Received: 30 April 2020 | Revised: 3 October 2020 | Accepted: 24 October 2020

DOI: 10.1002/bimj.202000126

Biometrical Journal

RESEARCH PAPER

A parametric quantile regression approach for modeling zero-or-one inflated double bounded data

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

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

[2] Departamento de Estatística, Universidade Estadual de Maringá, PR, Brazil

[3] Departamento de Estatística, Universidade Federal do Rio Grande do Norte, RN, Brazil

Correspondence
Marcelo Bourguignon, Universidade Federal do Rio Grande do Norte, Departamento de Estatística, RN, Brazil.
Email: m.p.bourguignon@gmail.com

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

Abstract

Over the last decades, the challenges in applied regression have been changing considerably, and full probabilistic modeling rather than predicting just means is crucial in many applications. Motivated by two applications where the response variable is observed on the unit-interval and inflated at zero or one, we propose a parametric quantile regression considering the unit-Weibull distribution. In particular, we are interested in quantifying the influence of covariates on the quantiles of the response variable. The maximum likelihood method is used for parameters estimation. Monte Carlo simulations reveal that the maximum likelihood estimators are nearly unbiased and consistent. Also, we define a residual analysis to assess the goodness of fit.

KEYWORDS

parametric quantile regression, proportions, unit-Weibull distribution, zero-or-one inflated models

1 | INTRODUCTION

Frequently, in real life data, fractions, rates, or proportions data contain zeros and/or ones. Therefore, the study of models to model data observed on the intervals $[0, 1)$ or $(0, 1]$ motivates a novel research branch with many practical applications, such as modeling the corporate capital structure decisions (Cook, Kieschnick, & McCullough, 2008), the mortality in traffic accidents (Ospina & Ferrari, 2012), the relative payment amount (Pereira, Botter, & Sandoval, 2013), the proportion of households with access to electricity (Santos & Bolfarine, 2015), leverage ratios (Bayes & Valdivieso, 2016), and Parkinson's disease (Di Brisco & Migliorati, 2020). In such cases, the usual linear models are not suitable for modeling these data.

In this context, for independent data, Ospina and Ferrari (2008) proposed inflated beta distributions as natural alternatives to the beta distributions for modeling data observed in $[0, 1)$, $(0, 1]$, or $[0, 1]$. Recently, Cribari-Neto and Santos (2019) established the inflated Kumaraswamy distributions. In order to accommodate explanatory variables in the modeling, Hoff (2007) introduced the one-inflated beta model. Cook et al. (2008) considered the beta regression models inflated at zero. Ospina and Ferrari (2012) proposed a general class of regression models for continuous proportions when the data contain zeros or ones. Santos and Bolfarine (2015) introduced zero-or-one inflated quantile model by considering suitiable transformation under the well known Koenker and Bassett (1978) quantile regression. These authors considered the formulation using the assymetric Laplace distribution discussed by Koenker and Machado (1999) in order to performed Bayesian inference. More recently, Liu, Kam Yuen, Wu, Tian, and Li (2020) studied the zero-or-one inflated simplex regression models for the analysis of continuous proportion data.

A substantial number of practical and theoretical studies have focused on the use of the mean reparameterized beta distribution (or its generalizations) as an integral of the model. However, it is well known that the widely popular mean regression model could be inadequate if the probability distribution of the observed responses do not follow a symmetric or multimodal distribution (Morales, Lachos, Cabral, & Cepero, 2017). Quantile regression, introduced by Koenker and Bassett (1978), is a very useful model for data analysis, and is an alternative approach to investigate the relationship between a response variable and covariates, in this context, that is, quantile regression is useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile. There are advantages to using quantile regression, such as the robustness to outliers, and can be more intuitive than the mean, especially for skewed distributions. According to Bayes, Bazán, and De Castro (2017), the main advantage of the quantile regression is its flexibility for modeling data with heterogeneous conditional distributions. Furthermore, in contrast to the mean regression model, quantile regression can provide an overall assessment of the covariate effects at different quantiles.

Noufaily and Jones (2013) explored a parametric approach to quantile regression, where the positive response variable, whose conditional distribution is modeled by the generalized gamma distribution. Bayes et al. (2017) studied a new quantile parametric mixed regression model for bounded response variables considering the Kumaraswamy distribution. Nascimento and Bourguignon (2020) defined and studied the quantile regression model in which the response variable is a generalized extreme value distribution. Sánchez, Leiva, Galea, and Saulo (2020) considered the Birnbaum–Saunders quantile regression model. Lemonte and Moreno-Arenas (2020) introduced a novel parametric quantile regression model for limited range response variables based on Johnson-t distribution. Furthermore, literature does exist on censored quantile regression models (Buchinsky & Hahn, 1998; Powell, 1986). Santos and Bolfarine (2015) proposed the use of Bayesian quantile regression for the analysis of proportion data present in a zero-or-one inflation using a two-part model approach. However, in the proposed model, the asymmetric Laplace distribution is assumed in the likelihood calculation. In this case, the approach based on a pseudo-likelihood through an asymmetric Laplace distribution (Yang, Wang, & He, 2016), and the conditional distribution of the response variable is unknown. Despite this, to the best of our knowledge, a specific parametric quantile regression model to describe data observed on the intervals $[0, 1)$ or $(0, 1]$ at different levels (quantiles) has never been considered in the literature.

Unfortunately, the cumulative distribution function (c.d.f.) of the beta distribution does not have an invertible closed form, which hinders its utilization with quantile regression purposes. In contrast to the beta distribution, the unit-Weibull (UW) distribution (Mazucheli, Menezes, Fernandes, Oliveira, & Ghitany, 2020) has a closed-form expression for the quantile function. The UW distribution, with support on the unit interval, was proposed by Mazucheli, Menezes, and Ghitany (2018). Recently, Mazucheli et al. (2020) proposed a parametric approach of quantile regression for limited range response variables. For this, Mazucheli et al. (2020) used a simple parameterization of UW distribution that is indexed by quantile and shape parameters.

Based on the above discussion, the main aim of this paper is to propose a parametric quantile regression model that is tailored for situations where the response variable is measured continuously on the intervals [0, 1) or (0, 1] based on reparameterized UW distribution (Mazucheli et al., 2020). In particular, the proposed model assumes that the response variable has a mixed continuous-discrete distribution with probability mass at zero or one. The reparameterized UW distribution (Mazucheli et al., 2020) is used to describe the continuous component of the model, since its density has a wide range of different shapes depending on the values of the two parameters that index the distribution. The quantile regression quantifies the association of the explanatory variables with a quantile of a dependent variable where the response variable is measured continuously on the intervals [0, 1) or (0, 1]. In fact, quantile regression is capable of providing a more complete statistical analysis of the stochastic relationships among random variables than the usual mean regression model. All the models cited above are not suitable for capturing this. Furthermore, the mixture parameter is modeled as functions of regression parameters. In the applications (this analysis is discussed in Section 5), note that the both data sets are asymmetry. In these cases, the mean is pulled in the direction of the tail, making it a less representative measure of central tendency. Thus, we will use of the proposed quantile model for fitting these two data sets.

This paper is organized as follows. Section 2 presents formulation of the zero-or-one inflated UW quantile regression model. In Section 3, the estimation method for the model parameters and diagnostic measures are discussed. Monte Carlo experiments regarding the parameter estimates and the empirical distribution of the residuals are presented with a discussion of the results in Section 4. In Section 5, we discuss an application to real data that demonstrates the usefulness of the proposed quantile regression model. Finally, in Section 6, we mention some concluding remarks.

2 | ZERO-OR-ONE INFLATED UNIT-WEIBULL QUANTILE REGRESSION MODELS

In a recent paper, Mazucheli et al. (2020) proposed a new parametric approach of quantile regression for limited range response variables. The regression model is based on the UW distribution, which was obtained by Mazucheli et al. (2018) as a transformation of a random variable with Weibull distribution. The probability density function (p.d.f.) and c.d.f. of the UW distribution re-parametrized in terms of the τ-th quantile are given, respectively, by

$$f(y \mid \mu, \phi, \tau) = \frac{\phi}{y}\left(\frac{\log \tau}{\log \mu}\right)\left(\frac{\log y}{\log \mu}\right)^{\phi-1} \tau^{\left(\frac{\log y}{\log \mu}\right)^{\phi}}, \qquad 0 < y < 1 \tag{1}$$

and

$$F(y \mid \mu, \phi, \tau) = \tau^{\left(\frac{\log y}{\log \mu}\right)^{\phi}}, \qquad 0 < y < 1, \tag{2}$$

where $\mu \in (0,1)$ is the τ-quantile of y, that is, the location parameter and $\phi > 0$ is the shape parameter.

The flexibility and advantage of the proposed quantile regression to model data in the unit interval were shown by the authors based on real applications and simulation studies (Mazucheli et al., 2020). However, the UW quantile regression model is not appropriate when the response variable contains observations at the extremes, either zeros or ones. In such situations, the underlying data generating process includes a discrete component that causes a given value (zero or one) to be observed with positive probability. Thus, a natural and well-known solution to combine the continuous and discrete data generating mechanisms into a more general law is to consider a mixture of two distributions.

In this paper, we assume that the continuous mechanism is described by the UW distribution, while the discrete component is a degenerate distribution in a known value c, either zero or one. Under this approach, the c.d.f. of the inflated UW distribution in c is given by

$$G(y \mid \nu, \mu, \phi, \tau) = \nu\, \mathbb{1}_c(y) + (1-\nu) F(y \mid \mu, \phi, \tau),$$

where $\mathbb{1}_A(y)$ is an indicator function that equals 1 when $y \in A$ and 0 when $y \notin A$; $\nu \in (0,1)$ is the mixture parameter and $F(y \mid \mu, \phi, \tau)$ is the c.d.f. of the UW distribution defined in (2). Notice that the random variable Y follows a UW distribution with probability $1-\nu$ and it follows a degenerate distribution in c with probability ν.

The corresponding p.d.f. of the inflated UW distribution in c is given by

$$g(y \mid \nu, \mu, \phi, \tau) = \begin{cases} \nu & \text{if} \quad y = c \\ (1-\nu)\, f(y \mid \mu, \phi, \tau) & \text{if} \quad y \in (0,1), \end{cases} \tag{3}$$

where $f(y \mid \mu, \phi, \tau)$ is the density of the UW distribution defined in (1). If $c = 0$, the density (3) is called zero-inflated UW distribution, and if $c = 1$ the density (3) is called one-inflated UW distribution. It should be mentioned that this approach for construction of inflated parametric distributions limited in the unit interval was considered in Ospina and Ferrari (2008) and Cribari-Neto and Santos (2019).

Now we can formulate a general class of zero-or-one inflated UW quantile regression model. Let $y_1, \ldots, y_n$ be an independent random variable such that each y_i, for $i = 1, \ldots, n$ has p.d.f. defined in (3) for a fixed (known) probability $\tau \in (0,1)$ associated with quantile of interest. We assume that the parameters μ_i and ν_i satisfy the following functional relations:

$$h_1(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}(\tau) \quad \text{and} \quad h_2(\nu_i) = \mathbf{w}_i^\top \boldsymbol{\alpha}, \qquad i = 1, \ldots, n, \tag{4}$$

where $\boldsymbol{\beta}(\tau) = (\beta_0(\tau), \ldots, \beta_{p-1}(\tau))^\top$ and $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_{q-1})^\top$ are vectors of unknown regression coefficients, which are assumed to be functionally independent, such that $\boldsymbol{\beta}(\tau) \in \mathbb{R}^p$ and $\boldsymbol{\alpha} \in \mathbb{R}^q$ with $p + q < n$. Also, $\mathbf{x}_i^\top = (1, x_{i1}, \ldots, x_{i(p-1)})$ and $\mathbf{w}_i^\top = (1, w_{i1}, \ldots, w_{i(q-1)})$ are observations on p and q known covariates, respectively. Moreover, we assume that the link functions $h_1(\cdot) : (0,1) \to \mathbb{R}$ and $h_2(\cdot) : (0,1) \to \mathbb{R}$ are strictly monotonic and twice differentiable. The main link functions for μ and ν are:

(i) logit: $g(\mu_i) = \log(\mu_i/(1-\mu_i))$;
(ii) probit: $g(\mu_i) = \Phi^{-1}(\mu_i)$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function;
(iii) complementary log-log: $g(\mu_i) = \log[-\log(1-\mu_i)]$.

3 | INFERENCE AND DIAGNOSTICS

This section is devoted to discuss inference based on the maximum likelihood method for the parameters of zero-or-one inflated UW quantile regression models. Finally, we proposed a residual analysis to detect departures from the assumed distribution.

3.1 | Maximum likelihood estimation

For a given $\tau \in (0,1)$, let $\theta = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}(\tau)^\top, \phi)^\top$ be the vector of unknown parameters to be estimated using the maximum likelihood method. Following the same idea presented in Ospina and Ferrari (2012) the likelihood function based on a sample of n independent observations of UW quantile regression model inflated at point c is given by

$$L(\theta \mid \boldsymbol{y}) = \prod_{i=1}^{n} g(y_i \mid \nu_i, \mu_i, \phi, \tau) = L_1(\boldsymbol{\alpha})\, L_2(\boldsymbol{\beta}(\tau), \phi), \tag{5}$$

where $g(\cdot \mid \nu_i, \mu_i, \phi, \tau)$ is the p.d.f. defined in (3),

$$L_1(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \nu_i^{\mathbb{1}_c(y_i)} (1-\nu_i)^{1-\mathbb{1}_c(y_i)}$$

and

$$L_2(\boldsymbol{\beta}(\tau), \phi) = \prod_{i\,:\,y_i \in (0,1)} f(y_i \mid \mu_i, \phi)$$

with $f(\cdot \mid \mu_i, \phi)$ being the p.d.f. of UW distribution defined in (1). The parameters $\mu_i = h_1^{-1}(\mathbf{x}_i^\top \boldsymbol{\beta}(\tau))$ and $\nu_i = h_2^{-1}(\mathbf{w}_i^\top \boldsymbol{\alpha})$ as defined in (4), are functions of the regression coefficients $\boldsymbol{\beta}(\tau)$ and $\boldsymbol{\alpha}$, respectively.

Note that the likelihood function factorizes in two terms: one depending only the $\boldsymbol{\alpha}$ (discrete component) and another one depending only on $(\boldsymbol{\beta}(\tau), \phi)^\top$ (continuous component). Thus, according Pace and Salvan (1997) the parameter vectors are separable, which implies that the maximum likelihood inference for $(\boldsymbol{\beta}(\tau), \phi)^\top$ can be performed separately as if $\boldsymbol{\alpha}$ were known and vice versa.

The corresponding log-likelihood function is given by

$$\ell(\theta \mid \boldsymbol{y}) = \ell_1(\boldsymbol{\alpha}) + \ell_2(\boldsymbol{\beta}(\tau), \phi),$$

where

$$\ell_1(\boldsymbol{\alpha}) = \sum_{i=1}^{n} [\mathbb{1}_c(y_i)\, \log(\nu_i) + (1-\mathbb{1}_c(y_i))\, \log(1-\nu_i)] \tag{6}$$

and

$$\ell_2(\boldsymbol{\beta}(\tau), \phi) = \sum_{i=1}^{n} \log\left(\frac{\phi}{y_i}\right) + \sum_{i=1}^{n} \log\left(\frac{\log\tau}{\log\mu_i}\right) + (\phi - 1) \sum_{i=1}^{n} \log\left(\frac{\log y_i}{\log \mu_i}\right) + \log(\tau) \sum_{i=1}^{n} \left(\frac{\log y_i}{\log \mu_i}\right)^{\beta}. \tag{7}$$

It is not possible to derive closed-form expression for the maximum likelihood estimator (MLE) of $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}(\tau)^\top, \phi)^\top$. Therefore, due the separability of $\boldsymbol{\alpha}$ and $(\boldsymbol{\beta}(\tau), \phi)^\top$, the MLE of $\boldsymbol{\alpha}$ are obtained by maximizing the log-likelihood function of discrete component $\ell_1(\boldsymbol{\alpha})$ defined in (6), while the MLEs of $(\boldsymbol{\beta}(\tau), \phi)^\top$ are obtained maximizing the log-likelihood function of the continuous component $\ell_2(\boldsymbol{\beta}(\tau), \phi)$ defined in (7). Numerical algorithms such as Newton–Raphson and quasi-Newton can be used.

The main advantage of the proposed model under the alternatives in literature is the ability to provide reasonable estimates for the quantiles of response variable. For instance, if the response variable is one-inflated, an estimate of the τth quantile would be

$$\widehat{y}_{\tau\text{th}} = \begin{cases} F\left(\tau(1-\widehat{\nu}_i)^{-1} \mid \widehat{\mu}_i, \widehat{\phi}, \tau\right) & \text{if} \quad 1-\widehat{\nu}_i \geq \tau \\ 1 & \text{if} \quad 1-\widehat{\nu}_i < \tau, \end{cases}$$

where $\widehat{\mu}_i$, $\widehat{\phi}$, and $\widehat{\nu}_i$ are the MLEs of μ_i, ϕ, and ν_i, respectively, obtained in the UW one-inflated τ-quantile regression model. This interesting feature allows a complete view of the conditional distribution of response variable in real applications rather than predicting just the mean (see Section 5.2).

Under suitable regularity conditions (see Cox & Hinkley, 1974, p. 107), the asymptotic distribution of the MLE $\widehat{\boldsymbol{\theta}}$ is a multivariate Normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}(\widehat{\boldsymbol{\theta}})$, which can be consistently estimated by the inverse of the observed Fisher information matrix, given by

$$\widehat{\boldsymbol{\Sigma}}\left(\widehat{\boldsymbol{\theta}}\right) = \left[-\frac{\partial\, \ell(\boldsymbol{\theta} \mid \boldsymbol{y})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}^\top}\right]^{-1}$$

evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$.

Let θ_r, $r = 1, 2, \ldots, p+q+1$, be the r-th component of $\boldsymbol{\theta}$. The asymptotic $100(1-\delta)\%$ confidence interval for θ_r is given by

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

where $z_{\delta/2}$ is the $\delta/2$ upper quantile of the standard normal distribution and $\text{se}(\widehat{\theta}_r)$ is the asymptotic standard error of $\widehat{\theta}_r$. Note that $\text{se}(\widehat{\theta}_r)$ is the square root of the r-th diagonal element of the matrix $\widehat{\boldsymbol{\Sigma}}(\widehat{\boldsymbol{\theta}})$.

In the next section, we evaluated the finite-sample behavior of the MLEs $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ with respect to bias, consistency, and probability coverage of the asymptotic interval.

3.2 | Randomized quantile (RQ) residuals

To evaluate and detect departures from the underlying inflated UW quantile regression model assumption, we propose to use the RQ residuals introduced by Dunn and Smyth (1996). It is defined as

$$\widehat{r}_i = \Phi^{-1}(u_i), \qquad i = 1, \ldots, n,$$

where $\Phi(\cdot)$ is the standard normal distribution function and u_i is a uniform random variable $[a_i, b_i]$, where the range depends on the inflation of the model. In the zero-inflated UW quantile regression, u_i is a uniform random variable on $(0, \widehat{\nu}_i]$ if $y_i = 0$ and $u_i = G(y_i \mid \widehat{\nu}_i, \widehat{\mu}_i, \widehat{\phi}, \tau)$ if $y_i \in (0,1)$. In the one-inflated UW quantile regression, u_i is a uniform random variable on $[\widehat{\nu}_i, 1)$ if $y_i = 1$ and $u_i = G(y_i \mid \widehat{\nu}_i, \widehat{\mu}_i, \widehat{\phi}, \tau)$ if $y_i \in (0,1)$. Apart from the variability due to the estimates of the parameters these residuals have standard normal distribution if the proposed model is correctly specified (Dunn & Smyth, 1996).

From the residuals, we can examine several graphics to detect departures from the model assumptions. For instance, the plot of the residuals versus the index of observations can be useful to detect patterns related of time. Link function misspecification can be revealed if a trend appears in the plot of residuals against predictors. The half-normal plot with a

TABLE 1 Average Monte Carlo proportion of zeros according to sample size and scenario

n	Scenario I	Scenario II
30	0.3084	0.4535
50	0.3114	0.4949
100	0.3117	0.5035
150	0.3106	0.4825
300	0.3109	0.4966

simulated envelope proposed by Atkinson (1981) is also a helpful diagnostic tool. Simulation studies are presented in the next section concerning the empirical distribution of the proposed residuals.

4 | SIMULATION STUDIES

In this section, we conducted simulation studies (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 RQ residuals proposed. Additionally, we evaluate the parameter estimates and residuals by inspecting five quantiles levels, namely, the first decile ($\tau = 0.10$), the first quartile ($\tau = 0.25$), the median ($\tau = 0.50$), the third quartile ($\tau = 0.75$), and the last decile ($\tau = 0.90$).

As discussed in (5) the maximum likelihood estimation can be performed separately, hence, without loss of generality, the zero-inflated UW quantile regression was considered.

The following scenarios are considered:

(a) Scenario I:

$$h_1(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0(\tau) + \beta_1(\tau)\, x_{1i},$$

$$h_2(\nu_i) = \log\left(\frac{\nu_i}{1-\nu_i}\right) = \alpha_0 + \alpha_1\, w_{1i}, \quad i = 1, \dots, n,$$

where the true values of the parameters were taken as $\beta_0(\tau) = 1.0, \beta_1(\tau) = 2.0, \alpha_0 = -1.0$, and $\alpha_1 = 0.4$, and the true value of the shape parameter is taken as $\phi = 2.0$. The covariate values of x_{1i} were generated from the standard normal distribution, while the values of w_{1i} were drawn from the standard uniform distribution.

(b) Scenario II:

$$h_1(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0(\tau) + \beta_1(\tau)\, x_{1i} + \beta_2(\tau)\, x_{2i} + \beta_3(\tau)\, x_{3i},$$

$$h_2(\nu_i) = \log\left(\frac{\nu_i}{1-\nu_i}\right) = \alpha_0 + \alpha_1\, w_{1i} + \alpha_2\, w_{2i}, \quad i = 1, \dots, n,$$

where the true values of the parameters were taken as $\beta_0(\tau) = -2.0, \beta_1(\tau) = 1.0, \beta_1(\tau) = 2.0, \beta_3(\tau) = 2.0, \alpha_0 = 2.0, \alpha_1 = -4.0, \alpha_2 = 1.0$, and $\phi = 2.0$. The covariate values of x_{1i} and x_{2i} were generated from two independent standard normal distribution, x_{3i} were generated from Bernoulli distribution with probability of success equal 0.5, while the values of w_{1i} and w_{2i} were drawn from the standard uniform and standard Normal distributions, respectively.

In all scenarios, the sample size were $n = 30, 50, 100, 150, 300$, and the covariate values were remained constant throughout the simulations. It should be mentioned that the proportion of zeros varies according to each Monte Carlo simulation and sample size. Table 1 shows the average Monte Carlo proportion of zeros according to each sample size and scenario.

All simulations were conducted in SAS using the quasi-Newton algorithm available in the `NLMIXED` procedure (SAS, 2010) to obtain the maximum likelihood estimates.

TABLE 2 Estimated RB, relative RMSE, and coverage probability for α_0 and α_1

n	RB		RMSE		CP$_{95\%}$	
	α_0	α_1	α_0	α_1	α_0	α_1
30	0.1685	0.4431	0.9997	16.7647	0.9665	0.9610
50	0.0866	0.2406	0.4333	8.6916	0.9550	0.9545
100	0.0442	0.1751	0.2173	4.0396	0.9575	0.9505
150	0.0243	0.1012	0.1333	2.3852	0.9600	0.9561
300	0.0141	0.0522	0.0657	1.1561	0.9570	0.9544

TABLE 3 Estimated RB, relative RMSE, and coverage probability for α_0, α_1, and α_2

n	RB			RMSE			CP$_{95\%}$		
	α_0	α_1	α_2	α_0	α_1	α_2	α_0	α_1	α_2
30	4.4392	3.9265	5.2550	543.2916	454.9435	641.9376	0.9686	0.9738	0.9648
50	0.7536	0.6948	0.6943	92.3060	85.0467	70.2188	0.9702	0.9686	0.9772
100	0.0527	0.0507	0.0691	0.1400	0.1135	0.0607	0.9680	0.9550	0.9640
150	0.0294	0.0266	0.0343	0.0801	0.0580	0.0360	0.9550	0.9620	0.9600
300	0.0198	0.0217	0.0163	0.0373	0.0301	0.0173	0.9450	0.9510	0.9420

For each combination of n, τ, and scenario, the Monte Carlo experiment was repeated 10,000 times.

4.1 | Parameter estimation

The aim of the simulation study presented in this subsection is to examine the small sample properties of the MLE previously described. For such evaluation, the estimated relative bias (RB), the estimated relative root-mean squared error (RMSE), and the coverage probability of 95% confidence interval (CP$_{95\%}$) were computed.

For a given simulated data set, the estimates of $\boldsymbol{\alpha}$ do not depend on τ, hence the above criteria were aggregate, on average, across all Monte Carlo simulations and all values of τ, as shown in Tables 2 and 3.

The results of the simulation experiments for Scenario I are presented in Figure 1 and Table 2. From these figures and tables we can observe the following:

(i) the intercept parameter presents higher RB and relative RMSE than $\beta_1(\tau)$ for all quantiles, especially in the right tails of the distribution ($\tau = 0.75$ and 0.90);
(ii) the $\beta_1(\tau)$ parameter presents high RB for $\tau = 0.75$ and 0.90;
(iii) the shape parameter, ϕ, has high RB in small sample sizes, but it is not affected by the quantiles;
(iv) the coverage probability of the 95% confidence intervals of parameter $\beta_0(\tau)$ are far from to the nominal level, especially for small sample size ($n = 30$ and 50) and tail quantiles ($\tau = 0.10$ and 0.90);
(v) for small sample size, parameters α_0 and α_1 present higher RB and relative RMSE. As expected by the asymptotically theory the estimators are unbiased and consistent, as sample size increases;
(vi) the coverage probability for α_0 and α_1 are higher than the nominal level for small sample size.

The results of the simulation experiments for Scenario II are presented in Figure 2 and Table 3. The same comments for Scenario I hold for Scenario II. However, it should interesting to mention that

(i) parameters $\beta_1(\tau)$ and $\beta_2(\tau)$ show lesser RB and relative RMSE than parameters $\beta_0(\tau)$ and $\beta_3(\tau)$, where the latter represent effect of dummy variable.
(ii) From these figures and tables, It can be seen that as proportion of zero increases the RB and the relative RMSE become bigger (for $\beta(\tau)$) for all quantiles, that is, the proportion of zero can influence the performance of the model.

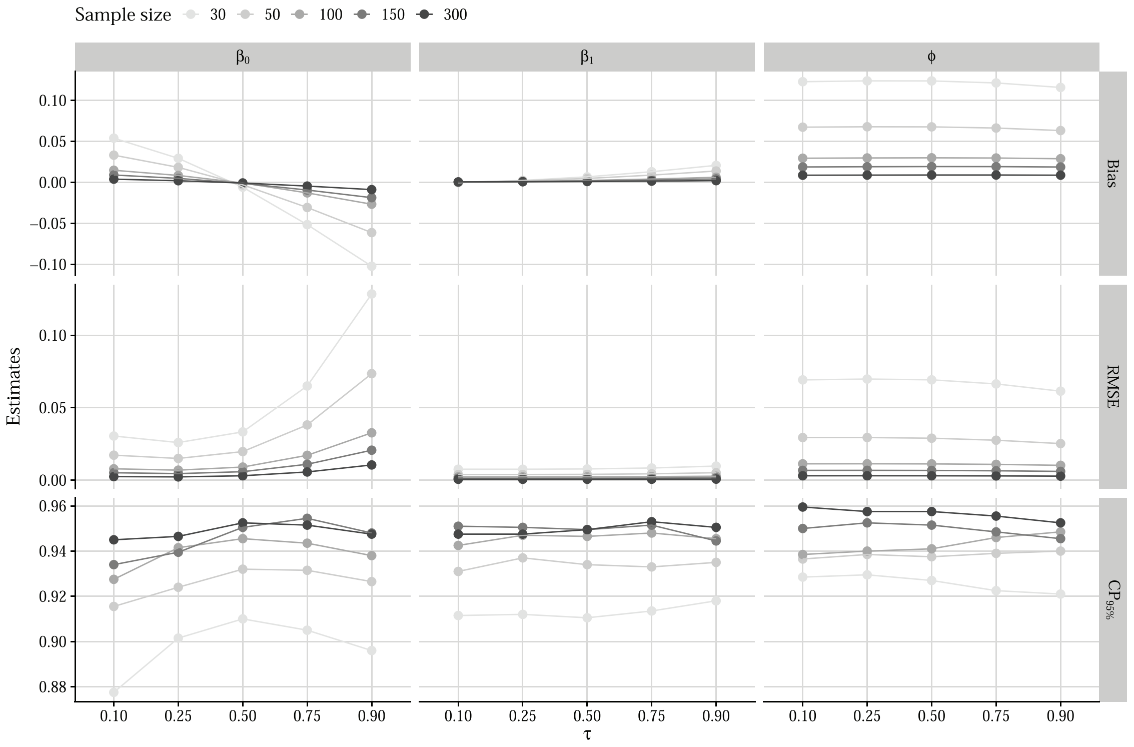

FIGURE 1 Estimated RB, relative RMSE, and coverage probability for β_0, β_1, and ϕ

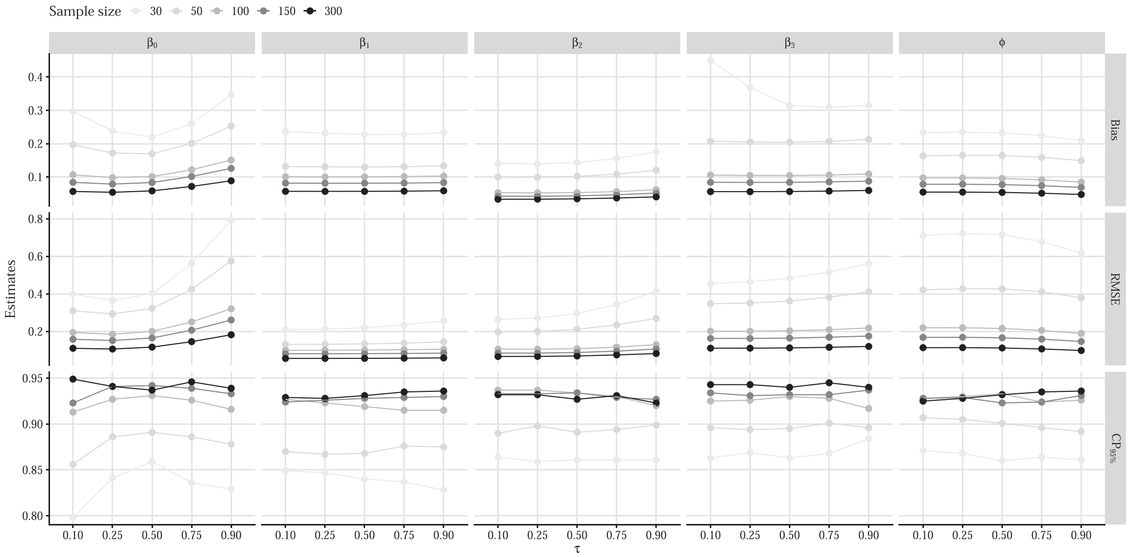

FIGURE 2 Estimated RB, relative RMSE, and coverage probability for β_0, β_1, β_2, β_3, and ϕ

4.2 | Residuals

In this subsection, we consider Monte Carlo experiments regarding the finite-sample behavior of the RQ residuals. The evaluation of the RQ residuals was based on the normal probability plots of the mean order statistics and descriptive measures.

Tables 4 and 5 reported the mean, standard deviation (StdDev), skewness (Skew), and kurtosis (Kurt) of RQ residuals. The descriptive measures of the RQ residuals are close to the theoretical true values of the standard normal distribution for all scenarios, that is, the residuals have approximately zero mean and unit standard deviation, have skewness close to zero, and the kurtosis is near three.

TABLE 4 Descriptive measures of the randomized quantile residuals—Scenario I

n	τ	Mean	StdDev	Skew	Kurt
30	0.10	0.00208	1.0079	0.0113	3.2749
	0.25	0.00362	1.0068	0.0153	3.2757
	0.50	0.00289	1.0077	0.0181	3.2750
	0.75	0.00268	1.0076	0.0222	3.2730
	0.90	0.00171	1.0079	0.0210	3.2888
50	0.10	0.00157	1.0045	0.0075	3.1492
	0.25	0.00199	1.0045	0.0084	3.1531
	0.50	0.00153	1.0054	0.0070	3.1619
	0.75	0.00173	1.0046	0.0117	3.1618
	0.90	0.00113	1.0040	0.0152	3.1636
100	0.10	0.00107	1.0016	0.0023	3.0755
	0.25	0.00118	1.0016	0.0042	3.0704
	0.50	0.00068	1.0026	0.0014	3.0742
	0.75	0.00069	1.0024	0.0027	3.0761
	0.90	0.00079	1.0016	0.0052	3.0797
150	0.10	0.00095	1.0005	0.0021	3.0429
	0.25	0.00073	1.0011	0.0000	3.0475
	0.50	0.00052	1.0012	0.0023	3.0400
	0.75	0.00034	1.0014	0.0019	3.0435
	0.90	0.00033	1.0013	0.0014	3.0499
300	0.10	0.00028	1.0006	0.0001	3.0241
	0.25	0.00042	1.0006	0.0008	3.0229
	0.50	0.00036	1.0007	0.0012	3.0218
	0.75	0.00042	1.0006	0.0018	3.0216
	0.90	0.00055	1.0005	0.0008	3.0282

Normal probability plots of the mean order statistics of RQ residuals are presented in Figures 3 and 4. It is observed that the residuals have good agreement with the standard normal distribution. Thus, we recommend the use of the RQ residual to check the goodness of fit of the proposed quantile regression model.

Figure 5 provides an example of a residual plot for simulated data sets considering various values of τ and n.

5 | MOTIVATING EXAMPLES

To evaluate the applicability of the proposed model, two real data sets with inflation in zero and one are considered. The two data sets have been taken from the recent book of Korosteleva (2019). As mentioned by the author the source of the data sets came from consulting projects, which she had involved.

Figure 6 shows the empirical cumulative distribution of the two response variables, the proportion of biked to campus data (left) and proportion of survived trees (right), respectively. It is observed that there is a considerable inflation of zeros and ones for the proportion of biked to campus data (left) and proportion of survived trees (right), respectively.

To identify the best subset of regressors, we employed the exhaustive search procedure that enumerates all possible subsets of regressions and evaluates the models in terms of criterion functions (AIC and BIC), residual analysis, and predictors of significance. Since both applications have four covariates and the zero-or-one UW quantile regression model has two components (discrete and continuous), there are $(2^4)^2 - 1 = 225$ subsets of regressors. Instead evaluate the models for several quantiles the median was chosen. All computations were performed in R software and the time elapsed to fit all models were 2.5 s per applications. An R package can be installed from https://github.com/AndrMenezes/uwquantreg to fit the models.

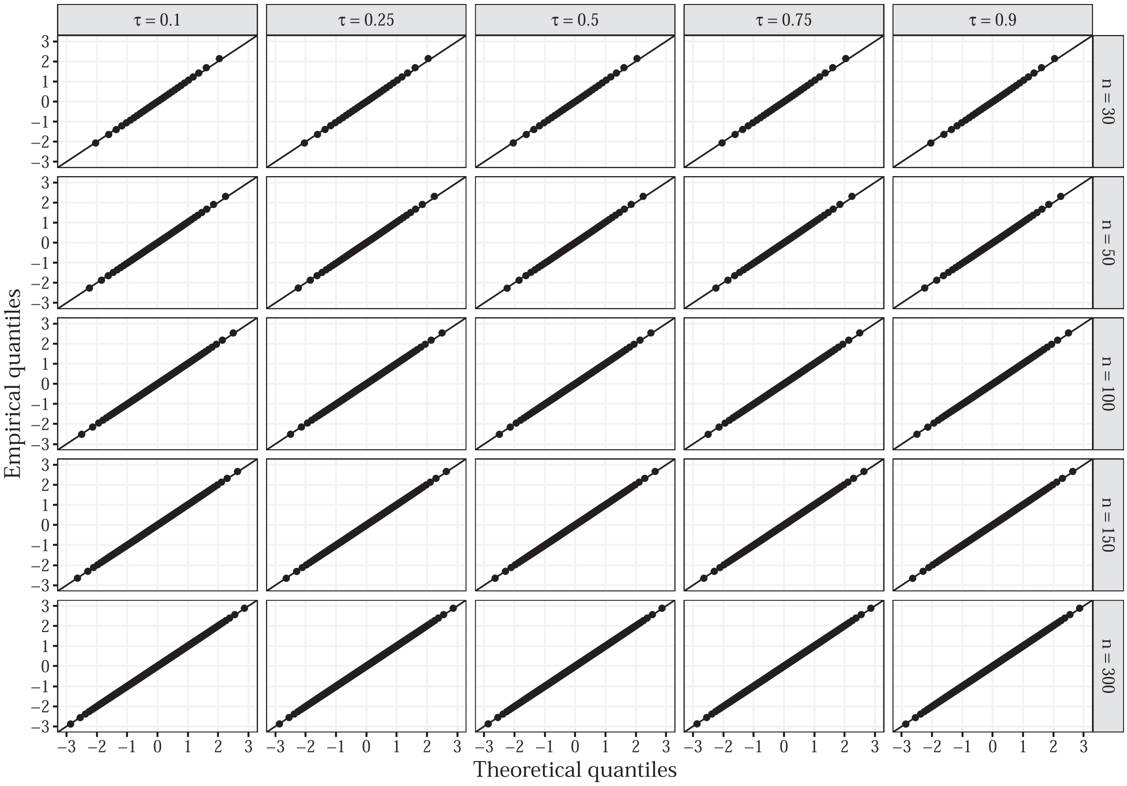

FIGURE 3 Normal probability plots of the mean order statistics for various values of τ and n—Scenario I

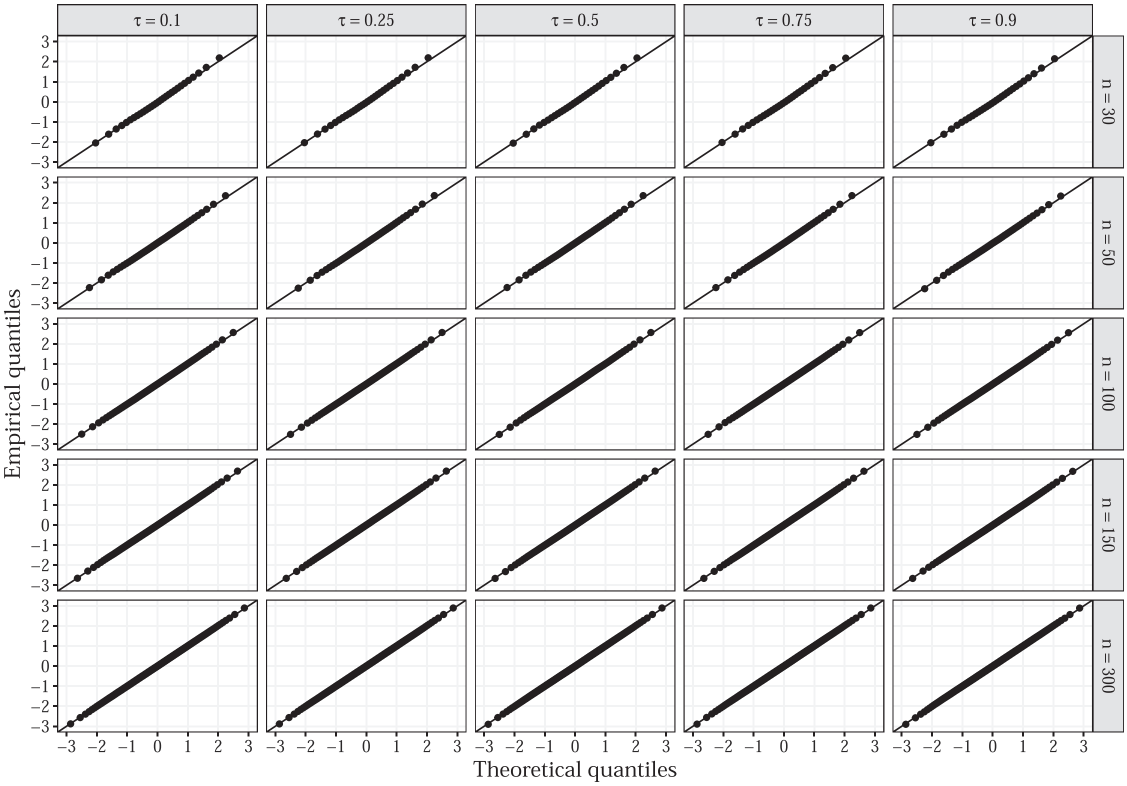

FIGURE 4 Normal probability plots of the mean order statistics for various values of τ and n—Scenario II

TABLE 5 Descriptive measures of the randomized quantile residuals—Scenario II

n	τ	Mean	StdDev	Skew	Kurt
30	0.10	0.00410	1.0128	0.0835	3.3033
	0.25	0.00411	1.0111	0.1004	3.2969
	0.50	0.00012	1.0109	0.0872	3.3512
	0.75	−0.00527	1.0067	0.1102	3.3228
	0.90	−0.01099	1.0022	0.1019	3.3563
50	0.10	0.00694	1.0038	0.0457	3.1769
	0.25	0.00022	1.0099	0.0458	3.1714
	0.50	0.00136	1.0054	0.0557	3.1613
	0.75	0.00066	1.0047	0.0617	3.1735
	0.90	−0.00604	1.0016	0.0417	3.2079
100	0.10	0.00092	1.0034	0.0250	3.0733
	0.25	−0.00027	1.0053	0.0190	3.0777
	0.50	0.00093	1.0028	0.0210	3.0911
	0.75	0.00071	1.0009	0.0213	3.0971
	0.90	−0.00584	1.0014	0.0217	3.1021
150	0.10	0.00202	1.0000	0.0170	3.0590
	0.25	0.00030	1.0043	0.0080	3.0458
	0.50	−0.00196	1.0037	0.0131	3.0512
	0.75	−0.00221	1.0031	0.0093	3.0526
	0.90	−0.00272	0.9982	0.0152	3.0666
300	0.10	0.00004	1.0023	−0.0005	3.0221
	0.25	−0.00103	1.0017	0.0037	3.0198
	0.50	−0.00204	1.0018	0.0030	3.0287
	0.75	−0.00033	1.0014	0.0032	3.0252
	0.90	−0.00193	1.0014	−0.0006	3.0381

TABLE 6 Descriptive statistics of response variable according to the covariate

Status	n	Mean	Median	Std.	$y = 0$	Parking	n	Mean	Median	Std.	$y = 0$
Faculty	23	0.36	0.33	0.38	0.43	6	20	0.40	0.28	0.36	0.30
Staff	12	0.24	0.17	0.25	0.33	9	19	0.34	0.25	0.35	0.37
Student	25	0.29	0.23	0.31	0.40	12	21	0.20	0.00	0.26	0.52
Gender	n	Mean	Median	Std.	$y = 0$	Distance	n	Mean	Median	Std.	$y = 0$
Female	30	0.32	0.23	0.32	0.33	[1; 2)	6	0.53	0.51	0.24	0.00
Male	30	0.30	0.20	0.34	0.47	[3; 5)	16	0.46	0.51	0.29	0.19
						[6; 7)	11	0.40	0.33	0.32	0.18
						[8; 15)	15	0.23	0.00	0.36	0.53
						[16; 60)	12	0.01	0.00	0.04	0.92

5.1 | Inflation of zeros: Transport in campus data

This data set is regarding the mode of transportation in a campus. According to Korosteleva (2019), a stratified sample of 60 respondents was drawn for the purpose of oversampling people who sometimes bike to campus. In this analysis, the interest lies in analyzing the association between the proportion of time a respondent biked to campus and a person's status (student/faculty/staff) and gender (F/M), duration of parking permit (6, 9, or 12 months) and distance to campus.

There are 24 (40%) respondents who never biked to campus. Table 6 presents some descriptive statistics according to the covariates. It is observed that the median of the response variable is less than the mean for all levels of covariates,

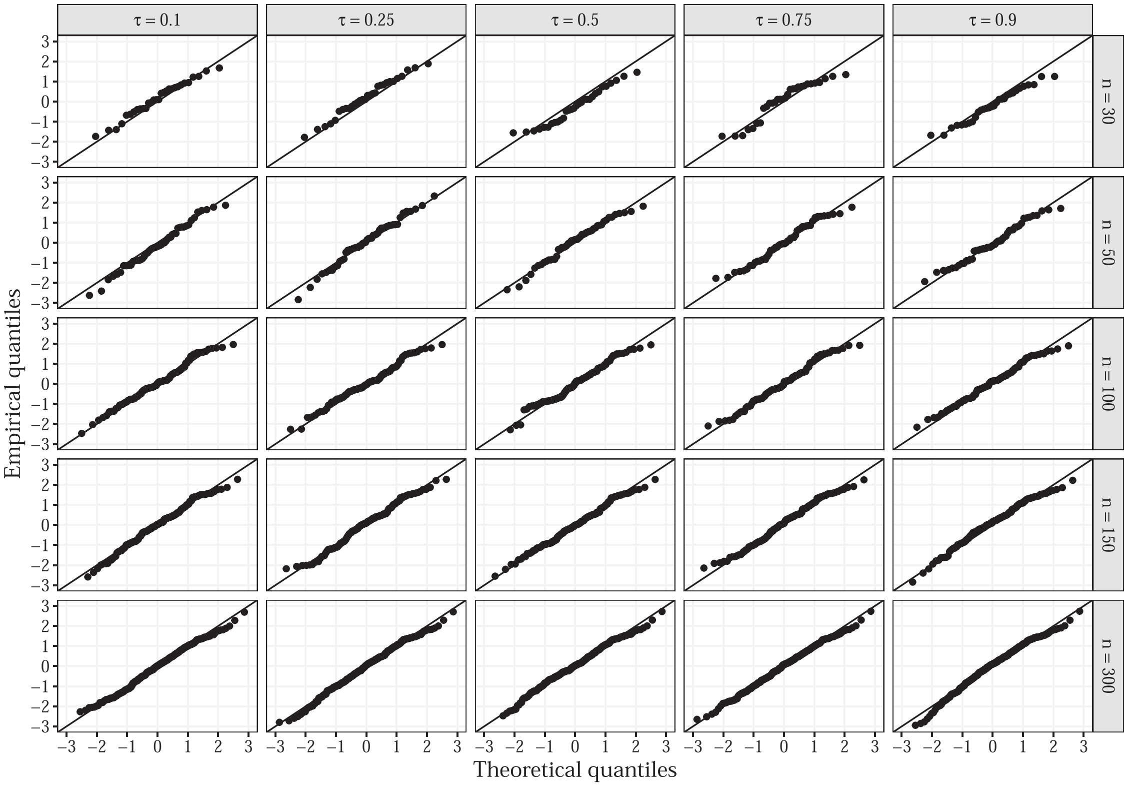

FIGURE 5 Examples of a residual plot for simulated data sets for various values of τ and n considering the model from Scenario II

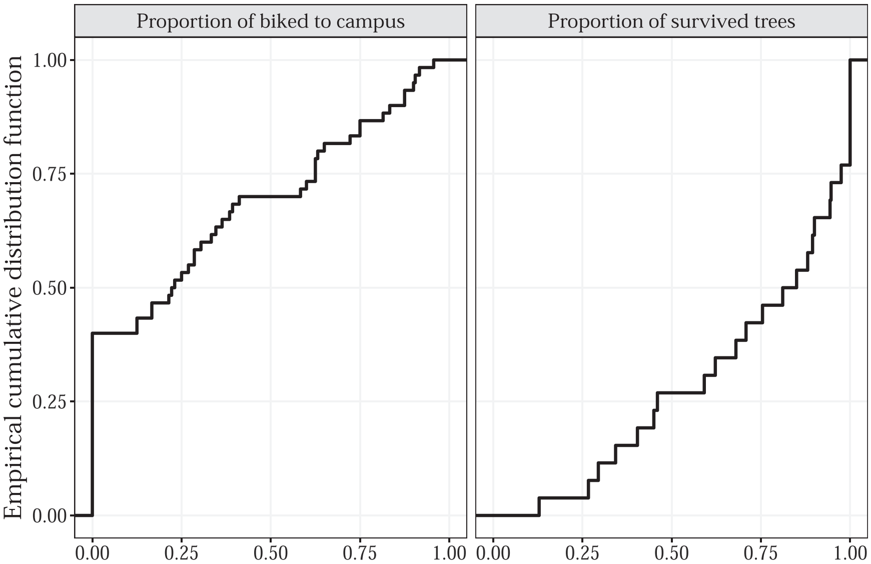

FIGURE 6 Empirical cumulative distribution function of the response variables

expect for distance between 3 and 5. Furthermore, it is observed that the proportion of zero ($y = 0$) is different across the levels of covariates.

The following model was selected:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1\,\text{male}_i + \beta_2\,\text{parking}_i + \beta_3\,\text{staff}_i + \beta_4\,\text{student}_i,$$

$$\text{logit}(\nu_i) = \alpha_0 + \alpha_1\,\text{male}_i + \alpha_2\,\text{distance}_i, \quad i = 1, \dots, 60,$$

where the status is transformed in dummy variables, staff and student, with the faculty level as reference.

TABLE 7 Parameter estimates, standard errors (S.E.), and *p*-value for zero-inflated UW median regression

Model	Parameter	Estimate	S.E.	*p*-value
Median	Intercept	3.1339	1.2357	0.0112
	Male	0.8592	0.3146	0.0063
	Parking	−0.2914	0.1281	0.0230
	Staff	−1.1511	0.4346	0.0081
	Student	−1.6803	0.5328	0.0016
	ϕ	1.5143	0.1982	—
Zero-inflated	Intercept	−4.4897	1.2454	0.0003
	Male	1.7728	0.8797	0.0439
	Distance	0.3396	0.0970	0.0005

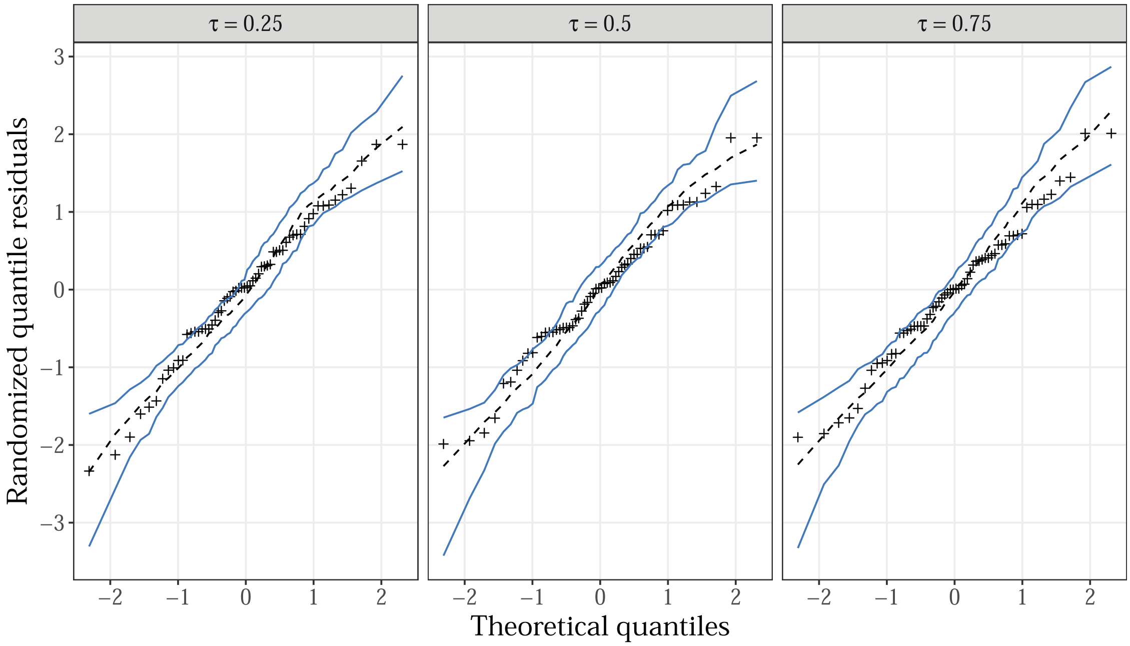

FIGURE 7 Randomized quantile residuals with simulated envelope for different τ

The parameter estimates, standard errors, and *p*-value for the zero-inflated UW median regression are shown in Table 7. It is noteworthy that as the distance to campus increases, the probability of people to bike to campus decreases. Also, men biked to campus lesser than women, in fact for a fixed distance the odds that men biked to campus decrease in $e^{\widehat{\alpha}_1} = 5.8871$. These results corroborate to the descriptive analysis presented in Table 6.

The results from median regression indicate that all covariates are statistically significant and the variables parking, staf,f and student have a negative effect on the response variable. This means, for example, that as the duration of parking increases, the proportion of people who biked to campus decreases, also student biked to campus lesser than faculty, on median.

To check the model assumption, the residual plots with a simulated envelope for different quantiles are presented in Figure 7. From these results, we can conclude that the proposed model provided a good fit for this data set.

Figure 8 displays the parameter estimates and their 95% confidence interval for the parameters of continuous part assuming different values for the quantiles. It is observed that all coefficients, except the intercept, became close to zero as the quantile level increases, indicating that these variables are more important to explain smaller quantiles. Although the coefficients increase, they decrease in magnitude.

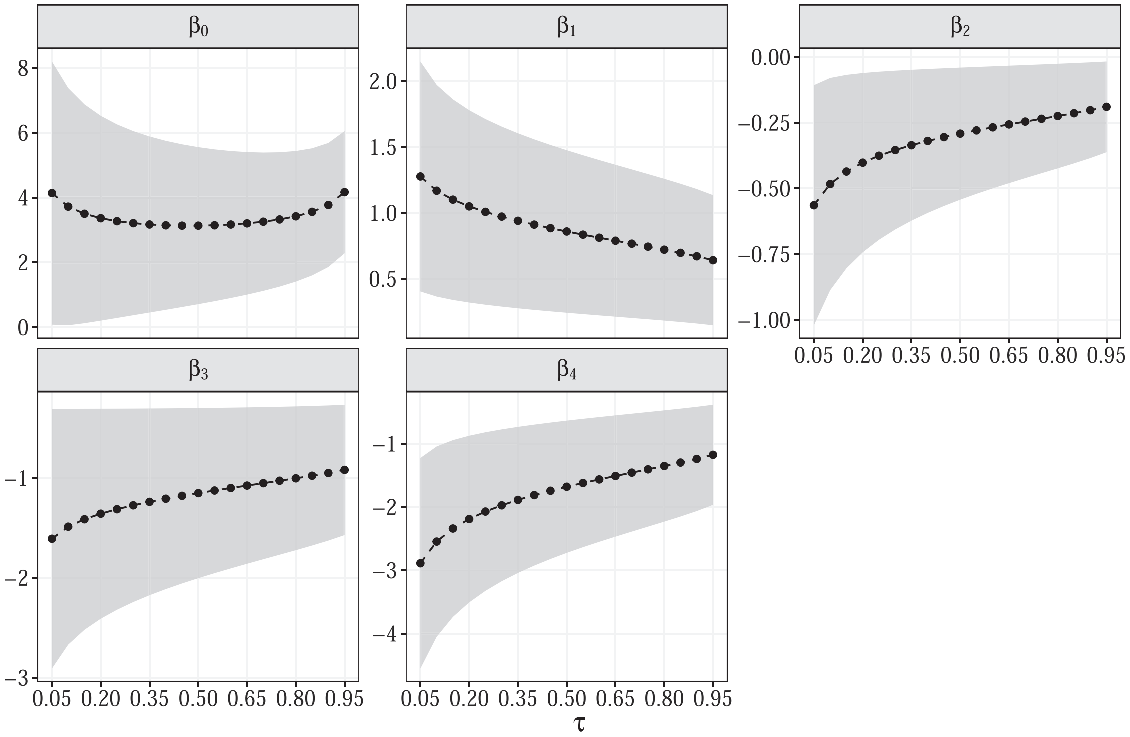

FIGURE 8 Parameter estimates and 95% confidence intervals for $\tau = 0.05, \dots, 0.95$ by 0.05

TABLE 8 Parameter estimates, standard errors (S.E.), and p-value for one-inflated UW median regression

Model	Parameter	Estimate	S.E.	*p*-value
Median	Intercept	3.0182	1.2015	0.0120
	Pest	0.4153	0.1095	0.0001
	Fert	1.0095	0.1772	0.0001
	Precip	−0.0906	0.0302	0.0027
	Wind	−0.2635	0.0722	0.0003
	ϕ	1.8683	0.3615	—
One-inflated	Intercept	3.3183	2.3526	0.1584
	Wind	−0.4671	0.2535	0.0654

5.2 | Inflation of ones: Mortality of young trees data

In this application we consider the data set described by Korosteleva (2019) related to a study conducted in 2 years on mortality of young trees planted in parks. The goal is to describe the association between the proportion of survived trees in 2 years and some climatological variables (average annual precipitation (in inches), and average annual wind speed (in miles per hour)) and soil variables (frequencies of pest control and soil fertilization). The number of parks investigated are 26, and in six (23%) all trees have survived.

The selected model is given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1\,\text{pest}_i + \beta_2\,\text{fert}_i + \beta_3\,\text{precip}_i + \beta_4\,\text{wind}_i$$

$$\text{logit}(\nu_i) = \alpha_0 + \alpha_1\,\text{wind}_i,$$

for $i = 1, \dots, 26$.

Table 8 gives the parameter estimates, standard errors, and p-values for the one-inflated UW median regression. For the probability of the proportion being equal to one, that is, all trees survived in 2 years we observed that the average annual

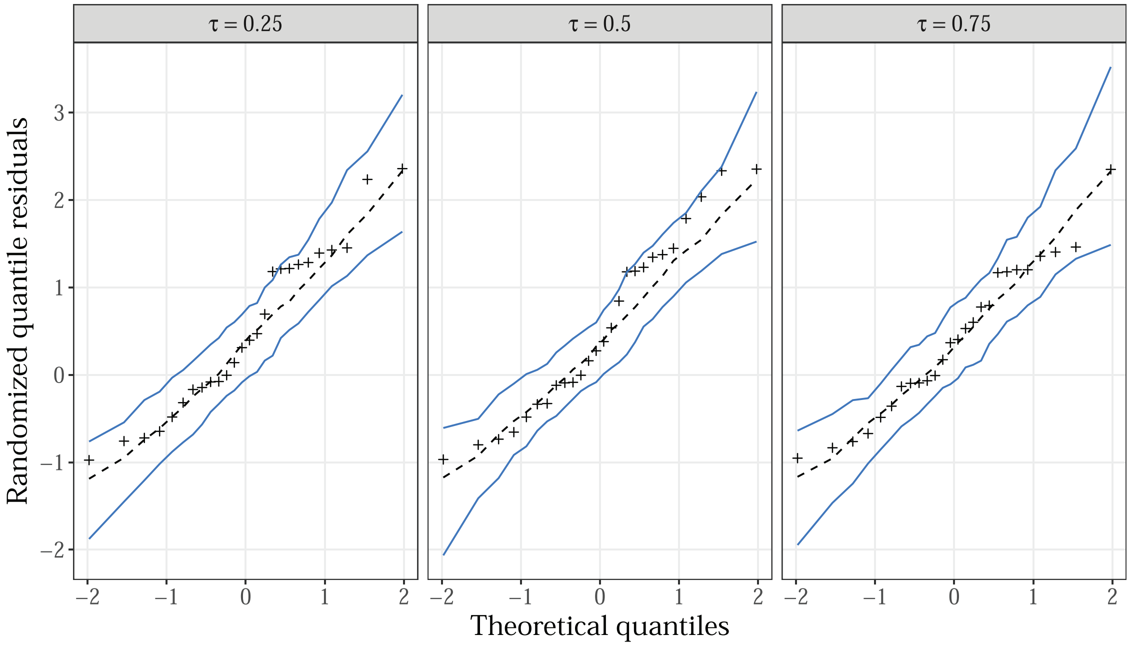

FIGURE 9 Randomized quantile residuals with simulated envelope for different τ

wind speed has a negative impact, which implies that as the average annual wind speed increases the probability of all trees survived in 2 years decreases.

For the continuous part, it is observed that all parameters are statistically significant. It is also interesting to observe that the estimated coefficient for precipitation (β_3) implies that higher rainfall is associated with a decrease in the median survival rate of young trees. The effect is analogous for the wind covariate. On the other hand, the estimated coefficients for pest control (β_1) and soil fertilization (β_2) indicate that places with higher frequencies of these variables increase the median survival rate of young trees.

To check the model assumption we presented in Figure 9 the RQ residuals with simulated envelope for different quantiles. The results indicate no departures about the model assumptions.

Figure 10 shows that the UW quantile regression is more informative than the usual conditional mean regression. A close inspection of the results reveals that the effects of average annual precipitation (β_3) and average annual wind speed (β_4) increase for higher quantiles, while, the effects of frequencies of pest control and soil fertilization decrease for higher quantiles.

Finally, a question of interest in this application is to predict the proportion of trees that would survive for 2 years in two specific areas, where neither pest control nor soil fertilization would be feasible. The goal is to decide between an area with lower precipitation (2 in) and stronger winds (12.5 mph), and an area with higher precipitation (25 in) and lower winds (6 mph). Thus, based on the fitted models (for various τ), we can estimate the quantiles and obtain a more informative view of the conditional distribution of the proportion of the survived trees. The results displayed in Figure 11 support the decision in favor of the area with higher precipitation and lower wind, since the estimated quantiles increase faster as τ increases.

6 | DISCUSSION AND CONCLUSION

Motivated by the presence of zeros or ones in proportion responses, we develop a parametric quantile regression model for double bounded response variables. Our model is built on the reparameterized UW distribution introduced by Mazucheli et al. (2020). We extend this model to the case when the proportion data present a considerable number of zeros or ones. In particular, the proposed model assumes that the response variable has a mixed continuous–discrete distribution with probability mass at zero or one, where the reparameterized UW distribution is used to describe the continuous component of the model. Inference is based on a frequentist approach, and the maximum likelihood inference is employed to estimate the model parameters. Its good performance, in terms of the bias and mean-squared error, has been evaluated by means of

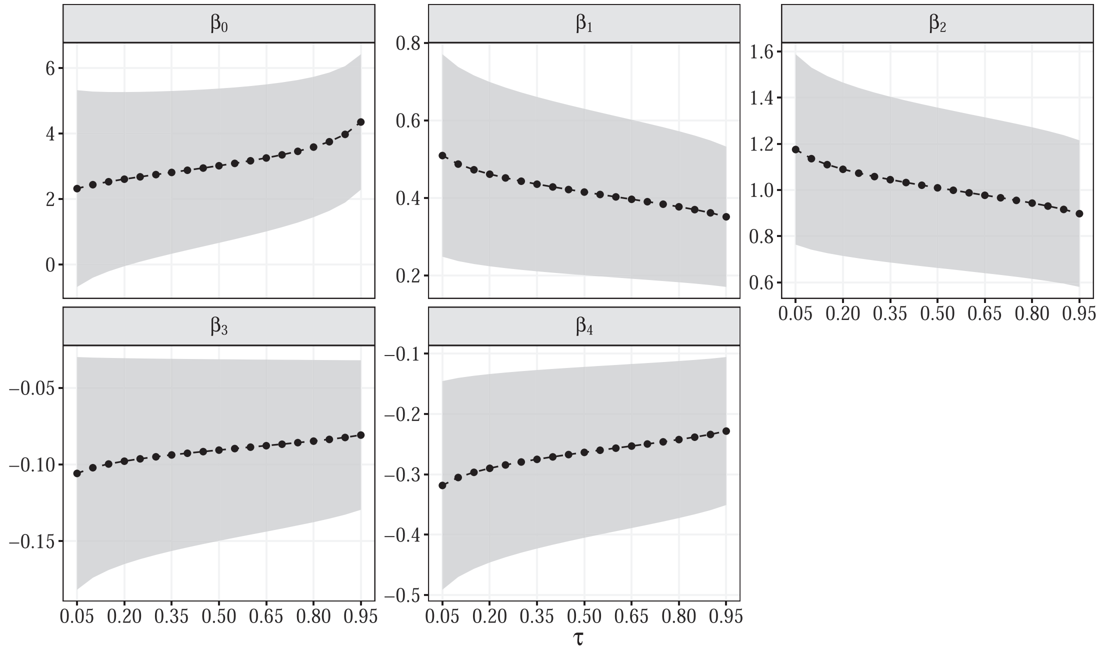

FIGURE 10 Parameter estimates and 95% confidence intervals for $\tau = 0.05, \ldots, 0.95$ by 0.05

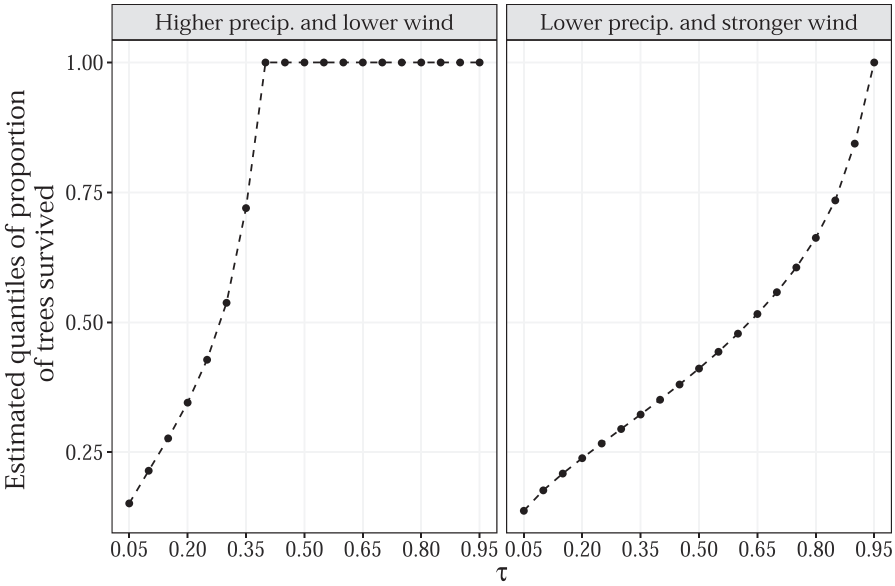

FIGURE 11 Estimated quantiles for two areas

Monte Carlo simulations. Furthermore, we introduce residuals for the proposed model based on the RQ and conducted a simulation study to establish their empirical properties in order to evaluate its performances. Two applications using real data sets were presented and discussed. Furthermore, as suggested by the referees, an extension of the methods developed in this paper would be to consider in (3) a much more general family of distributions; that is, consider models for zero-inflated and one-inflated data sets. The p.d.f. of the zero-and-one inflated UW distribution is given by

$$g(y \mid \nu, \mu, \phi, \tau) = \begin{cases} l\nu_0 & \text{if} \quad y = 0 \\ (1 - \nu_0 - \nu_1)\, f(y \mid \mu, \phi, \tau) & \text{if} \quad y \in (0, 1) \\ \nu_1 & \text{if} \quad y = 1, \end{cases}$$

where $f(y \mid \mu, \phi, \tau)$ is the density of the UW distribution defined in (1) and $0 < \nu_0 + \nu_1 < 1$.

ACKNOWLEDGMENTS

The authors are thankful to the referees for many valuable suggestions. Josmar Mazucheli gratefully acknowledges the partial financial support from Fundação Araucária (Grant 064/2019 - UEM/Fundação Araucária).

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

OPEN RESEARCH BADGES

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.
This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

ORCID

André F. B. Menezes https://orcid.org/0000-0002-3320-9834
Josmar Mazucheli https://orcid.org/0000-0001-6740-0445
Marcelo Bourguignon https://orcid.org/0000-0002-1182-5193

REFERENCES

Atkinson, A. C. (1981). Two graphical displays for outlying and influential observations in regression. *Biometrika*, *68*, 13–20.

Bayes, C. L., Bazán, J. L., & De Castro, M. (2017). A quantile parametric mixed regression model for bounded response variables. *Statistics and Its Interface*, *10*, 483–493.

Bayes, C. L., & Valdivieso, L. (2016). A beta inflated mean regression model for fractional response variables. *Journal of Applied Statistics*, *43*, 1814–1830.

Buchinsky, M., & Hahn, J. (1998). An alternative estimator for the censored quantile regression model. *Econometrica*, *66*, 653–671.

Cook, D., Kieschnick, R., & McCullough, B. (2008). Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*, *15*, 860–867.

Cox, D., & Hinkley, D. V. (1974). *Theoretical statistics*. Boca Raton. FL: Chapman & Hall/CRC.

Cribari-Neto, F., & Santos, J. (2019). Inflated Kumaraswamy distributions. *Annals of the Brazilian Academy of Sciences*, *91*, 1–18.

Di Brisco, A. M., & Migliorati, S. (2020). A new mixed-effects mixture model for constrained longitudinal data. *Statistics in Medicine*, *39*, 129–145.

Dunn, P. K., & Smyth, G. K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, *5*, 236–244.

Hoff, A. (2007). Second stage DEA: Comparison of approaches for modelling the dea score. *European Journal of Operational Research*, *181*, 425–435.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Koenker, R., & Machado, J. A. F. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the American Statistical Association*, *94*, 1296–1310.arXiv:https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1999.10473882.

Korosteleva, O. (2019). *Advanced regression models with SAS and R*. Boca Raton, FL: Taylor & Francis Group.

Lemonte, A., & Moreno-Arenas, G. (2020). On a heavy-tailed parametric quantile regression model for limited range response variables. *Computational Statistics*, *35*, 379–398.

Liu, P., Kam Yuen, K., Wu, L., Tian, G., & Li, T. (2020). Zero-one-inflated simplex regression models for the analysis of continuous proportion data. *Statistics and Its Interface*, *13*, 193–208.

Mazucheli, J., Menezes, A. F. B., Fernandes, L. B., Oliveira, R. P., & Ghitany, M. E. (2020). The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modelling of quantiles conditional on covariates. *Jounal of Applied Statistics*, *47*, 954–974.

Mazucheli, J., Menezes, A. F. B., & Ghitany, M. E. (2018). The unit-Weibull distribution and associated inference. *Journal of Applied Probability and Statistics*, *13*, 1–22.

Morales, C., Lachos, V., Cabral, C., & Cepero, L. (2017). Robust quantile regression using a generalized class of skewed distributions. *Stat*, *6*, 113–130.

Nascimento, F., & Bourguignon, M. (2020). Bayesian time-varying quantile regression to extremes. *Environmetrics*, *31*, e2596.

Noufaily, A., & Jones, M. (2013). Parametric quantile regression based on the generalized gamma distribution. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *62*, 723–740.

Ospina, R., & Ferrari, S. L. P. (2008). Inflated beta distributions. *Statistical Papers*, *51*, 111–126.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*, 609–623.

Pace, L., & Salvan, A. (1997). *Principles of statistical inference from a Neo-Fisherian perspective. Advanced series on statistical science and applied probability*. Vol. 4. World Scientific, Singapore.

Pereira, G., Botter, D., & Sandoval, M. (2013). A regression model for special proportions. *Statistical Modelling*, *13*, 125–151.

Powell, J. (1986). Censored regression quantiles. *Journal of Econometrics*, *32*, 143–155.

Sánchez, L., Leiva, V., Galea, M., & Saulo, H. (2020). Birnbaum-saunders quantile regression and its diagnostics with application to economic data. *Applied Stochastic Models in Business and Industry*, https://doi.org/10.1002/asmb.2556.

Santos, B., & Bolfarine, H. (2015). Bayesian analysis for zero-or-one inflated proportion data using quantile regression. *Journal of Statistical Computation and Simulation*, *85*, 3579–3593.

SAS (2010). *The NLMIXED procedure, SAS/STAT® user's guide, version 9.4*. Cary, NC: SAS Institute Inc.

Yang, Y., Wang, H. J., & He, X. (2016). Posterior inference in Bayesian quantile regression with asymmetric Laplace likelihood. *International Statistical Review*, *84*, 327–344.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Menezes André F. B, Mazucheli J, Bourguignon M. A parametric quantile regression approach for modeling zero-or-one inflated double bounded data. *Biometrical Journal*. 2021;1–18. https://doi.org/10.1002/bimj.202000126