mathematics

MDPI

Article

A New Regression Model on the Unit Interval: Properties, Estimation, and Application

Yury R. Benites [1,†], Vicente G. Cancho [1,†], Edwin M. M. Ortega [2,*,†], Roberto Vila [3,†] and Gauss M. Cordeiro [4,†]

[1] Department of Applied Mathematics and Statistics, University of São Paulo, São Carlos 13566-590, Brazil
[2] Department of Exact Sciences, University of São Paulo, Piracicaba 13418-900, Brazil
[3] Department of Statistics, University of Brasilia, Brasilia 70910-900, Brazil
[4] Department of Statistics, Federal University of Pernambuco, Recife 50670-901, Brazil
* Correspondence: edwin@usp.br
† These authors contributed equally to this work.

Abstract: A new and flexible distribution is introduced for modeling proportional data based on the quantile of the generalized extreme value distribution. We obtain explicit expressions for the moments, quantiles, and other structural properties. An extended regression model is constructed as an alternative to compete with the beta regression. Some simulations from the Bayesian perspectives are developed, and an illustrative application to real data involving the comparison of models and influence diagnostics is also addressed.

Keywords: Bayesian inference; generalized extreme value distribution; Johnson S_B distribution; regression model

MSC: 62J20

Citation: Benites, Y.R.; Cancho, V.G.; Ortega, E.M.M.; Vila, R.; Cordeiro, G.M. A New Regression Model on the Unit Interval: Properties, Estimation, and Application. *Mathematics* **2022**, *10*, 3198. https://doi.org/10.3390/math10173198

Academic Editor: Vasile Preda

Received: 29 July 2022
Accepted: 1 September 2022
Published: 4 September 2022

1. Introduction

In recent years, we have seen a considerable interest in formulating new families of distributions for modeling proportional data: for example, illiteracy and mortality rates, the proportion of eggs hatched in the production of cuttings, percentage of defective items, etc. One of the most common distributions in this context is the beta distribution [1], which was also extended to a regression [2]. There are several extensions of the beta regression in Simas et al. [3], Ospina & Ferrari [4], Ospina & Ferrari [5], Carrasco et al. [6], and Figueroa-Zúñiga et al. [7].

Some other alternatives have appeared in the literature. For example, Qiu et al. [8] defined a simplex regression [9], Bayes et al. [10] introduced a parametric quantile regression from the Kumaraswamy distribution [11], Lemonte and Bazán [12] proposed a regression based on an extended Johnson S_B distribution [13,14] introduced a family by compounding the cumulative distribution function (cdf) of a model with the quantile function (qf) of a second one. Cancho et al. [15] constructed a regression model by extending the Johnson S_B distribution with a shape parameter controlling the asymmetry.

Let $F_X(x)$ be the baseline cdf of a random variable (rv) X with real support $\mathbb{R}$, and the transformation:

$$Y = G\left(\frac{X-\gamma}{\delta}\right), \tag{1}$$

where $G(\cdot)$ is a cdf of an rv with support $\mathbb{R}$, $\gamma \in \mathbb{R}$, and $\delta > 0$. Then, $X = \gamma + \delta\, Q(Y)$ for $y \in (0,1)$, where $Q(y) = G^{-1}(Y)$. The probability density function (pdf) of Y follows from (1) as:

$$f_Y(y) = \delta\, f_X(\gamma + \delta Q(y))\left|\frac{dQ(y)}{dy}\right|, \tag{2}$$

where $f_X(x) = dF_X(x)/dx$.

Let $X \sim N(0,1)$ have the standard normal cdf $\Phi(x)$, and $G(y) = 1/(1+\mathrm{e}^{-y})$ the standard logistic cdf. Thus, the Johnson S_B density follows from (2), for $y \in (0,1)$,

$$f_Y(y;\gamma,\delta) = \frac{\delta\,\phi(\gamma+\delta Q(y))}{y(1-y)}, \tag{3}$$

where $\phi(\cdot)$ is the standard normal density, and:

$$Q(y) = \log\left(\frac{y}{1-y}\right), \tag{4}$$

is the logistic qf [14].

We define a new family by compounding a baseline standard normal with the qf of the generalized extreme value (GEV). Let Z be an rv having a GEV distribution [16] having cdf with zero location and unit scale parameter, namely:

$$G(z;\lambda) = \begin{cases} \exp\left[-(1+\lambda z)^{-1/\lambda}\right], & \lambda \neq 0; \\ \exp(-\mathrm{e}^{-z}), & \lambda = 0; \end{cases} \tag{5}$$

where $z \in [-1/\lambda, +\infty)$ for $\lambda > 0$, $z \in (-\infty, -1/\lambda]$ for $\lambda < 0$ and $z \in (-\infty, +\infty)$ for $\lambda = 0$.

The qf of Z becomes:

$$Q(y;\lambda) = \begin{cases} \frac{[-\log(y)]^{-\lambda}-1}{\lambda}, & \lambda > 0 \text{ and } y \in [0,1);\ \lambda < 0 \text{ and } y \in (0,1]; \\ -\log[-\log(y)], & \lambda = 0 \text{ and } y \in (0,1). \end{cases} \tag{6}$$

This article is organized in six sections. Section 2 defines a new model called the *normal-generalized extreme value* (Normal-GEV) distribution, and provides some of its structural properties. Section 3 constructs a new regression model from this distribution, and addresses Bayesian inferential procedures. Section 4 performs several simulations under different scenarios to study the behavior of the estimators. An application to colorectal cancer data in Section 5 shows the importance of the proposed regression. Some conclusions are given in Section 6.

2. The Normal-GEV Distribution

The pdf of the rv $Y \sim$ Normal-GEV(γ,δ,λ) follows by inserting (6) in Equation (3):

$$f_Y(y;\gamma,\delta,\lambda) = \frac{\delta\phi(\gamma+\delta Q(y;\lambda))}{y[-\log(y)]^{\lambda+1}}, \quad y \in (0,1). \tag{7}$$

Note that the above pdf includes both cases of Equation (6), i.e., the cases for $\lambda \neq 0$ and $\lambda \to 0$.

Proposition 1. *The limits below hold:*

$$\lim_{y\to 0^+} f_Y(y;\gamma,\delta,\lambda) = \begin{cases} 0, & \lambda < 0; \\ \infty, & \lambda \geqslant 0; \end{cases} \quad \textit{and} \quad \lim_{y\to 1^-} f_Y(y;\gamma,\delta,\lambda) = \begin{cases} \delta\phi(\gamma+\delta), & \lambda = -1; \\ 0, & \lambda \neq -1. \end{cases}$$

Proof. Setting $x = -\log(y)$, the Normal-GEV pdf (7) can be expressed as:

$$f_Y(y;\gamma,\delta,\lambda) = \frac{\delta}{\sqrt{2\pi}} \begin{cases} \frac{\exp\left\{-\frac{1}{2}\left[\gamma+\frac{\delta}{\lambda}\left(x^{-\lambda}-1\right)\right]^2\right\}}{\exp(-x)x^{\lambda+1}}, & \lambda \neq 0; \\ \frac{\exp\left\{-\frac{1}{2}\left[\gamma+\delta(-\log(x))\right]^2\right\}}{\exp(-x)x}, & \lambda = 0. \end{cases} \tag{8}$$

By using the known inequality:

$$\exp(x) \geqslant 1 + x, \quad x \in \mathbb{R}, \tag{9}$$

the expression on the right-hand side of (8) reduces to:

$$\geqslant \frac{\delta}{\sqrt{2\pi}} \begin{cases} \frac{1-\frac{1}{2}\left[\gamma+\frac{\delta}{\lambda}\left(x^{-\lambda}-1\right)\right]^2}{\exp(-x)x^{\lambda+1}}, & \lambda \neq 0; \\ \frac{1-\frac{1}{2}\left[\gamma+\delta(-\log(x))\right]^2}{\exp(-x)x}, & \lambda = 0; \end{cases} = g(x;\gamma,\delta,\lambda).$$

That is, $f_Y(y;\gamma,\delta,\lambda) \geqslant g(x;\gamma,\delta,\lambda)$. We have $\lim_{x\to\infty} g(x;\gamma,\delta,\lambda) = \infty, \forall\lambda \geqslant 0$, since the exponential grows faster than the polynomial. Then, since $y \to 0^+ \iff x \to \infty$ from the squeeze (or sandwich) theorem, we get $\infty = \lim_{x\to\infty} g(x;\gamma,\delta,\lambda) \leqslant \lim_{y\to 0^+} f_Y(y;\gamma,\delta,\lambda)$. This proves that $\lim_{y\to 0^+} f_Y(y;\gamma,\delta,\lambda) = \infty, \forall\lambda \geqslant 0$.

Again, by using (8) and inequality (9), we have:

$$f_Y(y;\gamma,\delta,\lambda) \leqslant \frac{\delta}{\sqrt{2\pi}} \begin{cases} \left(\frac{1-x}{x^{\lambda+1}}\right) \exp\left\{-\frac{1}{2}\left[\gamma+\frac{\delta}{\lambda}\left(x^{-\lambda}-1\right)\right]^2\right\}, & \lambda \neq 0; \\ \left(\frac{1-x}{x}\right) \exp\left\{-\frac{1}{2}\left[\gamma+\delta(-\log(x))\right]^2\right\}, & \lambda = 0; \end{cases} = h(x;\gamma,\delta,\lambda).$$

For $\lambda < 0$, from the inequality above and by the rapid growth of the exponential in comparison to polynomials, we have $\lim_{y\to 0^+} f_Y(y;\gamma,\delta,\lambda) \leqslant \lim_{x\to\infty} h(x;\gamma,\delta,\lambda) = 0$. Therefore, $\lim_{y\to 0^+} f_Y(y;\gamma,\delta,\lambda) = 0, \forall\lambda < 0$.

On the other hand, note that $y \to 1^- \iff x \to 0$. Again, from the inequality $f_Y(y;\gamma,\delta,\lambda) \leqslant h(x;\gamma,\delta,\lambda)$ and the rapid growth of the exponential, $\forall\lambda \neq 1$, we get $\lim_{y\to 1^-} f_Y(y;\gamma,\delta,\lambda) \leqslant \lim_{x\to 0} h(x;\gamma,\delta,\lambda) = 0$. Hence, $\lim_{y\to 1^-} f_Y(y;\gamma,\delta,\lambda) = 0$.

Finally, if $\lambda = -1$, then f_Y in (7) gives:

$$f_Y(y;\gamma,\delta,-1) = \frac{\delta\phi(\gamma+\delta[\log(y)+1])}{y}, \quad y \in (0,1).$$

Then, it is clear that (by the continuity of ϕ), $\lim_{y\to 1^-} f_Y(y;\gamma,\delta,-1) = \delta\phi(\gamma+\delta)$. □

The proof of the next result is immediate and hence omitted.

Proposition 2. *Let $Y \sim$ Normal-GEV(γ,δ,λ), and $T_{\gamma,\delta,\lambda}(y) = \gamma + \delta Q(y;\lambda)$. The cdf of Y is*

$$F_Y(y) = \Phi(T_{\gamma,\delta,\lambda}(y)), \quad y \in (0,1),$$

where $Q(y;\lambda)$ is as in (6).

2.1. Behavior of the Normal-GEV Distribution

In this subsection, some distributional properties such as unimodality and monotonicity of the Normal-GEV pdf are analyzed.

To determine the number of modes of a pdf, f, it is necessary to locate its critical points. By definition, a critical point of a function f is a point on the graph of f where the derivative is zero or infinite.

Proposition 3. *All critical points y of the new pdf (7) satisfy:*

$$[\lambda + 1 + \log(y)][-\log(y)]^\lambda - \delta[\gamma + \delta Q(y;\lambda)] = 0,$$

where $Q(y;\lambda)$ is given in (6).

Proof. Adopting the notation of Proposition 2, $f_Y(y) = f_Y(y;\gamma,\delta,\lambda)$ and $T(y) = T_{\gamma,\delta,\lambda}(y) = \gamma + \delta\, Q(y;\lambda)$, we have (dashes mean derivatives) $f_Y(y) = \phi(T(y))\, T'(y)$. Differentiating $f_Y(y)$ with respect to y gives:

$$f'_Y(y) = \phi(T(y)) \left\{ T''(y) - T(y)[T'(y)]^2 \right\}, \tag{10}$$

where:

$$T'(y) = \delta\, \frac{[-\log(y)]^{-(\lambda+1)}}{y} \quad \text{and} \quad T''(y) = T'(y)\, \frac{[-\log(y)]^{-1}[\lambda + 1 + \log(y)]}{y}. \tag{11}$$

By combining (10) and (11), we obtain:

$$f'_Y(y) = \frac{f_Y(y)[-\log(y)]^{-(\lambda+1)}}{y} \Big\{ [\lambda + 1 + \log(y)][-\log(y)]^{\lambda} - \delta[\gamma + \delta Q(y;\lambda)] \Big\}.$$

Then, the proof follows. □

Theorems 1 and 2 show that λ governs the shape of the new distribution.

Theorem 1. *If $Y \sim$ Normal-GEV(γ,δ,λ) with $\lambda \neq 0$, the pdf of Y is:*

1. *Decreasing-increasing-decreasing (DID) or decreasing (D) whenever $\lambda \in \mathbb{N}$.*
2. *Unimodal whenever $\lambda \in \mathbb{Z} \setminus (\mathbb{N} \cup \{-1, 0\})$ and $\gamma < \delta/\lambda$.*

Proof. By replacing $Q(y;\lambda)$ with $\lambda \neq 0$ in equation of Proposition 3, all critical points y of the pdf of Y satisfy:

$$p(x) = -x^{2\lambda+1} + (\lambda+1)x^{2\lambda} + \delta\left(\frac{\delta}{\lambda} - \gamma\right)x^{\lambda} - \frac{\delta^2}{\lambda} = 0, \quad \text{with } x = -\log(y).$$

If $\lambda \in \mathbb{N}$, then $p(x)$ is a polynomial of degree $2\lambda + 1$. By Descartes' rule of signs [17], $p(x)$ has two sign changes (regardless of the sign of $\delta/\lambda - \gamma$) and then two or zero positive roots. If $p(x)$ has two positive roots x_1 and x_2, the pdf of Y has two critical points $y_1 = \exp(-x_1)$ and $y_2 = \exp(-x_2)$ in $(0,1)$. On the other hand, if $p(x)$ has zero positive roots, it has no critical points in $(0,1)$. Finally, since $\lim_{y\to 0^+} f_Y(y) = \infty$ and $\lim_{y\to 1} f_Y(y) = 0$, the statement of Item 1 follows.

If $\lambda \in \mathbb{Z} \setminus (\mathbb{N} \cup \{-1, 0\})$, $p(x)$ can be written in terms of $w = x^{-1} = [-\log(y)]^{-1}$

$$q(w) = -w^{-2\lambda-1} + (\lambda+1)w^{-2\lambda} + \delta\left(\frac{\delta}{\lambda} - \gamma\right)w^{-\lambda} - \frac{\delta^2}{\lambda} = 0.$$

In this case, $q(w)$ is a polynomial of degree $-2\lambda - 1$. Again, by Descartes' rule of signs, $q(w)$ has only one sign change, and this polynomial has only one positive root, say w_0. Then, the pdf of Y has only one critical point $y_0 = \exp(-w_0^{-1})$ on $(0,1)$. Since $\lim_{y\to 0^+} f_Y(y) = 0$ and $\lim_{y\to 1} f_Y(y) = \delta\phi(\gamma + \delta)\delta_{\lambda,-1}$, where $\delta_{i,j}$ is the Kronecker delta, the unimodality stated in Item 2 follows. □

Theorem 2 provides the explicit critical points of the Normal-GEV pdf ($\lambda = 0$) whenever a constraint on parameters γ and δ is imposed. Further, this theorem shows that the form of the pdf is continuous monotone at three disjoint intervals.

Theorem 2. *If $Y \sim$ Normal-GEV(γ,δ,λ) and $\lambda = 0$, the pdf of Y is decreasing-increasing-decreasing (DID) whenever:*

$$\gamma \leqslant [2\log(\delta) - 1]\delta + \frac{1}{\delta}. \tag{12}$$

Moreover,

$$y_1 = \exp\left[\delta^2 W_{-1}\left(-\frac{\exp(\frac{\gamma}{\delta} - \frac{1}{\delta^2})}{\delta^2}\right)\right] \quad \textit{and} \quad y_2 = \exp\left[\delta^2 W_0\left(-\frac{\exp(\frac{\gamma}{\delta} - \frac{1}{\delta^2})}{\delta^2}\right)\right] \tag{13}$$

are the minimum and maximum points of the pdf of Y, respectively. For some integer k, $W_k(\cdot)$ denotes the Lambert W function.

Proof. By replacing the definition of $Q(y;\lambda)$ with $\lambda = 0$ in the equation of Proposition 3, all critical points y of the pdf of Y satisfy:

$$1 - x + \delta^2 \log(x) - \delta\gamma = 0, \quad \text{with } x = -\log(y).$$

Equivalently,

$$\left(-\frac{x}{\delta^2}\right)\exp\left(-\frac{x}{\delta^2}\right) = -\frac{\exp(\frac{\gamma}{\delta} - \frac{1}{\delta^2})}{\delta^2}.$$

Since the function $f(t) = t\exp(t)$ has a (global) minimum $-1/\mathrm{e}$ at the point $t = -1$, the above equation can be solved for $-x/\delta^2$ only if $z = -\exp(\frac{\gamma}{\delta} - \frac{1}{\delta^2})/\delta^2 \geqslant -1/\mathrm{e}$. Since $z < 0$, assuming the condition (12), the two values y_1 and y_2 in (13) are obtained. Finally, y_1 and y_2 are minimum and maximum points of the pdf of Y, respectively, because $\lim_{y\to 0^+} f_Y(y) = \infty$ and $\lim_{y\to 1} f_Y(y) = 0$, and $0 < y_1 < y_2 < 1$. □

Figures 1 and 2 reveal different types of asymmetrical. For negative values of the parameter λ we have positive asymmetry (and unimodality), for positive values of λ we have decreasing, increasing and decreasing behavior.

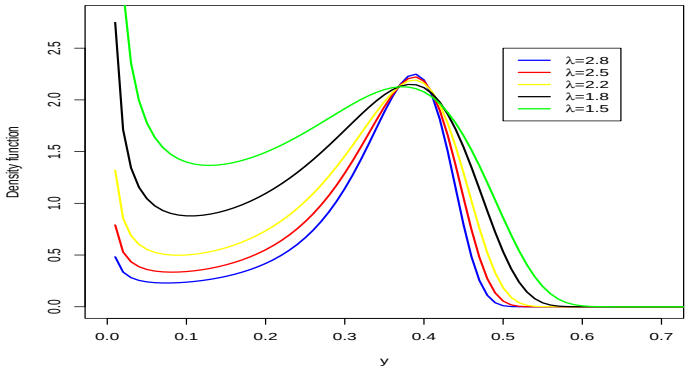

Figure 1. Normal-GEV density with $\gamma = 0.2$, $\delta = 2$, and λ varying.

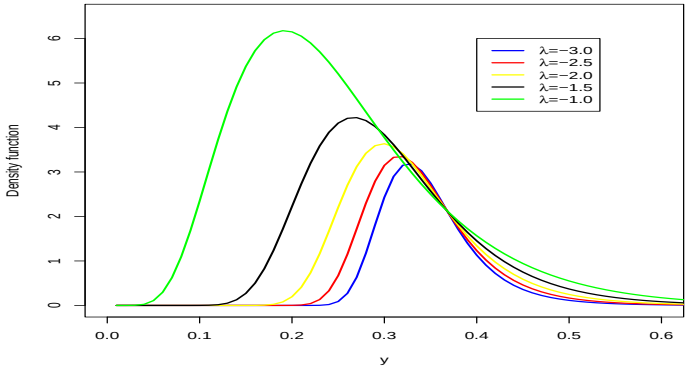

Figure 2. Normal-GEV density with $\gamma = -0.2$, $\delta = 2$, and λ varying.

Figure 3 shows that cases of decreasing strict monotonicity can appear in the Normal-GEV pdf.

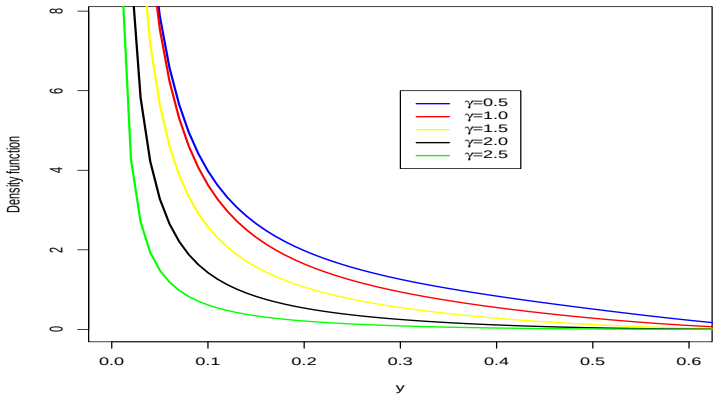

Figure 3. Normal-GEV density with $\delta = 1$, $\lambda = 1$, and γ varying.

2.2. Related Distributions

Proposition 4. *If $Y \sim$ Normal-GEV$(\gamma, \delta, \lambda)$, then (for any $a > 0$ and $b \in \mathbb{R}$):*

1. *Normal distribution:* $a\delta Q(Y;\lambda) + a\gamma + b \sim N(b, a^2)$.
2. *Log-normal distribution:* $\exp[a\delta Q(Y;\lambda) + a\gamma + b] \sim \log[N(b, a^2)]$.
3. *Folded normal distribution:* $|a\delta Q(Y;\lambda) + a\gamma + b| \sim N_f(b, a^2)$.
4. *χ distribution with one degree of freedom (df):* $|a\delta Q(Y;\lambda) + a\gamma|/a \sim \chi_1$.
5. *Noncentral χ^2 distribution:* $[a\delta\, Q(Y;\lambda) + a\gamma + b]^2/a^2 \sim \chi_1^2(b^2/a^2)$.
6. *Lévy distribution:* $[a\delta Q(Y;\lambda) + a\gamma]^{-2} \sim \text{Levy}(0, a^{-2})$.

Proof. The proof of this proposition follows by using well-known properties of the normal distribution with Equation (1). Hence, details are omitted. □

Proposition 5.

1. *χ^2 distribution with n df: If $Y_k \sim$ Normal-GEV$(\gamma_k, \delta_k, \lambda_k)$, $k = 1, \ldots, n$, are independent rvs, then:*

$$[\delta_1 Q(Y_1;\lambda_1) + \gamma_1]^2 + \cdots + [\delta_n Q(Y_n;\lambda_n) + \gamma_n]^2 \sim \chi_n^2.$$

2. *Student t-distribution with $n-1$ df: If $Y_k \sim$ Normal-GEV$(\gamma_k, \delta_k, \lambda_k)$, $k = 1, \ldots, n$, are independent rvs, then:*

$$\frac{\overline{X}}{\sqrt{\frac{1}{n(n-1)}[(\delta_1 Q(Y_1;\lambda_1) + \gamma_1 - \overline{X})^2 + \cdots + (\delta_n Q(Y_n;\lambda_n) + \gamma_n - \overline{X})^2]}} \sim t_{n-1},$$

where $\overline{X} = \{[\delta_1 Q(Y_1;\lambda_1) + \cdots + \delta_n Q(Y_n;\lambda_n)] + (\gamma_1 + \cdots + \gamma_n)\}/n$.

3. *F-distribution with (n, m) df: If $Y_k \sim$ Normal-GEV$(\gamma_k, \delta_k, \lambda_k)$, $k = 1, \ldots, n$, $Y'_j \sim$ Normal-GEV$(\gamma'_j, \delta'_j, \lambda'_j)$, $j = 1, \ldots, m$, are independent rvs, then:*

$$\frac{\{[\delta_1 Q(Y_1;\lambda_1) + \gamma_1]^2 + \cdots + [\delta_n Q(Y_n;\lambda_n) + \gamma_n]^2\}/n}{\{[\delta'_1 Q(Y'_1;\lambda'_1) + \gamma'_1]^2 + \cdots + [\delta'_m Q(Y'_m;\lambda'_m) + \gamma'_m]^2\}/m} \sim F_{n,m}.$$

Proof. The proof follows by combining known properties of the normal distribution with Equation (1). □

2.3. The New Model as a Limit Distribution

Let $\overline{X}$ be the sample mean of n samples drawn from a population with mean μ and standard deviation $\sigma \in (0, \infty)$. The central limit theorem leads to the well-known result:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{\mathscr{D}} Z \sim N(0,1),$$

where "$\xrightarrow{\mathscr{D}}$" denotes convergence in the distribution. Let $T_{\gamma,\delta,\lambda}$ be the transformation defined in Proposition 2. Since $T^{-1}_{\gamma,\delta,\lambda}$ is a continuous map, by applying the continuous mapping theorem, we have:

$$T^{-1}_{\gamma,\delta,\lambda}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right) \xrightarrow{\mathscr{D}} Y := T^{-1}_{\gamma,\delta,\lambda}(Z).$$

From Equation (1), $Y = T^{-1}_{\gamma,\delta,\lambda}(Z) \sim$ Normal-GEV$(\gamma, \delta, \lambda)$, where $T^{-1}_{\gamma,\delta,\lambda}(z) = G((z - \gamma)/\delta; \lambda)$ and $G(\cdot)$ is the GEV cdf given in (5).

Then, for $\lambda \neq 0$:

$$\exp\left[-\left(1 + \frac{\lambda}{\delta}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} - \gamma\right)\right)^{-1/\lambda}\right] \xrightarrow{\mathscr{D}} Y \sim \text{Normal-GEV}(\gamma, \delta, \lambda);$$

and, for $\lambda = 0$,

$$\exp\left[-\exp\left(-\frac{1}{\delta}\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} - \gamma\right)\right)\right] \xrightarrow{\mathscr{D}} Y \sim \text{Normal-GEV}(\gamma, \delta, 0).$$

2.4. Moments, Quantile and Other Measures

Proposition 6. *The rth real moment of* $Y \sim$ *Normal-GEV*$(\gamma, \delta, \lambda)$ *is (whenever it makes sense):*

$$E[Y^r] = M_{\varphi_\lambda(Z)}(r), \quad Z \sim N(0,1), \tag{14}$$

where,

$$\varphi_\lambda(z) = \begin{cases} -(1 - \frac{\lambda\gamma}{\delta} + \frac{\lambda}{\delta} z)^{-1/\lambda}, & \text{if } \lambda \neq 0; \\ -\exp(\frac{\gamma - z}{\delta}), & \text{if } \lambda = 0, \end{cases}$$

and $M_X(\cdot)$ *is the generating function of X. For example, for* $\lambda = -1$, $E[Y^r] = \exp(\frac{\lambda\gamma}{\delta} + \frac{r^2\lambda^2}{2\delta^2} - 1)$.

Proof. The proof is immediate since Y follows Equation (1) with $G(\cdot)$ given in (5). □

The moments of Y are finite since its support is limited, and the integral in (14) can be numerically computed via the software R, Mathematica, and Maple, among others. Specifically, the mean and variance are calculated using the integrate function of the R software. This approximation is based on the adaptive quadrature of functions of one variable over a finite interval; for more details, see Piessens et al. [18].

Table 1 reports the mean and variance of Y for some parameters obtained numerically using the *integrate* function of the R software. We note in the second and third columns that the mean and variance do not change for different values of λ. However, there is some variation of the variance for different values of γ with the other parameters fixed. So, the parameter γ is responsible for the location of the model and the parameter δ for the dispersion.

Table 1. Mean and variance of the Normal-GEV distribution.

Parameter	Mean	Variance	Parameter	Mean	Variance	Parameter	Mean	Variance
$\lambda = 0.5$ $\gamma = 0.2$ $\delta = 2$	0.3212	0.0279	$\lambda = 0.1$ $\gamma = 2$ $\delta = 2$	0.0921	0.0100	$\lambda = 0.5$ $\gamma = 0.2$ $\delta = 1$	0.3047	0.0622
$\lambda = 0.1$ $\gamma = 0.2$ $\delta = 2$	0.3341	0.0266	$\lambda = 0.1$ $\gamma = 1$ $\delta = 2$	0.2105	0.0213	$\lambda = 0.5$ $\gamma = 0.2$ $\delta = 2$	0.3212	0.0279
$\lambda = 0$ $\gamma = 0.2$ $\delta = 2$	0.3373	0.0265	$\lambda = 0.1$ $\gamma = 0.5$ $\delta = 2$	0.2856	0.0254	$\lambda = 0.5$ $\gamma = 0.2$ $\delta = 3$	0.3359	0.0144
$\lambda = -0.2$ $\gamma = 0.2$ $\delta = 2$	0.3437	0.0267	$\lambda = 0.1$ $\gamma = 0.2$ $\delta = 2$	0.3341	0.0266	$\lambda = 0.5$ $\gamma = 0.2$ $\delta = 6$	0.3532	0.0038
$\lambda = -0.5$ $\gamma = 0.2$ $\delta = 2$	0.3538	0.0281	$\lambda = 0.1$ $\gamma = 0$ $\delta = 2$	0.3666	0.0269	$\lambda = 0.5$ $\gamma = 0.2$ $\delta = 8$	0.3573	0.0021

Proposition 7. *Let $Y \sim$ Normal-GEV$(\gamma, \delta, \lambda)$. Then, the qth quantile of Y is:*

$$y_q = \begin{cases} \exp\left\{-[1 + \frac{\lambda}{\delta}(x_q - \gamma)]^{-\frac{1}{\lambda}}\right\}, & \lambda \neq 0; \\ \exp\left\{-\exp\left[-(\frac{x_q - \gamma}{\delta})\right]\right\}, & \lambda = 0, \end{cases}$$

where x_q is the qth standard normal quantile (for $0 < q < 1$).

Proof. The proof is immediate and then omitted. □

The median ν of the Normal-GEV distribution follows from Proposition 7:

$$\nu = \begin{cases} \exp\left\{-\left(1 + \frac{\lambda}{\delta}(-\gamma)\right)^{-\frac{1}{\lambda}}\right\}, & \lambda \neq 0; \\ \exp\left\{-\exp\left(\frac{-\gamma}{\delta}\right)\right\}, & \lambda = 0; \end{cases} \tag{15}$$

where $x_{0.5}$ is the median of the standard normal distribution. Furthermore, from Proposition 7, the random values for Y can be easily generated.

Further, the Bowley's skewness of Y is:

$$B = \frac{y_{0.75} + y_{0.25} - 2\nu}{y_{0.75} - y_{0.25}},$$

where $y_{0.25}, \nu$ and $y_{0.75}$ are the quantile values.

Figure 4 displays the skewness of Y for some parameters, which indicates that the distribution is symmetric when $\lambda \longrightarrow 0$ and $\gamma \longrightarrow 0$. Thus, the parameters λ and γ govern the skewness of Y.

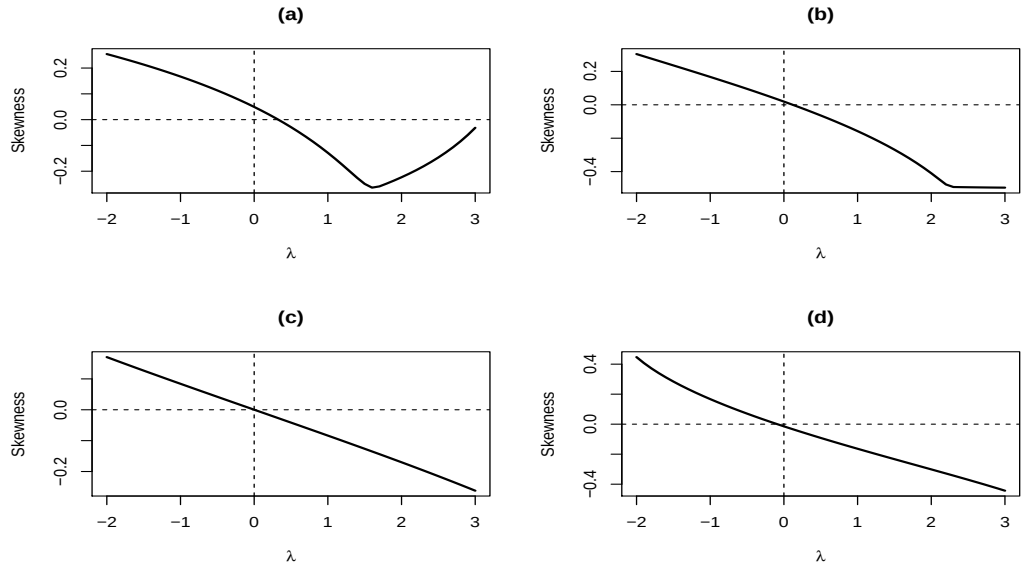

Figure 4. The skewness of Y. (**a**) $\delta = 2$, $\gamma = 0.5$, (**b**) $\delta = 2$, $\gamma = 0.2$, (**c**) $\delta = 4$, $\gamma = 0$, (**d**) $\delta = 2$, $\gamma = -0.2$.

3. Bayesian Inference for the Normal-GEV Regression

The Normal-GEV median can be expressed from (15) as:

$$\gamma = -\delta\, Q(\nu;\lambda),$$

where $Q(y;\lambda)$ is given by (6). Therefore, it has a simple form to construct a regression model. In this context, we obtain a reparameterized density of Y by replacing the above expression in Equation (2),

$$f(y;\lambda,\nu,\delta) = \frac{\delta\,\phi(\delta\,[Q(y;\lambda) - Q(\nu;\lambda)])}{y[-\log(y)]^{\lambda+1}}, \quad y \in (0,1), \tag{16}$$

where $\lambda \in \mathbb{R}$, $\nu \in (0,1)$, and $\delta > 0$ works as a dispersion parameter.

Let $\boldsymbol{y} = (y_1,\ldots,y_n)^\top$ be n observations from $Y_i \sim$ Normal-GEV(λ,ν_i,δ_i), where two systematic components are constructed for the median ν_i and dispersion δ_i. The Normal-GEV regression model is defined by (16) and the systematic components:

$$\eta_{1i} = h_1(\nu_i) = \mathbf{w}_i^\top \boldsymbol{\beta} \quad \text{and} \quad \eta_{2i} = h_2(\delta_i) = \mathbf{z}_i^\top \boldsymbol{\tau},$$

where $\mathbf{w}_i = (w_{1i},\ldots,w_{pi})^\top$ and $\mathbf{z}_i = (z_{1i},\ldots,z_{qi})^\top$ are vectors of covariates, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\boldsymbol{\tau} \in \mathbb{R}^q$ are vectors of unknown coefficients ($p+q < n$), and $h_1 : (0,1) \longrightarrow \mathbb{R}$ and $h_2 : (0,\infty) \longrightarrow \mathbb{R}$ are strictly monotonic and twice differentiable link functions. There are several possible choice for the link functions h_1 and h_2. For example, some useful link functions for the median are: logit $h_1(\nu) = \log\left(\frac{\nu}{1-\nu}\right)$; probit $h_1(\nu) = \Phi^{-1}(\nu)$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function; complementary log–log$h_1(\nu) = \log[-\log(1-\nu)]$; log–log $h_1(\nu) = -\log[-\log(\nu)]$; and Cauchy $h_1(\psi) = \tan[\pi(\psi - 0.5)]$. Some possible choice dispersion link are: logarithmic $h_2(\delta) = \log(\delta)$; square root $h_2(\delta) = \sqrt{\delta}$; identity $h_2(\delta) = \delta$ (with $\delta > 0$); among others. The relationship between ν and $\boldsymbol{\beta}$ and δ_i and $\boldsymbol{\tau}$ is equivalent to a canonical link for ν_i (location parameter) and δ_i (dispersion parameter) in setting generalized linear model.

Further, let $\boldsymbol{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)^\top$ and $\boldsymbol{Z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n)^\top$ be matrices of full ranks p and q, respectively. The likelihood function for the parameters given the observed data $\mathcal{D} = (\boldsymbol{y}, \boldsymbol{W}, \boldsymbol{Z})$ has the form:

$$L(\lambda, \boldsymbol{\beta}, \boldsymbol{\tau}|\mathcal{D}) = \prod_{i=1}^{n} \delta_i \, \phi(\delta_i[Q(y_i;\lambda) - Q(\nu_i;\lambda)]) \left[\prod_{i=1}^{n} y_i\right]^{-1} \left[\prod_{i=1}^{n} (-\log(y_i))^{\lambda+1}\right]^{-1}. \quad (17)$$

Maximizing (17) provides the maximum likelihood estimates (MLEs) of the parameters. However, we consider the Bayesian method with the common proper prior distributions:

$$\beta_j \sim \mathrm{Nl}(0, 100), j = 1, \ldots, p, \quad \tau_j \sim \mathrm{N}(0, 100), j = 1, \ldots q, \quad \text{and} \quad \lambda \sim \mathrm{N}(0, 1), \quad (18)$$

where $\boldsymbol{\beta}$, $\boldsymbol{\tau}$, and λ are assumed independent. Combining (17) and (18), the joint posterior density for $\boldsymbol{\vartheta} = (\lambda, \boldsymbol{\beta}, \boldsymbol{\tau}) \in \mathbb{R}^{p+q+1}$ reduces to:

$$\pi(\boldsymbol{\vartheta}|\mathcal{D}) \propto \prod_{i=1}^{n} \delta_i \, \phi(\delta_i[Q(y_i;\lambda) - Q(\nu_i;\lambda)]) \left[\prod_{i=1}^{n} (-\log(y_i))^{\lambda+1}\right]^{-1} \pi(\lambda) \, \pi(\boldsymbol{\beta}) \, \pi(\boldsymbol{\tau}).$$

The Metropolis–Hastings algorithm consists of the steps:

(1) Initialize from trial $\boldsymbol{\vartheta}_{(0)}$ and set $j = 0$;
(2) Construct the transitional kernel $K(\boldsymbol{\vartheta}', \boldsymbol{\vartheta}_j) = N_{p+q+1}(\boldsymbol{\vartheta}_j, \tilde{\Sigma})$ to generate a new point $\boldsymbol{\vartheta}'$, where $\tilde{\Sigma}$ is evaluated at $\boldsymbol{\vartheta}_j$;
(3) Update $\boldsymbol{\vartheta}_{(j)}$ to $\boldsymbol{\vartheta}_{(j+1)} = \boldsymbol{\vartheta}'$ with probability $p_j = \min\{1, \pi(\boldsymbol{\vartheta}'|\boldsymbol{\mathcal{D}})/\pi(\boldsymbol{\vartheta}_{(j)}|\boldsymbol{\mathcal{D}})\}$, or set $\boldsymbol{\vartheta}_{(j)}$ with probability $1 - p_j$;
(4) Steps (2) and (3) are repeated until the process becomes stationary.

The script can be obtained from the authors upon request. For more details on the Metropolis–Hastings algorithm, we refer to Chib et al. [19].

4. Simulation Study

We determine the accuracy of the Bayesian estimates in the new regression model. One thousand samples of sizes $n = 50, 100, 200$, and 400 are generated from $y_i \sim Normal - VEG(\lambda, \nu_i, \delta_i)$ under the systematic components $h_i(\nu_i) = \log(\frac{\nu_i}{1-\nu_i}) = \beta_0 + \beta_1 w_i$ and $h_2(\delta_i) = \log \delta_i = \tau_0 + \tau_1 w_i$, and $\lambda = -0.4, 0.4$. The covariate w_i is produced from the uniform $U(0, 1)$ distribution with $\beta_0 = -3$, $\beta_1 = -2$, $\tau_0 = 1$, and $\tau_1 = 1$.

We obtain the posterior summaries and 95% highest probability density (HPD) intervals of the parameters for each trial. We generate 25,000 MCMC posterior samples for the parameters, from which 5000 observations are discarded to eliminate the effect of the initial values. To avoid correlation between the generates values, we took a spacing of size 5, leading to samples of size 2000. Therefore, the final sample has size 2000 to record the convergence of the Gibbs samples [20]. For each configuration, we perform 1000 replicates to determine from the estimates: the average (MC mean), standard deviation (SD), mean root square error (MC RMSE), and coverage probability (CP).

Table 2 reports the simulations results, which reveal that the RMSEs decay when n increases (as expected), and the coverage probabilities approximate the nominal level.

Table 2. Simulation results of the Normal-GEV regression model from 1000 trials.

		$\lambda = -0.4$					$\lambda = 0.4$				
n	**Parameter**	**MC Mean**	**SD**	**Bias**	**MC RMSE**	**CP**	**MC Mean**	**SD**	**Bias**	**MC RMSE**	**CP**
50	λ	−0.368	0.820	0.032	0.821	0.983	0.345	0.292	−0.055	0.297	0.961
	β_0	−3.008	0.190	−0.008	0.190	0.935	−3.140	0.545	−0.140	0.562	0.938
	β_1	−1.990	0.272	0.010	0.272	0.945	−1.899	0.826	0.101	0.832	0.947
	τ_0	1.032	0.953	0.032	0.953	0.980	0.942	0.417	−0.059	0.421	0.959
	τ_1	1.030	0.558	0.030	0.558	0.965	0.966	0.408	−0.034	0.409	0.952
100	λ	−0.36	0.575	0.040	0.576	0.969	0.370	0.206	−0.031	0.208	0.954
	β_0	−3.00	0.131	0.004	0.131	0.940	−3.057	0.366	−0.057	0.370	0.943
	β_1	−2.00	0.194	−0.005	0.194	0.936	−1.969	0.583	0.031	0.583	0.939
	τ_0	1.05	0.672	0.049	0.673	0.953	0.969	0.292	−0.031	0.294	0.945
	τ_1	1.01	0.391	0.015	0.391	0.955	0.981	0.283	−0.019	0.283	0.932
200	λ	−0.355	0.388	0.045	0.391	0.963	0.383	0.133	−0.017	0.134	0.953
	β_0	−3.002	0.090	−0.003	0.090	0.944	−3.031	0.255	−0.031	0.256	0.944
	β_1	−1.996	0.131	0.004	0.131	0.941	−1.986	0.390	0.014	0.390	0.949
	τ_0	1.047	0.450	0.047	0.452	0.966	0.980	0.187	−0.020	0.188	0.965
	τ_1	1.022	0.263	0.022	0.264	0.960	0.987	0.192	−0.013	0.192	0.950
400	λ	−0.394	0.289	0.006	0.289	0.939	0.388	0.101	−0.012	0.101	0.941
	β_0	−3.004	0.067	−0.004	0.067	0.930	−3.022	0.177	−0.023	0.178	0.949
	β_1	−1.995	0.095	0.005	0.095	0.943	−1.983	0.271	0.017	0.271	0.955
	τ_0	1.008	0.337	0.008	0.337	0.938	0.981	0.142	−0.019	0.143	0.939
	τ_1	1.000	0.189	0.000	0.189	0.945	1.003	0.129	0.003	0.129	0.952

5. Application: Colorectal Cancer Data

We analyze data on patients with colorectal cancer [21] from 50 American States, where the mortality rate is the response variable. We consider $n = 220$ observations after deleting states with incomplete data. The variables below were collected:

- y_i: mortality rate ($i = 1, \ldots, 220$);
- x_{1i}: sex (0 = man, 1 = woman);
- x_{2i}: race (non-Hispanic white, non-Hispanic black, Hispanic).

Figure 5 displays boxplots of mortality rate by sex (left panel) and race (right panel). They indicate that it is different for men and women, and Hispanic patients have a lesser mortality rate than the other patients.

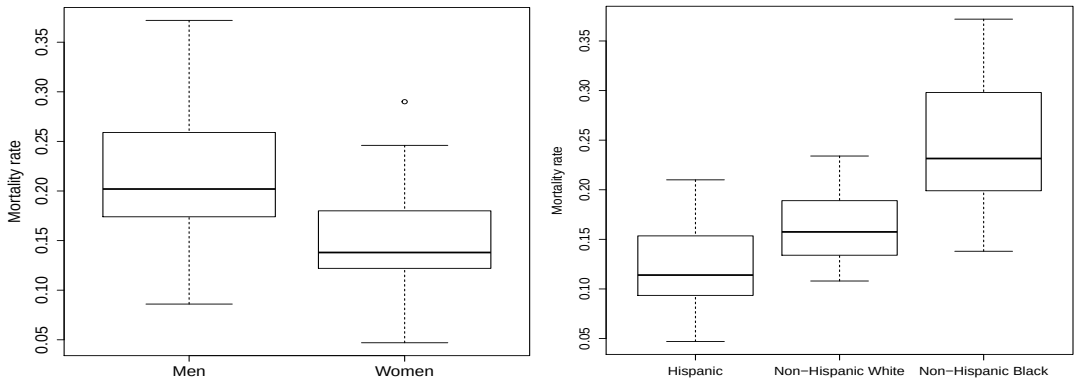

Figure 5. Boxplots of the mortality rates by sex (**left** panel) and race (**right** panel).

We fit the regression mode described in Section 3 to these data with all covariates on the median of the mortality rate (ν), and dispersion parameter (δ) with the link functions: logistic, probit, complement log–log for the median, and logarithmic for the dispersion, i.e.,

$$h_1(\nu_i) = \beta_0 + \beta_1 x_{1i} + \beta_{2_1} x_{2_1 i} + \beta_{2_2} x_{2_2 i}, \quad \text{and}$$
$$h_2(\nu_i) = \log(\delta_i) = \tau_0 + \tau_1 x_{1i} + \tau_{2_1} x_{2_1 i} + \tau_{2_2} x_{2_2 i},$$

where the race covariate (x_2) requires two dummy variables:

$$x_{2_1 i} = \begin{cases} 1, & \text{if non-hispanic white;} \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad x_{2_2 i} = \begin{cases} 1, & \text{if non-hispanic black;} \\ 0, & \text{otherwise.} \end{cases}$$

We consider 250,000 MCMC posterior samples from which 50,000 were excluded to eliminate the effect of the initial values. The autocorrelations of theses sampled values are reduced by taking a spacing of size five, yielding a final sample of size 4000. The trace plots for parameters of the new regression model with complementary log–log link are reported in Figure 6, thus indicating convergence of the chains [20].

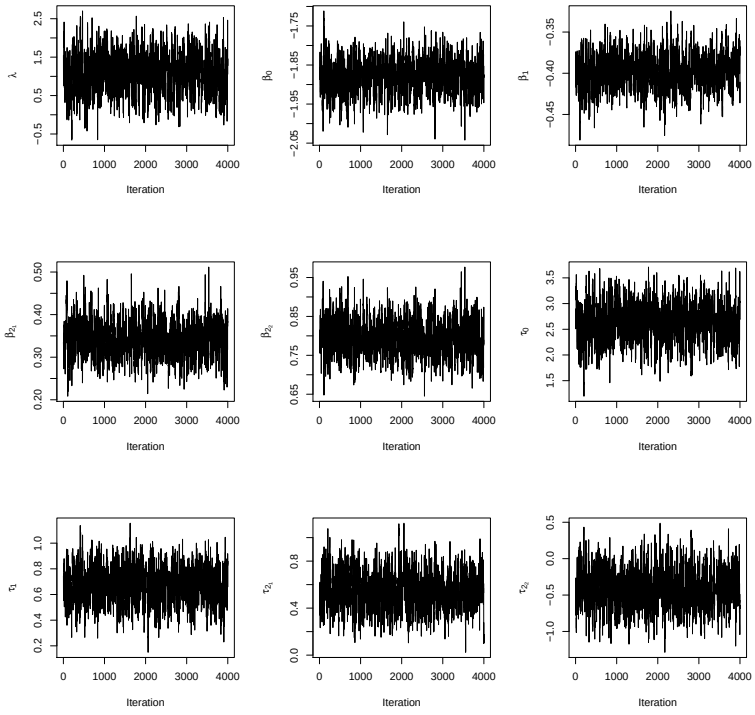

Figure 6. Trace plots for the parameters of the GEV-CLL regression model for colon rectal data.

For model comparison, we consider the deviance Information criterion (DIC [22]), the expected Akaike information criterion (EAIC, [23]), the expected Bayesian (or Schwarz) information criterion (EBIC, [24]), and the log pseudo marginal likelihood (LPML [25]). The last criteria is the one derived from the Conditional Predictive Ordinate (CPO) [26]. The Monte Carlo estimates of the DIC, EAIC, EBIC, and LPML criteria in Table 3 confirm that the proposed regression with complementary log–log link (com-log-log) (short Normal-GEV-CLL) is the best model.

Table 3. Some criteria for Normal-GEV regression models.

	Criteria			
Link Function	**DIC**	**EAIC**	**EBIC**	**LPML**
Logistic	−995.926	−986.771	−956.228	497.039
Probit	−993.660	−984.603	−954.061	496.038
Cauchy	−995.063	−985.563	−955.020	496.579
Com−log−log	−996.674	−987.814	−957.271	497.283
Log−Log	−991.124	−981.603	−951.060	494.889

The Bayesian estimates under quadratic and absolute losses of the parameters of the Normal-GEV-CLL and Johnson's S_B regression models and the 95% HPD intervals are reported in Table 4. All covariates are statistically significant at the significance level of 5% for all models. Figure 7 (left panel) displays the marginal posterior density of λ in the Normal-GEV-CLL regression model, which is symmetric. Table 4 reveals that the posterior mean of λ is 1.088, and a 95% HPD interval is $(0.0110, 2.103)$. We fit the GJS-Student-t regression model [12] with four degrees of freedom and log–log link to the current data. Table 5 reports the Monte Carlo estimates of DIC, EAIC, EBIC, and LPML for Jhonson's S_B and GJS-Student-t regression models. They indicate that the second regression is better than the first for these data, but it does not provide a better fit than the Normal-GEV-CLL regression model. The quantile–quantile (QQ) plot of the posterior normalized randomized quantile residuals for the last regression in Figure 7 (right panel) proves an acceptable fit [27,28].

Table 4. Estimates and 95% HPD intervals for the Johnson's S_B and Normal–GEV regression models with com-log-log link function.

	Normal-GVEV			Johnson's S_B		
Parameter	**Mean**	**Median**	**HPD (95%) Interval**	**Mean**	**Median**	**HPD (95%) Interval**
β_0	−1.879	−1.877	(−1.966, −1.796)	−1.905	−1.906	(−1.993, −1.826)
β_1	−0.398	−0.397	(−0.444, −0.363)	−0.391	−0.391	(−0.430, −0.348)
β_{2_1}	0.343	0.342	(0.265, 0.430)	0.359	0.361	(0.277, 0.440)
β_{2_2}	0.797	0.795	(0.716, 0.894)	0.801	0.802	(0.713, 0.888)
τ_0	2.641	2.644	(1.906, 3.352)	1.045	1.051	(0.810, 1.266)
τ_1	0.671	0.674	(0.376, 0.968)	0.336	0.339	(0.123, 0.513)
τ_{2_1}	0.545	0.544	(0.225, 0.846)	0.846	0.842	(0.596, 1.099)
τ_{2_2}	0.545	0.544	(0.225, 0.846)	0.846	0.842	(0.596, 1.099)
λ	1.088	1.088	(0.011, 2.103)			

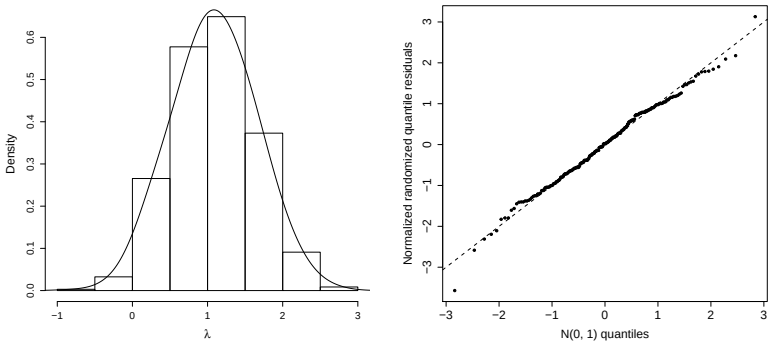

Figure 7. Marginal posterior density for λ (**left** panel) and QQ plot of the posterior normalized randomized quantile residuals (**right** panel) for the Normal-GEV-CLL regression model.

Table 5. Monte Carlo estimates of DIC, EAIC, EBIC, and LPML for Jhonson's S_B and GJS-Student-t regression models.

	Criteria			
Model	**DIC**	**EAIC**	**EBIC**	**LPML**
GJS-t Student	−984.764	−976.786	−949.637	492.665
Jhonson's	−986.343	−978.375	−951.226	491.750

We consider a Bayesian global influence methodology to identify the presence of outliers and/or influential observations under the general divergence measure [29]. Let $D_\psi(\pi, \pi_{(-i)})$ be the ψ-divergence between π and $\pi_{(-i)}$, where π denotes the posterior distribution of $\boldsymbol{\vartheta}$ for the full dataset , and $\pi_{(-i)}$ the posterior distribution of $\boldsymbol{\vartheta}$ without the ith observation, namely:

$$D_\psi(\pi, \pi_{(-i)}) = \int \psi\left(\frac{\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}^{(-i)})}{\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})}\right)\pi(\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}})\, d\boldsymbol{\vartheta} = E_{\boldsymbol{\vartheta}|\boldsymbol{\mathcal{D}}}\left[\psi\left(\frac{CPO_i}{f(y_i;\boldsymbol{\vartheta})}\right)\right],$$

where ψ is a convex function with $\psi(1) = 0$. Different choices of ψ are addressed by Dey and Birmiwal [30] and Pardo [31]. Here, $\psi(z) = -\log(z)$ defines the Kullback–Leibler (K-L) divergence, $\psi(z) = (z-1)\log(z)$ gives the J-distance, $\psi(z) = 0.5|z-1|$ provides L_1 norm, and $\psi(z) = z(1/z-1)^2$ yields the χ^2-square divergence. The divergence measure to verify whether a small subset of observations from the full data is influential or not follows the criterion by Peng and Dey [29] and Weiss [32]; see also Cancho et al. [15,33,34].

We calculate the Monte Carlo estimates of the divergence measures K-L, J, L_1, and χ^2 for the posterior distribution of the parameters of the Normal-GEV-CLL regression models to detect possible influential points.

They are plotted in Figure 8 and identify the cases 39, 54, and 122 as possible influential observations in the posterior distribution.

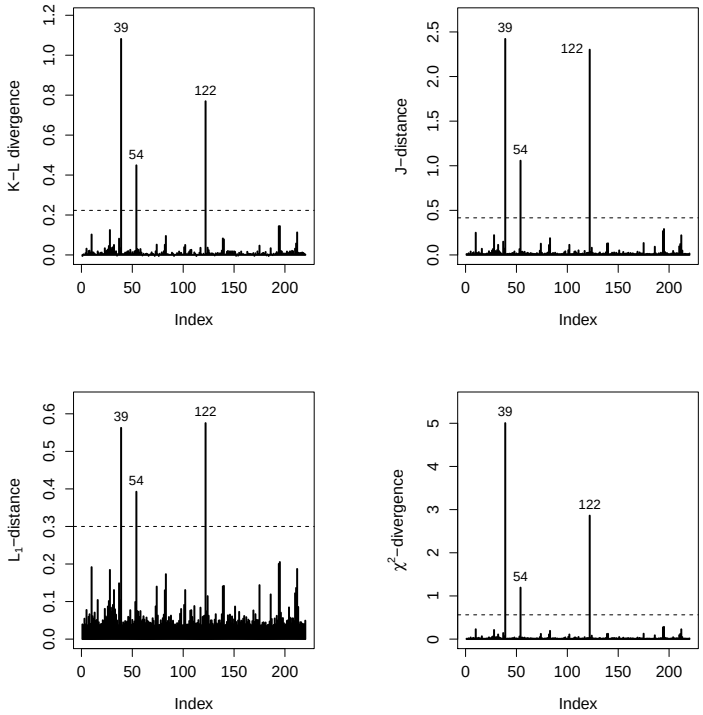

Figure 8. Index plot of ψ-divergence measures.

Table 6 presents subjects having large K-L, J, L_1, and χ^2.

Table 6. ψ-divergence measures for the Normal-GEV-CLL regression model.

Case	Mortality Rate	Sex	Race	State	K-L	J	L_1	χ^2
39	0.11	Men	non-Hispanic white	Dist-Columbia	1.081	2.422	0.563	5.004
54	0.05	Women	Hispanic	Georgia	0.449	1.057	0.392	1.193
122	0.29	Women	non-Hispanic black	Nebraska	0.769	2.301	0.575	2.861

For some cases, we refit the regression model to determine the impact of these observations on the posterior distribution of the parameters [15]. We eliminate each case individually and then two and three cases. The relative change (RC) of each estimate is $RC_{\vartheta_j} = (\widehat{\vartheta}_j - \widehat{\vartheta}_{j(I)})/\widehat{\vartheta}_j \times 100\%$, where $\widehat{\vartheta}_{j(I)}$ is the posterior mean of θ_j (for $j = 1, \ldots, 9$) when the set I of observations is removed.

Table 7 reports the RCs after removing some observations, and the lower (L) and upper (U) limits of the 95% HPD intervals of the new estimates. In general, the significance of the parameter estimates does not change after removing set I at the level of 5%.

Table 7. RCs (in %) and the L and U limits of the 95% HPD intervals after removing some cases.

Dropped		λ	β_0	β_1	β_{2_1}	β_{2_2}	τ_0	τ_1	τ_{2_1}	τ_{2_2}
none	Mean	1.088	−1.879	−0.398	0.343	0.797	2.641	0.671	0.545	−0.424
	L	0.011	−1.966	−0.444	0.265	0.716	1.906	0.376	0.225	−0.951
	U	2.103	−1.796	−0.363	0.430	0.894	3.352	0.968	0.846	0.101
{39}	RC	−39.0	−0.0	0.9	0.9	−0.3	−9.7	−21.4	25.2	−44.0
	L	−0.391	−1.960	−0.439	0.266	0.704	1.609	0.217	0.341	−0.744
	U	1.822	−1.797	−0.364	0.425	0.879	3.135	0.819	0.974	0.309
{54}	RC	−35.8	0.1	0.8	1.5	−0.1	−8.6	−20.7	22.4	−41.5
	L	−0.343	−1.964	−0.437	0.261	0.715	1.751	0.232	0.382	−0.777
	U	1.763	−1.802	−0.363	0.423	0.882	3.196	0.828	0.998	0.225
{122}	RC	23.3	−0.2	1.2	0.0	−1.4	5.2	17.2	−7.5	7.3
	L	0.305	−1.961	−0.442	0.260	0.692	2.040	0.475	0.195	−0.950
	U	2.324	−1.798	−0.363	0.420	0.867	3.435	1.059	0.801	0.050
{39,54}	RC	−50.7	−0.7	0.5	−3.2	−2.4	−10.3	−20.6	13.9	−37.8
	L	−0.528	−1.938	−0.436	0.251	0.693	1.693	0.262	0.333	−0.777
	U	1.602	−1.790	−0.363	0.402	0.859	3.179	0.835	0.929	0.243
{39,122}	RC	−17.0	−0.4	2.1	0.1	−1.8	−4.7	−5.2	17.1	−35.6
	L	−0.277	−1.955	−0.444	0.268	0.700	1.779	0.328	0.321	−0.807
	U	1.927	−1.794	−0.370	0.426	0.869	3.253	0.924	0.937	0.202
{54,122}	RC	12.1	−1.1	0.8	−5.6	−4.0	4.4	16.9	−17.5	11.8
	L	0.246	−1.933	−0.439	0.248	0.686	2.077	0.492	0.149	−0.978
	U	2.201	−1.789	−0.359	0.388	0.842	3.449	1.052	0.734	−0.034
{39,54,122}	RC	−29.7	−1.0	1.8	−3.7	−3.6	−5.5	−5.6	4.7	−31.0
	L	−0.388	−1.930	−0.444	0.261	0.694	1.782	0.338	0.270	−0.830
	U	1.848	−1.787	−0.370	0.402	0.846	3.323	0.947	0.883	0.223

We estimate the median of the mortality rate for eight patients A, B, C, D, E, and F with specified characteristics in Table 8. These numbers refer to the Bayes estimates under the quadratic and absolute loss functions and the 95% HPD intervals for the median mortality. For example, the median mortality rates are 0.147 and 0.276 for patient A of gender male and race Hispanic and patient E of gender male and race Hispanic, respectively. This difference can be seen in Figure 9, and in the posterior distribution of the median mortality rate of the other patients.

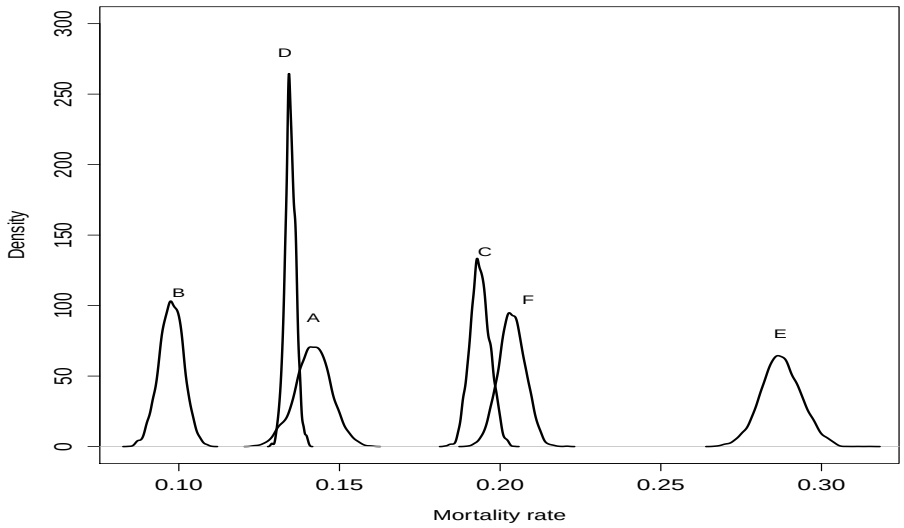

Figure 9. Posterior density of median of mortality rate for six hypothetical patients.

Table 8. Mortality rate estimates and the 95% HPD intervals for six colorectal cancer patients.

			Mortality Rate		
Patient	**Sex**	**Race**	**Mean**	**Median**	**HPD (95%) Interval**
A	Men	Hispanic	0.142	0.142	(0.131, 0.153)
B	Women	Hispanic	0.098	0.098	(0.090, 0.105)
C	Men	non-Hispanic white	0.194	0.194	(0.188, 0.200)
D	Women	non-Hispanic white	0.135	0.135	(0.131, 0.138)
E	Men	non-Hispanic black	0.287	0.287	(0.275, 0.299)
F	Women	non-Hispanic black	0.204	0.204	(0.195, 0.212)

6. Conclusions

We provided some mathematical properties of the new normal-generalized extreme value (Normal-GEV) distribution, and proposed a new and flexible regression model for proportional variables. This regression model is an alternative to the well-known beta regression model [2]. Some simulation studies and Bayesian procedures were developed to analyze a real dataset and they showed that the proposed regression is very competitive and useful for inferential and diagnostic problems involving bounded response variables and covariate variables.

Author Contributions: Conceptualization, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; methodology, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; software, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; validation, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; formal analysis, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; investigation, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; data curation, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; writing—original draft preparation, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; writing—review and editing, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; visualization, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C.; supervision, Y.R.B., V.G.C., E.M.M.O., R.V., G.M.C. All authors have read and agreed to the current version of the manuscript.

Funding: This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—Brasil (CAPES) (Finance Code 001).

Informed Consent Statement: Not applicable

Data Availability Statement: The authors confirm that the data supporting the findings of this study are available within the article.

Acknowledgments: We thank the anonymous reviewers whose comments/suggestions helped improve and clarify this manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Krysicki, W. On some new properties of the beta distribution. *Stat. Probab. Lett.* **1999**, *42*, 131–137.
2. Ferrari, S.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **2004**, *31*, 799–815.
3. Simas, A.B.; Barreto-Souza, W.; Rocha, A.V. Improved estimators for a general class of beta regression models. *Comput. Stat. Data Anal.* **2010**, *54*, 348–366.
4. Ospina, R.; Ferrari, S.L. Inflated beta distributions. *Stat. Pap.* **2010**, *51*, 111.
5. Ospina, R.; Ferrari, S.L. A general class of zero-or-one inflated beta regression models. *Comput. Stat. Data Anal.* **2012**, *56*, 1609–1623.
6. Carrasco, J.M.; Ferrari, S.L.; Arellano-Valle, R.B. Errors-in-variables beta regression models. *J. Appl. Stat.* **2014**, *41*, 1530–1547.
7. Figueroa-Zúñiga, J.I.; Arellano-Valle, R.B.; Ferrari, S.L. Mixed beta regression: A bayesian perspective. *Comput. Stat. Data Anal.* **2013**, *61*, 137–147.
8. Qiu, Z.; Song, P.X.-K.; Tan, M. Simplex mixed-effects models for longitudinal proportional data. *Scand. J. Stat.* **2008**, *35*, 577–596.
9. Barndorff-Nielsen, O.E.; Jørgensen, B. Some parametric models on the simplex. *J. Multivar. Anal.* **1991**, *39*, 106–116.
10. Bayes, C.; Bazan, J.L.; de Castro, M. A quantile parametric mixed regression model for bounded response variables. *Stat. Its Interface* **2017**, *10*, 483–493.
11. Kumaraswamy, P. A generalized probability density function for double-bounded random processes. *J. Hydrol.* **1980**, *46*, 79–88.
12. Lemonte, A.J.; Bazán, J.L. New class of johnson sb distributions and its associated regression model for rates and proportions. *Biom. J.* **2016**, *41*, 727–746.
13. Johnson, N.L. Systems of frequency curves generated by methods of translation. *Biometrika* **1949**, *36*, 149–176.
14. Smithson, M.; Shou, Y. Cdf-quantile distributions for modelling random variables on the unit interval. *Br. J. Math. Stat. Psychol.* **2017**, *70*, 412–438. [PubMed]
15. Cancho, V.; Bazán, J.L.; Dey, D.K. A new class of regression model for a bounded response with application in the study of the incidence rate of colorectal cancer. *Stat. Methods Med. Res.* **2020**, *29*, 2015–2033. [PubMed]
16. Jenkinson, A.F. The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Q. J. R. Meteorol. Soc.* **1955**, *81*, 158–171.
17. Griffiths, L. *Introduction to the Theory of Equations*; J. Wiley: New York, NY, USA, 1947.
18. Piessens, R.; de Doncker-Kapenga, E.; Uberhuber, C.W.; Kahaner, D.K. *QUADPACK A Subroutine Package for Automatic Integration*; Springer: Berlin, Germany, 1983.
19. Chib, S.; Greenberg, E. Understanding the metropolis-hastings algorithm. *Am. Stat.* **1995**, *49*, 327–335.
20. Cowles, M.K.; Carlin, B.P. Markov chain Monte Carlo convergence diagnostics: A comparative review. *J. Am. Stat. Assoc.* **1996**, *91*, 883–904.
21. Siegel, R.; DeSantis, C.; Jemal, A. Colorectal cancer statistics. *CA A Cancer J. Clin.* **2014**, *64*, 104–117.
22. Spiegelhalter, D.J.; Best, N.G.; Carlin, B.P.; van der Linde, A. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B* **2002**, *64*, 583–639.
23. Brooks, S.P. Discussion on the paper by Spiegelhalter, Best, Carlin, and van der Linde. *J. R. Stat. Soc. B* **2002**, *64*, 616–618.
24. Carlin, B.P.; Louis, T.A. *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed.; Chapman & Hall/CRC: Boca Raton, FL, USA, 2001.
25. Ibrahim, J.G.; Chen, M.-H.; Sinha, D. *Bayesian Survival Analysis*; Springer: New York, NY, USA, 2001.
26. Gelfand, A.; Dey, D.; Chang, H. Model determination using predictive distributions with implementation via sampling based methods (with discussion). *Bayesian Statistics 4*; Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., Eds.; Oxford University Press: Oxford, UK, 1992; Volume 1, pp. 7–167.
27. Dunn, P.K.; Smyth, G.K. Randomized quantile residuals. *J. Comput. Graph. Stat.* **1996**, *5*, 236–244.
28. Rigby, R.A.; Stasinopoulos, D.M. Generalized additive models for location, scale and shape (with discussion). *Appl. Stat.* **2005**, *54*, 507–554.
29. Peng, F.; Dey, D. Bayesian analysis of outlier problems using divergence measures. *Can. J. Stat.* **1995**, *23*, 199–213.
30. Dey, D.; Birmiwal, L.R. Robust bayesian analysis using divergence measures. *Stat. Probab. Lett.* **1994**, *20*, 287–294.
31. Pardo, L. *Statistical Inference Based on Divergence Measures*; Chapman & Hall/CRC: Boca Raton, FL, USA, 2006.
32. Weiss, R. An approach to Bayesian sensitivity analysis. *J. R. Stat. Soc. Ser. B* **1996**, *58*, 739–750.
33. Cancho, V.; Ortega, E.; Paula, G. On estimation and influence diagnostics for log-Birnbaum-Saunders student-t regression models: Full Bayesian analysis. *J. Stat. Plan. Inference* **2010**, *140*, 2486–2496.
34. Cancho, V.; Dey, D.; Lachos, V.; Andrade, M. Bayesian nonlinear regression models with scale mixtures of skew-normal distributions: Estimation and case influence diagnostics. *Comput. Stat. Data Anal.* **2011**, *55*, 588–602.