Journal of Statistical Theory and Practice (2022) 16:25
https://doi.org/10.1007/s42519-022-00256-w

ORIGINAL ARTICLE

Complementary Beta Regression Model for Fitting Bounded Data

André F. B. Menezes[1] · Marcelo Bourguignon[2] · Josmar Mazucheli[3]

Accepted: 25 February 2022

Abstract
The beta regression model is the commonly used approach for modeling data in the unit interval. However, there are in the literature some useful and interesting alternatives which often under-used. This paper proposes a novel regression model for bounded data, where the response variable is complementary beta distributed with mean and dispersion parameters. The proposed regression model is a natural strong competitor of the beta regression model. The maximum likelihood method is used for estimating the model parameters. A Monte Carlo experiment is conducted to evaluate the performances of these estimators in finite samples. The usefulness of the new regression model is illustrated by two real applications.

Keywords Bounded data · Beta distribution · Complementary beta distribution · Regression model

1 Introduction

Several natural or anthropogenic phenomena are measured as indicators, percentages, proportions, ratios and rates which are bounded on a certain interval, usually in the unit interval (0, 1). The need for modeling and analyzing bounded data occurs in many fields of real life such as politics [21], psychology [32], medicine [24] and so on. In such situations, to probabilistic modeling these phenomena, under a parametric paradigm, probability distributions limited to (0, 1) are indispensable. Certainly, the two parameter beta distribution is the most widely model used in the literature to describe data in the unit interval, especially because its flexibility [16].

✉ Marcelo Bourguignon
m.p.bourguignon@gmail.com

[1] Departamento de Estatística, Universidade Estadual de Campinas, Campinas, SP, Brasil

[2] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, RN, Brasil

[3] Departamento de Estatística, Universidade Estadual de Maringá, Maringa, PR, Brasil

Published online: 04 April 2022

In order to accommodate explanatory variable in the modeling, Cepeda-Cuervo [4] and Ferrari and Cribari-Neto [8] introduced the beta regression model, which has been successfully applied in several context, mainly because of its flexibility and direct parameter interpretation in terms of the mean and precision parameters. Recently, new families of distributions have been introduced for modeling bounded phenomena in the presence of covariates. For instance, the Kumaraswamy regression model [27], the Johnson S_B regression model [22], the unit-Lindley regression model [23] and the unit-Weibull regression model [24]. Nevertheless, in comparison with the rich variety of distributions for modeling positive random variables, there are few alternatives distribution for bounded variables. In addition, though the proposed distributions in the literature to describe bounded data there is still no agreement on preference and advantage of a specific model.

In particular, a probability distribution related to beta that has not received much attention in the literature was introduced by Jones [17] and results from reversing the roles of the distribution and quantile functions of the beta distribution. He called this distribution as complementary beta (CB) distribution and showed that it has several attractive properties that are complementary to those of the beta distribution. Despite of this, the CB distribution has been neglected in the literature. To the best of our knowledge, the only real data analysis using CB distribution appeared in Iacobellis [15]. More recently, Mazucheli et al. [23] compared the L-moment and maximum likelihood estimators by means of simulation studied and analysis of relative indices from annual temperature extremes. However, no attention has been paid for the CB distribution in the regression analysis, until now.

In this context, the goal of this paper is to propose a new regression model based on the CB distribution using a new parameterization that is indexed by mean and dispersion parameters and investigated as a useful alternative to the beta regression model, with the hope that the new model may have a “better fit” in certain practical situations. In fact, Iacobellis [15] wrote: “As a practical advantage, the CB model allows to evaluate, within a unique and coherent probabilistic framework, flow duration curves, annual minimum flow distribution, total annual flow distribution and other useful functions for design and management of structural and nonstructural water resources systems, from environmental minimum streamflow requirements to stream diversion and regulation structures”. Furthermore, the mean of the CB distribution has a simple formula allows us to directly incorporate the covariates in the mean in order to quantify their average influence on the response variable. It is noteworthy that in statistical modeling is important consider different mechanism to propose a solution rather than trusting in a single choice. Thus, one of the main motivation of these paper is to contribute with another attractive regression model for modeling of limited response variables.

Also, it should be highlighted that, compared with the beta model, the CB model has the following advantages [17]:

1. the CB distribution is much more amenable than the beta distribution to exact computations involving expectations of order statistics, including L-moments;
2. the CB distribution has the position in a wider family of distributions defined through the same simple form for their quantile density functions [20];

3. the CB distribution has unique sub-models as special cases, such as the cosine distribution, among others;
4. the CB distribution can provide better fits than beta distributions in some interesting situations (see Sect. 4).

The remainder of this paper is organized as follows. In Sect. 2, we study the CB distribution briefly based on Jones [17] and introduce a new parameterization of the CB distribution that is indexed by the mean and dispersion parameters. In Sect. 3, we formulate the CB regression model with varying dispersion and estimate its parameters. In Sect. 4, some numerical results of the estimators and the empirical distribution of the residuals are presented. In Sect. 5, we illustrate the proposed model and its diagnostics with two real-world data. In Sect. 6, we present our conclusions.

2 Complementary Beta Distribution

In this section, the CB distribution is reviewed and a new parameterization is introduced based on the mean and dispersion parameters.

The random variable Z follows a beta distribution with shape parameters $a, b > 0$ if its probability density function (p.d.f) is expressed as

$$f_B(z; a, b) = \frac{1}{B(a, b)} z^{a-1} (1-z)^{b-1}, \quad 0 < z < 1,$$

where $B(a, b) = \int_0^1 t^a (1-t)^{b-1} \mathrm{d}t$ is the beta function. The corresponding cumulative distribution function (c.d.f.) is the regularized incomplete beta function, given by

$$F_B(z; a, b) = \frac{B_z(a, b)}{B(a, b)}, \quad 0 < z < 1,$$

where $B_x(a, b)$ denotes the incomplete beta function, i.e., $B_x(a, b) = \int_0^x t^{a-1} (1-t)^{b-1} \mathrm{d}t$. For a detailed discussion of beta distribution, interested readers may refer to Johnson et al. [16], Gupta and Nadarajah [10] and Nadarajah and Kotz [28].

Regression models are typically constructed to model the mean of a distribution. In this context, a parameterization of the beta distribution in terms of mean and precision parameters was proposed by Jørgensen [19]. Let be $\mu = a/(a+b)$ and $\phi = a + b$, i.e., $a = \mu\phi$ and $b = (1-\mu)\phi$, then $\mathrm{E}(Z) = \mu$ and $\mathrm{Var}(Z) = \mu(1-\mu)/(1+\phi)$. This parameterization is useful to define the beta regression model since $0 < \mu < 1$ is the mean of Z and $\phi > 0$ is a precision parameter. Under this parameterization, Ferrari and Cribari-Neto [8] introduced the beta regression model.

By switching the roles of c.d.f. and quantile function of the beta distribution, Jones [17] introduced the CB distribution, which the c.d.f. is given by

$$F(y; a, b) = F_B^{-1}(y; a, b) = \mathscr{I}_y(a, b), \quad 0 < y < 1,$$

where $a > 0$ and $b > 0$ are shape parameters, and $\mathscr{I}_y(a, b)$ denotes the inverse of the incomplete beta ratio. The p.d.f. of CB is defined by

$$f(y; a, b) = B(a, b) \left[\mathscr{I}_y(a, b)\right]^{1-a} \left[1 - \mathscr{I}_y(a, b)\right]^{1-b}.$$

The CB distribution does not have explicit moments of order k; however, for $k = 1$ and $k = 2$, Jones [17] provided the following expressions

$$E(Y) = \frac{b}{a+b} \quad \text{and} \quad E\left(Y^2\right) = \frac{2\,B(2\,a, 2\,b+1)}{a\,[B(a,b)]^2}\,{}_3F_2\,(a+b, 1, 2\,a; a+1, 2\,(a+b)+1; 1)\,,$$

where ${}_3F_2$ denotes the generalized hypergeometric function (Gradshteyn and Ryzhik, [9], formula 7.512.5). From Dixon [6], the definition of ${}_3F_2(a, b, c; d, e; 1)$ is

$${}_3F_2(a, b, c; d, e; 1) = \frac{(a/2)!\,(a-b)!\,(a-c)!\,(a/2-b-c)!}{a!\,(a/2-b)!\,(a/2-c)!\,(a-b-c)!},$$

where $1 + a/2 - b - c$ has a positive real part, $d = a - b + 1$ and $e = a - c + 1$.

According to Jones [17], if Y has CB distribution with shape parameters a and b, then $1 - Y$ has a CB distribution with shape parameters b and a. Furthermore, some distributions are special cases of the CB distribution, such as the standard uniform ($a = b = 1$), power function ($b = 1$), and cosine ($a = b = 1/2$) distributions. For $a = 1$, we have the beta distribution with parameters 1 and $1/b$. More recently, Jones [18] studied theoretical structures underlying of families of complementary distributions that can take values on (0, 1).

In order to introduce a regression structure for the mean and the dispersion of the response variable, a different parameterization of the CB distribution is needed. Let be $\mu = b/(a + b)$ and $\phi = a + b$, i.e., $a = \phi(1 - \mu)$ and $b = \phi\mu$; hence, the p.d.f. and c.d.f. of CB distribution can be written, respectively, as follows

$$f(y; \mu, \phi) = B(\phi(1-\mu), \phi\mu) \left[\mathscr{I}_y(\phi(1-\mu), \phi\mu)\right]^{1-\phi(1-\mu)} \left[1 - \mathscr{I}_y(\phi(1-\mu), \phi\mu)\right]^{1-\phi\mu} \quad (1)$$

and

$$F(y; \mu, \phi) = \mathscr{I}_y(\phi(1-\mu), \phi\mu), \quad (2)$$

where $0 < \mu < 1$ and $\phi > 0$. From now on, the notation $Y \sim \mathrm{CB}(\mu, \phi)$ is used to indicate that Y is a random variable following a CB distribution with mean μ and dispersion parameter ϕ.

The L-moments, whose theory was unified by Hosking [13], are linear combinations of order statistics. The first L-moment is just the mean, while the second L-moment, λ_2 is a measure of spread, more precisely is one-half of Gini's mean difference. For the proposed parameterization of CB distribution, the second L-moment is given by

$$\lambda_2 = \frac{\phi\,\mu\,(1-\mu)}{1+\phi}. \quad (3)$$

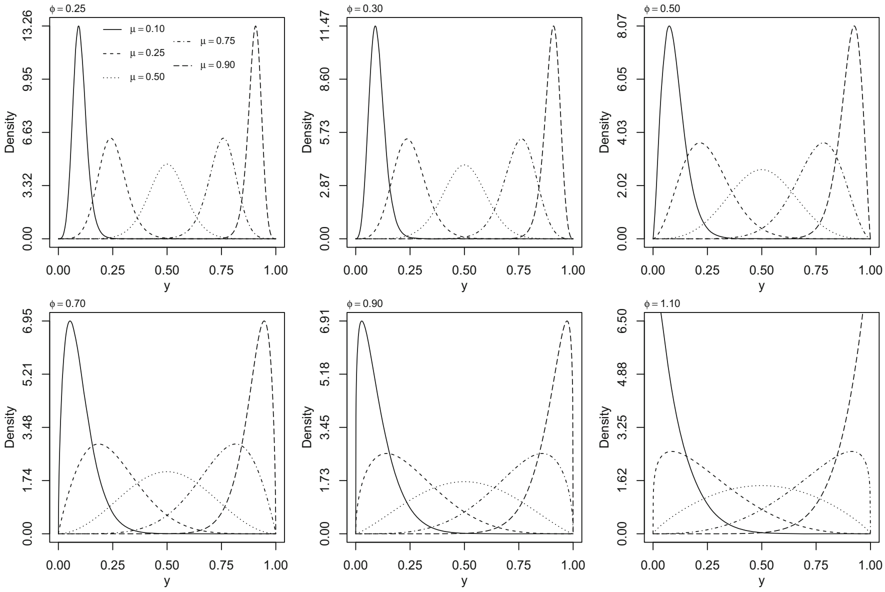

Fig. 1 Probability density function of the CB distribution for selected values of μ and ϕ

For fixed μ, the larger the value of ϕ, the greater the spread of CB distribution, justifying that ϕ can be interpreted as a dispersion parameter.

In Fig. 1, the p.d.f. of CB is displayed for several choices of parameter values, It is observed that CB density has the same basic shape properties of the Beta distribution, particularly, unimodal, U-shaped, J-shaped and reverse J-shaped.

The third and fourth L-moment ratios, $\tau_3 = \lambda_3/\lambda_2$ and $\tau_4 = \lambda_4/\lambda_2$, are important measures of the skewness and kurtosis of a distribution, respectively [14]. For the new CB parametrization, these quantities are given, respectively, by

$$\tau_3 = \frac{\phi}{\phi+2}(1-2\mu) \quad \text{and} \quad \tau_4 = \frac{\phi^2(1-2\,\mu)^2 - \phi^2\mu(1-\mu)+1}{(\phi+2)(\phi+3)}.$$

In Fig. 2, the behavior of variance, λ_2, τ_3 and τ_4 is illustrated as a function of mean for some values of ϕ. The variance values were computed by Monte Carlo simulation. It is observed that the variance and λ_2 reach the maximum at 0.25 for $\mu = 1/2$ and large values of ϕ. This property is also shared with beta distribution. Furthermore, it is observed that CB is a symmetric distribution for $\mu = 1/2$, positive skew for $\mu < 1/2$ and negative skew for $\mu > 1/2$, the intensity of skewness increases as ϕ increases.

In order to compare the CB and beta distributions used in regression analysis, values for the parameter ϕ were obtained to have different variance for fixed $\mu = 0.5$. The c.d.f.'s presented in Fig. 3 shows the flexibility of both distributions. In the extreme case for $\text{Var}(Y) = 0.2$, both distributions are very similar. Nevertheless, for smaller variances and mean far from 0.5 the distributions are quite different. These

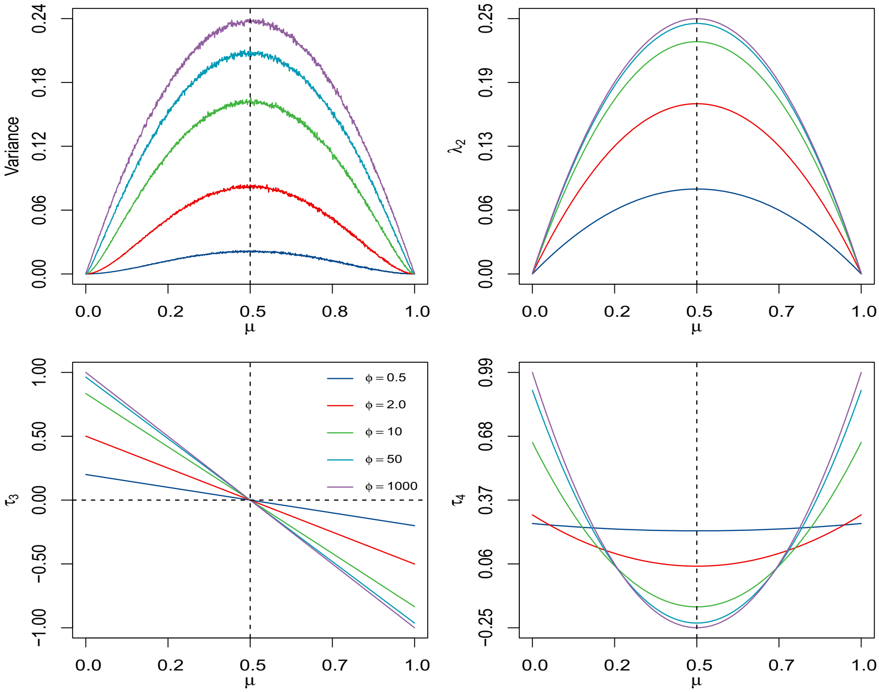

Fig. 2 Behavior of variance, spread (λ_2), skewness (τ_3) and kurtosis (τ_4) as function of mean (μ) for selected values of ϕ

results demonstrate the versatility of CB distribution under the practical point of view, meaning that the CB can fit better than beta depending on the phenomena.

3 Regression Model

Let $Y_1, \ldots, Y_n$ be n independent random variables, where each Y_i, $i = 1, \ldots, n$, follows the p.d.f. given in (1) with mean μ_i and dispersion parameter ϕ_i. Suppose that the mean and the precision parameter of Y_i satisfies the following functional relations

$$g_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad g_2(\phi_i) = \mathbf{w}_i^\top \boldsymbol{\alpha}, \tag{4}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_{p-1})^\top$ and $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_{q-1})^\top$ are vectors of unknown regression coefficients which are assumed to be functionally independent, $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\boldsymbol{\alpha} \in \mathbb{R}^q$, with $p + q < n$, and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\top$ and $\mathbf{w}_i = (1, w_{i1}, \ldots, w_{iq})^\top$ are observations on p and q known covariates, for $i = 1, \ldots, n$. Furthermore, assume that the covariate matrices $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top$ and $\mathbf{W} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)^\top$ have rank p and q, respectively. The link functions $g_1 : (0, 1) \rightarrow \mathbb{R}$ and $g_2 : \mathbb{R}^+ \rightarrow \mathbb{R}$ in (4) must be strictly monotone, positive and twice

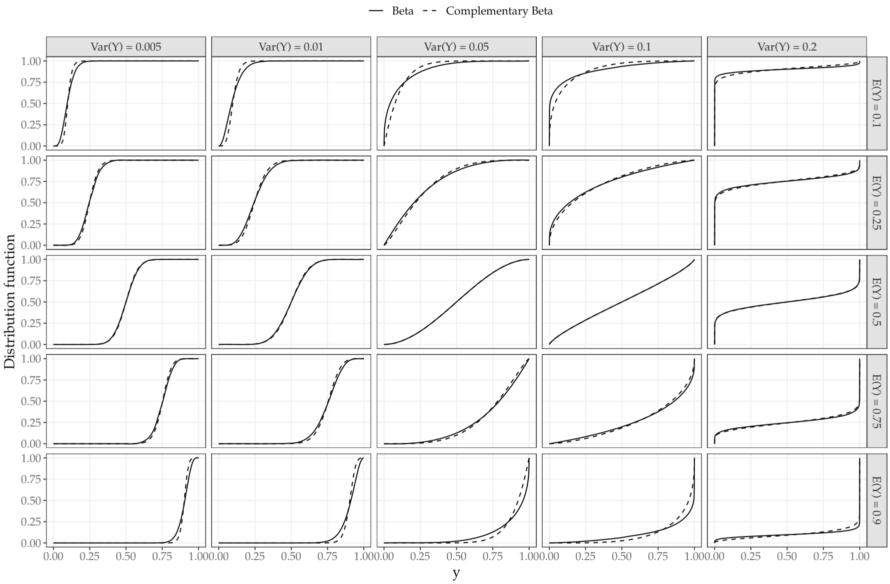

Fig. 3 Beta and CB c.d.f.'s for distinct values of variance and mean (fixed for $\mu = 0.5$)

differentiable, such that $\mu_i = g_1^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta})$ and $\phi_i = g_2^{-1}(\mathbf{w}_i^\top \boldsymbol{\alpha})$, with $g_1^{-1}(\cdot)$ and $g_2^{-1}(\cdot)$ being the inverse functions of $g_1(\cdot)$ and $g_2(\cdot)$, respectively.

There are several possibilities for the link functions $g_1(\cdot)$ and $g_2(\cdot)$. The most useful well-known link functions for $g_1(\cdot)$ are logit: $g(\mu_i) = \log(\mu_i/(1-\mu_i))$; probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function; and complementary log-log: $g(\mu_i) = \log\left[-\log(1-\mu_i)\right]$. Whereas, for $g_2(\cdot)$ are the logarithm: $g_2(\phi_i) = \log(\phi_i)$, and the square root $g_2(\phi_i) = \sqrt{\phi_i}$. An excellent discussion on link functions can be found in Atkinson [2] and McCullagh and Nelder [25]. Due to the direct interpretation of the parameters in terms of odds, in this paper we consider the logit link for $g_1(\cdot)$. Its interpretation when μ_i is the mean is given in [8]. For $g_2(\cdot)$, we consider the log link, since it is the most used for positive parameters.

3.1 Estimation and Inference

The log-likelihood function for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$ based on a sample of n independent observations is given by

$$\ell(\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{i=1}^{n} \ell(\mu_i, \phi_i), \tag{5}$$

where

$$\begin{aligned} \ell(\mu_i, \phi_i) = {} & \log[B(\phi_i(1-\mu_i), \phi_i\mu_i)] + (1-\phi_i(1-\mu_i)) \log\left[\mathscr{I}_{y_i}(\phi_i(1-\mu_i), \phi_i\mu_i)\right] \\ & + (1-\phi_i\mu_i) \log\left[1-\mathscr{I}_{y_i}(\phi_i(1-\mu_i), \phi_i\mu_i)\right]. \end{aligned} \tag{6}$$

The maximum likelihood estimator (MLE) $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\alpha}}^\top)^\top$ of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ is obtained by the maximization of the log-likelihood function (5). However, it is not possible to derive analytical solution for the MLE $\widehat{\boldsymbol{\theta}}$; hence, we must required to numerical solution using some optimization algorithm such as Newton–Raphson and quasi-Newton. Following Ferrari and Cribari-Neto [8], we suggested to use it as an initial guess for $\boldsymbol{\beta}$ the ordinary least squares estimates obtained from the linear regression of the transformed responses $g(y_1), \ldots, g(y_n)$ on $\mathbf{X}$, i.e., $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top z$, where $z = (g(y_1), \ldots, g(y_n))^\top$. As a initial guess for $\boldsymbol{\alpha}$, we suggest to use the q-dimensional vector of zeros ($\mathbf{0}_q$), which implies that the algorithm starts with CB regression model with constant dispersion.

Under mild regularity conditions [5] and when n is large, the asymptotic distribution of the MLE $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^\top, \widehat{\boldsymbol{\alpha}}^\top)^\top$ is approximately multivariate normal (of dimension $p+q$) with mean vector $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha}^\top)^\top$ and variance covariance matrix $\mathbf{K}^{-1}(\boldsymbol{\theta})$ where

$$\mathbf{K}(\boldsymbol{\theta}) = \mathbb{E}\left[-\frac{\partial \ell\,(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\;\partial \boldsymbol{\theta}^\top}\right],$$

is the expected Fisher information matrix. Unfortunately, there is no closed form expression for the matrix $\mathbf{K}(\boldsymbol{\theta})$. Nevertheless, a consistent estimator of the expected Fisher information matrix is given by

$$\mathbf{J}(\widehat{\boldsymbol{\theta}}) = -\left.\frac{\partial \ell\,(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\;\partial \boldsymbol{\theta}^\top}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}},$$

which is the estimated observed Fisher information matrix. Therefore, for large n, we can replace $\mathbf{K}(\boldsymbol{\theta})$ by $\mathbf{J}(\widehat{\boldsymbol{\theta}})$.

Let θ_r be the r-th component of $\boldsymbol{\theta}$. The asymptotic $100(1-\gamma)\%$ confidence interval for θ_r is given by

$$\widehat{\theta_r} \pm z_{\gamma/2}\ \mathrm{se}\left(\widehat{\theta_r}\right), \qquad r = 1, \ldots, p+q,$$

where $z_{\gamma/2}$ is the $\gamma/2$ upper quantile of the standard normal distribution and $\mathrm{se}\left(\widehat{\theta_r}\right)$ is the asymptotic standard error of $\widehat{\theta_r}$. Note that $\mathrm{se}\left(\widehat{\theta_r}\right)$ is the square root of the r-th diagonal element of the matrix $\mathbf{J}^{-1}(\widehat{\boldsymbol{\theta}})$.

Large sample inference can be conducted to test if the precision parameter does not vary across the observations. Consider the test of the null hypothesis $H_0 : \alpha_j = 0\,\forall\, j = 1, \ldots, q-1$ versus $H_0 : \alpha_j \neq 0$ for at least one j. We proposed the likelihood ratio (LR) test [29]. The LR statistic is given by

$$S_{LR} = 2\left[\ell(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}) - \ell(\widetilde{\boldsymbol{\beta}}, \widetilde{\alpha}_0)\right],$$

where $\ell(\boldsymbol{\beta}, \boldsymbol{\alpha})$ is the log-likelihood function given in (6) and $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\beta}}^\top, \widetilde{\alpha}_0)^\top$ is the restricted MLE of $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\alpha})^\top$. Under the usual regularity conditions and under H_0 S_{LR} converges in distribution to χ^2_{q-1}. The null hypothesis is rejected at level of significance γ if $S_{LR} > \chi^2_{1-\alpha,q-1}$, where $\chi^2_{1-\gamma,q-1}$ is the $1-\gamma$ upper quantile of Chi-square distribution with $q-1$ degrees of freedom.

3.2 Model Adequacy

In order to evaluate and study departures from the model assumption, we propose to use the randomized quantile residuals introduced by Dunn and Smyth [7]. These residuals for the CB regression model is defined by

$$r_i = \Phi^{-1}\{\mathscr{I}_{y_i}(\widehat{\mu}_i, \widehat{\phi}_i)\}, \quad i = 1, 2, \ldots, n,$$

where $\Phi^{-1}(\cdot)$ is the standard normal distribution function and $\mathscr{I}_{y_i}(\widehat{\mu}_i, \widehat{\phi}_i)$ is the c.d.f. of CB distribution given by (2) with $\widehat{\mu}_i = g_1^{-1}(\mathbf{x}_i^\top \widehat{\boldsymbol{\beta}})$ and $\widehat{\phi}_i = g_2^{-1}(\mathbf{w}_i^\top \widehat{\boldsymbol{\alpha}})$.

Apart from the variability due the estimates of parameters, these residuals have standard normal distribution if the proposed model is correctly specified [7]. Thus, to check if the model assumption is adequate, we can examine the residuals plots with simulated envelope proposed by Atkinson [3]. The simulated envelope can be construct as follows

(i) fit the model and generate sample set of n independent observations using the estimated parameters of the fitted model;
(ii) fit the model from the generated sample, calculate the values of the residuals and arrange them in order;
(iii) repeat steps (i) and (ii) B number of times;
(iv) consider the n sets of the B ordered statistics of the residuals, then for each set calculate the quantile $\gamma/2$, the median and the quantile $1-\gamma/2$;
(v) plot these values and the ordered residuals of the original sample set versus the expected order statistics of a normal distribution, which is approximated as

$$\Phi^{-1}\left(\frac{i+0.375}{n+0.25}\right).$$

4 Monte Carlo Studies

In this section, we conducted Monte Carlo simulations (i) to evaluate the finite-sample behavior of the maximum likelihood estimates of the regression coefficients and (ii) to investigate the empirical distribution of the randomized quantile residuals.

The Monte Carlo experiments were carried out considering the following regression structure

$$g_1(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_i, \quad i = 1, \ldots, n$$

$$g_2(\phi_i) = \log(\phi_i) = \alpha_0 + \alpha_1 w_i, \quad i = 1, \ldots, n,$$

Table 1 Estimated bias, mean-squared error and coverage probability

n	Bias				RMSE				$\text{CP}_{95\%}$			
	β_0	β_1	α_0	α_1	β_0	β_1	α_0	α_1	β_0	β_1	α_0	α_1
50	−0.007	0.011	−0.047	−0.004	0.127	0.146	0.332	0.523	93.22	93.34	93.30	93.72
100	−0.004	0.006	−0.021	−0.005	0.090	0.104	0.215	0.366	94.00	94.36	93.82	94.68
150	−0.003	0.004	−0.014	−0.001	0.073	0.079	0.180	0.297	94.30	94.74	94.04	94.36
200	−0.003	0.003	−0.011	−0.000	0.063	0.066	0.157	0.258	93.96	94.22	94.22	94.64
300	−0.002	0.003	−0.008	−0.001	0.051	0.054	0.129	0.215	94.28	94.28	94.02	94.24
500	−0.001	0.002	−0.004	−0.002	0.039	0.040	0.101	0.168	94.36	94.64	94.08	94.46

where the true values of the parameters were taken as $\beta_0 = -1.0$, $\beta_1 = 1.0$, $\alpha_0 = 0.5$ and $\alpha_1 = -1.0$. The covariate values of x_i were generated from the standard normal distribution, while the values of w_i were draws from the standard uniform distribution.

All simulations were conducted in SAS using the quasi-Newton algorithm available in the `NLMIXED` procedure [30] to obtain the maximum likelihood estimates. For each scenario, the Monte Carlo experiment was repeated 5, 000 times. These results are presented in the next subsection, and the SAS codes are available from the authors upon request.

4.1 Parameter Estimation

In this subsection, a small simulation study is presented to observe the finite sample performance of the proposed estimators. For such evaluation, the estimated bias, the estimated root-mean squared error (RMSE) and the coverage probability of 95% confidence interval ($\text{CP}_{95\%}$) were computed. The results are presented in Table 1 and Fig. 4.

Table 1 presents the biases, MSE and $\text{CP}_{95\%}$ of the estimators of the parameters β_0, β_1, α_0 and α_1. As expected, increasing the sample size reduces substantially both bias and RMSE. Furthermore, note that the asymptotic confidence intervals have an empirical coverage probability that is less than the nominal value 0.95. Overall, we observe that the asymptotic confidence intervals have a good performance. The previous findings are confirmed by the box plots shown in Fig. 4.

4.2 Residuals

The second simulation study was performed to examine how well the distribution of the randomized quantile residuals are approximated by the standard normal distribution. The evaluation of the randomized quantile residuals was based on the normal probability plots of the mean order statistics and descriptive measures. The results are presented in Table 2 and Fig. 5.

In Table 2, we present the mean, standard deviation (StdDev), skewness and kurtosis of randomized quantile residuals. Table 2 shows that there is a good overall agreement between the sample and population values for the randomized quantile residuals. The

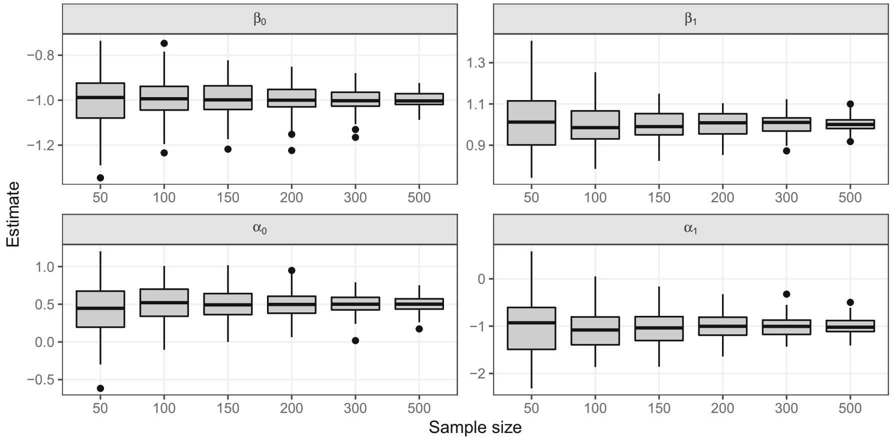

Fig. 4 Boxplots of the estimates parameters obtained in Monte Carlo experiments for different sample sizes

Table 2 Descriptive measures of the randomized quantile residuals

n	Mean	StdDev	Skewness	Kurtosis
50	−0.0004	0.9984	−0.2134	2.9988
100	0.0001	0.9993	−0.1139	3.0022
150	0.0000	0.9995	−0.0735	3.0030
200	0.0001	0.9997	−0.0537	3.0016
300	0.0000	0.9997	−0.0368	3.0034
500	−0.0000	0.9998	−0.0229	3.0039

residuals have approximately zero mean and unit standard deviation, have skewness close to zero, which indicates that these residuals are approximately symmetrical, and kurtosis is near three.

Figure 5 displays the empirical versus theoretical quantiles plots of the randomized quantile residuals. As expected, the simulations show that the sample distribution of randomized quantile residuals can be approximated by a standard normal distribution. Therefore, the randomized quantile residuals can be useful for model diagnostics.

5 Real-World Data Analysis

In this section, two empirical applications of the proposed model to real data are presented to compare the potentiality of the CB regression with the traditional beta regression model. Parameter estimates were performed under the maximum likelihood paradigm, as discussed in Sect. 3, by using the `SAS/NLMIXED` procedure [30]. The asymptotic standard errors and confidence intervals were computed using the observed Fisher information matrix.

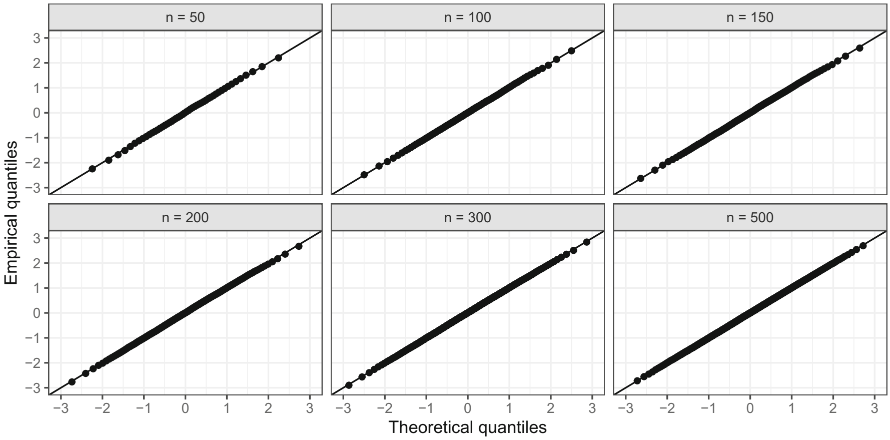

Fig. 5 Normal probability plots of the mean order statistics

The model selection is carried out using the AIC (Akaike information criterion) [1], the BIC (Bayesian information criterion) [31] and the HQIC (Hannan–Quinn information criteria) [11]

$$\mathrm{AIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2q, \quad \mathrm{BIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + q\,\log(n) \;\text{ and }\; \mathrm{HQIC} = -2\ell(\widehat{\boldsymbol{\theta}}) + 2\,q\,\log\left(\log(n)\right),$$

where $\ell(\widehat{\boldsymbol{\theta}})$ denotes the log-likelihood function evaluated at the MLE, q is the number of parameters, and n is the sample size. In all these criteria, the decision rule is favorable to the model with the lowest value [12].

Although both regression model are parameterized in terms of mean, and estimates for the mean regression coefficients should be similar in magnitude and sign, it should be highlighted that the parameter ϕ is clearly distinct between the two regression models. Particularly, for the beta distribution ϕ denotes the precision, while for the CB distribution denotes the dispersion.

5.1 Illiteracy Rate Data

In this application, the information of illiteracy rate of people between 25 e 29 years in the 141 cities of Mato Grosso, a state localized in the Midwest Brazil, is considered. The data set refers to the census of 2010, and it is available at http://atlasbrasil.org.br/2013/. The goal is to analyze the association between illiteracy rate and the municipal human development index (MDHI). The MHDI is used as explanatory variable since it is an important measure to guide authorities to assess progress and social reality as well as to define public policy priorities and comparisons of different cities [26].

Table 3 reports the mean, standard deviation (Std) and second sample L-moment (L_2) of the illiteracy rate according to MDHI of the cities in Mato Grosso (MT), Brazil. It is clear that the mean change as the MDHI increase. In contrast, the changes in Std and L_2 across the MDHI are more slight.

Table 3 Descriptive statistics of illiteracy rate according to different class of MDHI

MHDI	n	Mean	Std	L_2
[0.599; 0.653)	22	0.047	0.018	0.011
[0.655; 0.668)	23	0.043	0.014	0.008
[0.669; 0.686)	23	0.032	0.012	0.007
[0.687; 0.699)	21	0.032	0.011	0.006
[0.701; 0.716)	26	0.028	0.010	0.006
[0.718; 0.785)	23	0.020	0.008	0.005

Table 4 Summary of the fitted models—Illiteracy rate data

Parameter	CB			Beta		
	MLE	S.E.	95% C.I.	MLE	S.E.	95% C.I.
β_0	2.5012	0.6409	(1.2340, 3.7684)	1.4853	0.5853	(0.3281, 2.6426)
β_1	−8.5922	0.9319	(−10.435, −6.7497)	−7.1032	0.8613	(−8.8063, −5.4001)
ϕ	0.2620	0.0203	(0.2217, 0.3022)	217.92	26.4876	(165.54, 270.29)

Table 5 The likelihood-based statistics of fit—Illiteracy rate data

Criteria	CB	Beta
AIC	−843.1081	−836.5412
BIC	−834.3263	−827.7595
HQIC	−839.5394	−832.9725

This empirical behavior induced a regression only for the mean parameter since the L_2 has a little change according to the levels of MDHI. The likelihood ratio test computed for $H_0 : \phi_i = \phi$ versus $H_1 : \phi_i \neq \phi$ returned a statistic $S_{LR} = 1.2781$ and p value $= 0.7417$, corroborating the empirical analysis that for this data set the dispersion parameter ϕ can be constant across the covariate.

Thus, the regression structure for CB and Beta distribution is given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \text{MDHI}_i, \qquad i = 1, \ldots, 141.$$

Table 4 gives the MLE and the 95% confidence intervals for both models. Although the sign of coefficient β_1 is the same for both model in CB regression, the impact of MDHI is bigger than the Beta regression. Table 5 lists the values of the likelihood-based statistics for both models. It is observed that the CB regression model provides the best fit, since it has the lowest values of AIC, BIC and HQIC measures. These claim is also supported by the residuals plots with simulated envelopes shown in Fig. 6.

The inference results of CB regression model indicated that the MDHI has a negative impact on the illiteracy rate of the cities in MT. This means that cities with greater MDHI have less proportion of people that do not know to read and write. This fact is

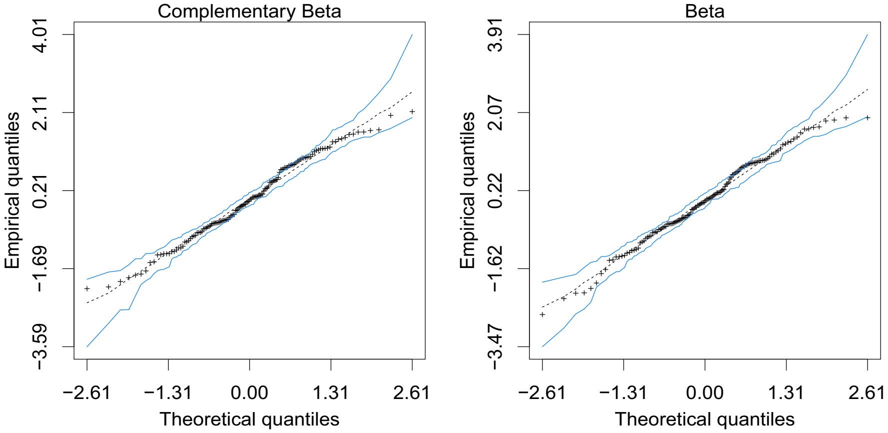

Fig. 6 Randomized quantile residuals with simulated envelope—Illiteracy rate data

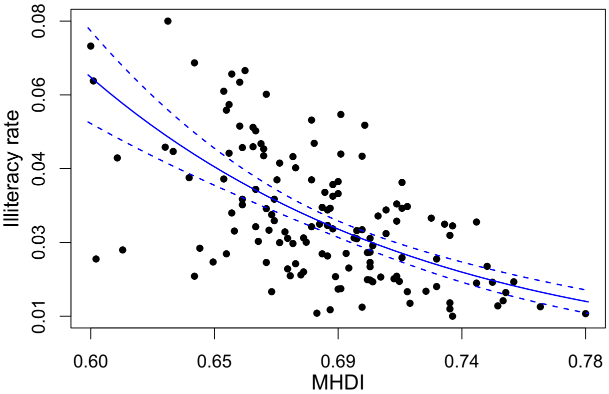

Fig. 7 Observed values of illiteracy rate versus MHDI with the fitted mean of CB and the 95% confidence interval

showed in Fig. 7, where the illiteracy rate versus MHDI is plotted along with the fitted mean and the 95% confidence interval from CB regression model.

5.2 Recovery Rate of CD34+ cells Data

This analysis corresponds to a study conducted with 239 patient between 2003 and 2008 at the Edmonton Hematopoietic Stem Cell Lab in Cross Cancer Institute – Alberta Health Services. The data set was extracted from Zhang et al. [33], and the goal is to model the recovery rate of CD34+ cells after peripheral blood stem cell (PBSC) transplants. The covariates associated with this response variable are: Sex: 0 for female, 1 for male; Chemo: 0 for receiving a chemotherapy on a one-day protocol, 1 for a 3-day protocol and Age: adjusted patient's age, i.e., the current age minus 40.

Table 6 Descriptive statistics of recovery rate of CD34+ cells according to the covariate

Covariate	n	Mean	Std	L_2
Age				
[0; 5)	65	0.776	0.123	0.069
[6; 12)	34	0.739	0.119	0.067
[13; 16)	26	0.803	0.086	0.049
[17; 20)	43	0.822	0.108	0.058
[21; 24)	37	0.801	0.112	0.062
[25; 31)	34	0.811	0.105	0.058
Chemo				
0	130	0.785	0.122	0.068
1	109	0.797	0.104	0.057
Sex				
Female	71	0.780	0.119	0.065
Male	168	0.795	0.112	0.062

Table 7 Summary of the fitted models —Recovery rate of CD34+ cells data

Parameter	CB			Beta		
	MLE	S.E.	95% C.I.	MLE	S.E.	95% C.I.
β_0	1.0002	0.1135	(0.7779, 1.2226)	1.0422	0.1137	(0.8193, 1.2650)
β_1	0.0166	0.0053	(0.0062, 0.0271)	0.0143	0.0054	(0.0038, 0.0248)
β_2	0.2354	0.1024	(0.0347, 0.4360)	0.2143	0.1037	(0.0111, 0.4175)
ϕ	0.6181	0.0408	(0.5382, 0.6981)	11.322	1.0159	(9.3303, 13.313)

Firstly, we considered the regression structure for mean and dispersion parameter with all covariate. Then, we conducted a likelihood ratio test for $H_0 : \phi_i = \phi$ versus $H_1 : \phi_i \neq \phi$ the test statistics returned for CB regression was $S_{LR} = 0.8094$ and the corresponding p-value $= 0.6671$. We also noted that the Sex covariate was not important for the analysis. Hence, the final regression model selected is given by

$$\operatorname{logit}(\mu_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Chemo}_i, \qquad i = 1, \ldots, 239.$$

The point estimates and the 95% confidence intervals for the parameters of the considered three regression models are given in Table 7. These results indicate that both models provide similar regression estimates for the mean parameters, leading to identical interpretations. It should be emphasized that the coefficients of CB regression have the smallest standard errors. Table 8 gives the values of the likelihood-based statistics for the models. The three information criteria indicate that the CB regression model presented a better fit than the Beta regression.

To check the model assumption, in Fig. 8 it is shown the residuals plots with simulated envelope. We can conclude that the CB regression model is a great alternative model to describe these dataset.

Table 8 The likelihood-based statistics of fit—Recovery rate of CD34+ cells data

Criteria	CB	Beta
AIC	−388.9434	−383.3042
BIC	−375.0375	−369.3983
HQIC	−383.3397	−377.7005

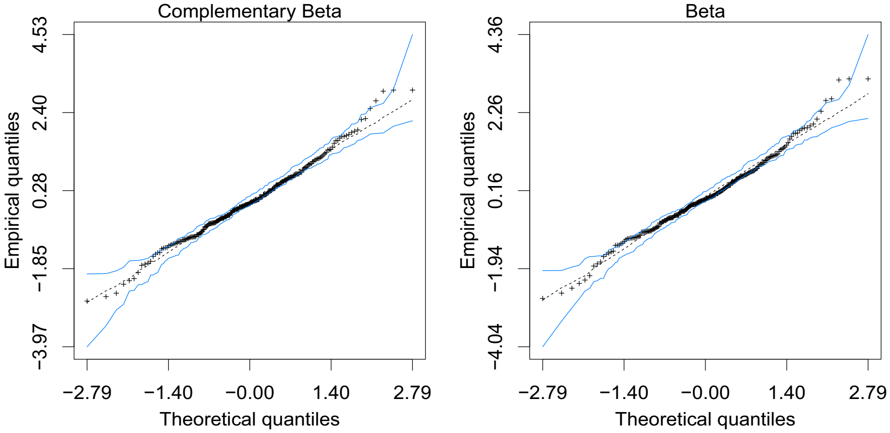

Fig. 8 Randomized quantile residuals with simulated envelope—Recovery rate of CD34+ cells data

6 Concluding Remarks

In this paper, we studied a new regression model for modeling bounded data. The new regression model is based on the unknown CB distribution. In particular, we have developed a new parameterized CB distribution in terms of the mean and dispersion parameters. The proposed regression model can be used for any application that involves unit interval data and is a natural strong competitor of the beta regression model. The major limitation of the proposed regression model is that the variance of CB distribution cannot be expressed in closed form (involves the generalized hypergeometric function). Maximum likelihood inference is implemented for estimating the model parameters, and its good performance has been evaluated by means of Monte Carlo simulations. Two real data sets were analyzed for illustrative and model comparison purposes. For these data sets, the proposed regression model has outperformed the usual beta model. Results of the applications showed that the proposed model is more adequate than usual beta regression model. We hope that this new model may attract wider applications for modeling and analyzing bounded data. As part of future research, we plan to extend the proposed regression model to the case of data contain zeros or ones.

References

1. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19(6):716–723
2. Atkinson A (1985) Plots. transformation and regression. Clarendon Press, Oxford
3. Atkinson AC (1981) Two graphical displays for outlying and influential observations in regression. Biometrika 68(1):13–20
4. Cepeda-Cuervo E (2001) Variability modeling in generalized linear models. Ph.D. thesis, Mathematics Institute, Universidade Federal do Rio de Janeiro
5. Cox DR, Hinkley DV (1979) Theoretical statistics. CRC Press, Boca Raton
6. Dixon AC (1902) Summation of a certain series. Proc London Math Soc 35(1):284–291
7. Dunn PK, Smyth GK (1996) Randomized quantile residuals. J Comput Gr Stat 5(3):236–244
8. Ferrari S, Cribari-Neto F (2004) Beta regression for modelling rates and proportions. J Appl Stat 31(7):799–815
9. Gradshteyn I, Ryzhik I (1994) Table of integrals, series, and products. Academic Press, Cambridge
10. Gupta AK, Nadarajah S (2004) Handbook of beta distribution and its applications. CRC Press, Boca Raton
11. Hannan EJ, Quinn BG (1979) The determination of the order of an Autoregression. J R Stat Soc Ser B (Methodol) 41(2):190–195
12. Held L, Sabanés Bové D (2014) Applied statistical inference - likelihood and bayes. Springer, Ney York
13. Hosking JRM (1990) L-Moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc Ser B (Methodol) 52(1):105–124
14. Hosking JRM (1992) Moments or L Moments? An example comparing two measures of distributional shape. Am Stat 46(3):186–189
15. Iacobellis V (2008) Probabilistic model for the estimation of T year flow duration curves. Water Resour Res 44(2):02413
16. Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions, vol 2, 2nd edn. Wiley, New York
17. Jones MC (2002) The complementary Beta distribution. J Stat Plan Inference 104(2):329–337
18. Jones MC (2018) Families of complementary distributions. Stat Probabil Lett 141:74–81
19. Jorgensen B (1997) The theory of dispersion models. Chapman & Hall, USA
20. Kamps U (1991) A general recurrence relation for moments of order statistics in a class of probability distributions and characterizations. Metrika 38:215–225
21. Kieschnick R, McCullough BD (2003) Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions. Stat Modell 3(3):193–213
22. Lemonte AJ, Bazán JL (2015) New class of Johnson distributions and its associated regression model for rates and proportions. Biometrical J 58(4):727–746
23. Mazucheli J, Menezes AFB, Chakraborty S (2019) On the one parameter unit-Lindley distribution and its associated regression model for proportion data. J Appl Stat 46(4):700–714
24. Mazucheli J, Menezes AFB, Fernandes LB, de Oliveira RP, Ghitany ME (2019) The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. J Appl Stat 47:954
25. McCullagh P, Nelder J (1989) Generalized linear models, 2nd edn. Chapman and Hall, USA
26. Menezes AFB, Furriel WO (2019) Beta and Simplex regression models in the analysis of the municipal human development index 2010. Rev Bras Biom 37(3):394–408
27. Mitnik PA, Baek S (2013) The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. Stat Papers 54(1):177–192
28. Nadarajah S, Kotz S (2007) Multitude of Beta distributions with applications. Stat: J Theor Appl Stat 41(2):153–179
29. Neyman J, Pearson ES (1928) On the use and interpretation of certain test criteria for purposes of statistical inference. Biometrika 20A(1/2):175–240
30. SAS, 2010. The NLMIXED Procedure, SAS/STAT® User's Guide, Version 9.4. Cary, NC: SAS Institute Inc
31. Schwarz G et al (1978) Estimating the dimension of a model. Annals Stat 6(2):461–464
32. Smithson M, Verkuilen J (2006) A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. Psychol Methods 11:54–71

33. Zhang P, Qiu Z, Shi C (2016) Simplexreg: an R package for regression analysis of proportional data using the Simplex distribution. J Stat Softw 71(11):1–21

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.