PLOS ONE

RESEARCH ARTICLE

Another unit Burr XII quantile regression model based on the different reparameterization applied to dropout in Brazilian undergraduate courses

Tatiane Fontana Ribeiro[1]*, **Fernando A. Peña-Ramírez**[2], **Renata Rojas Guerra**[3], **Gauss M. Cordeiro**[4]

1 Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, SP, Brazil, **2** Departamento de Estadística, Universidad Nacional de Colombia, Bogotá, Colombia, **3** Departamento de Estatística, Universidade Federal de Santa Maria, Santa Maria, RS, Brazil, **4** Departamento de Estatística, Universidade Federal de Pernambuco, Recife, PE, Brazil

* tatianefr@ime.usp.br

Citation: Ribeiro TF, Peña-Ramírez FA, Guerra RR, Cordeiro GM (2022) Another unit Burr XII quantile regression model based on the different reparameterization applied to dropout in Brazilian undergraduate courses. PLoS ONE 17(11): e0276695. https://doi.org/10.1371/journal.pone.0276695

Editor: Muhammad Amin, University of Sargodha, PAKISTAN

Received: March 13, 2022

Accepted: October 11, 2022

Published: November 3, 2022

Data Availability Statement: All relevant data are within the manuscript and its Supporting information files.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

In many practical situations, there is an interest in modeling bounded random variables in the interval (0, 1), such as rates, proportions, and indexes. It is important to provide new continuous models to deal with the uncertainty involved by variables of this type. This paper proposes a new quantile regression model based on an alternative parameterization of the unit Burr XII (UBXII) distribution. For the UBXII distribution and its associated regression, we obtain score functions and observed information matrices. We use the maximum likelihood method to estimate the parameters of the regression model, and conduct a Monte Carlo study to evaluate the performance of its estimates in samples of finite size. Furthermore, we present general diagnostic analysis and model selection techniques for the regression model. We empirically show its importance and flexibility through an application to an actual data set, in which the dropout proportion of Brazilian undergraduate animal sciences courses is analyzed. We use a statistical learning method for comparing the proposed model with the beta, Kumaraswamy, and unit-Weibull regressions. The results show that the UBXII regression provides the best fit and the most accurate predictions. Therefore, it is a valuable alternative and competitive to the well-known regressions for modeling double-bounded variables in the unit interval.

1 Introduction

University dropout is a problem with academic, social, and economic implications due to the high cost it inflicts on the students, their families, universities, and the country's growth [1]. Thus, it is necessary to extract relevant information to enable higher education institutions (HEIs) to understand this phenomenon and minimize the dropout proportion of their courses. In that idea, several authors studied how aspects of the organizational structure of universities affect student outcomes. See [2, 3], for instance. However, it is essential to look at appropriate

classes of regressions to model the dropout proportion, such as those based on distributions that lie on the standard unit interval.

The Beta [4] and Kumaraswamy [5, 6] regressions are the most widely used for modeling unit outcomes. The beta regression is useful to understand the influence of covariates on the response's mean. The Kumaraswamy is the classical alternative to the beta and allows modeling the quantile of a response in the unit interval. However, the search for alternative unit regressions has attracted many researchers' attention, especially those based on quantile approaches. For example, [7] introduced the unit-Weibull quantile regression. [8, 9] proposed the unit Burr XII and reflected unit Burr XII, respectively. Other quantile regressions were introduced by [10, 11], and [12]. One may also see [13–16] for unit regressions applied to educational measurements. These authors focus on comparing indicators from different countries, including educational attainment percentage, and school living conditions. However, to the author's knowledge, there is still a lack of information concerning the phenomenon of student dropout.

Under the above information, the goal of this paper is to propose a new alternative for unit quantile regression applied to the dropout proportion of undergraduate courses. We use an approach based on the unit Burr XII (UBXII) distribution, which was pioneered [8] by applying the transformation method in a Burr XII (BXII) random variable. Their choice was based on the versatility of the baseline, which has been applied in reliability analysis [17], regression modeling [18], generalized distributions [19, 20], and several other disciplines. Let Y be a unit random variable having the UBXII distribution. The cumulative distribution function (cdf) and probability density function (pdf) of Y are

$$F_Y(y; c, d) = (1 + \log^c y^{-1})^{-d}, \qquad 0 < y < 1, \tag{1}$$

and

$$f_Y(y; c, d) = c\, d\, y^{-1} \log^{c-1} y^{-1} (1 + \log^c y^{-1})^{-(d+1)}, \tag{2}$$

respectively, where $c > 0$ and $d > 0$ are shape parameters. The quantile function (qf) of Y follows by inverting Eq (1), namely

$$Q_Y(\tau) = \exp[\, -(\tau^{-1/d} - 1)^{1/c}\,], \qquad 0 < \tau < 1. \tag{3}$$

Henceforth, if Y is a random variable with pdf (2), we write $Y \sim$ UBXII (c, d). For $c = 1$, the UBXII distribution reduces to the unit Lomax distribution [21]. By taking $d = 1$, it is a special case of the unit log-logistic distribution [22]. Those models recently appeared in the literature, and the unit Lomax has not been studied in a regression context.

Our proposal is based on new reparametrization on Y by inverting its quantile function. We provide at least four motivations for this work. First, we propose a new reparametrization on Y and derive some useful statistical quantities that were not explored by [8]. Our investigation includes the computation of the score function and observed information matrix for distribution and also for the regression. Second, we consider a regression structure for the new quantile parameterby assuming that it can be expressed as a function of covariates and, hence, a more general class of regressions is obtained. The third motivation is to use a statistical learning tool for comparing the prediction performance of non-nested models and selecting the most suitable for the data at hand. The fourth motivation is referring to the usefulness of the new regression for modeling the dropout proportion of undergraduate courses. The motivating data set concern to Brazilian undergraduate animal science courses. This course has received attention in the literature; see, for instance, [23], who sought to identify demographic variables as well as their relation to students' performance and interest areas, and factors associated with enrollment in an introductory animal sciences course.

The rest of paper is outlined as follows. Section 2 proposes an alternative quantile parameterization for the UBXII distribution and investigates some of its mathematical and statistical properties. We obtain the maximum likelihood of the parameters in Section 3. We provide a simulation study in Section 4 to evaluate the performance of the estimators. In Section 5, we define a quantile regression model based on the new parameterization of the UBXII distribution. In addition, we discuss the estimation of the parameters, present some diagnostic analysis methods and regression selection criteria, and conduct simulation studies. In special, we present a statistical learning tool (cross-validation approach) to compare non-nested regressions. In Section 6, we perform an application of the new regression to dropout in Brazilian undergraduate animal sciences courses. We offer some conclusions in Section 7. Finally, we provide the observed matrix for the new distribution and Fisher's observed information matrix, and information about data's extraction used in application; see S1 Appendix and Supporting Information, respectively.

2 A new UBXII parametrization

Distributions with direct interpretation parameters are desirable in empirical applications, and for this purpose, several authors have adopted reparameterizations on well-known distributions; see [4, 7, 24, 25]. These reparameterizations generally seek to allow modeling of the random variable's mean, as in the [4]; and [25] proposal. However, the mean is an outlier-sensitive measure, and the UBXII distribution does not have a closed-form expression for it. Thus, modeling the quantiles is an interesting approach for asymmetric data because they can be outlier-resistant measures [24], besides being a smart alternative since the qf of Y (3) has a closed-form, and any quantile can be computed in explicit form. Further, one of the parameters of the UBXII distribution (under a quantile-parameterization) can be interpreted as the τth quantile of Y. Thus, we shall reparameterize Eq (1) in terms of the τth quantile $q = Q_Y(\tau)$. By inverting (3) and solving for d, we have

$$d = \log \tau^{-1} / \log(1 + \log^c q^{-1}). \tag{4}$$

By replacing (4) in Eqs (1) and (2), the cdf and pdf of the UBXII distribution (under this parametrization) have the forms

$$F_Y(y; q, c) = (1 + \log^c y^{-1})^{\log \tau / \log(1+\log^c q^{-1})}, \qquad 0 < y < 1, \tag{5}$$

and

$$f_Y(y; q, c) = \frac{\log \tau^{-c} \log^{c-1} y^{-1}}{y \log(1 + \log^c q^{-1})} (1 + \log^c y^{-1})^{\log \tau / \log(1+\log^c q^{-1})-1}, \tag{6}$$

respectively. Henceforth, we denote a random variable with density (6) by $Y \sim$ UBXII (c, q).

Some UBXII densities (for $\tau = 0.5$) are displayed in Fig 1, which reveal different shapes such as decreasing, increasing, reverse J-shaped, U-shaped, reverse tilde-shaped (decreasing-increasing-decreasing), non-skewed, and skewed-left. It is noteworthy that the UBXII density can accommodate several skew-left shapes and has a reverse tilde-shaped, which is not presented by classical unit distributions.

The qf of Y on the new parameterization has the form

$$Q_Y(u) = \exp\{-[u^{\log(1+\log^c q^{-1})/\log \tau} - 1]^{1/c}\}, \qquad 0 < \tau < 1. \tag{7}$$

So, the UBXII quantiles can be obtained from (7) by setting u values. Further, we can generate occurrences for this distribution using (7) by the inversion method.

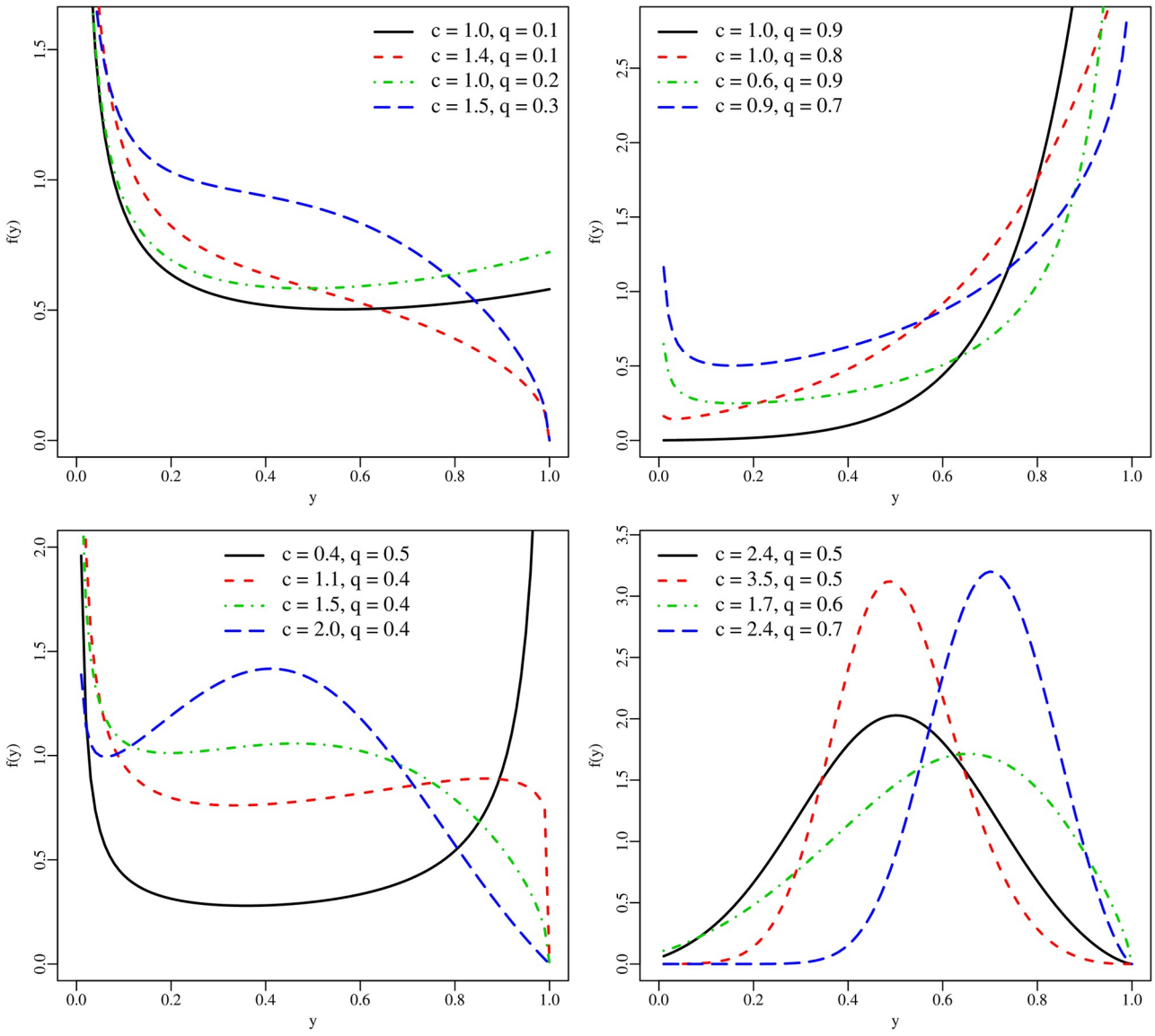

Fig 1. Plots of the UBXII density (with $\tau = 0.5$).

https://doi.org/10.1371/journal.pone.0276695.g001

Alternatively, the flexibility of the new distribution can be displayed from the Bowley skewness and Moors kurtosis formulas, namely

$$B = \frac{Q_Y(3/4) - 2Q_Y(1/2) + Q_Y(1/4)}{Q_Y(3/4) - Q_Y(1/4)}$$

and

$$M = \frac{Q_Y(7/8) - Q_Y(5/8) + Q_Y(3/8) - Q_Y(1/8)}{Q_Y(3/4) - Q_y(1/4)},$$

respectively, where $Q_Y(\cdot)$ is the qf given by (7). These measures provide a simple way to figure out the skewness and tail shapes of the distribution. Fig 2 displays plots for both measures B and M which show that they are sensible to variations of c and q for fixed $\tau = 0.5$.

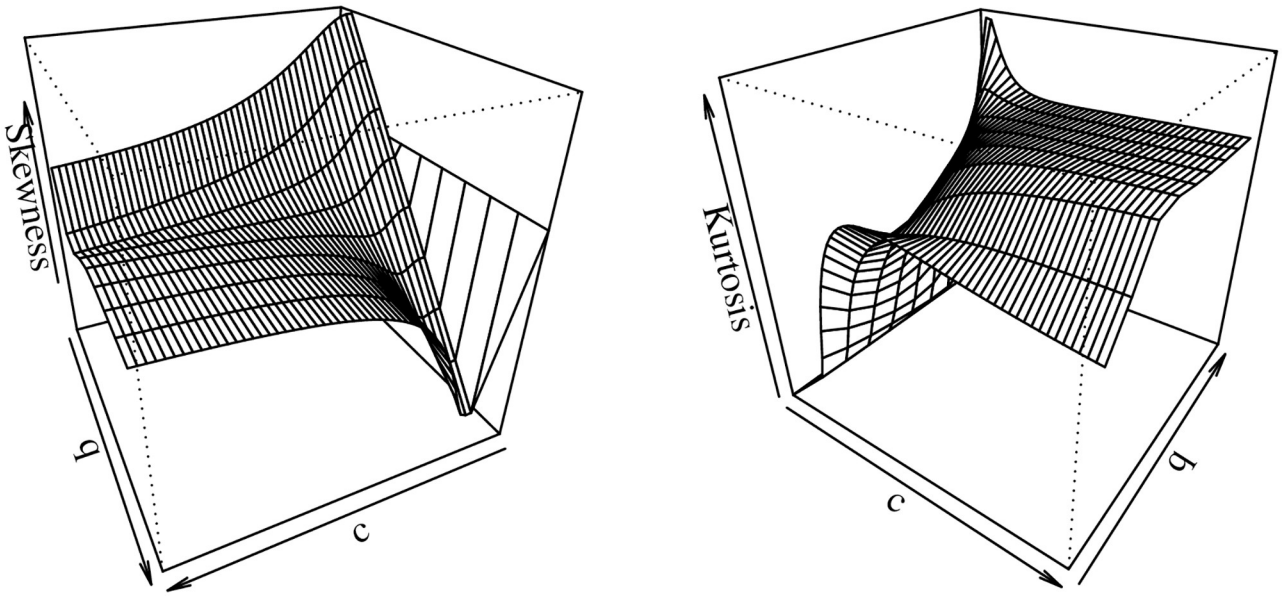

Fig 2. The Bowley skewness and Moors kurtosis of the UBXII distribution.

https://doi.org/10.1371/journal.pone.0276695.g002

3 Estimation

Various methods can be used to estimate the parameters of a distribution. The maximum likelihood (ML) method is the most commonly used. In what follows, we shall use this method for estimating the parameters of the UBXII distribution.

Let $y_1, \ldots, y_n$ be a random sample of size n from the UBXII distribution, the parameter vector $\boldsymbol{\theta} = (c, q)^\top$, and a known $\tau \in (0, 1)$ specified. Based on this sample, the log-likelihood function for $\boldsymbol{\theta}$, $\ell(\boldsymbol{\theta};y) \equiv \ell(\boldsymbol{\theta})$, has the form

$$\begin{aligned}\ell(\boldsymbol{\theta}) = \quad & n\log(\log\tau^{-c}) - n\log\{\log[t(q)]\} - \sum_{i=1}^{n}\log y_i + (c-1)\sum_{i=1}^{n}\log(\log y_i^{-1}) \\ & -\left[1 + \frac{\log\tau^{-1}}{\log[t(q)]}\right]\sum_{i=1}^{n}\log[t(y_i)],\end{aligned} \tag{8}$$

where $t(x) = 1 + \log^c x^{-1}$.

Eq (8) can be maximized either directly by using well-known plataforms such as the `R` (`optim` function), `SAS` (`PROC NLMIXED`), `Ox` program (`MaxBFGS` sub-routine) or by solving the nonlinear likelihood equations from the differentiation of $\ell(\boldsymbol{\theta})$. By maximizing (8), we obtain the MLE $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$.

Graphically, it is possible to show local maxima of the log-likelihood function ($\hat{\boldsymbol{\theta}}$) and that it is unimodal. Plots that illustrate this are constructed in four steps. First, we simulate data from UBXII(c, d), where $c = 1.5$ and $d = 3.4$ with $n = 100$. Second, we evaluate the log-likelihood function obtained from the pdf of Eq (2) in a range covering the respective ML estimate, that is, $c \in (0, 9)$ for d fixed at 3.4. After, the same is done for $d \in (0, 9)$ by fixing c at 1.5. Finally, we plot the log-likelihood function against the values range of the parameters c and d. Fig 3 displays the plots obtained. As expected, both log-likelihood functions are unimodal and their maxima points (ML estimates) are achieved on the true values of c and d, respectively.

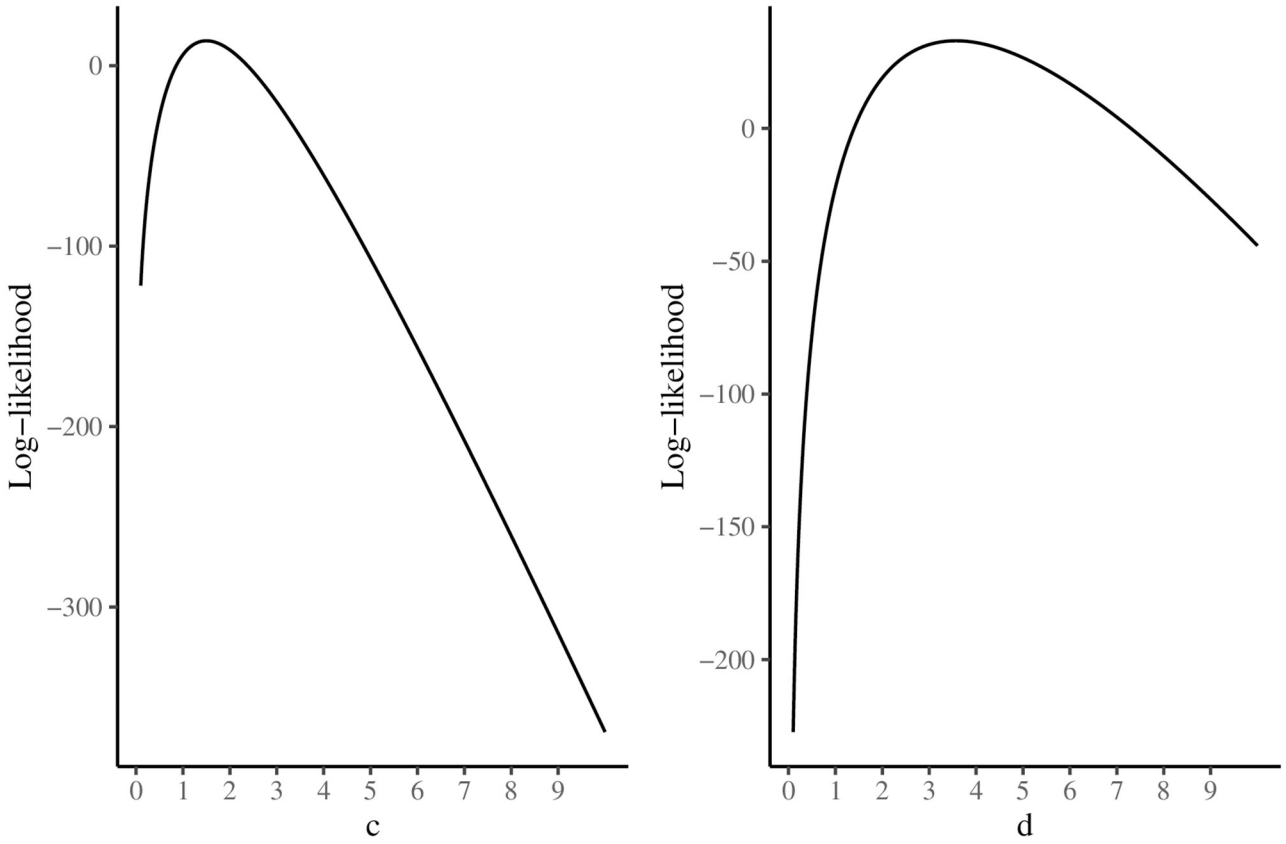

Fig 3. Plots of the log-likelihood function against the parameters *c* and *d*.

https://doi.org/10.1371/journal.pone.0276695.g003

The components of the score vector from Eq (8) are $U(\boldsymbol{\theta}) = [U_c(\boldsymbol{\theta}), U_q(\boldsymbol{\theta})]^\top$, where $U_c(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial c$ and $U_q(\boldsymbol{\theta}) = \partial\ell(\boldsymbol{\theta})/\partial q$. Setting these components to zero and solving them simultaneously gives $\hat{\boldsymbol{\theta}}$. The score components are

$$\begin{aligned} U_c(\boldsymbol{\theta}) = \quad & \frac{n}{c} + \sum_{i=1}^{n} \log(\log y_i^{-1}) - \frac{n \log(\log q^{-1})[t(q)-1]}{t(q)\log[t(q)]} - \sum_{i=1}^{n} \frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)} \\ & - \frac{\log \tau^{-1} \log[t(q)]}{\log^2[t(q)]} \sum_{i=1}^{n} [t(y_i)]^{-1}[t(y_i)-1]\log(\log y_i^{-1}) \\ & + \frac{\log\tau^{-1}[t(q)-1]\log(\log q^{-1})}{t(q)\log^2[t(q)]} \sum_{i=1}^{n} \log[t(y_i)], \end{aligned}$$

and

$$U_q(\boldsymbol{\theta}) = \frac{n\, c \log^{c-1} q^{-1}}{q\, t(q)\log[t(q)]} - \frac{\log \tau^{-c} \log^{c-1} q^{-1}}{q\, t(q)\log^2[t(q)]} \sum_{i=1}^{n} \log[t(y_i)].$$

The MLE of $\boldsymbol{\theta}$ can not be expressed in closed-form by setting $U(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}$. However, for fixed *c*, we note that a MLE semi-closed form of *q* follows by taking $U_q(\boldsymbol{\theta})|_{q=\hat{q}} = 0$. Hence, it is

the solution of

$$\hat{q}(c) = \exp\left(-\left\{\exp\left[\frac{1}{n}\log\tau^{-1}\sum_{i=1}^{n}\log[t(y_i)]\right]-1\right\}^{1/c}\right).$$

By replacing q by $\hat{q}(c)$ in Eq (8), we obtain the profile log-likelihood function

$$\begin{aligned}\ell(c) = -n + n\log(\log\tau^{-c}) - \sum_{i=1}^{n}\log y_i - \sum_{i=1}^{n}\log[t(y_i)] + (c-1)\sum_{i=1}^{n}\log(\log y_i^{-1}) \\ -n\log\left\{\frac{1}{n}\log\tau^{-1}\sum_{i=1}^{n}\log[t(y_i)]\right\}.\end{aligned} \tag{9}$$

We can compute the score function for c from (9)

$$U_c(c) = \frac{n}{c} + \sum_{i=1}^{n}\log(\log y_i^{-1}) - \sum_{i=1}^{n}\frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)} - \frac{n\sum_{i=1}^{n}\frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)}}{\sum_{i=1}^{n}\log[t(y_i)]}.$$

However, it is necessary to use a nonlinear optimization method to maximize numerically the profile log-likelihood function (9). Typically for the numerical computation of the MLEs, the quasi-Newton algorithm such as Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm is adopted.

Approximate confidence intervals and hypothesis tests for $\boldsymbol{\theta}$ can be constructed by considering its asymptotic distribution of the MLEs. For large samples, $\hat{\boldsymbol{\theta}} \sim \mathcal{N}(0, I^{-1}(\boldsymbol{\theta}))$ approximately assuming that standard regularity conditions (SRCs) hold, where $I(\boldsymbol{\theta})$ is the expected information matrix defined by

$$I(\boldsymbol{\theta}) = \mathbb{E}\left(-\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^{\top}}\right).$$

The computation of $I(\boldsymbol{\theta})$ may be cumbersome. Nevertheless, when the SRCs are valid, it follows that $I(\boldsymbol{\theta}) = \mathbb{E}[J(\boldsymbol{\theta})]$, where $J(\boldsymbol{\theta}) = -\partial^2\,\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}$ is the observed information matrix. For the UBXII distribution, we can write $J(\boldsymbol{\theta})$ as

$$J(\boldsymbol{\theta}) = -\begin{bmatrix} U_{cc}(\boldsymbol{\theta}) & U_{cq}(\boldsymbol{\theta}) \\ U_{qc}(\boldsymbol{\theta}) & U_{qq}(\boldsymbol{\theta}) \end{bmatrix},$$

where $U_{cc}(\boldsymbol{\theta}) = \partial^2\,\ell(\boldsymbol{\theta})/\partial c^2$, $U_{qq}(\boldsymbol{\theta}) = \partial^2\,\ell(\boldsymbol{\theta})/\partial q^2$, and $U_{cq}(\boldsymbol{\theta}) = \partial^2\,\ell(\boldsymbol{\theta})/(\partial c\partial q) = U_{qc}(\boldsymbol{\theta})$. The elements of the matrix $J(\boldsymbol{\theta})$ are given in S1 Appendix.

[26] proved that the estimated observed information matrix $J(\hat{\boldsymbol{\theta}})$ is a consistent estimator of $I(\boldsymbol{\theta})$ when the sample size is large. It is then possible to obtain the standard errors (SEs) of the MLEs by computing the square roots of the diagonal elements of $J(\hat{\boldsymbol{\theta}})^{-1}$. For instance, we can do large sample inference by building asymptotic confidence intervals with $100\%(1-\alpha)$ nominal coverage for $\boldsymbol{\theta}$ making $\hat{\boldsymbol{\theta}} \pm z_{1-\alpha/2}SE(\hat{\boldsymbol{\theta}})$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ standard normal quantile.

4 Simulation study

A Monte Carlo simulation study is carried out in the `R` programming language to evaluate the performance of the MLEs of the UBXII parameters that index the distribution. The `Optim`

Table 1. RB%s and RMSEs from the UBXII distribution.

Scenario	c	q	n	RB%		RMSE	
				$\hat{c}$	$\hat{q}(\hat{c})$	$\hat{c}$	$\hat{q}(\hat{c})$
1	1.5	0.3	25	7.3773	1.6170	0.3682	0.0751
			75	2.3671	0.8954	0.1823	0.0440
			150	1.1971	0.5372	0.1251	0.0310
			300	0.6399	0.3746	0.0872	0.0225
2	0.9	0.7	25	5.1296	−0.8436	0.1598	0.0758
			75	1.6937	−0.2415	0.0845	0.0434
			150	0.7708	−0.1126	0.0585	0.0311
			300	0.4013	−0.0741	0.0409	0.0221
3	1.1	0.4	25	6.3920	0.6669	0.2370	0.0967
			75	2.1153	0.6017	0.1216	0.0569
			150	1.1037	0.3580	0.0833	0.0402
			300	0.6429	0.3446	0.0583	0.0290
4	2.0	0.4	25	6.0735	0.2902	0.4203	0.0541
			75	1.9732	0.1424	0.2169	0.0313
			150	0.8997	0.0865	0.1496	0.0224
			300	0.4774	0.0212	0.1049	0.0159
5	1.7	0.6	25	5.0479	−0.1864	0.2940	0.0481
			75	1.6641	−0.0366	0.1555	0.0276
			150	0.7552	−0.0085	0.1075	0.0198
			300	0.3920	−0.0157	0.0755	0.0140
6	3.5	0.5	25	5.0270	0.0122	0.5975	0.0262
			75	1.6559	0.0203	0.3157	0.0150
			150	0.7520	0.0171	0.2182	0.0108
			300	0.3894	0.0012	0.1532	0.0076

https://doi.org/10.1371/journal.pone.0276695.t001

routine (with BFGS quasi-Newton nonlinear optimization algorithm and analytical derivative) is used for maximizing (9). The profile log-likelihood function involves a more straightforward numerical maximization than using (8) since it depends only on the parameter c. We start the root-finding algorithm using $c = 1$ for the shape parameter.

Different values for the parameter vector $\boldsymbol{\theta}$ are considered according to those presented in Fig 1. Therefore, various combinations of skewness and kurtosis coefficients and density shapes are contemplated. A total of six scenarios is considered for the sample size $n \in \{25, 75, 150, 300\}$. The inversion method is employed for generating observations, i.e., the qf (7) is evaluated in $u \sim \mathcal{U}(0, 1)$, being $Q_Y(u) = y$ and, hence, a sample of size n from $Y \sim$ UBXII (c, q) is generated. Each one of the sample sizes is replicated $R = 10,000$ times. We compute quantities as percentage relative bias (RB%) and root mean squared error (RMSE) of the MLEs.

Table 1 reports results from the simulation schemes. As expected, the consistency property of the MLEs holds, i.e., the RMSEs tend to decrease when the sample size increases. Also, it can be noted that the RB%s are smaller for sample size higher, thus indicating that the overall performance of the MLEs is appropriate, as well as they are more accurate and less biased when n increases. Notice that the biggest RB%s for $\hat{c}$ and $\hat{q}$ are less than 7.38 and 1.62, respectively, even with $n = 25$. In general, the estimate $\hat{q}$ is more accurate when compared with $\hat{c}$. In the scenarios two to six, all the RB%s of $\hat{q}$ are below of 0.84 in absolute value.

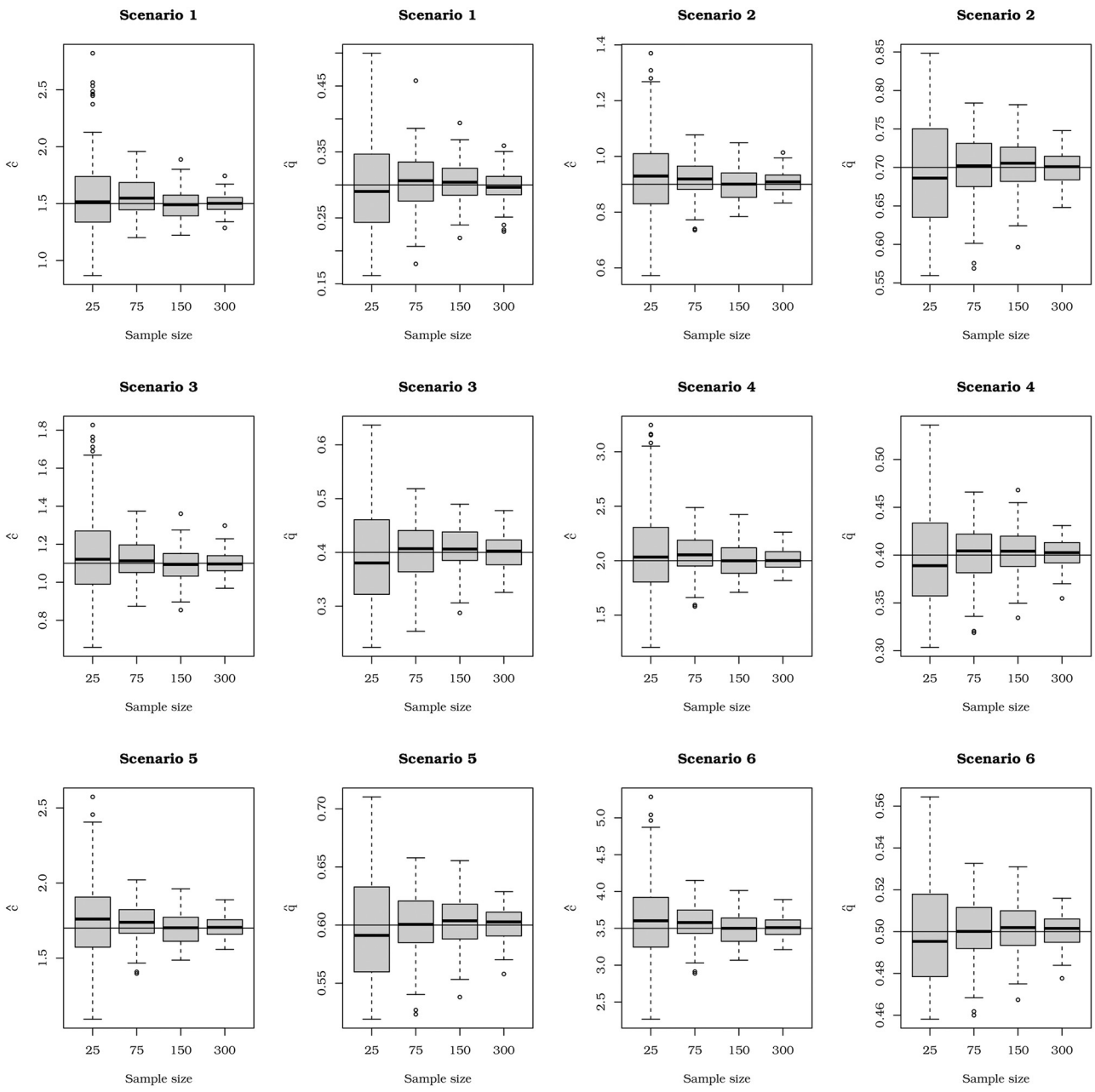

Fig 4. Boxplots of the first hundred estimates of the Monte Carlo simulation for some sample sizes.

https://doi.org/10.1371/journal.pone.0276695.g004

Fig 4 displays boxplots from the first 100 Monte Carlo replications (to favor easy viewing) of the eight current scenarios. We can note that, in most cases, the presence of outliers overestimates the estimates for small sample sizes. However, this fact is attenuated when n increases. Besides, the dispersion of the estimates decreases, and the precision is achieved for larger sample sizes.

Fig 5 contains plots of total absolute RB% and total RMSE versus sample sizes for all these scenarios. These quantities are obtained from the sum of the RB% and RMSE of both parameters for each sample size and scenario. Note that those measures decay to zero when n increases

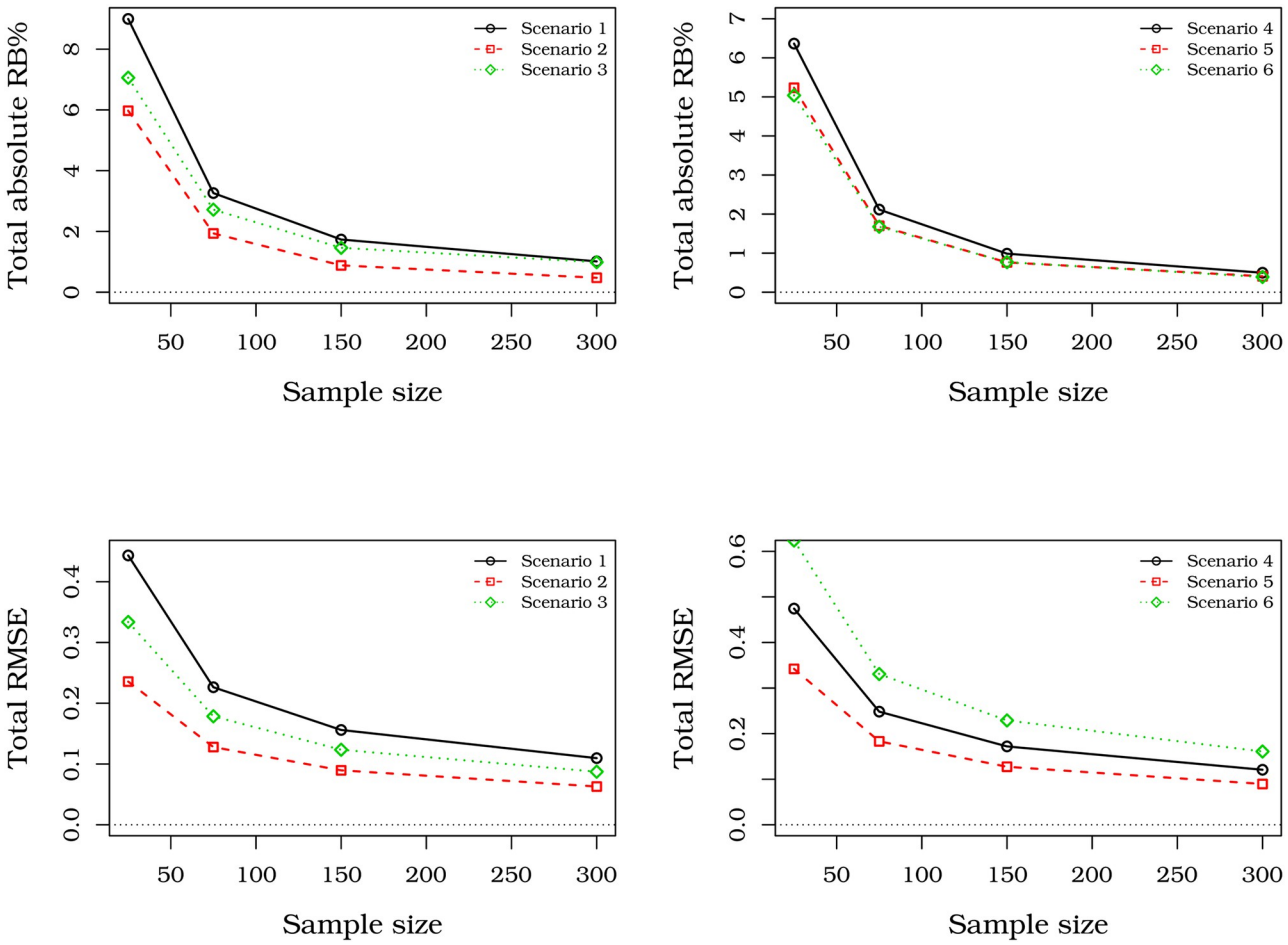

Fig 5. Total absolute RB%s and total RMSE of the MLEs from UBXII distribution with different sample sizes.

https://doi.org/10.1371/journal.pone.0276695.g005

in the six scenarios. This shows that the properties of the MLEs (such as asymptotically unbiased and consistent) are held.

5 The UBXII regression

Let $Y_1, \ldots, Y_n$ be n independent random variables, where $Y_i \sim$ UBXII (q_i, c) for $i = 1, \ldots, n$ with shape parameter c and quantile parameter q_i (both unknown) for $0 < \tau < 1$ assumed known. We propose the *UBXII regression* imposing that the quantile q_i of Y_i satisfies the functional relation

$$\boldsymbol{\eta} = g(\boldsymbol{q}) = \boldsymbol{X}\boldsymbol{\beta}, \tag{10}$$

where $\eta = (\eta_1, \ldots, \eta_n)^\top \in \mathbb{R}^n$ is the n-dimensional vector of linear predictors, $q = (q_1, \ldots, q_n)^\top$ is the vector of quantiles with $q_i \in (0, 1)$, $\beta = (\beta_1, \ldots, \beta_k)^\top \in \mathbb{R}^k$ is a k-dimensional vector of unknown regression coefficients $(k < n)$, $\boldsymbol{X} = (\boldsymbol{x}_1^\top, \ldots, \boldsymbol{x}_n^\top)^\top$ is the $n \times k$ full column rank matrix, $\boldsymbol{x}_i^\top = (x_{i1}, \ldots, x_{ik})$ denotes the ith observation on k covariates which are assumed known, and $x_{i1} = 1, \forall i$. Finally, we shall assume that $g(\cdot)$ is a strictly monotonic and twice differentiable link function which maps (0, 1) into $\mathbb{R}$. By inverting each component of (10), we

can write

$$q_i = g^{-1}\left(\sum_{j=1}^{k} x_{ij}\beta_j\right) = g^{-1}(\eta_i).$$

There are various possible choices for the link function $g(\cdot)$ such as

- logit: $g(q_i) = \log[q_i/(1-q_i)]$;
- probit: $g(q_i) = \Phi^{-1}(q_i)$, where $\Phi^{-1}(\cdot)$ is the qf of the standard normal random variable;
- complementary log-log: $g(q_i) = \log[-\log(1-q_i)]$;
- log-log: $g(q_i) = -\log[-\log(q_i)]$;
- Cauchy: $g(q_i) = \tan[\pi(q_t-1/2)]$.

The choice of the logit link function is the most common by practitioners since the interpretation of the regression parameters becomes quite interesting. Consider increasing the jth regressor at one unit, while the others are kept constant. Let q^* be the quantile of Y under the new value of x_j, whereas q denotes the quantile of Y under the original value of this regressor. It can be shown that with the logit link function, we have $\beta_j = \log\{q^*(1-q^*)/[q(1-q)]\}$, i.e., β_j is the log odds ratio [4]. In this context, we will consider the logit link function for $g(\cdot)$ in the UBXII regression. Then, the ith quantile of Y_i is $q_i = e^{\eta_i}/(1+e^{\eta_i})$.

5.1 Estimation

The parameters estimation in the UBXII regression can also be performed by the ML method. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, c)^\top$ be the vector of $k+1$ unknown parameters to be estimated. The log-likelihood function based on a sample of n independent observations having the UBXII distribution, i.e., $Y_i \sim$ UBXII (q_i, c), can be expressed as

$$\ell(\boldsymbol{\theta}) \equiv \ell(\boldsymbol{\beta}, c) = \sum_{i=1}^{n} \ell_i(q_i, c), \tag{11}$$

where $\ell_i(q_i, c)$ is the logarithm of $f_Y(y_i; q_i, c)$ given in Eq (6). Hence,

$$\ell_i(q_i, c) = \log(\log \tau^{-c}) - \log y_i + (c-1)\log(\log y_i^{-1}) - \log[t(y_i)] - \log\{\log[t(q_i)]\} - \frac{\log\tau^{-1}\log[t(y_i)]}{\log[t(q_i)]}.$$

The score vector, obtained by differentiating the log-likelihood function (11) with respect to the unknown parameters β_j, $j = 1, \ldots, k$, and c, is expressed as $\boldsymbol{U} = [U_\beta(\boldsymbol{\beta}, c)^\top, U_c(\boldsymbol{\beta}, c)]^\top$. The components of $\boldsymbol{U}$ can be written in matrix notation. For doing this, we now define some quantities.

Let $q_i^{\star} = \log^{c-1} q_i^{-1}/\{q_i\, t(q_i)\log[t(q_i)]\}$, $q_i^{\dagger} = \log\tau^{-1}\log^{c-1} q_i^{-1}/\{q_i\, t(q_i)\log^2[t(q_i)]\}$, $y_i^{\star} = \log[t(y_i)]$, and

$$\begin{aligned} y_i^{\sharp} = \quad & \frac{1}{c} + \log(\log y_i^{-1}) - \frac{\log(\log q_i^{-1})[t(q_i)-1]}{t(q_i)\log[t(q_i)]} - \frac{[t(y_i)-1]\log(\log y_i^{-1})}{t(y_i)} \\ & - \frac{\log\tau^{-1}\log[t(q_i)][t(y_i)]^{-1}[t(y_i)-1]\log(\log y_i^{-1})}{\log^2[t(q_i)]} \\ & + \frac{\log\tau^{-1}[t(q_i)-1]\log(\log q_i^{-1})\log[t(y_i)]}{t(q_i)\log^2[t(q_i)]}. \end{aligned}$$

Then, we have

$$U_{\boldsymbol{\beta}} \equiv U_{\boldsymbol{\beta}}(\boldsymbol{\beta}, c) = c\,\boldsymbol{X}^{\top}\boldsymbol{D}\,(\boldsymbol{q}^{\star} - \boldsymbol{q}^{\dagger}\boldsymbol{y}^{\star}), \tag{12}$$

and

$$U_c \equiv U_c(\boldsymbol{\beta}, c) = \mathrm{tr}(\boldsymbol{Y}^{\sharp}), \tag{13}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix whose ith row is $\boldsymbol{x}_i^{\top}$, $\boldsymbol{D} = \mathrm{diag}\,\{1/g'(q_1), \ldots, 1/g'(q_n)\}$, $\boldsymbol{q}^{\star} = (q_1^{\star}, \ldots, q_n^{\star})^{\top}$, $\boldsymbol{q}^{\dagger} = (q_1^{\dagger}, \ldots, q_1^{\dagger})^{\top}$, $\boldsymbol{y}^{\star} = (y_1^{\star}, \ldots, y_n^{\star})^{\top}$, and $\boldsymbol{Y}^{\sharp} = \mathrm{diag}\{y_1^{\sharp}, \ldots, y_n^{\sharp}\}$. We provide the calculations of the score components in S1 Appendix.

Again, the nonlinear Equations $U_{\boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = 0$ and $U_c|_{c=\hat{c}} = 0$ can not be expressed in closed-form. Hence, a nonlinear optimization method must be used for maximizing the function (11) and determine the MLEs $(\hat{\boldsymbol{\beta}}^{\top}, \hat{c})^{\top}$. We also provide the observed information matrix for $(\boldsymbol{\beta}^{\top}, c)^{\top}$.

To simplify the notation of its components, other quantities are defined as follows

$$m_i = \left\{\frac{c\log^c q_i^{-1}}{q_i\, t(q_i)} + \frac{c\log^c q_i^{-1}}{q_i\, t(q_i)\log[t(q_i)]} - \frac{\log q_i^{-1}}{q_i} - \frac{(c-1)}{q_i}\right\}\frac{c\log^{c-2} q_i^{-1}}{q_i\, t(q_i)\log[t(q_i)]},$$

$$p_i = \left\{\frac{(c-1)}{q_i\log[t(q_i)]} + \frac{\log q_i^{-1}}{q_i\log[t(q_i)]} - \frac{2\,c\log^c q_i^{-1}}{q_i\, t(q_i)\log^2[t(q_i)]} - \frac{c\log^c q_i^{-1}}{q_i\, t(q_i)\log[t(q_i)]}\right\}\frac{c\log\tau^{-1}\log^{c-2} q_i^{-1}}{q_i\, t(q_i)\log[t(q_i)]},$$

$$r_i = \left\{\log^{c-1} q_i^{-1} + \frac{\log^{c-1} q_i^{-1}}{c\log(\log q_i^{-1})} - \frac{\log^{2c-1} q_i^{-1}}{t(q_i)} - \frac{\log^{2c-1} q_i^{-1}}{t(q_i)\log[t(q_i)]}\right\}\frac{c\log(\log q_i^{-1})}{q_i\, t(q_i)\log[t(q_i)]},$$

$$u_i = \left\{\frac{2\log^{2c-1} q_i^{-1}}{t(q_i)\log^2[t(q_i)]} + \frac{\log^{2c-1} q_i^{-1}}{t(q_i)\log[t(q_i)]} - \frac{\log^{c-1} q_i^{-1}}{c\log(\log q_i^{-1})\log[t(q_i)]} - \frac{\log^{c-1} q_i^{-1}}{\log[t(q_i)]}\right\} \times \frac{c\log(\log q_i^{-1})\log\tau^{-1}}{q_i\, t(q_i)\log[t(q_i)]},$$

$$s_i = \frac{c\log\tau^{-1}\log^{c-1} q_i^{-1}}{q_i\, t(q_i)\log^2[t(q_i)]} \quad \text{and} \quad y_i^{\dagger} = \log(\log y_i^{-1})[t(y_i)-1][t(y_i)]^{-1}.$$

Therefore, the observed information matrix can be expressed as (see S1 Appendix)

$$\boldsymbol{J} = -\begin{pmatrix} \boldsymbol{J}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \boldsymbol{J}_{c\boldsymbol{\beta}} \\ \boldsymbol{J}_{\boldsymbol{\beta}c} & J_{cc} \end{pmatrix}.$$

The quantities $\boldsymbol{J}_{\boldsymbol{\beta\beta}} \equiv \partial^2 \ell(\boldsymbol{\beta}, c)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top$ and $\boldsymbol{J}_{\boldsymbol{\beta}\, c} = \boldsymbol{J}_{c\,\boldsymbol{\beta}}^\top \equiv \partial^2 \ell(\boldsymbol{\beta}, c)/\partial c\partial\boldsymbol{\beta}$, and $J_{cc} \equiv \partial^2 \ell(\boldsymbol{\beta}, c)/\partial c^2$ are

$$\boldsymbol{J}_{\boldsymbol{\beta\beta}} = \boldsymbol{X}^\top \left[(\boldsymbol{M} + \boldsymbol{P}\boldsymbol{Y}^\star)\boldsymbol{D} - c\,(\boldsymbol{Q}^\star - \boldsymbol{Q}^\dagger\boldsymbol{Y}^\star)\,\boldsymbol{T}\boldsymbol{D}^\top\boldsymbol{D} \right]\boldsymbol{D}\boldsymbol{X}, \tag{14}$$

$$\boldsymbol{J}_{c\,\boldsymbol{\beta}}^\top = (\boldsymbol{r} - \boldsymbol{s}\boldsymbol{y}^\ddagger + \boldsymbol{u}\boldsymbol{y}^\star)^\top \boldsymbol{D}\boldsymbol{X}, \tag{15}$$

and

$$J_{cc} = \mathrm{tr}(\boldsymbol{Y}^\diamond), \tag{16}$$

where $\boldsymbol{M} = \mathrm{diag}\,\{m_1, \ldots, m_n\}$, $\boldsymbol{P} = \mathrm{diag}\,\{p_1, \ldots, p_n\}$, $\boldsymbol{Q}^\star = \mathrm{diag}\,\{q_1^\star, \ldots, q_n^\star\}$, $\boldsymbol{Q}^\dagger = \mathrm{diag}\,\{q_1^\dagger, \ldots, q_n^\dagger\}$, $\boldsymbol{Y}^\star = \mathrm{diag}\,\{y_1^\star, \ldots, y_n^\star\}$, $\boldsymbol{T} = \mathrm{diag}\,\{g''(q_1), \ldots, g''(q_n)\}$, $\boldsymbol{r} = (r_1, \ldots, r_n)^\top$, $\boldsymbol{s} = (s_1, \ldots, s_n)^\top$, $\boldsymbol{y}^\ddagger = (y_1^\ddagger, \ldots, y_n^\ddagger)^\top$, and $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$.

As mentioned in Section 3, the matrix $\boldsymbol{J}$ is quite useful for interval estimation and hypothesis testing inference. Assuming that the SRCs hold and the sample size is large,

$$\begin{pmatrix} \hat{\boldsymbol{\beta}} \\ \hat{c} \end{pmatrix} \sim \mathcal{N}_{k+1}\left(\begin{pmatrix} \boldsymbol{\beta} \\ c \end{pmatrix}, \boldsymbol{I}^{-1} \right),$$

where $\boldsymbol{I}^{-1}$ is the inverse of $\boldsymbol{I} \equiv \mathbb{E}(\boldsymbol{J})$ is the expected information matrix. It can be estimated of the consistent way by $\hat{\boldsymbol{J}}$, which is computed after replacing the unknown parameters $(\boldsymbol{\beta}^\top, c)^\top$ by the corresponding MLEs.

5.2 Diagnostic measures and model selection

In order to check the goodness-of-fit and validate the UBXII regression assumptions, we adopt some well-known diagnostic tools that are now discussed. Initially, we used quantile residuals. These residuals verify if the model assumptions are satisfied and identify when the parameter estimations are considerably affected by the presence of atypical observations in the response. If the model is correctly specified, the quantile residuals are standard normally distributed. For the UBXII regression, they are given by

$$r_i = \Phi^{-1}[F_Y(y_i; \hat{q}_i, \hat{c})],$$

where $F_Y(\cdot)$ is the UBXII cdf given in Eq (5).

An incorrect functional form specification of the regression and the covariates omission can be identified through the RESET test. This test was initially introduced as a general misspecification test for the normal linear regression. Afterward, variants of the RESET test for classes of more general regressions were proposed by [27]. Thus, to determine whether a UBXII regression is misspecified, we propose using a RESET-like misspecification test. Next, we explain how this test can be performed.

The RESET-like test is carried out in two steps. Let $\hat{\boldsymbol{q}}$ be the predicted values vector obtained after fitting a UBXII regression. First, we build testing variables matrix as $\boldsymbol{T} = [\hat{\boldsymbol{q}}^2, \hat{\boldsymbol{q}}^3]$, where the vectors $\hat{\boldsymbol{q}}^2$ and $\hat{\boldsymbol{q}}^3$ are formed by $\hat{\boldsymbol{q}}$ squared and cubed components, respectively. We define the augmented regression

$$g(\boldsymbol{q}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{T}\boldsymbol{\delta}, \tag{17}$$

where $\boldsymbol{T}$ is the $n \times 2$ matrix of testing variables, and $\boldsymbol{\delta}$ is a 2×1 vector of parameters. Second, we estimate Eq (17) and test the null hypothesis $\mathcal{H}_0 : \boldsymbol{\delta} = \boldsymbol{0}$ against the alternative hypothesis $\mathcal{H}_1 : \boldsymbol{\delta} \neq \boldsymbol{0}$ by using the likelihood ratio (LR) statistic. We compute the LR statistic as $\omega =$

$2[\ell(\hat{\boldsymbol{\theta}} - \ell(\tilde{\boldsymbol{\theta}})]$, where $\ell(\cdot)$ is the log-likelihood function and $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\delta}}^\top, \hat{\boldsymbol{\beta}}^\top, \hat{c})^\top$ is the unrestricted MLE of $\boldsymbol{\theta}$, and $\tilde{\boldsymbol{\theta}} = (\mathbf{0}^\top, \tilde{\boldsymbol{\beta}}^\top, \tilde{c})^\top$ is the restricted MLE of $\boldsymbol{\theta}$ under the null hypothesis. Under $\mathcal{H}_0$ and the SRCs, ω converge in distribution to chi-square, χ^2_v, where v is the number of testing covariates added to the regression ($v = 2$ in this case). The non-rejection of the null hypothesis suggests that the regression is correctly specified.

The proportion of the response variable's variability explained by a fitted UBXII regression can be assessed using the generalized (pseudo) R-squared (R^2_G) defined by [28] as

$$R^2_G = 1 - \exp\{-2/n\,[\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_0)]\},$$

where $\ell(\hat{\boldsymbol{\theta}}_0)$ is the log-likelihood of the null regression, i.e., obtained from the modeling of the response in the covariates absence, and $\ell(\hat{\boldsymbol{\theta}})$ is the log-likelihood of the full regression. A regression with a higher value of R^2_G provides a larger explanation power of the response variable variation.

To select the more suitable model between several nested models, the information criteria such as Akaike information criterion (AIC) and Schwarz information criterion (BIC) can be considered. Both criteria are widely used in practical applications and they are defined by AIC $(\phi) = 2\,[p - \ell(\hat{\boldsymbol{\theta}})]$ and BIC $= p\log n - 2\ell(\hat{\boldsymbol{\theta}})$, where p is the number of estimated parameters.

A way of selecting the best one between different non-nested regressions is to assess its performance in the prediction of the response through statistical learning tools such as the cross-validation approach. Let $y = (y_1, \ldots, y_n)^\top$ the vector of n observations of a response variable and X the covariates matrix like in (10). In statistical learning methods, a training data set is the observations set in which a model is initially adjusted. An accuracy measure is the *test error*, that result from applying the model fitted to test observations that were not used in training. For example, if we use $(\boldsymbol{y}, \boldsymbol{X})$ as training observations, the test error is $\mathbb{E}[L(Y_0, \hat{y}_0)]$, where $L(\cdot)$ is the loss function and $\hat{y}_0$ is the predicted value using the fitted model from $(\boldsymbol{y}, \boldsymbol{X})$ evaluated in the predictors $\boldsymbol{x}_0{}^\top$ (that does not belong to $\boldsymbol{X}$). To estimate the test error with quadratic loss, we consider the mean square error (MSE) defined as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

where $\hat{y}_i$ is the ith predict value by the regression for the ith observation. This statistical measure is small if the predictions of the responses are very close to its true values, and it is large if for some of the observations, the predicted and true responses differ substantially [29].

As cross-validation method, we propose the use of the leave-one-out cross-validation (LOOCV). In this approach, we split the ith observation (ith row of a data set in which the response and covariates are disposed by columns) of the other n−1 observations that represent the training set whereas the row i is the validation set.

For each removed observation, we use the fitted model with the training set to predict the ith observation of the validation set. After, we estimate the test error by computing the MSEi. Repeating those procedure n times, we obtain $\text{MSE}_1, \ldots, MSE_n$. The final estimate of the test errors are computed through average of those n statistics as follows [29]

$$\text{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\text{MSE}_i.$$

Hence, we select the regression which provides smaller values for $\text{CV}_{(n)}$.

Finally, we perform an influence analysis to detect possible influential points as outliers. For this, the generalized Cook distance (GD) is considered. It is a measure of global influence,

which proposes eliminating the ith observation ($i = 1, \ldots, n$) to study its effect. The GD is computed as

$$GD_i = (\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}})^{\top}[J(\hat{\boldsymbol{\theta}})](\hat{\boldsymbol{\theta}}_{(i)} - \hat{\boldsymbol{\theta}}), \tag{18}$$

where $\hat{\boldsymbol{\theta}}_{(i)}$ is the MLE obtained when the ith observation is deleted, and $\boldsymbol{J}(\hat{\boldsymbol{\theta}})$ is the observed information matrix evaluated on the MLEs. We consider a general rule of thumb as a threshold for determining highly influential points. The rule is the following if $GD_i > 4/n$, then the observation is influential.

5.3 Simulation study

In this section, a Monte Carlo simulation study is conducted in order to numerically evaluate the finite sample behavior of the MLEs of the UBXII regression's parameters. The Monte Carlo experiments are performed using the `R` programming language [30]. Maximization of the log-likelihood function in (11) is carried out using the BFGS quasi-Newton nonlinear optimization algorithm implemented at the `optim` function available in `R`. We consider the ordinary least squares estimates (OLSEs) as an initial guess for $\boldsymbol{\beta}$ obtained from a linear regression of the transformed responses: $z = [g(q_1), \ldots, g(q_n)]^{\top}$, i.e., the initial point estimate of $\boldsymbol{\beta}$ is $\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^{\top}\boldsymbol{X})^{-1}\boldsymbol{X}^{\top}\boldsymbol{z}$. For the shape parameter c, we take the same initial guess in Section 4.

The simulations are based on the UBXII regression:

$$\text{logit}(q_i) = \beta_1 + \beta_2 x_{i2}, \qquad i = 1, \ldots, n. \tag{19}$$

The covariate $\boldsymbol{x_2}$ is randomly generated from a standard normal. We combine various values of the parameter vector $\boldsymbol{\theta} = (\beta_1, \beta_2, c)^{\top}$ at six different scenarios. The Monte Carlo replications number adopted and the sample sizes considered are the same from Section 4. In each Monte Carlo replication, the inversion method is used to generate n occurrences of a random variable $Y_i \sim$ UBXII (q_i, c). By assuming the regression structure defined in Eq (19), it follows that

$$q_i = \frac{\exp(\beta_1 + \beta_2 x_{i2})}{1 + \exp(\beta_1 + \beta_2 x_{i2})},$$

i.e., q_i is equal to the logistic cdf evaluated at $(\beta_1 + \beta_2 x_{i2})$. The statistical quantities computed are also the same of Section 4.

Table 2 presents the results of the Monte Carlo simulations. In general, the RB%s are smaller for larger sample sizes. We can note that the most RB% is equal to 10.02 in scenario four for the smallest sample size, and it refers to the estimate of c. For estimates of the parameters β_1 and β_2, all RB%s are below 6.25. In addition, even for $n = 25$, the RMSE values are quite low in any scheme.

Fig 6 displays plots for the total RB% and total RMSE versus sample sizes. They reveal that the MLEs are consistent, and their biases quickly tend to zero when the sample size grows. Further, the most RB% is about 20, but it decays to less than 5 to $n = 75$. Thus, as expected, the ML asymptotic properties remain.

We also investigate the behavior of the proposed model competing with the Kumaraswamy (Kw), unit-Weibull (UW) [7], and beta [4] regressions, which are well-known in the analysis of limited data. We aim to compare the performance of the maximum likelihood estimator for estimating the parameters and investigating their performance in case of misspecification of the distribution. We also evaluate the behavior of the AIC and BIC as selection criteria for models from different distributions.

Table 2. RB%s and RMSEs for the UBXII regression.

Scenario	β_1	β_2	c	n	RB%			RMSE		
					$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{c}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{c}$
1	1.3	1.4	2.0	25	−0.3323	0.7091	9.4106	0.1760	0.1499	0.4270
				75	−0.0978	0.3527	2.7208	0.0929	0.0913	0.1975
				150	−0.0392	0.1676	1.2937	0.0644	0.0570	0.1323
				300	−0.0068	0.1248	0.6621	0.0455	0.0451	0.0918
2	0.7	0.4	1.3	25	−1.6493	2.8842	7.8287	0.2680	0.2073	0.2504
				75	−0.1791	1.5523	2.5257	0.1462	0.1397	0.1246
				150	−0.0765	0.8871	1.1396	0.1043	0.0902	0.0843
				300	−0.0671	0.3218	0.5765	0.0732	0.0637	0.0584
3	−0.2	−0.6	1.8	25	6.2548	3.7917	9.8991	0.2599	0.1545	0.4274
				75	1.0153	0.9514	0.1454	0.1454	0.1264	0.1900
				150	0.0641	0.5987	1.3144	0.1001	0.0951	0.1324
				300	0.3399	0.2976	0.6825	0.0724	0.0590	0.0898
4	−0.7	0.4	2.3	25	1.7725	3.1922	10.0246	0.2625	0.2104	0.5838
				75	0.3187	2.0428	3.1417	0.1391	0.1240	0.2780
				150	0.0243	0.7296	1.4848	0.0979	0.0777	0.1894
				300	−0.0512	0.1449	0.7751	0.0691	0.0603	0.1317
5	1.2	−0.5	1.6	25	−0.1732	2.9301	7.9394	0.1860	0.1217	0.3028
				75	−0.0054	0.7715	2.5134	0.1039	0.1069	0.1470
				150	0.0099	0.3839	1.1571	0.0748	0.0663	0.0992
				300	−0.0200	0.2634	0.5976	0.0529	0.0486	0.0687
6	0.4	1.2	2.6	25	−1.4705	0.9087	9.6420	0.1666	0.1364	0.5731
				75	−0.2535	0.4271	2.7730	0.0854	0.0776	0.2673
				150	−0.0936	0.1443	1.3345	0.0597	0.0543	0.1779
				300	0.0397	0.0499	0.6723	0.0431	0.0417	0.1237

https://doi.org/10.1371/journal.pone.0276695.t002

Let Y be a random variable Kw distributed under a median-dispersion parameterization [5], say $Y \sim$ Kw(ω, d_p). The pdf of Y is

$$f(y;\omega,d_p) = \frac{\log 0.5}{d_p \log(1-\omega^{1/d_p})} y^{1/d_p}(1-y^{1/d_p})^{\log 0.5/\log(1-\omega^{1/d_p})-1}, \quad y \in (0,1)$$

where $0 < \omega < 1$ is the median of Y and $d_p > 0$ is a dispersion parameter.

The UW quantile regression was recently introduced by [7]. Let $Y \sim$ UW(q, γ) be a random variable having the UW distribution under the parameterization given in [7]. For $y \in (0, 1)$, the random variable Y has density

$$f(y;q,\gamma) = \frac{\gamma}{y}\left(\frac{\log\tau}{\log q}\right)\left(\frac{\log y}{\log q}\right)^{\gamma-1}\tau^{(\log y/\log q)^{\gamma}},$$

where $0 < q < 1$ is the τth quantile, $\gamma > 0$ is a shape parameter, and $\tau \in (0, 1)$ is assumed known. Here, it will be considered that $\tau = 0.5$ in order to model the median of Y.

[4] pioneered the beta regression. Different parameterizations can be considered for the beta distribution. We consider the mean-precision based parameterization. Let Y be a random

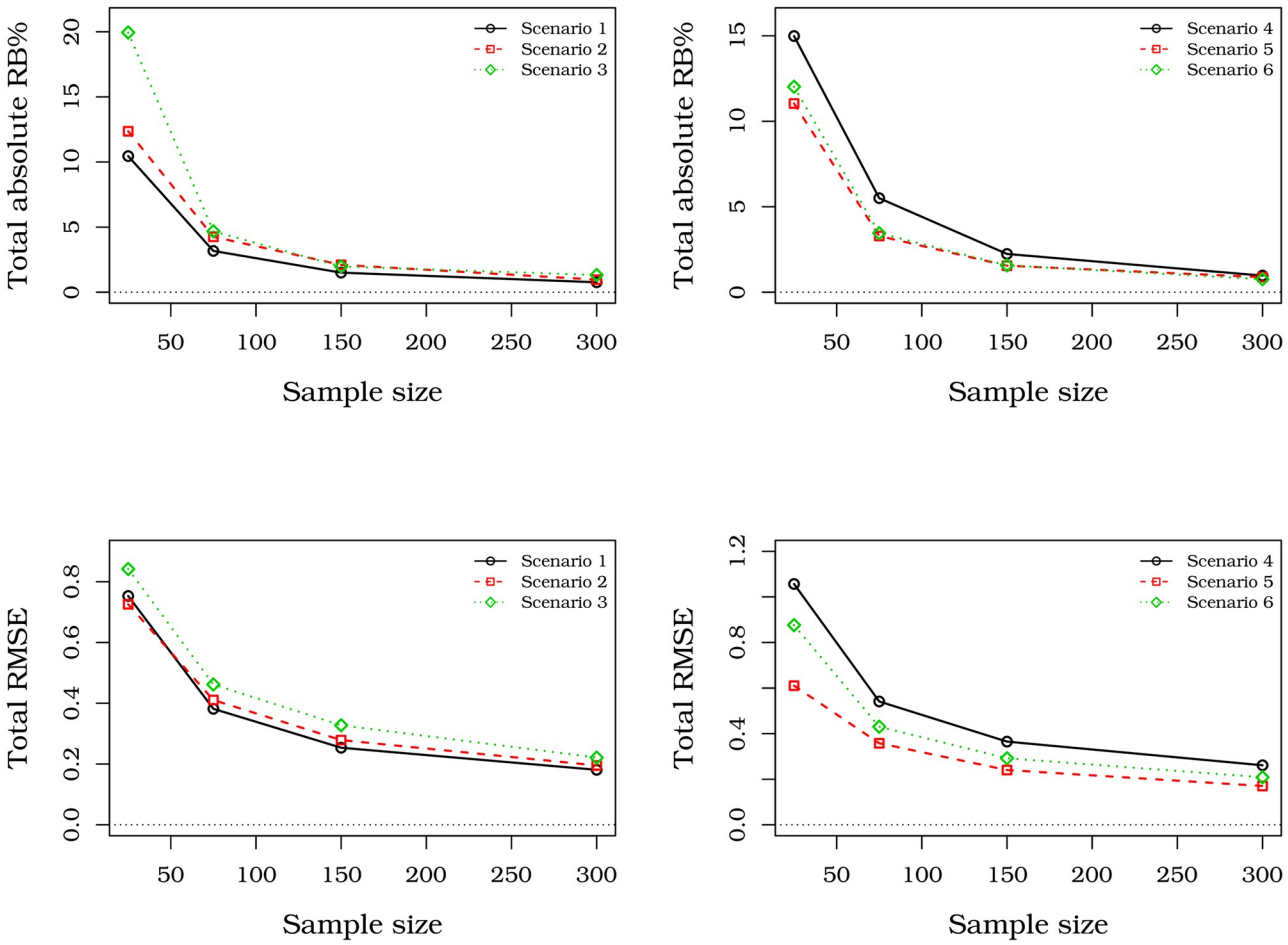

Fig 6. Total absolute RB%s and total RMSE of the MLEs from UBXII regression with different sample sizes.

https://doi.org/10.1371/journal.pone.0276695.g006

variable that follows a beta distribution, say $Y \sim$ Beta(μ, ϕ). For $y \in (0, 1)$, the Y density is

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1},$$

where $0 < \mu < 1$ is the mean of Y, $\phi > 0$ is a precision parameter and $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$ is the complete gamma function. Under this parameterization the variance of Y is $V(\mu)/(1+\phi)$, with $V(\mu) = \mu(1-\mu)$.

The regression structure for the Kw, UW, and beta distributions is analogous to (19). The main differences are the assumptions under the random components and modeled location parameters. To get the Kw regression, q must be replaced by the median (ω) in Eq (19) and supposed that $Y_i \sim$ Kw(ω_i, d_p). The UW regression is obtained by considering the structure (19) and assuming that $Y_i \sim$ UW(q_i, γ). In the beta regression, the location parameter is the mean (μ). Hence, in Eq (19), q must be replaced by μ and supposed that $Y_i \sim$ Beta(μ_i, ϕ). Thus, considering these regression structures, the simulation was performed in the following steps.

1. We generate a sample with $n = 100$ observations for each regression. The parameter values were selected from Scenario 1 in Table 2, replacing c for the respective shape parameter.

Table 3. Results of Monte Carlo simulations for Scenario 1 ($\beta_1 = 1.3, \beta_1 = 1.4$, and $c = 2.0$), with $n = 100$ and $R = 5000$ replications and MSE mean, AIC and BIC frequencies (%) of correct model selection.

Simulation	UBXII(q_i, c)				Kw(ω_i,d_p)		UW(q_i, γ)		Beta(μ_i, ϕ)	
Par./Meas.	UBXII	Kw	UW	Beta	UBXII	Kw	UBXII	UW	UBXII	Beta
$\hat{\beta}_1$	1.2984	0.7471	1.2972	0.9241	2.0396	1.3267	1.3238	1.2986	3.2141	1.2071
	(0.0802)	(0.4013)	(0.0949)	(0.1798)	(0.3435)	(0.2944)	(0.0752)	(0.0732)	(0.3661)	(0.1245)
$\hat{\beta}_2$	1.4040	1.0398	1.5873	1.2794	0.9411	1.4311	1.3597	1.4034	2.4634	1.2125
	(0.0670)	(0.2549)	(0.1246)	(0.2181)	(0.2664)	(0.2545)	(0.0624)	(0.0603)	(0.3905)	(0.1105)
$\hat{\boldsymbol{\theta}}$	2.0441	1.9806	1.3982	6.7714	0.4688	1.9139	2.1814	2.0451	0.4025	1.9896
	(0.1647)	(3.9387)	(0.3370)	(2.2466)	(0.0542)	(0.3801)	(0.1777)	(0.1661)	(0.0463)	(0.3092)
AIC (%)	92.1600	0.0000	7.8400	0.0000	0.0400	99.9600	3.6600	96.3400	0.0200	99.9800
BIC (%)	92.1600	0.0000	7.8400	0.0000	0.0400	99.9600	3.6600	96.3400	0.0200	99.9800
MSE	0.0053	1.0608	0.0136	0.3587	4.6296	1.0705	0.0044	0.0035	1.5143	0.0209

https://doi.org/10.1371/journal.pone.0276695.t003

2. We fit the true regression and the UBXII regression for each generating scheme. When the UBXII is the true model, we fit all the competitors.
3. We compute the MSE, AIC, and BIC for all fitted models.
4. For each scenario considered, 5,000 replications were performed.
5. For the MSE, we compute the average of all replications. For the AIC and BIC, the frequencies (%) of correct model selection are computed.

Table 3 displays the performance of the UBXII model when compared with the existing ones. The estimate $\hat{\boldsymbol{\theta}}$ is $\hat{c}$ for UBXII, $\hat{d}_p$ for Kw, $\hat{\gamma}$ for UW, and $\hat{\phi}$ for beta regression. We observe that the estimates obtained with the Beta and KW differ from those from the UBXII and UW distributions. The last two present estimates are close to each other. It shows that the traditional beta and Kw densities were not suitable to describe the data generated from the UBXII. The UW is the most competitive model but still presents a worse performance for fitting UBXII random variables. Concerning the model selection approaches, all measures were able to select the correct model. The results for AIC and BIC were very similar, their success rates exceeding 92% for all generating schemes. Thus, the information criteria are reliable for model selection among the considered competitive models.

6 Application

In this section, we assess the UBXII regression performance on real data. The analysis is carried out using the R statistical computing environment [30]. We fit the UBXII regression and compare it with the Kw, UW [7], and beta [4] regressions, which are well-known in the analysis of limited data and were also considered in the simulation experiment. The R codes of the simulation studies and application are available at https://github.com/tatianefribeiro/UBXII_regression. We get the data from the higher education census conducted yearly by the Brazilian National Institute for Educational Studies and Research "Anísio Teixeira". We are interested in the dropout proportion for animal sciences courses and factors associated with their enrollment and organizational structure. However, the response variable is not directly obtained from the original data set, and we use mining data techniques to obtain it from other reported variables. After preprocessing and cleaning steps, we select 40 covariates as possible predictors. A detailed description of the data mining tools employed and the final data set are available in Supporting information.

The UBXII, Kw, UW, and beta regressions also are used as data mining tools to select a subset of predictors that properly fits the dropout proportion. We test several combinations of predictors using the measures described in Section 6.3. to define the final regressions on each class. In what follows, we describe the response variable and predictive covariates used in our regression analysis.

The response variable is the dropout proportion from 2009 until 2017 of 77 Brazilian undergraduate animal sciences courses. For each course i ($i = 1, \ldots, 77$), we consider three covariates as follows: i) quantity of vacancies offered in the morning shift, denoted by x_{i2}; ii) a dummy variable that equals one if the course guarantees conditions of accessibility for people with disabilities, and zero otherwise, denoted by x_{i3}; and iii) a dummy variable, denoted by x_{i4}, that equals one if the course works on the night shift, and zero otherwise.

Let $\boldsymbol{y} = (y_1, \ldots, y_{77})^\top$ be the vector of the response variable and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4)$ the covariates matrix, where $\boldsymbol{x}_1$ is a vector column with 77 ones and $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{77j})^\top$, with $j = 2, \ldots, 4$. Table 4 provides a descriptive summary of the response variable ($\boldsymbol{y}$) and quantitative covariate ($\boldsymbol{x_2}$), revealing that $\boldsymbol{y}$ has negatively skewed distribution and lighter tails than a normal distribution. Further, its mean is close to the median, the standard deviation (SD) is low, and the values range is sizeable because the minimum and maximum are 0.1077 and 0.9714, respectively. The covariate x_2 presents different degrees of variability, skewness, and kurtosis.

To study the covariates' effects on the median dropout proportion, we set $\tau = 0.5$ and specify the UBXII regression as

$$\text{logit}(q_i) = \eta_i = \beta_1 + \beta_2\, x_{i2} + \beta_3\, x_{i3} + \beta_4\, x_{i4},$$

For comparison purposes, we also fit the Kw, UW, and beta regressions considering the same covariates combination and link function.

Table 5 brings some goodness-of-fit measures such as AIC, BIC, and R^2_G, the p-values of the Anderson-Darling test (AD) [31] to validate the null hypothesis that errors are normally distributed, the p-values from RESET-like test (RES), and the statistic obtained from the LOOCV approach ($CV_{(77)}$) that allows assessing the prediction performance of the fitted regressions. We consider $\alpha = 0.05$ as a significance level for all performed hypothesis tests. According to the RESET-like tests, all models are correctly specified. Similarly, the p-values from Anderson-Darling tests indicate is reasonable supposing normality of the errors at each class. It is noteworthy that most of some goodness-of-fit measures suggest that the UBXII regression is more suitable to fit the dropout proportion in the Brazilian zootechnics course between 2009 and

Table 4. Descriptive statistics from the response variable and quantitative covariates.

Var.	Statistics						
	Mean	Median	SD	Skewness	Kurtosis	Min.	Max.
y	0.5736	0.5965	0.1818	−0.3449	0.0854	0.1077	0.9714
x_2	13.7532	0.0000	29.5449	2.0533	3.2902	0.0000	120.0000

https://doi.org/10.1371/journal.pone.0276695.t004

Table 5. Goodness-of-fit measures and LOOCV statistic for the fitted regressions.

Regression	AIC	BIC	R^2_G	AD	RES	$CV_{(77)}$
UBXII	−55.8423	−44.1233	0.2348	0.8229	0.4334	0.0259
Kw	−48.8064	−37.0873	0.1898	0.2765	0.8354	0.0260
UW	−52.6329	−40.9139	0.2565	0.2795	0.5764	0.0266
BETA	−52.0595	−40.3405	0.2235	0.5433	0.8383	0.0285

https://doi.org/10.1371/journal.pone.0276695.t005

2017 than other considered class of regressions. Moreover, the $CV_{(77)}$ estimate for the fitted UBXII regression is the smallest among all other fitted regressions. This means that the proposed regression leads to better predictions than the classical regressions used in the context of restricted response to the unit interval. Indeed, in Fig 7 it is possible to note that the UBXII regression provides the best fit for this data set since about 97% of the points are under the red line in the QQ-plot of fitted UBXII regression's residuals.

In Table 6, we provide the estimates of the parameters, standard errors, t statistic value, and p-values for the UBXII regression. Results from other fitted regressions are given in Supporting information; see Table 2. The effect of the three considered covariates under the response's median is positive. Further, according to the estimate of β_4, the covariate x_{i4} presents the most impact on the median. That is, the odds ratio increases substantially if the course works on the night shift. This result may be related to the fact that many of the night students need to work during the day, making it challenging to persist [32]. However, the offer of night courses results from conquests achieved by popular pressure to meet the requirements of a population mainly consisting of workers [33]. Thus, our finding raises the discussion on the need to

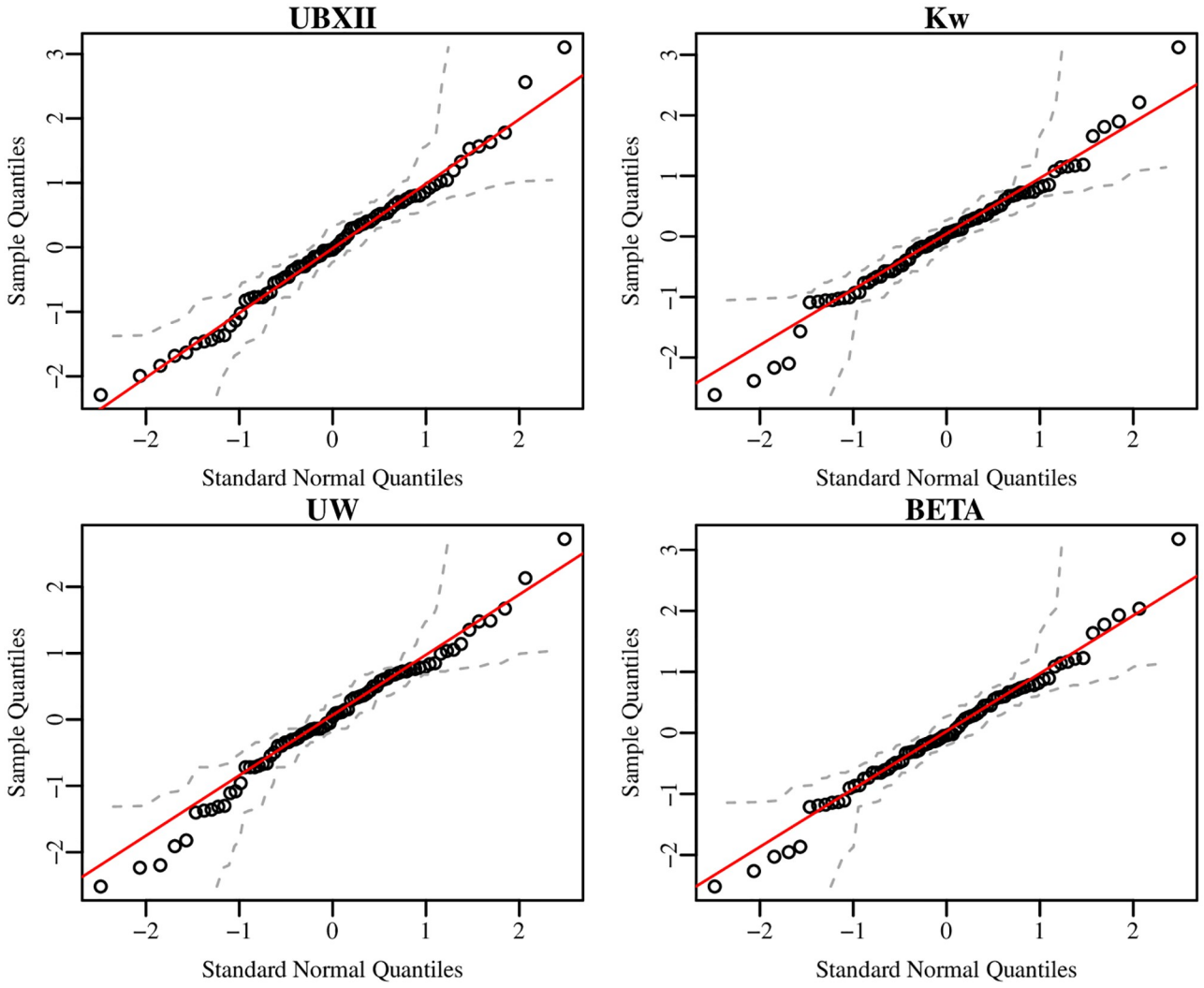

Fig 7. QQ-plots of the UBXII, Kw, UW, and beta regressions' residuals.

https://doi.org/10.1371/journal.pone.0276695.g007

Table 6. Fitted UBXII regression for the dropout proportion in the Brazilian zootechnics course.

Parameter	Estimate	Std. Error	t value	*p*-value
β_1	−0.0509	0.1294	−0.3932	0.6953
β_2	0.0082	0.0024	3.4429	0.0010
β_3	0.5389	0.1560	3.4535	0.0009
β_4	0.8310	0.2665	3.1183	0.0026
c	2.3780	0.2032	—	—

https://doi.org/10.1371/journal.pone.0276695.t006

provide a better service to this public. For example, the low offer of extracurricular activities for evening students is one of the problems reported by [33].

Fig 8 plots the GD for the UBXII regression. We can note that only observation 32 highlights the others. It corresponds to the *Faculdade de Estudos Superiores de Minas Gerais* and, with dropout proportion of 0.8163, is upper the 3th quartile of the data set. However, it is not potentially influential since the GD associated is smaller than $4/n$. Fig 9 assesses the impact of different τ values on the parameter estimates. We compute the 95% confidence intervals and point estimates for the UBXII regression by considering $\tau \in \{0.1, 0.2, \ldots, 0.9\}$. We observe that the intercept estimates become higher as τ increases, and the other regression coefficients are negatively related to the quantiles. It indicates that the covariates have a more substantial impact on explaining smaller quantiles of the dropout proportion. Finally, $\hat{c}$ does not seem to be affected by variations of τ values. It is worth noting that similar behavior is reported by [7] for the shape parameter of the UW regression.

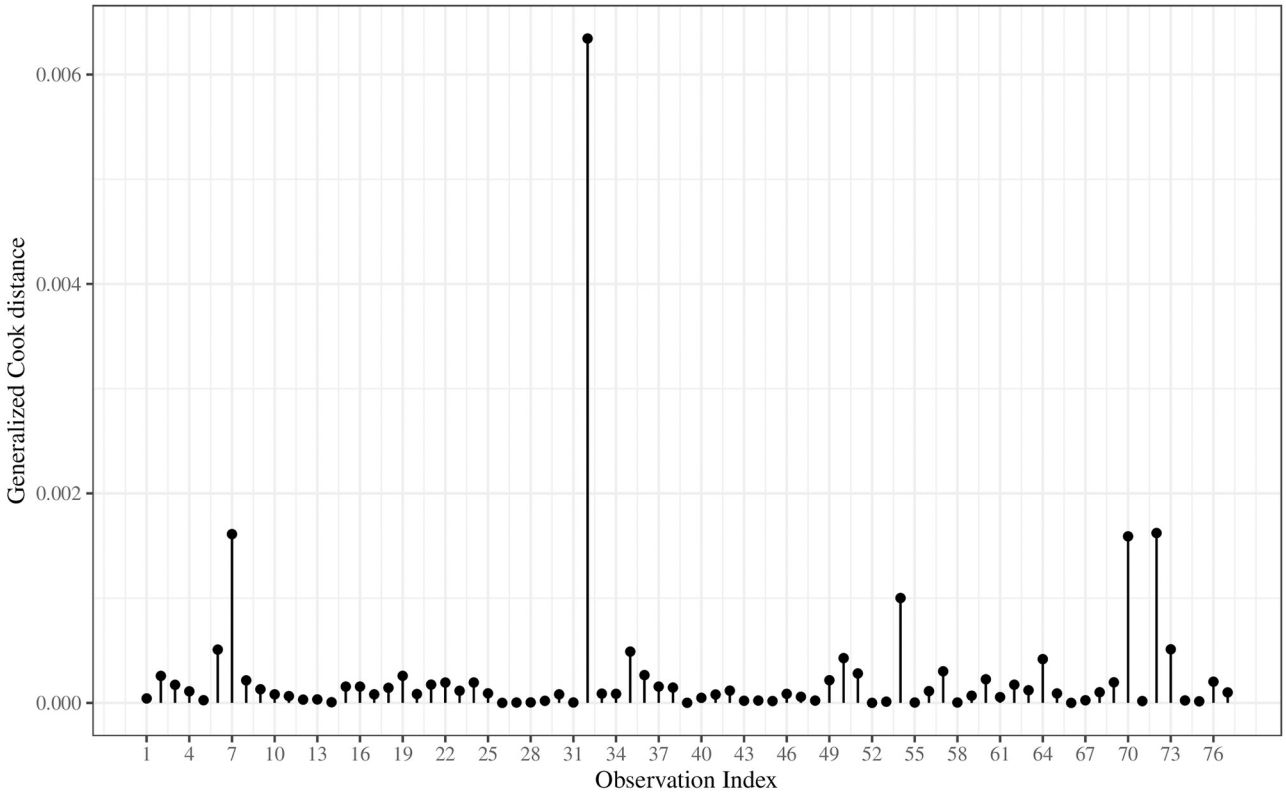

Fig 8. Generalized Cook distance for the UBXII regression.

https://doi.org/10.1371/journal.pone.0276695.g008

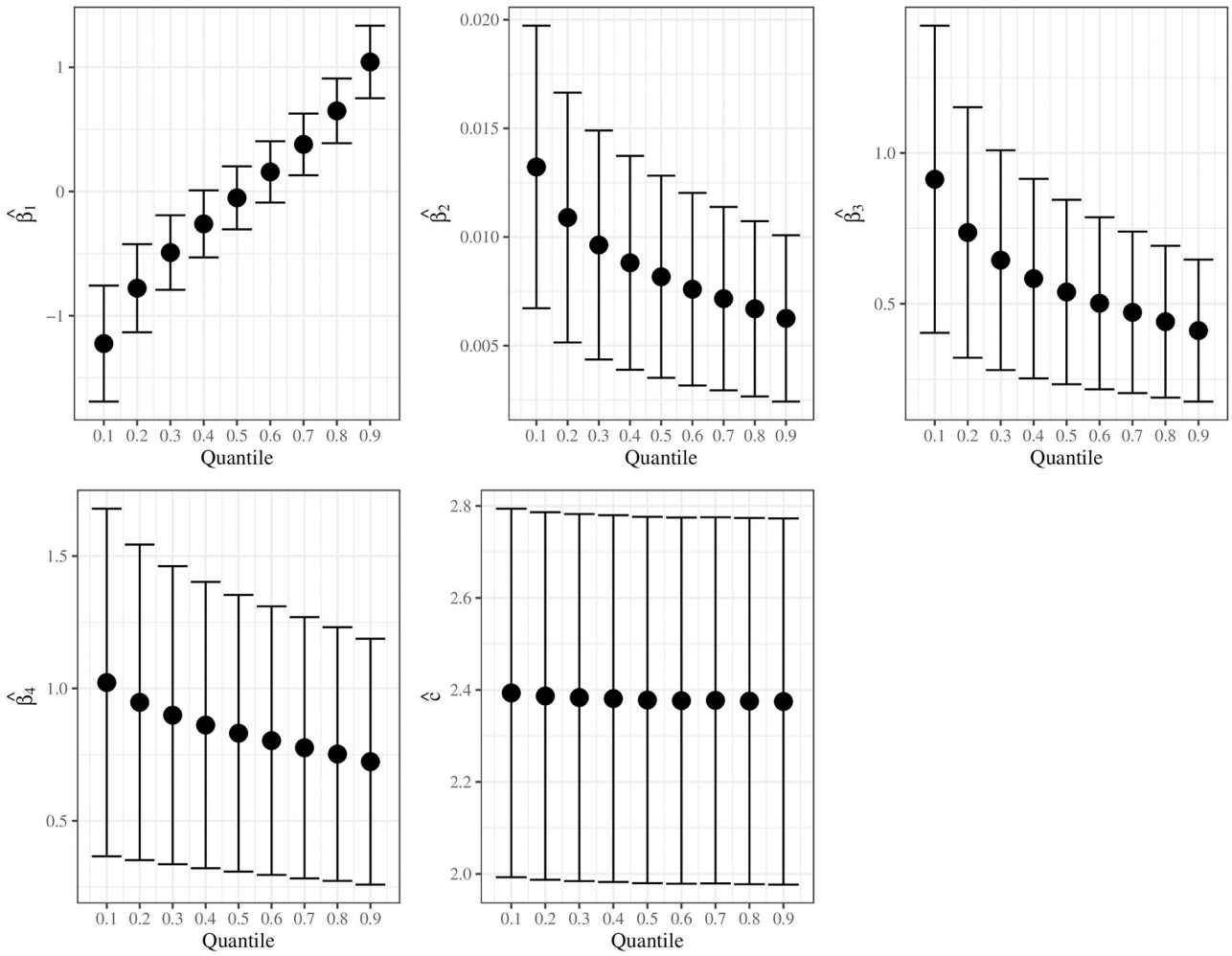

Fig 9. Parameter estimates and the 95% pointwise confidence intervals for the UBXII regression by considering $\tau \in \{0.1, 0.2, \ldots, 0.9\}$.

https://doi.org/10.1371/journal.pone.0276695.g009

7 Conclusions

We define a new unit quantile regression based on an alternative reparametrization for the unit Burr XII (UBXII) distribution pioneered by [8]. A highlight of the proposed parametrization is that one of its parameters, $q(\tau)$, represents the τth quantile of the random variable. The researcher defines the τ value and assumes a regression structure on $q(\tau)$. We investigated some additional statistical quantities to those explored by [8], namely the score functions, and observed information matrix. The maximum likelihood method is used for parameter estimation, and Monte Carlo simulations show that its properties remain. We adapt several diagnostic analysis and model selection techniques that can be employed to check the goodness-of-fit of the estimated model.

The utility of the proposed regression is illustrated with an application that targeted to explain the linear relation between the dropout proportion of Brazilian undergraduate animal sciences courses and some factors associated with their enrollment and organizational structure. An essential aspect of quantile-based regression is the possibility of separately analyzing the covariates' marginal effect on each response's quantile. That allowed us to find that the effects of some factors, such as the number of vacancies, accessibility, and night shift, are more

negligible on courses with fewer dropouts (those belonging to the lower quantiles). Another notable result is the positive effect between courses with night shifts and the dropout rate. This phenomenon is explained by the work carried out by the students during the morning shift, which makes persistence difficult. This situation must be considered by those who make educational policies since the opening of vacancies in the night shift must also be complemented by student attendance policies.

Additionally, we also fit the data set using other well-known regression models, such as the Kumaraswamy, unit-Weibull, and beta. The fit of the UBXII regression is superior to all of them since it provides better prediction performance. Thus, the UBXII regression is an alternative quite competitive for modeling data restricted to the unit interval and can be applied when the classical regressions are not unsuitable. That feature of capturing the nature of double-bonded variables makes the new model have a wide range of applications; for example, in the educational area, it can be helpful for modeling educational indicators such as graduation and persistence proportions of undergraduate and postgraduate courses. It may also be an alternative to educational measurements from different countries, such as in the applications [13–15], and [16] provided.

We end with some comments on possible future work. It should be noted that in conventional regression modeling, the non-existence of serial correlation between errors is assumed. In that sense, an extension of our proposal is the development of models that consider exogenous covariates in the median response with an Autoregressive Moving Average structure to handle serial dependence. It is important to highlight that the UBXII regression can be extended to the neutrosophic statistics analysis. This kind of analysis is applied when data or a part of it are indeterminate; that is, data have uncertain observations. Recently, some studies have been done in this context. [34] introduced the neutrosophic analysis of variance to test teaching methods using data collected from university students. [35] proposed a new Z-test for uncertainty events under neutrosophic statistics, which was applied to the Covid-19 data. In the regression context, [36] concluded that it is preferable to use Neutrosophic multiple regression over the classical regression models since this method is the most efficient for forecast the uncertainty observation data.

Supporting information

S1 File. Supplement to "The unit Burr XII regression: Properties, Simulation and application". It provides a detailed description of the data mining tools employed to obtain the final data set used in the application study in Section 7 and results from Kw, UW, and beta fitted regressions to the dropout proportion of Brazilian animal science courses.
(PDF)

S1 Data. Droupot proportion of Brazilian animal science courses data set. Data set used in the application study in Section 7.
(ODS)

S1 Appendix.
(PDF)

Author Contributions

Conceptualization: Tatiane Fontana Ribeiro.

Data curation: Tatiane Fontana Ribeiro, Fernando A. Peña-Ramírez, Renata Rojas Guerra.

Formal analysis: Tatiane Fontana Ribeiro, Fernando A. Peña-Ramírez.

Investigation: Tatiane Fontana Ribeiro.

Methodology: Tatiane Fontana Ribeiro, Fernando A. Peña-Ramírez.

Software: Tatiane Fontana Ribeiro, Fernando A. Peña-Ramírez.

Supervision: Fernando A. Peña-Ramírez.

Validation: Fernando A. Peña-Ramírez.

Visualization: Fernando A. Peña-Ramírez.

Writing – original draft: Tatiane Fontana Ribeiro, Fernando A. Peña-Ramírez.

Writing – review & editing: Fernando A. Peña-Ramírez, Renata Rojas Guerra, Gauss M. Cordeiro.

References

1. Rodríguez-Muñiz LJ, Bernardo AB, Esteban M, Díaz I. Dropout and transfer paths: What are the risky profiles when analyzing university persistence with machine learning techniques? Plos One. 2019; 14 (6):218–796. https://doi.org/10.1371/journal.pone.0218796 PMID: 31226158
2. Sneyers E, De Witte K. The interaction between dropout, graduation rates and quality ratings in universities. Journal of the Operational Research Society. 2017; 68(4):416–430. https://doi.org/10.1057/jors.2016.15
3. Srairi S. An Analysis of Factors Affecting Student Dropout: The Case of Tunisian Universities. International Journal of Educational Reform. 2022; 31(2):168–186. https://doi.org/10.1177/10567879211023123
4. Ferrari S, Cribari-Neto F. Beta regression for modelling rates and proportions. Journal of Applied Statistics. 2004; 31(7):799–815. https://doi.org/10.1080/0266476042000214501
5. 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. https://doi.org/10.1007/s00362-011-0417-y
6. Bayes CL, Bazán JL, De Castro M. A quantile parametric mixed regression model for bounded response variables. Statistics and its interface. 2017; 10(3):483–493. https://doi.org/10.4310/SII.2017.v10.n3.a11
7. 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. Journal of Applied Statistics. 2020; 47(6):954–974. https://doi.org/10.1080/02664763.2019.1657813 PMID: 35706917
8. Korkmaz MÇ, Chesneau C. On the unit Burr-XII distribution with the quantile regression modeling and applications. Computational and Applied Mathematics. 2021; 40(1):1–26. https://doi.org/10.1007/s40314-021-01418-5
9. Ribeiro TF, Cordeiro GM, Peña-Ramírez FA, Guerra RR. A new quantile regression for the COVID-19 mortality rates in the United States. Computational and Applied Mathematics. 2021; 40(7):1–16. https://doi.org/10.1007/s40314-021-01553-z
10. Korkmaz MÇ, Altun E, Alizadeh M, El-Morshedy M. The Log Exponential-Power Distribution: Properties, Estimations and Quantile Regression Model. Mathematics. 2021; 9(21):2634. https://doi.org/10.3390/math9212634
11. Korkmaz MÇ, Chesneau C, Korkmaz ZS. On the arcsecant hyperbolic normal distribution. Properties, quantile regression modeling and applications. Symmetry. 2021; 13(1):117. https://doi.org/10.3390/sym13010117
12. Mazucheli J, Alves B, Korkmaz MÇ, Leiva V. Vasicek Quantile and Mean Regression Models for Bounded Data: New Formulation, Mathematical Derivations, and Numerical Applications. Mathematics. 2022; 10(9):1389. https://doi.org/10.3390/math10091389
13. Korkmaz M, Chesneau C, Korkmaz ZS. Transmuted unit Rayleigh quantile regression model: Alternative to beta and Kumaraswamy quantile regression models. Univ Politeh Buchar Sci Bull Ser Appl Math Phys. 2021; 83:149–158.
14. Korkmaz MÇ, Chesneau C, Korkmaz ZS. A new alternative quantile regression model for the bounded response with educational measurements applications of OECD countries. Journal of Applied Statistics. 2021; p. 1–24. https://doi.org/10.1080/02664763.2021.2001442

15. Korkmaz MÇ, Chesneau C, Korkmaz ZS. The Unit Folded Normal Distribution: A New Unit Probability Distribution with the Estimation Procedures, Quantile Regression Modeling and Educational Attainment Applications. Journal of Reliability and Statistical Studies. 2022; p. 261–298.

16. Korkmaz MÇ, Korkmaz ZS. The unit log–log distribution: a new unit distribution with alternative quantile regression modeling and educational measurements applications. Journal of Applied Statistics. 2021; p. 1–20. https://doi.org/10.1080/02664763.2021.2001442

17. Saini S, Tomer S, Garg R. On the reliability estimation of multicomponent stress–strength model for Burr XII distribution using progressively first-failure censored samples. Journal of Statistical Computation and Simulation. 2022; 92(4):667–704. https://doi.org/10.1080/00949655.2021.1970165

18. Araújo FJMd, Guerra RR, Peña-Ramírez FA. The Burr XII quantile regression for salary-performance models with applications in the sports economy. Computational and Applied Mathematics;Accepted.

19. Guerra RR, Peña-Ramírez FA, Cordeiro GM. The Weibull Burr XII distribution in lifetime and income analysis. Anais da Academia Brasileira de Ciências. 2021; 93. PMID: 34105700

20. Bhatti FA, Hamedani GG, Korkmaz MÇ, Sheng W, Ali A. On the Burr XII-moment exponential distribution. Plos one. 2021; 16(2):e0246935. https://doi.org/10.1371/journal.pone.0246935 PMID: 33617564

21. Guerra RR, Peña-Ramírez FA, Bourguignon M. The unit extended Weibull families of distributions and its applications. Journal of Applied Statistics. 2021; 48(16):3174–3192. https://doi.org/10.1080/02664763.2020.1796936 PMID: 35707261

22. Ribeiro-Reis LD. Unit Log-Logistic Distribution and Unit Log-Logistic Regression Model. Journal of the Indian Society for Probability and Statistics. 2021; 22(2):375–388. https://doi.org/10.1007/s41096-021-00109-y

23. Peffer PAL. Demographics of an Undergraduate Animal Sciences Course and the Influence of Gender and Major on Course Performance. NACTA Journal. 2011; 55(1):26–31.

24. Lemonte AJ, Bazán JL. New class of Johnson distributions and its associated regression model for rates and proportions. Biometrical Journal. 2016; 58(4):727–746. https://doi.org/10.1002/bimj.201500030 PMID: 26659998

25. Mousa A, El-Sheikh A, Abdel-Fattah M. A gamma regression for bounded continuous variables. Advances and Applications in Statistics. 2016; 49(4):305–326. https://doi.org/10.17654/AS049040305

26. Lindsay BG, Li B. On second-order optimality of the observed Fisher information. The Annals of Statistics. 1997; 25(5):2172–2199. https://doi.org/10.1214/aos/1069362393

27. Pereira TL, Cribari-Neto F. Detecting model misspecification in inflated beta regressions. Communications in Statistics—Simulation and Computation. 2014; 43(3):631–656. https://doi.org/10.1080/03610918.2012.712183

28. Nagelkerke NJ, et al. A note on a general definition of the coefficient of determination. Biometrika. 1991; 78(3):691–692. https://doi.org/10.1093/biomet/78.3.691

29. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. vol. 112. Springer; 2013.

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

31. Stephens MA. EDF statistics for goodness of fit and some comparisons. Journal of the American Statistical Association. 1974; 69(347):730–737. https://doi.org/10.1080/01621459.1974.10480196

32. Costa FJd, Bispo MdS, Pereira RdCdF. Dropout and retention of undergraduate students in management: a study at a Brazilian Federal University. RAUSP Management Journal. 2018; 53:74–85. https://doi.org/10.1016/j.rauspm.2017.12.007

33. do Nascimento MdC, de Ribeiro Vieira PM, de Carvalho FMT, da Figueira MAS, Godoy GP. Perception of graduates about the quality of the night course in dentistry at a public institution in northeastern Brazil. Revista da ABENO. 2021; 21(1):1044–1044.

34. Aslam M. Neutrosophic analysis of variance: application to university students. Complex & intelligent systems. 2019; 5(4):403–407. https://doi.org/10.1007/s40747-019-0107-2

35. Aslam M. Design of a new Z-test for the uncertainty of Covid-19 events under Neutrosophic statistics. BMC Medical Research Methodology. 2022; 22(1):1–6. https://doi.org/10.1186/s12874-022-01593-x PMID: 35387604

36. Nagarajan D, Broumi S, Smarandache F, Kavikumar J. Analysis of neutrosophic multiple regression. Neutrosophic Sets and Systems. 2021; 43:44–53.