

# Parametric quantile regression models for fitting double bounded response with application to COVID-19 mortality rate data

Diego I. Gallardo<sup>1,\*</sup>, Marcelo Bourguignon<sup>2</sup>,  
Yolanda M. Gómez<sup>1</sup> and Christian Caamaño-Carrillo<sup>3</sup>

<sup>1</sup>Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile

<sup>2</sup>Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Natal, Brazil

<sup>3</sup>Departamento de Estadística, Facultad de Ciencias, Universidad del Bío-Bío, Concepción, Chile

## Abstract

In this paper, we develop two fully parametric quantile regression models, based on power Johnson  $S_B$  distribution Cancho et al. [Statistical Methods in Medical Research, 2020], for modeling unit interval response at different quantiles. In particular, the conditional distribution is modelled by the power Johnson SB distribution. The maximum likelihood method is employed to estimate the model parameters. Simulation studies are conducted to evaluate the performance of the maximum likelihood estimators in finite samples. Furthermore, we discuss residuals and influence diagnostic tools. The effectiveness of our proposals is illustrated with two data set given by the mortality rate of COVID-19 in different countries.

**Keywords:** COVID-19; Parametric quantile regression; Power Johnson  $S_B$  distribution; Proportion.

## 1 Introduction

The most commonly employed two-parameter distribution for modeling doubly bounded random variables on the unit interval is the beta distribution. In order to accommodate explanatory variable in the modeling, Ferrari and Cribari-Neto (2004) introduced the beta regression model based on a parameterization of the beta distribution in terms of the mean and precision parameters. A substantial number of practical and theoretical works have focused on the use of the mean reparameterized beta distribution as an integral of the model. For example, see Ospina and Ferrari (2008), Bayes et al. (2012) and Migliorati et al. (2018). However, there are limitations of the conditional mean models. For example, in an assymmetric distribution, or in the presence of outliers, the mean is pulled in the direction of the tail, making it a less representative measure of central tendency.

---

\*Corresponding author: Diego I. Gallardo. Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Copiapó, Chile. Email: diego.gallardo@uda.cl

Quantile regression, introduced by Koenker and Bassett (1978), is a methodology for understanding the conditional distribution of a response variable given the values of some covariates at different levels (quantiles), thus providing users with a more complete picture. In particular, several authors (Su, 2015; Lemonte and Moreno-Arenas, 2020) highlighted the robustness to outliers connected with quantile regression models. Furthermore, if the conditional dependent variable is skewed, the quantiles may be more appropriate when compared with the mean (Mazucheli et al., 2020).

However, parametric quantile regression models for limited range response variables has not received much attention in the literature. Lemonte and Bazán (2016) introduced a new class of distributions named the generalized Johnson  $S_B$  with bounded support on the basis of the symmetric family of distributions. In particular, Lemonte and Bazán (2016) provided the median re-parameterizations of the Johnson  $S_B$  distribution (Johnson, 1949) that facilitates its use in a regression setting. Unlike the beta regression, the median in the re-parameterized Johnson  $S_B$  distribution is related to a linear predictor. Cancho et al. (2020) generalized the Johnson  $S_B$  model to a general class of distributions. The authors introduced an extra parameter to model the shape of the Johnson  $S_B$  distribution, and studied a quantile regression model for limited range response variables. However, they consider the model only based on the normal distribution. Other quantile regression models for limited range response variables are presented in Bayes et al. (2017), Mazucheli et al. (2020) and Lemonte and Moreno-Arenas (2020).

In this paper, we formulate two rich classes of parametric quantile regression models for a bounded response, where the response variable is power Johnson  $S_B$  distributed (Cancho et al., 2020) using a new parametrization of this distribution that is indexed by quantile (not only for median regression) and shape parameters. The estimation and inference for the proposed quantile regression models can be carried out based on the likelihood paradigm (parametric approach). Also, we give full diagnostic tools for detecting possible outliers and discuss a type of residuals. The main motivations for these new parametric quantile regression models are fourfold: (i) the Johnson  $S_B$  and power Johnson  $S_B$  regression models are themselves special cases of the proposed quantile models; (ii) the first proposed model has a parameter which controls the shape and skewness of the distribution; (iii) the second proposed model has less computational cost; and (iv) we considered the model based on several models (logistic, Cauchy and normal) and several link functions.

The article is organized as follows. In Section 2, we construct two new quantile regression models for bounded response variables. Estimation, residuals and diagnostic measures are discussed in Section 3. Section 4 discusses some simulation results for the maximum likelihood (ML) estimation method. The effectiveness of our models is illustrated in Section 5 by using the mortality rate of COVID-19 in different countries. Final comments are presented in Section 6. This paper contains an additional application related to the reproductive activity of the anchoveta in Chile in a Supplementary Material.

## 2 The generalized Johnson $S_B$ distribution

Lemonte and Bazán (2016) introduced a new class of distributions named the generalized Johnson  $S_B$  (“GJS” for short) distribution. The class is defined by the transformation  $Y = Q^{-1}((X - \gamma)/\delta) \in (0, 1)$ , where  $\gamma \in \mathbb{R}$ ,  $\delta > 0$ ,  $Q(y) = \log(y/(1 - y))$  is the logit function (also representing the

quantile function for the standard logistic distribution) and  $X \sim S(0, 1; g)$ , i.e., the symmetrical family of distributions with pdf given by  $g(w)$ ,  $w \in \mathbb{R}$ , where  $g$  is a function such as  $g : \mathbb{R} \rightarrow [0, \infty)$ . Considering the reparametrization  $\gamma = -\delta Q(\xi)$ , the cdf of the GJS is given by

$$F(y; \xi, \delta) = \int_{-\infty}^{\delta[Q(y)-Q(\xi)]} g(u)du, \quad y, \xi \in (0, 1).$$

As  $F(\xi; \xi, \delta) = 1/2$ , the parameter  $\xi$  represents directly the median of the distribution. Additionally, the authors interpret  $\delta$  as a dispersion parameter. Therefore, a regression structure on  $\xi$  and  $\delta$  is studied by the authors, providing a rich class of median regression model with varying dispersion. Cancho et al. (2020) considered  $g(u) = \phi(u)$  (where  $\phi(\cdot)$  denotes the pdf of the standard normal model) and the power model transformation (Lehmann, 1953; Durrans, 1992) to extend this class of models (named as PJSB), which cdf is given by

$$F(y; \alpha, \gamma, \delta) = [\Phi(\gamma + \delta Q(y))]^\alpha, \quad y \in (0, 1), \alpha, \delta > 0, \gamma \in \mathbb{R}.$$

Besides the logistic model, the authors also considers  $Q(y)$  as the quantile function for the normal, Cauchy, Gumbel and reverse Gumbel models. Thus, the pdf of the PJSB model is

$$f(y; \gamma, \delta, \alpha) = \delta \alpha [\Phi(\gamma + \delta Q(y))]^{\alpha-1} \phi(\gamma + \delta Q(y)) \left| \frac{dQ(y)}{dy} \right|, \quad y \in (0, 1).$$

Defining  $x_q = \Phi^{-1}(q^{1/\alpha})$ , the authors considered the reparametrization  $\psi = Q^{-1}\left(\frac{x_{0.5}(\alpha) - \gamma}{\delta}\right)$ , which represents the median of the PJSB distribution (for any  $Q(\cdot)$  quantile function). As  $\gamma = x_{0.5}(\alpha) - \delta Q(\psi)$ , the pdf of the PJSB can be expressed as

$$f(y; \psi, \delta, \alpha) = \delta \alpha [\Phi(\delta[Q(y)-Q(\psi)] + x_{0.5}(\alpha))]^{\alpha-1} \phi(\delta[Q(y)-Q(\psi)] + x_{0.5}(\alpha)) \left| \frac{dQ(y)}{dy} \right|, \quad y \in (0, 1).$$

The authors proposed a regression model for  $\psi$  and  $\delta$  in this model. However, this model can be restrictive because considers the only normal distribution. For this reason, we consider the power model transformation of Lehmann (1953); Durrans (1992) for the GJS distribution of Lemonte and Bazán (2016), say the power generalized Johnson  $S_B$  (PGJSB) distribution, with cdf given by

$$F(y; \xi, \delta, \alpha) = \left( \int_{-\infty}^{\delta[Q(y)-Q(\xi)]} g(u)du \right)^\alpha = [G(\delta[Q(y) - Q(\xi)])]^\alpha = [G(\gamma + \delta Q(y))]^\alpha, \quad y \in (0, 1), \quad (1)$$

and pdf given by

$$f(y; \gamma, \delta, \alpha) = \delta \alpha [G(\gamma + \delta Q(y))]^{\alpha-1} g(\gamma + \delta Q(y)) \left| \frac{dQ(y)}{dy} \right|, \quad y \in (0, 1).$$

where  $G$  is the cdf related to  $g$ . Evidently, for  $G = \Phi$ , we recover the model in Cancho et al. (2020). However, we are interested in model a general quantile, say  $q$ , not only the median. In this work, we discuss two ways to model the  $100 \times q$ th quantile considering the PGJSB model.

1. We note that  $\psi = Q^{-1}\left(\frac{x_q^*(\alpha) - \gamma}{\delta}\right)$  is the  $100 \times q$ th quantile for the PGJSB model, where  $x_q^*(\alpha) = G^{-1}(q^{1/\alpha})$ . Based on this idea, we also can reparametrize the model noting defining  $\gamma = x_q^*(\alpha) - \delta Q(\psi)$ . The pdf for this reparametrization is

$$f(y; \psi, \delta, \alpha) = \delta \alpha [G(\delta[Q(y) - Q(\psi)] + x_q^*(\alpha))]^{\alpha-1} g(\delta[Q(y) - Q(\psi)] + x_q^*(\alpha)) \left| \frac{dQ(y)}{dy} \right|, \quad y \in (0, 1). \quad (2)$$

In this work, we will refer to this specific parametrization as  $\text{RPGJSB1}_q(\psi, \delta, \alpha)$ .

2. Despite the nature of  $\alpha$  is to be a parameter, we can consider  $\alpha(q) = -\log(q)/\log(2)$ ,  $q \in (0, 1)$  as fixed. With this definition, the cdf in (1) evaluated in  $\xi$  is given by  $F(\xi; \xi, \delta) = (1/2)^{\alpha(q)} = q$ . Therefore, fixing  $\alpha(q) = -\log(q)/\log(2)$ ,  $q \in (0, 1)$ , we have that  $\xi$  represents the  $100 \times q$ th quantile of the distribution and similarly to the work of Lemonte and Bazán (2016),  $\delta$  also can be interpreted as a dispersion parameter. We will refer to this parametrization as  $\text{RPGJSB2}_q(\xi, \delta)$ .

In both cases, the  $\text{RPGJSB1}_q$  and  $\text{RPGJSB2}_q$  models can be used to define a rich class to perform quantile regression for data in the  $(0, 1)$  interval (not only for median regression). The advantage of  $\text{RPGJSB1}_q$  model is that  $\alpha$ , for a fixed quantile  $\psi$ , controls the shape of the distribution (different  $\alpha$ 's produce different shapes). However, in this parametrization the shape of the model also depends on  $\psi$ . As we will perform regression on  $\psi$ , this indicates that the shape of the quantile depend on the covariates. A second problem is the computational costs, because evaluate 2 can be hard to compute for some combinations of  $g$  and  $Q$ . On the other hand, the advantage of  $\text{RPGJSB2}_q$  is the parsimonious (because one parameters is not estimated) and the reduction in the computational costs, because  $\alpha$  is considered fixed. However, in the  $\text{RPGJSB2}_q$  model the shape of the distribution is maintained (because the model belongs to the location-scale family of distributions) because such shape is ‘‘fixed’’.

Figure 1 shows the density function for the  $\text{RPGJSB1}_q(\psi, \delta = 1, \alpha)$  model with logit link and  $G = \Phi$  under different combinations of  $q$ ,  $\psi$  and  $\alpha$ . From Figure 1, note that the proposed model is very flexible since its density can assume different shapes.

### 3 Inference and its associated diagnostic analysis

In this section, we discuss some aspects related to the inference, residuals and diagnostic analysis of the  $\text{RPGJSB1}_q$  and  $\text{RPGJSB2}_q$  quantile regression models.

#### 3.1 Inference

Suppose the  $100 \times q$ th quantile  $\psi$  for the  $\text{RPGJSB1}_q$  model and the dispersion parameter  $\delta$  satisfies the following functional relations

$$Q(\psi_i) = \eta_{1i} = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \log(\delta_i) = \eta_{2i} = \mathbf{z}_i^\top \boldsymbol{\nu}, \quad (3)$$

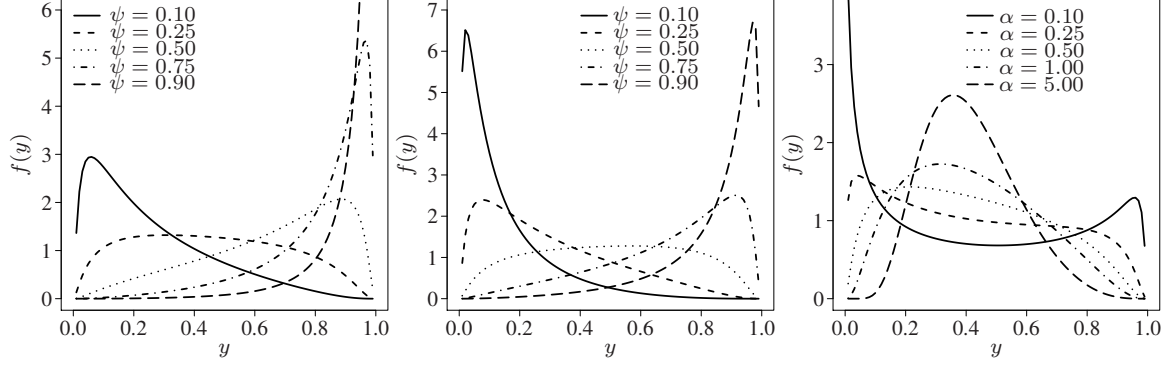


Figure 1: Pdf for  $\text{RPGJSB1}_q(\psi, \delta = 1, \alpha)$  model with logit link and  $G = \Phi$ . Left panel:  $q = 0.25$ ,  $\alpha = 0.5$  and varying  $\psi$ ; center panel:  $q = 0.5$ ,  $\alpha = 0.5$  and varying  $\psi$ ; right panel:  $q = 0.5$ ,  $\psi = 0.4$  and varying  $\alpha$ .

or

$$Q(\xi_i) = \eta_{1i} = \mathbf{x}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \log(\delta_i) = \eta_{2i} = \mathbf{z}_i^\top \boldsymbol{\nu}, \quad (4)$$

for the  $\text{RPGJSB2}_q$  model, where  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$  and  $\boldsymbol{\nu} = (\nu_1, \dots, \nu_r)^\top$  are vectors of unknown regression coefficients which are assumed to be functionally independent,  $\boldsymbol{\beta} \in \mathbb{R}^p$  and  $\boldsymbol{\nu} \in \mathbb{R}^r$ , with  $p + r < n$ ,  $\eta_{1i}$  and  $\eta_{2i}$  are the linear predictors, and  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^\top$  and  $\mathbf{z}_i = (z_{i1}, \dots, z_{ir})^\top$  are observations on  $p$  and  $r$  known regressors, for  $i = 1, \dots, n$ . Furthermore, we assume that the covariate matrices  $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^\top$  and  $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_n)^\top$  have rank  $p$  and  $r$ , respectively. The log-likelihood function for the  $\text{RPGJSB1}_q$  model is given by

$$\begin{aligned} \ell_1(\boldsymbol{\theta}) = \sum_{i=1}^n \left\{ \log(\delta_i) + \log(\alpha) + (\alpha - 1) \log \left\{ G \left( \delta_i [Q(y_i) - Q(\psi_i)] + x_q^*(\alpha) \right) \right\} \right. \\ \left. \log \left\{ g \left( \delta_i [Q(y_i) - Q(\psi_i)] + x_q^*(\alpha) \right) \right\} + \log \left| \frac{dQ(y_i)}{dy_i} \right| \right\}, \end{aligned} \quad (5)$$

whereas for the  $\text{RPGJSB2}_q$  is given by

$$\begin{aligned} \ell_2(\boldsymbol{\theta}) = \sum_{i=1}^n \left\{ \log(\delta_i) + \log(\alpha) + (\alpha - 1) \log \left[ G \left( \delta_i [Q(y_i) - Q(\xi_i)] \right) \right] \right. \\ \left. \log \left\{ g \left( \delta_i [Q(y_i) - Q(\xi_i)] \right) \right\} + \log \left| \frac{dQ(y_i)}{dy_i} \right| \right\}. \end{aligned} \quad (6)$$

Note that  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\nu}^\top, \alpha)$  and  $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{\nu}^\top)$  is the vector of parameters for the  $\text{RPGJSB1}_q$  and  $\text{RPGJSB2}_q$  models, respectively. The ML estimator of  $\boldsymbol{\theta}$ , say  $\hat{\boldsymbol{\theta}}$ , is obtained maximizing equation (5) or (6), depending on the considered model are presented in Section . We considered the maximization procedure based on the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method initialized with a vector of zeros. To validate a solution, we checked: i) If the convergence is attached and; ii) if the determinant of the hessian such matrix is positive. If the two conditions are not satisfied, we rerun the procedure based initialized with a random vector generated by independent standard normal vari-

ables until i) and ii) are satisfied. Under usual regularity conditions (see Cox and Hinkley, 1974)  $\boldsymbol{\theta}$  is consistent. Moreover,

$$\boldsymbol{\nu}^{-1}(\widehat{\boldsymbol{\theta}}) \left[ \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right] \xrightarrow{\mathcal{D}} N_{p+r}(\mathbf{0}_{p+r}, \mathbf{I}_{p+r}), \quad \text{as } n \rightarrow +\infty,$$

where  $\boldsymbol{\nu}(\widehat{\boldsymbol{\theta}}) = -\partial^2 \ell_l(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top \big|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$  is minus the estimated hessian matrix for the RPGJSB1<sub>q</sub> ( $l = 1$ ) and RPGJSB2<sub>q</sub> ( $l = 2$ ) models, respectively.

## 3.2 Residuals

In order to assess if the posited model is correct, we will consider the randomized quantile residuals (RQRs) proposed by Dunn and Smyth (1996). For the RPGJSB1<sub>q</sub> model, such residuals are given by

$$\widehat{r}_i = \Phi^{-1} \left( [G(\widehat{\delta}_i [Q(y_i) - Q(\widehat{\psi}_i)] + x_q^*(\widehat{\alpha}))]^{\widehat{\alpha}} \right), \quad i = 1, \dots, n,$$

whereas for the RPGJSB2<sub>q</sub> model, the RQRs are given by

$$\widehat{r}_i = \Phi^{-1} \left( [G(\widehat{\delta}_i [Q(y_i) - Q(\widehat{\xi}_i)])]^\alpha(q) \right), \quad i = 1, \dots, n.$$

$\widehat{\delta}_i$ ,  $\widehat{\xi}_i$  and  $\widehat{\psi}_i$ ,  $i = 1, \dots, n$ , correspond to the expressions in equations (3) and (4) evaluated in  $\widehat{\boldsymbol{\beta}}$  and  $\widehat{\boldsymbol{\nu}}$ , for each model, respectively. If the model is correctly specified, the distribution of  $\widehat{r}_1, \dots, \widehat{r}_n$  is standard normal, which can be validated considering different normality tests, such as Kolmogorov-Smirnov (KS), Shapiro-Wilks (SW), Anderson-Darling (AD) and the Cramér-Von-Mises (CVM) tests. See Yap and Sim (2011) for a discussion about such tests.

## 3.3 Local influence

The local influence method suggested by Cook (1986) evaluates the simultaneous effect of observations on the ML estimator without removing it from the data set, based on the curvature of the plane of the log-likelihood function. Consider  $\ell_1(\boldsymbol{\theta}_1; \boldsymbol{w})$  and  $\ell_2(\boldsymbol{\theta}_2; \boldsymbol{w})$  the log-likelihood functions corresponding to the RPGJSB1<sub>q</sub> and RPGJSB2<sub>q</sub> models, respectively, but now perturbed by  $\boldsymbol{w}$ , a vector of perturbations.  $\boldsymbol{w}$  belongs to a subset  $\Omega \in \mathbb{R}^n$  and  $\boldsymbol{w}_0$  is a non-perturbed  $n \times 1$  vector, such that  $\ell_l(\boldsymbol{\theta}; \boldsymbol{w}_0) = \ell_l(\boldsymbol{\theta})$ , for all  $\boldsymbol{\theta}$ ,  $l = 1, 2$ . In this case, the likelihood displacement (LD) is  $LD(\boldsymbol{\theta}) = 2(\ell_l(\widehat{\boldsymbol{\theta}}) - \ell_l(\widehat{\boldsymbol{\theta}}_{\boldsymbol{w}}))$ , where  $\widehat{\boldsymbol{\theta}}_{\boldsymbol{w}}$  denotes the ML estimate of  $\boldsymbol{\theta}$  on the perturbed regression models, that is,  $\widehat{\boldsymbol{\theta}}_{\boldsymbol{w}}$  is obtained from  $\ell_l(\boldsymbol{\theta}; \boldsymbol{w})$ . Note that  $\ell_l(\boldsymbol{\theta}; \boldsymbol{w})$  can be used to assess the influence of the perturbation of the ML estimate. Cook (1986) showed that the normal curvature for  $\widehat{\boldsymbol{\theta}}$  in the direction  $\boldsymbol{d}$ , with  $\|\boldsymbol{d}\| = 1$ , is expressed as  $C_{\boldsymbol{d}}(\widehat{\boldsymbol{\theta}}) = 2|\boldsymbol{d}^\top \nabla^\top \Sigma(\widehat{\boldsymbol{\theta}})^{-1} \nabla \boldsymbol{d}|$ , where  $\nabla$  is a  $(p+r) \times n$  matrix of perturbations with elements  $\nabla_{ji} = \partial^2 \ell_l(\boldsymbol{\theta}; \boldsymbol{w}) / \partial \boldsymbol{\theta}_j \partial \boldsymbol{w}_i$ , evaluated at  $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$  and  $\boldsymbol{w} = \boldsymbol{w}_0$ , for  $j = 1, \dots, p+r$  and  $i = 1, \dots, n$ . A local influence diagnostic is generally based on index plots. For example, denoting  $\Sigma(\boldsymbol{\theta})$  the observed Fisher information matrix, the index graph of the eigenvector  $\boldsymbol{d}_{max}$  corresponding to the maximum eigenvalue of  $\boldsymbol{B}(\boldsymbol{\theta}) = -\nabla^\top \Sigma(\boldsymbol{\theta})^{-1} \nabla$ , say  $C_{\boldsymbol{d}_{max}}(\boldsymbol{\theta})$ ,

evaluated at  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ , can detect those cases that, under small perturbations, exert a strong influence on  $\text{LD}(\boldsymbol{\theta})$ . Another important direction of interest is  $\mathbf{d}_i = \mathbf{e}_{in}$ , which corresponds to the direction of the case  $i$ , where  $\mathbf{e}_{in}$  is an  $n \times 1$  vector of zeros with value equal to one at the  $i$ th position, that is,  $\{\mathbf{e}_{in}, 1 \leq i \leq n\}$  is the canonical basis of  $\mathbb{R}^n$ . In this case, the normal curvature is  $C_i(\boldsymbol{\theta}) = 2|b_{ii}|$ , where  $b_{ii}$  is the  $i$ th diagonal element of  $\mathbf{B}(\boldsymbol{\theta})$  given above, for  $i = 1, \dots, n$ , evaluated  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ . If  $C_i(\hat{\boldsymbol{\theta}}) > 2\bar{C}(\hat{\boldsymbol{\theta}})$ , where  $\bar{C}(\hat{\boldsymbol{\theta}}) = \sum_{i=1}^n C_i(\hat{\boldsymbol{\theta}})/n$ , it indicates case  $i$  as potentially influential. This procedure is called total local influence of the case  $i$  and can be carried out for  $\boldsymbol{\theta}$ ,  $\boldsymbol{\beta}$  or  $\boldsymbol{\nu}$ , which are denoted by  $C_i(\boldsymbol{\theta})$ ,  $C_i(\boldsymbol{\beta})$  and  $C_i(\boldsymbol{\nu})$ , respectively. We calculate the matrix  $\nabla$  for three different perturbation schemes, namely: case weighting perturbation, response perturbation and explanatory variable perturbation.

### 3.3.1 Perturbation of the case weights

In this case the perturbed log-likelihood function is given by  $\ell_l(\boldsymbol{\theta}; \mathbf{w}) = \sum_{i=1}^n w_i \ell_l(\boldsymbol{\theta})$  for  $\text{RPGJSB1}_q$  ( $l = 1$ ) and  $\text{RPGJSB2}_q$  ( $l = 2$ ), respectively, with  $0 \leq w_i \leq 1$ , for  $i = 1, \dots, n$ , and  $\mathbf{w}_0 = \mathbf{1}^\top$  (all-ones vector). Hence, the perturbation matrices for the  $\text{RPGJSB1}_q$  and  $\text{RPGJSB2}_q$  models are given by

$$\hat{\nabla}_1 = \begin{pmatrix} \mathbf{X}^\top \hat{\mathbf{D}}_1 \hat{\mathbf{D}}_3 \\ \mathbf{Z}^\top \hat{\mathbf{D}}_2 \hat{\mathbf{D}}_4 \end{pmatrix} \quad \text{and} \quad \hat{\nabla}_2 = \begin{pmatrix} \mathbf{X}^\top \hat{\mathbf{D}}_5 \hat{\mathbf{D}}_7 \hat{\mathbf{D}}_9 \\ \mathbf{Z}^\top \hat{\mathbf{D}}_6 \hat{\mathbf{D}}_8 \hat{\mathbf{D}}_9 \end{pmatrix},$$

respectively, with  $\mathbf{D}_1 = [a_i \iota_{ij}]$ ,  $\mathbf{D}_2 = [b_i \iota_{ij}]$ ,  $\mathbf{D}_3 = [\dot{d}_\psi \iota_{ij}]$  and  $\mathbf{D}_4 = [\dot{d}_\delta \iota_{ij}]$  where  $a_i = \partial \psi_i / \partial \eta_{i1}$  and  $b_i = \partial \delta_i / \partial \eta_{i2}$  defined from (3);  $\dot{d}_\psi = \partial \ell_1(\psi_i, \delta_i) / \partial \psi_i$ ,  $\dot{d}_\delta = \partial \ell_1(\psi_i, \delta_i) / \partial \delta_i$  defined from the  $\text{RPGJSB1}_q$  model and  $\iota_{ij}$  is the Kronecker delta for  $i, j = 1, 2, \dots, n$ . Similarly,  $\mathbf{D}_5 = [c_i \iota_{ij}]$ ,  $\mathbf{D}_6 = [d_i \iota_{ij}]$ ,  $\mathbf{D}_7 = [\dot{d}_\xi \iota_{ij}]$ ,  $\mathbf{D}_8 = [\dot{d}_\delta \iota_{ij}]$  and  $\mathbf{D}_9 = [\dot{d}_\alpha \iota_{ij}]$  where  $c_i = \partial \xi_i / \partial \eta_{i1}$  and  $d_i = \partial \delta_i / \partial \eta_{i2}$  defined from (4);  $\dot{d}_\xi = \partial \ell_2(\xi_i, \delta_i, \alpha) / \partial \xi_i$ ,  $\dot{d}_\delta = \partial \ell_2(\xi_i, \delta_i, \alpha) / \partial \delta_i$  and  $\dot{d}_\alpha = \partial \ell_2(\xi_i, \delta_i, \alpha) / \partial \alpha$  defined from the  $\text{RPGJSB2}_q$  model.

### 3.3.2 Perturbation of the response

Now consider an multiplicative perturbation of the  $i$ th response by making  $y_i(w_i) = y_i w_i s_y$ , where  $s_y$  represents a scale factor and  $w_i \in \mathbb{R}$ , for  $i = 1, \dots, n$ . Then, under the scheme of response perturbation, the log-likelihood function is given by  $\ell_1(\boldsymbol{\theta}; \mathbf{w}) = \sum_{i=1}^n \ell_1(\psi_i, \delta_i, \alpha; \mathbf{w})$  for the  $\text{RPGJSB1}_q$  model and  $\ell_2(\boldsymbol{\theta}; \mathbf{w}) = \sum_{i=1}^n \ell_2(\xi_i, \delta_i; \mathbf{w})$  for the  $\text{RPGJSB2}_q$  model, where

$$\begin{aligned} \ell_1(\psi_i, \delta_i, \alpha; \mathbf{w}) &= (\alpha - 1) \log(G(\tau_{1i})) + \log(\alpha \delta_i) + \log(g(\tau_{1i})) + \log(|w_i s_y \dot{Q}_y(y_i w_i s_y)|) \\ \ell_2(\xi_i, \delta_i; \mathbf{w}) &= (\alpha - 1) \log(G(\tau_{2i})) + \log(\alpha \delta_i) + \log(g(\tau_{2i})) + \log(|w_i s_y \dot{Q}_y(y_i w_i s_y)|) \end{aligned}$$

with  $\tau_{1i} = \delta_i(Q(y_i w_i s_y) - Q(\psi_i))$  and  $\tau_{2i} = \delta_i(Q(y_i w_i s_y) - Q(\xi_i)) + x_q^*(\alpha)$ .

The disturbance matrices of the RPGJSB1<sub>q</sub> and RPGJSB2<sub>q</sub> models here take the form

$$\widehat{\nabla}_1 = \begin{pmatrix} \mathbf{X}^\top \widehat{\mathbf{D}}_1 \widehat{\mathbf{D}}_{10} \mathbf{S} \\ \mathbf{Z}^\top \widehat{\mathbf{D}}_2 \widehat{\mathbf{D}}_{11} \mathbf{S} \end{pmatrix} \quad \text{and} \quad \widehat{\nabla}_2 = \begin{pmatrix} \mathbf{X}^\top \widehat{\mathbf{D}}_5 \widehat{\mathbf{D}}_{12} \widehat{\mathbf{D}}_{14} \mathbf{S} \\ \mathbf{Z}^\top \widehat{\mathbf{D}}_6 \widehat{\mathbf{D}}_{13} \widehat{\mathbf{D}}_{14} \mathbf{S} \end{pmatrix}$$

where  $\mathbf{S} = [s_{y\ell_{ij}}]$ , the  $i$ th element of matrices  $\mathbf{D}_{10}$  and  $\mathbf{D}_{11}$  for model RPGJSB1<sub>q</sub> and matrices  $\mathbf{D}_{12}$ ,  $\mathbf{D}_{13}$  and  $\mathbf{D}_{14}$  for model RPGJSB2<sub>q</sub> are detailed in Section A.1 of the supplementary material.

### 3.3.3 Perturbation of the predictor

Now consider an multiplicative perturbation of the  $i$ th predictor by making  $x_i(w_i) = \mathbf{x}_i^\top w_i$  and  $z_i(w_i) = \mathbf{z}_i^\top w_i$ , for  $w_i \in \mathbb{R}$ ,  $i = 1, \dots, n$ . Then, under the scheme of prediction perturbation, the log-likelihood function is given by  $\ell_1(\boldsymbol{\theta}; \mathbf{w}) = \sum_{i=1}^n \ell_1(\psi_i^*, \delta_i^*)$  for the RPGJSB1<sub>q</sub> model and  $\ell_2(\boldsymbol{\theta}; \mathbf{w}) = \sum_{i=1}^n \ell_2(\xi_i^*, \delta_i^*, \alpha)$  for the RPGJSB2<sub>q</sub> model, where  $Q(\psi_i^*) = \mathbf{x}_i^\top \boldsymbol{\beta} w_i$  and  $\delta_i^* = \exp\{\mathbf{z}_i^\top \boldsymbol{\nu} w_i\}$  for the RPGJSB1<sub>q</sub> model and  $Q(\xi_i^*) = \mathbf{x}_i^\top \boldsymbol{\beta} w_i$  and  $\delta_i^* = \exp\{\mathbf{z}_i^\top \boldsymbol{\nu} w_i\}$  for the RPGJSB2<sub>q</sub> model.

The disturbance matrices of RPGJSB1<sub>q</sub> and RPGJSB2<sub>q</sub> models here take the form

$$\widehat{\nabla}_1 = \begin{pmatrix} \mathbf{X}^\top \widehat{\mathbf{D}}_{15} \\ \mathbf{Z}^\top \widehat{\mathbf{D}}_{16} \end{pmatrix} \quad \text{and} \quad \widehat{\nabla}_2 = \begin{pmatrix} \mathbf{X}^\top \widehat{\mathbf{D}}_{17} \widehat{\mathbf{D}}_{19} \\ \mathbf{Z}^\top \widehat{\mathbf{D}}_{18} \widehat{\mathbf{D}}_{19} \end{pmatrix}$$

where the  $i$ th elements of matrices  $\mathbf{D}_{15}$  and  $\mathbf{D}_{16}$  for RPGJSB1<sub>q</sub> model and matrices  $\mathbf{D}_{17}$ ,  $\mathbf{D}_{18}$  and  $\mathbf{D}_{19}$  for RPGJSB2<sub>q</sub> model are detailed in Section A.2. of the supplementary material.

## 4 Simulation studies

In this section, we present a simulation study to assess the performance of  $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\nu}, \alpha)^\top$  under different scenarios. First, we assume that  $G$  and the link function are correctly specified. The data were drawn motivated by the scheme for the anchoveta data set presented in Section C of the supplementary material. We considered  $\mathbf{x}_i = \mathbf{z}_i$ , where both matrices includes an intercept and a covariate. Such covariates were drawn from the  $U(-5.478, -2.305)$  distribution. We considered the logistic and normal models for  $G$  and the logit and loglog link functions. The true values for parameters were considered as the estimated parameters for three values for  $q = \{0.1, 0.5, 0.9\}$ . We also considered three sample sizes: 100, 200 and 500.

As mentioned previously, to validate a solution, we checked: If the convergence is attached and if the determinant of the hessian such matrix is positive. If the two conditions are not satisfied, we rerun the procedure initialized with a random vector generated by independent standard normal variables until both conditions are satisfied. For each combination of  $G$ , link,  $q$  and sample size, we considered 5,000 replicates and in each case the estimation is performed based on the same  $G$  and link function. Based on the 10,000 replicates, we report the bias for each estimator, the standard error of the estimates ( $SE_1$ ), the mean of the estimated standard errors ( $SE_2$ ) and the 95% coverage probabilities (CP). Tables 2 and 3 summarizes such results. Note that the bias of the parameters is



Table 1: True parameters used for simulation studies.

link	$q$	logistic					normal				
		$\beta_0$	$\beta_1$	$\nu_0$	$\nu_1$	$\log(\alpha)$	$\beta_0$	$\beta_1$	$\nu_0$	$\nu_1$	$\log(\alpha)$
logit	0.1	4.9	2.6	2.2	0.4	-0.7	4.4	2.4	1.5	0.3	-1.4
	0.5	4.8	2.1	2.2	0.4	-0.7	4.6	2.1	1.5	0.3	-1.4
	0.9	4.7	1.8	2.2	0.4	-0.7	4.8	1.9	1.5	0.3	-1.4
loglog	0.1	1.3	0.8	0.8	-0.3	0.1	1.2	0.7	-0.1	-0.3	1.1
	0.5	2.1	0.9	1.0	-0.2	0.1	2.0	0.9	0.0	-0.3	1.0
	0.9	2.8	1.0	1.1	-0.2	0.1	2.8	1.0	0.1	-0.2	1.0

reduced and the terms  $SE_1$  and  $SE_2$  are closer when  $n$  is increased, suggesting that the estimators are consistent in finite samples. Additionally, when the sample size is increased the CP are closer to the nominal value used. Finally, Table 4 presents the percentage of times where the algorithm converges when is initialized with a vector of zeros. Note that the maximization procedure converged at least in 89.43% of the generated samples and such percentages are increased when the sample size is increased.

Table 2: Recovery parameters when  $G$  and the link are correctly specified (case  $G$  is the cdf of the logistic distribution).

$G$	link	$q$	parameter	bias	$SE_1$	$SE_2$	CP	bias	$SE_1$	$SE_2$	CP	bias	$SE_1$	$SE_2$	CP
logistic	logit	0.1	$\beta_0$	-0.034	0.753	0.728	0.938	-0.017	0.538	0.529	0.946	-0.007	0.345	0.339	0.946
			$\beta_1$	-0.015	0.238	0.229	0.934	-0.007	0.166	0.163	0.942	-0.003	0.104	0.102	0.947
			$\nu_0$	0.041	0.381	0.367	0.935	0.020	0.269	0.263	0.942	0.009	0.171	0.170	0.947
			$\nu_1$	-0.001	0.088	0.085	0.939	0.000	0.061	0.060	0.946	0.000	0.039	0.038	0.946
			$\log(\alpha)$	-0.004	0.355	0.331	0.947	-0.002	0.232	0.224	0.946	-0.002	0.140	0.138	0.948
		0.5	$\beta_0$	-0.017	0.485	0.472	0.941	0.001	0.322	0.319	0.946	-0.003	0.204	0.205	0.950
			$\beta_1$	-0.005	0.146	0.142	0.939	0.000	0.096	0.095	0.946	-0.001	0.061	0.061	0.949
			$\nu_0$	0.046	0.452	0.443	0.946	0.027	0.296	0.294	0.948	0.007	0.183	0.182	0.949
			$\nu_1$	0.002	0.107	0.106	0.946	0.002	0.068	0.068	0.949	0.000	0.042	0.042	0.951
			$\log(\alpha)$	0.004	0.352	0.331	0.952	-0.001	0.231	0.224	0.948	0.000	0.142	0.139	0.947
		0.9	$\beta_0$	-0.001	0.620	0.591	0.930	-0.002	0.369	0.363	0.942	0.002	0.237	0.236	0.950
			$\beta_1$	0.004	0.177	0.169	0.932	0.002	0.112	0.111	0.943	0.001	0.072	0.072	0.948
			$\nu_0$	0.060	0.461	0.443	0.943	0.024	0.289	0.283	0.943	0.007	0.184	0.182	0.946
			$\nu_1$	0.006	0.103	0.100	0.938	0.002	0.066	0.065	0.946	0.000	0.043	0.043	0.945
			$\log(\alpha)$	0.010	0.362	0.334	0.946	0.003	0.234	0.224	0.949	0.002	0.140	0.139	0.949
loglog	0.1	0.1	$\beta_0$	0.008	0.175	0.168	0.931	0.002	0.116	0.113	0.938	0.000	0.071	0.071	0.949
			$\beta_1$	0.001	0.039	0.037	0.935	0.000	0.026	0.025	0.937	0.000	0.016	0.016	0.948
			$\nu_0$	0.020	0.413	0.398	0.944	0.005	0.280	0.275	0.946	-0.001	0.165	0.167	0.950
			$\nu_1$	0.000	0.096	0.092	0.939	-0.002	0.067	0.065	0.942	-0.001	0.039	0.039	0.949
			$\log(\alpha)$	0.153	1.175	2.515	0.964	0.035	0.349	0.324	0.961	0.014	0.178	0.174	0.956
		0.5	$\beta_0$	-0.002	0.130	0.128	0.944	-0.003	0.090	0.090	0.951	0.001	0.061	0.061	0.946
			$\beta_1$	0.000	0.031	0.030	0.945	-0.001	0.021	0.021	0.949	0.000	0.014	0.014	0.947
			$\nu_0$	0.007	0.386	0.376	0.944	0.003	0.264	0.261	0.947	0.006	0.175	0.175	0.950
			$\nu_1$	-0.003	0.093	0.091	0.945	-0.002	0.063	0.062	0.949	0.000	0.041	0.041	0.950
			$\log(\alpha)$	0.143	1.070	2.042	0.965	0.041	0.306	0.290	0.962	0.012	0.177	0.174	0.951
		0.9	$\beta_0$	-0.005	0.178	0.175	0.939	-0.004	0.141	0.139	0.942	-0.002	0.082	0.082	0.947
			$\beta_1$	-0.001	0.042	0.041	0.940	-0.001	0.033	0.032	0.943	0.000	0.019	0.019	0.947
			$\nu_0$	0.012	0.387	0.374	0.940	0.010	0.296	0.288	0.944	0.004	0.174	0.173	0.949
			$\nu_1$	-0.002	0.094	0.091	0.940	0.000	0.071	0.069	0.944	0.000	0.041	0.041	0.949
			$\log(\alpha)$	0.133	0.968	1.596	0.965	0.042	0.311	0.290	0.961	0.014	0.177	0.174	0.952

## 5 Data analysis

In this section, we present a real data set application related to the mortality rate of the COVID-19 in different countries to illustrate the performance of the  $RPGJSB1_q$  and  $RPGJSB2_q$  regression models. An additional application related to the reproductive activity of the anchoveta in Chile is presented in Section C of the supplementary material.

### 5.1 COVID-19 data set

The COVID-19 pandemic has unprecedentedly affected the entire world. Specifically, it has yielded high mortality rates since its emergence in December 2019, generating a disequilibrium societal, economic, cultural and political. It has been shown by early studies that statistical analysis can be applied to COVID-19 problems to build predictive models that can assess risk factors and mortality

Table 3: Recovery parameters when  $G$  and the link are correctly specified (case  $G$  is the cdf of the normal distribution).

$G$	link	$q$	parameter	bias	$SE_1$	$SE_2$	CP	bias	$SE_1$	$SE_2$	CP	bias	$SE_1$	$SE_2$	CP
normal	logit	0.1	$\beta_0$	-0.004	0.725	0.711	0.939	0.000	0.470	0.473	0.952	-0.002	0.289	0.291	0.951
			$\beta_1$	-0.005	0.202	0.198	0.939	-0.002	0.133	0.134	0.951	-0.002	0.081	0.082	0.949
			$\nu_0$	0.912	2.171	0.730	0.847	0.243	1.024	0.450	0.946	0.045	0.270	0.248	0.954
			$\nu_1$	0.000	0.083	0.079	0.932	0.001	0.055	0.054	0.945	0.000	0.032	0.032	0.951
		$\log(\alpha)$	-1.763	4.569	1.650	0.867	-0.462	2.167	1.019	0.960	-0.082	0.612	0.565	0.956	
		0.5	$\beta_0$	-0.006	0.464	0.452	0.942	-0.005	0.330	0.324	0.945	0.002	0.196	0.194	0.946
			$\beta_1$	-0.004	0.135	0.131	0.940	-0.002	0.094	0.092	0.941	0.000	0.056	0.056	0.946
			$\nu_0$	0.944	2.251	0.703	0.841	0.215	0.966	0.450	0.949	0.040	0.281	0.250	0.952
			$\nu_1$	0.002	0.082	0.079	0.939	0.001	0.056	0.055	0.947	0.000	0.034	0.033	0.950
		$\log(\alpha)$	-1.806	4.729	1.597	0.862	-0.398	2.046	1.012	0.961	-0.071	0.625	0.564	0.954	
		0.9	$\beta_0$	-0.028	0.595	0.550	0.910	-0.001	0.393	0.375	0.934	-0.004	0.244	0.242	0.947
			$\beta_1$	-0.002	0.165	0.153	0.912	0.003	0.111	0.106	0.933	0.000	0.069	0.069	0.949
$\nu_0$	0.923		2.248	0.712	0.852	0.235	1.009	0.450	0.947	0.047	0.279	0.253	0.949		
$\nu_1$	0.006		0.088	0.084	0.937	0.001	0.057	0.055	0.941	0.001	0.035	0.035	0.949		
$\log(\alpha)$	-1.733	4.706	1.576	0.871	-0.434	2.133	1.008	0.961	-0.083	0.614	0.563	0.956			
loglog	0.1	0.1	$\beta_0$	0.005	0.156	0.152	0.935	0.005	0.115	0.114	0.942	0.001	0.070	0.069	0.946
			$\beta_1$	0.000	0.035	0.034	0.936	0.001	0.026	0.026	0.945	0.000	0.016	0.015	0.946
			$\nu_0$	0.085	0.834	0.530	0.963	0.024	0.371	0.351	0.951	0.006	0.209	0.209	0.952
			$\nu_1$	-0.004	0.077	0.076	0.942	-0.001	0.059	0.058	0.946	-0.001	0.035	0.035	0.951
		$\log(\alpha)$	1.090	23.978	3.284	0.965	0.103	1.677	1.200	0.958	0.027	0.661	0.658	0.961	
		0.5	$\beta_0$	0.002	0.116	0.114	0.942	0.000	0.084	0.083	0.946	0.000	0.054	0.053	0.948
			$\beta_1$	0.000	0.026	0.025	0.942	0.000	0.019	0.019	0.947	0.000	0.012	0.012	0.947
			$\nu_0$	0.123	0.990	0.539	0.954	0.017	0.379	0.348	0.955	0.009	0.212	0.212	0.952
			$\nu_1$	-0.004	0.082	0.079	0.939	-0.002	0.059	0.059	0.946	-0.001	0.036	0.036	0.950
		$\log(\alpha)$	0.612	16.917	3.114	0.964	0.091	1.453	1.150	0.963	0.017	0.654	0.645	0.957	
		0.9	$\beta_0$	-0.016	0.224	0.219	0.935	-0.007	0.169	0.167	0.939	-0.004	0.095	0.095	0.943
			$\beta_1$	-0.002	0.051	0.050	0.937	-0.001	0.040	0.039	0.940	-0.001	0.022	0.022	0.946
$\nu_0$	0.125		0.934	0.538	0.958	0.029	0.386	0.357	0.951	0.008	0.208	0.207	0.951		
$\nu_1$	0.000		0.083	0.080	0.940	0.000	0.061	0.060	0.948	0.000	0.035	0.034	0.952		
$\log(\alpha)$	0.428	12.877	2.696	0.964	0.088	1.443	1.192	0.957	0.034	0.657	0.647	0.961			

Table 4: Percentage of time where the maximization algorithm converges with initial value as the vector zero.

G	link	$q = 0.1$			$q = 0.5$			$q = 0.9$		
		100	200	500	100	200	500	100	200	500
logistic	logit	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	loglog	99.71	100.00	100.00	99.83	100.00	100.00	99.85	100.00	100.00
normal	logit	90.77	98.40	100.00	89.43	98.65	99.99	90.38	98.59	99.99
	loglog	99.43	99.99	100.00	99.01	99.98	100.00	99.05	99.98	100.00

rates (Ji et al., 2020; Li et al., 2020; Du et al., 2020). Also the overall mortality rate has been about 5%, while the statistics showed a rate of around 20% for senior patients (Livingston and Bucher, 2020). We consider the following information for the countries with at least 1,000 reported cases of

COVID-19 and at least 100 deaths attributed to COVID-19, totalizing 123 countries at November 3, 2020.

- `mort`: mortality rate (reported death/reported cases). Mean=0.025, Median=0.020, standard deviation=0.028, minimum=0.002 and maximum=0.291.
- `surface`: surface of the country (in km<sup>2</sup>).
- `population`: official estimated population of the country.
- `cont`: continent to which the country belongs (categorized as 1: Africa, Asia u Oceania; 2: America; 3: Europe; with 56, 28 and 39 countries, respectively).

The information was taken from the World Health Organization (WHO, 2020). It is of interest to model the mortality rate in terms of the surface and the continent of each country (previous analysis suggest that the population is not significative to model the mortality rate). Figure 2 shows the plots for  $Q(\text{mort})$  for different link functions versus the  $\log(\text{surface})$  and separated by `cont`.

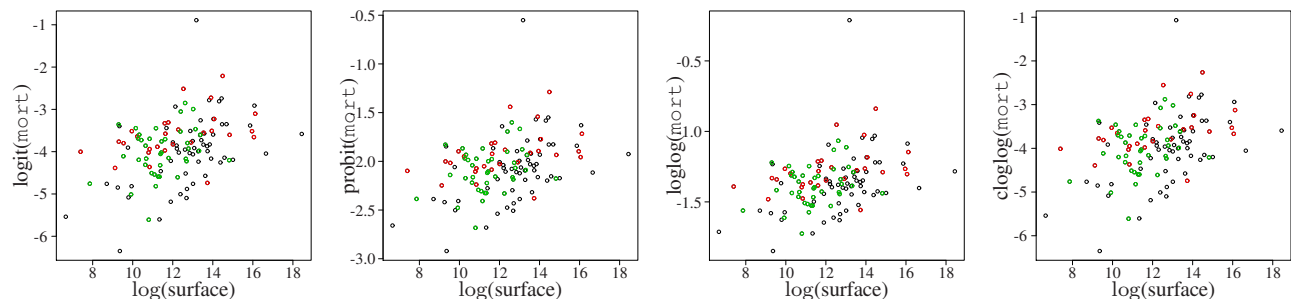


Figure 2: Descriptive plots for  $Q(\text{mort})$  versus  $\log(\text{surface})$  for different link functions: logit, probit, loglog and cloglog and separated by continent: Africa, Asia u Oceania (black), America (red) and Europe (green).

### 5.1.1 Estimation

In view of the above, we consider to model the mortality rate using  $\text{mort}_i \sim \text{RPGJSB1}_q(\psi_i, \delta_i, \alpha)$ , with

$$Q(\psi_i) = \beta_0 + \beta_1 \times \log(\text{surface}_i) + \beta_2 \times \text{America}_i + \beta_3 \times \text{Europe}_i \quad \text{and}$$

$$\log(\delta_i) = \nu_0 + \nu_1 \times \text{America}_i + \nu_2 \times \text{Europe}_i,$$

or alternatively,  $\text{mort}_i \sim \text{RPGJSB2}_q(\xi_i, \delta_i)$ , where  $Q(\xi_i) = \beta_0 + \beta_1 \times \log(\text{surface}_i) + \beta_2 \times \text{America}_i + \beta_3 \times \text{Europe}_i$  and  $\delta_i$  is modelled in the same way. In Section B.1 of the supplementary material, we present the AIC and BIC for  $q$  ranging in the set  $\{0.05, 0.10, \dots, 0.90, 0.95\}$  and the  $\text{RPGJSB1}_q$  and  $\text{RPGJSB2}_q$  models. Note that the  $\text{RPGJSB1}_q$  provides the lower AIC than the  $\text{RPGJSB2}_q$  for all the considered  $q$ . Then, hereinafter we focused in the  $\text{RPGJSB1}_q$  model, specifically where  $G$  is the cdf of the logistic model and the cloglog link (which provide the lower AIC

for all  $q$ ). Table 5 and Section B.2 of the supplementary material present the estimated parameter for such model for five selected quantiles. Also are presented the KS, SW, AD and CVM tests to check the normality of the RQRs. Note that the  $\log(\text{surface})$  is significant to model the quantile (with a nominal level of 5%) for all the considered  $q$ . This can be explained because countries with larger areas may have greater difficulties in providing medical coverage to their inhabitants in relation to countries with smaller areas. Also the parameter related to *America* is significant in both, quantile and scale parameters. However, the parameter related to *Europe* is significant to model the quantile of the mortality for COVID-19 only for small  $q$ . On the other hand, the four tests do not reject the normality assumption for the RQRs, suggesting that the  $\text{RPGJSB1}_q$  model with the logistic distribution for  $G$  and the cloglog link is appropriated to model all the considered quantiles of the mortality rate.

On the other hand, Figure 3 presented the point estimation and the 95% confidence interval (CI) for the parameters in terms of the quantile  $q$ . From 3, the intercept for the quantile increases as  $q$  increases, whereas the coefficients related to the quantile of *America* and *Europe* decreases when  $q$  is increased. Furthermore, the coefficients related to the quantile for  $\log(\text{surface})$  and the coefficients related to the scale of *America* and *Europe* remain similar for all  $q$ . Figure 4 presented the estimated quantiles 0.05, 0.25, 0.50, 0.75 and 0.95 for the mortality rate for different values of  $\log(\text{surface})$ .

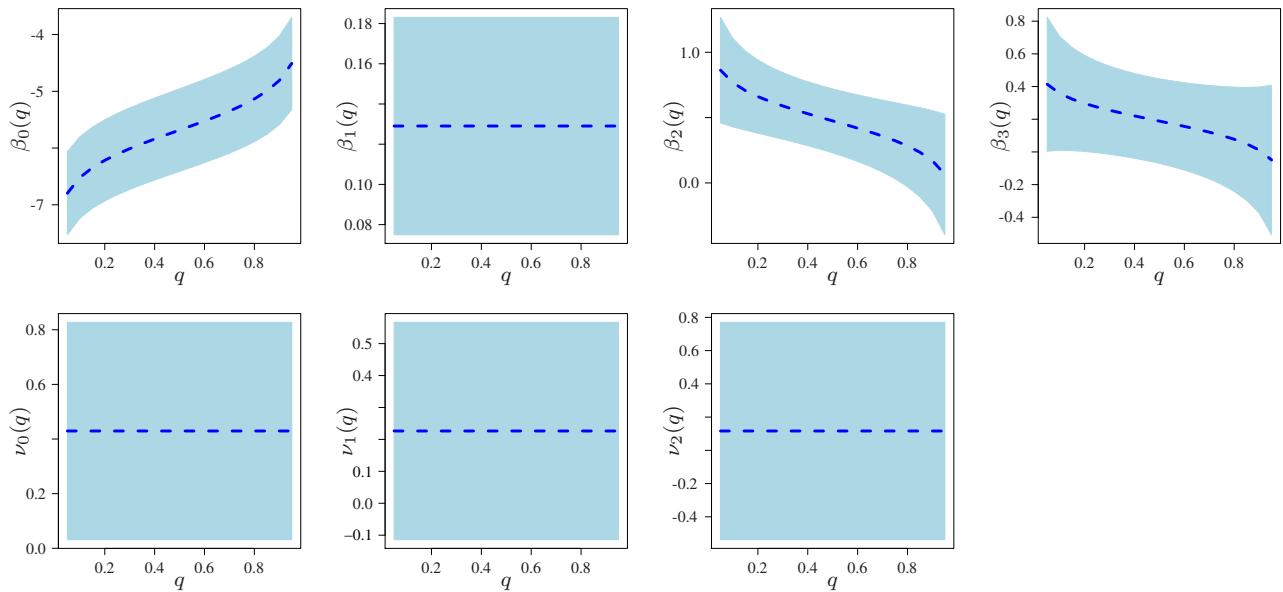


Figure 3: Point estimation and 95% confidence interval for parameters estimated in  $\text{RPGJSB1}_q$  model for different quantiles (cloglog link and  $G$  the cdf of the logistic model).

### 5.1.2 Local influence analysis

We also presented a local influence analysis for the selected model under the three perturbations schemes discussed in Section 3.3. Figure 5 shows such analysis for the  $\text{RPGJSB1}$  model with  $q = 0.5$

Table 5: Estimated parameters for different quantile in  $\text{RPGJSB1}_{q=0.5}$  model for the COVID-19 data set with  $G$  the cdf of the logistic model and cloglog link. Also are presented the p-values for the traditional normality test for RQRs.

$q$	parameter	estimated	s.e.	$t$ -value	$p$ -value	$p$ -values for quantile residuals			
						KS	SW	AD	CVM
0.50	$\beta_0$	-5.6835	0.3709	-15.32	<0.0001	0.995	0.820	0.915	0.969
	$\beta_1$	0.1290	0.0276	4.68	<0.0001				
	$\beta_2$	0.4749	0.1248	3.80	0.0001				
	$\beta_3$	0.1886	0.1320	1.43	0.0766				
	$\nu_0$	0.9060	0.1556	5.82	<0.0001				
	$\nu_1$	0.4294	0.2030	2.12	0.0172				
	$\nu_2$	0.2264	0.1737	1.30	0.0963				
	$\log \alpha$	0.1164	0.3337	0.35	0.3636				

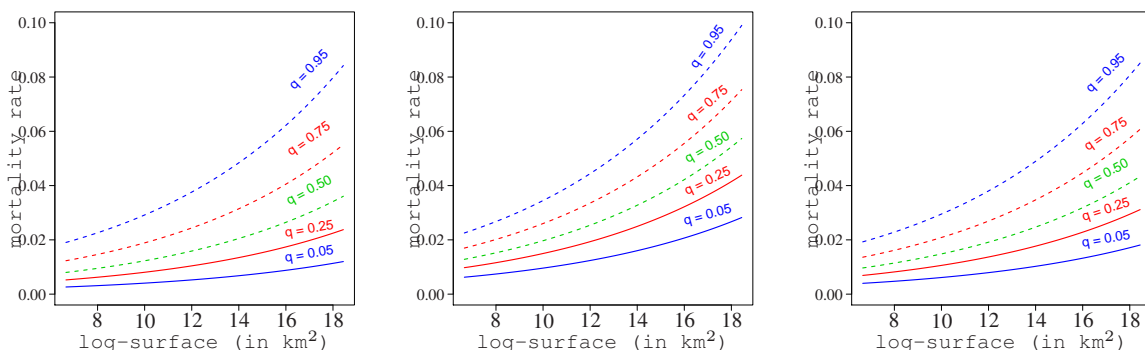


Figure 4: Estimated  $100 \times q$ th quantile in the  $\text{RPGJSB1}_q$  model varying the  $\log(\text{surface})$  for countries in Africa, Asia or Oceania (left panel), America (center panel) and Europe (right panel) considering the cloglog link and  $G$  the cdf from the logistic model.

using the cloglog link and  $G$  the cdf of the logistic model in the COVID-19 data set. In Section B.3 of the supplementary material is presented the same analysis for other selected quantiles. Note that, considering all the cases, the observation 121 appear in at least some case, which correspond to Yemen (Asia). Yemen reported a high mortality rate (29%, 601 accumulated deaths and 2067 accumulated cases, respectively). Evidently there is a problem in the handling of information about COVID-19 in the country. Table 6 presents the relative change for the parameters (RC), for its estimated standard errors (RCSE) and the respective  $p$ -value for the estimation without Yemen. We highlight that the greater variations are obtained for the parameters related to the scale and for  $\log \alpha$  (excepting the case for  $\beta_3(q = 0.90)$ ). However, the estimated quantiles presented in Figure 3 do not depend on those parameters. Therefore, such plot without the referred observations are similar. We highlight that the significance of the parameters related to the quantile is maintained for all the cases (excepting for  $\beta_3(q = 0.05)$ ), suggesting a robustness of the model to estimate the different quantiles in this problem.

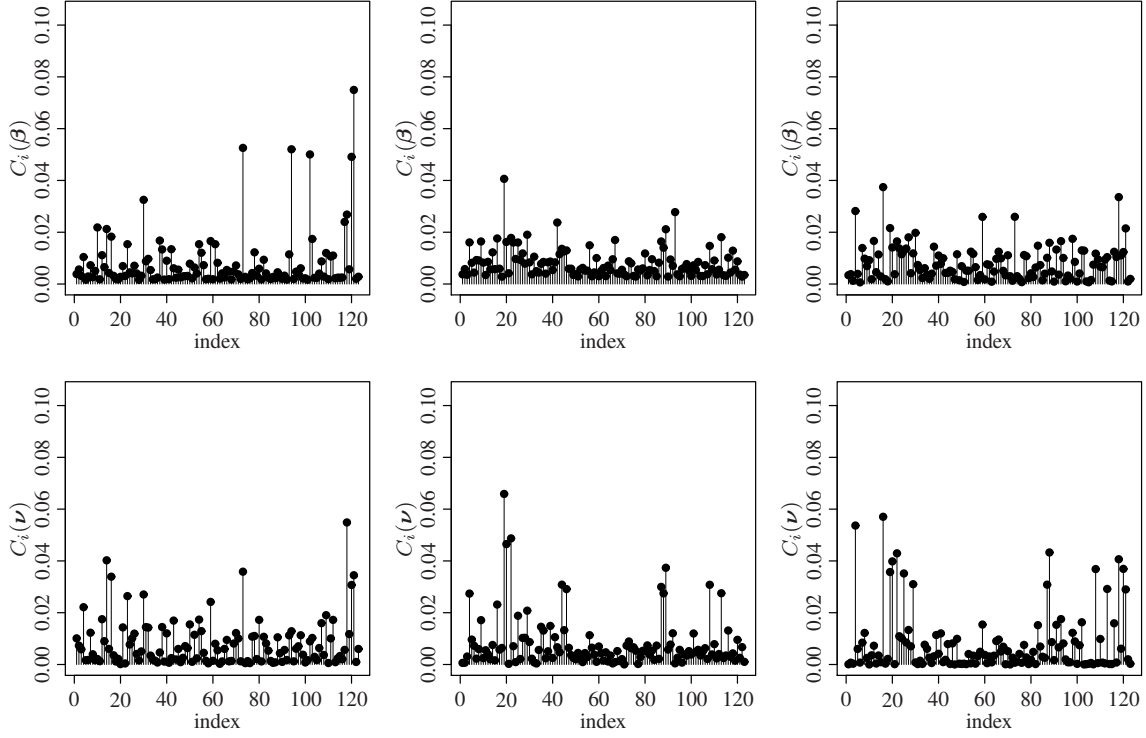


Figure 5: Index plots of  $C_i$  for  $\hat{\beta}$  (upper) and  $\hat{\nu}$  (lower) under the weight perturbation (left), response perturbation (center) and covariate perturbation (right) schemes for  $\text{RPGJSB1}_{q=0.5}$  model (cloglog link and  $G$  the cdf from the logistic model) in COVID-19 data set.

## 6 Conclusions

In this paper, we propose two classes of parametric quantile regression models for studying the association between a bounded response and covariates via inferring the conditional quantile of the response. The proposed quantile regression models was built based on power Johnson SB distribution (Cancho et al., 2020) using a new parameterization of this distribution that is indexed by quantile, dispersion ( $\psi$  and  $\delta$ ) and shape parameters ( $\text{RPGJSB1}_q(\psi, \delta, \alpha)$ ) or quantile and dispersion parameters ( $\text{RPGJSB2}_q(\psi, \delta)$ ). The first proposed quantile model has an extra-parameter  $\alpha > 0$  is associated with the “tailedness”, and the second proposed quantile model has a less computational costs. The ML inference was implemented to estimate the models parameters, which was satisfactory considering the simulation studies where parameters were recovered for different sample sizes. Furthermore, under each proposed quantile regression model, we have developed model diagnostic tools. In order to illustrate our approach, two applications using real data sets were presented and discussed. In particular, we analyze the mortality rate of COVID-19 and the reproductive activity of the Chilean anchoveta. Results of the applications showed that the proposed quantile models are adequate. Based on the results, the  $\text{RPGJSB1}_q$  regression model presents a better fit for the COVID-19 mortality rate and the anchoveta data sets. As part of future research, there are several extensions of the new models not considered in this paper that can be addressed in future research, in particular, an extension

Table 6: RCs (in %) in ML estimates and their corresponding SEs for the indicated parameter and respective p-values for COVID-19 data set when observation 121 is dropped.

		$q$				
parameter		0.10	0.25	0.50	0.75	0.90
RC		7.81	11.06	16.43	23.58	32.58
RCSE	$\beta_0(q)$	0.20	0.10	0.05	0.27	0.66
p-value		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
RC		15.58	15.58	15.58	15.58	15.58
RCSE	$\beta_1(q)$	0.28	0.28	0.28	0.28	0.28
p-value		<0.0001	<0.0001	<0.0001	<0.0001	<0.0001
RC		1.17	13.45	41.56	99.46	266.08
RCSE	$\beta_2(q)$	4.83	7.63	8.81	2.12	28.95
p-value		0.0003	0.0001	0.0001	<0.0001	<0.0001
RC		9.43	30.59	78.19	216.11	2502.37
RCSE	$\beta_3(q)$	6.92	12.61	16.00	1.10	450.43
p-value		0.0526	0.0431	0.0383	0.0359	0.0351
RC		12.54	12.54	12.54	12.54	12.54
RCSE	$\nu_0(q)$	28.63	28.63	28.63	28.63	28.63
p-value		0.0562	0.0562	0.0562	0.0562	0.0562
RC		45.7	45.7	45.7	45.7	45.7
RCSE	$\nu_1(q)$	8.75	8.75	8.75	8.75	8.75
p-value		0.1588	0.1588	0.1588	0.1588	0.1588
RC		35.73	35.73	35.73	35.73	35.73
RCSE	$\nu_2(q)$	11.86	11.86	11.86	11.86	11.86
p-value		0.3219	0.3219	0.3219	0.3219	0.3219
RC		951.62	951.63	951.63	951.61	951.64
RCSE	$\log \alpha(q)$	694.77	694.78	694.78	694.76	694.78
p-value		0.3855	0.3855	0.3855	0.3855	0.3855

of the methods developed in this paper would be to consider in (2) a much more general family of distributions; that is, consider models for zero-inflated and one-inflated data set. Directions related to random effects in the model also can be addressed in future works.

## Acknowledgements

The authors thank to “Instituto de fomento pesquero” (IFOP) to provide the anchoveta data set presented in the supplementary material.



## References

- Bayes, C.L., Bazán, J.L., García, C. (2012). A new robust regression model for proportions. *Bayesian Analysis* **7**, 841-866.
- 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.
- Cancho, V.G., Bazán, J.L., Dey, D.K. (2020). A new class of regression model for a bounded response with application in the study of the incidence rate of colorectal cancer. *Statistical Methods in Medical Research*. **29**, 2015-2033.
- Cook, R. D. (1986). Assessment of Local Influence. *Journal of the Royal Statistical Society: Series B (Methodological)* **48**, 133-155.
- Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London, UK.
- Du, R.H., Liang, L.R., Yang, C.Q., Wang, W., Cao, T.Z., Li, M., Guo, G.Y., Du, J., Zheng, C.L., Zhu, Q., Hu, M. (2020) Predictors of mortality for patients with COVID-19 pneumonia caused by SARS-CoV-2: a prospective cohort study. *Eur. Respir. J.*, **55**, 2000524
- Dunn, P.K., Smyth, G.K. (1996). Randomized quantile residuals. *Journal of Computational and Graphical Statistics* **5**, 236-244.
- Durrans, S.R. (1992). Distributions of fractional order statistics in hydrology. *Water Resources Research* **28**, 1649-1655.
- Ferrari, S. L., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics* **31**, 799-815.
- Ji, J.S., Liu, Y., Liu, R., Zha, Y., Chang, X., Zhang, L., Zhang, Y., Zeng, J., Dong, T., Xu, X., Zhou, L. (2020) Survival analysis of hospital length of stay of novel coronavirus (COVID-19) pneumonia patients in Sichuan, China medRxiv (2020) 10.1101/2020.04.07.20057299
- Johnson, N.L. (1949). Systems of frequency curves generated by the methods of translation. *Biometrika* **36**, 149-176.
- Koenker, R., Bassett, G. (1978). Regression quantiles. *Econometrica* **46**, 33-50.
- Lehmann, E.L. (1953). The power of rank tests. *Annals of Mathematical Statistics* **24**, 23-43.
- Lemonte, A.J., Bazán, J.L. (2016). New class of Johnson  $S_B$  distributions and its associated regression model for rates and proportions. *Biometrical Journal* **58**, 727-746.
- Lemonte, A.J., Moreno-Arenas, G. (2020). On a heavy-tailed parametric quantile regression model for limited range response variables. *Computational Statistics* **35**, 379-398.
- Livingston, E., Bucher, K. (2020) Coronavirus disease 2019 (COVID-19) in Italy. *Journal of the American Medical Association*. **323**, 1335.

- Li, X., Xu, S., Yu, M., Wang, K., Tao, Y., Zhou, Y., Shi, J., Zhou, M., Wu, B., Yang, Z. (2020) Risk factors for severity and mortality in adult COVID-19 inpatients in Wuhan. *J. Allergy Clin. Immunol.* **146**, 110-118
- 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. *Journal of Applied Statistics* **47**, 954-974.
- Migliorati, S., Di Brisco, A.M., Ongaro, A. (2018). A New Regression Model for Bounded Responses. *Bayesian Analysis* **13**, 845-872.
- Ospina, R., Ferrari, S.L.P. (2008). Inflated beta distributions. *Statistical Papers* **51**, 111-126.
- Su, S. (2015). Flexible parametric quantile regression model. *Statistics and Computing* **25**, 635-650.
- WHO coronavirus disease (COVID-19) dashboard (2020). Geneva: World Health Organization. Available online: <https://covid19.who.int/> (last cited: [03/11/2020]).
- Yap, B.W., Sim, C.H. (2011) Comparisons of various normality tests. *Journal of Statistical Computation and Simulation* **81**, 2141-2155.