Quasi-beta Longitudinal Regression Model Applied to Water Quality Index Data

Ricardo Rasmussen PETTERLE, Wagner Hugo BONAT, and Cassius Tadeu SCARPIN

We propose a new class of regression models to deal with longitudinal continuous bounded data. The model is specified using second-moment assumptions, and we employ an estimating function approach for parameter estimation and inference. The main advantage of the proposed approach is that it does not need to assume a multivariate probability distribution for the response vector. The fitting procedure is easily implemented using a simple and efficient Newton scoring algorithm. Thus, the quasi-beta longitudinal regression model can easily handle data in the unit interval, including exact zeros and ones. The covariance structure is defined in terms of a matrix linear predictor composed by known matrices. A simulation study was conducted to check the properties of the estimating function estimators of the regression and dispersion parameter estimators. The NORTA algorithm (NORmal To Anything) was used to simulate correlated beta random variables. The results of this simulation study showed that the estimators are consistent and unbiased for large samples. The model is motivated by a data set concerning the water quality index, whose goal is to investigate the effect of dams on the water quality index measured on power plant reservoirs. Furthermore, diagnostic techniques were adapted to the proposed models, such as DFFITS, DFBETAS, Cook's distance and half-normal plots with simulated envelope. The R code and data set are available in the supplementary material.

Key Words: Unit interval; Longitudinal data; Estimating function; Diagnostic techniques; Simulation study; NORTA algorithm.

1. INTRODUCTION

In many areas of research, it is common to analyze data with outcomes limited to the unit interval. These variables usually appear in the form of rates, proportions, index and percentages, being therefore limited to the interval (0, 1). For analysis of continuous bounded response variables, the beta (Ferrari and Cribari-Neto 2004) and simplex regression models (Barndorff-Nielsen and Jørgensen 1991) are usual choices. Further models have been

Ricardo Rasmussen Petterle (✉), Sector of Health Sciences, Medical School, Paraná Federal University, Curitiba, PR, Brazil (E-mail: *ricardopetterle@ufpr.br*). Wagner Hugo Bonat, Department of Statistics, Paraná Federal University, Curitiba, PR, Brazil. Cassius Tadeu Scarpin, Research Group of Technology Applied to Optimization (GTAO), Paraná Federal University (UFPR), Curitiba, Brazil.

Journal of Agricultural, Biological, and Environmental Statistics
https://doi.org/10.1007/s13253-019-00360-8

Published online: 12 March 2019

proposed, such as the unit gamma (Mousa et al. 2016) and Johnson S_B regression models (Lemonte and Bazán 2016). Additionally, Mitnik and Baek (2013) present a regression model based on Kumaraswamy distribution, which allows modeling the median of the response variable as a function of covariates. Recently, Bonat et al. (2018c) proposed a new class of regression models, based on second-moment assumptions, with variance in the form $\phi\mu^p(1-\mu)^p$, where μ is the expectation of the response variable and ϕ and p are the dispersion and power parameters, respectively.

Although the aforementioned models can be used in many applications, they are limited to the analysis of independent data. In many cases, it is common to analyze data resulting from experiments in which one or more response variables are measured repeatedly in a group of individuals (Fitzmaurice et al. 2008; Verbeke et al. 2014). Experiments with these characteristics are called longitudinal studies and are often performed in several areas of knowledge, examples including agriculture (Menarin et al. 2017), education (Kaya and Leite 2017) and medicine (Hunger et al. 2012; Mohd Din et al. 2014). In general, a longitudinal study allows us to evaluate changes in the response variable over time, in addition to investigating the effect of covariates (Diggle et al. 2002). Furthermore, we can describe the structure of correlation through the specification of a covariance matrix (Diggle et al. 2002; Fitzmaurice et al. 2011).

Thus, for the analysis of such data, it is essential that the proposed model considers the longitudinal and/or clustered nature of the data. Therefore, the orthodox generalized linear models (Nelder and Wedderburn 1972) are not suitable, since they assume independence between observations. In the last decades, a variety of methods have been proposed for the analysis of dependent data. Liang and Zeger (1986) and Zeger et al. (1988) proposed the popular method of generalized estimation equations, while Breslow and Clayton (1993) presented the traditional generalized linear models with random effects. For more details on these methods and their extensions, see Verbeke and Molenberghs (2001); Diggle et al. (2002); Molenberghs and Verbeke (2006) and Fitzmaurice et al. (2008).

For the analysis of longitudinal continuous bounded data, Song and Tan (2000), Song et al. (2004), Qiu et al. (2008) and Bonat et al. (2018b) presented a class of marginal models based on the simplex distribution, with fixed and varying dispersion. Verkuilen and Smithson (2012) use the beta regression model with random effects in the analysis of experiments in cognitive psychology, while Hunger et al. (2012) shows applications in medical research. Bonat et al. (2015b) discussed maximum likelihood inference for beta mixed models. Under the Bayesian paradigm, the beta mixed models are discussed in Figueroa-Zúñiga et al. (2013) and Bonat et al. (2015a). Masarotto and Varin (2012) proposed a class of marginal models based on Gaussian copulas, including the beta marginal model. Under the time-series framework, some models have been proposed (Grunwald et al. 1993; McKenzie 1985; Rocha and Cribari-Neto 2008; da Silva et al. 2011; Bayer et al. 2017). Recently, Zhao et al. (2018) proposed a partially linear additive model for analysis of correlated bounded data, while Zheng et al. (2017) used a similar approach to analyze quality of life data from cancer research.

Although the aforementioned methodology are used in numerous applications, they are limited. First, these models need specific algorithms that are not always available in non-commercial software packages. Second, estimation algorithms are computationally demanding, especially those used in linear (generalized) models with random effects, which require

methods to solve high-dimensional integrals. Third, a distributional assumption is required for the outcomes, which implies the selection of different distributions. Finally, it is not easy to specify the covariance structure to take into account different sources of dependence.

The main goal of this paper is to propose a regression model to deal with continuous bounded data in studies with repeated measures and clustered data. The longitudinal quasi-beta regression model is specified using second-moment assumptions, and parameter estimation is done based on the approach proposed by Bonat and Jørgensen (2016) using estimating functions. Thus, the proposed approach presents several advantages over standard estimation methods. First, there is no distributional assumption for the response vector. Thus, the proposed model does not present an explicit likelihood function. Second, the fitting procedure is easy to implement and can be summarized as a simple and efficient Newton scoring algorithm. Third, the covariance structure is specified by a matrix linear predictor composed of known matrices that allow different structures to be combined. Finally, the proposed model allows us to easily accommodate data in the unit interval including exacts zeros and ones and does not require equal number of observations per group.

The main contributions of this article are: (1) proposing a regression model to deal with continuous bounded data in the context of longitudinal data analyses; (2) performing a simulation study to check the properties of the regression and dispersion parameters estimators; (3) adapting diagnostic techniques such as Cook's distance, half-normal plots with simulated envelope, DFFITS and DFBETAS; (4) analyzing the water quality index data set.

The article is organized as follows. Section 2 presents the data set. Section 3 proposes the quasi-beta longitudinal regression model, and Sect. 4 discusses the estimation and inference procedures. The results of the simulation study are presented in Sect. 5, followed by the data analysis in Sect. 6. Finally, Sect. 7 discusses the main contributions of the article and presents suggestions for future work.

2. DATA SET

The water quality index (IQA, acronym in Portuguese) was developed in the USA in 1970 to assess the quality of water intended for supply after treatment. The IQA is calculated by means of nine physical–chemical and biological parameters, considered fundamental for water quality assessment (Abbasi and Abbasi 2012). They are fecal coliforms, total nitrogen, water pH, biochemical oxygen demand, total phosphorus, water temperature, turbidity, total residue and dissolved oxygen. According to the state or condition of each parameter, water quality variation curves were established showing a set of mean curves with their respective weights (w) and quality values (q). Hence, the IQA is calculated by the weighted output of the quality values of each parameter, resulting in a score between zero and one hundred. The higher the score, the better the water quality. Thus, the IQA calculation is defined by Eq. 1:

$$\text{IQA} = \prod_{i=1}^{9} q_i^{w_i}, \tag{1}$$

where q_i (value between 0 and 100) corresponds to the quality of the ith parameter, obtained from the result of the mean curves and the laboratory analysis and w_i (value between 0 and

1) refers to the weight of the ith parameter, such that $\sum_{i=1}^{9} w_i = 1$. Therefore, the IQA presents an easy interpretation, mainly, for a lay public and is comparable among different locations. For these reasons, the IQA is used by various companies linked to the environment.

The energy company COPEL operates 16 hydroelectric power plants in the State of Paraná, Brazil. The main purpose of these plants is the generation of electric energy through rivers and water reservoirs. In addition to power generation, reservoir water is also used for other purposes, such as fishing, navigation, recreation, agricultural irrigation and city supply. To meet the operating specifications of these hydroelectric plants, COPEL monitors the water quality index upstream, downstream and in the reservoirs of the dammed rivers. The main goal of this monitoring is to detect changes in water quality, possibly attributable to the presence of dams.

The study was conducted in 2004 and evaluated 12 measures (3 locations $\times$ 4 quarters) for each of the 16 power plants, resulting in a total of 190 observations with only two missing data. The main goal of the data analysis was to investigate the relationship of the IQA with the locations (upstream, reservoir and downstream) controlled by the effect of the quarters and power plants. Thus, we have a longitudinal study combined with grouped data or repeated measures. The first characteristic is related to the quarters, while the second is associated with the locations and since the measures are taken in the same power plant, we expect some correlation between them.

It is important to highlight that this data set was analyzed in other studies using different regression models (Bonat et al. 2015a,b, 2018b,c) and in this article it will be used to illustrate the proposed model. Figure 1 presents a histogram and boxplots for the IQA according to the covariates. Figure 1a suggests left asymmetric distribution for IQA, while Fig. 1b indicates higher IQA values during the second and third quarters. Figure 1c indicates that the IQA was larger in the reservoir than in the other locations.

Finally, the results presented in Fig. 1d show that the IQA is not homogeneous among the power plants, with greater variation in power plants 1, 2 and 10. The results presented in Fig. 1 refer only to the descriptive and exploratory analysis of the data, where hypotheses are created that will be confirmed only after adjustment of the regression model proposed in Sect. 3.

3. QUASI-BETA LONGITUDINAL REGRESSION MODELS

In this section, we use second-moment assumptions to specify the quasi-beta longitudinal regression model. Let $\mathbf{Y}_i = (\mathrm{Y}_{i1}, \ldots, \mathrm{Y}_{ij})^\top$ be a $J \times 1$ random vector and let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{ij})^\top$ be its corresponding vector of expected values for $j = 1, \ldots, J$ observations within the group $i = 1, \ldots, n$. Thus, the expectation of the jth observation within the ith group is given by

$$\mu_{ij} = g^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $g(\cdot)$ is a known link function, $\mathbf{x}_i$ and $\boldsymbol{\beta}$ are $J \times 1$ vectors of known covariates and unknown regression parameters, respectively. Thus, the quasi-beta longitudinal regression

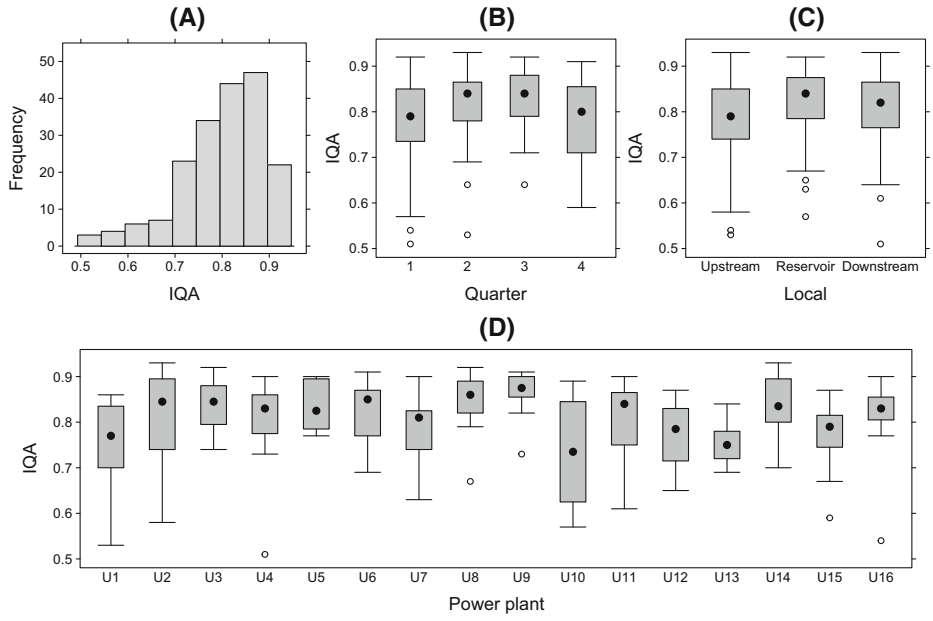

Figure 1. Histogram (**a**) and boxplots for the water quality index (IQA) by quarter (**b**), local (**c**) and power plant (**d**).

model is specified by

$$\begin{aligned} \mathrm{E}(\mathbf{Y}_i) &= \boldsymbol{\mu}_i \\ \mathrm{Var}(\mathbf{Y}_i) &= \boldsymbol{\Sigma}_i = \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\tau}) \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \end{aligned} \tag{2}$$

where $\boldsymbol{\Sigma}_i$ is a $J \times J$ matrix and $\mathrm{V}(\boldsymbol{\mu}_i)$ denote a diagonal matrix whose main diagonal entries are given by $\vartheta(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$. The matrix $\boldsymbol{\Omega}(\boldsymbol{\tau})$ describes the part of the covariance structure that does not depend on the expectation. The idea is similar to the generalized estimation equation method proposed by Liang and Zeger (1986) and Zeger et al. (1988), that uses a "working" correlation matrix for modeling dependent data. This structure is generally used in the analysis of longitudinal data, repeated measurements and clustered data. As this model is based only on second-moment assumptions, it can be expressed alternatively as follows

$$\begin{pmatrix} \mathrm{Y}_{i1} \\ \vdots \\ \mathrm{Y}_{ij} \end{pmatrix} \sim \bullet \left[\begin{pmatrix} \mu_{i1} \\ \vdots \\ \mu_{ij} \end{pmatrix}; \boldsymbol{\Sigma}_i \right], \quad i = 1, \ldots, n.$$

Note that the proposed model does not assume a multivariate probability distribution for the response vector, where the notation $\bullet$ replaces this assumption. In general, the covariance linear model has the following form:

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \Lambda_0 + \cdots + \tau_Q \Lambda_Q, \tag{3}$$

where Λ_q for $q = 0, \ldots, Q$ are known matrices that reflect the structure of interest and $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_Q)^\top$ is a vector of dispersion parameters. Bonat et al. (2017) following the arguments of Demidenko (2013) have shown that popular approaches for dealing with longitudinal data, such as the compound symmetry, moving average and first-order autoregressive are linear covariance models of the form (3). The specification of the matrix linear predictor (3) for some covariance structures will be discussed in the simulation study presented in Sect. 5, as well as in the data analysis in Sect. 6.

The regression model proposed in this paper follows the quasi-likelihood style presented by Wedderburn (1974), which combines the variance function of the binomial distribution with standard link functions for binary data such the logit and probit, as well as a structure of covariance specified by a linear combination of known matrices.

4. ESTIMATION AND INFERENCE

In this section, we shall present the quasi-score and Pearson estimation functions employed for the estimation of the regression and dispersion parameters, respectively. Thus, denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\tau}^\top)^\top$ a vector composed of two sets of parameters, where $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are vectors $K \times 1$ e $Q \times 1$ of parameters associated with the regression and dispersion coefficients, respectively. The quasi-score function for $\boldsymbol{\beta}$ is given by:

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{i=1}^{n} \mathbf{D}_i^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i),$$

where $\mathbf{D} = \nabla_{\boldsymbol{\beta}} \boldsymbol{\mu}_i$ is an $J \times K$ matrix and $\nabla_{\boldsymbol{\beta}}$ denote the gradient operator. The sensitivity matrix $K \times K$ of the $\psi_{\boldsymbol{\beta}}$ is given by

$$\mathrm{S}_{\boldsymbol{\beta}} = -\sum_{i=1}^{n} \mathbf{D}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i, \tag{4}$$

where the sum is element by element. In a similar way, the variability matrix $K \times K$ for $\psi_{\boldsymbol{\beta}}$ is given by

$$\mathrm{V}_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{D}_i^\top \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i.$$

The dispersion parameters are estimated based on the Pearson estimating function and according to Jørgensen and Knudsen (2004), Bonat and Jørgensen (2016), Bonat et al. (2018c) it has the following form:

$$\psi_{\boldsymbol{\tau}_q}(\boldsymbol{\tau}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \mathrm{tr} \left\{ W_{i\boldsymbol{\tau}_q} \left[\Delta_i^\top \Delta_i - \boldsymbol{\Sigma}_i \right] \right\}, \quad q = 1, \ldots, Q, \tag{5}$$

where the operator tr denotes the trace of the matrix, $\Delta_i = \mathbf{Y}_i - \boldsymbol{\mu}_i$ and $W_{i\boldsymbol{\tau}_q} = -\partial \boldsymbol{\Sigma}_i^{-1} / \partial \boldsymbol{\tau}_q$. For details on the computation of $W_{i\boldsymbol{\tau}_q}$, see Bonat and Jørgensen (2016) Section 3.1.

The entry (q, q') of the $Q \times Q$ sensitivity matrix for the dispersion parameters is given by

$$\mathrm{S}_{\boldsymbol{\tau}_{qq'}} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\tau}_q}\psi_{\boldsymbol{\tau}_{q'}}(\boldsymbol{\tau}, \boldsymbol{\beta})\right) = -\sum_{i=1}^{n}\mathrm{tr}\Big(W_{i\boldsymbol{\tau}_q}\boldsymbol{\Sigma}_i W_{i\boldsymbol{\tau}_{q'}}\boldsymbol{\Sigma}_i\Big). \tag{6}$$

In a similar way, the cross entries of the sensitivity matrix for $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are given by

$$\mathrm{S}_{\boldsymbol{\beta}_k\boldsymbol{\tau}_q} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\tau}_q}\psi_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{\tau})\right) = \mathbf{0} \tag{7}$$

and

$$\mathrm{S}_{\boldsymbol{\tau}_q\boldsymbol{\beta}_k} = \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\beta_k}}\psi_{\boldsymbol{\tau}_q}(\boldsymbol{\tau}, \boldsymbol{\beta})\right) = -\sum_{i=1}^{n}\mathrm{tr}\Big(W_{i\boldsymbol{\tau_q}}\boldsymbol{\Sigma}_i W_{i\boldsymbol{\beta}_k}\boldsymbol{\Sigma}_i\Big), \tag{8}$$

where $W_{i\boldsymbol{\beta}_k} = -\partial\boldsymbol{\Sigma}_i^{-1}/\partial\boldsymbol{\beta}_k$. The joint sensitivity matrix for the parameter vector $\boldsymbol{\theta}$ is given by

$$\mathrm{S}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{S}_{\boldsymbol{\beta}} & \mathbf{0} \\ \mathrm{S}_{\boldsymbol{\tau}\boldsymbol{\beta}} & \mathrm{S}_{\boldsymbol{\tau}} \end{pmatrix},$$

whose entries are defined by Eqs. (4)–(8).

The asymptotic variance of the estimating function estimators denoted by $\hat{\boldsymbol{\theta}}$ is obtained by the inverse of the Godambe information matrix, whose general form is $\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-\top}$, where $^{-\top}$ denotes inverse transpose, i.e., $\mathrm{S}_{\boldsymbol{\theta}}^{-\top} = (\mathrm{S}_{\boldsymbol{\theta}}^{-1})^{\top}$. The variability matrix for $\boldsymbol{\theta}$ has the form

$$\mathrm{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{V}_{\boldsymbol{\beta}} & \mathrm{V}_{\boldsymbol{\beta}\boldsymbol{\tau}} \\ \mathrm{V}_{\boldsymbol{\tau}\boldsymbol{\beta}} & \mathrm{V}_{\boldsymbol{\tau}} \end{pmatrix} \tag{9}$$

where $\mathrm{V}_{\boldsymbol{\tau}\boldsymbol{\beta}} = \mathrm{V}_{\boldsymbol{\beta}\boldsymbol{\tau}}^{\top}$ and $\mathrm{V}_{\boldsymbol{\tau}}$ depend on the third and fourth moments of $\mathbf{Y}_i$, respectively. In order to avoid such a dependence on high-order moments, we adopted the approach presented in Bonat and Jørgensen (2016) based on their empirical version. For details and equations, see Bonat and Jørgensen (2016).

Let $\hat{\boldsymbol{\theta}}$ denotes the estimating function estimator for $\boldsymbol{\theta}$. According to Godambe and Thompson (1978), Jørgensen and Knudsen (2004) the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} \sim \mathcal{N}(\boldsymbol{\theta}, \mathrm{J}_{\boldsymbol{\theta}}^{-1}),$$

where $\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-\top}$ is the inverse of the Godambe information matrix.

The chaser algorithm was proposed by Jørgensen and Knudsen (2004) to solve the system of equations $\psi_{\boldsymbol{\beta}} = \mathbf{0}$ and $\psi_{\boldsymbol{\tau}} = \mathbf{0}$ it is given by:

$$\begin{aligned} \boldsymbol{\beta}^{(i+1)} &= \boldsymbol{\beta}^{(i)} - \mathrm{S}_{\boldsymbol{\beta}}^{-1}\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)}, \boldsymbol{\tau}^{(i)}) \\ \boldsymbol{\tau}^{(i+1)} &= \boldsymbol{\tau}^{(i)} - \alpha\mathrm{S}_{\boldsymbol{\tau}}^{-1}\psi_{\boldsymbol{\tau}}(\boldsymbol{\beta}^{(i+1)}, \boldsymbol{\tau}^{(i)}). \end{aligned} \tag{10}$$

The main feature of the chaser algorithm is the insensitivity property (7), which allows us to separate $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ onto two equations to be updated in each step. In addition, the α parameter was introduced as a tuning constant to control the step length. Therefore,

the quasi-beta longitudinal regression model is fitted by the flexible algorithm presented by Bonat and Jørgensen (2016), in the context of multivariate covariance generalized linear models (McGLMs), which was also adapted for the analysis of count data using Poisson–Tweedie regression models (Bonat et al. 2018a; Petterle et al. 2019; Bonat et al. 2017) and for genetic data analysis presented in Bonat (2017). The computational implementation of the model was performed in the `R` statistical software (R Core Team 2018) and is available in the `mcglm` (Bonat 2016, 2018) package. The `R` code and data set are available as a supplementary material.[1]

5. SIMULATION STUDY

In this section, we present the main results of a simulation study conducted to verify the properties of the estimating function estimators for the regression and dispersion parameter estimators in the context of regression models for continuous bounded data in longitudinal studies. We designed a simulation study with 36 scenarios taking into account different covariance structures, correlation and dispersion levels.

The scenarios were designed by combining three covariance structures exchangeable, moving average and distance with three levels of correlation and four levels of dispersion. We fixed the dispersion parameter of the beta distribution at the values $\phi = (0.666, 4, 9, 23.99)$. The linear predictor of the mean structure was specified by

$$g(\mu_{ij}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},$$

where $g(\cdot)$ is the logit link function and $\boldsymbol{\beta} = (0, 2, 0.5)^\top$. The covariates x_{1i} and x_{2i} were generated from a Gaussian (mean zero and variance 0.5^2) and Bernoulli ($p = 0.5$) distributions, *respectively*. The matrix linear predictor of each covariance structure was specified by combining the identity matrix with known matrices that define the structure of interest. We consider $J = 5$ and $J = 10$, i.e., five and ten repeated measures in the same subject.

In order to describe the components of the matrix linear predictor for each case, suppose for simplicity $J = 3$. Thus, the specification of the matrix linear predictor for each of the above mentioned structures is given by:

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

in the exchangeable case, by

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1 & . \\ 1 & . & 1 \\ . & 1 & . \end{bmatrix},$$

[1] See http://www.leg.ufpr.br/doku.php/publications:papercompanions:quasibetaiqa.

in the MA1 structure, and by

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1/d_{12} & 1/d_{13} \\ 1/d_{12} & . & 1/d_{23} \\ 1/d_{13} & 1/d_{23} & . \end{bmatrix} + \tau_2 \begin{bmatrix} . & 1/d_{12}^2 & 1/d_{13}^2 \\ 1/d_{12}^2 & . & 1/d_{23}^2 \\ 1/d_{13}^2 & 1/d_{23}^2 & . \end{bmatrix},$$

for the structure based on distances (Dist^2).

For the covariance structures Exch and MA1, we define negative, null and positive correlations, while for the structure based on distances, we define null, weak and strong positive correlations. Therefore, for each of the 36 simulation scenarios, we generated 300 data sets with four sample sizes each (100, 250, 500 and 1000). For simulating correlated beta random variables, we adopt the NORTA algorithm (NORmal To Anything) (Li and Hammond 1975; Cario and Nelson 1997), available in the `NORTARA` (Su 2014) package.

In the case of the Exch covariance structure with negative correlation, we define $\tau_0 = 1.2$ and $\tau_1 = -0.2$ as the true values of the dispersion parameters, whereas for null and positive correlations the parameters were fixed at $(\tau_0 = 1, \tau_1 = 0)$ and $(\tau_0 = 0.3, \tau_1 = 0.7)$, respectively. For the covariance structures MA1 with negative, null and positive correlations the dispersion parameters were fixed at $(\tau_0 = 1, \tau_1 = -0.5)$, $(\tau_0 = 1, \tau_1 = 0)$ and $(\tau_0 = 1, \tau_1 = 0.5)$, respectively. Finally, the dispersion parameters $(\tau_0 = 1, \tau_1 = 0, \tau_2 = 0)$, $(\tau_0 = 1, \tau_1 = 0.1, \tau_2 = 0.2)$ and $(\tau_0 = 1, \tau_1 = 0.25, \tau_2 = 0.45)$ were fixed in the evaluation of the structure based on distances (Dist^2) for cases with null, weak and strong positive correlations, respectively.

Figure 2 shows the average bias plus and minus the average standard error for the dispersion parameters under each simulation scenario. In Fig. 2, the scales were standardized for each parameter by dividing the average bias and the limits of the confidence intervals by the standard error obtained from the sample of size 100. Moreover, this figure presents the results of the simulation study for $J = 5$ repeated measures. According to the results shown in Fig. 2, the estimates of τ_1 are biased under all simulation scenarios in which $\phi = 0.666$, except for the scenarios where the correlation is null. However, for the other simulation scenarios the proposed estimators for the dispersion parameters are consistent and unbiased as the sample size increases.

In the supplementary material online, we present a similar figure, which was constructed with $J = 10$ repeated measures. In addition, the results of the simulation study for the regression parameter estimates for $J = 5$ and $J = 10$ are also presented in the supplementary material. For the construction of this figure, some configurations were altered in the simulation study. In the covariance structure Exch with negative correlation, we set $\tau_0 = 1.1$ and $\tau_1 = -0.1$ as the true values of the dispersion parameters, and in the structure Dist^2, with strong positive correlation, only one value was changed from $\tau_2 = 0.45$ to $\tau_2 = 0.4$. Such modifications were necessary to ensure that the correlation matrix used in the NORTA algorithm is positive definite. Overall, the results are similar to the ones presented in this section. Basically, they show that our estimation method provides unbiased and consistent estimators for both regression and dispersion parameters for large samples and $\phi > 0.0666$.

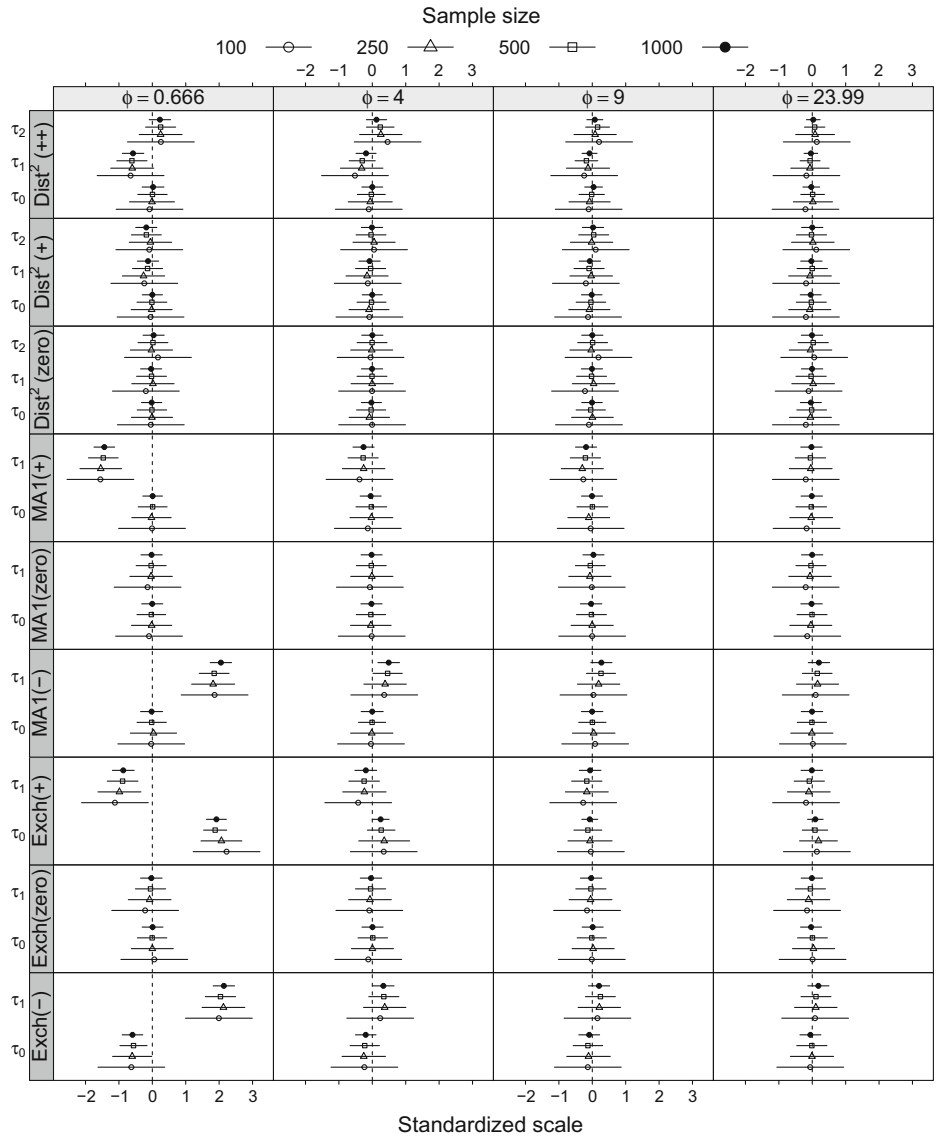

Figure 2. Average bias and confidence intervals on a standardized scale for the dispersion coefficients by sample size and covariance structures with different correlation levels.

6. APPLICATION

In this section, we present the main results of the analysis of the water quality index data presented in the Sect. 2. The data refer to the water quality index (IQA) of reservoirs of hydroelectric power plants operated by COPEL in the State of Paraná, Brazil. The main goal of the data analysis is to investigate the relationship of the IQA with the locations (upstream, reservoir and downstream) controlled by the effects of the fiscal quarters and power plants.

Therefore, the main challenge of this data analysis is the analysis of a continuous bounded response variable taking into account the longitudinal and repeated measures structures.

Let $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{i12})^\top$ be a random vector with 12 measures (3 locations $\times$ 4 quarters) associated with the water quality index of ith power plant, for $i = 1, \ldots, 16$. In this notation, $J = r \times k$, for $r = 1, 2, 3$ locations (upstream, reservoir and downstream) and $k = 1, \ldots, 4$ quarters. Let $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{i12})^\top$ be its respective vector of expected values and $g(\mu_{irk})$ the linear predictor associated with power plant i, location r and quarter k. Thus, its representation is given by:

$$g(\mu_{irk}) = \beta_0 + \beta_{1r}\, \texttt{location}_{ir} + \beta_{2k}\, \texttt{quarter}_{ik}, \tag{11}$$

where $g(\cdot) : (0, 1) \mapsto \mathbb{R}$ is a link function for bounded data, β_0 is the intercept and β_{1r}, for $r = 2$ and 3, evaluates the difference from the upstream to the reservoir and to the downstream, respectively. The coefficients β_{2k}, for $k = 2,\ 3$ and 4, measure the differences of quarter 1 to the other quarters.

Thus, the quasi-beta longitudinal regression model is defined as follows:

$$\begin{pmatrix} Y_{i1} \\ \vdots \\ Y_{i12} \end{pmatrix} \sim \bullet \left[\begin{pmatrix} \boldsymbol{\mu}_{i12} \\ \vdots \\ \boldsymbol{\mu}_{i12} \end{pmatrix} ; \boldsymbol{\Sigma}_i \right],$$

where $\boldsymbol{\Sigma}_i = \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}} \boldsymbol{\Omega}(\boldsymbol{\tau}) \mathrm{V}(\boldsymbol{\mu}_i)^{\frac{1}{2}}$ is a 12×12 matrix. In the analysis of repeated measures and longitudinal data analysis, the main interest is to model the matrix $\boldsymbol{\Omega}(\boldsymbol{\tau})$, which is the part of the covariance that does not depend on the mean structure. Thus, we propose four covariance structures for the matrix $\boldsymbol{\Omega}(\boldsymbol{\tau})$. The first structure assumes independence between observations, being composed by an identity matrix. It is important to highlight that the three structures presented below are composed of an identity matrix as well as other matrices that define the structure of interest. The second structure is known as exchangeable, defined by a matrix composed of 1's, according to Eq. 12.

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}. \tag{12}$$

The unstructured matrix used to evaluate the effect of the locations (clustered data) is given by

$$\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 \begin{bmatrix} 1 & . & . \\ . & 1 & . \\ . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1 & . \\ 1 & . & . \\ . & . & . \end{bmatrix} + \tau_2 \begin{bmatrix} . & . & 1 \\ . & . & . \\ 1 & . & . \end{bmatrix} + \tau_3 \begin{bmatrix} . & . & . \\ . & . & 1 \\ . & 1 & . \end{bmatrix}, \tag{13}$$

where τ_1 evaluates the covariance between upstream and reservoir, and τ_2 and τ_3 evaluate the covariance between upstream and downstream, and between reservoir and downstream,

Table 1. Pseudo-maximized log-likelihood values (`plogLik`), degrees of freedom (`df`) and pseudo-*Akaike* (`pAIC`) and Bayesian (`pBIC`) information criterion by covariance structures.

Structure	`plogLik`	`df`	`pAIC`	`pBIC`
Independent	212.06	7	−410.12	−387.39
Exchangeable	217.06	8	−418.12	−392.14
Unstructured 1	222.03	11	−422.06	−386.34
Unstructured 2	234.36	17	−434.72	−379.52

respectively. Therefore, the third structure, called unstructured 1, is composed of the combination of structure 2 with the specification presented in Eq. 13.

To assess the effect of the quarters (longitudinal data), the unstructured matrix has the following representation

$$\begin{aligned}\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 &\begin{bmatrix} 1 & . & . & . \\ . & 1 & . & . \\ . & . & 1 & . \\ . & . & . & 1 \end{bmatrix} + \tau_1 \begin{bmatrix} . & 1 & . & . \\ 1 & . & . & . \\ . & . & . & . \\ . & . & . & . \end{bmatrix} + \tau_2 \begin{bmatrix} . & . & 1 & . \\ . & . & . & . \\ 1 & . & . & . \\ . & . & . & . \end{bmatrix} \\ + \tau_3 &\begin{bmatrix} . & . & . & 1 \\ . & . & . & . \\ . & . & . & . \\ 1 & . & . & . \end{bmatrix} + \tau_4 \begin{bmatrix} . & . & . & . \\ . & . & 1 & . \\ . & 1 & . & . \\ . & . & . & . \end{bmatrix} + \tau_5 \begin{bmatrix} . & . & . & . \\ . & . & . & 1 \\ . & . & . & . \\ . & 1 & . & . \end{bmatrix} \\ + \tau_6 &\begin{bmatrix} . & . & . & . \\ . & . & . & . \\ . & . & . & 1 \\ . & . & 1 & . \end{bmatrix}. \end{aligned} \tag{14}$$

Note that the covariance between quarters is assessed by the coefficients τ_1 to τ_6. For example, the covariance between quarter 1 and 2 is evaluated by the coefficient τ_1. The covariance between quarter 1 and 3 and between quarter 1 and 4 is evaluated by the coefficients τ_2 and τ_3, respectively. For the other assessment, the interpretation follows similarly. Therefore, the fourth structure (unstructured 2) is defined by the combination of structure 3 with the specification shown in Eq. 14.

Then, the quasi-beta longitudinal regression model was fitted to the water quality index data set, considering the four structures mentioned above in addition to specifying the logit link function for the linear predictor (Eq. 11).

Table 1 shows the pseudo-log-likelihood values (plogLik), degrees of freedom (df) and pseudo *Akaike* (`pAIC`) and Bayesian (`pBIC`) information criterion for the proposed model, fitted under different covariance structures. The main difference between the four covariance structures is the number of parameters that are estimated, in addition to the way in which the longitudinal and clustered data are considered.

The results in Table 1 show that both the value of the pseudo-likelihood function (`plogLik` = 234.36) and the value of the pseudo *Akaike* information criterion (`pAIC`

Table 2. Wald statistics (W_s), degrees of freedom (df) and p values associated with quasi-beta longitudinal regression model.

Effects	W_s	df	p value
Location	8.56	2	0.01
Quarter	8.34	3	0.04

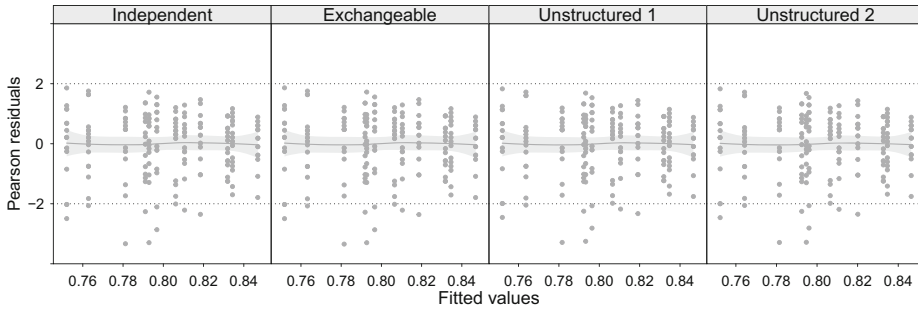

Figure 3. Pearson residuals for the quasi-beta regression models by covariance structure.

$= -434.72$) indicate the fourth structure (unstructured 2) as having the best fit to the data. Thus, this covariance structure was selected for the data analysis. The components of the matrix linear predictor for the selected covariance structure are shown in "Appendix."

Table 2 presents the results of the Wald statistic, degrees of freedom(df), and p values for each covariate that compose the linear predictor (Eq. 11).

The results in Table 2 show that both covariates are significantly different from zero and are therefore relevant in the analysis of the data. After selecting the covariance structure and analyzing the effect of the covariates, we present the residual analysis and the diagnosis, in order to evaluate the quality of the fitted model and investigate the presence of influential points or outliers.

Figure 3 presents the Pearson residuals versus fitted values for the quasi-beta longitudinal regression model, fitted using each covariance structure. In addition to the Pearson residuals, this figure also shows smoothing curves with confidence intervals estimated by loess method (Cleveland 1979).

The results in Fig. 3 indicate that for all covariance structures, the proposed model presented a satisfactory fit to the water quality index data, since the residuals vary between -2 and 2. Although some points are below the lower limit, the general fit seems suitable.

Figure 4 shows the Cook distance versus the observation index for the proposed model fitted under each covariance structure. The Cook's distance was initially proposed by Cook (1977) and was adapted in this article following the arguments of Venezuela et al. (2007). To evaluate Cook's distance, we define $2p/n$ as the cutoff, where p is the number of regression coefficients estimated by the model and n is the sample size. Thus, values above 0.063 are considered influential points.

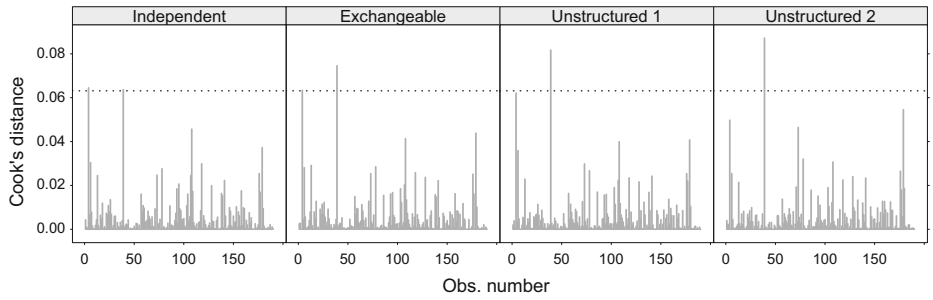

Figure 4. Cook's distance for the quasi-beta regression models by covariance structure.

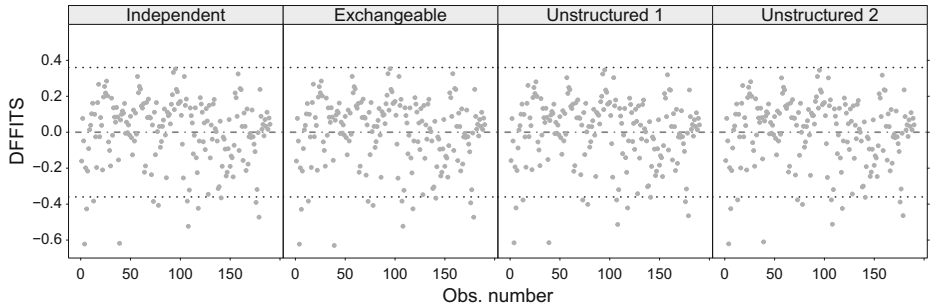

Figure 5. DFFITS for the quasi-beta regression models by covariance structure.

According to the results in Fig. 4, for independent and exchangeable structures, the observations 4 and 39 were indicated as influential points, while for others structures only the observation 39.

Figure 5 shows the measure DFFITS (Belsley et al. 1980) for the regression models fitted to the water quality index data. In general, this measure is used to evaluate the influence of the exclusion of the ith observation on its value estimated by the model. For the DFFITS measure, the cutoff is defined by $2\sqrt{p/n}$, where p and n refer to the number of regression coefficients and sample size, respectively. In this case, values outside of the range 0.355 should be investigated. Therefore, based on the DFFITS measure, there is no evidence that the fit of the four models is unsuitable.

The results associated with the measure DFBETA (Belsley et al. 1980) are shown in Fig. 6. Such a measure is commonly used to evaluate the influence of the ith observation on each regression coefficient. Its cutoff is given by $2/\sqrt{n}$, where n is the sample size. In special, for the water quality index data, values outside of the range 0.145 are considered as influential points on the regression parameters. According to the results presented in Fig. 6, there are some influential points in the adjustment of the regression models. In addition, it is observed that these points are the same in the adjustment of each model.

Figure 7 presents the half-normal plots with simulated envelope for the proposed model adjusted to the water quality index data, under each covariance structure. This diagnostic technique is commonly used to evaluate the fit of model, as well as to identify the presence of outliers. In this paper, we adapted the half-normal plots with simulated envelope based on

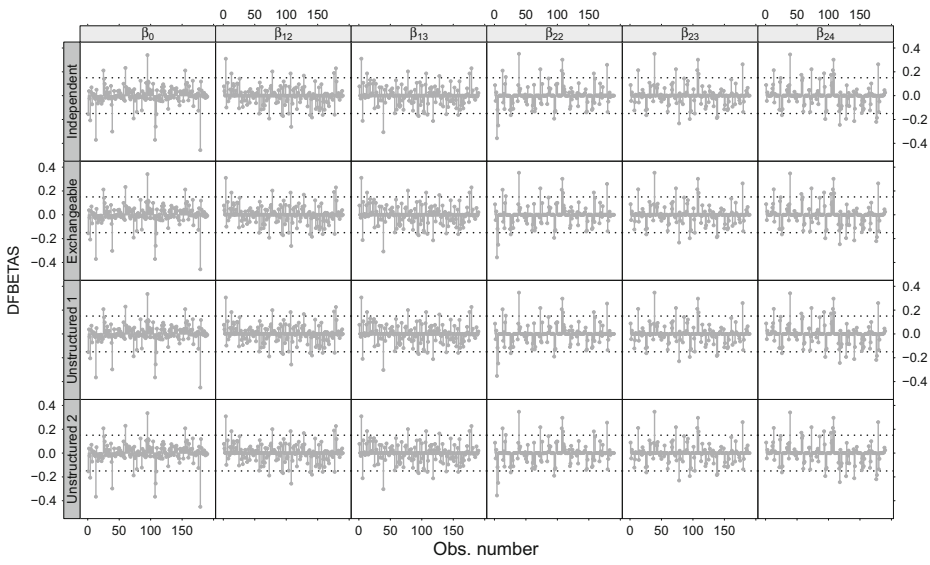

Figure 6. DFBETAS for the quasi-beta regression models by covariance structure.

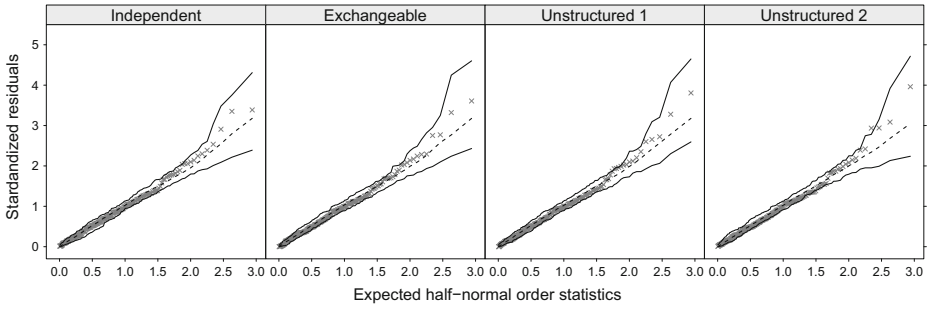

Figure 7. Half-normal plots for the quasi-beta regression models by covariance structure.

the results presented in Venezuela et al. (2007). It is important to highlight that the random variables are simulated from the beta distribution with parameters estimated by the proposed model.

The results in Fig. 7 indicate that few points were outside of the confidence intervals. These results indicate that the fit of the quasi-beta longitudinal regression model is suitable. It should be remembered that in this procedure, beta probability distribution was assumed for the simulated responses variables. Thus, it was shown that the regression model proposed in this article provides a suitable approximation for data sets generated from marginal beta distribution.

In general, both Pearson's residuals and the other measurements (Cook's distance, DFFITS, DFBETAS, and half-normal plots) indicate that the quasi-beta longitudinal regression model provides a suitable fit to water quality data.

After residual analysis and diagnosis, the interpretations of the parameters estimated by the model will be presented. Table 3 shows the estimates of the regression parameters,

Table 3. Regression parameter estimates, standard errors (SE), odds ratio (OR) with 95% confidence intervals (CI), Z-statistics and p values.

Effects	Estimates	SE	OR (CI 95%)	Z-statistics	p value
β_0: Intercept	1.111	0.103	–	10.745	<0.001
β_{12}: Reservoir	0.251	0.098	1.286 (1.060−1.559)	2.552	0.011
β_{13}: Downstream	0.163	0.102	1.178 (0.964−1.438)	1.604	0.109
β_{22}: Quarter 2	0.243	0.131	1.275 (0.985−1.650)	1.848	0.065
β_{23}: Quarter 3	0.345	0.132	1.412 (1.090−1.828)	2.617	0.009
β_{24}: Quarter 4	0.065	0.106	1.068 (0.867−1.315)	0.618	0.537

standard errors, odds ratio, p values in addition to other information associated with the proposed model. It is important to note that the odds ratio and the confidence intervals shown in Table 3 are calculated in the usual way.

According to the results in Table 1, the IQA in the reservoir was 28.6% higher than in the upstream. During quarter 3, the IQA estimates were 41.2% higher than the values obtained from quarter 1. On the other hand, IQA values for quarters 2 and 4 were similar to those from quarter 1, since the parameters associated with these effects did not present relevance (p value > 0.05). In addition, the difference in the IQA between upstream and downstream was also not significant in the data analysis (p value $= 0.109$). Therefore, the results provided by the quasi-beta longitudinal regression model are concordant with the descriptive and exploratory analysis presented in Fig. 1 (Sect. 2).

Table 4 presents the estimates of the dispersion parameters and their respective standard errors, Z-statistics and p values. The structures associated with the locations and quarters are modeled by the parameters (τ_2, τ_3, τ_4) and $(\tau_5, \tau_6, \tau_7, \tau_8, \tau_9, \tau_{10})$, respectively. For the grouped data structure, none of the parameters presented at 5% significance level. With respect to the longitudinal structure, only the parameters τ_5, τ_6 and τ_9 are significantly different from zero. Although some dispersion parameters are not relevant, the interpretation related to them will be shown below as an illustration. Therefore, the main interest in dispersion parameters is the evaluation of intra-class correlations between locations and quarters.

The correlation between upstream and the reservoir was estimated at $\hat{\rho}_{12} = 0.32$, being calculated by $(\hat{\tau}_1 + \hat{\tau}_2)/(\hat{\tau}_0 + \hat{\tau}_1)$. Thus, to make this idea more general, we constructed the correlation matrix for the effect of the locations, given by

$$\hat{\boldsymbol{\Omega}}(\boldsymbol{\tau})_{\text{Location}} = \begin{bmatrix} 1.00 & & \\ 0.32(0.10) & 1.00 & \\ 0.29(0.11) & 0.56(0.03) & 1.00 \end{bmatrix}, \tag{15}$$

where the numbers in parentheses denote the standard errors calculated by the delta method. From this matrix, it is observed that the correlation between the reservoir and the downstream was estimated at $\hat{\rho}_{23} = 0.56$. The lowest correlation was between upstream and downstream $(\hat{\rho}_{13} = 0, 29)$. Such result is expected, since the water of the river first passes through the

Table 4. Dispersion parameter estimates, standard errors (SE), Z-statistics and p values associate to the quasi-beta longitudinal regression model.

Parameters	Estimate	SE	Z-statistics	p value
τ_0	0.019	0.003	5.460	< 0.001
τ_1	0.020	0.005	3.913	< 0.001
τ_2	-0.007	0.004	-1.662	0.097
τ_3	-0.008	0.004	-1.832	0.067
τ_4	0.002	0.001	1.697	0.089
τ_5	-0.015	0.007	-2.191	0.028
τ_6	-0.015	0.006	-2.235	0.025
τ_7	-0.008	0.005	-1.648	0.099
τ_8	-0.011	0.005	-1.821	0.069
τ_9	-0.016	0.007	-2.286	0.022
τ_{10}	-0.008	0.005	-1.616	0.106

upstream, enters the reservoir leaving in the downstream direction. In other words, the distance between locations induces stronger or weaker correlations.

In the following, we present the correlation matrix for the quarter effects:

$$\hat{\boldsymbol{\Omega}}(\boldsymbol{\tau})_{\text{Quarter}} = \begin{bmatrix} 1.00 & & & \\ 0.12(0.16) & 1.00 & & \\ 0.13(0.15) & 0.24(0.14) & 1.00 & \\ 0.29(0.13) & 0.10(0.16) & 0.29(0.13) & 1.00 \end{bmatrix}. \tag{16}$$

The correlation between quarters 1 and 2 was estimated at $\hat{\rho}_{12} = 0.12$, which is obtained by $(\hat{\tau}_1+\hat{\tau}_5)/(\hat{\tau}_0+\hat{\tau}_1)$. Note that the results presented in the above matrix show weak correlations between quarters. The highest correlation was between quarter 1 and 4, estimated at $\hat{\rho}_{14} = 0.29$. This result is expected, since the quarters are cyclical and therefore quarter 1 and 4 are close.

Figure 8 shows the results of the fitted longitudinal quasi-beta regression model under different covariance structures and link functions. The results are compared by means of the regression coefficients and their respective 95% confidence intervals. The coefficient β_0 was omitted to avoid problems with the graphic scale.

Figure 8 shows that the results obtained by the four covariance structures are concordant for all link functions. The coefficients β_{13} and β_{22} were not statistically different from zero when evaluated by the unstructured 2 model, the one chosen for analysis of the data. In addition to the coefficient β_{22} not to be significant, the length of its confidence interval was greater in the covariance structure used for the data analysis, considering all the link functions evaluated. These results show the importance of evaluating different covariance structures, since the interpretations of the parameters significance varying between the structures. It is important to note that the coefficients β_{13} and β_{22} measure the effect of downstream and quarter 2, respectively. Thus, when comparing the reference categories (upstream and quarter 1) with these coefficients, their effects are zero, i.e., the effect between quarter 1 and 2 are similar, as are the effects between upstream and downstream.

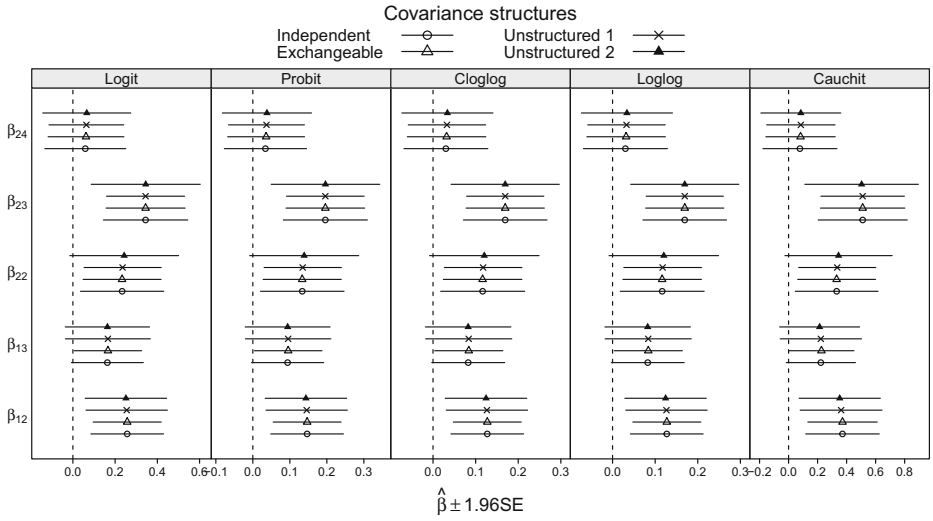

Figure 8. Parameter estimates and 95% confidence intervals for the quasi-beta regression models by link functions and covariance structures.

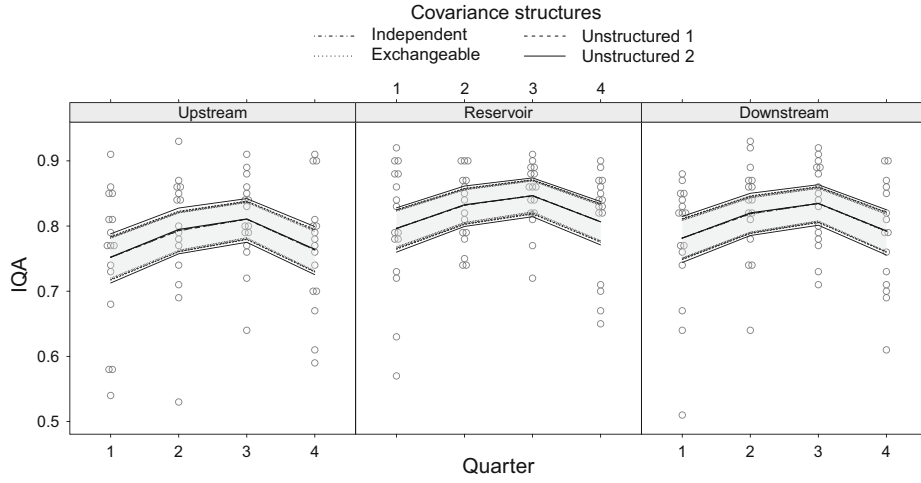

Figure 9. Curves of fitted values with 95% confidence intervals by quarters and locations for the quasi-beta regression models by covariance structures.

Figure 9 presents curves of fitted values for the expected values of the response variable with 95% confidence intervals obtained by the quasi-beta longitudinal regression model, fitted for each covariance structure.

The results in Fig. 9 show small differences between the fitted models, both in point estimates and confidence intervals. Moreover, the results presented in this figure confirm the hypotheses raised in Fig. 1 and indicate that the IQA was higher for the data collected in the third quarter, as well as in the reservoir.

7. DISCUSSION

In this paper, we presented a new class of regression models for the analysis of continuous bounded data in studies with repeated measurements and clustered data. The model is specified using second-moment assumptions, and the method used for parameter estimation and inference is based on estimating functions. Thus, the proposed approach combines the quasi-score and Pearson estimating functions for estimation of the regression and dispersion parameters, respectively. The model proposed in this article follows the quasi-likelihood style presented by Wedderburn (1974), which combines the variance function of the binomial distribution with standard link functions for binary data. In addition, the covariance structure is specified using a linear combination of known matrices. This specification allows us to consider several covariance structures, such as exchangeable, unstructured, moving averages, in addition to a structure based on distances.

The results of the simulation study showed that the proposed estimators for both regression and dispersion parameters are unbiased and consistent under different simulation scenarios. It is important to highlight that in the simulation study, we used the logit link function, since Bonat et al. (2012), Bonat et al. (2018b) show that there is no great difference in the choice of the link functions. Furthermore, in our data analysis we compared the fit of different link functions and overall the results are similar. However, in practice the logit link function has been more often used, mainly because of its easy practical interpretation. The model was motivated by the data analysis of the water quality index of reservoirs of hydroelectric power plants. The main question of the analysis was to model the covariance structure to accommodate longitudinal and clustered data. In this way, different covariance structures were proposed, which were compared by measures of goodness-of-fit proposed in Bonat (2018). The results of this evaluation showed that an exchangeable covariance structure combined with two other structures used to evaluate the longitudinal and clustered data was important in the data analysis. Furthermore, we adapted diagnostic techniques for the proposed model, showing its application and utility during data analysis. Finally, it is important to note that the results presented in Sect. 6 agree with previous analysis of the IQA data set presented, for instance in Bonat et al. (2015b) and Bonat et al. (2018b) using beta and simplex mixed models, respectively.

As future work, it is suggested to include a linear predictor combined with link functions to model the dispersion structure as a function of covariates, as well as to include a power parameter to give more flexibility in the modeling of the mean and variance relationship in the analysis of continuous bounded dependent data. We also suggest comparing our model with some existing methods (mainly the ones based on random effects) for dealing with continuous bounded data in the context of longitudinal data analyses, as well as investigating how excess zeros and ones impact marginal modeling.

[Received November 2018. Accepted March 2019.]

APPENDIX

$$
\begin{aligned}
\boldsymbol{\Omega}(\boldsymbol{\tau}) = \tau_0 &\left[\begin{array}{cccccccccccc}
1 & . & . & . & . & . & . & . & . & . & . & . \\
. & 1 & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & . & . & . & . & . \\
. & . & . & . & 1 & . & . & . & . & . & . & . \\
. & . & . & . & . & 1 & . & . & . & . & . & . \\
. & . & . & . & . & . & 1 & . & . & . & . & . \\
. & . & . & . & . & . & . & 1 & . & . & . & . \\
. & . & . & . & . & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & 1
\end{array}\right]
+ \tau_1 \left[\begin{array}{cccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{array}\right]
+ \tau_2 \left[\begin{array}{ccc|ccc|ccc|ccc}
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right] \\
&+\tau_3 \left[\begin{array}{ccc|ccc|ccc|ccc}
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\ \hline
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\ \hline
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & . \\ \hline
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & 1 & . & . & 1 & . & . & 1 & . & .
\end{array}\right]
+ \tau_4 \left[\begin{array}{ccc|ccc|ccc|ccc}
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & 1 & . & . & 1 & . & . & 1 \\
. & 1 & . & . & 1 & . & . & 1 & . & . & 1 & .
\end{array}\right]
+ \tau_5 \left[\begin{array}{cccc|cccc|cccc}
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right] + \\
&+\tau_6 \left[\begin{array}{cccc|cccc|cccc}
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right]
+ \tau_7 \left[\begin{array}{cccc|cccc|cccc}
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\ \hline
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & . \\ \hline
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
1 & . & . & . & 1 & . & . & . & 1 & . & . & .
\end{array}\right]
+ \tau_8 \left[\begin{array}{cccc|cccc|cccc}
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right] \\
&+\tau_9 \left[\begin{array}{cccc|cccc|cccc}
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & 1 & . & . & . & 1 & . & . & . & 1 & . & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & 1 & . & . & . & 1 & . & . & . & 1 & . & .
\end{array}\right]
+ \tau_{10} \left[\begin{array}{cccc|cccc|cccc}
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & 1 & . & . & . & 1 & . & . & . & 1 & . \\ \hline
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & 1 & . & . & . & 1 & . & . & . & 1 \\
. & . & 1 & . & . & . & 1 & . & . & . & 1 & .
\end{array}\right].
\end{aligned}
$$

REFERENCES

Abbasi, T. and Abbasi, S. A. (2012). *Water quality indices*, Elsevier.

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). Some parametric models on the simplex, *Journal of Multivariate Analysis* **39**(1): 106–116.

Bayer, F. M., Bayer, D. M. and Pumi, G. (2017). Kumaraswamy autoregressive moving average models for double bounded environmental data, *Journal of Hydrology* **555**: 385–396.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression diagnostics. J*, Wiley & Sons, New York, New York.

Bonat, W. H. (2016). mcglm: Multivariate covariance generalized linear models, http://git.leg.ufpr.br/wbonat/mcglm. R package version 0.4.0.

——— (2017). Modelling mixed types of outcomes in additive genetic models, *The International Journal of Biostatistics* **13**(2): 1–16.

——— (2018). Multiple response variables regression models in R: The mcglm package, *Journal of Statistical-Software* **84**(1): 1–30.

Bonat, W. H. and Jørgensen, B. (2016). Multivariate covariance generalized linear models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(5): 649–675.

Bonat, W. H., Jørgensen, B., Kokonendji, C. C., Hinde, J. and Demétrio, C. G. (2018). Extended Poisson–Tweedie: properties and regression models for count data, *Statistical Modelling* **18**(1): 24–49.

Bonat, W. H., Lopes, J. E., Shimakura, S. E. and Ribeiro Jr, P. J. (2018). Likelihood analysis for a class of simplex mixed models., *Chilean Journal of Statistics* **9**(2).

Bonat, W. H., Petterle, R. R., Hinde, J. and Demétrio, C. G. (2018). Flexible quasi-beta regression models for continuous bounded data, *Statistical Modelling* p. (published online).

Bonat, W. H., Ribeiro Jr, P. J. and Shimakura, S. E. (2015). Bayesian analysis for a class of beta mixed models, *Chilean Journal of Statistics* **6**(1): 3–13.

Bonat, W. H., Ribeiro Jr, P. J. and Zeviani, W. M. (2012). Regression models with responses on the unit interval: specification, estimation and comparison, *Biometric Brazilian Journal* **30**(4): 415–431.

——— (2015). Likelihood analysis for a class of beta mixed models, *Journal of Applied Statistics* **42**(2): 252–266.

Bonat, W., Olivero, J., Grande-Vega, M., Farfán, M. and Fa, J. (2017). Modelling the covariance structure in marginal multivariate count models: Hunting in bioko island, *Journal of Agricultural, Biological and Environmental Statistics* pp. 1–19.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American statistical Association* **88**(421): 9–25.

Cario, M. C. and Nelson, B. L. (1997). Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix, *Technical report*, Citeseer.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots, *Journal of the American Statistical Association* **74**(368): 829–836.

Cook, R. D. (1977). Detection of influential observation in linear regression, *Technometrics* **19**(1): 15–18.

da Silva, C., Migon, H. and Correia, L. (2011). Dynamic Bayesian beta models, *Computational Statistics & Data Analysis* **55**(6): 2074–2089.

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*, Wiley.

Diggle, P., Heagerty, P., Liang, K.-Y. and Zeger, S. (2002). *Analysis of Longitudinal Data (Second edition)*, Oxford University Press, United Kingdom.

Ferrari, S. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions, *Journal of Applied Statistics* **31**(7): 799–815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. (2013). Mixed beta regression: A Bayesian perspective, *Computational Statistics & Data Analysis* **61**(0): 137–147.

Fitzmaurice, G., Davidian, M., Verbeke, G. and Molenberghs, G. (2008). *Longitudinal data analysis*, CRC Press.,

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2011). *Applied Longitudinal Analysis (Second edition)*, John Wiley and Sons Inc., New Jersey.

Godambe, V. P. and Thompson, M. (1978). Some aspects of the theory of estimating equations, *Journal of Statistical Planning and Inference* **2**(1): 95–104.

Grunwald, G. K., Raftery, A. E. and Guttorp, P. (1993). Time series of continuous proportions, *Journal of the Royal Statistical Society, Series B* **55**(1): 103–116.

Hunger, M., Döring, A. and Holle, R. (2012). Longitudinal beta regression models for analyzing health-related quality of life scores over time, *BMC Medical Research Methodology* **12**(1): 144.

Jørgensen, B. and Knudsen, S. J. (2004). Parameter orthogonality and bias adjustment for estimating functions, *Scandinavian Journal of Statistics* **31**(1): 93–114.

Kaya, Y. and Leite, W. L. (2017). Assessing change in latent skills across time with longitudinal cognitive diagnosis modeling: An evaluation of model performance, *Educational and Psychological Measurement* **77**(3): 369–388.

Lemonte, A. J. and Bazán, J. L. (2016). New class of Johnson SB distributions and its associated regression model for rates and proportions, *Biometrical Journal* **58**(4): 727–746.

Li, S. T. and Hammond, J. L. (1975). Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients, *IEEE Transactions on Systems, Man, and Cybernetics* (5): 557–561.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models, *Biometrika* **73**(1): 13–22.

Masarotto, G., Varin, C. et al. (2012). Gaussian copula marginal regression, *Electronic Journal of Statistics* **6**: 1517–1549.

McKenzie, E. (1985). An autoregressive process for beta random variables, *Management Science* **31**(8): 988–997.

Menarin, V., Lara, I. A. R. d. and Silva, S. C. d. (2017). Longitudinal model for categorical data applied in an agriculture experiment about elephant grass, *Scientia Agricola* **74**(4): 265–274.

Mitnik, P. A. and Baek, S. (2013). The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation, *Statistical Papers* **54**(1): 177–192.

Mohd Din, S. H., Molas, M., Luime, J. and Lesaffre, E. (2014). Longitudinal profiles of bounded outcome scores as predictors for disease activity in rheumatoid arthritis patients: a joint modeling approach, *Journal of Applied Statistics* **41**(8): 1627–1644.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data, Springer Series in Statistics*, Springer New York.

Mousa, A. M., El-Sheikh, A. A. and Abdel-Fattah, M. A. (2016). A gamma regression for bounded continuous variables, *Advances and Applications in Statistics* **49**(4): 305.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A* **135**(3): 370–384.

Petterle, R. R., Bonat, W. H., Kokonendji, C. C., Seganfredo, J. C., Moraes, A. and Gomes-da Silva, M. M. (2019). Double Poisson–Tweedie regression models, *to appear* .

Qiu, Z., Song, P. X.-K. and Tan, M. (2008). Simplex mixed-effects models for longitudinal proportional data, *Scandinavian Journal of Statistics* **35**(4): 577–596.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rocha, A. V. and Cribari-Neto, F. (2008). Beta autoregressive moving average models, *Test* **18**(3): 529–545.

Song, P. X.-K., Qiu, Z. and Tan, M. (2004). Modelling heterogeneous dispersion in marginal models for longitudinal proportional data, *Biometrical Journal* **46**(5): 540–553.

Song, P. X.-K. and Tan, M. (2000). Marginal models for longitudinal continuous proportional data, *Biometrics* **56**(2): 496–502.

Su, P. (2014). *NORTARA: Generation of Multivariate Data with Arbitrary Marginals*. R package version 1.0.0.

Venezuela, M. K., Aparecida Botter, D. and Carneiro Sandoval, M. (2007). Diagnostic techniques in generalized estimating equations, *Journal of Statistical Computation and Simulation* **77**(10): 879–888.

Verbeke, G., Fieuws, S., Molenberghs, G. and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review, *Statistical Methods in Medical Research* **23**(1): 42–59.

Verbeke, G. and Molenberghs, G. (2001). *Linear Mixed Models for Longitudinal Data, Springer Series in Statistics*, Springer New York.

Verkuilen, J. and Smithson, M. (2012). Mixed and mixture regression models for continuous bounded responses using the beta distribution, *Journal of Educational and Behavioral Statistics* **37**(1): 82–113.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method, *Biometrika* **61**(3): 439–447.

Zeger, S. L., Liang, K.-Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach, *Biometrics* **44**(4): 1049–1060.

Zhao, W., Lian, H. and Bandyopadhyay, D. (2018). A partially linear additive model for clustered proportion data, *Statistics in Medicine* **37**(6): 1009–1030.

Zheng, X., Qin, G. and Tu, D. (2017). A generalized partially linear mean-covariance regression model for longitudinal proportional data, with applications to the analysis of quality of life data from cancer clinical trials, *Statistics in Medicine* **36**(12): 1884–1894.