Statistical Modelling 2018; **xx(x)**: 1–17

Flexible quasi-beta regression models for continuous bounded data

Wagner H Bonat[1], Ricardo R Petterle[2], John Hinde[3] and Clarice GB Demétrio[4]
[1]Laboratory of Statistics and Geoinformation, Department of Statistics, Paraná Federal University, Curitiba, Brazil.
[2]Sector of Health Sciences, Medical School, Paraná Federal University, Curitiba, PR, Brazil.
[3]School of Mathematics, Statistics and Applied Mathematics National University of Ireland Galway, Galway, Ireland.
[4]Departamento de Ciências Exatas, Escola Superior de Agricultura Luiz de Queiroz, São Paulo University, Piracicaba, Brazil.

Abstract: We propose a flexible class of regression models for continuous bounded data based on second-moment assumptions. The mean structure is modelled by means of a link function and a linear predictor, while the mean and variance relationship has the form $\phi\mu^p(1-\mu)^p$, where μ, ϕ and p are the mean, dispersion and power parameters respectively. The models are fitted by using an estimating function approach where the quasi-score and Pearson estimating functions are employed for the estimation of the regression and dispersion parameters respectively. The flexible quasi-beta regression model can automatically adapt to the underlying bounded data distribution by the estimation of the power parameter. Furthermore, the model can easily handle data with exact zeroes and ones in a unified way and has the Bernoulli mean and variance relationship as a limiting case. The computational implementation of the proposed model is fast, relying on a simple Newton scoring algorithm. Simulation studies, using datasets generated from simplex and beta regression models show that the estimating function estimators are unbiased and consistent for the regression coefficients. We illustrate the flexibility of the quasi-beta regression model to deal with bounded data with two examples. We provide an `R` implementation and the datasets as supplementary materials.

Key words: bounded data, estimating functions, beta distribution, simplex distribution, regression models

Received January 2018; **revised** May 2018; **accepted** July 2018

1 Introduction

For the analysis of continuous bounded data, the simplex (Kieschnick and McCullough, 2003) and beta (Ferrari and Cribari-Neto, 2004) regression models are two frequent approaches. Both models are based on the principles of generalized linear models (Nelder and Wedderburn, 1972), relating the expected value of the response variable to covariates through a suitable link function. Such approaches are

Address for correspondence: Wagner Hugo Bonat, Department of Statistics, Paraná Federal University, Centro Politécnico, Curitiba, 81531980 CP19081, Brazil.
E-mail: wbonat@ufpr.br

 10.1177/1471082X18790847

nowadays easily employed for practitioners, mainly because of the well-developed `simplexreg` (Zhang et al., 2016) and `betareg` (Cribari-Neto and Zeileis, 2010) packages for the statistical software `R` (R Development Core Team, 2017).

The literature on the analysis of continuous bounded data in the form of rates and proportions is rich. Standard simplex and beta regression models for independent data were presented in Ferrari and Cribari-Neto (2004), Kieschnick and McCullough (2003) and Paolino (2001) to cite but a few. Simas et al. (2010), Cepeda and Gamerman (2005) and Cepeda (2001) extend the beta regression model by regressing both mean and dispersion parameters on potential covariates. A similar extension has been done for simplex regression models in Song et al. (2004).

For the analysis of longitudinal and repeated measures data, Bonat et al. (2015a) and Figueroa-Zúñiga et al. (2013) presented beta mixed models within the Bayesian paradigm. Bonat et al. (2015b) proposed a similar model class, in a pure likelihood framework. Similarly, Qiu et al. (2008) discussed the specification of simplex mixed models for longitudinal data, while a marginal version of the simplex model for modelling longitudinal data was proposed in Song and Tan (2000). Time series models for continuous bounded data were presented in Rocha and Cribari-Neto (2008), Grunwald et al. (1993) and McKenzie (1985). Influence and residuals analysis for beta regression models were discussed in Rocha and Simas (2011), Espinheira et al. (2008b) and Espinheira et al. (2008a). Dynamic beta regression models were proposed by da Silva et al. (2011). For further discussions and references on beta regression models see Smithson and Verkuilen (2006) and Verkuilen and Smithson (2012).

Additionally, new classes of probability density functions have been proposed for modelling continuous bounded data. Lemonte and Bazàn (2016) proposed a new class of Johnson S_B distribution and associated regression models. Abdel-Fattah et al. (2016) presented a gamma regression model for bounded continuous response variables. Regression models based on the Kumaraswamy distribution has been proposed as an alternative to the beta regression model (Mitnik and Baek, 2013). Meanwhile, Kotz and van Dorp (2004) present many alternative probability density functions with bounded support.

Estimation and inference for regression models are generally done based on the standard method of maximum likelihood, which in turn provides efficient estimation for both regression and dispersion parameters. Although of broad use and efficient, the method of maximum likelihood requires a full model specification, that is, we need to specify a probability density function as the data generator. Given the plethora of available approaches to model continuous bounded data in the literature, it is difficult to decide, with conviction, which is the best choice for a particular dataset. The standard approach seems to take a small set of models, such as the beta, simplex, Kumaraswamy, etc., fit all of them and then choose the best fit by using some measures of goodness-of-fit, such as the Akaike or Bayesian information criteria. A typical example of this approach can be found in Bonat et al. (2012), where the authors compared the fit of four different distributions for the analysis of four datasets.

Although reasonable, such an approach is challenging to implement in practical data analysis. First, we should define the set of models to be fitted. Second, each bounded model can require specific fitting algorithms and give its own set of fitting problems, in general due to difficult aspects of the likelihood function. Third, the choice of the best fit may not be obvious, with different information criteria leading to different selected models. Finally, the uncertainty around the choice of the distribution is not taken into account when choosing the best fit. Thus, we claim that it is very useful and attractive to have a unified model that can automatically adapt to the underlying bounded data distribution and that can be easily implemented in practice.

The main goal of this article is to propose a new class of regression models for continuous bounded data. The proposed flexible quasi-beta regression model is based on second-moment assumptions, only. The expectation is modelled in the orthodox way by means of a link function and a linear predictor. The variance is specified by $\phi\mu^p(1-\mu)^p$, where μ is the mean and ϕ is the dispersion parameter. Finally, the extra power parameter p is introduced to give more flexibility in the modelling of the mean and variance relationship. As we shall show in Section 2, such a mean and variance relationship is often found in bounded distributions, which implies that the proposed model can easily mimic many well-known regression models, such as the beta and simplex regression models. Our approach for model specification resembles Wedderburn's quasi-likelihood (Wedderburn, 1974) method and has been recently used in the context of continuous and count data by Bonat and Kokonendji (2017) and Bonat et al. (2017), respectively. In this framework, we do not specify a full probability distribution for the bounded response variable and consequently a likelihood function is not available. Thus, the models are fitted by an estimating function approach as in Bonat and Jørgensen (2016) and Jørgensen and Knudsen (2004) obtained by combining the quasi-score and Pearson estimating functions for the estimation of the regression and dispersion parameters, respectively.

In the next section, we provide some background on regression models for continuous bounded data and present the flexible quasi-beta regression models. We focus on simplex and beta regression models because they are easily available in `R`. Section 3 discusses the estimating function approach employed for estimation and inference. Section 4 presents the main results from our simulation study. In Section 5 we illustrate the application of the flexible quasi-beta regression model through the analysis of two datasets. Finally, Section 6 gives some final remarks. Datasets and `R` code are available in the supplementary material.

2 Regression models for continuous bounded data

In this section, we shall explore the mean and variance relationship induced by the simplex and beta distributions. Furthermore, we use it as motivation to propose a new class of regression models to deal with continuous bounded data based on second-moment assumptions.

Let $Y \sim S^{-}(\mu, \phi)$ denote a simplex distributed random variable with probability density function given by

$$f(y;\mu,\phi) = \left(2\pi\phi^2\{y(1-y)^3\}\right)^{-1/2} \exp\left\{-\frac{1}{2\phi^2}\left\{\frac{(y-\mu)^2}{y(1-y)\mu^2(1-\mu)^2}\right\}\right\}, \tag{2.1}$$

where $y, \mu \in (0, 1)$ and $\phi > 0$ is a dispersion parameter.

Jørgensen (1997) showed that $\mathrm{E}(Y) = \mu$ and

$$\mathrm{var}(Y) = \mu(1-\mu) - \sqrt{\frac{1}{2\phi^2}} \exp\left\{\frac{1}{2\phi^2\mu^2(1-\mu)^2}\right\} \Gamma\left\{\frac{1}{2}, \frac{1}{2\phi^2\mu^2(1-\mu)^2}\right\}.$$

We note in passing that $\mathrm{var}(Y) \to \mu(1-\mu)$ for $\phi \to \infty$, that is, the simplex mean and variance relationship corresponds to the Bernoulli mean and variance relationship.

Similarly, let $Y \sim B(\mu, \phi)$ denote a beta distributed random variable with probability density function as follows

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \tag{2.2}$$

where $y, \mu \in (0, 1)$ and $\phi > 0$ is now a precision parameter (inverse of dispersion). Ferrari and Cribari-Neto (2004) showed that $\mathrm{E}(Y) = \mu$ and $\mathrm{var}(Y) = \frac{\mu(1-\mu)}{1+\phi}$. Thus, it is easy to see that when $\phi \to 0$, the $\mathrm{var}(Y) \to \mu(1-\mu)$, which in turn also corresponds to the mean and variance relationship of the Bernoulli distribution. Furthermore, for both simplex and beta distributions, the largest possible variance is 0.25, and it is reached when $\mu = 0.5$ for extreme values of ϕ.

Figure 1 shows the mean and variance relationship of the simplex and beta distributions for different values of ϕ. We fixed the values of ϕ to have different levels of variance when $\mu = 0.5$. Thus, for $\mathrm{var}(Y) = 0.25, 0.15, 0.05, 0.025$ and 0.01, the values of the simplex dispersion parameter were $\phi = 10\,000\,000$, 9.1500, 2.4067, 1.4651, 0.8487, while for the beta distribution the precision parameter $\phi = 0.00001, 0.6660, 4, 9, 23.9989$.

The mean and variance relationship shows that the main differences between the distribution shapes appear as μ moves away from 0.5. In general, the beta tails are heavier than the simplex tails. To better illustrate these differences, Figure 2 presents the cumulative probability function of the simplex and beta distributions for different values of the expectation and variance of Y.

The cumulative probability functions presented in Figure 2 highlight the flexibility of both distributions to deal with continuous bounded data. In the extreme case of $\mathrm{var}(Y) = 0.25$, both distributions present virtually the same mean and variance relationship and consequently very similar cumulative probability function (i.e., Bernoulli). For smaller variances the distributions are really similar for values of the mean around 0.5, while quite different in the tails. Such results emphasize that in practice one distribution can fit better than the other for an observed dataset. Thus, a data analysis should consider both of these, or even other alternative distributions.

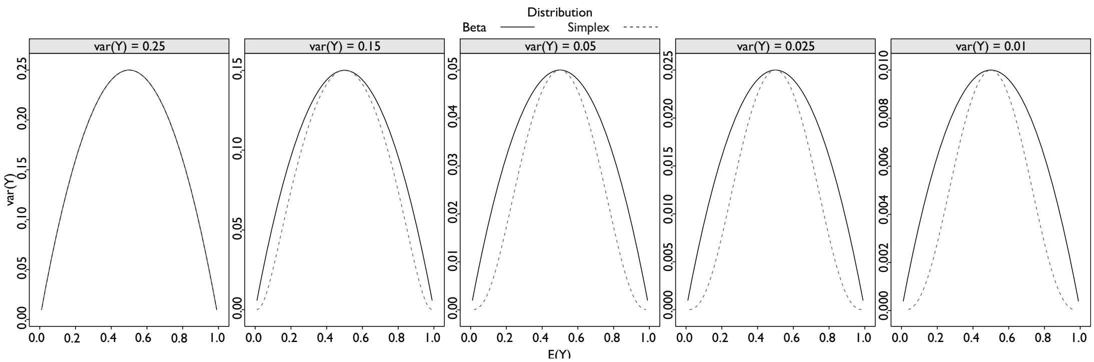

Figure 1 Mean and variance relationship beta and simplex distributions

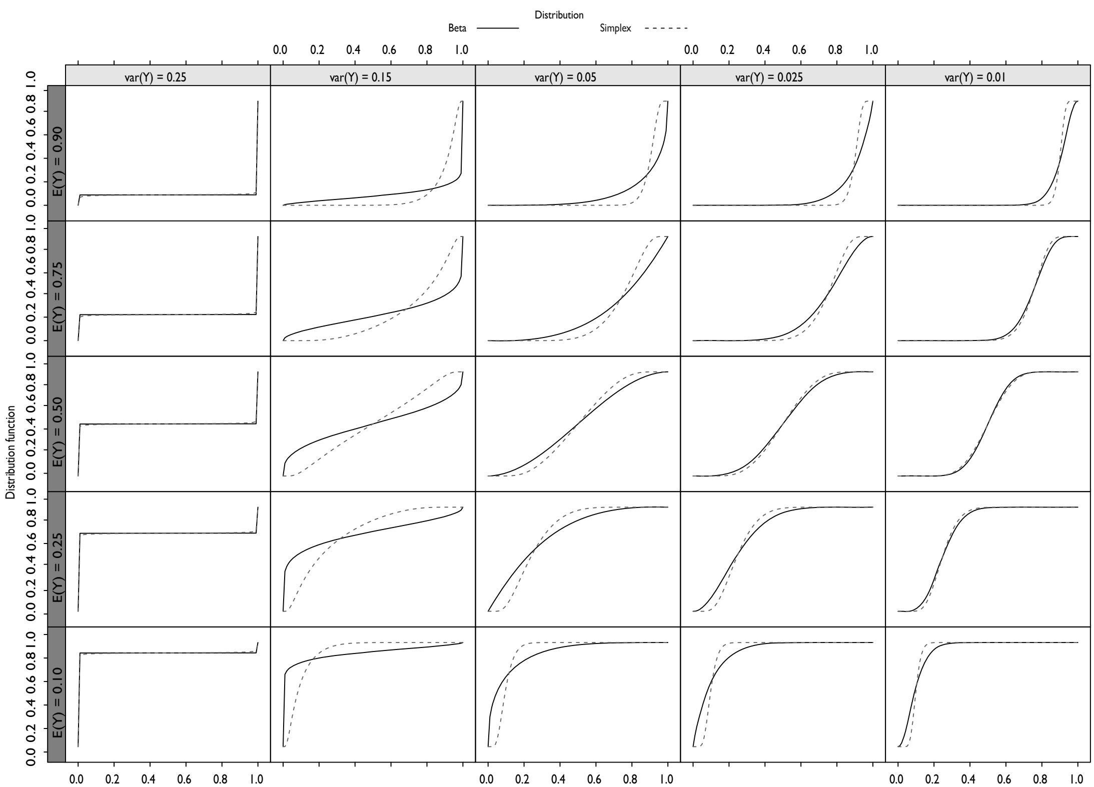

Figure 2 Beta and simplex cumulative probability function by values of the variance (fixed for $\mu = 0.5$) and expectation

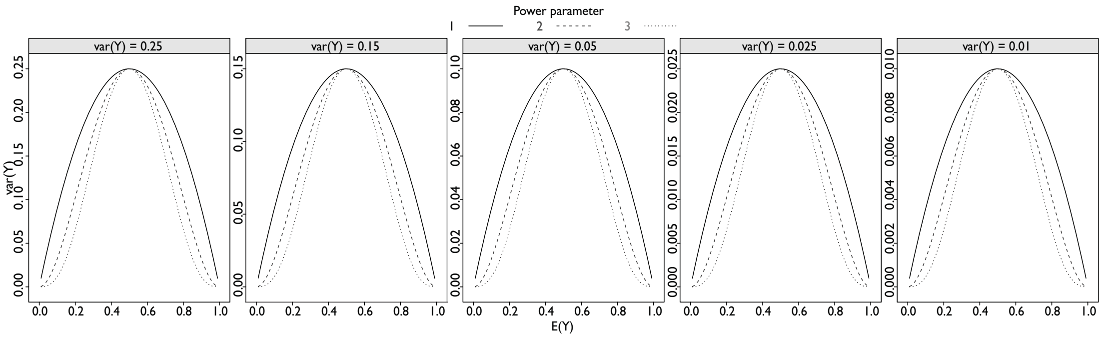

Figure 3 Mean and variance relationship modelled by the function var(Y) $= \phi\mu^p(1-\mu)^p$ by values of the var(Y) and power parameter p

In spite of the differences in the mean and variance relationship between the simplex and beta distributions, such a relationship can be well modelled by a simple function of the expected values μ. This fact motivates us to specify a regression model by using only second-moment assumptions. Thus, consider a cross-sectional dataset, $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$, where y_is are independent and identically distributed realizations of Y_i according to an unspecified distribution, whose expectation and variance are given by

$$\begin{aligned} \mathrm{E}(Y_i) &= \mu_i = g^{-1}(\boldsymbol{x}_i^\top \boldsymbol{\beta}) \\ \mathrm{var}(Y_i) &= \sigma_i = \phi\mu_i^p(1-\mu_i)^p, \end{aligned} \tag{2.3}$$

where $\boldsymbol{x}_i$ and $\boldsymbol{\beta}$ are $(q \times 1)$ vectors of known covariates and unknown regression parameters, respectively. Moreover, g is a standard link function, for which here we adopt the logit link function to give mean values in the interval (0, 1), but potentially any other suitable link function could be adopted. The regression model specified in (2.3) is parametrized by $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\lambda}^\top)^\top$, where $\boldsymbol{\lambda} = (\phi, p)$ with the usual parameter space and can mimic the mean and variance relationship of both simplex and beta models as shown in Figure 3.

As expected for $p = 1$, our model corresponds to the beta regression model in a slightly different parametrization. The plots in Figure 3 show that the simplex mean and variance relationship is well approximated by using $p = 3$ when $\mathrm{var}(Y) \leq 0.15$, although there is no direct equivalence with the simplex distribution. However, in practice, any simplex or beta distributions can be the best choice for a particular dataset. Thus, the algorithm we shall present in Section 3 allows us to estimate the power parameter, which in turn works as an automatic model selection. In this sense, the model proposed in (2.3) unifies the simplex and beta regression models and provides a broader class of regression models for continuous bounded data.

3 Estimation and inference

In this section, we present the estimating function approach used for fitting the flexible quasi-beta regression models using terminologies and results from Jørgensen and Knudsen (2004) and Bonat and Jørgensen (2016). We adopt the quasi-score and Pearson estimating functions for estimation of the regression and dispersion parameters, respectively. The quasi-score function for $\boldsymbol{\beta}$ is given by,

$$\psi_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \left(\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_1} \sigma_i^{-1} (Y_i - \mu_i), \ldots, \sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \beta_q} \sigma_i^{-1} (Y_i - \mu_i) \right)^{\top},$$

where $\partial \mu_i / \partial \beta_j = \mu_i(1 - \mu_i) x_{ij}$ for $j = 1, \ldots, q$.

The entry (j, k) of the $q \times q$ sensitivity matrix $S_{\boldsymbol{\beta}}$ for $\psi_{\boldsymbol{\beta}}$ is given by

$$S_{\boldsymbol{\beta}_{jk}} = E\left(\frac{\partial}{\partial \beta_k} \psi_{\boldsymbol{\beta}_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}) \right) = -\sum_{i=1}^{n} \mu_i(1 - \mu_i) x_{ij} \sigma_i^{-1} x_{ik} \mu_i (1 - \mu_i). \tag{3.1}$$

In a similar way, the entry (j, k) of the $q \times q$ variability matrix $V_{\boldsymbol{\beta}}$ for $\psi_{\boldsymbol{\beta}}$ is given by

$$V_{\boldsymbol{\beta}_{jk}} = \mathrm{Cov}(\psi_{\boldsymbol{\beta}_j}(\boldsymbol{\beta}, \boldsymbol{\lambda}), \psi_{\boldsymbol{\beta}_k}(\boldsymbol{\beta}, \boldsymbol{\lambda})) = \sum_{i=1}^{n} \mu_i(1 - \mu_i) x_{ij} \sigma_i^{-1} x_{ik} \mu_i (1 - \mu_i).$$

The Pearson estimating functions for the dispersion parameters have the following form,

$$\psi_{\boldsymbol{\lambda}}(\boldsymbol{\lambda}, \boldsymbol{\beta}) = \left(-\sum_{i=1}^{n} \frac{\partial \sigma_i^{-1}}{\partial \phi} \left[(Y_i - \mu_i)^2 - \sigma_i \right], -\sum_{i=1}^{n} \frac{\partial \sigma_i^{-1}}{\partial p} \left[(Y_i - \mu_i)^2 - \sigma_i \right] \right)^{\top}.$$

It is important to highlight that the Pearson estimating functions are unbiased estimating functions for $\boldsymbol{\lambda}$ based on the squared residuals $(Y_i - \mu_i)^2$ with expected value σ_i.

The entry (j, k) of the 2×2 sensitivity matrix $S_{\boldsymbol{\lambda}}$ for the dispersion parameters is given by

$$S_{\boldsymbol{\lambda}_{jk}} = E\left(\frac{\partial}{\partial \lambda_k} \psi_{\boldsymbol{\lambda}_j}(\boldsymbol{\lambda}, \boldsymbol{\beta}) \right) = -\sum_{i=1}^{n} \frac{\partial \sigma_i^{-1}}{\partial \lambda_j} \sigma_i \frac{\partial \sigma_i^{-1}}{\partial \lambda_k} \sigma_i, \tag{3.2}$$

where λ_j and λ_k denote either ϕ or p.

The cross entries of the sensitivity matrices $\mathrm{S}_{\boldsymbol{\beta}\boldsymbol{\lambda}}$ and $\mathrm{S}_{\boldsymbol{\lambda}\boldsymbol{\beta}}$ are given by

$$\mathrm{S}_{\beta_j\lambda_k} = \mathrm{E}\left(\frac{\partial}{\partial\lambda_k}\psi_{\beta_j}(\boldsymbol{\beta},\boldsymbol{\lambda})\right) = 0 \tag{3.3}$$

and

$$\mathrm{S}_{\lambda_j\beta_k} = \mathrm{E}\left(\frac{\partial}{\partial\beta_k}\psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})\right) = -\sum_{i=1}^{n}\frac{\partial\sigma_i^{-1}}{\partial\lambda_j}\sigma_i\frac{\partial\sigma_i^{-1}}{\partial\beta_k}\sigma_i. \tag{3.4}$$

Finally, the joint sensitivity matrix for the parameter vector $\boldsymbol{\theta}$ is given by

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

whose entries are defined by equations (3.1), (3.2), (3.3) and (3.4).

The asymptotic variance of the estimating function estimators denoted by $\hat{\boldsymbol{\theta}}$ is obtained from the inverse Godambe information matrix, whose general form for a vector parameter $\boldsymbol{\theta}$ 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. The variability matrix for $\boldsymbol{\theta}$ has the form

$$\mathrm{V}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathrm{V}_{\boldsymbol{\beta}} & \mathrm{V}_{\boldsymbol{\beta}\boldsymbol{\lambda}} \\ \mathrm{V}_{\boldsymbol{\lambda}\boldsymbol{\beta}} & \mathrm{V}_{\boldsymbol{\lambda}} \end{pmatrix}, \tag{3.5}$$

where $\mathrm{V}_{\boldsymbol{\lambda}\boldsymbol{\beta}} = \mathrm{V}_{\boldsymbol{\beta}\boldsymbol{\lambda}}^{\top}$ and $\mathrm{V}_{\boldsymbol{\lambda}}$ depend on the third and fourth moments of Y_i, respectively. In order to avoid this dependence on higher-order moments, we adopt the approach proposed in Bonat and Jørgensen (2016) that consists of using the empirical versions of $\mathrm{V}_{\boldsymbol{\lambda}}$ and $\mathrm{V}_{\boldsymbol{\lambda}\boldsymbol{\beta}}$ as given by

$$\tilde{\mathrm{V}}_{\lambda_{jk}} = \sum_{i=1}^{n}\psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})_i\psi_{\lambda_k}(\boldsymbol{\lambda},\boldsymbol{\beta})_i \quad \text{and} \quad \tilde{\mathrm{V}}_{\lambda_j\beta_k} = \sum_{i=1}^{n}\psi_{\lambda_j}(\boldsymbol{\lambda},\boldsymbol{\beta})_i\psi_{\beta_k}(\boldsymbol{\lambda},\boldsymbol{\beta})_i.$$

Finally, the approximate distribution of $\hat{\boldsymbol{\theta}}$ (Jørgensen and Knudsen, 2004; Godambe and Thompson, 1978) is the multivariate Gaussian distribution with expectation $\boldsymbol{\theta}$ and covariance matrix $\mathrm{J}_{\boldsymbol{\theta}}^{-1} = \mathrm{S}_{\boldsymbol{\theta}}^{-1}\mathrm{V}_{\boldsymbol{\theta}}\mathrm{S}_{\boldsymbol{\theta}}^{-\top}$.

To solve the system of equations $\psi_{\boldsymbol{\beta}} = \mathbf{0}$ and $\psi_{\boldsymbol{\lambda}} = \mathbf{0}$, we adopted the modified chaser algorithm proposed by Jørgensen and Knudsen (2004) and implemented in a very general form in the package `mcglm` (Bonat, 2018) for the statistical software `R`. The modified chaser algorithm uses the insensitivity property (3.3), which allows us to use two separate equations to update $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ as follows

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

where α is a tuning constant used to control the step-length. This algorithm was used by Bonat and Kokonendji (2017) and Bonat et al. (2017) for estimation and inference in the context of Tweedie and Poisson–Tweedie regression models, respectively. Furthermore, this algorithm is a special case of the flexible algorithm presented by Bonat and Jørgensen (2016) in the context of multivariate covariance generalized linear models.

4 Simulation study

In this section, we present a simulation study conducted to verify the properties of the estimating function estimators. We designed five simulation scenarios to explore the flexibility of the proposed regression model to deal with bounded data, as generated from the simplex and beta distributions. For each setting, we considered five different sample sizes, 50, 100, 250, 500 and 1 000, generating 1 000 datasets in each case. The dispersion parameter of the simplex distribution was fixed at the values $\phi = 10\,000\,000, 9.1500, 2.4067, 1.4651, 0.8487$ and the corresponding precision parameter for the beta distribution at $\phi = 0.00001, 0.6660, 4, 9, 23.9989$. These values correspond to $\text{var}(\text{Y}) = 0.25, 0.15, 0.05, 0.025, 0.01$ when $\mu = 0.5$. Thus, we can explore the behaviour of the fitting algorithm as well as the properties of the estimators from an extreme and challenging ($\text{var}(\text{Y}) = 0.25$) to an easy ($\text{var}(\text{Y}) = 0.01$) scenario.

For all scenarios, we consider a standard regression model where the linear predictor is composed of a continuous and a categorical covariate. The continuous covariate was generated from a standard Gaussian distribution and the categorical covariate from a Bernoulli distribution with probability fixed at 0.5. The regression coefficients were fixed at $\beta_0 = 0$, $\beta_1 = 3$ for the continuous covariate, and $\beta_2 = -1.5$ for the categorical indicator variable. Such values were chosen in order to cover the whole range of values for the expectation of the random variable, that is, the unit interval. Figure 4 shows the average bias plus and minus the average standard error for the regression parameter estimators under each scenario.

The results in Figure 4 show clearly that the bias and standard error (SE) of the estimating function estimators tend to zero as the sample size increases. Thus, we can conclude that the estimating function estimators are unbiased and consistent for large samples, as expected. For small sample sizes, the results in Figure 4 show that the bias tends to increase from the easy $\text{var}(\text{Y}) = 0.01$ to the challenging $\text{var}(\text{Y}) = 0.25$ scenario, again as expected. In general, the biases are larger when using the beta distribution than the simplex distribution as the data generator, mainly for the parameter β_1. Figure 5 presents the coverage rate for the individual parameter confidence intervals by sample size and simulation scenario.

The results in Figure 5 show that the empirical coverage rates are close to the nominal level of 95% for large samples. On the other hand, for small sample sizes and large variance scenarios the empirical coverage rate for the continuous covariate regression coefficient β_1 presented values around 80%. This result shows that for

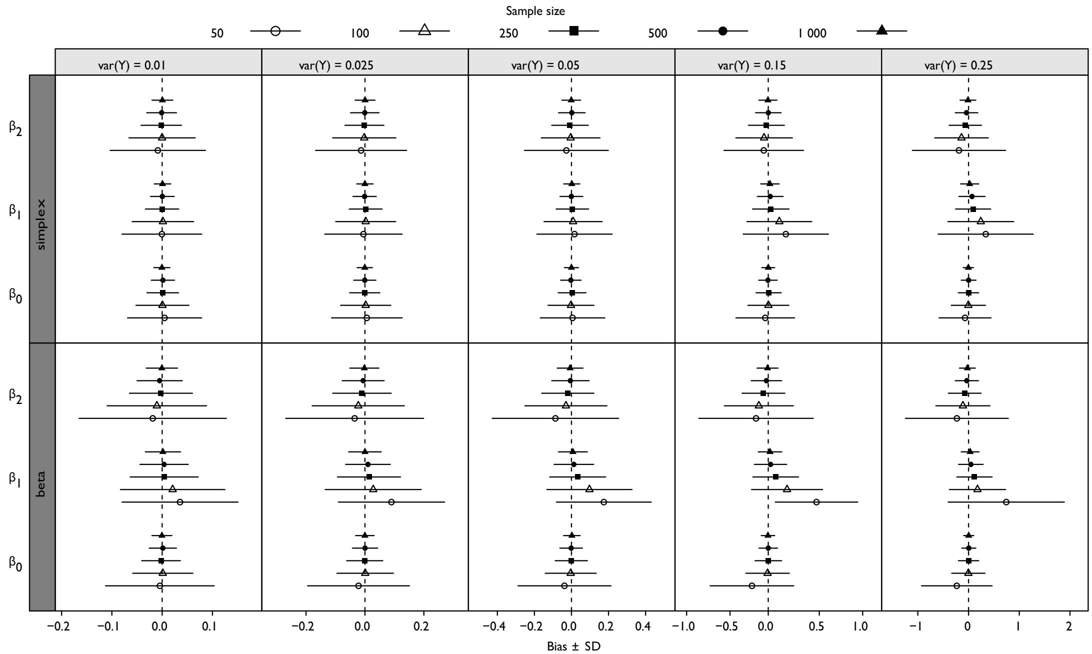

Figure 4 Average bias and confidence intervals by sample size and simulation scenario

these scenarios, the estimation of the standard errors associated with the regression coefficients is challenging and thus requires large samples for their correct estimation.

5 Data analyses

In this section, we illustrate the application of the flexible quasi-beta regression through the analysis of two datasets. The first dataset concerns measures of the water quality index collected at 16 hydroelectric power plants in Paraná State, Brazil. This dataset was analysed in Bonat et al. (2015a,b, 2018) using beta and simplex mixed models. Bonat et al. (2018) showed that the simplex distribution fits better than the beta distribution to this dataset. The second dataset corresponds to observations of the stress and anxiety indices among non-clinical women in Townsville, Queensland, Australia. This dataset is part of the `betareg` Cribari-Neto and Zeileis (2010) package and the beta distribution clearly fits better than the simplex distribution to this dataset. Thus, we have a case where the simplex fits better and another for which the beta distribution provides the better fit. Our goal is to show that the proposed model fits well for both cases. The datasets and the `R` scripts used for their analysis can be obtained `http://www.leg.ufpr.br/doku.php/publications:papercompanions:quasibeta`.

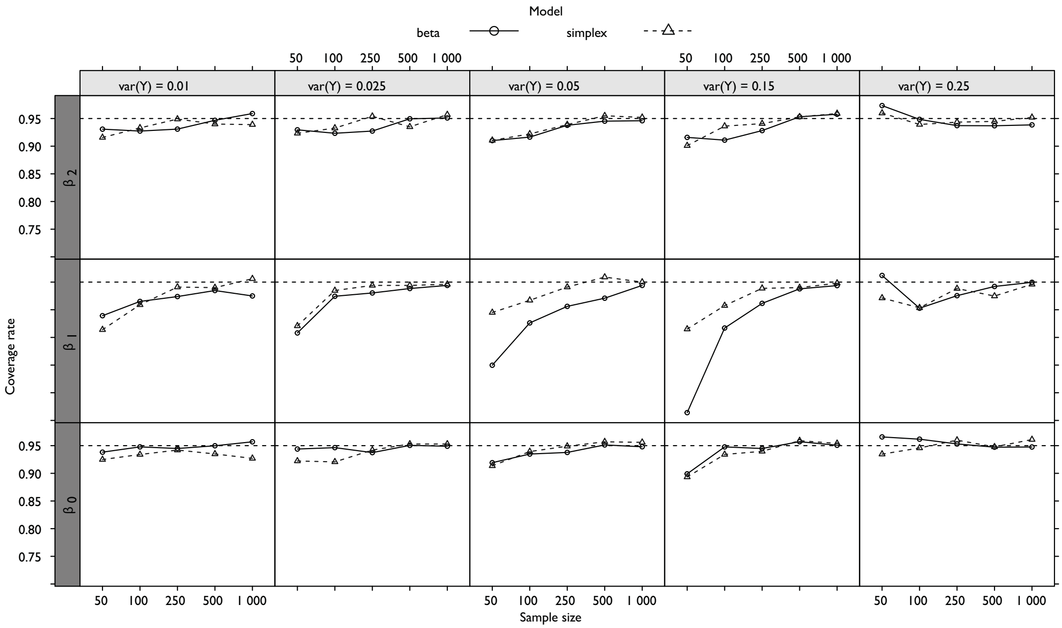

Figure 5 Coverage rate for each regression parameter confidence interval by sample size and simulation scenario

5.1 Dataset 1: Water quality on power plant reservoirs

The water quality dataset corresponds to observations of the water quality index measured quarterly during 2004 at 16 operating hydroelectric power plants in Paraná State, Brazil. The main goal of this study was to detect changes in water quality, possibly due to the presence of the dams. The water quality index was measured at locations considered affected and unaffected by the reservoir and then compared, that is, measurements taken upstream in the main river are considered unaffected reference values, whereas measurements taken at the reservoir and downstream are considered potentially affected. Thus, the main goal is to assess the effect of the covariate `LOCAL`, with levels upstream, reservoir and downstream, controlled for the effects of the `QUARTER` of data collection. In this article, for simplicity, we do not consider the possible effect of the individual power plants. The dataset has 190 observations with 12 measurements (4 quarters $\times$ 3 locations) for each of the 16 power plants with only two missing values. Figure 6 presents a histogram and boxplots relating the response variable to the potential covariates.

Figure 6(A) shows a left asymmetry in the marginal distribution, while Figure 6(B) suggests that upstream observations present smaller values than reservoir and downstream. Furthermore, Figure 6(C) indicates that the water quality index presents smaller values on the first and fourth quarters (warmer periods).

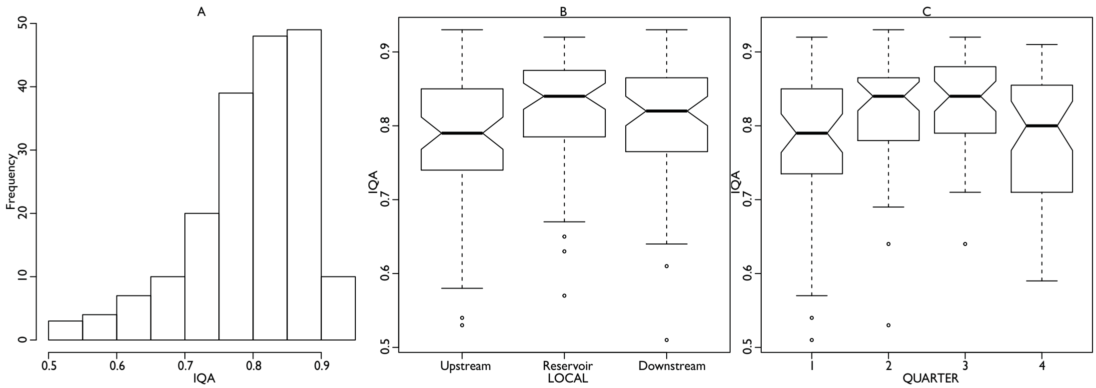

Figure 6 Histrogram (A) and boxplots for the LOCAL(B) and QUARTER(C) for the water quality index (IQA)

Table 1 Regression parameter estimates, standard error (SE) and measures of goodness-of-fit by models: Water quality dataset

	Beta		Simplex		Flexible quasi-beta	
Parameter	Est (SE)	Z-value	Est (SE)	Z-value	Est (SE)	Z-value
β_0	1.14(0.08)	14.08	1.10(0.09)	12.37	1.11(0.10)	11.44
β_{12}	0.23(0.08)	2.72	0.25(0.09)	2.92	0.25(0.09)	2.88
β_{13}	0.15(0.08)	1.77	0.16(0.09)	1.86	0.17(0.09)	1.87
β_{22}	0.21(0.10)	2.13	0.24(0.10)	2.39	0.23(0.10)	2.22
β_{23}	0.29(0.10)	3.01	0.35(0.10)	3.57	0.34(0.10)	3.42
β_{24}	0.05(0.09)	0.50	0.07(0.10)	0.69	0.06(0.11)	0.53
ϕ	25..79(2.62)*	–	0.26(0.05)	–	13.48(1.44)	–
p	–	–	–	–	4.17(0.06)	–
LogLik	224.62		232.43		–	
plogLik	211.25		215.95		216.81	
pAIC	−408.5		−417.90		−417.62	
pBIC	−385.77		−395.17		−391.64	

Note: *Precision parameter estimate as usual in the beta regression model.

We fitted the flexible quasi-beta regression model to the water quality dataset using a logit link function. The linear predictor is composed of the LOCAL and QUARTER covariates, that is,

$$\eta_{it} = \beta_0 + \beta_{1i} + \beta_{2t}, \tag{5.1}$$

where β_0 is the usual intercept, and β_{1i}, for $i = 2, 3$, measures the changes from upstream to reservoir and downstream, respectively. Similarly, β_{2t}, for $t = 2, 3$ and 4, quantifies the changes from quarter 1 to 2 up to 4, respectively. Table 1 shows the parameter estimates for the beta, simplex and the flexible quasi-beta regression models along with the maximized log-likelihood (beta and simplex) and pseudo-Gaussian log-likelihood values (Bonat et al. (2018)) and the associated pseudo Akaike and Bayesian information criteria.

The results presented in Table 1 show a large difference in terms of log-likelihood values in favour of the simplex regression model. In spite of such a large difference, both models provide very similar regression estimates and standard errors, leading to identical interpretations. Similarly, the pseudo log-likelihood, Akaike and Bayesian information criterion indicates that the simplex model provides a better fit than the beta model. The flexible quasi-beta model presents regression estimates and standard errors similar to both simplex and beta regression models. The power parameter estimate was $\hat{p} = 4.17(0.06)$ showing that probably the beta distribution ($p = 1$) is not the best choice for this water quality dataset. This result agrees with the log-likelihood criterion. Furthermore, the pseudo log-likelihood criterion indicates that the flexible quasi-beta model provides an even better fit than the simplex model. Such a result is expected, since the proposed model has a more flexible mean and variance relationship than the beta and simplex regression models and consequently can mimic the fit of the other two models. Note that, the use of information criterion to compare the quasi-beta with the beta and simplex models can be misleading because we have one more parameter in the flexible quasi-beta that is fixed when fitting the beta and simplex regression models.

5.2 Dataset 2: Stress anxiety dataset

The second dataset concerns an experiment conducted to assess the relationship between stress and anxiety. Both variables were assessed through the Depression Anxiety Stress Scales, with scores ranging from 0 to 42. Subsequently, they were linearly transformed to the open unit interval, for details see Smithson and Verkuilen (2006). The main goal of the analysis is to assess the effect of the anxiety score on the response variable stress score. We fitted the beta, simplex and flexible quasi-beta regression models, where the linear predictor is composed of an intercept (β_0) and the effect of the covariate anxiety score (β_1), again using the standard `logit` link function. Table 2 shows the parameter estimates for the beta, simplex and the flexible quasi-beta regression models along with the maximized log-likelihood (beta and simplex) and pseudo-Gaussian log-likelihood values and the associated pseudo Akaike and Bayesian information criteria.

Based on the log-likelihood criterion, the beta regression model provides a better fit than the simplex model for this stress anxiety dataset. The regression coefficient associated with the covariate anxiety score is approximately 20% larger when fitting the simplex model. The pseudo log-likelihood criterion agrees with the log-likelihood criterion showing the better fit of the beta model. The fit of the flexible quasi-beta regression model also suggests that the beta regression is a suitable choice for this dataset. The estimated value of the power parameter $\hat{p} = 1.17(0.06)$ indicates a mean and variance relationship close to that of the beta distribution. The regression coefficient obtained based on the beta and flexible quasi-beta regression models are quite similar. Finally, the pseudo log-likelihood criterion also indicates that the beta and the flexible quasi-beta regression models provide a very similar fit.

Table 2 Regression parameter estimates, standard error (SE) and measures of goodness-of-fit by models: Stress anxiety dataset

	Beta		Simplex		Flexible quasi-beta	
Parameter	Est (SE)	Z-value	Est (SE)	Z-value	Est (SE)	Z-value
β_0	−1.57(0.08)	−18.94	−1.80(0.10)	−17.66	−1.53(0.08)	−19.76
β_1	4.94(0.47)	10.57	5.92(0.62)	9.58	4.88(0.45)	10.74
ϕ	6.90(0.73)*	–	1.40(0.05)	–	0.14(0.02)	–
p	–	–	–	–	1.17(0.06)	–
LogLik	109.07		57.93		–	
plogLik	93.11		81.16		94.00	
pAIC	−180.20		−156.32		−180.00	
pBIC	−175.99		−152.09		−167.55	

Note: *Precision parameter estimate as usual in the beta regression model.

6 Discussion

We described a new class of regression models to deal with continuous bounded data. Our approach is based on second-moment assumptions, where we extend the beta mean and variance relationship by the inclusion of an additional power parameter. Thus, the flexible quasi-beta regression model can automatically adapt to the unknown underlying bounded data distribution by the estimation of this power parameter. Furthermore, the second-moment assumptions allow us to use an estimating functions approach for parameter estimation and inference. The main technical advantage of the proposed models is the simplicity of the fitting algorithm, which amounts to finding the root of a set of non-linear equations.

We designed five simulation scenarios to explore and show the performance of our fitting algorithm. The simulation scenarios range from very easy $\text{var}(Y) = 0.01$ to extreme and challenging $\text{var}(Y) = 0.25$ scenarios. It is important to highlight that in the last scenario, the simulated values are numerically 0s and 1s only. Thus, standard implementations of the beta and simplex regression models fail for fitting the model in this case. On the other hand, the flexible quasi-beta regression model and the associated estimating function approach proposed for model fitting delivered approximately unbiased and consistent estimates for all simulation scenarios even for small sample sizes. Furthermore, the coverage rate of the obtained confidence intervals presented values close to the nominal level of 95% for large samples.

We illustrated the application of the flexible quasi-beta regression models through the analysis of two datasets. The datasets were chosen to show a situation where the simplex regression model is better than the beta regression model and vice-versa. In the data analyses, the pseudo log-likelihood criterion assists us to indicate that the flexible quasi-beta regression model, proposed in this article, automatically adapted to the underlying data distribution, and in turn provides more reliable results even when the data distribution is unknown.

The basic model discussed in this article can be extended in many ways, including incorporating penalized splines and the use of regularization for high dimensional data, with important applications in covariate selection. We plan to develop improved methods for model checking, such as residual analysis, leverage and outlier detection. Finally, we can extend the model to deal with multiple bounded response variables, with important applications for the analysis of longitudinal and spatial data. These extensions will form the basis of future work.

Supplementary material

`http://www.leg.ufpr.br/doku.php/publications:papercompanions:quasibeta`

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The last author was partially supported by CNPq, a Brazilian science funding agency.

References

Abdel-Fattah MA, El-Sheikh AA and Mousa AM (2016) A gamma regression for bounded continuous variables. *Advances and Applications in Statistics*, **49**, 305–26.

Bonat WH (2018) Multiple response regression models in R: The mcglm package. *Journal of Statistical Software*, **85**, 1–30.

Bonat WH and Jørgensen B (2016) Multivariate covariance generalized linear models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **65**, 649–75.

Bonat WH, Jørgensen B, Kokonendji CC, Hinde J and Demétrio CGB (2017) Extended Poisson-Tweedie: Properties and regression models for count data. *Statistical Modelling*, **87**, 2138–52.

Bonat WH and Kokonendji CC (2017) Flexible Tweedie regression models for continuous data. *Journal of Statistical Computation and Simulation*, **87**, 2138–52.

Bonat WH, Lopes JE, Ribeiro Jr PJ and Shimakura SE (2018) Likelihood analysis for a class of simplex mixed models. *Chilean Journal of Statistics*. Accepted.

Bonat WH, Ribeiro Jr PJ and Shimakura SE (2015a) Bayesian analysis for a class of beta mixed models. *Chilean Journal of Statistics*, **6**, 3–13.

Bonat WH, Ribeiro Jr PJ and Zeviani WM (2012) Regression models with response on the unit interval: Specification, estimation and comparison. *Biometric Brazilian Journal*, **30**, 415–31.

——— (2015b) Likelihood analysis for a class of beta mixed models. *Journal of Applied Statistics*, **42**, 252–66.

Cepeda E (2001) Variability modeling in generalized linear models. *PhD thesis*, Mathematics Institute, Universidade Federal do Rio de Janeiro, Rio de Janeiro.

Cepeda E and Gamerman D (2005) Bayesian methodology for modeling parameters in the two parameter exponential family. *Estadística*, **57**, 93–105.

Cribari-Neto F and Zeileis A (2010) Beta regression in R. *Journal of Statistical Software*, **34**, 1–24.

da Silva CQ, Migon HS and Correia L (2011) Dynamic Bayesian beta models. *Computational Statistics and Data Analysis*, **55**, 2074–89.

Espinheira PL, Ferrari SLP and Cribari-Neto F (2008a) On beta regression residuals. *Journal of Applied Statistics*, **35**, 407–19.

——— (2008b) Influence diagnostics in beta regression. *Computational Statistics and Data Analysis*, **52**, 4417–31.

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

Figueroa-Zùñiga JI, Arellano-Valle RB and Ferrari SLP (2013) Mixed beta regression: A Bayesian perspective. *Computational Statistics and Data Analysis*, **61**, 137–47.

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

Grunwald GK, Raftery AE and Guttorp P (1993) Time series of continuous proportions. *Journal of the Royal Statistical Society, Series B*, **55**, 103–16.

Jørgensen B (1997) *The Theory of Dispersion Models*. London: Chapman and Hall.

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

Kieschnick R and McCullough BD (2003) Regression analysis of variates observed on (0; 1): Percentages, proportions and fractions. *Statistical Modelling*, **3**, 193–213.

Kotz S and van Dorp JR (2004). *Beyond Beta: Other Continuous Families of Distributions with Bounded Support and Applications*. Singapore: World Scientific Press.

Lemonte AJ and Bazàn JL (2016) New class of Johnson SB distributions and its associated regression model for rates and proportions. *Biometrical Journal*, **58**, 727–46.

McKenzie E (1985) An autoregressive process for beta random variables. *Management Science*, **31**, 988–97.

Mitnik PA and Baek S (2013) The Kumaraswamy distribution: Median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers*, **54**, 177–92.

Nelder JA and Wedderburn RWM (1972) Generalized linear models. *Journal of the Royal Statistical Society, Series A*, **135**, 370–84.

Paolino P (2001) Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis*, **9**, 325–46.

Qiu Z, Song PX-K and Tan M (2008) Simplex mixed-effects models for longitudinal proportional data. *Scandinavian Journal of Statistics*, **35**, 577–96.

R Development Core Team (2017) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.

Rocha AV and Cribari-Neto F (2008) Beta autoregressive moving average models. *Test*, **18**, 529–45.

Rocha A and Simas A (2011) Influence diagnostics in a general class of beta regression models. *Test*, **20**, 95–119.

Simas AB, Barreto-Souza W and Rocha AV (2010) Improved estimators for a general class of beta regression models. *Computational Statistics and Data Analysis*, **54**, 348–66.

Smithson M and Verkuilen J (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**, 54–71.

Song PX-K, Qiu Z and Tan M (2004) Modelling heterogeneous dispersion in marginal models for longitudinal proportional data. *Biometrical Journal*, **46**, 540–53.

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

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**, 82–113.

Wedderburn RWM (1974) Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, **61**, 439–47.

Zhang P, Qiu Z and Shi C (2016) simplexreg: An R package for regression analysis of proportional data using the simplex distribution. *Journal of Statistical Software*, **71**, 1–21.