MODELING TRIVARIATE CORRELATED BINARY DATA

AHMED MOHAMED MOHAMED EL-SAYED

Department of High Institute for Specific Studies Management Information Systems Nazlet Al-Batran, Giza, Egypt.

ABSTRACT

This paper provides the estimation and the test procedures for the measures of association in the correlated binary data associated with covariates. The generalized linear model (GLM) using the serial dependence and the developed alternative quadratic exponential form (AQEF) procedures are employed for the trivariate binary correlated outcome variables. The log-odds ratios, as measures of association, are estimated and the appropriate tests are suggested. For comparison between the two procedures, we used the simulation study and an application to an ecology problem which involves the estimation of the measures of association and their tests. The over-dispersion criteria is investigated for these procedures. Finally, the deviance and scaled deviance are used as tests for the goodness of fit of the model to determine the best procedure.

Keywords*: Trivariate Bernoulli Distribution; Markov model; Generalized linear model; Deviance; Likelihood ratio test; Maximum likelihood estimators; Alternative quadratic exponential form.*

1 INTRODUCTION

The dependence between the response and the explanatory variables is of interest increasingly in the recent studies specially with correlated outcome variables associated with covariates. A quasi-maximum likelihood estimate, also known as a pseudo-likelihood estimate or a composite likelihood estimate, is an estimate of a parameter θ in a statistical model that is formed by maximizing a function that is related to the logarithm of the likelihood function, but is not equal to it. In contrast, the maximum likelihood estimate maximizes the actual log-likelihood function. In likelihood analysis we must specify the actual form of the distribution. In quasi-likelihood we specify only the relationships between the mean of outcome and covariates, and between the mean and the variance functions. By adopting a quasi-likelihood approach and specifying only the mean-variance structure, we can develop methods that are applicable to several types of outcomes variables. To use the regression model using the quasi-likelihood method, we must use the link function as a transform between the natural parameters and regression parameters. These link functions are different from case to case according to the distribution of correlated outcome variables.

In this study, our focus is on the trivariate case for the correlated binary data because few authors are devoted with the trivariate case. Islam et al. [5] developed a new simple procedure to take account of the bivariate binary model with covariate dependence. This model is based on the integration of conditional and marginal models. Qaqish [7] presented a family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Zhao and Prentice [12] discussed the pseudo-maximum likelihood for analyzing correlated binary responses. Their parametrization is based on a simple pairwise model in which the association between responses is modeled in terms of correlations. Also, Heagerty and Zeger [3], Heagerty [4] presented the conditional log-odds interpretation, and developed a general parametric class of the serial dependence models that permits the likelihood based marginal regression analysis of binary response data. El-Sayed et al. [2] introduced an alternative quadratic exponential form (AQEF), in the bivarite case, to make the quadratic exponential form, which is presented by Zhao and Prentice [12], more realistic in terms of defining the underlying pseudo-likelihood function, by modifying the normalizing procedure in the bivariate case.

In this paper, the major work is modeling the GLM and the AQEF procedures associated with one covariate. These procedure can be extended for more than one covariate without any loss of generality, McCullagh and Nelder [6]. The generalization of the association parameters can be

done with specified link functions for the trivariate correlated binary responses variables. Hence, the bivariate AQEF will be extended to the trivariate case in simple form also by modifying the normalizing process. Also, to compare with the AQEF procedure for the log-odds ratios as measures of association and the regression parameters, we will use the GLM procedure using the serial dependence and the first-order Markov model. Section (2) presents the trivariate Bernoulli distribution, namely the joint probabilities and the log-odds. Sections (3) presents the trivariate AQEF procedure and section (4) presents the trivariate GLM procedure using serial dependence criteria. Each section contains the estimation of natural parameters, the estimation of regression parameters, the testing hypothesis, the goodness of fit of the model and over-dispersion property. Finally, Section (5) displays the numerical examples, using R program, for the simulation study and an application to an ecology problem using the Hunua Ranges Data.

2 Trivariate Bernoulli Distribution

In this section, we will present the joint probability function and the log-likelihood function for three correlated binary variables having the Bernoulli distribution. In this case, we can extend for the trivariate Bernoulli distribution. If Y_1, Y_2 and Y_3 have a Bernoulli marginals, each of which takes the value of either 0 or 1, then it must be that (Y_1, Y_2, Y_3) has only eight possible values $(0,0,0),(0,0,1),(0,1,0),(0,1,1)$ and $(1,0,0),(1,0,1),(1,1,0),(1,1,1)$.

For the trivariate binary data with correlated binary outcomes, the joint mass function is

$$f(y_1, y_2, y_3) = p_{000}^{\prod_{j=1}^{3}(1-y_j)} \times p_{100}^{y_1\prod_{j=2}^{3}(1-y_j)} \times \ldots \times p_{111}^{\prod_{j=1}^{3} y_j}. \quad (1)$$

where, $p_{ijs} = P(Y_1 = i, Y_2 = j, Y_3 = s)$ are the joint probabilities.

The corresponding log-likelihood function of the joint mass function (1) , for n observations, is

$$\ell(y_i; p) = \sum_{i=1}^{n}\prod_{j=1}^{3}(1-y_j)\log p_{000} + y_{1i}\prod_{j=2}^{3}(1-y_j)\log p_{100} + \ldots + \prod_{j=1}^{3} y_j \log p_{111}. \quad (2)$$

Let us define the following parameters using the relationships between the expectaions and both of the marginals p_j, the joint probabilities p_{ij} & p_{ijs} and the covariances σ_{ij} as:

$$\begin{aligned} &p_1 = E(Y_1) \quad p_2 = E(Y_2) \quad p_3 = E(Y_3) \quad q_1 = 1-p_1, \quad q_2 = 1-p_2, \quad q_3 = 1-p_3, \\ &E(Y_1Y_2) = p_{12}, \quad E(Y_1Y_3) = p_{13}, \quad E(Y_2Y_3) = p_{23}, \quad E(Y_1Y_2Y_3) = p_{123}, \\ &\sigma_{12} = p_{12} - p_1p_2, \quad \sigma_{13} = p_{13} - p_1p_3, \quad \sigma_{23} = p_{23} - p_2p_3, \\ &K = E[(Y_1 - p_1)(Y_2 - p_2)(Y_3 - p_3)] = p_{123} - p_{13}p_2 - p_{23}p_1 - p_{12}p_3 + 2p_1p_2p_3; \end{aligned} \quad (3)$$

Note that: if $\sigma_{12} = 0$, then the variables Y_1 and Y_2 are independent, also if $\sigma_{13} = 0$, then the variables Y_1 and Y_3 are independent, and finally if $\sigma_{23} = 0$, then the variables Y_2 and Y_3 are independent. For example, using the expectaion property, in the bivartiate case, we have:

$$E[(Y_1 - p_1)(Y_2 - p_2)] = \sigma_{12} = E(Y_1Y_2) - p_1E(Y_2) - p_2E(Y_1) + p_1p_2$$

$$= p_{12} - 2p_1p_2 + p_1p_2 = p_{12} - p_1p_2$$

Then, Teugels [8] used this property to present the joint probabilities for the three correlated binary variables as:

$$\begin{aligned} p_{000} &= q_1q_2q_3 + q_3\sigma_{12} + q_2\sigma_{13} + q_1\sigma_{23} - K \\ p_{001} &= q_1q_2p_3 + p_3\sigma_{12} - q_2\sigma_{13} - q_1\sigma_{23} + K \\ p_{010} &= q_1p_2q_3 - q_3\sigma_{12} + p_2\sigma_{13} - q_1\sigma_{23} + K \\ p_{011} &= q_1p_2p_3 - p_3\sigma_{12} - p_2\sigma_{13} + q_1\sigma_{23} - K \\ p_{100} &= p_1q_2q_3 - q_3\sigma_{12} - q_2\sigma_{13} + p_1\sigma_{23} + K \\ p_{101} &= p_1q_2p_3 - p_3\sigma_{12} + q_2\sigma_{13} - p_1\sigma_{23} - K \\ p_{110} &= p_1p_2q_3 + q_3\sigma_{12} - p_2\sigma_{13} - p_1\sigma_{23} - K \\ p_{111} &= p_1p_2p_3 + p_3\sigma_{12} + p_2\sigma_{13} + p_1\sigma_{23} + K \end{aligned} \quad (4)$$

The next sections explain the parameters estimation and test procedures for the AQEF and the GLM procedures in the trivariate case as following:

3 Trivariate AQEF Procedure

In this section, we will extend the bivariate Alternative Quadratic Exponential form (AQEF) which is proposed by El-Sayed et al. [2] to the trivariate case. This function reformulates the joint mass function (1) in simple form. So, the joint mass function (1) for the three correlated binary variables Y_1, Y_2 and Y_3 can be written in the alternative quadratic exponential form as:

$$f(y_1,y_2,y_3)=\exp\{y_1\theta_1+y_2\theta_2+y_3\theta_3+y_1y_2\psi_{12}+y_1y_3\psi_{13}+y_2y_3\psi_{23}+y_1y_2y_3\psi_{123}-\log[c(\theta,\psi)]\}, \quad (5)$$

where, $\theta_1=\log\frac{p_1}{1-p_1}$, $\theta_2=\log\frac{p_2}{1-p_2}$,

$\theta_3=\log\frac{p_3}{1-p_3}$ are the natural parameters.

The associated parameters in the function (5) can be written as:

$$\psi_{12}=\log\frac{P(Y_2=1\mid Y_1=1)}{P(Y_2=1\mid Y_1=0)}=\log\frac{P(Y_1=1,\ Y_2=1)\ P(Y_1=0,\ Y_2=0)}{P(Y_1=1,\ Y_2=0)\ P(Y_1=0,\ Y_2=1)},$$

$$\psi_{13}=\log\frac{P(Y_3=1\mid Y_1=1)}{P(Y_3=1\mid Y_1=0)}=\log\frac{P(Y_1=1,\ Y_3=1)\ P(Y_1=0,\ Y_3=0)}{P(Y_1=1,\ Y_3=0)\ P(Y_1=0,\ Y_3=1)}$$

$$\psi_{23}=\log\frac{P(Y_3=1\mid Y_2=1)}{P(Y_3=1\mid Y_2=0)}=\log\frac{P(Y_2=1,\ Y_3=1)\ P(Y_2=0,\ Y_3=0)}{P(Y_2=1,\ Y_3=0)\ P(Y_2=0,\ Y_3=1)},$$

$$\psi_{123}=\frac{P(Y_2=1\mid Y_1=1,\ Y_3=1)}{P(Y_2=1\mid Y_1=0,\ Y_3=1)}=\log\left[\frac{P(Y_1=1,\ Y_2=1,\ Y_3=1)}{P(Y_1=1,\ Y_2=0,\ Y_3=1)}\div\frac{P(Y_1=0,\ Y_2=1,\ Y_3=1)}{P(Y_1=0,\ Y_2=0,\ Y_3=1)}\right].$$

If $P(Y_2\mid Y_1,Y_3)=P(Y_2\mid Y_3)$, this means that Y_1 and Y_2 are conditionally independent, [1], for given Y_3. This can be written as $\rho_{13}.\rho_{23}=\rho_{12}$, where

$$\rho_{jk}=\frac{\sigma_{jk}}{\sqrt{p_jp_kq_jq_k}},\ j,k\in\{1,2,3\},$$

are the correlation coefficients, [8]. To obtain the normalizing term, $c(\theta,\psi)$, in the joint function (5), we can use the probability constraint:

$$\sum_{y_1=0}^{1}\sum_{y_2=0}^{1}\sum_{y_3=0}^{1}f(y_1,y_2,y_3)=1 \quad (6)$$

In this case, the normalizing term can be obtained as

$$c(\theta,\psi)=1+e^{\theta_1}+e^{\theta_2}+e^{\theta_3}+e^{\theta_1+\theta_2+\psi_{12}}+e^{\theta_1+\theta_3+\psi_{13}}+e^{\theta_2+\theta_3+\psi_{23}}+e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}, \quad (7)$$

Simplifying the results of Tuegels [8], for the joint probabilities (4), we can use the joint function (5) to obtain the joint probabilities as:

$$p_{000}=\frac{1}{c(\theta,\psi)},$$

$$p_{001}=\exp(\theta_3-\log[c(\theta,\psi)]),$$

$$p_{010}=\exp(\theta_2-\log[c(\theta,\psi)]),$$

$$p_{011}=\exp(\theta_2+\theta_3+\psi_{23}-\log[c(\theta,\psi)]),$$

$$p_{100}=\exp(\theta_1-\log[c(\theta,\psi)]),$$

$$p_{101}=\exp(\theta_1+\theta_3+\psi_{13}-\log[c(\theta,\psi)]),$$

$$p_{110}=\exp(\theta_1+\theta_2+\psi_{12}-\log[c(\theta,\psi)]),$$

$$p_{111}=\exp(\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}-\log[c(\theta,\psi)]). \quad (8)$$

As shown from the previous equations (8), the joint probabilities can be obtained easily rather than Tuegels [8], in the the equations (4). The next subsection presents the parameters estimation of the AQEF procedures as follows:

3.1 Natural Parameters Estimation

Using the joint mass function (5), the log-likelihood function, for n observations, can be written as

$$\ell(\theta,\psi)=\sum_{i=1}^{n}\{y_{1i}\theta_1+y_{2i}\theta_2+y_{3i}\theta_3+y_{1i}y_{2i}\psi_{12}+y_{1i}y_{3i}\psi_{13}+y_{2i}y_{3i}\psi_{23}+\psi_{123}y_{1i}y_{2i}y_{3i}-\log[c(\theta,\psi)]\}, \quad (9)$$

Where the normalizing term, $c(\theta,\psi)$, is defined as shown in (7). The first derivatives for the log-likelihood function (9) with respect to $\theta_1,\theta_2,\theta_3,\psi_{12},\psi_{13},\psi_{23}$ and ψ_{123}, and put it equal to zero, are

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \theta_1} = \sum_{i=1}^{n} y_{1i} - \sum_{i=1}^{n} \frac{e^{\theta_1} + e^{\theta_1+\theta_2+\psi_{12}} + e^{\theta_1+\theta_3+\psi_{13}} + e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0$$

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \theta_2} = \sum_{i=1}^{n} y_{2i} - \sum_{i=1}^{n} \frac{e^{\theta_2} + e^{\theta_1+\theta_2+\psi_{12}} + e^{\theta_2+\theta_3+\psi_{23}} + e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0$$

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \theta_3} = \sum_{i=1}^{n} y_{3i} - \sum_{i=1}^{n} \frac{e^{\theta_3} + e^{\theta_1+\theta_3+\psi_{13}} + e^{\theta_2+\theta_3+\psi_{23}} + e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0$$

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \psi_{12}} = \sum_{i=1}^{n} y_{1i}y_{2i} - \sum_{i=1}^{n} \frac{e^{\theta_1+\theta_2+\psi_{12}} + e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0$$

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \psi_{13}} = \sum_{i=1}^{n} y_{1i}y_{3i} - \sum_{i=1}^{n} \frac{e^{\theta_1+\theta_3+\psi_{13}} + e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0$$

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \psi_{23}} = \sum_{i=1}^{n} y_{2i}y_{3i} - \sum_{i=1}^{n} \frac{e^{\theta_2+\theta_3+\psi_{23}} + e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0$$

$$\frac{\partial \ell(y_i;\theta,\psi)}{\partial \psi_{123}} = \sum_{i=1}^{n} y_{1i}y_{2i}y_{3i} - \sum_{i=1}^{n} \frac{e^{\theta_1+\theta_2+\theta_3+\psi_{12}+\psi_{13}+\psi_{23}+\psi_{123}}}{c(\theta,\psi)} = 0 \tag{10}$$

Solving the estimating equations (10), numerically, we have the estimates of natural parametrs, $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\psi}_{12}, \hat{\psi}_{13}, \hat{\psi}_{23}$ and $\hat{\psi}_{123}$. Then, we can use these estimates to obtain the estimates of the joint probabilities in the equations (8). The number of natural parameters in this procedure, AQEF, are 14 parameters.

3.2 Regression Parameters Estimation

We can use the next link functions, which are used to transform the natural paprameters to the regression parameters, to specify the regression model as a function of regression parameters as following:

$$\theta_1 = \beta_1'x, \quad \beta_1' = (\beta_0 \quad \beta_1), \quad b_1 = \frac{1+e^{\beta_1'x}}{e^{\beta_1'x}},$$

$$\theta_2 = \beta_2'x, \quad \beta_2' = (\beta_0 \quad \beta_2), \quad b_2 = \frac{1+e^{\beta_2'x}}{e^{\beta_2'x}}, \tag{11}$$

$$\theta_3 = \beta_3'x, \quad \beta_3' = (\beta_0 \quad \beta_3), \quad b_3 = \frac{1+e^{\beta_3'x}}{e^{\beta_3'x}}, \quad x_1 = 1$$

Then, the joint function (5), using the regression parameters, is become:

$$f(y_1,y_2,y_3 \mid x) = \exp\{y_1\beta_1'x + y_2\beta_2'x + y_3\beta_3'x + y_1y_2\alpha_1'x + y_1y_3\alpha_2'x + y_2y_3\alpha_3'x + y_1y_2y_3\alpha_4'x - \log[c(\beta,\alpha)]\}, \tag{12}$$

where,

$$c(\beta,\alpha) = 1 + e^{\beta_1'x} + e^{\beta_2'x} + e^{\beta_3'x} + e^{\beta_1'x+\beta_2'x+\alpha_1'x} + e^{\beta_1'x+\beta_3'x+\alpha_2'x} + e^{\beta_2'x+\beta_3'x+\alpha_3'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}.$$

Consequently, the log-likelihood function, for n observations, can be expressed as:

$$\ell(\beta,\alpha) = \sum_{i=1}^{n}\{y_{1i}\beta_1'x + y_{2i}\beta_2'x + y_{3i}\beta_3'x + y_{1i}y_{2i}\alpha_1'x + y_{1i}y_{3i}\alpha_2'x + y_{2i}y_{3i}\alpha_3'x + y_{1i}y_{2i}y_{3i}\alpha_4'x - \log[c(\beta,\alpha)]\}, \quad (13)$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\beta_1} = \sum_{i=1}^{n}(y_{1i} - \frac{e^{\beta_1'x} + e^{\beta_1'x+\beta_2'x+\alpha_1'x} + e^{\beta_1'x+\beta_3'x+\alpha_2'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\beta_2} = \sum_{i=1}^{n}(y_{2i} - \frac{e^{\beta_2'x} + e^{\beta_1'x+\beta_2'x+\alpha_1'x} + e^{\beta_2'x+\beta_3'x+\alpha_3'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\beta_3} = \sum_{i=1}^{n}(y_{3i} - \frac{e^{\beta_3'x} + e^{\beta_1'x+\beta_3'x+\alpha_2'x} + e^{\beta_2'x+\beta_3'x+\alpha_3'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\alpha_1} = \sum_{i=1}^{n}(y_{1i}y_{2i} - \frac{e^{\beta_1'x+\beta_2'x+\alpha_1'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\alpha_2} = \sum_{i=1}^{n}(y_{1i}y_{3i} - \frac{e^{\beta_1'x+\beta_3'x+\alpha_2'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\alpha_3} = \sum_{i=1}^{n}(y_{2i}y_{3i} - \frac{e^{\beta_2'x+\beta_3'x+\alpha_3'x} + e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0$$

$$\frac{\partial\ell(\beta,\alpha)}{\partial\alpha_4} = \sum_{i=1}^{n}(y_{1i}y_{2i}y_{3i} - \frac{e^{\beta_1'x+\beta_2'x+\beta_3'x+\alpha_1'x+\alpha_2'x+\alpha_3'x+\alpha_4'x}}{c(\beta,\alpha)})x = 0 \quad (14)$$

The first derivative for the log-likelihood function (13) with respect to $\beta_1,\beta_2,\beta_3,\alpha_1,\alpha_2,\alpha_3$ and α_4 and putting these equal to zero, respectively, we have estimation equations:

Solving the equations (14), numerically, we get the vectors of estimates $\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3$ and $\hat{\alpha}_4$. Then, using the equations (11), we have the estimates of natural parameters $\hat{\theta}_1,\hat{\theta}_2,\hat{\theta}_3,\hat{\psi}_{12},\hat{\psi}_{13},\hat{\psi}_{23}$ and $\hat{\psi}_{123}$. The number of regression parameters in this procedure, AQEF, are 14 parameters.

3.3 Testing Hypothesis for Regression Parameters

The Likelihood ratio test (LRT) can be used to test the regression parameters. Approximately LRT following Chi-square distribution with one degree of freedom.

We will use the LRT to test the null hypothesis $H_0:\alpha_1 = 0$ against the alternative hypothesis $H_1:\alpha_1 \neq 0$. The LRT can be written as:

$$LRT = -2\left[\ell(y_i;\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\beta}_3,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4) - \ell(y_i;\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4)\right] \quad : \quad \chi_1^2 \tag{15}$$

Also, we can use the LRT to test the null hypothesis $H_0 : \alpha_2 = 0$ against the alternative hypothesis $H_1 : \alpha_2 \neq 0$. The LRT can be written as:

$$LRT = -2\left[\ell(y_i;\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\beta}_3,\tilde{\alpha}_1,\tilde{\alpha}_3,\tilde{\alpha}_4) - \ell(y_i;\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4)\right] \quad : \quad \chi_1^2 \tag{16}$$

Similarly, we can use the LRT to test the null hypothesis $H_0 : \alpha_3 = 0$ against the alternative hypothesis $H_1 : \alpha_3 \neq 0$. The LRT can be written as:

$$LRT = -2\left[\ell(y_i;\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\beta}_3,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_4) - \ell(y_i;\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4)\right] \quad : \quad \chi_1^2 \tag{17}$$

Finally, we can use the LRT to test the null hypothesis $H_0 : \alpha_4 = 0$ against the alternative hypothesis $H_1 : \alpha_4 \neq 0$. The LRT can be written as:

$$LRT = -2\left[\ell(y_i;\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\beta}_3,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3) - \ell(y_i;\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\alpha}_1,\hat{\alpha}_2,\hat{\alpha}_3,\hat{\alpha}_4)\right] \quad : \quad \chi_1^2 \tag{18}$$

The estimates of $\tilde{\beta}_1,\tilde{\beta}_2,\tilde{\beta}_3,\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\alpha}_4$ under H_0 can be obtained by solving the equations (14) numerically.

3.4 Goodness of Fit of Model

The deviance can be used to determine the goodness of fit of the model. We can define the deviance function as:

$$D(y_i;\hat{\theta},\hat{\psi}) = 2\left[\ell(y_i,y_i) - \ell(y_i,\hat{\theta},\hat{\psi})\right] \quad : \quad \chi^2_{n-p} \tag{19}$$

where p is the number of estimated regression parameters,

$$\ell(y_i;\hat{\theta},\hat{\psi}) = \sum_{i=1}^{n}\left(y_{1i}\hat{\theta}_1 + y_{2i}\hat{\theta}_2 + y_{3i}\hat{\theta}_3 + y_{1i}y_{2i}\hat{\psi}_{12} + y_{1i}y_{3i}\hat{\psi}_{13} + y_{2i}y_{3i}\hat{\psi}_{23} + y_{1i}y_{2i}y_{3i}\hat{\psi}_{123} - \log[c(\hat{\theta},\hat{\psi})]\right),$$

is the log-likelihood function as a function of the natural parameters,

$$c(\hat{\theta},\hat{\psi}) = 1 + e^{\hat{\theta}_1} + e^{\hat{\theta}_2} + e^{\hat{\theta}_3} + e^{\hat{\theta}_1+\hat{\theta}_2+\hat{\psi}_{12}} + e^{\hat{\theta}_1+\hat{\theta}_3+\hat{\psi}_{13}} + e^{\hat{\theta}_2+\hat{\theta}_3+\hat{\psi}_{23}} + e^{\hat{\theta}_1+\hat{\theta}_2+\hat{\theta}_3+\hat{\psi}_{12}+\hat{\psi}_{13}+\hat{\psi}_{23}+\hat{\psi}_{123}},$$

is the normalizing term as a functtion of natural parameters,

$$\ell(y_i;y_i) = \sum_{i=1}^{n}\left(y_{1i} + y_{2i} + y_{3i} + y_{1i}y_{2i} + y_{1i}y_{3i} + y_{2i}y_{3i} + y_{1i}y_{2i}y_{3i} - \log[c(y_i,y_i)]\right),$$

is the log likelihood function as a function of y.

and

$$c(y_i,y_i) = 1 + e^{y_{1i}} + e^{y_{2i}} + e^{y_{3i}} + e^{y_{1i}+y_{2i}+y_{1i}y_{2i}} + e^{y_{1i}+y_{3i}+y_{1i}y_{3i}} + e^{y_{2i}+y_{3i}+y_{2i}y_{3i}} + e^{y_{1i}+y_{2i}+y_{3i}+y_{1i}y_{2i}+y_{1i}y_{3i}+y_{2i}y_{3i}+y_{1i}y_{2i}y_{3i}}$$

is the normalizing term as a functtion of binary data y.

3.5 Over-Dispersion Criteria

The over-dispersion is happened when Var (Y) > Var (μ). So, the over-dispersion parameter ϕ can be obtained from the relation $Var(Y) = \phi Var(\mu)$. The estimation of dispersion parameter, ϕ, can be used as a good measure for the over-dispersion criteria.

So, let us define:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}, \quad \hat{p} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \hat{p}_3 \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} Var(Y_1) & Cov(Y_1,Y_2) & Cov(Y_1,Y_3) \\ Cov(Y_2,Y_1) & Var(Y_2) & Cov(Y_2,Y_3) \\ Cov(Y_3,Y_1) & Cov(Y_3,Y_2) & Var(Y_3) \end{pmatrix},$$

The quantity $(Y-\hat{p})'\hat{\Sigma}^{-1}(Y-\hat{p})$ follows the non-central χ^2 distribution. Under independence, the estimator of dispersion parameter ϕ is

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j=1}^{3}\frac{(y_j-\hat{p}_j)^2}{Var(\hat{p}_j)}, \quad (20)$$

where p is the number of estimated regression parameters and $\hat{p}_j = \frac{e^{\hat{\beta}_j'x}}{1+e^{\hat{\beta}_j'x}}$, $j=1,2,3$ are the estimated marginals.

The value of $\hat{\phi}$ is close to 1, for the Bernoulli data, may reflect absence of over-dispersion.

Also, we can use the scaled deviance function

$$Scaled\ D. = \frac{D(y_i;\hat{\theta},\hat{\psi})}{\hat{\phi}}, \quad (21)$$

as a measure of the goodness of fit of the model.

The lower value is good, and surely it is better than the deviance function (19) and both of them equals when $\hat{\phi}$=1.

4 Trivariate GLM Procedure

In this section, we will use the serial dependence property using the first-order Markov model. According to the conditional logs-odds interpretation of Heagerty and Zeger [3], Heagerty [4].

The conditional probability of $(Y_2 = y_2)$ given that $(Y_1 = y_1)$ is:

$$P(Y_2 = y_2 \mid Y_1 = y_1) = \left[\frac{e^{\theta_2 y_2}}{1+e^{\theta_2}}\right] \times \left[\frac{1+e^{\theta_2}}{1+e^{\theta_2+\psi_{12}}}\right]^{y_1} \times e^{\psi_{12}\, y_1 y_2}, \quad (22)$$

Also, the conditional probability of $(Y_3 = y_3)$ given that $(Y_2 = y_2)$ is:

$$P(Y_3 = y_3 \mid Y_2 = y_2) = \left[\frac{e^{\theta_3 y_3}}{1+e^{\theta_3}}\right] \times \left[\frac{1+e^{\theta_3}}{1+e^{\theta_3+\psi_{23}}}\right]^{y_2} \times e^{\psi_{23}\, y_2 y_3}, \quad (23)$$

Using the equations (22), (23) and the following serial dependence relationship:

$$f(y_1,y_2,y_3) = P(Y_3 = y_3 \mid Y_2 = y_2) \times P(Y_2 = y_2 \mid Y_1 = y_1) \times P(Y_1 = y_1) \quad (24)$$

we can obtain the joint mass function for the correlated binary variables, Y_1, Y_2 and Y_3, in the exponential family form, as:

$$\begin{aligned} f(y_1,y_2,y_3) = \exp\{&\theta_1 y_1 + \theta_2 y_2 + \theta_3 y_3 + \psi_{12}\, y_1 y_2 + \psi_{23}\, y_2 y_3 - \log[1+e^{\theta_1}] - \log[1+e^{\theta_2}] \\ &- \log[1+e^{\theta_3}] - y_1(\log[1+e^{\theta_2+\psi_{12}}] - \log[1+e^{\theta_2}]) \\ &- y_2(\log[1+e^{\theta_3+\psi_{23}}] - \log[1+e^{\theta_3}])\} \end{aligned} \quad (25)$$

4.1 Natural Parameters Estimation

In this section, we will present the estimation of parameters of the trivariate Bernoulli model.

For n observations, we can get the log-likelihood function as:

$$\begin{aligned} \ell(\theta,\psi) = \sum_{i=1}^{n}\{&\theta_1 y_{1i} + \theta_2 y_{2i} + \theta_3 y_{3i} + \psi_{12}\, y_{1i} y_{2i} + \psi_{23}\, y_{2i} y_{3i} - \log[1+e^{\theta_1}] - \log[1+e^{\theta_2}] \\ &- \log[1+e^{\theta_3}] - y_{1i}(\log[1+e^{\theta_2+\psi_{12}}] - \log[1+e^{\theta_2}]) \\ &- y_{2i}(\log[1+e^{\theta_3+\psi_{23}}] - \log[1+e^{\theta_3}])\} \end{aligned} \quad (26)$$

Taking the first order derivatives for (26) with respect to $\theta_1, \theta_2, \theta_3, \psi_{12}$ and ψ_{23}, and putting equal to zero, we have the estimating equations:

$$\frac{\partial \ell(\theta,\psi)}{\partial \theta_1} = \sum_{i=1}^{n}(y_{1i} - \frac{e^{\theta_1}}{1+e^{\theta_1}}) = 0$$

$$\frac{\partial \ell(\theta,\psi)}{\partial \theta_2} = \sum_{i=1}^{n}(y_{2i} - \frac{e^{\theta_2}}{1+e^{\theta_2}}) - \sum_{i=1}^{n} y_{1i}(\frac{e^{\theta_2+\psi_1}}{1+e^{\theta_2+\psi_1}} - \frac{e^{\theta_2}}{1+e^{\theta_2}}) = 0$$

$$\frac{\partial \ell(\theta,\psi)}{\partial \theta_3} = \sum_{i=1}^{n}(y_{3i} - \frac{e^{\theta_3}}{1+e^{\theta_2}}) - \sum_{i=1}^{n} y_{2i}(\frac{e^{\theta_3+\psi_3}}{1+e^{\theta_3+\psi_3}} - \frac{e^{\theta_3}}{1+e^{\theta_3}}) = 0$$

(27)

$$\frac{\partial \ell(\theta,\psi)}{\partial \psi_1} = \sum_{i=1}^{n} y_{1i}y_{2i} - \sum_{i=1}^{n} y_{1i}(\frac{e^{\theta_2+\psi_1}}{1+e^{\theta_2+\psi_1}}) = 0$$

$$\frac{\partial \ell(\theta,\psi)}{\partial \psi_3} = \sum_{i=1}^{n} y_{2i}y_{3i} - \sum_{i=1}^{n} y_{2i}(\frac{e^{\theta_3+\psi_3}}{1+e^{\theta_3+\psi_3}}) = 0$$

Solving the estimating equations (27), numerically, we have the estimates $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\psi}_1$ and $\hat{\psi}_3$. The number of natural parameters in this procedure, GLM, are 10 parameters.

4.2 Regression Parameters Estimation

We need to study the effect of covariates on the joint function (25) which can be expressed as

$$\begin{aligned} f(y_1, y_2, y_3 \mid x) = \exp\{&\beta_1' x_1 + \beta_2' x_2 + \beta_3' x_3 + \alpha_1' x_1 y_2 \\ &+ \alpha_3' x_2 y_3 - \log[1 + e^{\beta_1' x}] - \log[1 + e^{\beta_2' x}] \\ &- \log[1 + e^{\beta_3' x}] - y_1(\log[1 + e^{\beta_2' x + \alpha_1' x}] - \log[1 + e^{\beta_2' x}]) \\ &- y_2(\log[1 + e^{\beta_3' x + \alpha_3' x}] - \log[1 + e^{\beta_3' x}])\} \end{aligned} \quad (28)$$

where, the natural parameters $(\theta_1, \theta_2$ and $\theta_3)$ and the regression parameters $(\beta_1, \beta_2, \beta_3, \alpha_1$ and $\alpha_3)$

are defined as shown in the equations (11). For n observations, we can get the log-likelihood function as

$$\begin{aligned} \ell(\theta,\psi) = \sum_{i=1}^{n}\{&\beta_1' x_{1i} + \beta_2' x_{2i} + \beta_3' x_{3i} + \alpha_1' x_{1i} y_{2i} \\ &+ \alpha_3' x_{2i} y_{3i} - \log[1 + e^{\beta_1' x}] - \log[1 + e^{\beta_2' x}] \\ &- \log[1 + e^{\beta_3' x}] - y_{1i}(\log[1 + e^{\beta_2' x + \alpha_1' x}] - \log[1 + e^{\beta_2' x}]) \\ &- y_{2i}(\log[1 + e^{\beta_3' x + \alpha_3' x}] - \log[1 + e^{\beta_3' x}])\} \end{aligned} \quad (29)$$

Taking the first derivatives for (29) with respect to $\beta_1, \beta_2, \beta_3, \alpha_1$ and α_3, and putting these equal to zero, we have

$$\frac{\partial \ell(\beta,\alpha)}{\partial \beta_1} = \sum_{i=1}^{n}(y_{1i} - \frac{e^{\beta_1' x}}{1+e^{\beta_1' x}})x = 0$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \beta_2} = \sum_{i=1}^{n}\Big(y_{2i} - \frac{e^{\beta_2' x}}{1+e^{\beta_2' x}} - y_{1i}(\frac{e^{\beta_2' x+\alpha_1' x}}{1+e^{\beta_2' x+\alpha_1' x}} - \frac{e^{\beta_2' x}}{1+e^{\beta_2' x}})\Big)x = 0$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \beta_3} = \sum_{i=1}^{n}\Big(y_{3i} - \frac{e^{\beta_3' x}}{1+e^{\beta_3' x}} - y_{2i}(\frac{e^{\beta_3' x+\alpha_3' x}}{1+e^{\beta_3' x+\alpha_3' x}} - \frac{e^{\beta_3' x}}{1+e^{\beta_3' x}})\Big)x = 0 \tag{30}$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \alpha_1} = \sum_{i=1}^{n}(y_{1i}y_{2i} - y_{1i}\frac{e^{\beta_2' x+\alpha_1' x}}{1+e^{\beta_2' x+\alpha_1' x}})x = 0$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \alpha_3} = \sum_{i=1}^{n}(y_{2i}y_{3i} - y_{2i}\frac{e^{\beta_3' x+\alpha_3' x}}{1+e^{\beta_3' x+\alpha_3' x}})x = 0$$

Solving the equations (30), numerically, we get the estimates $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\alpha}_1$ and $\hat{\alpha}_3$.

The number of regression parameters in this procedure, GLM, are 10 parameters.

4.3 Testing Hypothesis for Regression Parameters

We can use the LRT to test the null hypothesis $H_0 : \alpha_1 = 0$ against the alternative hypothesis $H_1 : \alpha_1 \neq 0$.

The LRT test can be written as:

$$LRT = -2[\ell(y_i; \widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\beta}_3, \widetilde{\alpha}_3) - \ell(y_i; \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\alpha}_1, \hat{\alpha}_3)] \quad : \quad \chi_1^2 \tag{31}$$

Also, we can use the LRT to test the null hypothesis $H_0 : \alpha_3 = 0$ against the alternative hypothesis $H_1 : \alpha_3 \neq 0$.

The LRT test can be written as:

$$LRT = -2[\ell(y_i; \widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\beta}_3, \widetilde{\alpha}_1) - \ell(y_i; \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\alpha}_1, \hat{\alpha}_3)] \quad : \quad \chi_1^2 \tag{32}$$

The estimates $\widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\beta}_3, \widetilde{\alpha}_1, \widetilde{\alpha}_3$ under H_0, can be obtained by solving the equations (30) numerically.

4.4 Goodness of Fit of Model

The deviance can be used to determine the goodness of fit of the model. So, we can define the deviance function as

$$D(y_i; \hat{\theta}, \hat{\psi}) = -2[\ell(y_i, y_i) - \ell(y_i, \hat{\theta}, \hat{\psi})] \quad : \quad \chi^2_{n-p}, \tag{33}$$

where,

$$\ell(y_i, y_i) = \sum_{i=1}^{n}\{y_{1i} + y_{2i} + y_{3i} + y_{1i}y_{2i} + y_{2i}y_{3i} - \log[1 + e^{y_{1i}}] - \log[1 + e^{y_{2i}}] - \log[1 + e^{y_{3i}}]$$

and $y_{1i}(\log[1 + e^{y_{2i}+y_{1i}y_{2i}}] - \log[1 + e^{y_{2i}}]) - y_{2i}(\log[1 + e^{y_{3i}+y_{2i}y_{3i}}] - \log[1 + e^{y_{3i}}])\}$,

$$\ell(y_i;\hat{\theta},\hat{\psi}) = \sum_{i=1}^{n}\{\hat{\theta}_1 y_{1i} + \hat{\theta}_2 y_{2i} + \hat{\theta}_3 y_{3i} + \hat{\psi}_1 y_{1i}y_{2i} + \hat{\psi}_2 y_{2i}y_{3i} - \log[1 + e^{\hat{\theta}_1}] - \log[1 + e^{\hat{\theta}_2}]$$
$$- \log[1 + e^{\hat{\theta}_3}] - y_{1i}(\log[1 + e^{\hat{\theta}_2+\hat{\psi}_1}] - \log[1 + e^{\hat{\theta}_2}]) - y_{2i}(\log[1 + e^{\hat{\theta}_3+\hat{\psi}_2}] - \log[1 + e^{\hat{\theta}_3}])\}.$$

Also, the scaled deviance as a goodness of fit of the model can be used, using the equation (21).

4.5 Over-Dispersion Criteria

The estimator of dispersion parameter ϕ can be used as a good measure of the over-dispersion as shown in (20).

In the next section, we will use numerical examples to explain the differences between the AQEF and GLM procedures, using the R program, for the simulation and application studies.

5 Numerical Examples

In this section, we have two subsections, the first one explains the simulation study using the generation of multivariate binary data, and the second one demonstrates the application study using the Hunua Ranges Data on an ecological field ,Yee [10,11].

5.1 Simulation Study

In the simulation study, we use the binarySimCLF package of the R program to generate the multivariate binary data with exchangeable correlation matrix with parameter value , $\rho = 0.25$, and the marginals $(p_1 = 0.30, p_2 = 0.40, p_3 = 0.20, p_4 = 0.30)$. The first three columns from the generated data are specified to the correlated binary responses, (Y_1, Y_2 and Y_3). The fourth columns is specified to the explanatory variable X . In this study, we will use large sample size, $n = 500$.

The estimates in Table (1) are obtained for the GLM and AQEF procedures, using the BB-package of R program [9].

Table (1) explains the results for the AQEF and GLM procedures as follows:

Hence, the LRTs will be compared with $\chi^2(0.05,1) = 3.8415$.

So, we can summarize the results from Table (1) as following:

Table 1. Results of the AQEF and GLM procedures

Estimate	AQEF	GLM	Estimate	AQEF	GLM
$\hat{p}_1$	0.2281	0.3080	$\hat{\alpha}_0$	0.7762	0.7949
$\hat{p}_2$	0.3357	0.3605	$\hat{\alpha}_1$	-0.4986	-0.3169
$\hat{p}_3$	0.1130	0.1608	$\hat{\alpha}_2$	1.0296	-
$\hat{\beta}_0$	-1.7228	-1.2417	$\hat{\alpha}_3$	-0.3511	-
$\hat{\beta}_1$	1.3319	1.2417	$\hat{\alpha}_0$	0.8621	0.6785
$\hat{\beta}_2$	-0.9193	-0.7949	$\hat{\alpha}_1$	-0.3803	0.0268
$\hat{\beta}_3$	0.7251	0.6896	$\hat{\alpha}_4$	-0.3513	-
$\hat{\beta}_0$	-2.2336	-1.8099	$\hat{\alpha}_4$	0.7024	-
$\hat{\beta}_1$	0.4990	0.4665	$\hat{\phi}$	3.0425	2.1492
Scaled Deviance				28.3746	14.0415
Log likelihood Value				-847.8763	-854.743
LRT ($H_0: \alpha_1 = 0$)				8.0801	11.3126
LRT ($H_0: \alpha_2 = 0$)				5.9868	-
LRT ($H_0: \alpha_3 = 0$)				6.0343	17.5674
LRT ($H_0: \alpha_4 = 0$)				0.5107	-

For the AQEF procedure:

The LRTs demonstrate significant association between the correlated binary pairwise variables, associated with explanatory variable, X. However, there is no significant association between all the correlated binary variables. This indicates that there is a pair-wise first-order dependence between the pairs of binary variables but the log-odds ratios for three Bernoulli variables demonstrates that there is no second order association among the three Bernoulli outcome variables. The estimate of dispersion parameter reflects the over-dispersion case ($\hat{\phi} = 3.0425 > 1$).

The scaled deviance reflects the goodness of fit of the model,

[Scaled deviance $= 28.3746 < \chi^2(0.05, n - p = 486) = 538.393$, p =14].

For the GLM procedure:

The LRTs demonstrate significant association between the binary variables Y_1 and Y_2 and significant association between the binary variables Y_2 and Y_3, both are associated with explanatory X. The estimate of dispersion parameter reflects the over-dispersion case ($\hat{\phi} = 2.1492 > 1$).

The scaled deviance also reflects the goodness of fit of the model,

[Scaled deviance $= 14.0415 < \chi^2(0.05, n - p = 490) = 542.604$, p =10].

The regression models are shown below as follows:

For the AQEF procedure, we have the regression model

$$\begin{aligned}
\log\frac{p_{1i}}{1-p_{1i}} &= -1.7228 + 1.3319x_i \\
\log\frac{p_{2i}}{1-p_{2i}} &= -0.9193 + 0.7251x_i \\
\log\frac{p_{3i}}{1-p_{3i}} &= -2.2336 + 0.4990x_i \\
\psi_{12i} &= 0.7762 - 0.4986x_i \\
\psi_{13i} &= 1.0296 - 0.3511x_i \\
\psi_{23i} &= 0.8621 - 0.3803x_i \\
\psi_{123i} &= -0.3513 + 0.7024x_i
\end{aligned} \tag{34}$$

Similarly, for the GLM procedure, we have the regression model

$$\begin{aligned}
\log\frac{p_{1i}}{1-p_{1i}} &= -1.2417 + 1.2417x_i \\
\log\frac{p_{2i}}{1-p_{2i}} &= -0.7949 + 0.6896x_i \\
\log\frac{p_{3i}}{1-p_{3i}} &= -1.8099 + 0.4665x_i \\
\phi_{12i} &= 0.7949 - 0.3169x_i \\
\phi_{23i} &= 0.6785 + 0.0268x_i
\end{aligned} \tag{35}$$

5.2 Application Study

Source: R-program (Dr Neil Mitchell, University of Auckland) and Yee [10,11].

These data were collected from the Hunua Ranges, a small forest in southern Auckland, New Zealand. At 392 sites in the forest, the presence/absence of 17 plant species was recorded, as well as the altitude. Each site was of area size 200 m^2. The Hunua Ranges Data frame has (392) rows and (18) columns. The Altitude represents the the continuous independent column, and the (cyadea, beitaw and kniexc) columns are correlated binary responses (presence=1, absence =0) for the 17 plant species. For these data, we use the columns (cyadea, beitaw and kniexc) as the dependent correlated binary variables Y_1, Y_2 and Y_3 respectively. On the other hand, we will use the column altitude (meters above sea level), as the continuous explanatory variable X.

The estimates of the regression parameters in both the procedures, as explained in Table (2), can be obtained by solving the estimating equations using the BB-package in R program [9].

Table (2) explains the results for the AQEF and GLM procedures as follows:

Hence, the LRT's will be compared with $\chi^2(0.05,1) = 3.8415$.

So, as we observe from Table (2), we have the estimates of regression parameters and the tests which are based on the Hunua Ranges Data on ecological observations.

For the AQEF procedure:

Table 2. Results of the AQEF and GLM procedures

Estimate	AQEF	GLM	Estimate	AQEF	GLM
$\hat{p}_1$	0.3415	0.3416	$\hat{\alpha}_{10}$	-0.1245	-0.0493
$\hat{p}_2$	0.4061	0.3960	$\hat{\alpha}_{11}$	-0.0014	0.0029
$\hat{p}_3$	0.5565	0.5575	$\hat{\alpha}_{20}$	-0.1180	-
$\hat{\beta}_{10}$	-0.2910	-0.5747	$\hat{\alpha}_{21}$	-0.0007	-
$\hat{\beta}_{11}$	-0.0023	-0.0005	$\hat{\alpha}_{30}$	0.0443	0.6033
$\hat{\beta}_{20}$	-0.5336	-0.8459	$\hat{\alpha}_{31}$	0.0006	-0.0007
$\hat{\beta}_{21}$	0.0009	0.0026	$\hat{\alpha}_{40}$	0.0438	-
$\hat{\beta}_{30}$	-0.0139	-0.1326	$\hat{\alpha}_4$	0.0053	-
$\hat{\beta}_{31}$	0.0015	0.0023	$\hat{\phi}$	1.8424	1.8324
Scaled Deviance				248.8728	185.1379
Log likelihood Value				-762.1282	-752.3798
LRT (H0 : $\alpha_1 = 0$)				34.6890	19.8268
LRT (H0 : $\alpha_2 = 0$)				2.7690	-
LRT (H0 : $\alpha_3 = 0$)				6.4283	21.5709
LRT (H0 : $\alpha_4 = 0$)				23.9190	-

The LRT's demonstrate significant association between the binary variables Y_1 and Y_2, also no significant association between the binary variables Y_1 and Y_3, both are associated with explanatory variable, X. It is also evident that there is significant association between the binary variables Y_2 and Y_3, also significant association between the binary variables Y_1, Y_2 and Y_3, associated with explanatory variable, X. The estimate of dispersion parameter reflects the over-dispersion case ($\hat{\phi} = 1.8424 > 1$). The scaled deviance reflects the goodness of fit of the model,

[Scaled deviance $= 248.8728 < \chi^2(0.05, n-p=378) = 333.939$, p =14].

For the GLM procedure:

The LRT's demonstrates significant association between the correlated binary variables Y_1 and Y_2 and significant association between the correlated binary variables Y_2 and Y_3, both are associated with explanatory variable, X. The estimate of dispersion parameter reflects the over-dispersion case ($\hat{\phi} = 1.8324 > 1$). The scaled deviance also reflects the goodess of fit of the model,

[Scaled deviance $= 185.1379 < \chi^2(0.05, n-p = 382) = 337.700, p = 0$]. The regression models can be shown as follows:

For the AQEF procedure, we have the regression model:

$$\log\frac{p_{1i}}{1-p_{1i}} = -0.2910 - 0.0023x_i$$
$$\log\frac{p_{2i}}{1-p_{2i}} = -0.5336 + 0.0009x_i$$
$$\log\frac{p_{3i}}{1-p_{3i}} = -0.0139 + 0.0015x_i$$
$$\psi_{12i} = -0.1245 - 0.0014x_i \qquad (36)$$
$$\psi_{13i} = -0.1179 - 0.0007x_i$$
$$\psi_{23i} = 0.0443 + 0.0006x_i$$
$$\psi_{123i} = 0.0438 + 0.0053x_i$$

Similarly, for the GLM procedure, we have the regression model:

$$
\begin{aligned}
\log\frac{p_{1i}}{1-p_{1i}} &= -0.5747 - 0.0005x_i \\
\log\frac{p_{2i}}{1-p_{2i}} &= -0.8459 + 0.0026x_i \\
\log\frac{p_{3i}}{1-p_{3i}} &= -0.1326 + 0.0023x_i \\
\psi_{12i} &= -0.0493 + 0.0029x_i \\
\psi_{13i} &= 0.6033 - 0.0007x_i
\end{aligned} \tag{37}
$$

Acknowledgment

Many thanks for all My professors.

REFERENCES

[1] Agresti, A. (2002). Categorical data analysis (second edition). Wiley, New York.

[2] El-Sayed, A. M., Islam, M. A. and Alzaid, A. A. (2013). Estimation and test of measures of association for correlated binary data. Bulletin of the Malaysian Mathematical Sciences Society 2, 36, 4 , 985-1008.

[3] Heagerty, P. J. and Zeger, S. L. (2002). Marginalized multi-level models and likelihood inference (with discussion). Statistical Science 15, 1-26.

[4] Heagerty, P. J. (2002). Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics 58, 342-351.

[5] Islam, M. A., Chowdhury, R. I. and Briollais, L. (2012). A bivariate binary model for testing dependence in outcomes. Bulletin of the Malaysian Mathematical Sciences Society 2, 35, 4, 845-858.

[6] McCullagh, P. and Nelder, J. A. (1989). Generalized linear models (second edition). Chapman and Hall, London.

[7] Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. Biometrika 92, 455-463.

[8] Teugels, J. L. (1990). Some representations of the multivariate Bernoulli and binomial distributions. Journal of Multivariate Analysis 32, 256-268.

[9] Varadhan, R. and Gilbert, P. D. (2009). BB: An R package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. Journal of Statistical Software 32, 4, 1-26.

[10] Yee, T. W. (2008). The VGAM package. R News 8, 2, 28-39.

[11] Yee, T. W. (2010). The VGAM package for categorical data analysis. Journal of Statistical Software 32, 10, 1-34.

[12] Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a generalized quadratic model. Biometrika 77, 642-648.