Submitted to the Brazilian Journal of Probability and Statistics

L-Logistic regression models: Prior sensitivity analysis, robustness to outliers and applications

Rosineide F. da Paz[a] **Narayanaswamy Balakrishnan**[b] **and Jorge Luis Bazán**[c]

[a] *Universidade Federal do Ceará, Campus de Russas, CE, Brazil.*
[b] *McMaster University, Hamilton, Ontario, Canada.*
[c] *Uninversidade de São Paulo, São Carlos, SP, Brazil.*

Abstract. Tadikamalla & Johnson (1982) developed the L_B distribution to variables with bounded support by considering a transformation of the standard Logistic distribution. In this manuscript, a convenient parametrization of this distribution is proposed in order to develop regression models. This distribution, referred to here as L-Logistic distribution, provides great flexibility and includes the uniform distribution as a particular case. Several properties of this distribution are studied, and a Bayesian approach is adopted for the parameter estimation. Simulation studies, considering prior sensitivity analysis, recovery of parameters and comparison of algorithms, and robustness to outliers are all discussed showing that the results are insensitive to the choice of priors, efficiency of the algorithm MCMC adopted, and robustness of the model when compared with the beta distribution. Applications to estimate the vulnerability to poverty and to explain the anxiety are performed. The results to applications show that the L-Logistic regression models provide a better fit than the corresponding beta regression models.

1 Introduction

Modeling data that are restricted to the interval (0,1), as for example the proportion of children vulnerable to poverty or anxiety as a function of the stress, is frequently encouraged by researchers. Many different regression models have been proposed in the past two decades for modeling this type of data. For example, Buckley (2003), Ferrari & Cribari-Neto (2004), Paz *et al.* (2015), Lemonte & Bazán (2016), Gómez-Déniz *et al.* (2014) and Bayes *et al.* (2017) have all proposed regression models. Yet, there are still continuous distributions with bounded support that need further study. This is the case of the L-Logistic distribution, which was originally proposed by Tadikamalla & Johnson (1982) through a transformation of the standard Logistic distribution. This construction is similar to the S_B system proposed by Johnson (1949). This distribution was studied by, Tadikamalla & Johnson (1990) and Johnson & Tadikamalla (1991), among others, who proposed the method of moments and the percentile point method to fit this distribution, and by Wang & Rennolls (2005), who considered the Maximum Likelihood estimation. However, regression models have not been studied based on this distribution.

Keywords and phrases. Bayesian analysis, L-Logistic distribution, Regression analysis, beta distribution, Sensibility analysis

BJPS Accepted Manuscript

In this work, we discuss the properties of this distribution by considering a new parametrization. Another parametrization is also preserved in the Online Suplementary Material. Here, we present the parametrization of the L-Logistic distribution that considers the median as one parameter and the dispersion as another parameter. Therefore, we propose a new regression model considering this distribution in the context of quantile regression (QR) models, which were introduced by Koenker & Bassett (1978). Specifically, we propose a median regression model, which may represent the relationship between the median (central location) of the response and a set of covariates as well as of the dispersion parameter and another or the same) set of covariates, by using a convenient link function. If the data are highly skewed, since the median is a natural robust measure of location, the conditional median modeling can be more useful than the usual conditional mean modeling adopted usually in beta regression models (Buckley, 2003; Ferrari & Cribari-Neto, 2004). Different from previous studies of L-Logistic distribution, here we propose a Bayesian approach employing a Markov chain Monte Carlo (MCMC) method for the modeling framework. The issues of model fitting are addressed by means of a hybrid algorithm that combines *Metropolis-Hasting* algorithm with *Gibbs sampling*. In the Online Supplementary Material, we report results from the first studies with simulated data sets to investigate prior sensitivity analysis of the dispersion parameter of the L-logistic distribution and the median L-logistic regression, parameter recovery and comparison of algorithms, and to evaluate the robustness to outliers of the L-logistic distribution in comparison with beta distribution. These results display that the proposed estimation method works well, and that the model proposed is more robust than the beta models in the presence of outliers. Also, real application of social and psychological data is considered to show the advantages of proposed approach. Firstly, we show that the proportion of children vulnerable to poverty of the municipalities of the state of Alagoas in Brazil, for the 2010 season, is best fitted by the L-Logistic distribution as compared to the beta distribution. Additionally, we show that L-Logistic regression models provide a better fit than the corresponding beta regression models for analyzing the anxiety as a function of the stress using a sample of nonclinical women in Townsville, Queensland, Australia.

The rest of this manuscript is organized as follows. Section 2 is dedicated to L-Logistic distribution, and we study some characteristics of this distribution like alternative parametrizations, some related distributions, moments, the skewness and kurtosis measures. In Section 3, we propose the L-Logistic median regression model. Section 4 is dedicated to the Bayesian estimation of the distribution parameters, and to the parameter of the proposed median regression model. Section 5 presents the results of three simulation studies that examine a prior sensitivity analysis, parameter recovery, comparison of algorithms, and robustness to outliers of the L-Logistic distribution. Section 6 discusses applications of the proposed distribution, including the applicability of regression models to real data sets. Finally, some concluding remarks are made in Section 7.

BJPS-Accepted Manuscript

2 The L-Logistic Distribution

We say that the random variable (r.v.) Y follows a L-Logistic distribution, denoted by $Y \sim LL(m, b)$, if its probability density function (pdf) is given by

$$f(y|m,b) = \frac{b(1-m)^b m^b y^{b-1}(1-y)^{b-1}}{\left[(1-m)^b y^b + m^b(1-y)^b\right]^2}, \quad 0 < y < 1, 0 < m < 1, \quad b > 0. \tag{1}$$

Depending on the parameters m and b, the L-Logistic distribution takes on a variety of shapes (see, for example, Figures 1 and 2). Note that when we set $m = 0.5$ and $b = 1$ in (1), the pdf of the L-Logistic distribution simply becomes the pdf of the standard uniform distribution. Here, m is the median of the distribution, which scales the graph to the left or right on the horizontal axis, and consequently it is a location parameter. The L-Logistic density is uni-modal (or "uni-antimodal"), increasing, decreasing, or constant, depending on the values of its parameters. Additionally, we note that for a fixed value of the parameter m, the dispersion of the distribution decreases as b increases, and so b is a parameter that governs the dispersion of the distribution. In the Online Supplementary Material, we show that the parameter b is in fact a dispersion parameter.

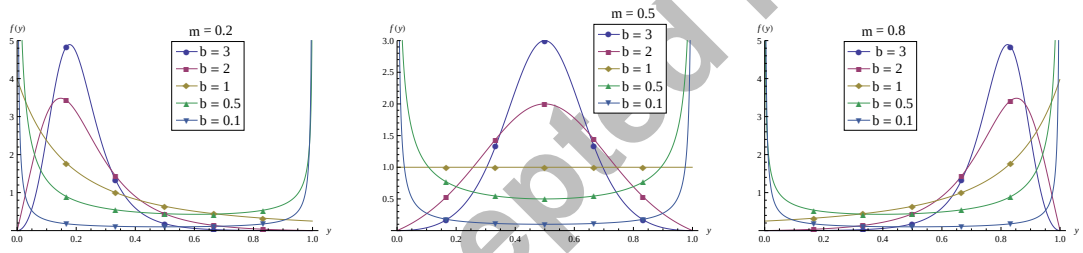

Figure 1: L-Logistic probability density functions for $m = 0.2, 0.5$ and 0.8 and some choices of parameter b.

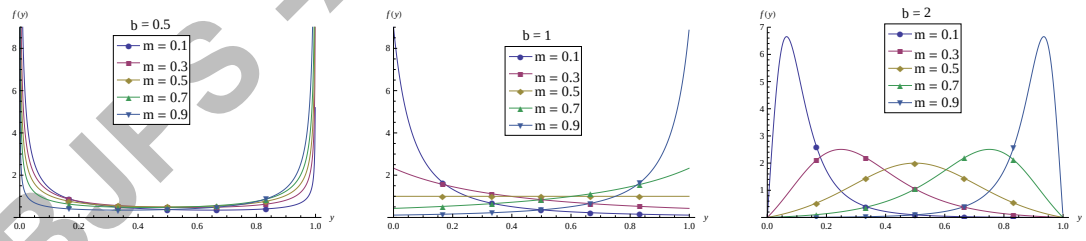

Figure 2: L-Logistic probability density function for dispersion parameter $b = 0.5, 1$ and 2 and some choices of parameter m.

BJPS - Accepted Manuscript

The cumulative distribution function (cdf) of the L-Logistic distribution is given by

$$F_Y(y|m,b) = \left(1 + \left(\frac{m(1-y)}{y(1-m)}\right)^b\right)^{-1}, \quad 0 < y < 1, \tag{2}$$

which can be readily inverted to yield the quantile function

$$Q_Y(p) = F_Y^{-1}(p) = \frac{\left(\frac{m}{1-m}\right)}{\left(\frac{m}{1-m}\right) + \left(\frac{1-p}{p}\right)^{1/b}}, \quad 0 \le p \le 1. \tag{3}$$

This would readily enable a quantile-based analysis of this model. Note that if $p = 1-p = 0.5$, then $Q(p) = m$, which means that the parameter m is indeed the 50^{th} percentile or the median of the L-Logistic distribution.
Equation (3) facilitates simple r.v. generation. Specifically, if $U \sim uniform(0,1)$, then

$$X = Q(U) = \frac{\left(\frac{m}{1-m}\right)}{\left(\frac{m}{1-m}\right) + \left(\frac{1-U}{U}\right)^{1/b}} \sim LL(m,b). \tag{4}$$

Additionally, we can express the inter-quartile range (IQR) as

$$IQR = Q(0.75) - Q(0.25) = \frac{m3^{1/b}}{(1-m) + 3^{1/b}m} - \frac{m}{3^{1/b}(1-m) + m}. \tag{5}$$

The IQR has a breakdown point of 50%, and this measure is often preferred over range. When the distribution is symmetric, half IQR equals the median absolute deviation, and is often used in the detection of outliers in data.

2.1 Mode

Property 2.1. *For $b > 1$, the mode y_0 of the L-Logistic distribution is the solution of the equation*

$$\left(\frac{1-m}{m}\right)^b = \left(\frac{1-y_0}{y_0}\right)^b \frac{b + 2y_0 - 1}{b - 2y_0 + 1}. \tag{6}$$

Note that, upon taking $\delta = -b\log\left(\frac{m}{1-m}\right)$, the mode y_0 can be obtained by solving the equation

$$\delta = \log\left(\left(\frac{1-y_0}{y_0}\right)^b \frac{b + 2y_0 - 1}{b - 2y_0 + 1}\right). \tag{7}$$

BJPS-Accepted Manuscript

In addition, from (6) and (7), if $y_0 = m = 0.5$, then $\delta = 0$ for all values of b. Thus, we can study the behavior of the mode by studying the function in (7). For this purpose, we take the derivative of the right-hand side of (7) with respect to y_0 to obtain the equation

$$\frac{\partial \delta}{\partial y_0} = \frac{b\left(b^2-1\right)}{(y_0-1)y_0\left\{(b^2-1)+4y_0-4{y_0}^2\right\}}. \tag{8}$$

(6) is negative for $b > 1$, the situation where δ decreases as y_0 (mode) increases (first derivative test), then the mode lies in the interval $< 0, 0.5 >$ if $\delta > 0$ (or $m < 0.5$) and for $\delta < 0$ (or $m > 0.5$) the mode is in the interval $< 0.5, 1 >$. If $b < 1$, (8) is positive whenever $\left\{(b^2-1)+4y_0-4y_0^2\right\} > 0$, that is, whenever $\frac{1-b}{2} < y < \frac{1+b}{2}$, the situation where δ increases as y increases. Thus, from (7) and (8), the minimum of the pdf lies in the interval $< \frac{1-b}{2}, 1/2 >$ for $\delta < 0$ or $m > 0.5$, and in the interval $< 0.5, \frac{1+b}{2} >$ for $\delta > 0$ or $m < 0.5$.

2.2 Moments

The following proposition gives an expression for the moments of the L-Logistic distribution.

Property 2.2. *If $Y \sim LL(m, b)$, then the moments of Y about zero are given by*

$$E[Y^t] = \int_0^1 \left[1 + \left(\frac{1-v}{v}\right)^{1/b}\left(\frac{1-m}{m}\right)\right]^{-t} dv. \tag{9}$$

The integral in (9) cannot be expressed in an analytical form. However, we can use numerical integration to evaluate some moments as $E_Y(Y)$, $E_Y(Y^2)$ and $Var_Y(Y) = E_Y(Y^2) - E_Y(Y)^2$. Table 1 shows some values of the first and second moments and the variance of the L-Logistic distribution. In addition, Figure 3 shows the graphs of the mean and variance as functions of the dispersion parameter b, for some choices of the parameter m. For this purpose, the integral in (9) was evaluated by the Gaussian quadrature.

BJPS - Accepted Manuscript

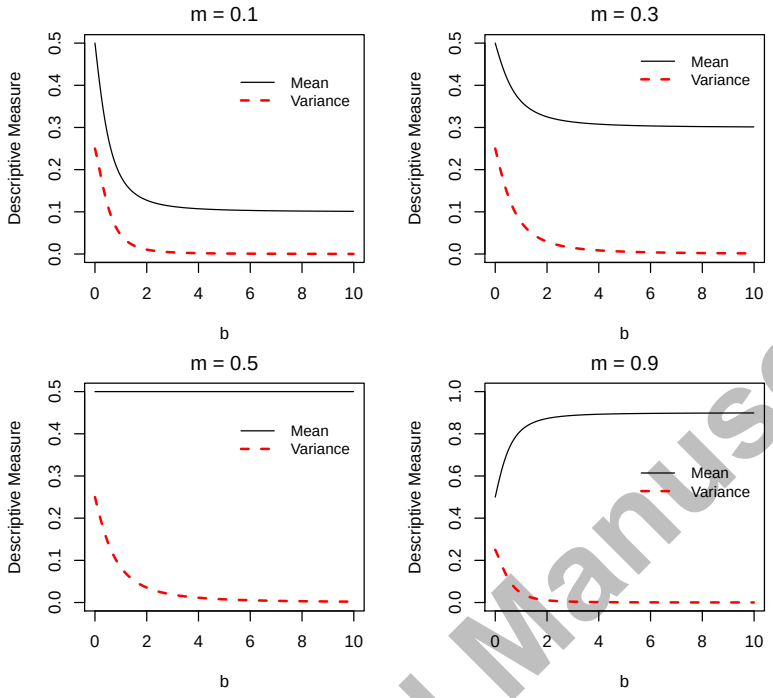

Figure 3: Descriptive measures of the L-Logistic distributions for some values of the parameters

Table 1 *$E_Y[Y]$, $E_Y[Y^2]$, and $Var_Y(X)$ of the L-Logistic distribution for some choices of b and m.*

m	0.2	0.5	0.8	0.2	0.5	0.8
b	1	1	1	3	3	3
$E_Y[Y]$	0.283	0.5	0.717	0.216	0.5	0.784
$E_Y[Y^2]$	0.145	0.333	0.579	0.056	0.269	0.625
$Var_Y[Y]$	0.065	0.083	0.065	0.010	0.019	0.01

2.3 Skewness and Kurtosis Measures

First, we have the following symmetry property.

Property 2.3. *The L-Logistic density is symmetric when $m = 0.5$ for all values of b.*

For the case when the L-Logistic density is asymmetric, the degree of skewness can be quantified by some measures of skewness. Since the L-Logistic distribution is related to the Logistic distribution, the skewness measure introduced by Arnold & Groeneveld (1995), denoted by γ_M, seems to be a suitable skewness measure. The measure γ_M is based on the

BJPS-Accepted Manuscript

mode of distribution and is given by

$$\gamma_M = 1 - 2F(M), \tag{10}$$

where M is the mode of the distribution and $F(.)$ is the distribution function. The value of γ_M lies in $(-1, 1)$, and if γ_M is near 1, it indicates extreme right skewness. On the other hand, if γ_M is near -1, it indicates extreme left skewness.

We also consider another measure of skewness, called quantile skewness (denoted here by γ_p), first proposed by Hinkley (1975). This skewness measure is given by

$$\gamma_p = \frac{Q(1-p) + Q(p) - 2m}{Q(1-p) - Q(p)}, \tag{11}$$

which is a function of high and low percentiles defined by $p \in (0, 0.5)$ with $Q(.)$ being as in (3). The maximum value of γ_p is 1, representing extreme right skewness, while the minimum is -1 representing extreme left skewness. This measure is also zero for any symmetric distribution. However, the function in (11) depends on the value of p. We can remove this dependence by integrating over p, or to decide which value of p is appropriate for use. In Brys *et al.* (2003), there is a comparison between several robust skewness measures in which accuracy, robustness, and computational complexity are all considered. The most interesting skewness measure of all the measures investigated is octile skewness. Octile skewness takes $p = 0.125$ in (11), that is, it is given by

$$\gamma_O = \frac{O_7 - O_4 + O_1 - O_4}{O_7 - O_1} = \frac{Q(0.875) + Q(0.125) - 2m}{Q(0.875) - Q(0.125)}. \tag{12}$$

For the L-Logistic distribution, we made use of this particular skewness measure instead of removing the dependence over p through integration.

Moreover, the kurtosis of the L-Logistic distribution can also be derived easily by using the quantiles. The kurtosis measure introduced by Moors (1988) is given by

$$\kappa_O = \frac{O_7 - O_6 + O_3 - O_1}{O_6 - O_2} = \frac{Q(0.875) - Q(0.625) + Q(0.375) - Q(0.125)}{Q(0.75) - Q(0.25)}, \tag{13}$$

with $\kappa_O \in (0, \infty)$. Moors (1988) justified the use of the kurtosis measure in (13) by the interpretation that the two terms in the numerator of (13) are large (small) if relatively little (much) probability mass is concentrated in the neighborhood of $Q(0.75)$ and $Q(0.25)$. This corresponds to large (small) dispersion around (roughly) $E_Y[Y] \pm Var_Y[Y]$ where $E_Y[Y]$ and $Var_Y[Y]$ are the mean and variance of Y, respectively.

Figure 4 presents the results of the measures of skewness and kurtosis described here for some values of the parameter m as a function of the dispersion parameter b, for $b > 1$. From this figure, we can see that when $m < 0.5$, the two measures of skewness decrease as b increases. However, when $m > 0.5$, the two measures of skewness increase as b increases.

BJPS - Accepted Manuscript

For $m = 0.5$, we can see that these two measures are constant. In addition, the mode of the L-Logistic distribution increases as b increases when $m < 0.5$, is constant when $m = 0.5$, and decreases as b increases when $m > 0.5$. For the measure of kurtosis, we see no pattern in this figure.

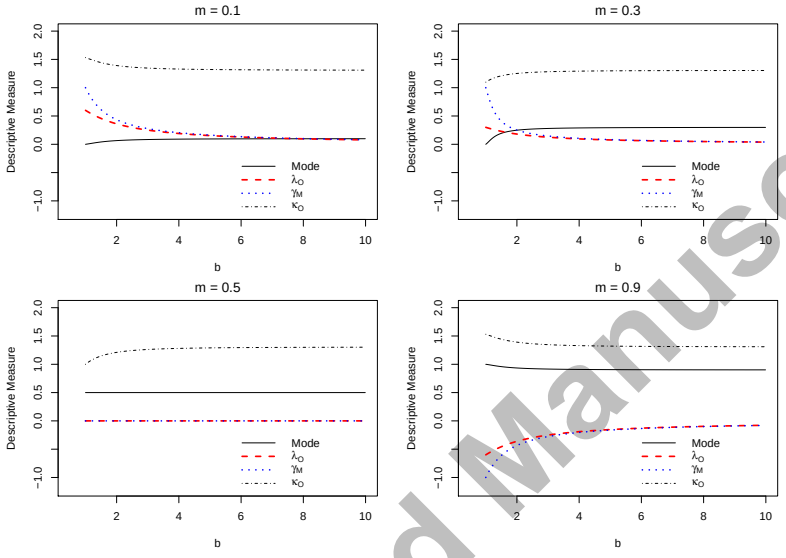

Figure 4: The mode, two measure of skewness (γ_M, γ_O), and kurtosis (κ_O) of the L-Logistic distribution for some choice of the parameters.

3 L-Logistic median regression model

Regression analysis estimates the potential differential effect of a covariate on the mean or quantiles of the conditional distribution (Hao & Naiman, 2007). Here, we are interested in studying the conditional (or regression) median as a function of the covariates when the response variable takes values in a bounded interval. Our goal is to define a median regression model for a r.v. that assumes values in the standard unit interval. Let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be a vector of independent r.v.'s following the distribution in (1) with median m_i and dispersion parameter b_i, and consider that $\boldsymbol{x}_{1i}^T$ and $\boldsymbol{x}_{2i}^T$ are $q-$ and $d-$dimensional vectors, respectively, containing the explanatory variables both with 1 as the first component. Thus, in the regression analysis with the L-Logistic distribution, we assume that conditional on the explanatory variables (covariates), the r.v.'s Y_i, $i = 1, ..., n$, are mutually independent with L-Logistic distribution, i.e.,

$$Y_i \sim LL(m_i, b_i), \tag{14}$$

BJPS Accepted Manuscript

with

$$h_1(m_i) \quad = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \text{ and } h_2(b_i) = \boldsymbol{x}_{2i}^T\boldsymbol{\delta}, \tag{15}$$

where $\boldsymbol{\beta} = (\beta_0, ..., \beta_{q-1})$ and $\boldsymbol{\delta} = (\delta_0, ..., \delta_{d-1})$ ($\beta_i, \delta_j \in \mathbb{R}$ for $i = 0, ..., (q-1)$ and $j = 0, ..., (d-1)$) represent, respectively, $q-$ and $d-$dimensional vectors of unknown regression parameters. In (15), h_1 and h_2 are strictly monotone and twice differentiable real link functions. This method allows to fit the model adequately with a variety of link functions, which ensures parameter m is in the interval (0, 1) and the dispersion parameter b is in the interval $(0, \infty)$. Some choices of link functions for a parameter are discussed in Ferrari & Cribari-Neto (2004). A common link function for the parameter m is the logit function,

$$\text{logit}(m_i) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \quad \Rightarrow \quad m_i = \frac{\exp\{\boldsymbol{x}_{1i}^T\boldsymbol{\beta}\}}{1+\exp\{\boldsymbol{x}_{1i}^T\boldsymbol{\beta}\}}. \tag{16}$$

For the dispersion parameter, a common link function is the log-linear link function. For easy interpretation, here we follow Smithson & Verkuilen (2006) and take $h_2 = -log(b_i)$, that is,

$$\log(b_i) = -\boldsymbol{x}_{2i}^T\boldsymbol{\delta} \quad \text{or} \quad b_i = \exp\left\{-\boldsymbol{x}_{2i}^T\boldsymbol{\delta}\right\}. \tag{17}$$

4 Bayesian estimation

In this section, we describe the Bayesian approach for the estimation of parameters of the L-Logistic distribution, and also of the L-Logistic regression model.

4.1 Bayesian estimation of the L-logistic distribution

If we consider a random sample $\boldsymbol{Y} = (Y_1, ..., Y_n)$ from the distribution in (2), then the likelihood function is given by

$$L(m, b|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{b(1-m)^b m^b y_i^{b-1}(1-y_i)^{b-1}}{\left[(1-m)^b y_i^b + m^b(1-y_i)^b\right]^2}, \tag{18}$$

where $0 < m < 1$ and $b > 0$. To complete the Bayesian specification of the model, since parameters m and b have different behavior, we assume independence between them, and the following structure is then considered:

$$\pi(m, b) = \pi(m)\pi(b), \tag{19}$$

where $\pi(m)$ and $\pi(b)$ are the prior densities for m and b, respectively.

The prior $\pi(b)$ can be, for example, the pdf of the Gamma distribution with parameter vector (ϵ, ϵ), ϵ being a small value. The choice of this prior seems to be reasonable as a Gamma prior has large variance. The prior sensitivity analysis considering other priors for the parameter b presented in Section 7 shows the robustness of results in the selection of

BJPS - Accepted Manuscript

prior distributions. For the parameter m, we can take $m \sim unifom(0,1)$ or $m \sim beta(1,1)$, where $beta(a,b)$ represents the beta distribution with parameters a and b. This is a "flat" prior where all values in the range are equally likely. This choice can be considered weakly informative because in other cases (i.e., when a or $b \neq 1$) we will expect a value of the median m to be greater than 0.5 or otherwise. Then, a more informative, subjective or expert prior by interviewing experts can be considered in order to elicit parameters in a parametric family of priors following methods discussed, for example, by Albert *et al.* (2012). Another choice for m can be through specifying other values of the hyper parameters a and b in the beta distribution. From our results in simulation studies, we did not find any problem with this choice for the model without covariate. By considering the specifications above, the joint posterior distribution for (m,b) is given by

$$\pi(m,b|\boldsymbol{y}) \propto \prod_{i=1} \frac{b(1-m)^b m^b y_i^{b-1}(1-y_i)^{b-1}}{\left[(1-m)^b y_i^b + m^b(1-y_i)^b\right]^2}\pi(b), \tag{20}$$

where $0 < m < 1$ and $b > 0$.

Since the posterior distribution is not available in a closed-form, the Markov Chain Monte Carlo (MCMC) approach (Gelman *et al.*, 2013, pp. 259 – 349) is used to estimate the model parameters. Initially, we consider the full conditional posterior distributions for the parameters (m,b) given by

$$\pi(m|b,\boldsymbol{y}) = K_1^{-1} \frac{(1-m)^{nb} m^{nb}}{\prod_{i=1}^n \left[(1-m)^b y_i^b + m^b(1-y_i)^b\right]^2}, \tag{21}$$

$$\pi(b|m,\boldsymbol{y}) = K_2^{-1} \prod_{i=1}^n \left(\frac{b(1-m)^b m^b y_i^{b-1}(1-y_i)^{b-1}}{\left[(1-m)^b y_i^b + m^b(1-y_i)^b\right]^2} \right) \pi(b), \tag{22}$$

$$\tag{23}$$

where $0 < m < 1$ and $b > 0$ with K_1 and K_2 being normalizing constants.

Thus, a hybrid algorithm that combines Metropolis-Hastings algorithm and *Gibbs sampling* was implemented in R language (R Development Core Team, 2015) to obtain a sample from the posterior distribution of model parameters (m,b). These codes are available upon request from the authors.

As suggested by a referee, we also implemented a MH algorithm. Based on different scenarios, we did not find difference in the recovery of parameters based on this two algorithms.

4.2 Bayesian estimation of the L-logistic regression model

Now, let $\boldsymbol{Y} = (Y_1, ..., Y_n)$ be a vector of independent r.v.'s following the L-Logistic distribution with median m_i and dispersion parameter b_i given by (16) and (17), respectively. Then, the likelihood of the observed sample $\boldsymbol{y} = (y_1, .., y_n)$ of $\boldsymbol{Y}$ can be written as

$$L(\boldsymbol{\beta}, \boldsymbol{\delta}|\boldsymbol{y}, \boldsymbol{X}) = \prod_{i=1}^n \frac{b_i(1-m_i)^{b_i} m_i^{b_i} y_i^{b_i-1}(1-y_i)^{b_i-1}}{\left[(1-m_i)^{b_i} y_i^b + m^{b_i}(1-y_i)^{b_i}\right]^2}, \tag{24}$$

BJPS-Accepted Manuscript

where $\boldsymbol{X}$ is the matrix containing all the explanatory variables, and $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are the regression parameters such that

$$\text{logit}(m_i) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \text{ and } \log(b_i) = -\boldsymbol{x}_{2i}^T\boldsymbol{\delta}. \tag{25}$$

Suppose we have no prior information from historical data for the regression parameters. In many applications, a normal prior distribution centered at zero with a standard error equal to 100 for the regression coefficient will be sufficiently noninformative. So, we assign these weakly informative prior distributions to the parameters, i.e., we adopt prior normal distributions with large variance such that

$$\begin{array}{rcl} \beta_j & \sim & normal(0, 100), \text{ for } j = 0, .., q-1, \\ \delta_l & \sim & normal(0, 100), \text{ for } l = 0, .., d-1. \end{array} \tag{26}$$

The prior distributions of parameters are chosen here under the assumption that they are independent of each other.

Assuming the prior distributions in (26) for the parameters, the posterior density takes on the form

$$\pi(\boldsymbol{b}, \boldsymbol{\beta}, \boldsymbol{\delta}|\boldsymbol{y}) \propto L\left(\boldsymbol{\beta}, \varphi|\boldsymbol{y}, \boldsymbol{X}\right)\pi(\boldsymbol{\beta})\pi(\delta) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \delta)\pi(\boldsymbol{\beta})\pi(\boldsymbol{\delta}). \tag{27}$$

Therefore, the full conditional posterior distributions for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ can be written as

$$\begin{array}{rcl} \pi(\boldsymbol{\beta}|\boldsymbol{x}_i, \boldsymbol{\delta}) & \propto & \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\delta})\pi(\boldsymbol{\beta}), \\ \pi(\boldsymbol{\delta}|\boldsymbol{x}_{ij}, \boldsymbol{\beta}) & \propto & \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}, \boldsymbol{\delta})\pi(\boldsymbol{\delta}). \end{array} \tag{28}$$

Using the posterior distribution in (27), we can use several Bayesian mechanisms for estimating the parameters. Here, a *Gibbs sampling* algorithm with *Metropolis-Hasting* step inside becomes useful since the full conditional posterior distribution for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ in (28) does not have a closed-form. The algorithm described above was implemented in R language (R Development Core Team, 2015), and these codes are available upon request from the authors.

4.3 Model comparison criteria and Posterior predictive checking

In order to compare alternative models, we made use of some model comparison criteria. Specifically, we considered the Expected Akaike information criterion (EAIC), the expected Bayesian information criterion (EBIC), the deviance information criterion (DIC), and the Watanabe-Akaike information criterion, which are all described in Section 2 of Suplementary Material. Additionally, in this section, using MCMC techniques, we show how to simulate values of the posterior predictive distribution for checking the model.

BJPS - Accepted Manuscript

5 Simulation studies

This section presents three simulation studies, one that examines a prior sensitivity analysis, another investigates the recovery of the parameters of the model by the proposed estimation method and, finally, another that compares the proposed approach and existing approaches to model data with outliers. For this purpose, the Bayesian method is applied on simulated data sets from the L-Logistic distribution, considering different scenarios. For the estimation of parameters, we generated $20,000$ values from the posterior distribution in (20), then the first $10,000$ values were discarded and sequences of 10 observations were eliminated, and finally the resulting sample of size $1,000$ were used for inference.

5.1 Prior sensitivity analysis

Prior sensitivity analysis plays an important role in applied Bayesian analysis. This is especially true for Bayesian models used for new distribution, wherein the interpretability of the corresponding parameters becomes important. In this section, we consider a simulation study to evaluate the sensitivity of different choices of prior distributions for parameter b since this is different from parameter m, which is clearer in its interpretation. Specifically, we assume prior independence between parameters b and m, considering a unit uniform distribution for parameter m.

We considered five different prior distributions for b, considering simulated data sets from the L-Logistic distribution for some pairs of parameters m and b. The values of m and b used are as follows: $b \in \{0.5, 1, 5\}$ and $m \in \{0.2, 0.5, 0.9\}$, leading to nine scenarios or pairs of parameters, corresponding to nine models simulated. We simulated samples of size $n = 100$, $y_1, ...y_n$, from the L-Logistic distribution based on these pairs of parameters, then nine distinct simulated datasets were considered in the analysis.

Based on the works of Figueroa-Zúñiga *et al.* (2013), we consider for the parameter b three relatively non-informative and two informative prior distributions. The non-informative prior distributions are the gamma distribution with parameter vector $(0.001, 0.001)$ $(b \sim Gamma(0.001, 0.001))$, denoted by prior A, the uniform distribution with parameter vector $(0, 100)$ for U $(U \sim uniform(0, 100))$ with $b = U^2$, denoted by prior B, and the central Student t distribution with parameter vector $(10, 0, 2)$ $(L \sim St(10, 0, 2))$ for L with $\log(b) = L$, denoted by prior C. The prior B is chosen because it is less informative than the usual gamma with parameter vector (ϵ, ϵ). For the informative prior distributions, we consider $b \sim Gamma(2.5, 1)$, denoted by prior D, and $b \sim Gamma(50, 1)$, denoted by prior E. Note that prior E provides incorrect information about parameter b, while prior D provides almost correct information. In all the cases, the prior distribution for parameter m is taken as the uniform distribution with parameters 0 and 1, that is, $m \sim uniform(0, 1)$.

In order to compare the models with different prior distributions, we made use of some model comparison criteria. Specifically, we considered the Expected Akaike information criterion (EAIC), the expected Bayesian information criterion (EBIC), the deviance information criterion (DIC), and the Watanabe-Akaike information criterion. For a review of

BJPS-Accepted Manuscript

these criteria, one may refer to Gelman *et al.* (2013) (a brief description of these criteria is also given in the Online Supplementary Material 2). Based on these criteria, for all the simulated datasets in the nine considered scenarios, we found that prior E provided the model with the worst fit among all the fitted models. However, for the models using all other prior distributions, the values of WAIC, EAIC, EBIC, and DIC are all quite close, showing no significant difference, giving evidence that the estimated models provide almost the same quality of fit for the analyzed samples. Thus, for these cases, the posterior distribution does not seem to be sensitive with respect to the specification of these prior distributions. The values of WAIC, EAIC, EBIC and DIC for the fitted models, considering these different prior distributions, are presented in Tables A and B of the Supplementary Material 3.1.

For a more detailed analysis, additionally, we choose the non-informative priors A and C, and the worst informative prior E to present HPD intervals and point estimates. Prior A was chosen for this second analysis because it is simplest among the non-informative priors considered before, while prior C presents lower values for EAIC, EBIC and DIC than prior B in most of the studied cases. In this analysis, we observe that when prior E is used, the HPD interval does not contain the true value of b. On the other hand, the non-informative A and C priors provide intervals containing the true value of the parameters for all the cases analyzed. However, prior A provides the estimated values for the parameter b (posterior mean) closer to the true value than prior C in most cases. The posterior mean and the 95% HPD interval (obtained from the package of Martin *et al.* (2011)) can be seen in Table C of the Supplementary Material 3.1.

Considering the results discussed before, we choose priors A and C for developing a sensitivity analysis in the context of the median regression model. Here, we also consider prior independence between the parameters in which $\beta_j \sim N(0, 100)$, for $j = 0, 1, ..., q-1$. The simulated data sets for this analysis were generated from L-Logistic distribution such that

$$\begin{array}{rcl} Y_i & \sim & LL(m_i, b) \\ \text{logit}(m_i) & = & \boldsymbol{x}_i^T \boldsymbol{\beta}, \end{array} \tag{29}$$

for $i = 1, ..., n$, where $\boldsymbol{\beta} = (\beta_0, ..., \beta_{q-1})$. The $x_i's$ were generated independently from beta distributions ($beta(2,5), beta(5,1)$ and $beta(1,3)$, respectively) and their values were centered at their respective sample averages, in order to improve the convergence of the MCMC algorithm. We considered two different median regression models, the first with just one covariate (q=2) and the second with three covariates (q=4). For each median regression model, we considered three true values for the dispersion parameter b leading to six scenarios which are presented in Table 2. For each scenario, we simulated samples of size $n \in \{20, 50, 100\}$ from these models. The true values of the coefficients and the dispersion parameter for each considered model are shown in Table 2. As in the previous analysis, considering the results presented in Table 2, we can see no difference between the models fitted with priors A and C for the dispersion parameter b.

BJPS-Accepted Manuscript

Table 2 *Statistics for model comparison with different prior distributions for parameter b, using sample size of* 100, 50 *and* 20 *observations simulated from L-Logistic distribution for different values of b and the coefficient of the median regression model.*

Coefficient	Dispersion	Prior A				Prior C			
$\boldsymbol{\beta} = (\beta_0, ..., \beta_p)$	b	WAIC	EAIC	EBIC	DIC	WAIC	EAIC	EBIC	DIC
Sample size $n = 100$									
$\boldsymbol{\beta} = (-1.5, 1.5)$	0.5	99.60	-192.25	-182.44	-193.20	99.64	-192.37	-182.56	-193.44
	1	69.77	-132.59	-122.77	-133.52	69.81	-132.62	-122.81	-133.58
	5	187.88	-368.84	-359.03	-369.82	187.89	-368.73	-358.92	-369.60
$\boldsymbol{\beta} = (-3, -1.5, 1.5, 3)$	0.5	206.55	-400.50	-385.48	-403.45	206.65	-400.41	-385.38	-403.19
	1	220.62	-428.65	-413.62	-431.61	220.72	-428.73	-413.70	-431.70
	5	366.60	-720.56	-705.53	-723.46	366.60	-720.72	-705.69	-723.79
Sample size $n = 50$									
$\boldsymbol{\beta} = (-1.5, 1.5)$	0.5	42.56	-78.25	-70.52	-79.21	42.63	-78.36	-70.62	-79.40
	1	35.25	-63.59	-55.86	-64.52	35.30	-63.75	-56.01	-64.80
	5	98.36	-189.83	-182.10	-190.78	98.35	-189.92	-182.18	-190.96
$\boldsymbol{\beta} = (-3, -1.5, 1.5, 3)$	0.5	112.62	-212.86	-201.30	-215.78	112.81	-213.01	-201.45	-215.94
	1	124.40	-236.41	-224.85	-239.28	124.56	-236.63	-225.07	-239.61
	5	198.21	-384.03	-372.47	-386.89	198.22	-383.97	-372.41	-386.79
Sample size $n = 20$									
$\boldsymbol{\beta} = (-1.5, 1.5)$	0.5	30.10	-52.91	-47.92	-53.80	30.38	-53.04	-48.05	-54.00
	1	19.22	-31.18	-26.20	-32.10	19.46	-31.40	-26.42	-32.43
	5	39.81	-72.35	-67.36	-73.21	39.90	-72.47	-67.48	-73.43
$\boldsymbol{\beta} = (-3, -1.5, 1.5, 3)$	0.5	39.34	-66.33	-59.35	-69.17	39.85	-66.64	-59.66	-69.51
	1	33.03	-53.67	-46.69	-56.48	33.48	-54.10	-47.12	-57.06
	5	56.97	-101.49	-94.51	-104.19	57.10	-101.66	-94.68	-104.48

5.2 Parameter recovery and comparison of algorithms

A study of parameter recovery for the parameters of the L-Logistic distribution using prior A for the dispersion parameter b and the unit uniform prior for the parameter m was conducted and can be seen in Supplementary Material 3.2. This study showed that the proposed estimation method for the parameters of the L-Logistic distribution works quite well. Additionally, following a recommendation of the Associate Editor, we compared the proposed hybrid algorithm (Metropolis-Hastings algorithm within the Gibbs sampler) with an adaptive Metropolis-Hastings algorithm in order to estimate the median L-logistic regression. The results presented in Supplementary Material reveal that there is no difference in the recovery of the parameters of the model by both these methods.

5.3 Robustness to outliers of L-Logistic distribution

Now, we discuss a simulation study carried out to examine the robustness to outliers of the L-Logistic distribution, i.e., we discuss a study of the relative performance of the procedure for estimating the beta and L-Logistic models, with data coming from a beta distribution with outliers.

The contaminated beta data were generated following Bayes *et al.* (2012) in two steps. First, the datasets were generated from a beta distribution with location parameter $\mu = 0.2$, considering two values of the dispersion parameter, $\phi = 10, 30$, and three sample sizes, $n = 50, 100, 200$. Second, these data were contaminated with outliers generated from a

BJPS - Accepted Manuscript

uniform distribution with parameters 0.999 and 1. The proportions of outliers considered were 0.02, 0.05 and 0.08 for each dataset, i.e., $r = 2\%, 5\%, 8\%$ of the data in each dataset were randomly replaced by outliers. This gave $r \times n/100$ total outliers in each dataset containing n values. The combination of values of ϕ, n and r provides $2 \times 3 \times 3 = 18$ scenarios to be analyzed.

In order to compare the fit of beta and L-Logistic models to each of the contaminated datasets, WAIC, EAIC, EBIC and DIC were obtained for beta and L-Logistic models for 100 replications in each scenario. Thus, the percentage of cases in which the L-Logistic model achieved a lower value for WAIC, EAIC, EBIC and DIC than the beta model was determined. We found no significant difference between the two analyzed models when the DIC is used to select the model. However, the L-Logistic model performed better than beta models in all analyzed cases based on WAIC, EAIC and EBIC (see Table F of Suplementary Material 3.3).

The bias and MSE of the estimators of m and μ obtained by replicating in each scenario, considering 0.2 as the true value for the parameters m and μ, were obtained (see Table F of Supplementary Material 3.3). The bias and MSE were always smaller for the m estimator than the μ estimator showing that for any scenario with outliers, there was an improvement in the accuracy (bias and MSE decrease) for the estimation of the model parameters when using an L-Logistic model rather than the beta model for a contaminated dataset. In order to illustrate the results, the estimated densities for the scenario in which $n = 100$, $r = 5\%$ and $\phi = 10$ is shown in Figure 5, where the L-Logistic model is seem to fit the data better than the beta model.

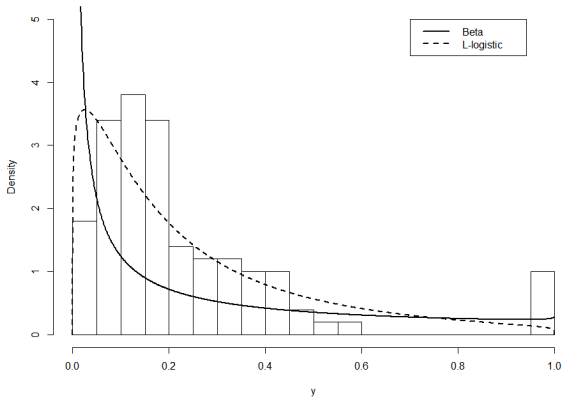

Figure 5: Estimated densities for beta and L-Logistic models for the scenario in which $n = 100$, $\phi = 10$ and $r = 5\%$.

BJPS Accepted Manuscript

6 Applications of L-Logistic distribution to real data

In order to illustrate the advantages of the use of the L-Logistic distribution in comparison to beta distribution, in sub-section 6.1 we estimate the distribution of the vulnerability to poverty in Alagoas state, Brazil. Later in sub-section 6.2, we propose different regression models to explain the anxiety as a function of stress, considering a know data set in the literature.

6.1 Estimating the distribution of the vulnerability to poverty in Alagoas state

In this subsection, we consider a real dataset, which contains the proportion of children (0-14 year olds) vulnerable to poverty. The data came from the municipalities of the state of Alagoas in Brazil, and was collected in 2010. The state of Alagoas is located in the eastern part of the Northeastern Region of Brazil and is made up of 102 municipalities. This state is one of the poorest states of Brazil and its HDI (Human Development Index) is the country's worst, based on information available in PNUD *et al.* (2013). Thus, we are interested in modeling the proportion of children vulnerable to poverty (PCVP). Here, a child is considered vulnerable to poverty if the per capita household income is at most BRL 255, in 2010. The PCVP data set comprises 102 observations and is modeled here using the L-Logistic distribution and the beta distribution that is often used to model data when a distribution over some finite interval is needed; see Gupta & Nadarajah (2004). Here, we use the re-parametrized beta distribution discussed by Ferrari & Cribari-Neto (2004) in the context of regression analysis.

The Bayesian methodology was used to estimate the parameters of both models. For the L-Logistic distribution with parameters m and b, we considered prior A discussed earlier in Section 5. Since the beta distribution has parameters $0 < \mu < 1$ and $\phi > 0$, we considered the same prior A in this model as well.

Table 3 *Estimates and 95% HPD intervals for the parameters of the L-Logistic and beta models, and model comparison criteria.*

Model	Parameter		Criteria			
			WAIC	EAIC	EBIC	DIC
L-Logistic	m	0.86(0.85, 0.87)	155.1322	-304.2996	-299.0496	-306.3422
	b	4.04(3.42, 4.72)				
beta	μ	0.85(0.84, 0.86)	150.8993	-295.3312	-290.0813	-297.3437
	ϕ	37.81(27.55, 47.83)				

The final result on the estimation is presented in Table 3. This table also shows the values of statistics for model comparison in order to evaluate the ability of L-Logistic and beta models to fit the data. According to this table, it is clear that the L-Logistic model is better for modeling the PCVP data than the beta model. In addition, Figure 6 shows two graphs with the mean values and error bars with 95% credible intervals plotted against

BJPS - Accepted Manuscript

the corresponding observed value of the data. The error bars were constructed from 1000 values (ordered, and of size 102) generated from the L-Logistic and beta distributions, respectively, for each graph, with the estimated parameters. In the case of the L-Logistic model, the bars crossed by the diagonal line $y = x$ indicate that the model is quite suitable for the data. On the other hand, in the case of the beta model, we observe high deviations between the predicted and observed data, mainly in the tail of the distribution. In this case, an observation is flagged as an outlier, since the corresponding posterior interval does not contain the observed value. Thus, Figure 6 provides evidence that the beta model is not suitable for these data. Finally, the estimated and the observed histograms of the PCVP data are presented in Figure 7, which confirms that the L-Logistic model provides a better fit for these data than the beta model.

BJPS - Accepted Manuscript

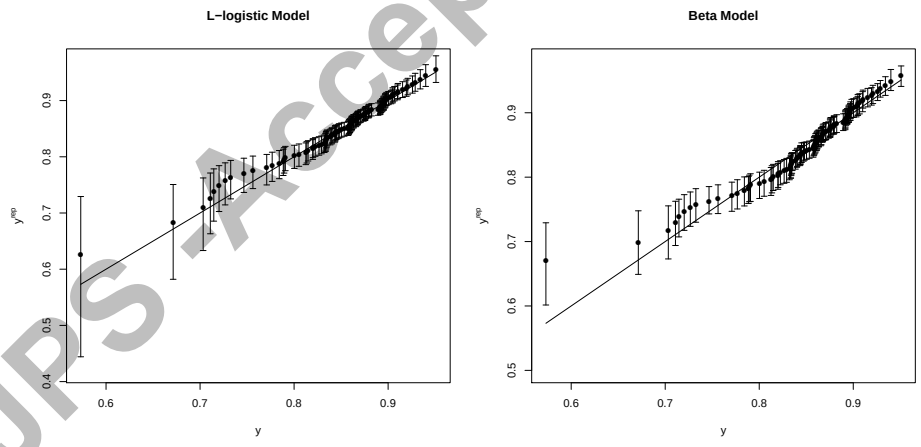

Figure 6: Posterior predictive error bars with 95% confidence intervals of the generated values $y_{(i)}^{rep}$ versus ordered observed data $y_{(i)}$ for the PCVP data, using L-Logistic and beta models.

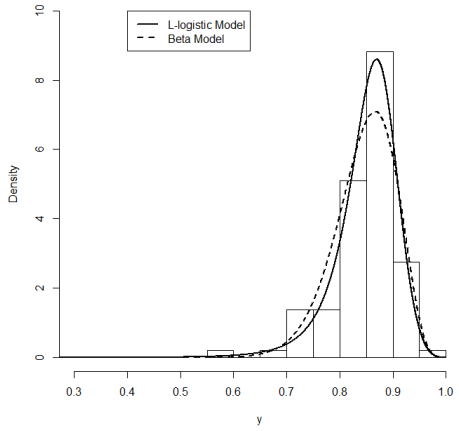

Figure 7: Observed and estimated density of PCVP data.

BJPS -Accepted Manuscript

Assuming the L-Logistic model as a final model, we can see in Table 3 that the median of proportion of children vulnerable to poverty is close to 1 with low dispersion. Therefore, based on these results we can conclude that the children vulnerable to poverty in the Alagoas state is higher and systematically present lower dispersion.

6.2 Application of the L-Logistic regression model

In order to illustrate the regression analysis considering the L-Logistic distribution, we analyzed a known data set in the literature that come from a sample of nonclinical women in Townsville, Queensland, Australia. The data set contains 166 observations on two variables, namely, the stress score and the anxiety score. Both variables were assessed on the Depression Anxiety Stress Scales, ranging from 0 to 42, but linearly transformed to the open unit interval by Smithson & Verkuilen (2006). The scatter plot of the anxiety versus stress variable, and the histograms of the data, are presented in Figure 8. The histogram given in this figure suggest that the anxiety is strongly skewed.

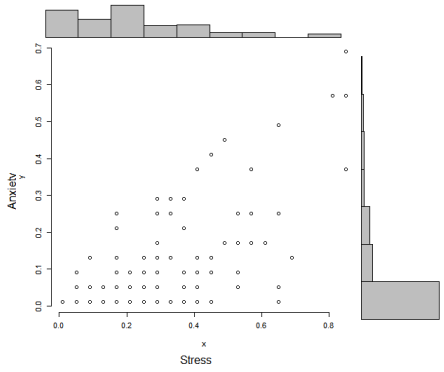

Figure 8: Scatterplot and histograms of the data in Application 6.2.

For this data, we propose four possible regression sub-models using L-Logistic distribution. We consider a null regression model without any covariate, a regression model considering only covariate effects in parameter m, a dispersion regression model considering only covariate effects in the dispersion parameter b, and full regression model considering both effects, as follows:

$$Y_i \sim LL(m_i, b_i) \text{ and } \begin{cases} \text{null model } (L_0): & logit(m_i) = \beta_0 \text{ and } \log(b_i) = -\boldsymbol{\delta}, \\ \text{median-model } (\boldsymbol{L_1}): & logit(m_i) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \text{ and } log(b_i) = -\delta_0, \\ \text{dispersion model } (L_2): & logit(m_i) = \beta_0 \text{ and } \log(b_i) = -\boldsymbol{x}_{1i}^T\boldsymbol{\delta}, \\ \text{full model } (L_3): & logit(m_i) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \text{ and } \log(b_i) = -\boldsymbol{x}_{2i}^T\boldsymbol{\delta}, \end{cases}$$

for $i = 1, ..., 166$. In addition, we also consider equivalent regression models using the beta distribution as follows:

$$Y_i \sim beta(\mu_i, \phi_i) \text{ and } \begin{cases} \text{null model } (B_0): & logit(\mu_i) = \beta_0 \text{ and } \log(\phi_i) = -\boldsymbol{\delta}, \\ \text{mean-model } (B_1): & logit(\mu_i) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \text{ and } log(b_i) = -\delta_0, \\ \text{dispersion model } (B_2): & logit(\mu_i) = \beta_0 \text{ and } \log(\phi_i) = -\boldsymbol{x}_{1i}^T\boldsymbol{\delta}, \\ \text{full model } (B_3): & logit(\mu_i) = \boldsymbol{x}_{1i}^T\boldsymbol{\beta} \text{ and } \log(\phi_i) = -\boldsymbol{x}_{2i}^T\boldsymbol{\delta}, \end{cases}$$

for $i = 1, ..., 166$.

Here, the Bayesian approach is considered for the inference process with a prior distribution for the unknown regression parameters, as given in (26). All the algorithms were prepared in the R language and we report the results corresponding to $10,000$ iterations following a burn-in period also of $10,000$ iterations. In order to eliminate dependence, we

BJPS - Accepted Manuscript

eliminated a sequence of 10 observations every 11 simulations in the sample of size 10,000, resulting in a final sample of 1,000 elements. Finally, the convergence of MCMC chain was assessed by using the separated partial means test of Geweke (1992), which provided evidence for the chains to have converged.

Table 4 *Model comparison criteria for model comparison.*

	L-Logistic model				beta model			
Sub model	WAIC	EAIC	EBIC	DIC	WAIC	EAIC	EBIC	DIC
0	259.34	-512.94	-506.72	-514.92	239.45	-472.90	-466.67	-474.90
1	277.67	-545.51	-536.17	-548.56	243.28	-478.06	-468.72	-481.07
2	316.57	-624.78	-615.44	-627.83	283.41	-556.95	-547.62	-559.92
3	319.65	-627.47	-615.02	-631.48	301.91	-591.85	-579.41	-595.82

The regression models investigated were compared by the use of EAIC, EBIC, DIC and WAIC criteria, and the obtained results are shown in Table 4. The parameter estimates for these models are shown in Table 5. Considering Table 4, we can observe that the regression models considering L-Logistic distribution provides a better fit than the corresponding beta regression models, for all the criteria considered. These results also give evidence that L_3 and L_2 are the best models among the ones based on the L-Logistic distribution. Though there is no significant difference between the L_2 and L_3 regressions, we consider the L_3 regression model to be a reasonable choice for this data set, due to the expected influence on covariates in the dispersion parameter (Smithson & Verkuilen, 2006).

Moreover, a posterior distribution of residuals was obtained and a posterior mean of this distribution was computed (Gelman *et al.*, 2014). That is, for $i = 1, ..., 166$, we have $\hat{r}_i = G^{-1} \sum_{g=1}^{G} \frac{y_i - \hat{y}^g}{SD(Y_i|\boldsymbol{\beta}^g)}$, where $\boldsymbol{\beta}^1, ..., \boldsymbol{\beta}^G$ are obtained from the posterior distribution, $\hat{y}$ is the estimated value of a data point y_i, and $SD(Y|\boldsymbol{\beta}^g)$ is the standard deviation of posterior values of Y, both obtained given a single random draw $\boldsymbol{\beta}^g$ of the posterior distribution. Figure 9 shows the standard residual versus the estimated values in which we can see that the L_3 regression model provides a better fit than the B_3 model, which confirms that L_3 model are better than the corresponding B_3 model.

BJPS - Accepted Manuscript

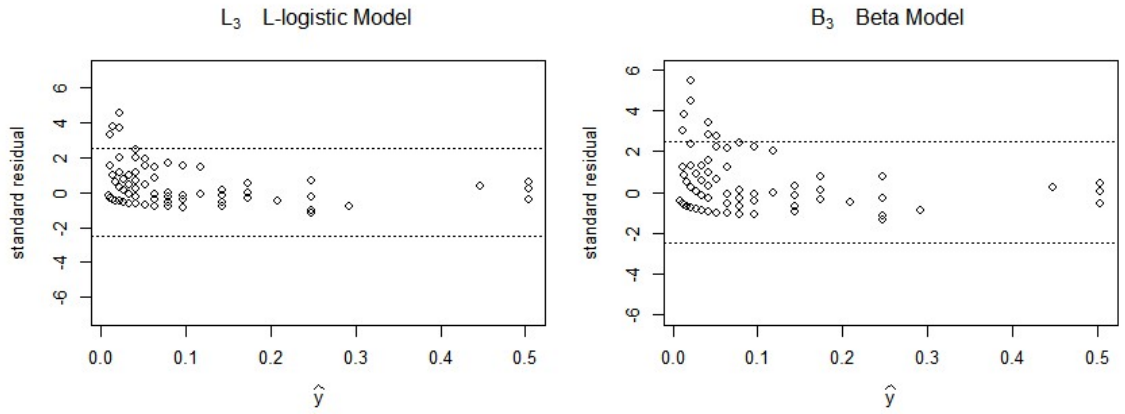

Figure 9: Standard residual versus adjusted values for the L-Logistic and beta models.

Table 5 *Parameter estimates and 95% HPD intervals for the L-Logistic and beta models.*

Model		Coefficient			
		$\hat{\beta_0}$(HPD)	$\hat{\beta_1}$(HPD)	$\hat{\delta_0}$(HPD)	$\hat{\delta_1}$(HPD)
L-Logistic	L_0	-3.354 (-3.615, -3.106)	-	-0.08(-0.20, 0.04)	-
	L_1	-4.78(-5.04, -4.52)	5.78(4.99, 6.58)	-0.44(- 0.57,-0.31)	-
	L_2	-4.03(-4.27, -3.78)	-	-0.87(-1.14,-0.58)	2.56(1.62,3.42)
	L_3	-4.77(-5.00, -4.53)	5.64(4.68, 6.61)	-0.76(-1.03, -0.48)	1.14(0.19, 2.05)
beta	B_0	-2.239(-2.430, -2.04)	-	-1.78(-2.02,-1.54)	-
	B_1	-3.47(-3.75, -3.18)	3.74(3.11, 4.37)	-2.44(-2.7,-2.20)	-
	B_2	-2.54(-2.80, -2.27)	-	-2.49(-2.98,-1.95)	1.53(0.42,2.55)
	B_3	-4.02 (-4.30, -3.72)	4.95(4.09, 5.83)	-3.94(-4.45,-3.44)	4.28 (2.78,5.79)

Finally, from Table 5 giving the 95% HPD intervals for all the coefficients of the models under analysis, we can see that the estimates are quite precise. For the model chosen, that is, L_3 model, we observe that the HPD intervals for the estimates of the parameters β_1 and δ_1 do not contain zero giving evidence that the parameters in the model are significant in the model. In other words, stress is important in both parameters of the distribution of anxiety.

7 Final remarks

The L-Logistic distribution, introduced by Tadikamalla & Johnson (1982), is a bounded continuous distribution that possesses some nice properties, as discussed in Section 2. Considering the parametrization introduced in this manuscript, we propose a Bayesian estimator by considering an MCMC method as an alternative to the moment and maximum likelihood methods developed previously in the literature. In the Bayesian context, a non-informative prior distribution can be adopted for the parameter m since it lies in the

BJPS -Accepted Manuscript

unit interval, enabling the use of unit uniform distribution as a non-informative prior distribution.

The main motivation of the parametrization introduced here is the development of regression models based on the L-Logistic distribution. We also introduce conditional median regression models, which is a special case of quantile regression wherein a conditional quantile is modeled as a function of covariates.

Two applications have been considered in this work. Firstly, we consider an application to social data, wherein the proportion of children vulnerable to poverty of the municipalities of the state of Alagoas in Brazil, for the 2010 season, is modeled. Secondly, we analyze a known data set, previously analyzed using beta distribution by Smithson & Verkuilen (2006), which contains the stress score and the anxiety score. Here, the anxiety variable is modeled as a function of the stress. In the case of the L-Logistic distribution, we use a regression model proposed in this work. In these applications, we observe that the L-Logistic distribution seems to fit better than the beta model for both this analyzed cases. Considering the application to the anxiety data set (Anxiety explained by stress), we show that the L-Logistic regression models can be a good alternative to the beta model. An advantage of this approach is the possibility of modeling other quantiles in order to describe a non-central position of a distribution. So, one may choose a position specifically for his/her needs. For example, it is possible to consider a regression model to explain other quantiles to the Anxiety considering the influence of the stress in our application. Thus, conditional quantile models offer the flexibility to focus on these population segments, whereas conditional mean models do not. However, since quantile regression curves are estimated individually, the quantile curves can cross, leading to an invalid distribution for the response. Thus, this problem, referred to as crossing in the literature, needs to be studied carefully. Some authors have proposed methods to deal with this problem; see, for example, Cai & Jiang (2015).

In the future, we aim to develop techniques for mixed quantile regression for the L-Logistic distribution. Moreover, we intend to explore mixtures of L-Logistic distributions in a Bayesian framework as well as a multivariate version of this distribution.

8 Acknowledgments

This work was supported in part by the Coordenao de Aperfeioamento do Pessoal de Ensino Superior (CAPES-Brazil). The first author thanks the support from CAPES-Brazil. The last author was partially supported by FAPESP-Brazil 2017/15452-5. The authors also thank the Editor, the Associate Editor, and the Referees for their useful comments and suggestions, which resulted in an improvement in the original version of the article.

BJPS -Accepted Manuscript

References

Albert, I., Donnet, S., Guihenneuc-Jouyaux, C., Low-Choy, S., Mengersen, K. & Rousseau, J. (2012). Combining expert opinions in prior elicitation. *Bayesian Anaysisl.*, **7**(3), 503–532.

Arnold, B. C. & Groeneveld, R. A. (1995). Measuring skewness with respect to the mode. *The American Statistician*, **49**(1), 34–38.

Bayes, C., Bazán, J. L. & García, C. (2012). A new robust regression model for proportions. *Statistics and Its Interface*, **7**(4), 841–866.

Bayes, L. C., Bazán, J. L. & Castro, M. (2017). A quantile parametric mixed regression model for bounded response variables. *Statistics and Its Interface*, **10**(3), 483–493.

Brys, G., Hubert, M. & Struyf, A. (2003). A comparison of some new measures of skewness. In R. Dutter, P. Filzmoser, U. Gather, & P. J. Rousseeuw, editors, *Developments in Robust Statistics*, pages 98–113. Springer-Verlag, New York.

Buckley, J. (2003). Estimation of models with beta-distributed dependent variables: A replication and extension of paolino's study. *Political Analysis*, **11**(2), 204–205.

Cai, Y. & Jiang, T. (2015). Estimation of non-crossing quantile regression curves. *Australian and New Zealand Journal of Statistics*, **57**(1), 139–162.

Ferrari, S. & Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, **31**(7), 799–815.

Figueroa-Zúñiga, J. I., Arellano-Valle, R. B. & Ferrari, S. L. P. (2013). Mixed beta regression: A Bayesian perspective. *Computational Statistics and Data Analysis*, **61**(1), 137–147.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013). *Bayesian Data Analysis*. Third Edition, Taylor & Francis, Philadelphia, PA.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC Press, Boca Raton, FL, USA.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernardo, J. Berger, A. P. Dawid, & J. F. M. Smith, editors, *Bayesian Statistics*, pages 169–193. Oxford University Press, London, England.

Gómez-Déniz, E., Sordo, M. A. & Calderín-Ojeda, E. (2014). The log-Lindley distribution as an alternative to the beta regression model with applications in insurance. *Insurance: Mathematics and Economics*, **54**(1), 49–57.

Gupta, A. & Nadarajah, S. (2004). *Handbook of Beta Distribution and Its Applications*. Taylor & Francis, Philadelphia.

Hao, L. & Naiman, D. (2007). *Quantile Regression*. SAGE Publications, New Jersey.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika*, **62**(1), 101–111.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, **36**(1/2), 149–176.

Johnson, N. L. & Tadikamalla, P. R. (1991). Translated family of distribution. In N. Balakrishnan, (editor). In *Handbook of the Logistic Distribution, Chapeter 8*, pages 189–208. Marcel Derrer, New York.

Koenker, R. & Bassett, J. G. (1978). Regression quantiles. *Econometrica*, **46**(1), 33–50.

Lemonte, A. J. & Bazán, J. L. (2016). New class of Johnson SB distributions and its associated regression model for rates and proportions. *Biometrical Journal*, **58**(4), 727–746.

Martin, A. D., Quinn, K. M. & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, **42**(9), 1–22.

Moors, J. J. A. (1988). A quantile alternative for kurtosis. *Journal of the Royal Statistical Society, Series B*, **37**(1), 25–32.

Paz, R. F., Bazán, J. L. & Milan, L. A. (2015). Bayesian estimation for a mixture of simplex distributions with an unknown number of components: HDI analysis in Brazil. *Journal of Applied Statistics*, **44**(9), 1–14.

PNUD, IPEA & FJP. (2013). *Atlas do Desenvolvimento Humano no Brasil*. PNUD, Brasilia, Brazil.

BJPS : Accepted Manuscript

Disponible in: http://www.atlasbrasil.org.br/2013/pt/.

R Development Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smithson, M. & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**(1), 1–54.

Tadikamalla, P. R. & Johnson, N. L. (1982). Systems of frequency curves generated by transformations of logistic variables. *Biometrika*, **69**(2), 461.

Tadikamalla, P. R. & Johnson, N. L. (1990). Tables to facilitate fitting Tadikamalla and Johnson's LB distributions. *Communications in Statistics - Simulation and Computation*, **19**(4), 1201–1229.

Wang, M. & Rennolls, K. (2005). Tree diameter distribution modelling: introducing the logitlogistic distribution. *Canadian Journal of Forest Research*, **35**(6), 1305–1313.

BJPS -Accepted Manuscript