Thailand Statistician
April 2025; 23(2): 316-330
http://statassoc.or.th
Contributed paper

The Unit Weibull Regression Model with Variable Shape Parameter

Lucas David Ribeiro-Reis*
Faculty of Statistics, Institute of Exact and Natural Sciences,
Federal University of Pará, Belém-PA, Brazil
*Corresponding author; e-mail: econ.lucasdavid@gmail.com

Received: 7 April 2022
Revised: 26 September 2022
Accepted: 27 November 2022

Abstract

In this paper a generalization of the unit Weibull regression model is introduced. Here, both the median and the shape parameter are modelled through covariates. The parameters are estimated by maximum likelihood. Analytical expressions for the score vector and the Fisher's observed information matrix are demonstrated. A simulation study is performed to show the consistency of the maximum likelihood estimators. Finally two applications to real data from Brazil are considered. These applications show the usefulness of the proposed model.

Keywords: Bounded data analyses, maximum likelihood, median, regression model, Northeast of the Brazil, Covid-19

1. Introduction

In applied statistics, many phenomena lie on the interval $(0, 1)$, such as inequality index, homicide rates, proportion of people on the poverty line, etc. The distributions most commonly used for analysis of these types of data are the beta and Kumaraswamy distributions. Recently, several other new distributions for the interval $(0, 1)$ have been proposed, some of them being: log-Lindley (Goméz-Déniz et al., 2014), log-xgamma (Altun and Hamedani, 2018), unit Weibull (Mazucheli et al., 2018), unit Gompertz (Mazucheli et al., 2019), unit Burr-III (Modi and Gill, 2020), unit modified Burr-III (Haq et al., 2020) and unit Burr-XII (Korkmaz and Chesneau, 2021).

Often, it is sought to analyze the behavior of such variables in the presence of exogenous variables. This analysis is made through regression models. The pioneering work was the beta regression model, introduced by Ferrari and Cribari-Neto (2004). In this work, the authors assume that the dependent variable follows the beta distribution. And then, through a link function, a regression model that models the mean of the distribution is proposed.

More recently, Mitnik and Baek (2013) introduced the Kumaraswamy regression model. The authors use the same idea as the model proposed by Ferrari and Cribari-Neto (2004), but the difference is that instead of the mean, what is modelled is the median of the distribution. Thus, in the presence of very asymmetric data or data with atypical values, the Kumaraswamy regression model is more robust than the beta regression model.

From unit Weibull (Mazucheli et al., 2018) distribution, (Mazucheli et al., 2020) proposed a regression model, called the unit Weibull regression model. In this model, the response variable is assumed to have unit Weibull distribution. Like Kumaraswamy regression model, the unit Weibull regression model has modeling on the median of the distribution. Thus, these two models are quite competitive.

This paper aims to extend the unit Weibull regression model by including exogenous variables in the modeling of the shape parameter. Thus, this new regression model has covariates to model the median and the shape parameter. When no covariates are included in the shape parameter modeling, the model with fixed shape parameter is obtained. Parameter estimates are performed by maximum likelihood. Analytical expressions for the score vector and for the Fisher's observed information matrix are demonstrated. In Mazucheli et al. (2020) these expressions were not shown.

This paper is organized as follows. In Section 2, the unit Weibull regression model with variable shape parameter is introduced. In Section 3, the maximum likelihood estimation method, score vector, Fisher's observed information matrix and calculation of the confidence intervals for the parameters are demonstrated. In Section 4, Monte Carlo simulation to show the performance of the maximum likelihood estimators is presented. In Section 5, two applications to real data is considered. Finally, the paper is concluded in Section 6.

2. Regression model

The cumulative distribution function (cdf) and probability density function (pdf) of the unit Weibull (UW) distribution with parameterization in the median, as defined by Mazucheli et al. (2020), are given by

$$F(y;\tau,\phi) = 2^{-\left(\frac{\ln y}{\ln \tau}\right)^{\phi}}, \quad 0 < y < 1 \tag{1}$$

and

$$f(y;\tau,\phi) = \frac{\phi \ln 0.5}{y \ln \tau}\left(\frac{\ln y}{\ln \tau}\right)^{\phi-1} 2^{-\left(\frac{\ln y}{\ln \tau}\right)^{\phi}}, \quad 0 < y < 1, \tag{2}$$

respectively, where $0 < \tau < 1$ is the median parameter and $\phi > 0$ is shape parameter. The random variable Y with cdf (1) is denoted as $Y \sim \text{UW}(\tau,\phi)$.

The quantile function corresponding to cdf (1) is given by

$$Q(u;\tau,\phi) = \tau^{\left(\frac{\ln u}{\ln 0.5}\right)^{1/\phi}}, \quad 0 < u < 1.$$

Using the quantile function, if V is a random variable having standard uniform distribution, say $V \sim \mathcal{U}(0,1)$, then the random variable

$$Y = \tau^{\left(\frac{\ln V}{\ln 0.5}\right)^{1/\phi}} \tag{3}$$

has UW distribution with density function (2).

Let the independent random variables $Y_i \sim \text{UW}(\tau_i,\phi_i)$, with observed values y_i, $i = 1, \cdots, n$. The proposed model for the τ_i (median) and the ϕ_i (shape parameter) of y_i is given by

$$\eta_i = g(\tau_i) = \mathrm{x}_i^\top \beta = \sum_{m=1}^{k} x_{im}\beta_m \tag{4}$$

and

$$\nu_i = h(\phi_i) = \mathrm{z}_i^\top \delta = \sum_{r=1}^{q} z_{ir}\delta_r, \tag{5}$$

respectively, where $\beta = (\beta_1, \cdots, \beta_k)^\top$ and $\delta = (\delta_1, \cdots, \delta_q)^\top$ are k-vector and q-vector of unknown parameters, $\mathrm{x}_i = (x_{i1}, \cdots, x_{ik})^\top$ and $\mathrm{z}_i = (z_{i1}, \cdots, z_{iq})^\top$ are exogenous variables $(k + q < n)$, which are assumed fixed and known. η_i and ν_i are the linear predictors. For model with intercepts, it is assumed that $x_{i1} = 1$ and $z_{i1} = 1$, $\forall\, i$. The $g(\cdot)$ and $h(\cdot)$ are link functions strictly monotonic and twice differentiable, such that $g : (0,1) \to \mathbb{R}$ and $h : \mathbb{R}^+ \to \mathbb{R}$. Examples of some $g(\cdot)$ link functions can be:

- logit: $g(\tau_i) = \ln[\tau_i/(1-\tau_i)]$;
- Gumbel type I: $g(\tau_i) = -\ln(-\ln \tau_i)$;
- Gumbel type II: $g(\tau_i) = \ln[-\ln(1-\tau_i)]$;
- Cauchy: $g(\tau_i) = \tan(\pi(\tau_i - 0.5))$.

For the $h(\cdot)$ link function the logarithmic function can be used, thus $h(\phi_i) = \ln \phi_i$. If the logit specification for τ_i and the logarithmic specification for ϕ_i are used, then

$$\tau_i = g^{-1}(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)} \quad \text{and} \quad \phi_i = h^{-1}(\nu_i) = \exp(\nu_i).$$

Note that, if $\delta_2 = \cdots = \delta_q = 0$, the UW regression model with fixed shape parameter (Mazucheli et al., 2020) is obtained, with $\phi_i = \phi = \exp(\delta_1)$.

3. Maximum Likelihood Estimation

From pdf (2), the log-likelihood function for a sample of n independent observations of $y_i \sim \mathrm{UW}(\tau_i, \phi_i)$, under the structure of the regression model (4)-(5), is given by

$$\mathcal{L}(\beta,\delta) = \sum_{i=1}^{n} \mathcal{L}_i(\tau_i,\phi_i),$$

where
$$\mathcal{L}_i(\tau_i,\phi_i) = \ln\phi_i - \ln\left(\frac{\ln 0.5}{\ln \tau_i}\right) - \ln y_i + (\phi_i - 1)\ln\left(\frac{\ln y_i}{\ln\tau_i}\right) - \ln(2)\left(\frac{\ln y_i}{\ln \tau_i}\right)^{\phi_i}.$$

The differential total of $\mathcal{L}(\beta,\delta)$ is given by

$$U_{\beta_j}(\beta,\delta) = \frac{\partial \mathcal{L}(\beta,\delta)}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i(\tau_i,\phi_i)}{\partial \tau_i}\frac{\mathrm{d}\tau_i}{\mathrm{d}\eta_i}\frac{\partial \eta_i}{\partial \beta_j},$$
$$U_{\delta_p}(\beta,\delta) = \frac{\partial \mathcal{L}(\beta,\delta)}{\partial \delta_p} = \sum_{i=1}^{n} \frac{\partial \mathcal{L}_i(\tau_i,\phi_i)}{\partial \phi_i}\frac{\mathrm{d}\phi_i}{\mathrm{d}\nu_i}\frac{\partial \nu_i}{\partial \delta_p}.$$

Note that, $\mathrm{d}\tau_i/\mathrm{d}\eta_i = 1/g'(\tau_i)$, $\partial\eta_i/\partial\beta_j = x_{ij}$, $\mathrm{d}\phi_i/\mathrm{d}\nu_i = 1/h'(\phi_i)$ and $\partial\nu_i/\partial\delta_p = z_{ip}$, where $r'(w) = \mathrm{d}r(w)/\mathrm{d}w$. Already, the first derivatives of $\mathcal{L}_i(\tau_i,\phi_i)$ with respect to τ_i and ϕ_i are

$$\begin{aligned}\frac{\partial \mathcal{L}_i(\tau_i,\phi_i)}{\partial \tau_i} &= -\frac{\phi_i}{\tau_i \ln \tau_i} + \frac{\phi \ln(2)}{\tau_i \ln\tau_i}\left(\frac{\ln y_i}{\ln \tau_i}\right)^{\phi_i} \\ &= -b_i(1-\dot{y}_i) \end{aligned} \tag{6}$$

and
$$\begin{aligned}\frac{\partial \mathcal{L}_i(\tau_i,\phi_i)}{\partial \phi_i} &= \frac{1}{\phi_i} + \ln\left(\frac{\ln y_i}{\ln \tau_i}\right) - \ln(2)\left(\frac{\ln y_i}{\ln\tau_i}\right)^{\phi_i}\ln\left(\frac{\ln y_i}{\ln \tau_i}\right) \\ &= \frac{1}{\phi_i} + \ddot{y}_i - \dot{y}_i\ddot{y}_i, \end{aligned} \tag{7}$$

respectively, where $b_i = \frac{\phi_i}{\tau_i \ln \tau_i}$, $\dot{y}_i = \ln(2)\left(\frac{\ln y_i}{\ln \tau_i}\right)^{\phi_i}$ and $\ddot{y}_i = \ln\left(\frac{\ln y_i}{\ln\tau_i}\right)$.

Thus, the score vector of β_j and δ_p are given by

$$U_{\beta_j}(\beta,\delta) = \sum_{i=1}^{n} \frac{d_i}{g'(\tau_i)} x_{ij},$$
$$U_{\delta_p}(\beta,\delta) = \sum_{i=1}^{n} \frac{v_i}{h'(\phi_i)} z_{ip},$$

where $d_i = -b_i(1-\dot{y}_i)$ and $v_i = 1/\phi_i + \ddot{y}_i - \dot{y}_i\ddot{y}_i$.

In matrix form, the score vector of β and δ can be written as $U_\beta(\beta,\delta) = X^\top G\, d$ and $U_\delta(\beta,\delta) = Z^\top H\, v$, respectively, where X is a $n \times k$ matrix whose ith row is $\mathrm{x}_i^\top$, Z is a $n \times q$ matrix whose ith row is $\mathrm{z}_i^\top$, $G = \mathrm{diag}\{1/g'(\tau_i)\}$, $H = \mathrm{diag}\{1/h'(\phi_i)\}$, with $\mathrm{diag}\{q_i\}$ denoting the $n \times n$ diagonal matrix with typical element q_i, $d = (d_1, \cdots, d_n)^\top$ and $v = (v_1, \cdots, v_n)^\top$.

The maximum likelihood estimators (MLEs) of β and δ, says $\hat{\beta}$ and $\hat{\delta}$, are the solutions of

$$\begin{cases} U_\beta(\beta,\delta) = 0, \\ U_\delta(\beta,\delta) = 0. \end{cases}$$

The above system does not have a closed-form solution. Thus, non-linear numerical optimization methods are needed to obtain the MLEs of β and δ. These methods require initial values for the parameters. For the sub-model of the median, initial guesses for β is the ordinary least squares estimator of the regression of $g(y)$ on X, which is given by $\beta^{(0)} = (X^\top X)^{-1} X^\top g(y)$. Already, for the sub-model of the shape parameter, initial guess for δ is $\delta^{(0)} = 0$, i.e., $\delta_r^{(0)} = 0$, $r = 1, \cdots, q$.

Some software that can be used to obtain these MLEs are: Ox Console (Doornik, 2018) with `MaxBFGS` or `MaxSQP` functions and R Project (R Core Team, 2020) with `optim` or `nlminb` functions.

The second derivatives of $\mathcal{L}(\beta,\delta)$ with respect to β_l and δ_s are given by

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}(\beta,\delta)}{\partial\beta_j\partial\beta_l} &= \sum_{i=1}^n \frac{\partial}{\partial\tau_i}\left(\frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i}\frac{1}{g'(\tau_i)}x_{ij}\right)\frac{\mathrm{d}\tau_i}{\mathrm{d}\eta_i}\frac{\partial\eta_i}{\partial\beta_l} \\
&= \sum_{i=1}^n \left[\frac{\partial^2\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i^2}\frac{1}{g'(\tau_i)}x_{ij} + \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i}\frac{\partial}{\partial\tau_i}\left(\frac{1}{g'(\tau_i)}\right)x_{ij}\right]\frac{1}{g'(\tau_i)}x_{il} \\
&= \sum_{i=1}^n \left[\frac{\partial^2\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i^2}\frac{1}{g'(\tau_i)^2}x_{ij}x_{il} - \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i}\frac{g''(\tau_i)}{g'(\tau_i)^3}x_{ij}x_{il}\right], \\
\frac{\partial^2 \mathcal{L}(\beta,\delta)}{\partial\beta_j\partial\delta_s} &= \sum_{i=1}^n \frac{\partial}{\partial\phi_i}\left(\frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i}\frac{1}{g'(\tau_i)}x_{ij}\right)\frac{\mathrm{d}\phi_i}{\mathrm{d}\nu_i}\frac{\partial\nu_i}{\partial\delta_s} \\
&= \sum_{i=1}^n \left[\frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i\partial\phi_i}\frac{1}{g'(\tau_i)}x_{ij} + \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i}\frac{\partial}{\partial\phi_i}\left(\frac{1}{g'(\tau_i)}\right)x_{ij}\right]\frac{1}{h'(\phi_i)}z_{is} \\
&= \sum_{i=1}^n \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i\partial\phi_i}\frac{1}{g'(\tau_i)}\frac{1}{h'(\phi_i)}x_{ij}z_{is}, \\
\frac{\partial^2 \mathcal{L}(\beta,\delta)}{\partial\delta_p\partial\delta_s} &= \sum_{i=1}^n \frac{\partial}{\partial\phi_i}\left(\frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\phi_i}\frac{\mathrm{d}\phi_i}{\mathrm{d}\nu_i}\frac{\partial\nu_i}{\partial\delta_p}\right)\frac{\mathrm{d}\phi_i}{\mathrm{d}\nu_i}\frac{\partial\nu_i}{\partial\delta_s} \\
&= \sum_{i=1}^n \left[\frac{\partial^2\mathcal{L}_i(\tau_i,\phi_i)}{\partial\phi_i^2}\frac{1}{h'(\phi_i)}z_{ip} + \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\phi_i}\frac{\partial}{\partial\phi_i}\left(\frac{1}{h'(\phi_i)}\right)z_{ip}\right]\frac{1}{h'(\phi_i)}z_{is} \\
&= \sum_{i=1}^n \left[\frac{\partial^2\mathcal{L}_i(\tau_i,\phi_i)}{\partial\phi_i^2}\frac{1}{h'(\phi_i)^2}z_{ip}z_{is} - \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\phi_i}\frac{h''(\phi_i)}{h'(\phi_i)^3}z_{ip}z_{is}\right].
\end{aligned}$$

Note that, from Equations (6) and (7),

$$\begin{aligned}
t_i &= \frac{\partial^2\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i^2} = \frac{\phi_i(1+\ln\tau_i)}{(\tau_i\ln\tau_i)^2}(1-\dot{y}_i) - b_i^2\dot{y}_i, \\
r_i &= \frac{\partial\mathcal{L}_i(\tau_i,\phi_i)}{\partial\tau_i\partial\phi_i} = -\frac{1}{\tau_i\ln\tau_i}(1-\dot{y}_i) + b_i\dot{y}_i\ddot{y}_i \\
\text{and} \qquad w_i &= \frac{\partial^2\mathcal{L}_i(\tau_i,\phi_i)}{\partial\phi_i^2} = -\frac{1}{\phi_i^2} - \dot{y}_i\ddot{y}_i^2.
\end{aligned}$$

So,

$$J_{\beta_j\beta_l}(\beta,\delta)=\frac{\partial^2\mathcal{L}(\beta,\delta)}{\partial\beta_j\partial\beta_l}=\sum_{i=1}^{n}\left[\frac{t_i}{g'(\tau_i)^2}x_{ij}x_{il}-\frac{d_i g''(\tau_i)}{g'(\tau_i)^3}x_{ij}x_{il}\right],$$

$$J_{\beta_j\delta_s}(\beta,\delta)=\frac{\partial^2\mathcal{L}(\beta,\delta)}{\partial\beta_j\partial\delta_s}=\sum_{i=1}^{n}r_i\frac{1}{g'(\tau_i)}\frac{1}{h'(\phi_i)}x_{ij}z_{is},$$

$$J_{\delta_p\delta_s}(\beta,\delta)=\frac{\partial^2\mathcal{L}(\beta,\delta)}{\partial\delta_p\partial\delta_s}=\sum_{i=1}^{n}\left[\frac{w_i}{h'(\phi_i)^2}z_{ip}z_{is}-\frac{v_i h''(\phi_i)}{h'(\phi_i)^3}z_{ip}z_{is}\right].$$

Let $T=\text{diag}\{t_i\}$, $R=\text{diag}\{r_i\}$, $W=\text{diag}\{w_i\}$, $D=\text{diag}\{d_i g''(\tau_i)\}$, $V=\text{diag}\{v_i h''(\phi_i)\}$, in matrix form, this expressions can be written as $J_{\beta\beta}(\beta,\delta)=X^\top TG^2X-X^\top DG^3X$, $J_{\beta\delta}(\beta,\delta)=J_{\delta\beta}(\beta,\delta)^\top=X^\top RGHZ$ and $J_{\delta\delta}(\beta,\delta)=Z^\top WH^2Z-Z^\top VH^3Z$. Thus, Fisher's observed information matrix is

$$\mathcal{J}(\beta,\delta)=-\begin{bmatrix} J_{\beta\beta}(\beta,\delta) & J_{\beta\delta}(\beta,\delta)\\ J_{\delta\beta}(\beta,\delta) & J_{\delta\delta}(\beta,\delta)\end{bmatrix}.$$

For large n, and assuming that $\mathcal{J}(\beta,\delta)$ is a non-singular matrix, the MLEs have the following asymptotic distribution

$$\begin{pmatrix}\hat{\beta}\\ \hat{\delta}\end{pmatrix}\quad\overset{a}{\sim}\quad\mathcal{N}_{k+q}\left(\begin{pmatrix}\beta\\ \delta\end{pmatrix},\mathcal{J}(\beta,\delta)^{-1}\right).$$

Thus, confidence intervals and hypothesis testing can be performed using the normal distribution. Based on asymptotic distribution, the $100(1-\gamma)\%$ confidence intervals for β_m $(m=1,\cdots,k)$ and δ_r $(r=1,\cdots,q)$ are given by

$$\hat{\beta}_m\pm z_{(1-\gamma/2)}\sqrt{L_{jj}}\quad\text{and}\quad\hat{\delta}_r\pm z_{(1-\gamma/2)}\sqrt{L_{(k+r)(k+r)}},$$

respectively, where $z_{(1-\gamma/2)}$ is the $(1-\gamma/2)$ quantile of the standard normal distribution and L_{ii} denotes the ith diagonal element of the matrix $\mathcal{J}(\beta,\delta)^{-1}$.

3.1. Likelihood ratio test

Since that the UW regression model with fixed shape parameter (restricted model) is a special case of the UW regression model with variable shape parameter (unrestricted model) when $\delta_2=\cdots=\delta_q=0$, the likelihood ratio (LR) test between these models can be is used. Thus, the null hypothesis ($\mathcal{H}_0$) and the alternative hypothesis ($\mathcal{H}_1$) of the LR test are given by $\mathcal{H}_0:\delta_2=\cdots=\delta_q=0$ and $\mathcal{H}_1:\mathcal{H}_0$ *is false*, respectively. Let $\Theta=(\beta,\delta)^\top$ and $\Theta_0=(\beta,\delta_1)^\top$ the vector of the unrestricted and restricted models, respectively. Note that, Θ is a vector of dimension $k+q$, while Θ_0 is a vector of dimension $k+1$. Thus, the difference of parameters between the unrestricted and the restricted model is of $q-1$. The LR test is given by

$$\mathcal{T}=2[\mathcal{L}(\hat{\Theta})-\mathcal{L}(\tilde{\Theta}_0)],$$

where $\hat{\Theta}$ and $\tilde{\Theta}_0$ are the MLEs under the unrestricted and restricted models, respectively. For n large and under $\mathcal{H}_0$, $\mathcal{T}\overset{a}{\sim}\chi^2_{q-1}$, where χ^2_{q-1} is the chi-square distribution with $q-1$ degree-freedom. Thus, $\mathcal{H}_0$ is rejected when $\mathcal{T}>\chi^2_{q-1}(1-\gamma)$, in which $\chi^2_{q-1}(1-\gamma)$ is the quantile $(1-\gamma)$ of χ^2_{q-1}. Once $\mathcal{H}_0$ is rejected, then the unrestricted model is more appropriate.

3.2. Diagnostic analysis

In this Section, some diagnostic measures will be described. The Cox-Snell and the Dunn-Smith residuals are often used to verify the adequacy of the estimated model.

The Cox-Snell residuals (Cox and Snell, 1968) are defined as

$$\hat{e}_i = -\ln[1 - F(y_i; \hat{\tau}_i, \hat{\phi})],$$

where $F(y_i; \hat{\tau}_i, \hat{\phi})$ is the cdf (1) evaluated in $\hat{\tau}_i$ and $\hat{\phi}_i$. If the model is well adjusted, the Cox-Snell residuals follow the exponential distribution with scale parameter one (Altun, 2019).

The Dunn-Smyth residuals (Dunn and Smyth, 1996) are given by

$$\hat{r}_i = Q_{\mathcal{N}}(F(y_i; \hat{\tau}_i, \hat{\phi})),$$

where $Q_{\mathcal{N}}(\cdot)$ is the quantile function of the standard normal distribution. If the model is well estimated, the Dunn-Smith residuals follows approximately the standard normal distribution. Thus, it is expected that the plot of the Dunn-Smith residuals *versus* the index of the observations will have a behavior around zero.

4. Numerical Evaluation

To show the consistency of the MLEs of the proposed regression model, a numerical evaluation is performed. The simulation is done for the sample sizes of $n = \{40, 80, 120, 200, 300\}$ with $10,000$ repetitions for each sample size. The performance of the MLEs is evaluated through the mean estimates (AEs), the mean square errors (MSEs) and the coverage rates (CRs) for the confidence intervals.

For the sub-model of the median, the logit link function was adopted and for the sub-model of the shape parameter, the link function adopted is the logarithmic. The simulated model is given by

$$\ln\left(\frac{\tau_i}{1-\tau_i}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4}, \qquad i, \cdots, n.$$
$$\ln \phi_i = \delta_1 + \delta_2 z_{i2} + \delta_3 z_{i3}, \qquad i, \cdots, n.$$

The covariates x_i and z_i are generated from the standard uniform distribution and are kept fixed in the simulation. The true parameters chosen are: $\beta_1 = 3.3$, $\beta_2 = 0.6$, $\beta_3 = -1.2$, $\beta_4 = -2.7$, $\delta_1 = -0.9$, $\delta_2 = 1.5$ and $\delta_3 = 0.5$. The observations of the dependent variable are generated from Equation (3). The simulations were made using the matrix programming language Ox Console (Doornik, 2018) through of the `MaxBFGS` function and with analytical derivatives.

Table 1 presents the AEs, the MSEs and the CRs for the different sample sizes. Note that when the sample size increases, the AEs and the CRs converge to the true parameters and the true nominal levels, respectively. The MSEs in turn decrease when n grows. Based on these results, it can be concluded that the proposed regression model is in agreement with what is expected from the asymptotic theory.

Table 1: Simulation results

n	Par	True	AE	MSE	CR 90%	CR 95%	CR 99%
40	β_1	3.3	3.310893	0.228578	81.76	88.02	94.84
	β_2	0.6	0.606451	0.311398	82.41	88.82	95.76
	β_3	-1.2	-1.197279	0.301123	82.05	88.46	95.41
	β_4	-2.7	-2.729951	0.276904	82.30	88.62	95.42
	δ_1	-0.9	-0.897488	0.161930	87.05	92.66	97.98
	δ_2	1.5	1.711295	0.344758	83.82	90.42	96.96
	δ_3	0.5	0.502851	0.275793	84.67	90.99	97.14
80	β_1	3.3	3.312084	0.120692	86.24	92.40	97.91
	β_2	0.6	0.606698	0.139369	86.22	92.06	97.41
	β_3	-1.2	-1.209529	0.129315	86.03	91.95	97.58
	β_4	-2.7	-2.720909	0.156781	86.73	92.52	97.85
	δ_1	-0.9	-0.915689	0.069287	88.38	93.77	98.28
	δ_2	1.5	1.593236	0.110455	86.69	92.32	97.95
	δ_3	0.5	0.549582	0.122223	86.93	92.69	97.91
120	β_1	3.3	3.316369	0.084779	87.09	92.90	97.94
	β_2	0.6	0.597643	0.082928	87.21	92.70	98.09
	β_3	-1.2	-1.210115	0.106329	87.52	92.87	98.31
	β_4	-2.7	-2.718772	0.091570	87.36	93.04	98.04
	δ_1	-0.9	-0.908942	0.027404	88.44	94.18	98.67
	δ_2	1.5	1.561059	0.064437	87.52	93.46	98.40
	δ_3	0.5	0.537193	0.059755	88.22	93.49	98.43
200	β_1	3.3	3.300901	0.034352	88.59	94.11	98.54
	β_2	0.6	0.604019	0.045987	89.55	94.59	98.71
	β_3	-1.2	-1.202217	0.044271	88.50	93.85	98.60
	β_4	-2.7	-2.704659	0.045837	88.35	93.74	98.45
	δ_1	-0.9	-0.902211	0.017960	90.61	95.37	99.09
	δ_2	1.5	1.536930	0.033471	88.86	94.35	98.85
	δ_3	0.5	0.510345	0.029524	89.19	94.70	99.06
300	β_1	3.3	3.301142	0.027814	88.86	94.25	98.77
	β_2	0.6	0.601567	0.031930	89.08	94.64	98.83
	β_3	-1.2	-1.202195	0.030273	89.58	94.53	98.79
	β_4	-2.7	-2.702235	0.032188	88.79	94.42	98.75
	δ_1	-0.9	-0.900099	0.011127	90.25	95.06	98.99
	δ_2	1.5	1.519624	0.020977	89.36	94.65	98.90
	δ_3	0.5	0.507744	0.019613	89.83	95.08	99.04

5. Applications

To show in practice the use of the proposed regression model two applications to real data is considered. The UW regression model with variable shape parameter will be compared with the UW regression model with fixed shape parameter. The LR test is used to discriminate between the two models.

The Kumaraswamy regression model (Mitnik and Baek, 2013) is also introduced as a competitive model. This model has a density function given by

$$f_{\text{K}}(y;\tau,\phi) = \frac{\phi \ln(0.5)}{\ln(1-\tau^{\phi})} y^{\phi-1}(1-y^{\phi})^{\frac{\ln(0.5)}{\ln(1-\tau^{\phi})}-1},$$

where $0 < \tau < 1$ denotes the median and $\phi > 0$ is a shape parameter, that can be interpreted as precision parameter. This model is estimated considering fixed and variable shape parameter.

The comparison between the models is done by means of the following measures: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan–Quinn Information

Criterion (HQIC) statistics can be used.

The AIC, BIC and HQIC measures are given by

$$\begin{aligned}\text{AIC} &= -2\mathcal{L}(\hat{\boldsymbol{\eta}}) + 2p,\\ \text{BIC} &= -2\mathcal{L}(\hat{\boldsymbol{\eta}}) + p\ln(n)\\ \text{HQIC} &= -2\mathcal{L}(\hat{\boldsymbol{\eta}}) + 2p\ln(\ln(n)),\end{aligned}$$

where $\mathcal{L}(\hat{\boldsymbol{\eta}})$ is the log-likelihood evaluated in the maximum likelihood estimates $\hat{\boldsymbol{\eta}}$, p is the number of parameters in the model and n is the number of observations. The model that presents the smallest for the AIC, BIC and HQIC statistics is considered the best model.

In addition to these adequacy measures, the Vuong test (Vuong, 1989) is performed to compare the UW and Kumaraswamy regression models with variable shape parameter. For the rival densities $f_1(x_i;\boldsymbol{\eta})$ and $f_2(x_i;\boldsymbol{\nu})$, the Vuong statistic (Vuong, 1989) is given by

$$V_s = \frac{1}{\hat{w}^2\sqrt{n}}\sum_{i=1}^{n}\ln\left(\frac{f_1(x_i;\hat{\boldsymbol{\eta}})}{f_2(x_i;\hat{\boldsymbol{\nu}})}\right),$$

where
$$\hat{w}^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\ln\left(\frac{f_1(x_i;\hat{\boldsymbol{\eta}})}{f_2(x_i;\hat{\boldsymbol{\nu}})}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\ln\left(\frac{f_1(x_i;\hat{\boldsymbol{\eta}})}{f_2(x_i;\hat{\boldsymbol{\nu}})}\right)\right]^2,$$

and $\hat{\boldsymbol{\eta}}$ and $\hat{\boldsymbol{\nu}}$ are the MLEs of the $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$, respectively. The null hypothesis $\mathcal{H}_0$ of the Vuong test is that the pdfs $f_1(x_i;\hat{\boldsymbol{\eta}})$ and $f_2(x_i;\hat{\boldsymbol{\nu}})$ are equivalent. According to Vuong (1989) (Theorem 5), when $n \to \infty$, it follows that

i. Under $\mathcal{H}_0$: $V_s \xrightarrow{D} \mathcal{N}(0,1)$,

ii. Under $\mathcal{H}_{f_1}$: $V_s \to +\infty$,

iii. Under $\mathcal{H}_{f_2}$: $V_s \to -\infty$.

Then, from the standard normal distribution, a value c must be chosen for some level of significance. If V_s is greater than c, then $\mathcal{H}_0$ is rejected in favor of $f_1(x_i;\hat{\boldsymbol{\eta}})$, and $f_1(x_i;\hat{\boldsymbol{\eta}})$ is better that $f_2(x_i;\hat{\boldsymbol{\nu}})$. On the other hand, if $V_s < -c$, then the null hypothesis of equivalence of the models is rejected in favor of $f_2(x_i;\hat{\boldsymbol{\nu}})$, and $f_2(x_i;\hat{\boldsymbol{\nu}})$ is better than $f_1(x_i;\hat{\boldsymbol{\eta}})$. Now if $|V_s| < c$, then $\mathcal{H}_0$ is not rejected, and then the models are equivalent for the data in question. For the significance level of 1%, 5% and 10%, the value of c is of 2.58, 1.96 and 1.64, respectively.

5.1. Firjan data

In the regression model, the exogenous variables of the sub-model of the shape parameter are the same for the sub-model of the median, i.e., $x_i = z_i$. The data refers to 1793 cities belonging to the Northeast Region of Brazil. It is worth remembering that this number does not correspond to the totality of cities in the Northeast Region. Due to the lack of observations in many cities, a smaller number was obtained, which is 1793 cities. The response variable (y) refers to the Health Indice Firjan for cities in 2010. This index is between 0 and 1. The closer to 1, the better the health condition. While values close to 0, it indicates that health conditions are more precarious. The exogenous variable (x_2) refers to the Gross Domestic Product (GDP) *per capita* (in R$) of these cities in the year 2010. The variables y and x_2 were collected in http://www.firjan.com.br (Firjan, 2020) and https://www.ibge.gov.br/ (IBGE, 2020), respectively.

Table 2 presents descriptive statistics for the study variables. As can be seen, y varies between 0.1535 and 0.9556, showing a moderate behavior, since the mean and median are close and the standard deviation is low. On the other hand, the variable x_2 is quite heterogeneous, ranging between 2256.42 and 160494.21, with a mean far from the median and high standard deviation. This behavior is expected, since GDP is a variable that is highly distorted between cities.

Table 2: Statistical summary of the variables for Firjan data

Description	y	x_2
Minimum	0.1535	2256.4200
1st quartile	0.4682	3843.4200
Median	0.5800	4668.4200
Mean	0.5749	6142.8922
3rd quartile	0.6855	6073.0000
Maximum	0.9556	160494.2100
Standard deviation	0.1465	6634.9781
Skewness	−0.1849	11.0091
Kurtosis	2.4427	198.2800

The model to be estimated is given by

$$\begin{aligned} \ln\left(\tfrac{\tau_i}{1-\tau_i}\right) &= \beta_1 + \beta_2 x_{i2} \\ \ln\phi_i &= \delta_1 + \delta_2 x_{i2} \end{aligned} \quad , \quad i, \cdots, 1793.$$

Note that here $q = 2$, then the model with shape parameter fixed has one parameter less. Interest is to test if $\delta_2 = 0$, through of the LR test $\mathcal{H}_0 : \delta_2 = 0$. If $\mathcal{H}_0$ is not rejected, then the model with fixed shape parameter is more suitable.

Table 3 presents the results of the estimates for the unrestricted and restricted models. Note that for all four models, all estimates are highly significant. The positive signal of β_2 indicates that the higher the *per capita* GDP of a city, this has a positive effect on the median of Health Indice Firjan. This is as expected, since cities that have higher *per capita* income levels, their inhabitants seek to pay a health plan, not being exclusively dependent on government assistance measures. In addition, another factor that can be placed, is the fact that richer people have a greater educational level, and consequently has more hygiene conditions. On the other hand, poorer cities, there are problems such as lack of basic sanitation, poor service in the health posts of the municipalities, among others.

The LR statistic that tests the UW model with variable shape parameter *versus* the UW model with fixed shape parameter is $\mathcal{T} = 21.51$, while that $\chi^2_1(0.99) = 6.63$. Thus, $\mathcal{H}_0$ is rejected at the level of significance of 1%. With this, the model with variable shape parameter is chosen.

The AIC, BIC and HQC statistics are given in Table 4. These statistics indicate that the Kumaraswamy model with fixed shape parameter is appropriate. The Vuong statistic, which tests the UW and Kumaraswamy models with variable shape parameter is of $V_s = -10.76$. Thus, the null hypothesis of equivalence of the models is rejected, at the 1% significance level, in favor of the Kumaraswamy model, indicating that this model has a better fit.

Figures 1 and 2 shows the Dunn-Smith residuals and probability-probability plots (PP-plots) of the Cox-Snell residuals for the UW and Kumaraswamy regression models, with variable shape parameter and fixed shape parameter, respectively. In all models, the Dunn-Smith residuals show a behavior around zero and the Cox-Snell residuals are close to the diagonal line. Thus, these plots show that all models are well estimated.

Table 3: Summary of the estimates for the unrestricted and restricted models for Firjan data

Par	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
	UW			Kumaraswamy		
	variable shape parameter					
β_1	0.078821	0.033316	0.017987	0.184022	0.019228	<0.000001
β_2	0.000032	0.000005	<0.000001	0.000027	0.000003	<0.000001
δ_1	0.905210	0.026525	<0.000001	1.404459	0.009553	<0.000001
δ_2	−0.000015	0.000003	0.000003	0.000003	0.000001	0.004513
	fixed shape parameter					
β_1	0.208225	0.017905	<0.000001	0.179136	0.019413	<0.000001
β_2	0.000010	0.000001	<0.000001	0.000028	0.000003	<0.000001
δ_1	0.808075	0.017714	<0.000001	1.422821	0.006792	<0.000001

Table 4: Information criteria for Firjan data

Model	AIC	BIC	HQIC
variable shape parameter			
UW	−1824.3303	−1802.3637	−1816.2199
Kumaraswamy	−1895.9770	−1874.0104	−1887.8667
fixed shape parameter			
UW	−1804.8237	−1788.3487	−1798.7409
Kumaraswamy	−1897.0569	−1880.5820	−1890.9742

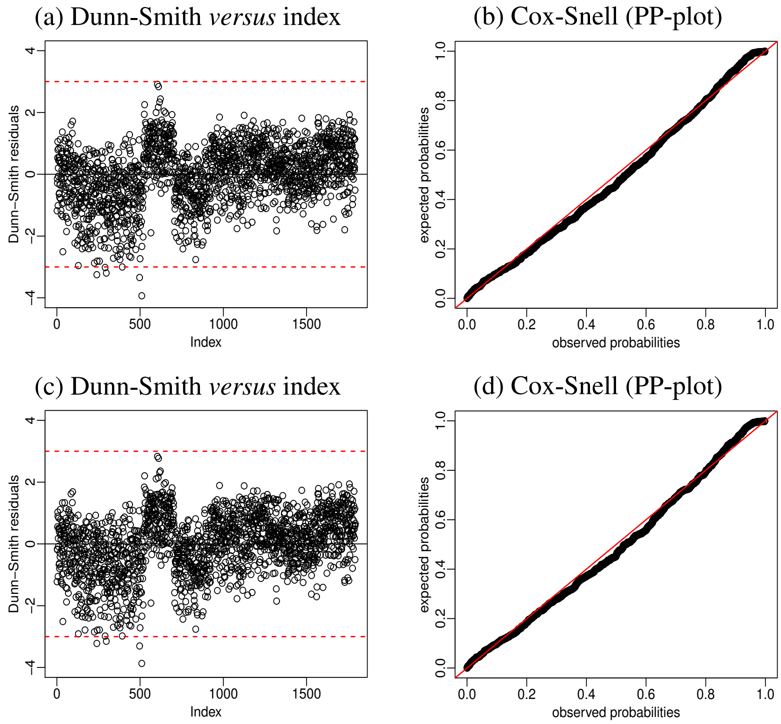

Figure 1: UW residuals (Firjan data): (a) and (b) variable shape parameter, and (c) and (d) fixed shape parameter

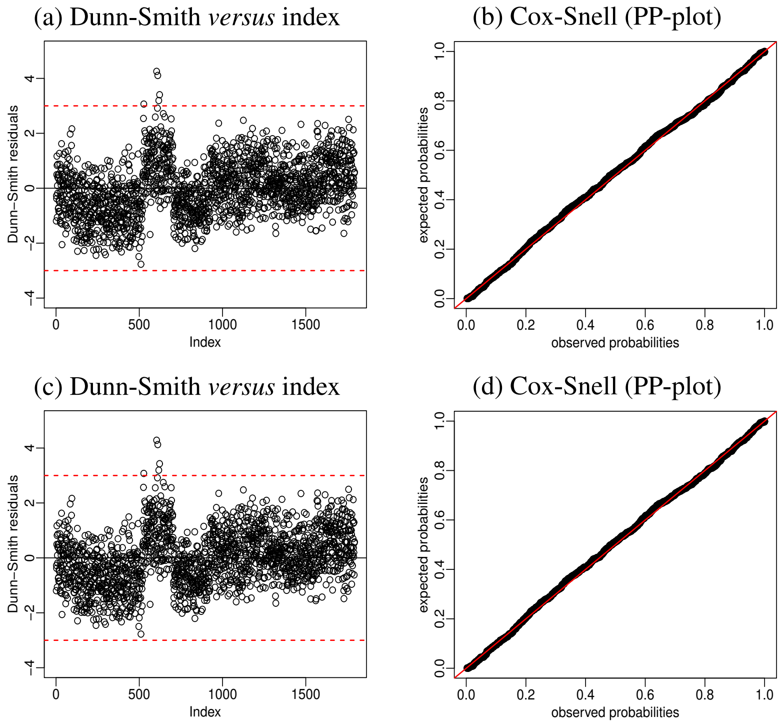

Figure 2: Kumaraswamy residuals (Firjan data): (a) and (b) variable shape parameter, and (c) and (d) fixed shape parameter

5.2. Covid-19 data

This study will be based on the deaths caused by Covid-19 in the capitals of the Brazil and in the Federal District. Thus, the number of observations is 27.

The variables are:

- y_i: Covid deaths per 100 thousand inhabitants (collected at December 6, 2021) (https://brasil.io/covid19/);
- x_{2i}: Social Vulnerability Index (SVI) is the result of the arithmetic mean of the sub-indices: Urban Infrastructure SVI, Human Capital SVI and Income and Labor SVI, each of which enters the calculation of the final SVI with the same weight. The SVI index varies between 0 and 1. A value of 0 indicates perfect social condition, and 1 indicates a worse situation (http://ivs.ipea.gov.br/).
- x_{3i}: Municipal Human Development Index (MHDI) that is a measure used to classify the degree of economic development and quality of life. This index varies between 0 and 1. The closer to 0, the lower the degree of development. Already for values close to 1, it indicates a high level of development (http://ivs.ipea.gov.br/).
- x_{4i}: Demographic density (DD), that is the ratio between total population and square kilometers (inhabitants/km^2) (https://www.ibge.gov.br/);
- x_{5i}: GDP *per capita*, at current prices (R$ 1.00) (https://www.ibge.gov.br/).

Table 5 presents the descriptive statistics of these variables. Variables y, x_2 and x_3 present a very close mean and median. These measures are further apart for variables x_4 and x_5. Variable x_5 presents the greatest dispersion.

Table 5: Statistical summary of the variables for Covid-19 data

Description	y	x_2	x_3	x_4	x_5
Minimum	0.0022	0.1780	0.7210	15.8210	20821.4600
1st quartile	0.0029	0.2655	0.7515	191.9552	23878.2400
Median	0.0034	0.2910	0.7700	2013.1941	30924.8900
Mean	0.0035	0.2953	0.7765	2813.4138	34351.8633
3rd quartile	0.0039	0.3260	0.8020	4321.7504	39824.1250
Maximum	0.0057	0.3930	0.8470	8601.2044	80502.4700
Standard deviation	0.0009	0.0499	0.0345	2758.0515	14122.0973
Skewness	0.6590	−0.1083	0.3613	0.7274	1.4909
Kurtosis	3.0999	3.0178	2.2331	2.2784	5.0500

The estimated model is given as

$$\ln\left(\frac{\tau_i}{1-\tau_i}\right) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} \qquad i, \cdots, 27.$$

$$\ln \phi_i = \delta_1 + \delta_2 x_{i4} + \delta_3 x_{i5} \qquad i, \cdots, 27.$$

Since $q = 3$, then the model with fixed shape parameter has two fewer parameters. The interest of the LR test is to test whether $\delta_2 = \delta_3 = 0$, i.e., $\mathcal{H}_0 : \delta_2 = \delta_3 = 0$. If $\mathcal{H}_0$ is rejected, then the model with variable shape parameter should be chosen.

Table 6 presents the estimates results for the models with fixed and variable shape parameter. In the median sub-model, the coefficient β_4 was not significant in any model.

Table 6: Summary of the estimates for the unrestricted and restricted models Covid-19 data.

Par	Estimate	Std. Error	p-value	Estimate	Std. Error	p-value
	UW			Kumaraswamy		
	variable shape parameter					
β_1	−0.040712	1.894400	0.982854	−0.906520	2.479361	0.714644
β_2	−2.501656	1.141156	0.028364	−2.550128	1.452680	0.079180
β_3	−6.917846	2.200549	0.001668	−5.787671	2.930949	0.048305
β_4	0.000017	0.000015	0.277393	−0.000012	0.000015	0.423226
β_5	0.000011	0.000005	0.033871	0.000015	0.000006	0.006590
δ_1	3.921062	0.726825	<0.000001	1.865013	0.020910	<0.000001
δ_2	0.000131	0.000071	0.065932	0.000029	0.000002	<0.000001
δ_3	−0.000022	0.000021	0.292205	−0.000011	0.000001	<0.000001
	fixed shape parameter					
β_1	−3.226926	0.105799	<0.000001	−3.401520	0.487423	<0.000001
β_2	0.018340	0.049023	0.708319	0.103615	0.270823	0.702022
β_3	5.753387	0.131940	<0.000001	5.944077	0.581574	<0.000001
β_4	−0.000001	0.000001	0.347515	−0.000001	0.000003	0.684374
β_5	<0.000001	<0.000001	0.121240	<0.000001	0.000001	0.972748
δ_1	5.001304	0.171510	<0.000001	4.715370	0.006205	<0.000001

The LR statistic that tests the UW model with variable shape parameter *versus* the UW model with fixed shape parameter is $\mathcal{T} = 52.67$, while that $\chi^2_2(0.99) = 9.21$. Thus, $\mathcal{H}_0$ is rejected at the level of significance of 1%. Then, the UW regression model with variable shape parameter is better.

The AIC, BIC and HQC statistics are given in Table 7. According to these statistics, the UW model with variable shape parameter has a better fit. The Vuong statistic, which tests the UW and Kumaraswamy models with variable shape parameter is of $V_s = 2.30$. Thus, the null hypothesis of equivalence of these models is rejected in favor of the UW model at the 5% significance level. This indicates that the UW model is a better fit for the data.

Table 7: Information criteria for Covid-19 data

Model	AIC	BIC	HQIC
variable shape parameter			
UW	−303.3323	−292.9656	−300.2498
Kumaraswamy	−297.2503	−286.8836	−294.1677
fixed shape parameter			
UW	−254.6576	−246.8826	−252.3457
Kumaraswamy	−200.8498	−193.0748	−198.5379

Figures 3 and 4 shows the Dunn-Smith residuals and PP-plots of the Cox-Snell residuals for the UW and Kumaraswamy regression models, with variable shape parameter and fixed shape parameter, respectively.

The Dunn-Smith residuals, except for Kumaraswamy model with fixed shape parameter, show a behavior around zero. The Cox-Snell residuals, except for Kumaraswamy model with fixed shape parameter, are close to the diagonal line. So, these graphs reveal that only the Kumaraswamy model with fixed shape parameter does not have a good fit.

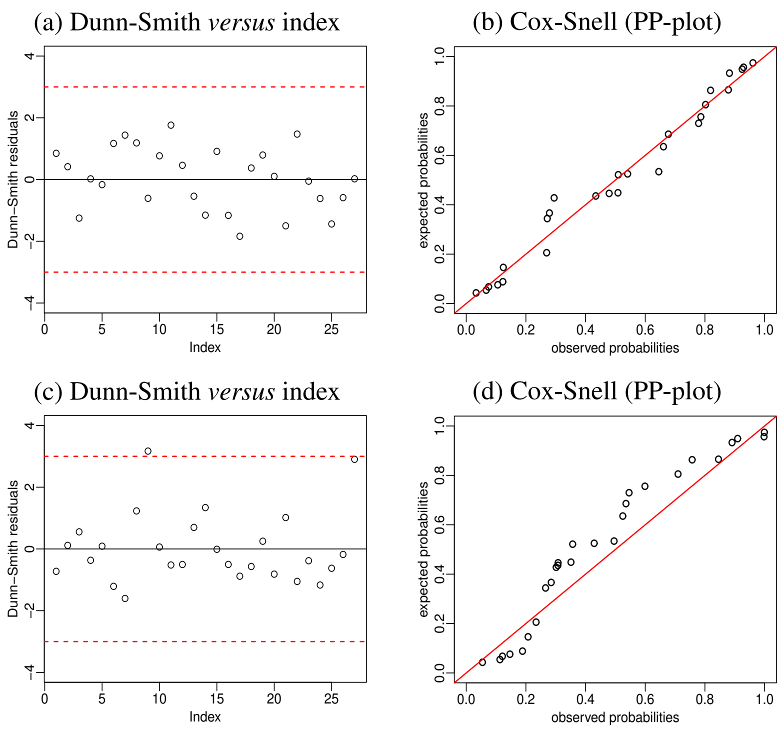

Figure 3: UW residuals (Covid-19 data): (a) and (b) variable shape parameter, and (c) and (d) fixed shape parameter

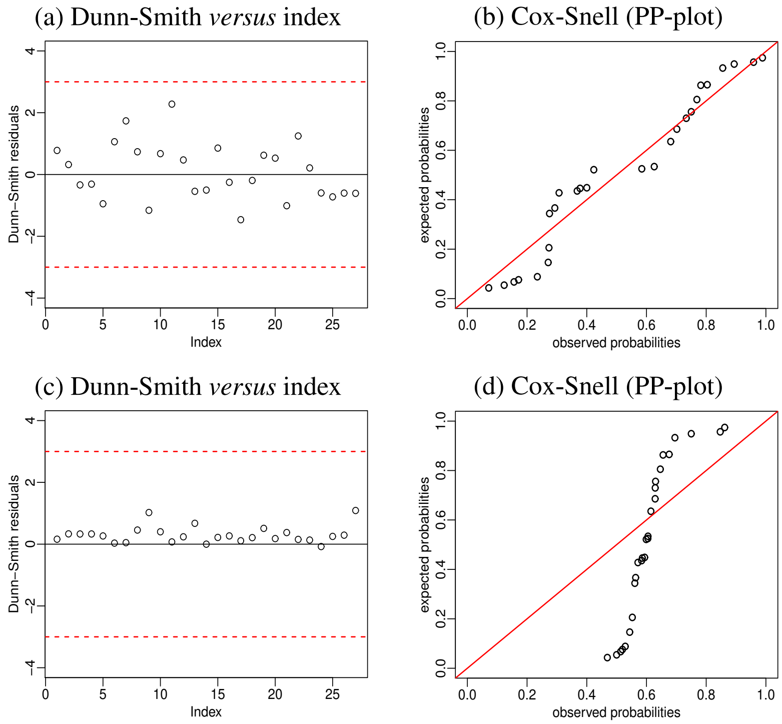

Figure 4: Kumaraswamy residuals (Covid-19 data): (a) and (b) variable shape parameter, and (c) and (d) fixed shape parameter

6. Conclusions

In this paper, an extension of the unit Weibull regression model, proposed by Mazucheli et al. (2020) is introduced. Here it is assumed that the shape parameter is related through of covariates. When no covariates is assumed, the regression model with fixed shape parameter is obtained. The estimates of the unknown parameters are performed by maximum likelihood. Analytical expressions of the score vector and the Fisher observed information matrix are demonstrated. Simulations of Monte Carlo have shown the good performance of the maximum likelihood estimators of the proposed model.

The usefulness of the model is demonstrated through of two applications to real data from Brazil. The results show that the model with variable shape parameter performs better than the model with fixed shape parameter.

Acknowledgements

I would like to thank the editor-in-chief and the anonymous reviewers for their comments and suggestions on the manuscript.

References

Altun E, Hamedani GG. The log-xgamma distribution with inference and application. Journal de la Société Française de Statistique. 2018; 159(3): 40–55.

Altun E. The log-weighted exponential regression model: alternative to the beta regression model. Commun Stat Theory. 2019; 50(10): 2306–2321.

Cox DR, Snell EJ. A general definition of residuals. J Roy Stat Soc B Met. 1968; 30(2): 248–265.

Doornik JA. Ox: an object-oriented matrix programming language. London, United Kingdown; 2018. Available from: https://www.doornik.com/.

Dunn PK, Smyth GK. Randomized quantile residuals. J Comput Graph Stat. 1996; 5(3): 236–244.

Ferrari SLP, Cribari-Neto F. Beta regression for modelling rates and proportions. J Appl Stat. 2004; 31(7): 799–815.

Firjan. Federação das Indústrias do Estado do Rio de Janeiro. Índice FIRJAN de Desenvolvimento Municipal ; 2020. Available from: http://www.firjan.com.br.

Gómez-Déniz E, Sordo MA, Calderín-Ojeda E. The Log-Lindley distribution as an alternative to the beta regression model with applications in insurance. Insur Math Econ. 2014; 54: 49–57.

Haq MA, Hashmi S, Aidi K, Ramos PL, Louzada F. Unit modified Burr-III distribution: Estimation, characterizations and validation test. Ann Data Sci. 2020; 20: 1–26.

IBGE. Instituto de Brasileiro Geografia e Estatística; 2020. Available from: https://www.ibge.gov.br/.

Korkmaz MÇ, Chesneau C. On the unit Burr-XII distribution with the quantile regression modeling and applications. Comput Appl Math. 2021; 40(1): 1–26.

Mazucheli J, Menezes AFB, Dey S. Unit-Gompertz distribution with applications. Statistica. 2019; 79(1): 25–43.

Mazucheli J, Menezes AFB, Fernandes LB, de Oliveira RP, Ghitany ME. The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates. J Appl Stat. 2020; 47(6): 954–974.

Mazucheli J, Menezes AFB, Ghitany ME. The unit-Weibull distribution and associated inference. J Appl Probab Stat. 2018; 13: 1–22.

Mitnik PA, Baek S. The Kumaraswamy distribution: median-dispersion re-parameterizations for regression modeling and simulation-based estimation. Statistical Papers. 2013; 54(1): 177–192.

Modi K, Gill V. Unit Burr-III distribution with application. Journal of Statistics and Management Systems. 2020; 23(3): 579–592.

R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2020. Available from: https://www.R-project.org/.

Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica. 1989; 57: 307–333.