Statistics and Its Interface Volume 10 (2017) 483–493

A quantile parametric mixed regression model for bounded response variables

Cristian L. Bayes[*], Jorge L. Bazán, and Mário de Castro

Bounded response variables are common in many applications where the responses are percentages, proportions, or rates. New regression models have been proposed recently to model the relationship among one or more covariates and the conditional mean of a response variable based on the beta distribution or a mixture of beta distributions. However, when we are interested in knowing how covariates impact different levels of the response variable, quantile regression models play an important role. A new quantile parametric mixed regression model for bounded response variables is presented by considering the distribution introduced by [27]. A Bayesian approach is adopted for inference using Markov Chain Monte Carlo (MCMC) methods. Model comparison criteria are also discussed. The inferential methods can be easily programmed and then easily used for data modeling. Results from a simulation study are reported showing the good performance of the proposed inferential methods. Furthermore, results from data analyses using regression models with fixed and mixed effects are given. Specifically, we show that the quantile parametric model proposed here is an alternative and complementary modeling tool for bounded response variables such as the poverty index in Brazilian municipalities, which is linked to the Gini coefficient and the human development index.

Keywords and phrases: proportions, Kumaraswamy distribution, HDI, Bayesian inference, MCMC methods, Mixed models, RStan.

1. INTRODUCTION

Regression models for response variables in the unit interval, including regression models for percentages, proportions or rates have been introduced recently in the literature. Among them the beta regression model introduced by [23] and [11], the beta-rectangular regression model proposed by [1], and the beta mixed regression model proposed by [13].

Examples of dependent response variables in these models include the percentage of time devoted to an activity during a certain period of time, the fraction of income spent on food, the unemployment rate, the poverty rate, the score achieved in a test, the fraction of "good" cholesterol (HDL/total cholesterol), the proportion of sand in the soil, and the fraction of a surface covered by vegetation.

In the beta regression model, the regression parameters are interpretable in terms of the mean response, and in many aspects are similar to generalized linear models. Estimation can be performed by maximum likelihood [11] or Bayesian methods [3]. The beta regression model is sufficiently documented in several publications such as [8], [9], [12], and [6] and in several applications like in [22] and [37]. In addition, the beta rectangular model proposed by [1] is more robust to outliers (comparatively large or influential values of the response variable) than the beta regression model. This new model, based on a mixture of a beta distribution and a uniform distribution, includes the beta regression and the variable dispersion beta regression model [12] as particular cases.

Linear mixed models have been used to analyze repeated measures data or clustered data. The popularity of these models can be explained by the flexibility to model the within-subject correlation by handling both balanced and unbalanced data. However, such models are not adequate when the response variable is restricted to the unit interval. For these situations, [36] and [13] proposed a beta mixed regression model considering random effects in both the mean and the dispersion parameters. While the first authors employed a maximum likelihood methodology, the last ones opted for a Bayesian approach. Using a logit transformation of the Student-t distribution, [43] developed a robust mixed-effect models for longitudinal response variables in the unit interval.

In the cited literature, the authors concentrated on the relationship between one or more covariates and the conditional mean of a response variable given the covariates and random effects. However, in many applications the quantiles of the response variable are of central interest. Quantile regression, introduced by [24], has attracted the attention of many researchers in recent years, as can be seen in the works by [42], [7], and [28], to name just a few.

Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile. The main advantage is its flexibility for modeling data with heterogeneous conditional distributions. Data of this type occurs in many fields, including Econometrics, Survival Analysis,

[*] Corresponding author.

and Ecology [see, for example, 25]. Thus, quantile mixed regression models for proportions can be useful to model the relationship between the covariates and the conditional quantiles of the response variable given the covariates and random effects. Quantile mixed regression models also provide a more complete picture of the conditional distribution of the response variable given the covariates and random effects. Consider, for example, a model for the quantiles of a socioeconomic level or the achievement in an educational test. The interest might rest on the upper quantiles.

From a Bayesian perspective, [41] proposed to assume an asymmetric Laplace distribution (ALD) in a parametric quantile regression model for an unbounded response variable. [26] provided an useful stochastic representation for the ALD that facilitates the implementation of a Gibbs sampling scheme for this model. This approach has been extended to a random intercept regression model by [16] and to a mixed regression model by [17]. Our proposal is similar to the one in [17], but for a response variable in the unit interval, which can be easily extended to any bounded response variables on the interval (c_1, c_2), with $c_1 < c_2$.

Since the cumulative distribution function (cdf) of the beta distribution does not have a closed form, quantile regression models built upon this distribution pose some difficulties. In contrast, the Kumaraswamy distribution [27, 21] is a continuous probability distribution defined on the $(0, 1)$ interval that is similar to the beta distribution, but with the advantage of having a simple closed form for both the probability density function (pdf) and the cdf. In order to formulate a quantile mixed regression models, we consider a convenient reparameterization of the Kumaraswamy distribution in terms of a precision parameter and the q-th quantile $\kappa = \kappa(q) \in (0, 1)$, which is a location parameter.

The main goal of this paper is to propose a parametric quantile mixed regression model for proportions assuming that the response variable follows a Kumaraswamy distribution. The proposed model can be easily extended to any bounded response variable.

The paper is organized as follows. In Section 2 we present a short account of the Kumaraswamy distribution and a convenient parameterization that is introduced in order to formulate our general mixed quantile regression model. In Sections 3 and 4 we formulate and develop a Bayesian approach for the proposed regression model including model comparison criteria. In Section 5 we present results from simulation studies. Two real data sets are analysed in Section 6 using our proposed models. Final comments are presented in Section 7.

2. THE KUMARASWAMY DISTRIBUTION

A random variable Y follows the Kumaraswamy distribution if its pdf is given by

$$f(y|\alpha,\beta) = \alpha\beta y^{\alpha-1}(1-y^{\alpha})^{\beta-1}, \quad 0 < y < 1,\ \alpha,\beta > 0. \tag{1}$$

The cdf has closed expression and is given by $F(y|\alpha,\beta) = 1-(1-y^{\alpha})^{\beta}$. The mean and variance of this distribution are

$$\begin{aligned} E(Y|\alpha,\beta) &= \beta B\left(1+\frac{1}{\alpha},\beta\right) \quad \text{and} \\ Var(Y|\alpha,\beta) &= \beta B\left(1+\frac{2}{\alpha},\beta\right) - \beta^2 B^2\left(1+\frac{1}{\alpha},\beta\right), \end{aligned} \tag{2}$$

where $B(\cdot,\cdot)$ denotes the beta function.

As pointed out by [30], the expressions for $E(Y)$ and $Var(Y)$ make a mean-variance based reparameterization unfeasible. However, we can find a simple expression for the quantile function, given by $\kappa(q) = F^{-1}(q) = \{1-(1-q)^{1/\beta}\}^{1/\alpha}$, for $0 < q < 1$. In particular, the median is given by $\kappa(0.5) = (1-0.5^{1/\beta})^{1/\alpha}$.

In order to propose a quantile regression analysis, we consider a reparameterization of the Kumaraswamy distribution in terms of the q-th quantile $\kappa(q)$ and the precision parameter $\varphi = \varphi(q)$ following the ideas presented in [30]. Thus, to obtain a more appropriate regression structure for the Kumaraswamy distribution, we take

$$\kappa = \{1-(1-q)^{1/\beta}\}^{1/\alpha} \quad \text{and} \quad \varphi = -\log\left(1-(1-q)^{1/\beta}\right) \tag{3}$$

as a new parameterization. In this case, q is assumed to be known and the parameter space of $(\kappa,\varphi)^T$ is given by $(0, 1) \times (0, \infty)$.

Under this parameterization, the pdf and the cdf of the Kumaraswamy distribution turn out to be

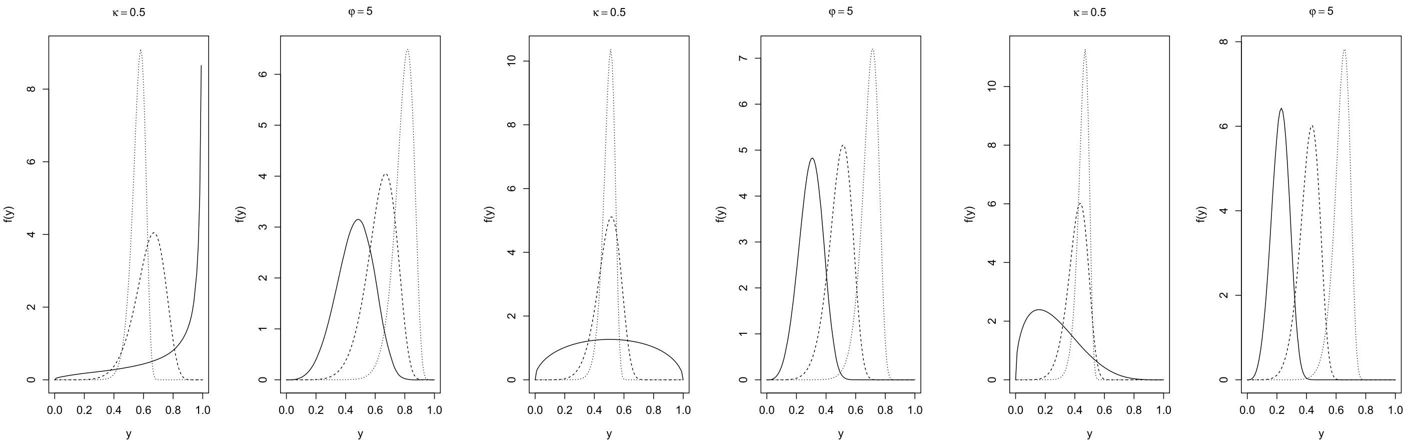
$$\begin{aligned} f(y|\kappa,\varphi) = &-\frac{\log(1-q)\varphi}{\log\left(1-e^{-\varphi}\right)\log(\kappa)} y^{-\frac{\varphi}{\log(\kappa)}-1} \\ &\times \left\{1-y^{-\frac{\varphi}{\log(\kappa)}}\right\}^{\frac{\log(1-q)}{\log(1-e^{-\varphi})}-1} \end{aligned} \tag{4}$$

and

$$F(y|\kappa,\varphi) = 1-\left\{1-y^{-\frac{\varphi}{\log(\kappa)}}\right\}^{\frac{\log(1-q)}{\log(1-e^{-\varphi})}}.$$

We consider the notation $Y \sim \mathrm{K}(\kappa,\varphi,q)$ with the quantile parameter $\kappa \in (0, 1)$ as a location parameter, $\varphi > 0$ as a precision parameter, and the probability q is assumed to be fixed according to the quantile of interest.

Figure 1 depicts the pdf of the reparameterized version of the Kumaraswamy distribution in (4) for different values of κ and φ. We pick the first decile, the median, and the last decile. When κ is fixed, we note that φ is a parameter that controls the precision of the distribution. For larger values of φ we observe less dispersion. On the other hand, when φ is fixed we note that κ acts as a parameter that controls the location of the distribution. For instance, for larger values of φ the mode tends to move to the right. In general, since κ is the q-th quantile of Y, we interpret it as a location parameter in the range of values of the variable being modeled.



(a) 0.1-quantile

(b) 0.5-quantile

(c) 0.9-quantile

Figure 1. Kumaraswamy pdf for different values of κ and φ. Left panel: $\kappa = 0.5$ and different values of φ: 1 (solid line), 5 (dashed line), and 10 (dotted line). Right panel: $\varphi = 5$ and different values of κ: 0.3 (solid line), 0.5 (dashed line), and 0.7 (dotted line).

3. THE KURAMASWAMY QUANTILE DISPERSION MIXED REGRESSION MODEL

As indicated by [40], we introduce random effects for each sample unit (individual or cluster) to incorporate a correlation between the repeated measurements within the unit sample. In this situation, there are two sources of variation in the data: the between-unit variation and the within-unit variation.

Let $\boldsymbol{Y}_i = (y_{i1}, \ldots, y_{in_i})^T$ be a vector of responses for the sample unit i, where each component y_{ij} takes values in the $(0, 1)$ interval. The Kuramaswamy quantile dispersion mixed regression model is given by

$$\begin{aligned} &y_{ij} \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_{ij}, \varphi_{ij}, q), \\ &g_1(\kappa_{ij}) = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \text{ and } g_2(\varphi_{ij}) = -\boldsymbol{w}_{ij}^T\boldsymbol{\delta} - \boldsymbol{h}_{ij}^T\boldsymbol{d}_i, \end{aligned} \tag{5}$$

for $j = 1, \ldots, n_i$ and $i = 1, \ldots, n$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_l)^T$ are vectors of regression coefficients (fixed effects) associated with the location and the precision parameters, respectively, $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{ip})^T$ and $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{ir})^T$ are the random effects of the location and precision parameters, respectively, $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijk})^T$, $\boldsymbol{w}_{ij} = (w_{ij1}, \ldots, w_{ijl})^T$, $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijp})^T$, and $\boldsymbol{w}_{ij} = (h_{ij1}, \ldots, h_{ijr})^T$ are covariate vectors (possibly overlapping or even identical) and $q \in (0, 1)$ is the fixed probability associated to the quantile of interest.

We assume that the random effects $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n$ and $\boldsymbol{d}_1, \ldots, \boldsymbol{d}_n$ are all independent and normally distributed, i.e., $\boldsymbol{b}_i \sim N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_b)$ and $\boldsymbol{d}_i \sim N_r(\boldsymbol{0}, \boldsymbol{\Sigma}_d)$, for $i = 1, \ldots, n$, being $\boldsymbol{\Sigma}_b$ and $\boldsymbol{\Sigma}_d$ positive definite matrices.

In general, the link function $g_1(\cdot)$ relating the quantile κ_{ij} with the covariates and random effects can be the inverse of any cdf corresponding to a continuous distribution. Some examples are the logit, probit, and complementary log-log functions. In this paper we adopt the logit link, but other link functions might be explored. Similarly, $g_2(\cdot)$ is a link function relating the precision parameter φ_{ij} with the covariates and random effects. Since the φ_{ij} must be strictly positive, we will use the log link. We take the negative sign, similar to [32], to ease the interpretation of the coefficients. Since φ is a precision parameter, a positive-signed δ_{ij} indicates smaller variability, which can be confusing. It seems more natural to model the dispersion rather than the precision, and the negative sign enables us to do so.

Under the parameterization in (4), the augmented likelihood function can be written as

$$\begin{aligned} L(\boldsymbol{\theta}, &\boldsymbol{b}, \boldsymbol{d} | \boldsymbol{Y}) = \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} f(y_{ij}|\kappa_{ij}, \varphi_{ij})\phi_p(\boldsymbol{b}_i \mid \boldsymbol{0}, \boldsymbol{\Sigma}_b)\phi_r(\boldsymbol{d}_i \mid \boldsymbol{0}, \boldsymbol{\Sigma}_d) \\ &= \prod_{i=1}^{n}\prod_{j=1}^{n_i} \frac{-\log(1-q)\varphi_{ij}}{\log(1-e^{-\varphi_{ij}})\log(\kappa_{ij})} y_{ij}^{-\frac{\varphi_{ij}}{\log(\kappa_{ij})}-1} \\ &\times \left\{1 - y_{ij}^{-\frac{\varphi_{ij}}{\log(\kappa_{ij})}}\right\}^{\frac{\log(1-q)}{\log(1-e^{-\varphi_{ij}})}-1} \\ &\times \phi_p(\boldsymbol{b}_i \mid \boldsymbol{0}, \boldsymbol{\Sigma}_b)\phi_r(\boldsymbol{d}_i \mid \boldsymbol{0}, \boldsymbol{\Sigma}_d), \end{aligned} \tag{6}$$

where $\boldsymbol{\theta}$ encapsulates $\boldsymbol{\beta}$, $\boldsymbol{\delta}$, $\boldsymbol{\Sigma}_b$, $\boldsymbol{\Sigma}_d$, $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)^T$, $\boldsymbol{d} = (\boldsymbol{d}_1^T, \ldots, \boldsymbol{d}_n^T)^T$, $\kappa_{ij} = 1/\{1 + \exp(-\boldsymbol{x}_{ij}^T\boldsymbol{\beta} - \boldsymbol{z}_{ij}^T\boldsymbol{b}_i)\}$, $\varphi_{ij} = \exp(-\boldsymbol{w}_{ij}^T\boldsymbol{\delta} - \boldsymbol{h}_{ij}^T\boldsymbol{d}_i)$, and $\phi_s(\cdot \mid \boldsymbol{m}, \boldsymbol{S})$ denotes the pdf of the s-variate normal distribution with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{S}$.

4. BAYESIAN INFERENCE

With independent observations, the likelihood function for the Kumaraswamy quantile mixed regression model is obtained from (5) and (6). In this way, the augmented posterior distribution of $\boldsymbol{\theta}$, $\boldsymbol{b}$, and $\boldsymbol{d}$, denoted by $p(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{d} \mid \boldsymbol{Y})$, is

$$p(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{d} \mid \boldsymbol{Y}) \propto L(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{d} \mid \boldsymbol{Y}) p(\boldsymbol{\theta}), \tag{7}$$

where $p(\boldsymbol{\theta})$ stands for the prior distribution of $\boldsymbol{\theta}$. To complete the Bayesian specification of the model, we set the prior as

$$p(\boldsymbol{\theta}) = p(\boldsymbol{\beta}) p(\boldsymbol{\delta}) p(\boldsymbol{\Sigma}_b) p(\boldsymbol{\Sigma}_d). \tag{8}$$

We propose the multivariate normal distributions as the prior distribution for the fixed effects, i.e., $\boldsymbol{\beta} \sim N_k(\boldsymbol{0}, \boldsymbol{A})$ and $\boldsymbol{\delta} \sim N_l(\boldsymbol{0}, \boldsymbol{C})$. For the covariance matrices of the random effects, we adopt inverse Wishart distributions, i.e., $\boldsymbol{\Sigma}_b \sim IW_p(\psi_b, \boldsymbol{\Psi}_b)$ and $\boldsymbol{\Sigma}_d \sim IW_r(\psi_d, \boldsymbol{\Psi}_d)$, where $\boldsymbol{A}$, $\boldsymbol{C}$, ψ_b, $\boldsymbol{\Psi}_b$, ψ_d, and $\boldsymbol{\Psi}_b$ are specified hyperparameters.

After combining the likelihood function in (6) with the prior distribution in (8), we get the posterior distribution

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{b}, \boldsymbol{d} \mid \boldsymbol{Y}) = & \prod_{i=1}^{n} \prod_{j=1}^{n_i} \frac{-\log(1-q)\varphi_{ij}}{\log(1-e^{-\varphi_{ij}})\log(\kappa_{ij})} y_{ij}^{-\frac{\varphi_{ij}}{\log(\kappa_{ij})}-1} \\ & \times \left\{1 - y_{ij}^{-\frac{\varphi_{ij}}{\log(\kappa_{ij})}}\right\}^{\frac{\log(1-q)}{\log(1-e^{-\varphi_{ij}})}-1} \\ & \times \phi_p(\boldsymbol{b}_i \mid \boldsymbol{0}, \boldsymbol{\Sigma}_b) \phi_r(\boldsymbol{d}_i \mid \boldsymbol{0}, \boldsymbol{\Sigma}_d) \\ & \times \phi_k(\boldsymbol{\beta} \mid \boldsymbol{0}, \boldsymbol{\Sigma}_b) \phi_l(\boldsymbol{\delta} \mid \boldsymbol{0}, \boldsymbol{\Sigma}_d) \\ & \times g(\boldsymbol{\Sigma}_b \mid \psi_b, \boldsymbol{\Psi}_b) g(\boldsymbol{\Sigma}_d \mid \psi_d, \boldsymbol{\Psi}_d), \end{aligned} \tag{9}$$

where $g(\cdot \mid \psi, \boldsymbol{\Psi})$ denotes the pdf of the inverse Wishart distribution. In the particular case of the fixed effects regression model with a constant precision parameter φ in Sections 5.2 and 6.1, the prior distribution for $(\boldsymbol{\beta}^T, \varphi)^T$ will be $p(\boldsymbol{\beta}, \varphi) = p(\boldsymbol{\beta}) p(\varphi)$, where $\log(\varphi) \sim N(0, \sigma_1^2)$, with σ_1^2 being sufficiently large to ensure vague prior knowledge.

The posterior distribution in (9) is intractable. Hence, a possible approximation is obtained through MCMC methods to draw samples from the posterior density. A simple way is to trust in the capabilities of the WinBUGS software [29] or the RStan package [35] in R [34] (see in Appendix A an example of RStan code). For the particular case of a location-regression model without modeling the precision parameter, the MCMC computations in Section 5.2 were implemented using the FORTRAN language.

4.1 Model comparison criteria

There are several criteria for comparing different models fitted to a given data set and for selecting the one that best fits the data. First, we mention the deviance information criterion (DIC) proposed by [33]. The DIC is built upon the deviance $\mathcal{D}(\boldsymbol{\vartheta}) = -2\ell(\boldsymbol{\vartheta}|\boldsymbol{Y})$, with $\boldsymbol{\vartheta} = (\boldsymbol{\theta}^T, \boldsymbol{b}^T, \boldsymbol{d}^T)^T$ and $\ell(\boldsymbol{\vartheta}|\boldsymbol{Y})$ denoting the logarithm of the likelihood function in (6). From G samples $\boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_G$ generated by the Gibbs sampler, the DIC is computed as $DIC = \mathcal{D}(\overline{\boldsymbol{\vartheta}}) + 2p_D$, where $p_D = \overline{\mathcal{D}}(\boldsymbol{\vartheta}) - \mathcal{D}(\overline{\boldsymbol{\vartheta}})$ is termed the effective number of parameters, with $\overline{\mathcal{D}}(\boldsymbol{\vartheta}) = \sum_{g=1}^{G} \mathcal{D}(\boldsymbol{\vartheta}_g)/G$ and $\overline{\boldsymbol{\vartheta}} = \sum_{g=1}^{G} \boldsymbol{\vartheta}_g/G$. We also consider the expected Akaike's information criteria ($EAIC$) and its Bayesian version ($EBIC$), all of them detailed for example in [15]. These criteria are defined as $EAIC = \mathcal{D}(\overline{\boldsymbol{\vartheta}}) + 2p^*$ and $EBIC = \mathcal{D}(\overline{\boldsymbol{\vartheta}}) + p^* \log(n)$, respectively, where p^* is the number of parameters in the model and n is the sample size. Additionally, we used the criterion proposed by [39], named Wanatabe's information criterion ($WAIC$). Wanatabe's proposal can be viewed as an approximation to cross-validation [15] and the only difference with respect to $EAIC$ and $EBIC$ is the computation of the complexity penalty p_{WAIC}. In this work, we adopt the variance version due to its stability properties. Hence, in this case, $WAIC = \mathcal{D}(\overline{\boldsymbol{\vartheta}}) + 2p_{WAIC}$, with $p_{WAIC} = \sum_{i=1}^{n} Var\big(\log(p(y_i|\boldsymbol{\vartheta}))\big)$ computed from the output of the Gibbs sampler. Given a set of candidate models, the model yielding the smallest value of these criteria is the one that best fits the data.

5. SIMULATIONS

Based on the methodology described in Sections 3 and 4, simulation studies were carried out and two real data sets were analyzed. For all situations, we simulate a large number of samples, discarding the first of them as a burn-in period. To avoid correlation problems, we consider a spacing of size equal to five or more. The convergence of the chains was monitored by the Geweke's statistic [18] and graphical inspection of the chains. The highest posterior density (HPD) intervals were computed following the steps described in [5, Section 7.3.1].

5.1 Prior sensitivity analysis

In this section, we conduct a sensitivity analysis of the prior specification for the precision parameter (φ). We consider a procedure similar to the one developed by [13] for the mixed beta regression model. First, a single data set is generated from a Kumaraswamy mixed regression model given by $y_{ij} \mid \boldsymbol{b}_i, \varphi, \boldsymbol{\beta} \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_{ij}, \varphi, 0.5)$, with $\mathrm{logit}(\kappa_{ij}) = (\beta_1 + b_{i1}) + (\beta_2 + b_{i2}) x_{ij2} + \beta_3 x_{ij3}$, for $j = 1, \ldots, 5$ and $i = 1, \ldots, 200$, where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$. The covariates were generated from a uniform distribution in the unit interval and $\boldsymbol{b}_i = (b_{i1}, b_{i2})^T \sim N_2(\boldsymbol{0}, \boldsymbol{\Sigma}_b)$. The parameters were set at $\varphi = 50$, $\boldsymbol{\beta} = (-2, 1, 2)^T$, and $\boldsymbol{\Sigma}_b$ has elements $Var(b_{i1}) = 1$, $Var(b_{i2}) = 0.2$, and $Cov(b_{i1}, b_{i2}) = -0.3$.

Next, we use the generated data set to perform a sensitivity analysis under different prior specifications for the precision parameter. Specifically, we consider three priors previously used in the literature: (i) $\varphi \sim IG(0.01, 0.01)$ (Ma), (ii) $\log(\varphi) \sim N(0, 25)$ (Mb), and (iii) $\log(\varphi) \sim N(0, 10^4)$ (Mc).

Table 1. Model comparison criteria for models fitted to a synthetic data set with different prior specifications for the precision parameter under a Kumaraswamy mixed quantile regression model

Model	Prior for φ	DIC	$EAIC$	$EBIC$	$WAIC$
Ma	$\varphi \sim IG(0.01, 0.01)$	-2680.2	-2669.1	-1326.7	-2789.8
Mb	$\log(\varphi) \sim N(0, 25)$	-2684.6	-2669.9	-1326.5	-2795.3
Mc	$\log(\varphi) \sim N(0, 10^4)$	-2684.2	-2668.9	-1326.5	-2792.7

The prior distributions for the remaining parameters were specified as $\boldsymbol{\beta} \sim N_3(\mathbf{0}, 10^4\boldsymbol{I}_3)$ and $\boldsymbol{\Sigma}_b \sim IW_2(5, 20\boldsymbol{I}_2)$. Then, we use the MCMC methods outlined in Section 4 to draw samples from the posterior distribution. We discarded the first 10,000 of 20,000 iterations and a thinning equal to 10 resulted in 2000 samples upon which the posterior inference is based on.

Table 1 reports the values of DIC, $EAIC$, $EBIC$, and $WAIC$ for the fitted models with different prior distributions for φ. We observe that the models Mb and Mc outperform model Ma. Hence, the vaguer prior $\log(\varphi) \sim N(0, 10^4)$ is adopted in the models with constant precision (Sections 5.2 and 6.1).

5.2 Parameter recovery

We conduct a brief simulation study to assess the performance of our Bayesian approach in estimating the parameters of the model given in (5). In this study we consider a location regression model without random effects and we assume a constant precision parameter φ. We choose three quantiles levels; namely, a lower quantile ($q = 0.1$), the median ($q = 0.5$), and a higher quantile ($q = 0.9$), that is,

$$y_i \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_i, \varphi, q), \quad \kappa_i = \frac{1}{1 + \exp(-\beta_1 - \beta_2 x_{i2} - \beta_3 x_{i3})},$$

$i = 1, \ldots, n,\ q = 0.1, 0.5, 0.9.$

To generate the data, we first draw n independent $x_{ik} \sim N(0, 1)$ covariates with $k = 2, 3$ and $x_{i1} = 1$ corresponding to the intercept. These values remain fixed throughout the 500 repetitions of the simulations and the three sample sizes: small ($n = 40$), intermediate ($n = 100$), and relatively big ($n = 300$). Therefore, the simulation study comprises nine scenarios.

The prior distributions were specified as $\boldsymbol{\beta} \sim N_3(\mathbf{0}, 10^4\boldsymbol{I}_3)$ and $\log(\varphi) \sim N(0, 10^4)$, with $\boldsymbol{I}_3$ denoting the 3×3 unity matrix. For each replication, after discarding the first 1000 iterations of the Gibbs sampler, we used 10,000 iterations with thinning equal to 5, leading to 2000 samples for each parameter.

Some posterior results together with the true parameters are summarized in Table 2. We observe that, for all the considered scenarios, our method performs well. In particular, the bias is negligible, even when the sample size is as small as $n = 40$, and the coverage probability of the 95% HPD intervals differs from the nominal value by at most 2.8%. Since the datasets were generated from several scenarios covering different quantile values including extremes, this study indicates that the Bayesian estimator yields good results irrespective of the scenario. Furthermore, as expected, for a given quantile, the average of the posterior standard deviations (SD) and the root mean squared errors of the posterior means (RMSE) are close and decrease when the sample size increases.

Table 2. Posterior results from 500 replications (Par: parameter to be estimated, True: true value of the parameter, Est: average of the posterior means, SD: average of the posterior standard deviations, RMSE: root mean squared error of the posterior means and CP: coverage probability of the 95% HPD interval)

$n = 40$					0.1-quantile						0.5-quantile						0.9-quantile
Par	True	Est	SD	RMSE	CP	Par	True	Est	SD	RMSE	CP	Par	True	Est	SD	RMSE	CP
β_1	0.5	0.16	0.99	0.89	0.962	β_1	0.5	0.47	0.32	0.29	0.970	β_1	0.5	0.51	0.20	0.21	0.922
β_2	-1.5	-1.67	1.41	1.26	0.950	β_2	-1.5	-1.52	0.43	0.39	0.972	β_2	-1.5	-1.51	0.23	0.24	0.934
β_3	0.9	1.01	0.61	0.58	0.958	β_3	0.9	0.91	0.22	0.23	0.934	β_3	0.9	0.90	0.14	0.14	0.932
φ	1.2	1.19	0.15	0.16	0.942	φ	1.2	1.21	0.20	0.20	0.954	φ	1.2	1.23	0.27	0.29	0.942
$n = 100$					0.1-quantile						0.5-quantile						0.9-quantile
Par	True	Est	SD	RMSE	CP	Par	True	Est	SD	RMSE	CP	Par	True	Est	SD	RMSE	CP
β_1	0.5	0.44	0.43	0.41	0.948	β_1	0.5	0.50	0.17	0.17	0.962	β_1	0.5	0.49	0.13	0.13	0.952
β_2	-1.5	1.56	0.64	0.60	0.966	β_2	-1.5	-1.50	0.24	0.23	0.960	β_2	-1.5	-1.49	0.16	0.16	0.938
β_3	0.9	0.96	0.38	0.37	0.948	β_3	0.9	0.90	0.13	0.13	0.938	β_3	0.9	0.90	0.09	0.09	0.962
φ	1.2	1.20	0.09	0.09	0.952	φ	1.2	1.20	0.12	0.12	0.956	φ	1.2	1.22	0.16	0.16	0.946
$n = 300$					0.1-quantile						0.5-quantile						0.9-quantile
Par	True	Est	SD	RMSE	CP	Par	True	Est	SD	RMSE	CP	Par	True	Est	SD	RMSE	CP
β_1	0.5	0.46	0.24	0.23	0.950	β_1	0.5	0.47	0.09	0.09	0.946	β_1	0.5	0.50	0.07	0.07	0.952
β_2	-1.5	-1.50	0.34	0.33	0.946	β_2	-1.5	-1.52	0.13	0.14	0.934	β_2	-1.5	-1.50	0.09	0.09	0.952
β_3	0.9	0.91	0.18	0.17	0.956	β_3	0.9	0.88	0.07	0.08	0.936	β_3	0.9	0.90	0.05	0.04	0.954
φ	1.2	1.2	0.05	0.05	0.950	φ	1.2	1.19	0.07	0.07	0.952	φ	1.2	1.21	0.09	0.09	0.960

6. REAL DATA ANALYSIS

6.1 Location quantile regression model

In this section we analyze a real data set using quantile regression models without random effects and a fixed precision parameter φ. [10] studied the attitudes toward Statistics of 146 Elementary School teachers of primary education taking into account some characteristics of them such as whether they have a specialty or not, named here as specialty (covariate x_2 with Sciences as a baseline, and categories Social Sciences, SS, and Elementary School without specialty, ES), the country where the teachers live (covariate x_3 with categories Spain as baseline and Perú, P), and gender (covariate x_4 with categories female as baseline and male, M), whereas $x_1 = 1$ in order to accommodate the intercept. The scale of attitudes consists of 25 items from a five-point Likert scale ranging from "strongly disagree" (level 1) to "strongly agree" (level 5). The responses of a subject are added together to form a score S with values in the set $\{5, 6, \ldots, 125\}$. For this data set, the score ranges from 48 to 102, with mean and standard deviation equal to 77.9 and 11.0, respectively. In our application, we take the transformed score Y, given by $Y = (S - 25)/(125 - 25)$, as the response variable, ranging from 0.23 to 0.77, with mean

Table 3. Posterior summaries for the fitted models (Mean: mean and SD: standard deviation)

		Full model			Reduced model		
Model	Parameter	Mean	SD	95% HPD interval	Mean	SD	95% HPD interval
0.25-quantile Kumaraswamy regression	Intercept	0.15	0.09	(-0.02, 0.32)	0.10	0.06	(-0.01, 0.20)
	Specialty: SS	-0.15	0.09	(-0.32, 0.03)			
	Specialty: ES	-0.35	0.14	(-0.60, -0.08)	-0.23	0.11	(-0.43, 0.01)
	Country: P	-0.33	0.09	(-0.50, -0.15)	-0.38	0.08	(-0.54, -0.22)
	Gender: M	0.04	0.08	(-0.11, 0.20)			
	φ	4.58	0.22	(4.16, 5.01)	4.56	0.22	(4.14, 4.98)
	DIC	-263.4			-264.6		
	$EAIC$	-257.5			-260.6		
	$EBIC$	-239.6			-248.7		
	$WAIC$	-262.9			-264.5		
0.5-quantile Kumaraswamy regression	Intercept	0.43	0.08	(0.28, 0.58)	0.38	0.05	(0.29, 0.48)
	Specialty: SS	-0.14	0.08	(-0.30, 0.02)			
	Specialty: ES	-0.32	0.13	(-0.57, -0.08)	-0.21	0.11	(-0.41, 0.00)
	Country: P	-0.31	0.08	(-0.48, -0.16)	-0.35	0.08	(-0.50, -0.20)
	Gender: M	0.04	0.08	(-0.11, 0.19)			
	φ	3.72	0.22	(3.29, 4.12)	3.69	0.21	(3.28, 4.12)
	DIC	-263.4			-264.6		
	$EAIC$	-257.6			-260.6		
	$EBIC$	-239.7			-248.7		
	$WAIC$	-262.9			-264.5		
0.75-quantile Kumaraswamy regression	Intercept	0.68	0.08	(0.54, 0.83)	0.64	0.05	(0.55, 0.74)
	Specialty: SS	-0.13	0.08	(-0.29, 0.02)			
	Specialty: ES	-0.31	0.12	(-0.54, -0.08)	-0.20	0.10	(-0.39, -0.01)
	Country: P	-0.29	0.08	(-0.45, -0.15)	-0.34	0.07	(-0.48, -0.19)
	Gender: M	0.03	0.07	(-0.10, 0.18)			
	φ	3.04	0.21	(2.59, 3.43)	3.01	0.21	(2.59, 3.41)
	DIC	-263.4			-264.6		
	$EAIC$	-257.6			-260.6		
	$EBIC$	-239.7			-248.7		
	$WAIC$	-262.9			-264.5		
Mean beta regression	Intercept	0.41	0.07	(0.27, 0.56)	0.36	0.05	(0.26, 0.45)
	Specialty: SS	-0.12	0.08	(-0.28, 0.03)			
	Specialty: ES	-0.28	0.11	(-0.50, -0.06)	-0.18	0.09	(-0.34, -0.01)
	Country: P	-0.34	0.08	(-0.50, -0.19)	-0.38	0.07	(-0.51, -0.24)
	Gender: M	0.01	0.07	(-0.13, 0.14)			
	φ	26.37	3.08	(20.70, 32.73)	26.31	3.02	(20.37, 32.19)
	DIC	-267.1			-268.9		
	$EAIC$	-267.3			-268.9		
	$EBIC$	-249.4			-256.9		
	$WAIC$	-266.1			-268.6		

and standard deviation equal to 0.53 and 0.11, respectively.

We have interest in capturing the effect of covariates on different levels of attitude assuming a common effect in the dispersion, that is, we have interest in a quantile regression model for the attitude toward Statistics. Thus, the Kumaraswamy location quantile regression model was considered. We fit the model to assess the effect of the covariates on three levels of attitude, i.e., lower (0.25-quantile), middle (0.5-quantile), and upper (0.75-quantile). The fitted model is specified as $y_i \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_i, \varphi, q)$, for $q = 0.25, 0.5, 0.75$, with

$$\operatorname{logit}(\kappa_i) = \beta_1 + \beta_2 I(x_{2i} = \mathrm{SS}) + \beta_3 I(x_{2i} = \mathrm{ES}) + \beta_4 I(x_{3i} = \mathrm{P}) + \beta_5 I(x_{4i} = \mathrm{M}),$$

for $i = 1, \ldots, 146$, where $I(x = \mathrm{A}) = 1$, if $x = \mathrm{A}$; 0, otherwise. As in Section 5.2, we set the prior distributions as $\boldsymbol{\beta} \sim N_5(0, 10^4 \boldsymbol{I}_5)$ and $\log(\varphi) \sim N(0, 10^4)$. For comparison, we include also the usual beta regression model. In this case we adapted the BUGS code provided by [3]. We ran the Gibbs sampler under the same conditions, discarding the first 10,000 iterations and performing 25,000 additional iterations with thinning equal to 5, leading to 5000 samples for each parameter. The convergence analysis of the MCMC chains provide strong indication of chain convergence in all fitted models. We fitted also the location-precision model linking the precision parameter φ to the covariates. However, according to the information criteria in Section 4.1 (not shown), the location model is preferred.

Table 3 collects posterior summaries of the full (with all covariates) and the reduced models (without non-significant coefficients, i.e., regression coefficients such that the 95% HPD intervals include 0). Since the three quantile models represent different parameterizations of the same model, the differences in the information criteria are due only to round-off errors. Taking into account the information criteria in Table 3, we select for simplicity the reduced models as our working models. The results indicate that the effect of the covariates is similar at the three levels of attitude, differing only in the estimate of the intercept. Note also that the median regression and the beta regression models have similar estimates of the coefficients. This is not surprising because mean and median are comparable measures of location. However, note that the estimate of the correspondent dispersion parameter in the median regression model takes lower values in comparison to the corresponding dispersion parameters in the mean regression model.

Considering the results of the four models, we conclude that Elementary School teachers without specialty and Peruvian teachers present significant lower attitudes toward Statistics in comparison with the baseline categories (Sciences/Social Sciences teachers and Spanish teachers, respectively). The results are in accordance with the findings in [10].

Not surprisingly, the posterior means in Tables 3 have the same sign. We observe that the larger the quantile, the smaller the estimate of the shape parameter φ, indicating less precision and higher kurtosis for the quantile of attitude. We also note that the estimates for the 0.5-quantile model and the beta regression model are similar. The effect of specialty slightly changes from non-significant to significant, indicating that our model can reveal that the role of a covariate is not necessarily important at different levels (quantiles) of the response variable.

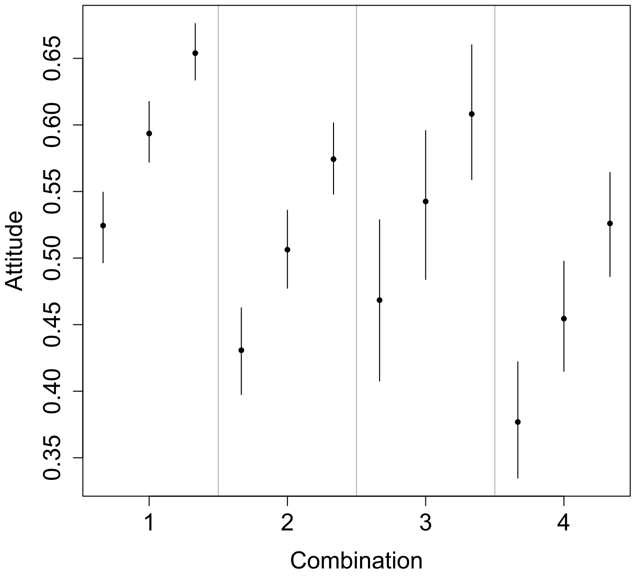


Figure 2. Posterior means and 95% HPD intervals for the 0.25-, 0.5-, and 0.75-quantiles of attitude, from left to right. 1: Spanish teachers of Sciences/Social Sciences, 2: Elementary School Spanish teachers without specialty, 3: Peruvian teachers of Sciences/Social Sciences, and 4: Elementary School Peruvian teachers without specialty.

From (5) we obtain $\kappa = \operatorname{logit}(\boldsymbol{x}^T \boldsymbol{\beta})$. For a given q, using the output of the Gibbs sampler we get samples of the q-quantile. Figure 2 shows posterior summaries for the 0.25, 0.5, and 0.75 quantiles of attitude. This figure synthesizes the effects of the covariates on the quantiles. There are four different combinations of the levels of specialty and country. Note that for Spanish teachers, comparing the teachers of Sciences/Social Sciences and Elementary School teachers without specialty, the three 95% HPD intervals of attitude do not overlap. Note also the high variability in the attitude for Peruvian teachers. As pointed out by [10], this can be explained, at least partially, since in Spain there is a greater effort in Statistics teachers' formation, curriculum organization, and didactics.

According to Table 3, the beta model yields a better fit to the data. However, since we are interested in the quantiles, the Kuramaswamy model was taken as our working model. Moreover, we are aware of the possible quantile crossing issue as pointed out by [2]. In this example this problem was not observed.

Table 4. Model comparison criteria for the fitted models

Model	p^*	p_D	p_{WAIC}	$\overline{\mathcal{D}}(\boldsymbol{\vartheta})$ (Dbar)	DIC	$EAIC$	$EBIC$	$WAIC$
M1	4	4.0	5.7	-18050.9	-18042.9	-18042.9	-18016.4	-18041.1
M2	6	6.0	7.8	-18429.9	-18418.0	-18417.9	-18378.2	-18416.1
M3	32	32.1	34.3	-19608.9	-19544.8	-19544.9	-19332.9	-19541.7
M4	62	61.7	62.0	-20026.6	-19903.1	-19902.6	-19491.9	-19899.3

6.2 Mixed quantile dispersion regression model

In this section we analyze data collected from a nationwide household survey in Brazil, the so-called PNAD (National Household Survey) [19], carried out in 2013. The data comprises 5563 municipalities in Brazil (two of them were excluded because they present missing values). The response variable is a poverty index defined as the proportion of people with per capita income equal to or less than R$ 140.00 (Brazilian currency) per month (ranging from 0.002 to 0.786, with mean and standard deviation equal to 0.232 and 0.179, respectively). As covariates in (5), we take the Gini coefficient ($x_2 = w_2$: Gini) representing the income distribution inequality (ranging from 0.28 to 0.80, with mean and standard deviation equal to 0.49 and 0.07, respectively) and the municipal human development index (MHDI) proposed by the United Nations Development Programme (UNDP) ($x_3 = w_3$: MHDI) (ranging from 0.42 to 0.86, with mean and standard deviation equal to 0.66 and 0.07, respectively). The MHDI is built on per capita income, education, and life expectancy at birth. Our interest lies in the relationship between the poverty index and the inequality and human development indexes. Since in Brazil the municipalities are grouped in 27 states, the following models for the median of the poverty index were formulated (in increasing order of complexity):

M1: quantile model

$$y_i \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_i, \varphi_i, 0.5),$$
$$\mathrm{logit}(\kappa_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \log(\varphi_i) = -\delta_1.$$

M2: quantile and dispersion model

$$y_i \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_i, \varphi_i, 0.5),$$
$$\mathrm{logit}(\kappa_i) = \boldsymbol{x}_i^T \boldsymbol{\beta}, \quad \log(\varphi_i) = -\boldsymbol{w}_i^T \boldsymbol{\delta},$$

for $i = 1, \ldots, 27$ states.
M3: random intercept quantile model

$$y_{ij} \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_{ij}, \varphi_{ij}, 0.5),$$
$$\mathrm{logit}(\kappa_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_i, \quad b_i \sim N(0, \sigma_b^2), \quad \log(\varphi_i) = -\delta_1.$$

M4: random intercepts quantile and dispersion model

$$y_{ij} \overset{\text{indep.}}{\sim} \mathrm{K}(\kappa_{ij}, \varphi_{ij}, 0.5),$$
$$\mathrm{logit}(\kappa_{ij}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + b_i, \quad b_i \sim N(0, \sigma_b^2),$$
$$\log(\varphi_{ij}) = -\boldsymbol{w}_{ij}^T \boldsymbol{\delta} - d_i, \quad d_i \sim N(0, \sigma_d^2),$$

for $j = 1, \ldots, n_i$ municipalities and $i = 1, \ldots, 27$ states, with the number of municipalities ranging from 1 to 853 (median = 143, mean = 206).

M1 and M2 correspond to regression models where the covariates have an effect on the median ($q = 0.5$) and on both the median and the dispersion parameter, respectively. On the other side, M3 and M4 are obtained from M1 and M2 by introducing random effects in the intercept in order to capture within-state dependence.

We adopt the prior specifications $\boldsymbol{\beta} \sim N_3(\boldsymbol{0}, 10^4 \boldsymbol{I}_3)$, $\delta_1 \sim N(0, 10^4)$ (M1 and M3), and $\boldsymbol{\delta} \sim N_3(\boldsymbol{0}, 10^4 \boldsymbol{I}_3)$ (M2 and M4) for the regression coefficients. Furthermore, $\sigma_b^2 \sim$ Inv-$Gamma(0.01, 0.01)$ and $\sigma_d^2 \sim$ Inv-$Gamma(0.01, 0.01)$. We use the RStan package in R taking 20,000 iterations and four chains. By default, only the second half of the chains was used leading to a MCMC sample size of 40,000. The convergence analysis of the chains (not shown) provides strong indication of convergence for all fitted models.

Model comparison criteria are displayed in Table 4. Notice that there is a good agreement between the number of parameters and the effective number of parameters. We see that mixed models are preferable to models without random effects. Overall, model M4 is the one that yields the best fit to the data set. Posterior summaries for this model are shown in Table 5. In what follows, all the discussions pertain to the results from model M4.

Notice that for both components of the model (quantile and dispersion parameter), the Gini coefficient (Gini) has a positive effect while the human development index (MHDI) has a negative effect. Since both Gini and MHDI are centered and are in the same scale, we can say that MHDI has a greater impact on the location and dispersion of the poverty index than Gini. Moreover, we see that the effect of the state, measured by the variance of random intercept, is lower when compared to the effect of the selected covariates.

Looking a little bit closer at the predictions of the random intercept of the quantile component (b) in Table 5, we can split the states in three clusters of states, as follows: (i) cluster 1, with negative predictions, is formed by Santa Catarina, Rio Grande do Sul, São Paulo, Mato Grosso do Sul, Paraná, Goiás, Rio de Janeiro, Rondônia, and Mato Grosso states, (ii) cluster 3, with positive predictions, is formed by Pará, Amazonas, Alagoas, Piauí, Roraima, Bahia, Pernambuco, Paraíba, Sergipe, Tocantins, Rio Grande do Norte,

Table 5. Posterior summaries for model M4 (Mean: mean, SD: standard deviation, and 95% CI: 95% credible interval)

		Quantile ($\boldsymbol{\beta}$)			Dispersion parameter ($\boldsymbol{\delta}$)		
		Mean	SD	95% CI	Mean	SD	95% CI
	Coefficients: Intercept	-1.45	0.05	(-1.55, -1.34)	-2.00	0.05	(-2.10, -1.91)
	Gini	3.86	0.09	(3.68, 4.04)	0.62	0.18	(0.27, 0.96)
	MHDI	-11.18	0.13	(-11.43, -10.94)	-2.25	0.24	(-2.72, -1.77)
	σ^2	0.07	0.02	(0.04, 0.13)	0.05	0.02	(0.03, 0.09)
Random intercepts (b and d)	States						
	Santa Catarina	-0.46	0.06	(-0.58, -0.34)	0.30	0.07	(0.18, 0.43)
	Rio Grande do Sul	-0.39	0.06	(-0.51, -0.28)	0.31	0.06	(0.20, 0.43)
	São Paulo	-0.38	0.06	(-0.49, -0.26)	0.28	0.06	(0.17, 0.39)
	Mato Grosso do Sul	-0.31	0.06	(-0.44, -0.18)	-0.04	0.09	(-0.21, 0.13)
	Paraná	-0.29	0.06	(-0.40, -0.18)	0.13	0.06	(0.01, 0.25)
	Goiás	-0.26	0.06	(-0.38, -0.14)	0.18	0.06	(0.06, 0.31)
	Rio de Janeiro	-0.20	0.06	(-0.31, -0.08)	-0.37	0.09	(-0.54, -0.20)
	Rondônia	-0.19	0.06	(-0.31, -0.06)	-0.19	0.10	(-0.37, 0.01)
	Mato Grosso	-0.14	0.06	(-0.26, -0.02)	0.06	0.07	(-0.08, 0.21)
	Distrito Federal	-0.08	0.22	(-0.50, 0.35)	-0.03	0.22	(-0.47, 0.41)
	Acre	-0.07	0.08	(-0.23, 0.09)	0.06	0.13	(-0.18, 0.32)
	Minas Gerais	-0.07	0.06	(-0.18, 0.04)	0.24	0.05	(0.13, 0.34)
	Espírito Santo	-0.01	0.06	(-0.13, 0.10)	-0.27	0.09	(-0.45, -0.09)
	Pará	0.01	0.06	(-0.11, 0.13)	0.20	0.07	(0.06, 0.35)
	Amazonas	0.05	0.07	(-0.09, 0.20)	0.33	0.09	(0.16, 0.51)
	Alagoas	0.07	0.06	(-0.04, 0.19)	-0.14	0.08	(-0.30, 0.02)
	Piauí	0.13	0.06	(0.02, 0.24)	-0.05	0.06	(-0.18, 0.07)
	Roraima	0.20	0.09	(0.02, 0.37)	0.06	0.16	(-0.23, 0.38)
	Bahia	0.15	0.05	(0.04, 0.25)	-0.13	0.06	(-0.24, -0.02)
	Pernambuco	0.17	0.06	(0.06, 0.28)	-0.05	0.07	(-0.18, 0.09)
	Paraíba	0.18	0.06	(0.07, 0.29)	-0.11	0.06	(-0.23, 0.01)
	Sergipe	0.20	0.06	(0.09, 0.32)	-0.27	0.09	(-0.44, -0.08)
	Tocantins	0.24	0.06	(0.13, 0.35)	-0.08	0.07	(-0.22, 0.06)
	Rio Grande do Norte	0.25	0.06	(0.14, 0.36)	-0.17	0.07	(-0.31, -0.04)
	Amapá	0.33	0.08	(0.17, 0.49)	-0.04	0.15	(-0.32, 0.26)
	Maranhão	0.29	0.06	(0.18, 0.40)	0.02	0.06	(-0.11, 0.14)
	Ceará	0.55	0.06	(0.44, 0.66)	-0.25	0.07	(-0.38, -0.12)

Amapá, Maranhão, and Ceará states (all of them located at the north part of the country), and (iii) cluster 2, with predictions around 0, is formed by Distrito Federal, Acre, Minas Gerais, and Espírito Santo states. In other words, cluster 1 is formed by states with low levels of poverty, whereas cluster 3 is composed by states with high levels of poverty. Taking into account the states in clusters 1 and 3 and the literature on gross domestic product studies [see, for example, 20], we can say that our approach clearly identifies the behaviour of the states with respect to their positions in the poverty index ranking.

With respect to the predictions of the random intercept of the dispersion parameter component, there is no a clear-cut pattern.

7. FINAL COMMENTS

In this paper a new quantile parametric mixed regression model for bounded response variables is proposed. Our model is built on the distribution introduced by [27]. A reparameterization of this distribution in terms of a given quantile and the precision parameter enables us to link any quantile of the distribution to covariates. Inference is based on a Bayesian approach with proper (and vague) prior distributions.

Since the posterior distribution is not amenable to analytical treatment, we rely on Markov Chain Monte Carlo methods. Results from a simulation study shows that even in case of extreme quantiles (0.25 and 0.75), our Bayesian proposal yields estimators with a good performance. Furthermore, two real data sets are analyzed using the proposed methodologies. Besides the study in Section 5.1, we ran the Gibbs sampler with different values of the hyperparameters in (8). The differences in the results are not important when compared with our conclusions in Sections 5 and 6.

We envision future works exploring different link functions in (5), possibly asymmetric ones. Bayesian diagnostic tools [31] are also of interest. Models for zero-inflated and one-inflated data sets [14] and with a spatial component [4] would extend the present paper, as well as extensions to censored data [38] with a bounded response variable. Another point of interest for future research is a Bayesian solution to deal with the quantile crossing problem.

Finally, the model and inferential methods can be easily implemented using standard software, as can be seen in Appendix A, and then used for data modeling.

ACKNOWLEDGMENTS

We would like to thank a Co-Editor and two reviewers whose comments lead to an improved presentation of our paper. This work was partially funded by the Dirección de Gestión de la Investigación at PUCP, Perú, through the grants DGI-2014-0017/0070 and DGI-2014-0077/0065. The third author acknowledges partial support from CNPq, Brazil.

APPENDIX A. RSTAN CODE

The code for the M4 model in Table 4 is given below. We adopted the logit link for the quantile parameter and the log link for the precision parameter. The hyperparameters in the prior distributions for all the parameters need to be specified by the user, as well as the probability q corresponding to the quantile of interest.

```
data {
int<lower = 0> n; // number of observations
int<lower = 0> M; // number of subjects
real<lower = 0,upper = 1> y[n]; // response variable
real x1[n]; // covariate
real x2[n]; // covariate
int<lower = 0> id[n]; // id variable
real<lower = 0,upper = 1> q; // quantile
}
parameters {
real delta0;
real delta1;
real delta2;
real beta0;
real beta1;
real beta2;
real<lower = 0> sigma2b;
real<lower = 0> sigma2d;
real bib[M];
real bid[M];
}
transformed parameters {
real<lower = 0> sigmab;
real<lower = 0> sigmad;
sigmab <- sqrt(sigma2b);
sigmad <- sqrt(sigma2d);
}
model {
real kappa[n];
real phi[n];
real a[n];
real b[n];
for(j in 1:M){
bib[j] ~ normal(0, sigmab);
bid[j] ~ normal(0, sigmad);
}
beta0 ~ normal(0, 100);
beta1 ~ normal(0, 100);
beta2 ~ normal(0, 100);
delta0 ~ normal(0, 100);
delta1 ~ normal(0, 100);
delta2 ~ normal(0, 100);
sigma2b ~ inv_gamma(0.01, 0.01);
sigma2d ~ inv_gamma(0.01, 0.01);
for(i in 1:n){
phi[i] <- exp(-delta0 - delta1 * x1[i] - delta2 * x2[i]
- bid[id[i]]);
b[i] <- log(1 - q) / log(1 - exp(-phi[i]));
kappa[i] <- inv_logit(beta0 + beta1 * x1[i] + beta2 * x2[i]
+ bib[id[i]]);
a[i] <- -phi[i] / log(kappa[i]);
increment_log_prob(log(a[i]) + log(b[i])
+ (a[i] - 1) * log(y[i]) +
(b[i] - 1) * log1m(pow(y[i], a[i])));
}
}
```

Received 19 August 2015

REFERENCES

[1] Bayes, C. L., Bazán, J. L. and García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis* **7** 771–796. MR3000016

[2] Bondell, H. D., Reich, B. J. and Wang, H. (2010). Noncrossing quantile regression curve estimation. *Biometrika* **97** 825–838. MR2746154

[3] Branscum, A. J., Johnson, W. O. and Thurmond, M. C. (2007). Bayesian beta regression: Application to household data and genetic distance between foot-and-mouth disease viruses. *Australian & New Zealand Journal of Statistics* **49** 287–301. MR2405396

[4] Cepeda-Cuervo, E. and Núñez-Antón, V. (2013). Spatial double generalized beta regression models. Extensions and application to study quality of education in Colombia. *Journal of Educational and Behavioral Statistics* **38** 604–628.

[5] Chen, M. H., Shao, Q. M. and Ibrahim, J. G. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer, New York. MR1742311

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

[7] Dehbi, H. M., Cortina-Borja, M. and Geraci, M. (2016). Aranda-Ordaz quantile regression for student performance assessment. *Journal of Applied Statistics* **43** 58–71. MR3437024

[8] Espinheira, P., Ferrari, S. L. P. and Cribari-Neto, F. (2008). Influence diagnostics in beta regression. *Computational Statistics & Data Analysis* **52** 4417–4431. MR2432471

[9] Espinheira, P., Ferrari, S. L. P. and Cribari-Neto, F. (2008). On beta regression residuals. *Journal of Applied Statistics* **35** 407–419. MR2420486

[10] Estrada, A., Bazán, J. L. and Aparicio, A. (2010). Un estudio comparativo de las actitudes hacia la estadística en profesores españoles y peruanos (in Spanish). *Unión – Revista Iberoamericana de Educación Matemática* **24** 45–56.

[11] Ferrari, S. L. P. and Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31** 799–815. MR2095753

[12] Ferrari, S. L. P., Espinheira, P. L. and Cribari, F. (2011). Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica* **65** 337–351. MR2857878

[13] Figueroa-Zuñiga, J. I., Arellano-Valle, R. B. and Ferrari, S. L. P. (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics and Data Analysis* **61** 137–147. MR3063006

[14] Galvis, D. M., Bandyopadhyay, D. and Lachos, V. H. (2014). Augmented mixed beta regression models for periodontal proportion data. *Statistics in Medicine* **33** 3759–3771. MR3260658

[15] Gelman, A., Hwang, J. and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing* **24** 997–1016. MR3253850

[16] Geraci, M. and Bottai, M. (2007). Quantile regression for longitudinal data using asymmetric Laplace distribution. *Biostatistics* **8** 140–154.

[17] Geraci, M. and Bottai, M. (2014). Linear quantile mixed models. *Statistics and Computing* **24** 461–479. MR3192268

[18] Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Proceedings of the Fourth Valencia International Meeting, Peñíscola, 1991* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 169–193. Oxford, New York. MR1380276

[19] IBGE-Brazil (2014). Pesquisa Nacional por Amostra de Domicílios, PNAD: Síntese de indicadores (in Portuguese). http://www.ibge.gov.br. Fundação Instituto Brasileiro de Geografia e Estatística, Departamento de Emprego e Rendimento, Brazil.

[20] IBGE-Brazil (2014). Regional Accounts 2012. http://www.ibge.gov.br/english/estatistica/economia/contasregionais/2012/default.shtm. November 2014, retrieved November 24, 2014.

[21] Jones, M. C. (2009). Kumaraswamy's distribution: A beta-type distribution with some tractability advantages. *Statistical Methodology* **6** 70–81. MR2655540

[22] Kelly, G. O., Garabed, R., Branscum, A., Perez, A. and Thurmond, M. (2007). Prediction model for sequence variation in the glycoprotein gene of infectious hematopoietic necrosis virus in California, USA. *Diseases of Aquatic Organisms* **78** 97–104.

[23] Kieschnick, R. and McCullough, B. D. (2003). Regression analysis of variates observed on (0,1): Percentages, proportions and fractions. *Statistical Modelling* **3** 193–213. MR2005473

[24] Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica* **46** 33–50. MR0474644

[25] Koenker, R. and Hallock, K. F. (2001). Quantile regression: An introduction. *Journal of Economic Perspectives* **15** 143–156.

[26] Kozumi, H. and Kobayashi, G. (2011). Gibbs sampling methods for Bayesian quantile regression. *Journal of Statistical Computation and Simulation* **81** 1565–1578. MR2851270

[27] Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random process. *Journal of Hidrology* **46** 79–88.

[28] Lachos, V. H., Chen, M. H., Abanto-Valle, C. A. and Azevedo, C. L. N. (2015). Quantile regression for censored mixed-effects models with applications to HIV studies. *Statistics and Its Interface* **8** 203–215. MR3322167

[29] Lunn, D. J., Thomas, A., Best, N. and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing* **10** 325–337.

[30] Mitnik, P. A. and Baek, S. (2013). The Kumaraswamy distribution: Median-dispersion re-parameterizations for regression modeling and simulation-based estimation. *Statistical Papers* **54** 177–192. MR3016961

[31] Peng, F. and Dey, D. K. (1995). Bayesian analysis of outlier problems using divergence measures. *Canadian Journal of Statistics* **23** 199–213.

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

[33] Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society. Series B* **64** 583–639. MR1979380

[34] R Core Team (2015). R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria.

[35] Stan Development Team (2014). RStan: The R interface to Stan, version 2.5.0.

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

[37] Wallis, E., Mac Nally, R. and Lake, S. (2009). Do tributaries affect loads and fluxes of particulate organic matter, inorganic sediment and wood? Patterns in an upland river basin in south-eastern Australia. *Hydrobiologia* **636** 307–317.

[38] Wang, H., Zhou, J., and Li, Y. (2013). Variable selection for censored quantile regression. *Statistica Sinica* **23** 145–167. MR3076162

[39] Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* **11** 3571–3594. MR2756194

[40] Wu, L. (2009). *Mixed Effects Models for Complex Data*. Chapman and Hall/CRC. MR2598844

[41] Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. *Statistics and Probability Letters* **54** 437–447. MR1861390

[42] Yu, K. M., Chen, C. W. S., Reed, C. and Dunson, D. B. (2013). Bayesian variable selection in quantile regression. *Statistics and Its Interface* **6** 261–274. MR3066690

[43] Zhang, P., Qiu, Z., Fu, Y. and Song, P. X. K. (2009). Robust transformation mixed-effects models for longitudinal continuous proportional data. *The Canadian Journal of Statistics* **37** 266–281. MR2531831

Cristian L. Bayes
Departamento de Ciencias
Pontificia Universidad Católica del Perú
Lima
Perú
E-mail address: cbayes@pucp.edu.pe

Jorge L. Bazán
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, SP
Brazil
E-mail address: jlbazan@icmc.usp.br

Mário de Castro
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
São Carlos, SP
Brazil
E-mail address: mcastro@icmc.usp.br