American Journal of Theoretical and Applied Statistics
2016; 5(4): 225-233
http://www.sciencepublishinggroup.com/j/ajtas
doi: 10.11648/j.ajtas.20160504.19
ISSN: 2326-8999 (Print); ISSN: 2326-9006 (Online)

Modeling Multivariate Correlated Binary Data

Ahmed Mohamed Mohamed El-Sayed

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

Email address:
dr.ahmedelsayed4@yahoo.com

To cite this article:
Ahmed Mohamed Mohamed El-Sayed. Modeling Multivariate Correlated Binary Data. *American Journal of Theoretical and Applied Statistics.* Vol. 5, No. 4, 2016, pp. 225-233. doi: 10.11648/j.ajtas.20160504.19

Received: June 13, 2016; **Accepted**: June 22, 2016; **Published**: July 13, 2016

Abstract: This paper provides the model, estimation and test procedures for the measures of association in the correlated binary data associated with covariates in multivariate case. The generalized linear model (GLM) which satisfies the Markov properties for serial dependence, and the alternative quadratic exponential form (AQEF) are employed for multivariate Bernoulli outcome variables. The log-odds ratios as measures of association have been estimated, and the appropriate test procedures are suggested. The over-dispersion measure is investigated for the multivariate correlated binary outcomes. The scaled deviance is used as a goodness of fit of the model. For comparison, we have used the data on the respiratory disorder. In such situation, we indicate that the vectorized generalized linear models (VGLM) and AQEF procedures have the same estimates of regression parameters in the bivariate case.

Keywords: Multivariate Bernoulli Distribution, Generalized Linear Model, Scaled Deviance Test, Likelihood Ratio Test, Maximum Likelihood Estimators, Alternative Quadratic Exponential Form

1. Introduction

The dependence between the responses and the explanatory variables have been focused in the recent studies specially with one and two correlated outcomes variables associated with covariates. These studies make an attempt to focus on the multivariate correlated binary outcomes. Lovison [10] proposed a matrix-valued Bernoulli distribution, based on the log-linear representation introduced by Cox [6], for the multivariate Bernoulli distribution with correlated components. The model is based on the integration of conditional and marginal models. Teugels [12] used the concept of the Kronecker product to give some relationships between the correlated variables, namely, the correlation and odds ratios as measures of association. Zhao and Prentice [16] 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 [7], Heagerty and Zeger [8] presented the 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. Islam et al. [9] developed a new simple procedure to take account of the bivariate binary model with covariate dependence. Many of the vectorized generalized additive model (VGAM) features come from generalized linear model (GLM) and generalized additive model (GAM), so that readers with these functions can be returned to Chambers and Hastie [4]. Additionally, Yee and Wild [15] and the VGAM user R-manual, [14], should be consulted for general instructions about the software. General books dealt with log-linear model are referred as well, especially Christensen [5], Agresti [1] and McCullagh and Nelder [11]. Finally, El-Sayed et al. [3] introduced an alternative measure, based on the quadratic exponential form in the bivariate case, to make it more realistic, in terms of defining the underlying pseudo likelihood function, by modifying the normalizing term and developed Zhao and Prentice model [16] in the bivariate case, and this work also is devoloped in the trivariate case by El-Sayed [2].

In this paper, the major work is modeling the GLM with serial dependence, and the AQEF procedures associated with covariates. The estimations and tests of the association parameters are specified with appropriate link functions for the multivariate correlated binary case. Hence, the bivariate and trivariate AQEF will be extended to the multivariate case by modifying the 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 approach which demonstrates the serial dependence with the first-order Markov model. Section (2) presents the introduction to the multivariate Bernoulli distribution, namely, the joint probabilities and the log-odds ratios as measures of association explaining the relationship between the marginal, conditional and joint probabilities. Sections (3) and (4) present the modeling of the GLM and the AQEF procedures, in the multivariate case, respectively. Section (5) present simple introduction to VGLM procedure. Section (6) explained the numerical examples using the respiratory disorder data.

2. Multivariate Bernoulli Distribution

In this section, we will present the joint probability and the log-likelihood function for K correlated binary outcomes variables each following the Bernoulli distribution.

Let $Y=(Y_1,Y_2,...,Y_k)$ be a K dimensional vector of possibly correlated Bernoulli outcomes variables. The most general form of the joint mass function for Y is

$$f(y_1,y_2,...,y_k)=p_{000...0}^{\prod_{j=1}^{k}(1-y_j)}\times p_{100...0}^{y_1\prod_{j=2}^{k}(1-y_j)}\times...\times p_{111...1}^{\prod_{j=1}^{k}y_j}. \tag{1}$$

The corresponding log-likelihood function, for n observations, is

$$\ell(y_i;p)=\sum_{i=1}^{n}\Big(\prod_{j=1}^{k}(1-y_{ji})\log p_{000...0}+y_{1i}\prod_{j=2}^{k}(1-y_{ji})\log p_{100...0}+...+\prod_{j=1}^{k}y_{ji}\log p_{111...1}\Big). \tag{2}$$

For special case, $k=2$, we have the joint mass function for the correlated Bernoulli outcomes variables, Y_1 and Y_2, as

$$f(y_1,y_2)=p_{00}^{(1-y_1)(1-y_2)}+p_{01}^{(1-y_1)y_2}+p_{10}^{y_1(1-y_2)}+p_{11}^{y_1y_2}, \tag{3}$$

and the log-likelihood function, for n observations, is

$$\ell(y;p)=\sum_{i=1}^{n}\Big((1-y_{1i})(1-y_{2i})\log p_{00}+(1-y_{1i})y_{2i}\log p_{01}+y_{1i}(1-y_{2i})\log p_{10}+y_{1i}y_{2i}\log p_{11}\Big), \tag{4}$$

The next sections explain the parameters estimation and appropriate test procedures for both the AQEF and GLM procedures for the multivariate Bernoulli distribution as following:

3. Multivariate AQEF Procedure

In this section, we will extend the bivariate alternative quadratic exponential form which proposed by El-Sayed et al. [3] to the multivariate case. So, the joint mass function for K correlated binary variables $Y_1,Y_2,...,Y_k$ is

$$f(y_1,y_2,...,y_k)=\exp\Big\{\sum_{j=1}^{k}\theta_j y_j+\sum_{1\le j<l\le k}\psi_{jl}y_jy_l+\sum_{1\le j<l<m\le k}\psi_{jlm}y_jy_ly_m + ...+\psi_{123...k}y_1y_2y_3...y_k-\log c(\theta,\psi)\Big\}, \tag{5}$$

where, $\theta_j=\log\frac{p_j}{1-p_j}$, $j=1,2,..,k$, are natural parameters, and

$$\psi_{jl}=\log\frac{P(Y_l=1\mid Y_j=1)}{P(Y_l=1\mid Y_j=0)}=\log\frac{P(Y_j=1,Y_l=1)P(Y_j=0,Y_l=0)}{P(Y_j=1,Y_l=0)P(Y_j=0,Y_l=1)},\ 1\le j<l\le k,$$

are associated parameters, and so on.

To obtain the normalizing term, $c(\theta,\psi)$, in the function (5), we can use this constraint

$$\sum_{y_1=0}^{1}\sum_{y_2=0}^{1}\cdots\sum_{y_k=0}^{1} f(y_1, y_2, ..., y_k) = 1. \tag{6}$$

In this case, the normalizing constant can be obtained as

$$c(\theta,\psi) = \sum \exp\Big\{\sum_{j=1}^{k}\theta_j y_j + \sum_{1\le j<l\le k}\psi_{jl}y_j y_l + \sum_{1\le j<l<m\le k}\psi_{jlm}y_j y_l y_m + ... + \psi_{123...k}y_1 y_2 y_3 ... y_k\Big\}, \tag{7}$$

the summation over all $\mathbf{2}^k$ possible values of Y. Then, the normalizing constant is

$$\begin{aligned} c(\theta,\psi) &= 1 + \sum_{j=1}^{k}\exp(\theta_j) + \sum_{1\le j<l\le k}\exp(\theta_j+\theta_l+\psi_{jl}) \\ &+ \sum_{1\le j<l<m\le k}\exp(\theta_j+\theta_l+\theta_m+\psi_{jl}+\psi_{jm}+\psi_{lm}+\psi_{jlm}) \\ &+ ... + \exp\Big(\sum_{j=1}^{k}(\theta_j) + \sum_{1\le j<l\le k}(\theta_j+\theta_l+\psi_{jl}) + ... + \psi_{123...k}\Big) \end{aligned} \tag{8}$$

For special case, $k = \mathbf{2}$, the joint probability mass function for Y_1 and Y_2 is

$$f(y_1, y_2) = \exp\left\{\theta_1 y_1 + \theta_2 y_2 + \psi_{12} y_1 y_2 - \log(1 + e^{\theta_1} + e^{\theta_2} + e^{\theta_1+\theta_2+\psi_{12}})\right\}. \tag{9}$$

3.1. Natural Parameters Estimation

The log-likelihood function, for n observations, can be written as

$$\begin{aligned} \ell(\theta,\psi) = \sum_{i=1}^{n}\Big\{&\sum_{j=1}^{k}\theta_j y_{ji} + \sum_{1\le j<l\le k}\psi_{jl}y_{ji}y_{li} + \sum_{1\le j<l<m\le k}\psi_{jlm}y_{ji}y_{li}y_{mi} + ... + \\ &\psi_{12...k}y_{1i}y_{2i}...y_{ki} - \log\, c(\theta,\psi)\Big\}, \end{aligned} \tag{10}$$

where $c(\theta,\psi)$ is defined as shown in (8).

Taking the first derivatives for (10) with respect to $\theta_j, \psi_{jl}, ..., \psi_{123...k}$, and put it equal to zero, we have:

$$\begin{aligned} &\frac{\partial\ell(\theta,\psi)}{\partial\theta_j} = \sum_{i=1}^{n}\Big(y_{ji} - \frac{e^{\theta_j} + e^{\theta_j+\theta_l+\psi_{jl}}}{c(\theta,\psi)}\Big) = 0, \quad l > j = 1,2,.., k, \\ &\frac{\partial\ell(\theta,\psi)}{\partial\psi_{jl}} = \sum_{i=1}^{n}\Big(y_{ji}y_{li} - \frac{e^{\theta_j+\theta_l+\psi_{jl}}}{c(\theta,\psi)}\Big) = 0, \quad l > j = 1,2,.., k, \\ &\vdots \\ &\frac{\partial\ell(\theta,\psi)}{\partial\psi_{12...k}} = \sum_{i=1}^{n}\Big(y_{1i}y_{2i}...y_{ki} - \frac{\exp\{\sum_{j=1}^{k}(\theta_j) + \sum_{1\le j<l\le k}(\theta_j+\theta_l+\psi_{jl}) + ... + \psi_{123...k}\}}{c(\theta,\psi)}\Big) = 0, \end{aligned} \tag{11}$$

Solving the equations (11), numerically, we can get the estimates $\hat{\theta}_j, \hat{\theta}_l, .., \hat{\psi}_{jl}, .., \hat{\psi}_{123...k}$, respectively.

3.2. Testing Hypothesis for Natural Parameters

We can test the null hypothesis $H_0 : \theta_j = 0$ against the alternative hypothesis $H_0 : \theta_j \neq 0$, $j = 1, 2.., k$. To test the significance of association parameters, we can test the null hypothesis $H_0 : \psi_{jl} = 0$ against the alternative hypothesis $H_0 : \psi_{jl} \neq 0$, $1 \leq j < l \leq k$. Also, we can test the null hypothesis $H_0 : \psi_{jlm} = 0$, $1 \leq j < l < m \leq k$, and so on. All tests can be done using the Likelihood ratio test (LRT).

3.3. Modeling Multivariate AQEF Procedure

In this section, we will use the next link functions to generalize the model, with correlated dependent binary variables associated with some covariates, x (not always binary variables). The marginal probabilities $p_j (j = 1, 2, .., k)$ is given by the the regression model

$$\theta_j = \log \frac{p_j(x)}{1 - p_j(x)} = \beta_j' x, \quad x' = (1 \quad x_1 \quad x_2 \quad ... \quad x_k), \quad \beta_j' = (\beta_{j0} \quad \beta_{j1} \quad \beta_{j2} \quad ... \quad \beta_{jk}). \tag{12}$$

A regression model expresses the association between these responses, associated with some covariates, x, can be given by

$$\psi_{jl} = \alpha_{jl}' x, \quad \psi_{jlm} = \alpha_{jlm}' x, \quad \psi_{12...k} = \alpha_{12...k}' x. \tag{13}$$

The covariates, x, which are selected show some significant association with the variables, $Y_1, Y_2, ..., Y_k$, in multivariate analysis.

Now, we will study the effect of covariates x on the log-likelihood function (10), using the equations (12) and (13).

3.4. Regression Parameters Estimation

The log-likelihood function can be expressed as follows:

$$\ell(\alpha, \beta) = \sum_{i=1}^{n} \Big(\sum_{j=1}^{k} \beta_j' x y_{ji} + \sum_{1 \leq j < l \leq k} \alpha_{jl}' x y_{ji} y_{li} + \sum_{1 \leq j < l < m \leq k} \alpha_{jlm}' x y_{ji} y_{li} y_{mi} + ... + \alpha_{12...k}' x y_{1i} y_{2i} ... y_{ki} - \log c(\beta, \alpha) \Big), \tag{14}$$

where $c(\beta, \alpha)$ is defined as shown below

$$\begin{aligned} c(\beta, \alpha) &= 1 + \sum_{j=1}^{k} \exp(\beta_j' x) + \sum_{1 \leq j < l \leq k} \exp(\beta_j' x + \beta_l' x + \alpha_{jl}' x) \\ &+ \sum_{1 \leq j < l < m \leq k} \exp(\beta_j' x + \beta_l' x + \beta_m' x + \alpha_{jl}' x + \alpha_{jm}' x + \alpha_{lm}' x + \alpha_{jlm}' x) + ... \\ &+ \exp\Big(\sum_{j=1}^{k} (\beta_j' x) + \sum_{1 \leq j < l \leq k} (\beta_j' x + \beta_l' x + \alpha_{jl}' x) + ... + \alpha_{123...k}' x\Big). \end{aligned} \tag{15}$$

Taking the first derivatives for (14) with respect to $\beta_j, .., \alpha_{jl}, ..., \alpha_{123...k}$, and put it equal to zero, we have:

$$\frac{\partial \ell(\beta,\alpha)}{\partial \beta_j} = \sum_{i=1}^{n}\Big(y_{ji} - \frac{e^{\beta_j'x} + e^{\beta_j'x+\beta_l'x+\alpha_{jl}'x}}{c(\beta,\alpha)}\Big)x = 0, \quad l > j = 1,2,..,k,$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \alpha_{jl}} = \sum_{i=1}^{n}\Big(y_{ji}y_{li} - \frac{e^{\beta_j'x+\beta_l'x+\alpha_{jl}'x}}{c(\beta,\alpha)}\Big)x = 0, \quad l > j = 1,2,..,k, \tag{16}$$

$$\vdots$$

$$\frac{\partial \ell(\beta,\alpha)}{\partial \alpha_{12...k}} = \sum_{i=1}^{n}\Big(y_{1i}y_{2i}...y_{ki} - \frac{\exp\Big\{\sum_{j=1}^{k}(\beta_j'x) + \sum_{1\le j<l\le k}(\beta_j'x+\beta_l'x+\alpha_{jl}'x) + ... + \alpha_{123...k}'x\Big\}}{c(\beta,\alpha)}\Big)x = 0.$$

Solving the equations (16), numerically, we can get the estimates $\hat{\beta}_j,..,\hat{\alpha}_{jl},..,\hat{\alpha}_{123...k}$, respectively.

3.5. Testing Hypothesis for Regression Parameters

We can test the null hypothesis, $H_0 : \beta_j = 0 (j = 1,2,..,k)$, using

$$LRT = -2[\ell(\beta_j = 0, \tilde{\alpha}) - \ell(\hat{\beta}_j, \hat{\alpha})] \quad \sim \quad \chi_1^2 \tag{17}$$

Finally, we can test the null hypothesis, $H_0 : \alpha = 0$ ($\alpha = \alpha_{jl}$ or α_{jlm} or... or $\alpha_{12...k}$), using

$$LRT = -2[\ell(\tilde{\beta}_j, \alpha = 0) - \ell(\hat{\beta}_j, \hat{\alpha})] \quad \sim \quad \chi_1^2 \tag{18}$$

The estimated dispersion parameter φ can be used as a measure for the over-dispersion. So, let us define

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \end{pmatrix}, \quad \hat{p} = \begin{pmatrix} \hat{p}_1 \\ \hat{p}_2 \\ \vdots \\ \hat{p}_k \end{pmatrix}, \quad \hat{\Sigma} = \begin{pmatrix} Var(Y_1) & Cov(Y_1,Y_2) & \cdots & Cov(Y_1,Y_k) \\ Cov(Y_2,Y_1) & Var(Y_2) & \cdots & Cov(Y_2,Y_k) \\ \vdots & \vdots & \ddots & \cdots \\ Cov(Y_k,Y_1) & Cov(Y_k,Y_2) & \cdots & Var(Y_k) \end{pmatrix},$$

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 can be defined as

$$\hat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\sum_{j=1}^{k}\frac{(y_{ji}-\hat{p}_j)^2}{Var(\hat{p}_j)}, \tag{19}$$

the value of $\hat{\varphi}$ should be closed to one for a Bernoulli data. To evaluate $\hat{\varphi}$, we must obtain the estimate of marginals, $\hat{p}_j$, using the equation (12), as

$$p_j(\beta) = \frac{e^{\beta_j'x}}{1+e^{\beta_j'x}}, \quad j = 1,2,...,k. \tag{20}$$

Also, to specify the goodness of fit model, we can use the scaled deviance function

$$S.D(y_i, \hat{\beta}_j, \hat{\alpha}) = \frac{D(y_i, \hat{\beta}_j, \hat{\alpha})}{\hat{\phi}} = 2\big[\ell(y_i, y_i) - \ell(y_i, \hat{\beta}_j, \hat{\alpha})\big] \quad \sim \quad \chi_{n-p}^2, \tag{21}$$

where p is the number of estimated parameters, and $\hat{\varphi}$ is the dispersion parameter estimate as defined in (19). Since, the deviance function is

$$D(y_i,\hat{\beta}_j,\hat{\alpha}) = S.D(y_i,\hat{\beta}_j,\hat{\alpha})\times\hat{\phi}. \tag{22}$$

4. Multivariate GLM Procedure

The Markov structures of dependence often adequately describe serial stochastic dependence in specified data. This pattern of dependence has been studied and so only a few remarks will be made here. Markov dependence of first order implies

$$Pr(Y_1 = y_1, Y_2 = y_2,.., Y_k = y_k) = Pr(Y_1 = y_1)\prod_{j=2}^{k} Pr(Y_j \mid Y_{j-1}). \tag{23}$$

Using the conditional logg-odds interpretation, Heagerty [7], and Heagerty and Zeger [8], and the Markov property, the joint mass function for the variables Y can be defined as

$$\begin{aligned} f(y_1, y_2,.., y_k) = \exp\Big\{&\sum_{j=1}^{k}\theta_j y_j - \sum_{j=1}^{k}\log(1+e^{\theta_j}) + \sum_{j=2}^{k}\psi_{j-1,j}y_{j-1}y_j \\ &- \sum_{j=2}^{k} y_{j-1}(\log[1+e^{\theta_j+\psi_{j-1,j}}] - \log[1+e^{\theta_j}])\Big\}. \end{aligned} \tag{24}$$

For special case, $k = 2$, the joint probability mass function for Y_1 and Y_2 is

$$f(y_1,y_2) = \exp\Big\{\theta_1 y_1 + \theta_2 y_2 + \psi_{12}y_1y_2 - \log(1+e^{\theta_1}) - \log(1+e^{\theta_2}) - y_1(\log[1+e^{\theta_2+\psi_{12}}] - \log[1+e^{\theta_2}])\Big\}. \tag{25}$$

4.1. Natural Parameters Estimation

In this section, we present the estimation of parameters of the multivariate Bernoulli distribution. For n observations, we can get the log-likelihood function as

$$\begin{aligned} \ell(\theta,\psi) = \sum_{i=1}^{n}\Big\{&\sum_{j=1}^{k}\theta_j y_{ji} - \sum_{j=1}^{k}\log(1+e^{\theta_j}) + \sum_{j=2}^{k}\psi_{j-1,j}y_{j-1,i}y_{ji} \\ &- \sum_{j=2}^{k} y_{j-1,i}(\log[1+e^{\theta_j+\psi_{j-1,j}}] - \log[1+e^{\theta_j}])\Big\}. \end{aligned} \tag{26}$$

Taking the first derivatives for (26) with respect to θ_j and $\psi_{j-1,j}$, and put it equal to zero, we have

$$\begin{aligned} &\frac{\partial\ell(\theta,\psi)}{\partial\theta_j} = \sum_{i=1}^{n}(y_{ji} - \frac{e^{\theta_j}}{1+e^{\theta_j}}) = 0, \quad j = 1, \\ &\frac{\partial\ell(\theta,\psi)}{\partial\theta_j} = \sum_{i=1}^{n}(y_{ji} - \frac{e^{\theta_j}}{1+e^{\theta_j}}) - \sum_{i=1}^{n} y_{j-1,i}(\frac{e^{\theta_j+\psi_{j-1,j}}}{1+e^{\theta_j+\psi_{j-1,j}}} - \frac{e^{\theta_j}}{1+e^{\theta_j}}) = 0, \quad j = 2,..,k, \\ &\frac{\partial\ell(\theta,\psi)}{\partial\psi_{j-1,j}} = \sum_{i=1}^{n}(y_{j-1,i}y_{ji} - y_{j-1,i}\frac{e^{\theta_j+\psi_{j-1,j}}}{1+e^{\theta_j+\psi_{j-1,j}}}) = 0. \end{aligned} \tag{27}$$

Solving the equations (27), numerically, we have the estimates $\hat{\theta}_j (j = 1,2,..,k)$ and $\hat{\psi}_{j-1,j} (j = 2,3,..,k)$.

4.2. Testing Hypothesis for Natural Parameters

We can test the null hypothesis $H_0 : \theta_j = 0$ against the alternative hypothesis $H_0 : \theta_j \neq 0$, $j = 1,2..,k$. To test the

association parameters, we can test the null hypothesis $H_0 : \psi_{j-1,j} = 0$ against the alternative hypothesis $H_0 : \psi_{j-1,j} \neq 0$, $j = 2,3,..,k$. All tests can be done using the Likelihood ratio test (LRT).

4.3. Modeling Multivariate GLM Procedure

In this section, we will use the same link functions similar to the AQEF to determine the regression model. A regression model which expresses the link functions and the association between the correlated binary responses, Y, associated with covariates, x, can be given by the equations (12) and (13).

4.4. Regression Parameters Estimation

Now, we study the effect of covariates on the log-likelihood function (26) which is become

$$\begin{aligned}\ell(\beta,\alpha) = \sum_{i=1}^{n}\Big\{\sum_{j=1}^{k}\beta_j' x y_{ji} - \sum_{j=1}^{k}\log(1+e^{\beta_j' x}) + \sum_{j=2}^{k}\alpha_{j-1,j}' x y_{j-1,i} y_{ji} \\ - \sum_{j=2}^{k} y_{j-1,i}(\log[1+e^{\beta_j' x+\alpha_{j-1,j}' x}] - \log[1+e^{\beta_j' x}])\Big\}.\end{aligned} \tag{28}$$

Taking the first derivative for (28) with respect to $\beta_j, \alpha_{j-1,j}$, and putting it equal to zero, we get the estimating equations

$$\begin{aligned}\frac{\partial \ell(\beta,\alpha)}{\partial \beta_j} &= \sum_{i=1}^{n}\Big(y_{ji} - \frac{e^{\beta_j' x}}{1+e^{\beta_j' x}}\Big)x = 0, \quad j = 1, \\ \frac{\partial \ell(\beta,\alpha)}{\partial \beta_j} &= \sum_{i=1}^{n}\Big(y_{ji} - \frac{e^{\beta_j' x}}{1+e^{\beta_j' x}}\Big)x - \sum_{i=1}^{n} y_{j-1,i}\Big(\frac{e^{\beta_j' x+\alpha_{j-1,j}' x}}{1+e^{\beta_j' x+\alpha_{j-1,j}' x}} - \frac{e^{\beta_j' x}}{1+e^{\beta_j' x}}\Big)x = 0, \quad j = 2,..,k, \\ \frac{\partial \ell(\beta,\alpha)}{\partial \alpha_{j-1,j}} &= \sum_{i=1}^{n}\Big(y_{j-1,i}y_{ji} - y_{j-1,i}\frac{e^{\beta_j' x+\alpha_{j-1,j}' x}}{1+e^{\beta_j' x+\alpha_{j-1,j}' x}}\Big)x = 0, \quad j = 2,..,k.\end{aligned} \tag{29}$$

Solving the equations (29), numerically, we have the vectors estimates $\hat{\beta}_j (j = 1,2,..,k), \hat{\alpha}_{j-1,j} (j = 2,3,..,k)$.

4.5. Hypothesis Test for Regression Parameters

We can test the regression parameters using the null hypothesis $H_0 : \beta_j = 0 (j = 1,2,..,k)$, by the function

$$LRT = -2[\ell(\beta_j = 0, \tilde{\alpha}_{j-1,j}) - \ell(\hat{\beta}_j, \hat{\alpha}_{j-1,j})] \quad \sim \quad \chi_1^2 \tag{30}$$

Finally, we can test the association parameters using the the null hypothesis $H_0 : \alpha_{j-1,j} = 0$, by the function

$$LRT = -2[\ell(\tilde{\beta}_j, \alpha_{j-1,j} = 0) - \ell(\hat{\beta}_j, \hat{\alpha}_{j-1,j})] \quad \sim \quad \chi_1^2 \tag{31}$$

The estimate of dispersion parameter φ can be defined as shown in the equation (19). Also, to specify the goodness of fit model, we can use the scaled deviance function (21).

5. Multivariate VGAM Procedure

The conditional distribution of vectorized generalized linear models (VGAM), Yee and Wild [15], for multivariate correlated binary responses ($Y_1, Y_2, ..., Y_k$), given that some covariates, x, is given by the function:

$$\log f(y_1, y_2, ..., y_k \mid x) = u_0(x) + \sum_{j=1}^{k} u_j(x) y_j + \sum_{j<l}^{k} u_{jl}(x) y_j y_l.$$

Where, $u_0(x)$ is the normalizing term. Similar to the GLM and AQEF procedures, we can get the estimate of natural parameters, the estimate of regression parameters, the estimate of dispersion parameters, the scaled deviance and the LRTs.

6. Numerical Examples

Respiratory Disorder Data: Source: Stokes, Davis, and Koch (1995), SAS and R programs.

These data is taken from a clinical trial of patients comparing two treatments for a respiratory illness. The data contains (111) patients from two different clinics (centers) which were randomized to receive either placebo = 0 or active = 1 treatment. Patients were examined at baseline (represent the baseline respiratory status) and at four visits during the treatment. At each examination, the respiratory status was determined. A data frame are (444) observations and (8) variables which are: outcome variable (represent the respiratory status at each visit [categorized as good = 1, poor = 0]), center (center 1=1, center 2 = 2), id (repetition), age (age at time of entry into the study which represents a continuous variable), baseline (baseline respiratory status good or not, hence [good = 1, poor = 0]), treatment (placebo = P, active = A), hence to be binary data we can put P = 0 and A = 1, sex (female = 1, male = 0) and visit (four visits). We suppose that, for the bivariate case $(K = 2)$, the response variables in this model are two variables: the "outcome" variable represented by the binary variable Y_1 and the "treatment" variable represented by the binary variable Y_2. Explanatory variables in this model are six variables: center, age, baseline, sex and visit. In this example, the two dependent correlated binary variables Y_1 and Y_2, represent the outcome and the treatment variables respectively. One explanatory variable X, represents the visit. In the next examples, we use the VGLM procedure, Yee [14], Yee and Wild [15], which depends on the log-linear approach in the bivariate case. The estimates obtained using the BB-package of R program, [13].

Table 1 explains the results for the GLM, QEF and AQEF procedures as following:

Table 1. *Results of VGAM, AQEF and GLM procedures.*

Estimate	VGAM	AQEF	GLM
$\hat{\beta}_{10}$	-0.1402	-0.1402	0.3818
$\hat{\beta}_{11}$	-0.0356	-0.0356	-0.0585
$\hat{\beta}_{20}$	-0.7328	-0.7328	-0.7318
$\hat{\beta}_{21}$	0.0481	0.0481	0.0477
$\hat{\alpha}_0$	1.1331	1.1331	1.1323
$\hat{\alpha}_1$	-0.0583	-0.0583	-0.0579
$\hat{\psi}$	2.6901	2.6901	2.6900
$\hat{\varphi}$	1.5251	1.5251	1.3915
Scaled Deviance	276.9002	276.9002	269.2120
Log likelihood Value	-599.1579	-599.1579	-599.1584
LRT (H_0: $\alpha = 0$)	25.8172	25.8172	25.8162

$\chi^2(0.05, n-p = 438) = 487.7930$, $\chi^2(0.05,1) = 3.8415$, $p = 6$ parameters.

From Table 1, we have found that:

The VGLM and AQEF procedures have the same estimates, but the GLM procedure has different estimates.

For the scaled deviance measure as a goodness of fit of the model, we found all measures have values less than $\chi^2(0.05, n-p = 438) = 487.7930$, $p = 6$ parameters.

This means that all measures have a good fit.

For the estimate of dispersion parameter φ, the procedure GLM has the smallest value.

For the LRT to test the null hypothesis $H_0 : \alpha = 0$, we find, for all procedures, the value of LRT is more than $\chi^2(0.05,1) = 3.8415$. This means that, for all procedures, the two correlated dependent variables Y_1 and Y_2 are affected significantly with the explanatory variable X.

Then, the patient respiratory status, contributed and the treatment, are affected significantly by each visit. Hence the test of associated parameters reflect the significant association between Y_1 and Y_2 associated with x covariates.

In sum, the previous results proved that the same results are obtained for the VGLM and AQEF procedures. Then, we can use the Wald statistic to test the significance of the parameters of regression model as shown below.

The results in Table 1, are demonstrated in the regression model shown below:

For the GLM procedure, we have the regression model:

$$\begin{aligned} logit\ (p_{1i}) &= 0.3818 - 0.0585\ x_i \\ logit\ (p_{2i}) &= -0.7318 + 0.0477\ x_i \\ \psi_{12i} &= 1.1323 - 0.0579\ x_i \end{aligned} \tag{32}$$

Also, for the VGAM and AQEF procedures, we have the regression model:

$$\begin{aligned} logit\ (p_{1i}) &= -0.1402 - 0.0356\ x_i \\ logit\ (p_{2i}) &= -0.7328 + 0.0481\ x_i \\ \psi_{12i} &= 1.1331 - 0.0583\ x_i \end{aligned} \tag{33}$$

Table 2 reflects the estimates, standard error and Wald statistic for regression parameters for the procedures VGAM and AQEF, which have the same results.

Table 2. *Estimates, Standard error and Wald statistic.*

	$\hat{\beta}_{10}$	$\hat{\beta}_{11}$	$\hat{\beta}_{20}$	$\hat{\beta}_{21}$	$\hat{\alpha}_0$	$\hat{\alpha}_1$
Estimate	-0.1402	-0.0356	-0.7328	0.0481	1.1331	-0.0582
Standard Error	0.3262	0.1193	0.3740	0.1340	0.4881	0.1770
Wald Statistic	-0.4300	-0.02982	-1.9593	0.3592	2.3217	-0.3292

$Z_{0.025} = \pm 1.95996$

From Table 2, the Wald statistics reflect the dependent variables Y_1 and Y_2 together are affected significantly with the explanatory variable, X. This confirms the results obtained for the LRT in Table 1. Also, we can use the VGAM-package to fit the model using more than one covariates. Applying that on the respiratory disorder data,

considering the dependent correlated binary variables are outcome (Y_1) and treatment (Y_2), and the the explanatory variables are: center (X_1), sex (X_2), age (X_3), visit (X_4) and baseline (X_5).

Table 3 represents the results associated with more than one covariates:

Table 3. *Logits, Measure of association, Standard error and Wald statistic.*

Model	Intercept	Center X_1	Sex X_2	Age X_3	Visit X_4	Baseline X_5
logit P_1	-0.5026	0.5189	0.0376	-0.03596	-0.0504	1.6967
Standard Error	0.6194	0.3125	0.3531	0.0119	0.1279	0.3067
Wald Statistic	-0.8114	1.6605	0.1065	-3.0237	-0.3938	5.5316
logit P_2	0.2095	-0.0784	-2.0342	-0.0098	0.0553	-0.9838
Standard Error	0.6663	0.3494	0.5669	0.0123	0.1393	0.4221
Wald Statistic	0.3143	-0.2245	-3.5883	-0.7932	0.3971	-2.3309
$\hat{\psi}_{12}$	-0.4083	0.3308	0.8610	0.0340	-0.0619	0.4303
Standard Error	0.8806	0.4448	0.6715	0.0161	0.1804	0.5060
Wald Statistic	-0.4637	0.7438	1.2822	2.1082	-0.3430	0.8504

Log-likelihood: -531.1790 , $Z_{0.025} = \pm 1.95996$.

From Table 3, we have found that:

The two dependent correlated binary variables, outcome (Y_1), and treatment (Y_2) are together affected significantly by the explanatory variable age (X_3).

The dependent variable outcome (Y_1) is affected significantly by the explanatory variables, baseline (X_5) and age (X_3).

The dependent variable treatment (Y_2) is affected significantly by the explanatory variables, baseline (X_5) and sex (X_2).

From Table 3, we have the regression model:

$$\begin{aligned} logit\,(p_{1i}) &= -0.5026 + 0.5189\,x_{1i} + 0.0376\,x_{2i} - 0.03596\,x_{3i} - 0.0504\,x_{4i} + 1.6967\,x_{5i} \\ logit\,(p_{2i}) &= 0.2095 - 0.0784\,x_{1i} - 2.0342\,x_{2i} - 0.0098\,x_{3i} + 0.0553\,x_{4i} - 0.9838\,x_{5i} \\ \psi_{12i} &= -0.4083 + 0.3308\,x_{1i} + 0.8610\,x_{2i} + 0.0340\,x_{3i} - 0.0619\,x_{4i} + 0.4303\,x_{5i} \end{aligned} \quad (34)$$

References

[1] Agresti A. Categorical data analysis (second edition). New Jersey, United States: John Wiley & Sons; 2002.

[2] El-Sayed A M. M. Modeling trivariate binary data. Al-Azhar University, Journal of College of Science 2016; Accepted.

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

[4] Chambers J M, Hastie TJ. Statistical Models in Solomon. New York: Chapman and Hall; 1993.

[5] Christensen R. Log-linear Models and Logistic Regression (second edition). New York, United States: Springer-Verlag; 1997.

[6] Cox D R. The analysis of multivariate binary data. Journal of the Royal Statistical Society, Series C (Applied Statistics) 1972; 21: 113-120.

[7] Heagerty P J. Marginalized transition models and likelihood inference for longitudinal categorical data. Biometrics 2002; 58: 342-351.

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

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

[10] Lovison G. A matrix-valued Bernoulli distribution. Journal of Multivariate Analysis 2006; 97: 1573-1585.

[11] McCullagh P, Nelder J A. Generalized linear models (second edition). London, United Kingdom: Chapman & Hall; 1989.

[12] Teugels J L. Some representations of the multivariate Bernoulli and Binomial distributions. Journal of multivariate analysis 1990; 32: 256-268.

[13] Varadhan R, Gilbert P D. 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 2009; 32, 4: 1-26.

[14] Yee T W. The VGAM package, R News 2008; 8, 2: 28-39.

[15] Yee T W, Wild C J. Vector generalized additive models. Journal of the Royal Statistical Society, Series B, Methodological 1996; 58: 481-493.

[16] Zhao L P, Prentice R L. Correlated binary regression using a generalized quadratic model. Biometrika 1990; 77: 642-648.