernoulli marginals then (X, Y) has only four possible values $(0,0)(0,1),(1,0)$,and$(1,1)$. Let

$$P(X,Y) = (0,0) = P_{00} \qquad P(X,Y) = (0,1) = P_{01}$$
$$P(X,Y) = (1,0) = P_{10} \qquad P(X,Y) = (1,1) = P_{11}$$

Islam *et al.*(2013) initiated the proposed model as follows:
The bivariate Bernoulli distribution for outcomes Y_1 and Y_2 can be expressed as

$$P(Y_1 = y_1, Y_2 = y_2) = P_{00}^{(1-y_1)(1-y_2)} P_{01}^{(1-y_1)y_2} P_{10}^{y_1(1-y_2)} P_{11}^{y_1 y_2} \tag{1}$$

The bivariate Bernoulli distribution can be expressed in a 2x2 contingency table as follows:

Table 3.1: 2 x 2 Contingency table for the bivariate Bernoulli distribution

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

The joint probability can be derived from the marginal and conditional and probabilities as:

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

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

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

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

$$P(Y_1 = y_1, Y_2 = y_2) = exp[(1 - y_1)(1 - y_2)logP_{00} + (1 - y_1)y_2logP_{01} + y_1(1 - y_2)log_{10} + y_1y_2logP_{11}] \tag{4}$$

$$= exp[(logP_{00} - y_2logP_{00} - y_1logP_{00} + y_1y_2logP_{00} + y_2logP_{01} - y_1y_2logP_{01} + y_1logP_{10} - y_1y_2logP_{10} + y_1y_2logP_{11}] \tag{5}$$

$$= exp[(y_1logP_{10} - y_1logP_{00} + y_2logP_{01} - y_2logP_{00} + y_1y_2logP_{00} - y_1y_2P_{01} - y_1y_2P_{10} + y_1y_2logP_{11} + logP_{00})] \tag{6}$$

$$P(Y_1 = y_1, Y_2 = y_2) = exp\left[y_1log\left(\frac{P_{10}}{P_{00i}}\right) + y_2log\left(\frac{P_{01}}{P_{00i}}\right) + y_1y_2log\left(\frac{P_{00}P_{11}}{P_{01}P_{10}}\right) + logP_{00}\right] \tag{7}$$

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

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n}\left[y_{1i}log\left(\frac{P_{10i}}{P_{00i}}\right) + y_{2i}log\left(\frac{P_{01i}}{P_{00i}}\right) + y_{1i}y_{2i}log\left(\frac{P_{00i}P_{11i}}{P_{01i}P_{10i}}\right) + logP_{00i}\right] \tag{8}$$

It follows that the components of the link function can be denoted as follows:

$$\eta_0 = (logP_{00}), \eta_1 = log\left(\frac{P_{01}}{P_{00}}\right), \eta_2 = log\left(\frac{P_{10}}{P_{00}}\right), \eta_3 = log\left(\frac{P_{00}P_{11}}{P_{01}P_{10}}\right)$$

where η_0 is the base line link function,η_2 is the link function for Y_1 , η_1 is the link function for Y_2 and η_3 is the link function for dependence between Y_1 and Y_2.
We have demonstrated the probabilities without function of covariates in the previous expressions. Now let us consider $X = (1, X_1, X_2, ..., X_p)$ and $x = (1, x_1, x_2, ..., x_p)$ where $X^* = (1, X_1, X_2, ..., X_p)$ and $x^* = (1, x_1, x_2, ..., x_p)$ are the vector of covariates and their corresponding covariates values. We can now express the conditional probabilities in terms of the logit link functions as follows:

$$P(Y_2 = 1/Y_1 = 0, x) = \frac{e^{x\beta_{01}}}{1 + e^{x\beta_{01}}} = \pi_{01}(x) \tag{9}$$

$$P(Y_2 = 1/Y_1 = 1, x) = \frac{e^{x\beta_{11}}}{1 + e^{x\beta_{11}}} = \pi_{11}(x) \tag{10}$$

$$P(Y_2 = 0/Y_1 = 0, x) = \frac{1}{1 + e^{x\beta_{01}}} = \pi_{00}(x) \tag{11}$$

$$P(Y_2 = 0/Y_1 = 1, x) = \frac{1}{1 + e^{x\beta_{11}}} = \pi_{10}(x) \tag{12}$$

where

$$\beta_{01} = (\beta_{010}, \beta_{011}, \beta_{012}, ..., \beta_{01p})\prime \; and \; \beta_{11} = (\beta_{110}, \beta_{111}, \beta_{112}, ..., \beta_{11p})\prime$$

The marginal probabilities are as follows

$$P(Y_1 = 1/X = x) = \pi_1(x) \; and \; P(Y_1 = 0/X = x) = 1 - \pi_1(x) \tag{13}$$

Now, we assume that

$$P(Y_1 = 1/x) = \frac{e^{x\beta_1}}{1 + e^{x\beta_1}} = \pi_1(x) \; and \; P(Y_1 = 0/x) = \frac{1}{1 + e^{x\beta_1}} = 1 - \pi_1(x) \tag{14}$$

where

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

Also we can write

$$\begin{aligned}
P_{01}(x) &= P(Y_2 = 1/Y_1 = 0, X = x).P(Y_1 = 0/X = x) = \frac{e^{x\beta_{01}}}{1+e^{x\beta_{01}}}.\frac{1}{1+e^{x\beta_1}},\\
P_{00}(x) &= P(Y_2 = 0/Y_1 = 0, X = x).P(Y_1 = 0/X = x) = \frac{1}{1+e^{x\beta_{01}}}.\frac{1}{1+e^{x\beta_1}},\\
P_{11}(x) &= P(Y_2 = 1/Y_1 = 1, X = x).P(Y_1 = 0/X = x) = \frac{e^{x\beta_{11}}}{1+e^{x\beta_{11}}}.\frac{e^{x\beta_1}}{1+e^{x\beta_1}},\\
P_{10}(x) &= P(Y_2 = 0/Y_1 = 1, X = x).P(Y_1 = 1/X = x) = \frac{1}{1+e^{x\beta_{11}}}.\frac{e^{x\beta_1}}{1+e^{x\beta_1}}
\end{aligned} \tag{15}$$

Now we can show that

$$\eta_0(x) = In(P_{00}(x)) = In\left(\frac{1}{1+e^{x\beta_{01}}}.\frac{1}{1+e^{x\beta_1}}\right) = In\left(\frac{1}{1+e^{x\beta_{01}}}\right) + In\left(\frac{1}{1+e^{x\beta_1}}\right)$$

$$\eta_0(x) = -In(1+e^{x\beta_{01}}) - In(1+e^{x\beta_1}); \tag{16}$$

$$\eta_1(x) = In\left(\frac{P_{01}(x)}{P_{00}(x)}\right) = In\left(\frac{e^{x\beta_{01}}}{1+e^{x\beta_{01}}}.\frac{1}{1+e^{x\beta_1}}.1+e^{x\beta_{01}}.1+e^{x\beta_1}\right)$$

$$\eta_1(x) = x\beta_{01}; \tag{17}$$

$$\eta_2(x) = In\left(\frac{1}{1+e^{x\beta_{11}}}.e^{x\beta 1}.1+e^{x\beta_{01}}\right)$$

$$\eta_2(x) = In\left(\frac{1}{1+e^{x\beta_{11}}}\right) + Ine^{x\beta_1} + In(1+e^{x\beta_{01}})$$

$$\eta_2(x) = x\beta_1 + In(1+e^{x\beta_{01}}) - In(1+e^{x\beta_{11}}); \tag{18}$$

$$\eta_3(x) = In\left(\frac{P_{00}(x)P_{11}(x)}{P_{01}(x)P_{10}(x)}\right) = In\left[\frac{\left(\frac{1}{1+e^{x\beta_{01}}}.\frac{1}{1+e^{x\beta_1}}.\frac{e^{x\beta_{11}}}{1+e^{x\beta_{11}}}.\frac{e^{x\beta_1}}{1+e^{x\beta_1}}\right)}{\left(\frac{e^{x\beta_{01}}}{1+e^{x\beta_{01}}}.\frac{1}{1+e^{x\beta_1}}.\frac{1}{1+e^{x\beta_{11}}}.\frac{e^{x\beta_1}}{1+e^{x\beta_1}}\right)}\right]$$

$$\eta_3(x) = In\left(\frac{e^{x\beta_{11}}}{e^{x\beta_{01}}}\right)$$

$$\eta_3(x) = x(\beta_{11} - \beta_{01}) \tag{19}$$

Which indicates that if there is no association between Y_1 and Y_2 then this is true for $\beta_{01} = \beta_{11}$. β_{01} and β_{11} are the estimated covariate coefficients for the conditional logit models ,given covariates X for $Y_1 = 0$ and $Y_2 = 1$, respectively. The assumption underlying the covariates is that they are time independent.
We can test for the overall significance of a model using the likelihood ratio test and the dependence can be examined on the basis of η_3.
This study, came out with a method of estimating the parameters .First, substituting η_0(x) ,η_1(x) ,η_2(x) and η_3(x) into the likelihood equation (8) will give us :

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \Big[- In(1 + e^{x\beta_{01}}) - In(1 + e^{x\beta_1} + y_{1i}(x\beta_1 + In(1 + e^{x\beta_{01}})) \\ -In(1 + e^{x\beta_{11}}) + y_{2i}x\beta_{01} + y_{1i}y_{2i}x(\beta_{11} - \beta_{01}) \Big] \quad (20)$$

The estimating equations will be

$$\frac{\delta l}{\delta \beta_{01}} = \sum_{i=1}^{n} -\frac{X_i e^{X_i\beta_{01}}}{1 + e^{X_i\beta_{01}}} + \frac{Y_{1i}X_i e^{X_i\beta_{01}}}{1 + e^{X_i\beta_{01}}} + Y_{2i}X + Y_{1i}Y_{2i}X_i, \quad (21)$$

$$\frac{\delta l}{\delta \beta_{11}} = \sum_{i=1}^{n} -\frac{X_i e^{X_i\beta_{11}}}{1 + e^{X_i\beta_{11}}} + Y_{1i}Y_{2i}X_i \quad (22)$$

$$\frac{\delta l}{\delta \beta_{1}} = \sum_{i=1}^{n} -\frac{X_i e^{X_i\beta_{1}}}{1 + e^{X_i\beta_{1}}} + Y_{1i}X_i \quad (23)$$

Now, because of the complexity of solving for the estimated values of the parameters, this study adopted the use of the Nelder Mead algorithm. The Nelder Mead algorithm iterates on a simplex and then replaces the worst simplex.

3. Simulation

In this section, simulation is undertaken to verify the usefulness of the proposed model.

3.1. Approach

Simulation was undertaken to investigate the usefulness of the bivariate Bernoulli model. Correlated binary data were generated for simulations, using a technique suggested by Leisch *et al.*(1990, cited in Islam *et al.* 2012, p852) known as "bindata package for R.

Three variables were simulated; two dependent response variables Y_1, Y_2 and one covariate, X. Different correlation combinations between the two response variables and pairwise correlations, that is, their relationship with the covariate, was considered. Simulation was performed 1000 times with different sample sizes (i.e 20, 50 100,500 and 1000).

In generating the required data for simulation, this study used the following inputs:

(i) Marginal probabilities for X , and Y ,

(ii) ρ, this describes the correlation between the response variables.

Specifically, five different sets of models were generated using different marginal probability combinations for and as follows (it is worth stating that the function rmvbin was used in generating correlated binary data. It required the user to provide the marginal probabilities whereas the function estimates its own pairwise probabilities, hence only marginal probabilities were provided):

(i) Low marginal probabilities i.e. $(P_1 = 0.1, P_2 = 0.1)$ with $\rho = 0.2$

(ii) Average marginal probabilities i.e $(P_1 = 0.5, P_2 = 0.5)$ with $\rho = -0.3$

(iii) A high and a low marginal probability i.e $(P_1 = 0.1, P_2 = 0.8)$ with $\rho = 0.1$

(iv) An average and a low marginal probability i.e $(P_1 = 0.5, P_2 = 0.3)$ with $\rho = 0.3$

(v) Above average and below average marginal probabilities i.e $(P_1 = 0.2, P_2 = 0.6)$ with $\rho = 0$

The true pairwise correlations between(Y_1, X) and (Y_2, X) were also shown for each model
The following were extracted to form one complete table:

(i) the average estimates of the parameters from the models,

(ii) the average values of the proposed tests and total number of $p < 0.05$,

(iii) the average values of the marginal and conditional models approach and corresponding number of $p < 0.05$

(iv) A check on the overall fit of the bivariate Bernouli model, the likelihood ratio test was employed. The log-likelihood of the conditional and marginal models were estimated and compared to the log-likelihood of the proposed test.

3.2. Findings

First, exploratory analysis is done to examine the characteristics of the data simulated .Second, with the same data simulated, the conditional logistic and marginal model under GEE are thoroughly examined before the proposed model. Finally, the overall-fit of the proposed model compared to the conditional logistic and the GEE methods are then critically examined.

3.3. Dependence value based on η_3

All but one model had a non- zero value indicating that there exist dependence between the response variables. The model with marginal probability 0.5 for both response variables had a dependence value of approximately 0. Sample size 50 for instance, produced a value of -0.02 whereas sample size of 1000 produced a value of 0.0001.This indicates existence of independence when the data was simulated with marginal probabilities of 0.5 for both response variables. This is rightly so because the marginal probabilities chosen to model the data, add up to one (1) meaning that the pairwise probabilities will be zero, hence the independence.

3.4. Conditional $(Y_1 = 0)$

The fitted conditional model is of the form $logit(\hat{\pi_{01}}) = log\left(\frac{\hat{\pi_{01}}}{1-\hat{\pi_{01}}}\right) = \hat{\alpha} + \hat{\beta_{01}}X$, where $\hat{\pi_{01}}$ represents the probability of recording a 0 response at the first measurement and a 1 at the second measurement $1 - \hat{\pi_{01}}$. represents the probability of not recording a 0 response at the first measurement and a 1 at the second measurement. $\hat{\beta_{01}}$ represents the estimated coefficient parameter for covariate X . $\hat{\alpha}$ is the value of the logit when the covariate is zero.

The simulation exercise revealed that, with a sample size of 50, only a small percentage of the parameter estimates were significant .In addition ,only a small percentage of the models fit the data well after running the simulation for 1000 times. Since we are interested in predicting accurate predictions as much as possible, we need a model that fits the data well enough. A case in point is Model 1 ($P_1 = 0.1, P_2 = 0.1$) which had only 22% of parameter estimates being significant and 33% of all the models run fitted the data well.

As the size of the data grew larger, the percentage of significant estimated parameter also grew bigger as well as the percentage of models that fitted the data well. For instance, with a sample size of 500, Model 4 ($P_1 = 0.5, P_2 = 0.3$) had all its parameter estimates being significant. Moreover, all the models simulated fitted the available data well.

In contrast, Model 2 ($P_1 = 0.5, P_2 = 0.5$) had all estimated parameters being

significant as well as all models fitting the available data. We must be cautious not to jump to conclusion here, because it has been established that Model 2 ($P_1 = 0.5, P_2 = 0.5$) has independent response variables; a possible reason for the perfect fit is as follows:
When binary data are randomly generated, the covariance of the outcome variables will follow the binomial model (i.e. two possible outcomes; an occurrence of an event of interest or non-occurrence) with constant probability. However, when binary data are not sampled randomly then there is the likelihood that the outcome variances will not follow the binomial model. An example is picking samples from clusters, this will result in different probabilities and hence increased variance compared to variances observed under the binomial model. This phenomenon is known as over dispersion and the effects are that , the standard errors and the conclusions might be affected.

3.5. Conditional $(Y_1 = 1)$

The fitted conditional model is of the form $\hat{logit(\pi_{11})} = log\left(\frac{\hat{\pi_{11}}}{1-\hat{\pi_{11}}}\right) = \hat{\alpha} + \hat{\beta_{11}}X.$
Just like the previous conditional case where $(Y_1 = 0)$, as the sample size increased, the percentage of significant parameter coefficient increased .On the same hand, the percentage of models that fitted the available data also increased. Model 3 ($P_1 = 0.1, P_2 = 0.8$) for instance, had only 19% of the models fitting the available data well but by sample 500 , 94% of all models simulated under the stated marginal probabilities fitted the data well. The same model had none of the estimated parameters attaining significance at sample size 50. However, by size 500, 66% of the estimated parameters were significant.
Model 5 ($P_1 = 0.2, P_2 = 0.6$) in general had very low percentage of estimated parameters and well fit models respectively, irrespective of sample size. At sample size 50, none of the estimated parameters was significant and only 5% of the models had a good fit to the simulated data. At sample size 1000, none of the estimated parameters was significant and none of the models had a good fit.

3.6. Marginal model

The fitted marginal model is of the form $\hat{logit(\pi_1)} = log\left(\frac{\hat{\pi_1}}{1-\hat{\pi_1}}\right) = \hat{\alpha} + \hat{\beta}_1 X.$
With sample size of 50, only Model 3 ($P_1 = 0.1, P_2 = 0.8$) had 14% of its parameter coefficient not equal to zero(i.e. using the wald test).All other Models had all the estimated parameter coefficient equal to zero, meaning that they will not serve as a good predictor model for the available data.
For a sample size of 100, none of the models had a non-zero estimated parame-

ter coefficient. However, with a sample size 200, Model 1 ($P_1 = 0.1, P_2 = 0.1$) had 28% of simulated models with a non-zero parameter coefficient, Model 5 ($P_1 = 0.2, P_2 = 0.6$) had 14%,all others equaled zero.
For a sample size of 500, none of the models had a non-zero estimated parameter whereas with a sample size of 1000, only Model 3 ($P_1 = 0.1, P_2 = 0.8$) had 29% of models fitting the available data.
In short, the marginal model was poor at fitting the available data compared to the conditional model.

3.7. Bivariate Bernoulli model

The Nelder-Mead maximization method was used to estimate the actual parameter that maximizes the likelihood function. To make it easier for the selection of initial parameter estimates, the estimates for the conditional and marginal were referred to as to give an idea of where the best fit parameter might fall.
A more important test was that of the overall fit of proposed model.

3.8. Overall fit of the proposed model

The log likelihoods for the conditional as well as the marginal were calculated and then compared to the log likelihood of the proposed model. Results showed that the bivariate Bernoulli model came out as a better model than the rest of the models.
The likelihood ratio test was employed by comparing the fit of the model of the proposed model to that of the conditional and marginal models respectively. A model with more parameters is likely to fit a model better .However, it is important to find out whether the right model fits significantly, hence the use of the test. All tests produced highly significant results i.e. significant at 5% level of significance, meaning that the proposed model fits better.(See Tables A1 to A5 under Appendix A)

4. Application to real-life data

4.1. Approach

This study illustrated the application to real life data by using data from the National Income Dynamics Study (NIDS), a panel study conducted in South Africa. The panel study takes the personal records of 28,000 South Africans starting 2008 to answer policy and research questions. The data helps to put the spotlight on who is getting ahead and who is falling behind and what factors might be contributing to their state. The NIDS data constitute areas such as health, education, labour market and birth history. Two periods were chosen for this study 2008 representing

Wave 1 and 2011 representing Wave 2.The datasets generated and analysed during the current study are available via "http://www.nids.uct.ac.za/nids-data/data-access".

This study focused on the employment status of persons selected for the exercise. The binary response variable measured therefore, was whether an individual was employed at the time of visit. Specifically, (Not employed=0, Employed=1).

Covariates chosen for this exercise include:

(i) Sex of respondent (Male= 0, Female=1)

(ii) Ever been to school (No =0,Yes =1) and

(iii) Age (17-25 years =0, 26-51years = 1)

The covariates chosen were tested to find out whether there is significant impact for conditional, marginal and joint models.Extracts done included finding out the number of persons who had

(i) Not been employed for the two periods,

(ii) not been employed in the first period but employed in the second period,

(iii) been employed in the first period but not in the second

In this section, real-life data was used to examine the usefulness of the proposed model. First, exploratory analysis was conducted.All the models were examined under the same dataset and the overall fit of the proposed model was also examined. The study went ahead to estimate parameters for all the models under study and compared results.

Table 5.1: Transition counts and respective probabilities on the employment status for the two periods

	Wave 2						
	Transition count				**Transition probability**		
Wave 1	0	1	Total		0	1	Total
0	4132	707	4839		0.8539	0.1461	1.000
1	1442	1559	3001		0.4805	0.5195	1.000

Dependence value = 1.84

Table 5.1 displays the transition counts and probabilities on the employment status of persons for the two periods. It is evident that 85.4% of persons between the ages of 17 and 51 interviewed remained unemployed whereas 14.6% gained employment by the second period. Moreover, within the same period, 48.1% moved from being employed to not being employed whereas 52% remained employed.

Dependence value based on η_3

Dependence value of 1.84 clearly indicating dependence between the response variables i.e employment status at the two periods.

Table 5.2a: Fitted models using data from data from NIDS: Traditional models

	Conditional Models						Marginal Model		
	Model $\hat{\beta_{01j}}$			Model $\hat{\beta_{1j}}$			Model $\hat{\beta_{01j}}$-GEE		
Covariates	$\hat{\beta_{01j}}$	se	p-value	$\hat{\beta_{11j}}$	se	p-value	$\hat{\beta_{1j}}$	se	p-value
Constant	-4.115	0.436	$2e^{-16***}$	2.517	0.287	$2e^{-16***}$	1.012	0.200	$4.44\ e^{-07***}$
Age	-0.069	0.081	0.386	-0.750	0.069	$2e^{-16***}$	-1.429	0.053	$2e^{-16***}$
Education	0.826	0.206	$6.38\ e^{-05***}$	0.026	0.136	0.846	0.589	0.090	$5.53\ e^{-11***}$
Sex	0.043	0.085	0.615	-0.310	0.062	$5.73e^{-07**}$	-0.740	0.047	$2e^{-16***}$
Log-likelihood	-2520			-3240			8755.65		

Table 5.2b: Fitted models using data from data from NIDS: Generalized bivariate Bernoulli model

	Generalized bivariate Bernoulli model								
	Model $\hat{\beta_{01j}}$			Model $\hat{\beta_{11j}}$			Model $\hat{\beta_{1j}}$		
Covariates	$\hat{\beta_{01j}}$	se	p-value	$\hat{\beta_{11j}}$	se	p-value	$\hat{\beta_{1j}}$	se	p-value
Constant	0.07	0.00	$<2e^{-16***}$	0.11	0.00	$<2e^{-16***}$	-0.55	0.0004	$<2e^{-16***}$
Age	-5.558	0.005	$<2e^{-16***}$	0.145	0.0008	$<2e^{-16***}$	-4.892	0.0005	$<2e^{-16***}$
Education	0.679	0.002	$<2e^{-16***}$	0.309	0.001	$<2e^{-16***}$	4.663	0.0003	$<2e^{-16***}$
Sex	-4.092	0.004	$<2e^{-16***}$	-2.006	0.002	$<2e^{-16***}$	-3.250	0.0009	$<2e^{-16***}$
Log-likelihood	2892940			2786759			10672276		
LRT	*p< 0.05*			*p< 0.05*			*p< 0.05*		

LRT means Likelihood ratio test between the proposed model and traditional methods

4.3. Traditional models

Table 5.2a provides the following details:

4.3.1. Marginal model-GEE

Using the marginal model , all three covariates were significant , meaning that they contribute significantly to the state of employment status.

The fitted marginal model is of the form $logit(\hat{\pi_1}) = log\left(\frac{\hat{\pi_1}}{1-\hat{\pi_1}}\right) = 1.012 -$

$1.429age + 0.589edu - 740sex$, where $\hat{\pi_1}$ represents the estimated probability of recording a employed response at the second measurement. $1 - \hat{\pi_1}$ represents the estimated probability of not recording an employed response at the second measurement. The estimated intercept is 1.012 representing the estimated logit when age=0,edu=0 and sex=0. This means that the respondent had no age group, no sex and no educational status which does not really make sense in this particular study. The estimated coefficient for the variable age is -1.429 meaning that for respondents who are in age range 17-25 versus those in the age bracket 26-51years, the expected change in the log odds is 1.429,given that education and sex stays constant.
In terms of probabilities, the probability of an individual in the 17-25 year group to be employed at the time of the second visit will be 2.366/(1+2.366) = 0.70 and that of an individual in the age group 26-51 years will be 0.30.
All three covariates were significant at 5% level of significance indicating that if the same sample were run for 100 times, 95 of them will have all three estimated coefficients being significant.
The interpretation for education status and sex of the respondents was left undone, since that is not the main objective for this study.

4.3.2. Conditional model $(Y_1 = 1)$

Given that a respondent at the first time of visit was employed, only two of the covariates produced significant values, i.e age and sex. In other words, only age and sex had significant impact on the probability of being employed at the second time of visit given that a respondent was initially employed.

The fitted marginal model is of the form $\hat{logit(\pi_{11})} = log\left(\frac{\hat{\pi_{11}}}{1-\hat{\pi_{11}}}\right) = 2.517 - 0.750age + 0.026edu - 0.310sex$, where $\hat{\pi_{11}}$ represents the estimated probability of recording an "employed" response at the second measurement, given that the first response recorded "employed" and $1 - \hat{\pi_{11}}$ represents the estimated probability of not recording an "employed" response at the second measurement . The estimated intercept is 2.517 representing the estimated logit when age=0,edu=0 and sex=0. The estimated coefficient for the variable age is -0.750 meaning that for respondents who are in age range 17-25 versus those in the age bracket 26-51years, the expected change in the log odds is 0.750,given that education and sex is constant. In probability terms, the probability of an individual in the 17-25 year group to be employed at the time of the second visit will be 9.327/(1+9.327) =0.90 and that of an individual in the age group 26-51 years will be 0.10.

4.3.3. Conditional model $(Y_1 = 0)$

Given that the respondent was unemployed at the first time of visit, the fitted model will be $logit(\hat{\pi_{01}}) = log\left(\frac{\hat{\pi_{01}}}{1-\hat{\pi_{01}}}\right) = -4.115 - 0.069age + 0.826edu + 0.043sex$.This time, only the variable education turned out be to be significant. $\hat{\pi_{01}}$ represents the estimated probability of recording an employed response at the second measurement, given that the first response recorded "unemployed".$1 - \hat{\pi_{01}}$ represents the estimated probability of not recording "employed" response at the second measurement, given that the first response recorded "unemployed".
The estimated coefficient for the variable education is 0.826 .Also the odds of the group that had never being to school is exp (-4.115-0.069-0.043) = 0.014. This means that the group that had never been to school are 0.01 more likely to employed or the probability of the group that had never been to school to be employed will be (0.014/1+0.014) = 0.01

4.4. Generalized bivariate Bernoulli model

Unifying the marginal and conditional probabilities into a joint distribution as specified by the generalized bivariate Bernoulli model, we can estimate parameters by conditioning on the initial response or using the marginal model. Table 5.2b provides the following:

4.4.1. Marginal

The fitted logistic model is $logit(\hat{\pi_1}) = log\left(\frac{\hat{\pi_1}}{1-\hat{\pi_1}}\right) = -0.55 - 4.892age + 4.663edu - 3.250sex$.The estimated population parameters follow the usual interpretation as stated under the traditional models. All the estimated population parameters were significant with corresponding very small standard errors.
On the odds ratio scale, the odds for a 17-25 year group to be employed is exp (-0.55+4.663-3.250) = 2.370 which means that the persons in the 17-25 year group are 2.37 times more likely to be employed at the second visit than the age group 26-51.
The probability of an individual in the 17-25 year group to be employed at the time of the second visit will be 2.37/ (1+2.37) = 0.70 and that of age group 26-51 years will be 0.30.

4.4.2. Conditional model $(Y_1 = 0)$

Given that a respondent was employed at the first time of visit, the fitted model will be $logit(\hat{\pi_{11}}) = log\left(\frac{\hat{\pi_{11}}}{1-\hat{\pi_{11}}}\right) = 0.11 - 0.851age + 1.186edu - 0.954sex$.Again, all the estimated parameters were significant with very small standard errors and

the estimated population parameters follow the usual interpretation.

4.4.3. Conditional model ($Y_1 = 1$)

Given that a respondent was unemployed at the first time of visit, then by the generalized Bernoulli model,the logistic model will be $logit(\hat{\pi_{01}}) = log\left(\frac{\hat{\pi_{01}}}{1-\hat{\pi_{01}}}\right) = 0.07 - 0.2887age + 0.7805edu - 0.4651sex$.
All the estimated parameters were close to zero.

4.4.4. Overall fit of the model

Even though the proposed model produced estimated parameters with small very small errors it is necessarily to find out how good the proposed model fits the available data in comparison with the traditional methods. The likelihood ratio test showed that the generalized bivariate Bernoulli model fits the available data significantly better at 5% level of significance than the traditional methods.

Table 5.3: Comparison of results: probability of a covariate being employed at the time of second visit

		Conditional				Marginal	
Covariate	Categories	Model $\hat{\beta}_{01j}$	GbB	Model $\hat{\beta}_{11j}$	GbB	Model $\hat{\beta}_{1j}$-GEE	GbB
Age	17-25	0.96	0.97	0.90	0.83	0.70	70
	26-51	0.04	0.03	0.10	0.17	0.30	30
Education	ES	0.99	0.9999	0.81	0.86	0.77	0.82
	NS	0.01	0.00001	0.19	0.14	0.23	0.18
Sex	Male	0.97	0.992	0.86	0.64	0.54	0.69
	Female	0.03	0.008	0.14	0.36	0.46	0.31

GbS-Generalized bivariate Bernoulli, ES-Ever been to school, NS-Never been to school

Table 5.3 shows that the probabilities for the traditional as well as the proposed model were generally similar, but first, under the marginal models, the GEE estimated that the probability of a male employed at the second time of visit will be 0.54 whereas the proposed model estimated a probability of 0.69.
Second, given that a respondent was employed at the first time of study, the conditional logistic model estimated the probability of a male employed at the second time of visit to be 0.86 whereas the proposed model estimated 0.64.

5. Results and discussion

Key findings from the simulation as well as the application to real life exercise are discussed in this section.

5.1. Dependence

The simulation study revealed also that the measure of dependence (η_3) which was derived from the bivariate Bernoulli model was able to identify independence between two response variables when independent response variables were simulated. This is an added plus to the use of the proposed model because it helps to confirm dependence or otherwise of response variables before further tests are carried out.

5.2. Investigation of the traditional models

In this investigation, the marginal model, which represents the probability of obtaining an occurrence of interest with no conditionality to previous responses and the conditional model which says that the probability of an event occurring depends on the previous state of employment status were examined.
The simulation exercise revealed that the conditional models fitted the available data far better than the marginal. This finding is not so surprising because taking into consideration the response in an initial investigation before taking the second response (or saying the probability of an event is conditional on a past response) is likely to result in a better finding than disregarding the first responses.
The measure used to find out how well a model fits the available data is the Wald test. The Wald test determines if the estimated coefficients are simultaneously equal to zero, against at least one estimated coefficient not being zero.

5.3. The Bivariate Bernoulli model

The model proposed for this study unified the marginal and conditional models into one model. This resulted in a joint model which was then used to estimate probabilities. The good thing about this model is that its distribution is uniquely identified. Hence, finding the likelihood function was straight forward compared to the traditional marginal model which does not employ the use of a known distribution but rather uses the probabilities of success and their correlations of the vector of binary responses.

5.3.1. Parameter estimates

All the parameter estimates produced were significant under the bivariate Bernoulli model. This does not in any way prove that the model is good, but it means is that the covariates selected for this study have an impact on the probability of the response variable.

5.3.2. Comparison of results

All three covariates had significant impact under the GEE marginal model and the bivariate Bernoulli model. However, only some of the covariates had significant impact when the conditional logistic models were used.

Specifically, age and sex significantly impacted the employment status when the probability of being employed was conditioned on the respondent being employed at the first time of visit. Also, only education significantly impacted the employment status when the probability of being employed was conditioned on the respondent not being employed at the first time of visit.
In some cases, there seem to be differences in results; for instance, in the probability of being employed at the second time of visit. Under the GEE marginal model, whereas the probability of male getting employed in the second time of visit was 0.54, the bivariate Bernoulli estimated the probability to be 0.69.

5.3.3. Structure of models

It is also important to note that the GEE uses various correlation structures such as "independence", "unstructured", "exchangeable" and "user defined" to estimate covariate coefficients. This might lead to inadequate results because a joint modelling approach is not used. The joint modelling provides explicitly the distribution for the available data, they enhance the fit of the data and hence enhance the efficiency of parameter estimates. This study has shown that with the use of the proposed model, all three covariates have significant impact on the probability of the response variable, irrespective of whether there was conditioning or not .

In conclusion, this study has shown that the bivariate Bernoulli model fits bivariate binary response data significantly better than the GEE marginal model and the conditional logistic models.

References

[1] Bonney, G. E., Regressive logistic models for familial disease and other binary traits, Biometrics, (1986) 611-625.

[2] Bonney, G. E., Logistic regression for dependent binary observations. Biometrics, (1987) 951-973.

[3] Islam, M. A., Alzaid, A. A., Chowdhury, R. I., and Sultan, K. S., A generalized bivariate Bernoulli model with covariate dependence. Journal of Applied Statistics, 40(5), (2013) 1064-1075.

[4] Islam, M. A., Chowdhury, R. I., and Briollais, L., A bivariate binary model for testing dependence in outcomes, Bull. Malays. Math. Sci. Soc.(2), 35(4), (2012) 845-858.a

[5] Leisch, F., Weingessel, A., and Hornik, K., On the generation of correlated artificial binary data (1998).

[6] Liang, K. Y., and Zeger, S. L., Longitudinal data analysis using generalized linear models, Biometrika, (1986) 13-22.

[7] Marshall, A. W., and Olkin, I., A family of bivariate distributions generated by the bivariate Bernoulli distribution, Journal of the American Statistical Association, 80(390), (1985) 332-338

[8] Molenberghs, G., and Lesaffre, E. (1994), Marginal modeling of correlated ordinal data using a multivariate Plackett distribution, Journal of the American Statistical Association, 89(426), 633-644.

[9] Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2008, Wave 1 [dataset], Version 6.0. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2015, Cape Town: DataFirst [distributor], 2015.

[10] Southern Africa Labour and Development Research Unit. National Income Dynamics Study 2010-2011, Wave 2 [dataset], Version 3.0. Cape Town: Southern Africa Labour and Development Research Unit [producer], 2015, Cape Town: DataFirst [distributor], 2015.