ISSN 0104-0499

PROGRAMA INTERINSTITUCIONAL DE PÓS-GRADUAÇÃO EM ESTATÍSTICA
UFSCar-USP

DEs-UFSCar e SME-ICMC-USP

CORRELATED BINOMIAL REGRESSION: MODELING, ESTIMATION AND DIAGNOSTICS

Rubiane M. Pires
Carlos A. R. Diniz
Carol C. M. Paraíba

RELATÓRIO TÉCNICO

TEORIA E MÉTODO – SÉRIE A

Setembro/2013
nº 256

Correlated binomial regression: modeling, estimation and diagnostics

Rubiane M. Pires and Carlos A. R. Diniz Carol C. M. Paraiba

Departamento de Estatística
Universidade Federal de São Carlos
C.P. 676, 13565-905, São Carlos-SP, Brasil
email: dcad@ufscar.br

Abstract: This article considers a new approach, an alternative class of correlated binomial regression models, to analyze data set involving correlated binary variables. The proposed methodology is illustrated considering a real data set containing information about a Brazilian health plan operator. The main interest of the analysis is to fit a regression model that can be used to determine the probability of high-cost health service occurrence in a company. A decision about the renewal or not of the health plan can be taken based on the magnitude of this probability. The data set presents a portfolio of companies (clusters) and the occurrence or not of high-cost health service for the employee (individual) inside the company. The available data in the ith company, $i = 1, 2, \ldots, 160$, with n_i employees, consists of $W_{i1}, W_{i2}, \ldots, W_{in_i}$, each one assuming value 0 or 1, depending on the status of the employee. The response variable for the ith company, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, assumes values at $\{0, 1, \ldots, n_i\}$. A dependence structure between the Bernoulli variables inside the company can be explained by the fact that the employees are exposed to the same environment. Some cluster covariates are available in the data set. The probability of high-cost health service occurrence is determined using the alternative class of correlated binomial regression models which is based on a generalized binomial distribution.

Keywords and phrases: Generalized binomial distribution, overdispersed regression models, residuals, local influence.

1. Introduction

In a real practical situation, it is common to observe data sets where the response variable represents the sum of dependent Bernoulli random variables. McCullagh & Nelder (1989) argue that, unless there are good reasons for relying on the binomial assumption, it seems to be wise to be cautious and to assume that overdispersion is present in these types of data sets. The overdispersion phenomenon occurs when a higher variability than that assigned to the usual binomial model is observed in the data and can be attributed to several causes such as correlation between the binary responses, the absence of relevant explanatory variables, etc.

As a motivational example of overdispersed binomial data modeling, we consider a data set from a Brazilian health plan operator. The data comprise a portfolio of companies (clusters) for which the occurrence or not of high-cost health services - such as oncological surgery, prosthesis, chemotherapy and hemodialysis - is/are observed for each employee. The available data for the ith company with n_i employees, $i = 1, 2, \ldots, 160$, consists of $W_{i1}, W_{i2}, \ldots, W_{in_i}$, each one assuming value 0 or 1, depending on the status of the employee ($0 =$ not occurrence; $1 =$ occurrence). Thus, the response variable for the ith company, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, assumes values in $\{0, 1, \ldots, n_i\}$ according to the number of employees who have used the high-cost health services. For this particular data set, a dependence structure between the Bernoulli variables inside the company could be assumed and explained by the fact it is reasonable to consider that employees within the same company are exposed to the same environment conditions. This data set also features the following covariates: average number of medical appointments per employee; average cost of medical test; occurrence of surgical procedure; number of therapies; number of emergency procedures; number of days between the beginning of the plan period and the first high-cost health service occurrence per each employee and specific information about the companies (size, number of employees, business activity). From the point of view of the operator, the main interest while analyzing this data set would be to fit a regression model able to precisely determine the probability of a high-cost health service occurrence in a company, which would be taken into consideration at the time of renewing or not of the health plan. Since we have assumed correlations between the employees inside the company, the probability of a high-cost health service occurrence could be found using overdispersed binomial regression models for independent data.

The modified logistic-linear model (Williams, 1982) could be fitted if employees within the same cluster had a correlation parameter common to all companies. The beta-binomial regression model (Altham, 1978; Prentice, 1986; Efron, 1986; Lindsey, 1995; Lindsey & Altham, 1998), the multiplicative binomial regression model (Lindsey & Altham, 1998), and the double-binomial regression model belonging to Efron's double-exponential family (Efron, 1986) could also be considered for the health plan data set. For theses models, both the probability of a high-cost health service occurrence and the correlation parameter could be found using two separate regression equations. The multivariate probit regression model (Ochi & Prentice, 1984), which also allows pairwise correlation between binary variates would also be a possibility to model the discussed data. Moreover, if each binary observation had its own covariates, we could use the regression methods proposed by Prentice (1988). Another useful class of distributions proposed for binomial data modeling in the presence of overdispersion is the class of finite mixed models, particularly correlated binomial regression models, which also can be interpreted as inflated models (see Lambert, 1992). A Bayesian approach for correlated binomial regression models is presented in Pires & Diniz (2012).

Correlated binomial regression models are based on the generalized binomial distribution, proposed by Luceño (1995) and Luceño & De Ceballos (1995) and discussed in detail in Diniz *et al.* (2010), which represents a form to write the distribution of sums of dependent Bernoulli random variables equicorrelated using the mixture of the distributions of two variables. In this paper, we develop an alternative class of correlated binomial regression models by jointly modeling the probability of success using four different link functions, the logit, the complementary log-log, the log-log and the probit links, and the dependency between individuals within the same cluster using correlated functions (Jennrich & Schluchter, 1986; Zimmerman & Harville, 1991; Cressie, 1993; Russell, 1996; Sherman, 2011), while taking into account the available covariates. The maximum likelihood estimator are obtained by direct maximization of the likelihood function.

As with any modeling procedure, we need to make some initial assumptions in order to fit the model to the data. Section 3 presents some underlying assumptions made when constructing the correlated binomial regression model. Residuals based on the predicted values and deviance residuals are defined to check the assumptions in the model. A sensitivity study to detect outliers or influential cases that can change the inferential results is performed. A case-deletion influence diagnostic (Cook & Weisberg, 1982) based on the generalized Cook's distance and the likelihood distance (Zhu *et al.*, 2001) are considered to evaluate the sensitivity of the observations when estimating the parameters. Two predictive model selection criteria, the Akaike information criteria (AIC) (Akaike, 1974) and the Bayesian information criteria (BIC) (Schwarz, 1978), are used.

2. Correlated binomial regression model

Assume $Y_1, Y_2, \ldots, Y_m$ are independent random variables such that each Y_i follows a correlated binomial distribution, denoted by $Y_i \sim CB(n_i, p_i, \rho_i), i = 1, \ldots, m$. The correlated binomial distribution (Luceño, 1995; Luceño & De Ceballos, 1995) is a form to write the distribution of sums of equicorrelated Bernoulli random variables. It is given by the mixture of the distributions of two variables. One of them follows a binomial distribution, $B(n_i, p_i)$, with mixing probability $(1-\rho_i)$, and the other one follows a modified Bernoulli distribution, $MBern(p_i)$, taking values 0 or n_i (Fu & Sproule, 1995), rather than the conventional values 0 or 1, with mixing probability ρ_i. Taking this into account, Y_i, the number of successes in n_i trials of Bernoulli, $i = 1, 2, \ldots, m$, is the sum of equicorrelated binary responses with a probability of success constant p_i and a common correlation coefficient equal to ρ_i. Thus, $Y_i = \sum_{j=1}^{n_i} W_{ij}$, where $W_{ij} = 0, 1$, $j = 1, \ldots, n_i$, is a binary variable with $E(W_{ij}) = p_i$, $\mathrm{Var}(W_{is}) = \mathrm{Var}(W_{it}) = p_i(1-p_i)$ and $\mathrm{Corr}(W_{is}, W_{it}) = \rho_i$, for all s and t, $s \neq t$. The probability distribution of Y_i, given n_i, p_i and ρ_i is then given by

$$P(Y_i = y_i | n_i, p_i, \rho_i) = \binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i - y_i}(1-\rho_i) I_{A_{1_i}}(y_i) + p_i^{\frac{y_i}{n_i}}(1-p_i)^{\frac{n_i - y_i}{n_i}} \rho_i I_{A_{2i}}(y_i), \quad (2.1)$$

where $A_{1_i} = \{0, 1, \ldots, n_i\}$, $A_{2_i} = \{0, n_i\}$, $n_i \in \mathbb{N} - \{0\}$, $0 < p_i < 1$ and $0 \leq \rho_i \leq 1$. The mean and variance of Y_i are $n_i p_i$ and $p_i(1-p_i)\{n_i + \rho_i\, n_i(n_i - 1)\}$, respectively. Note that the binomial model is a particular case of the $CB(n_i, p_i, \rho_i)$ model, when $\rho_i = 0$. This distribution can be interpreted as a zero-n_i inflated distribution (see Lambert, 1992). The values zero and n_i, which occur with greater frequency than expected by binomial distribution, are captured by the modified Bernoulli distribution. The occurrence of various values zero and n_i can be explained by the positive correlation between the individuals inside the cluster.

2.1. Inference

Let $\boldsymbol{y} = (y_1, y_2, \dots, y_m)^\top$ be a set of observed values of response variables $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_m)$ and $\boldsymbol{n} = (n_1, n_2, \dots, n_m)^\top$, a vector with the cluster sizes. Then, the likelihood function of $\boldsymbol{p} = (p_1, p_2, \dots, p_m)^\top$, a vector with the success probabilities for each cluster, and $\boldsymbol{\rho} = (\rho_1, \rho_2, \dots, \rho_m)^\top$, a vector with the correlation between any two individuals within the cluster, may be written as

$$\begin{aligned}\mathcal{L}(\boldsymbol{p}, \boldsymbol{\rho}; m, \boldsymbol{n}, \boldsymbol{y}) \quad = \quad & \prod_{i=1}^{m} \Bigg\{ a_i \bigg((1-p_i)^{n_i}(1-\rho_i) + (1-p_i)\rho_i \bigg) + b_i \bigg(p_i^{n_i}(1-\rho_i) + p_i\rho_i \bigg) \\ & + (1-a_i-b_i) \bigg(\binom{n_i}{y_i} p_i^{y_i}(1-p_i)^{n_i-y_i}(1-\rho_i) \bigg) \Bigg\}, \end{aligned} \tag{2.2}$$

where $a_i = 1$ if $y_i = 0$, and $a_i = 0$ otherwise; $b_i = 1$ if $y_i = n_i$, and $b_i = 0$ otherwise. Note that a_i and b_i are known values, with $i = 1, ..., m$.

To define a correlated binomial regression model, the success probability, p_i, and the correlation parameter, ρ_i, are jointly modeled using the sets of covariates available for the clusters and for the individuals inside the clusters. The parameters p_i are modeled using the link functions Q_i, specified in Table 1, with $\eta_i = \sum_{r=0}^{k} \beta_r x_{ir}$. The coefficients $\beta_0, \beta_1, \dots, \beta_k$ are unknown regression parameters to be estimated; $x_{i0} = 1$, for all i; and $x_{1i}, x_{2i}, \dots, x_{ik}$ represent the values of the k covariates for the ith cluster.

TABLE 1
Link functions used to model p_i.

Link function	Q_i
Logito	$\exp\{\eta_{ir}\} / [1 + \exp\{\eta_{ir}\}]$
Comp. log-log	$1 - \exp\{-\exp\{\eta_{ir}\}\}$
Log-log	$\exp\{-\exp\{-\eta_{ir}\}\}$
Probito	$\Phi(\eta_{ir})$

$\Phi(\cdot)$ is the cumulative distribution function for the normal distributions.

The correlated structure is modeled considering a specific function of the available covariates that are able to relate the dependence between individuals inside the cluster. The correlated structure can be written, in general, as

$$R_i = h(v(\boldsymbol{r}_i), \gamma), \tag{2.3}$$

where $h(v(\boldsymbol{r}_i), \gamma)$, an appropriate nonlinear, monotonic and differentiable function, is the correlation between any two individuals within the ith cluster; $v(\boldsymbol{r}_i)$ represents a function of the individual covariates values, assuming positive values; $\boldsymbol{r}_i = (r_{i11}, \dots, r_{i1n_i}, r_{i21}, \dots, r_{i2n_i}, r_{iq1}, \dots, r_{iqn_i})^\top$, with r_{ilj} representing the value of the lth covariate for jth individual inside the ith cluster, $i = 1, \dots, m$, $l = 1, \dots, q$ and $j = 1, \dots, n_i$; γ is the parameter which determines the rate of decay of correlation as a function of $v(\boldsymbol{r}_i)$ (Sherman, 2011). Using spatial ideas of correlation structures, the possible choices of the function $v(\boldsymbol{r}_i)$ can be made considering, for instance, continuous functions of some distance between position vectors or between other available vectors which allow us to characterize the relationship among the individuals within the cluster (Sherman, 2011). Therefore, candidates for $v(\boldsymbol{r}_i)$, using only the covariates r_{i1} and r_{i2}, could be the Euclidean distance metric, defined as $\sqrt{\sum_{l=1,2} \sum_s \sum_{s<t} (r_{ils} - r_{ilt})^2}$, the Manhattan distance, defined as $\sum_{l=1,2} \sum_s \sum_{s<t} |r_{ils} - r_{ilt}|$, maximum distance, defined as $\max_{s,t} |r_{i1s} - r_{i1t}|$, minimum distance as $\min_{s,t} |r_{i2s} - r_{i2t}|$, with $s, t = 1, \dots, n_i$.

Subsequently, the likelihood can be rewritten as a function of the regression coefficients $\beta_0, \beta_1, \dots, \beta_k$, associated with the covariates, and of the coefficient γ, associated with the correlated structures. Let the observed data set be $D = (m, \boldsymbol{n}, \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{r})^\top$, where $\boldsymbol{n} = (n_1, n_2, \dots, n_m)^\top$, $\boldsymbol{y} = (y_1, y_2, \dots, y_m)^\top$, $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_m)^\top$, $\boldsymbol{x}_i = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{ik})^\top$, and $\boldsymbol{r} = (\boldsymbol{r}_{11}, \dots, \boldsymbol{r}_{1q}, \boldsymbol{r}_{21}, \dots, \boldsymbol{r}_{mq})^\top$. Using a link function Q_i and a correlated structure R_i, the likelihood function (2.2) can be expressed as a function for $\boldsymbol{\theta} = (\beta_0, \beta_1, \dots, \beta_k, \gamma)^\top$. Thus,

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; D) \quad = \quad & \prod_{i=1}^{m} \Bigg\{ a_i \Bigg((1 - Q_i)^{n_i}(1 - R_i) + (1 - Q_i)R_i \Bigg) + b_i \Bigg(Q_i^{n_i}(1 - R_i) + Q_i R_i \Bigg) \\ & + (1 - a_i - b_i) \Bigg(\binom{n_i}{y_i} Q_i^{y_i}(1 - Q_i)^{n_i - y_i}(1 - R_i) \Bigg) \Bigg\}, \end{aligned} \tag{2.4}$$

where $a_i = 1$ if $y_i = 0$, and $a_i = 0$ otherwise; $b_i = 1$ if $y_i = n_i$, and $b_i = 0$ otherwise, with $i = 1, ..., m$. When R_i assumes value zero, we need to consider $R_i = \zeta$, where ζ is a fixed value very close to zero.

The maximum likelihood estimators can be obtained by direct maximization of the log-likelihood function ($\ell(\boldsymbol{\theta}; D) = \log \mathcal{L}(\boldsymbol{\theta}; D)$) using for instance the BFGS algorithm (Nocedal & Wright (2006)). The advantage of this procedure is that it runs easily from a statistical packages such R (Team (2008)). The code in R used in this procedure is available by request from the first author or on the website
http://www.ufscar.br/~des/docente/carlos/Dados/MRBC_EMV.txt.

3. Diagnostics

Two different types of residuals, the standardized residual and the deviance residual, and two global influence measures, the generalized Cook's distance and the likelihood distance are considered to identify the presence of outliers and/or influential observations. To check the underlying model assumption in which the response variables follow a correlated binomial distribution $\text{BC}(n_i, p_i, \rho_i)$, with a positive correlation between the Bernoulli variables in the cluster, $\rho_i > 0$, the significance of the correlated structure parameter γ is observed using confidence intervals obtained in the inferential process. If $\gamma = 0$ or $\gamma = 1$, the usual binomial regression model can be considered in the analysis.

3.0.1. Standardized residuals

The standardized residual for the correlated binomial regression model is defined as

$$r_i = \frac{y_i - n_i \hat{p}_i}{\sqrt{\hat{p}_i(1 - \hat{p}_i)\{n_i + \hat{\rho}_i n_i(n_i - 1)\}}}, \qquad i = 1, \dots, m, \tag{3.1}$$

where $\hat{p_i} = \hat{Q}_i$, $\hat{\rho_i} = \hat{R}_i$ and $\hat{\gamma}$ and $\hat{\boldsymbol{\beta}}$ are, respectively, the maximum likelihood estimates of the parameters γ and $\boldsymbol{\beta}$.

3.1. Deviance residuals

The deviance residual for the correlated binomial regression model is defined as

$$r_i^d = \text{signal}(y_i - n_i \hat{p}_i) \sqrt{2\ell(y_i; D, \hat{\gamma}) - 2\ell(\hat{\boldsymbol{\beta}}; D, \hat{\gamma})}, \tag{3.2}$$

where $\hat{p}_i$, $\hat{\rho}_i$, $\hat{\gamma}$ and $\hat{\boldsymbol{\beta}}$ are as in the previous case, $\ell(y_i; D, \hat{\gamma})$ is the saturated log-likelihood function, with $\hat{p}_i = y_i/n_i$ and the correlated structure parameter γ substituted by the maximum likelihood estimate $\hat{\gamma}$ and $\ell(\hat{\boldsymbol{\beta}}; D, \hat{\gamma})$ is the log-likelihood function evaluated at the maximum likelihood estimates.

3.2. Global influences

Two metrics can be used for assessing the influence on a correlated binomial regression model: the generalized Cook's distance and the likelihood distance (Zhu *et al.*, 2001). These methodologies are effective when there is only one outlier (She & Owen, 2011). She and Owen She & Owen (2011) suggest an alternative method for the presence of multiple outliers. However, this tool was not adapted to our model yet.

3.2.1. Generalized Cook distance

The generalized Cook distance (Zhu *et al.*, 2001) can be used to quantify the impact of the ith observation on the maximum likelihood estimator of $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$. It is given by

$$C_i = \left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right)^\top J(\hat{\boldsymbol{\theta}}) \left(\hat{\boldsymbol{\theta}}_{(-i)} - \hat{\boldsymbol{\theta}}\right),$$

where $\hat{\boldsymbol{\theta}}_{(-i)}$ is the maximum likelihood estimator of $\boldsymbol{\theta}$ based on $\mathcal{L}(\boldsymbol{\theta}; D)$ with the ith observation $(n_i, y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)^\top$ deleted and $J(\hat{\boldsymbol{\theta}})$ is the Fisher observed information matrix. When the number of clusters, m, is large, Cook & Weisberg (1982) suggest the following approximation for $\hat{\boldsymbol{\theta}}_{(-i)}$:

$$\hat{\boldsymbol{\theta}}_{(-i)} = \hat{\boldsymbol{\theta}} + J(\hat{\boldsymbol{\theta}})^{-1} U(\hat{\boldsymbol{\theta}}_{(-i)}), \tag{3.3}$$

where

$$U(\hat{\boldsymbol{\theta}}_{(-i)}) = \left.\frac{\partial \ell(\boldsymbol{\theta}; D_{(-i)})}{\partial \boldsymbol{\theta}_{(-i)}}\right|_{\boldsymbol{\theta}_{(-i)} = \hat{\boldsymbol{\theta}}_{(-i)}}.$$

The vector of scores, $U(\boldsymbol{\theta}_{(-i)})$, with the ith observation deleted, has a dimension $(k+2)$. Using the approximation present in (3.3), the generalized Cook distance is rewritten as

$$C_i = U(\hat{\boldsymbol{\theta}}_{(-i)})^\top J(\hat{\boldsymbol{\theta}}) U(\hat{\boldsymbol{\theta}}_{(-i)}).$$

3.2.2. Likelihood distance

The likelihood distance (Zhu *et al.*, 2001) can also be used to measure the difference between $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{(-i)}$. This natural measure is given by:

$$LD_i = 2\left\{\ell(\hat{\boldsymbol{\theta}}; D) - \ell(\hat{\boldsymbol{\theta}}_{(-i)}; D)\right\}, \tag{3.4}$$

where $\ell(\hat{\boldsymbol{\theta}}; D)$ and $\ell(\hat{\boldsymbol{\theta}}_{(-i)}; D)$ are the log-likelihood functions evaluated at the usual maximum likelihood estimate, $\hat{\boldsymbol{\theta}}$, and at the maximum likelihood estimate with the ith observation $(n_i, y_i, \boldsymbol{x}_i, \boldsymbol{r}_i)^\top$ deleted, $\hat{\boldsymbol{\theta}}_{(-i)}$, respectively. Note that, as $\ell(\boldsymbol{\theta}; D)$, for fixed D, is maximized for $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, for whatever any other $\boldsymbol{\theta} \neq \hat{\boldsymbol{\theta}}$, $\ell(\boldsymbol{\theta}; D)$ will be less than $\ell(\hat{\boldsymbol{\theta}}; D)$, so the expression in (3.4) is always positive.

The ith observation is considered as influential if the value of the generalized Cook distance or the likelihood distance is large. This value can be compared to the critical points of the χ^2_{k+2} distribution.

4. Beta-binomial regression models

In order to compare the results of the fitted model proposed in this work with those given by other methods, an analysis is considered for the beta-binomial regression models (Prentice, 1986; Lindsey & Altham, 1998).

Assuming p arises from a conjugate beta distribution, $Beta(\alpha_1, \alpha_2)$, $\alpha_1 > 0$ and $\alpha_2 > 0$, and the parameterization $p = \alpha_1/(\alpha_1 + \alpha_2)$ and $\rho = 1/(\alpha_1 + \alpha_2 + 1)$, such that $\alpha_1 = p/\zeta$ and $\alpha_2 = (1-p)/\zeta$, where $\zeta = \rho/(1-\rho)$. The beta-binomial distribution can be written as

$$P(Y = y | n, p, \zeta) = \binom{n}{y} \prod_{j=0}^{y-1} (p + \zeta j) \prod_{j=0}^{n-y-1} ((1-p) + \zeta j) \left[\prod_{j=0}^{n-1} (1 + \zeta j)\right]^{-1}, \tag{4.1}$$

where $\prod_{j=0}^{x} c_j = 0$, for any $x < 0$, $y = 0, 1, \ldots, n$, $n \in \mathbb{N} - \{0\}$, $0 < p < 1$ and $-1 \leq \rho \leq 1$. The mean and variance of this model are $E(Y) = np$ and $\mathrm{Var}(Y) = np(1-p)(1 + (n-1)\rho)$ (Prentice, 1986).

Let $y_1, y_2, \ldots, y_m$ be a set of observed values of $Y_1, Y_2, \ldots, Y_m$. The log-likelihood function is given by

$$\ell_{BB}(\boldsymbol{p}, \boldsymbol{\zeta}; m, \boldsymbol{n}, \boldsymbol{y}) = \sum_{i=1}^{m} \left\{ \log \binom{n_i}{y_i} + \sum_{j=0}^{y_i-1} \log (p_i + \zeta_i j) + \sum_{j=0}^{n_i-y_i-1} \log ((1-p_i) + \zeta_i j) - \sum_{j=0}^{n_i-1} \log (1 + \zeta_i j) \right\}, \qquad (4.2)$$

where $p_i = Q_i$ and $\rho_i = R_i$.

The maximum likelihood estimators are obtained by direct maximization of the log-likelihood function (4.2). The observed Fisher information matrix is similar to that presents in Prentice (1986).

In the next section, the results of the fitted correlated binomial regression model, using the complementary log-log link function and continuous AR correlated structure, are compared with these given by the fitted Prentice models.

5. Health plan data set

We consider a health plan operator problem in Brazil for which the data are available from http://www.ufscar.br/~des/docente/carlos/Dados/Dados2.txt (information about health plans in Brazil can be found at http://www.ans.gov.br/). The health plan operator needs to determine the probability of high-cost health service occurrence in a company, which would be taken into consideration at the time of renewing or not of the health plan. This problem was discussed in the introduction section as our motivating example of overdispersed binomial data modeling. Two covariates are considered in the analysis: the average number of medical appointments per employee, x_{i1}, and the average cost of a medical test, x_{i2}. The covariate number of days between the beginning of the plan period and the first high-cost health service occurrence per each employee, r_{ij}, is used to account for the dependence between the Bernoulli variables inside the company. In fact, we consider the variable $\min_{s,t}|r_{is} - r_{it}|$, the minimum of days between the employee s and t, which assumes values between zero, where both employees use the service on the same day, and 365, where there is no use of the plan by the employees. This variable is standardized in the interval [0,1] by the transformation $\min_{s,t}|r_{is} - r_{it}| = \min_{s,t}|r_{is} - r_{it}|/365$. It is intuitive to assume that the greater the difference between the times of using the plan, the lower the relation between the use of the service. For this reason, the continuous AR correlated structure, given by $R_i = \gamma^{\frac{\min_{s,t}|r_{is}-r_{it}|}{365}}$, with $i = 1, \ldots, m$ and $s, t = 1, \ldots, n_i$, is considered in the analysis. A correlated binomial regression model is fitted to the data for each of the four link functions. The results obtained by the model selection method *AIC* and *BIC* are shown in Table 2. As can be observed by these results, the model with the complementary log-log link function is identified as the best choice.

Table 2
AIC and BIC values for correlated binomial regression fitted models with different link functions.

Criterion	Logit	Complementary log-log	Log-log	Probit
AIC	397.861	397.811	398.390	398.143
BIC	410.161	410.112	410.691	410.444

The maximum likelihood estimators of the parameters for the fitted model with complementary log-log link function are shown in Table 3. Note that the confidence intervals of the correlated structure parameter, γ, does not contain zero, corroborating with the fact that a correlated binomial regression model should be used in the analysis.

Table 3
The maximum likelihood estimators (MLE) and asymptotic confidence interval (ACI) of the parameters γ, β_0, β_1 and β_2, for the business health plan data set.

Parameter	γ	β_0	β_1	β_2
MLE	0.223	-3.833	0.206	0.322
95% ACI	(0.121 , 0.325)	(-4.411 , -3.254)	(0.032 , 0.381)	(0.161 , 0.484)

The assumption of independence and the presence of outliers can be observed by examining the residual plotted in time order, if the order is available. The standardized residuals, based on the predicted values of Y_i, and the deviance residuals, based on the log-likelihood function, are presented in Figures 1*a* and 1*b*. Figure 1*b* indicates cases 36, 43 and 85 as possible outliers. The model specification and, again, the presence of outliers are observed by examining the residuals plotted against predicted values. Both plots indicate a good specification of the model. The greatest values of the generalized Cook distance, 1.147, and the likelihood distance, 0.557, are given by case 85.

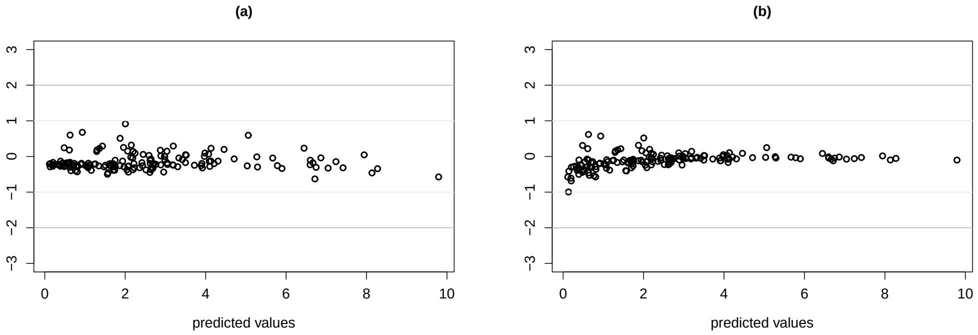

FIGURE 1. *(a) Standardized residuals versus predicted values; (b) Deviance residual versus predicted values.*

To reinforce this need of using the proposed model in this data set, we also fitted four other models, the usual binomial regression model and negative binomial regression model fitted using the complementary log-log link function, the usual Poisson regression model fitted using the log link function, and the beta-binomial regression model fitted using the complementary log-log link function and the continuous AR correlated structure. The binomial regression fitted model had an AIC value of 538.495 and BIC value of 547.720, the poisson regression fitted model had an AIC value of 668.343 and BIC value of 677.569, the negative binomial regression fitted model had an AIC value of 466.431 and BIC value of 478.731, the beta-binomial regression fitted model had an AIC value of 440.545 and BIC value of 452.846, while the correlated binomial regression fitted model had an AIC value of 397.811 and BIC value of 410.112. Besides the smallest AIC and BIC in its favor, the analysis described in this section indicates that the alternative model, the correlated binomial regression model fitted using the complementary log-log link function and the continuous AR correlated structure, provides a very good fit for this data set.

The decision regarding the renewal of contracts, based on the analysis conducted in this work (considering the complete data set), establishes that the probability of high-cost health service in the i*th* company for this real data set is given by $\hat{p}_i = 1 - \exp\{-\exp\{-3.833 + 0.206x_{i1} + 0.322x_{i2}\}\}$, with x_{i1}: average number of medical appointments per employee and x_{i2}: average cost of medical tests. The correlation between any two individuals within the ith company for this real data set is given by $\hat{\rho}_i = 0.223^{v(\boldsymbol{r_i})}$, with $v(\boldsymbol{r_i})$ the variable the minimum of days between the employees / 365.

6. Conclusions

A usual methodology to model a data set, whose response variable is the frequency of events, is the binomial regression model. An important assumption in the setting of the binomial regression model is the independence among the Bernoulli trials. For situations when this independence is not feasible, we propose a correlated binomial regression model. This regression structure, besides modeling the probability of success of an event of interest in a particular cluster, allows us to insert a correlated structure to model the dependence between the Bernoulli trials within the clusters. In the present article, a correlated binomial regression model is proposed to model the probability of high-cost health service used in a real data set. The developed model can make the analysis more realistic in the sense that it assumes that the employees inside the company are not necessarily independent.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Altham, P. M. E. (1978). Two generalizations of the binomial distribution. *Journal of the Royal Statistical Society. Series C*, **27**(2), 162–167.

Cook, R. & Weisberg, S. (1982). *Residuals and influence in regression*. Monographs on statistics and applied probability. Chapman and Hall, London.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Wiley-Interscience, New York.

Diniz, C. A. R., Tutia, M. H. & Leite, J. G. (2010). Bayesian analysis of a correlated binomial model. *Brazilian Journal of Probability and Statistics*, **24**(1), 68–77.

Efron, B. (1986). Double exponential families and their use in generalized linear regression. *Journal of the American Statistical Association*, **81**(395), 709–721.

Fu, J. & Sproule, R. (1995). A generalization of the binomial distribution. *Communications in Statistics - Theory and Methods*, **24**(10), 2645–2658.

Jennrich, R. I. & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, **42**(4), 805–820.

Lambert, D. (1992). Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics*, **34**(1), 1–14.

Lindsey, J. K. (1995). *Modelling frequency and count data*. Oxford Statistical Science. Oxford University.

Lindsey, J. K. & Altham, P. M. E. (1998). Analysis of the human sex ratio by using overdispersion models. *Journal of the Royal Statistical Society. Series C*, **1**(47), 149–157.

Luceño, A. (1995). A family of partially correlated poisson models for overdispersion. *Computational Statistics and Data Analysis*, **20**(5), 511–520.

Luceño, A. & De Ceballos, F. (1995). Describing extra-binomial variation with partially correlated models. *Communications in Statistics - Theory and Methods*, **24**(6), 1637–1653.

McCullagh, P. & Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, second edition.

Nocedal, J. & Wright, S. J. (2006). *Numerial Optimization*. Springer-Verlag, New York, second edition.

Ochi, Y. & Prentice, R. L. (1984). Likelihood inference in a correlated probit regression model. *Biometrika*, **71**(3), 531–543.

Pires, R. M. & Diniz, C. A. R. (2012). Correlated binomial regression models. *Computational Statistics and Data Analysis*, **56**(8), 2513–2525.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *Journal of the American Statistical Association*, **81**(394), 321–327.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**(4), 1033–1048.

Russell, D. W. (1996). Heterogeneous variance: Covariance structures for repeated measures. *Journal of Agricultural, Biological, and Environmental Statistics*, **1**(2), 205–230.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**(2), 461–464.

She, Y. & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*, **106**(494), 626–639.

Sherman, M. (2011). *Spatial Statistics and Spatio-Temporal Data: Covariance Functions and Directional Properties*. Wiley Series in Probability and Statistics. John Wiley and Sons.

Team, R. D. C. (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Journal of the Royal Statistical Society. Series C*, **31**(2), 144–148.

Zhu, H., Lee, S.-Y., Wei, B.-C. & Zhou, J. (2001). Case-deletion measures for models with incomplete data. *Biometrika*, **88**(3), 727–737.

Zimmerman, D. L. & Harville, D. A. (1991). A random field approach to the analysis of field-plot experiments and other spatial experiments. *Biometrics*, **47**(1), 223–239.