MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

The Pezeta regression model: an alternative to unit Lindley regression model

Abstract

A new probability distribution is proposed in this paper. This new distribution has support on the interval (0,1) and was obtained after transforming the random variable with exponential distribution. The mode, quantile function, median, ordinary moments and density function belongs to exponential family of distributions are demonstrated. The maximum likelihood method is used to obtain the parameter estimate. A regression model for the median of the distribution is also proposed. Closed-form expressions for the score vector and Fisher's information matrix are demonstrated. A simulation study and an application to real data showed the good performance of the proposed regression model.

Keywords: unit interval, exponential family, exponential distribution, mode, ordinary moments, regression model

Volume 12 Issue 4 - 2023

Lucas D Ribeiro Reis
Department of Economics, Federal University of Alagoas, Brazil

Correspondence: Lucas D Ribeiro Reis, Department of Economics, Federal University of Alagoas, Brazil,
Email econ.lucasdavid@gmail.com

Received: July 18, 2023 | **Published:** August 04, 2023

Introduction

The probability density function (pdf) of a random variable W with exponential distribution is given by

$$r(w;\lambda) = \lambda e^{-\lambda w}, w > 0,$$

where $\lambda > 0$ is scale parameter.

Taking $Y = 1/(1+W)$, the cdf and pdf of Y are

$$F(y;\lambda) = \exp\left[\lambda\left(1 - y^{-1}\right)\right], 0 < y < 1 \text{ and}$$

$$f(y;\lambda) = \frac{\lambda}{y^2}\exp\left[\lambda\left(1 - y^{-1}\right)\right], 0 < y < 1, \quad (1)$$

respectively.

Here, we will call the random variable with pdf (1) of Pezeta distribution, and denote this random variable as $Y \sim \text{pezeta}(\lambda)$. The Figure 1 shows some forms of the density function (1) for selected values of λ. This figure reveals that the peseta distribution is unimodal, and may also present positive (when λ approaches 0) and negative (when λ moves away from 0) asymmetry.

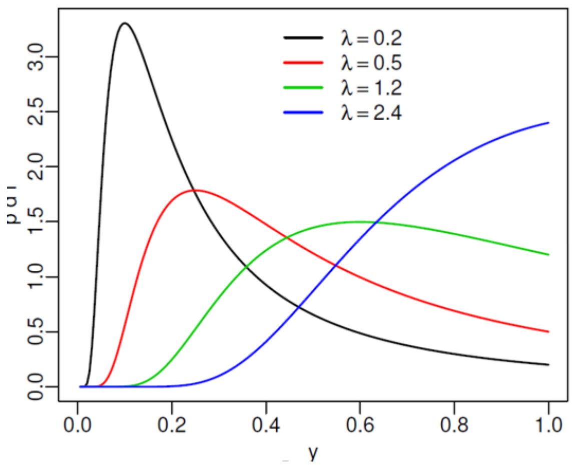


Figure 1 Some forms of the pdf (1), for special cases.

The first derivative of the log-pdf is

$$\zeta(y) = \frac{d}{dy}\ln\left(f(y;\lambda)\right) = -\frac{2}{y} + \frac{\lambda}{y^2}.$$

Solving $\zeta(y) = 0$, the mode of Y is

$$\text{mode}(Y) = \begin{cases} \frac{\lambda}{2}, & \lambda < 2, \\ 1, & \lambda \ge 2. \end{cases}$$

The r th ordinary moment of Y is

$$\mathbb{E}\left(Y^r\right) = \int_0^1 \lambda y^{r-2}\exp\left[\lambda\left(1 - y^{-1}\right)\right]dy$$

$$= \lambda e^{\lambda} E_r(\lambda),$$

where $E_n(x) = \int_1^{\infty} z^{-n} e^{-xz} dz$ denotes the exponential integral function.[1]

By inverting $F(y;\lambda) = p$, the quantile function is given by

$$Q(p;\lambda) = \left[1 - \lambda^{-1}\ln(p)\right]^{-1}, 0 < p < 1.$$

The median is obtained when $p = 0.5$. So, the median of Y is

$$median(Y) = \left[1 - \lambda^{-1}\ln(0.5)\right]^{-1}.$$

Using the quantile function, the random variable $Y = \left[1 - \lambda^{-1}\ln(V)\right]^{-1}$ has density function (1), where V is a uniform random variable over the interval (0,1).

The paper is structured as follows. In Section 4, it is shown that the distribution belongs to the exponential distribution family. The mean and variance of the sufficient statistic are also presented. In Section 5, the maximum likelihood method to obtain the parameter estimate is presented. Analytical expressions for the bias correction of the maximum likelihood estimator are also presented. In Section 6, a new regression model is introduced. In Sections 7 and 8, numerical and empirical results are presented, respectively. Finally, Section 9 concludes the paper.

Exponential family

Let the random variable Y with pdf $f(y;\theta)$, in which θ is the parameter that indexes the distribution. This random variable belongs to the exponential family if its pdf can be written as

$$f(y;\theta) = h(y)\exp\left[\eta(\theta)t(y) - b(\theta)\right], \quad (2)$$

where the functions $\eta(\theta)$, $b(\theta)$, $t(y)$ and $h(y)$ assume values in subsets of the reals.

Note that, the pdf (1) can be written as

$$f(y;\lambda)=\frac{1}{y^2}\exp\left[\lambda\left(1-y^{-1}\right)-\left(-\ln(\lambda)\right)\right].$$

that belongs to exponential family (2), where $\eta(\lambda)=\lambda$, $t(y)=1-y^{-1}$, $b(\lambda)=-\ln(\lambda)$ and $h(y)=1/y^2$. Thus, by the factorization criterion $t(y)$ is sufficient statistics for λ. The fact that Y belongs to exponential family, the mean and variance of $t(Y)$ are given by $\mathbb{E}\left[t(Y)\right]=-\frac{1}{\lambda}$ and $\mathbb{V}\left[t(Y)\right]=\frac{1}{\lambda^2}$, respectively.

Maximum likelihood estimation

For a random sample of size n of the random variable Y with density function (1), the log-likelihood function for λ is given by

$$\mathcal{L}_0(\lambda)=n\ln(\lambda)+\lambda\sum_{i=1}^{n}\left(1-y_i^{-1}\right)-2\sum_{i=1}^{n}\ln(y_i).$$

The maximum likelihood estimator (MLE) of λ is the solution of

$$\frac{\partial\mathcal{L}_0(\lambda)}{\partial\lambda}=\frac{n}{\lambda}+\sum_{i=1}^{n}\left(1-y_i^{-1}\right)=0.$$

So, the MLE of λ is

$$\hat{\lambda}=-\frac{n}{\sum_{i=1}^{n}\left(1-y_i^{-1}\right)}.$$

The second derivative of $\mathcal{L}_0(\lambda)$ is given as

$$\frac{\partial^2\mathcal{L}_0(\lambda)}{\partial\lambda^2}=-\frac{n}{\lambda^2}<0,$$

showing that $\hat{\lambda}$ really is the point that a maximizes the function $\mathcal{L}_0(\lambda)$. It can be further shown that the variance and standard error of $\hat{\lambda}$ are expressed as $\mathbb{V}(\hat{\lambda})=\hat{\lambda}^2/n$ and $\mathrm{se}(\hat{\lambda})=\hat{\lambda}/\sqrt{n}$, respectively.

MLE bias correction

Generally, when n is small, the MLEs tends to be biased. Here, a bias correction of the MLE of the parameter that indexes the Pezeta distribution will be presented. Here, the bias of $\hat{\lambda}$ can be expressed[2] as

$$B(\hat{\lambda})=\mathbb{V}(\lambda)^2\left(\frac{1}{2}\kappa_{\lambda\lambda\lambda}+\kappa_{\lambda\lambda,\lambda}\right),$$

where $\kappa_{\lambda\lambda\lambda}=\mathbb{E}\left[\frac{d^3\mathcal{L}_0(\lambda)}{d\lambda^3}\right]$ and $\kappa_{\lambda\lambda,\lambda}=\mathbb{E}\left[\frac{d^2\mathcal{L}_0(\lambda)}{d\lambda^2}\frac{d\mathcal{L}_0(\lambda)}{d\lambda}\right]$.

Note that $\kappa_{\lambda\lambda\lambda}=2n/\lambda^3$ and

$$\frac{\partial^2\mathcal{L}_0(\lambda)}{\partial\lambda^2}\frac{\partial\mathcal{L}_0(\lambda)}{\partial\lambda}=-\frac{n^2}{\lambda^3}-\frac{n}{\lambda^2}\sum_{i=1}^{n}\left(1-y_i^{-1}\right).$$

From Section 4, follows that

$$\sum_{i=1}^{n}\mathbb{E}\left[\left(1-y_i^{-1}\right)\right]=\sum_{i=1}^{n}\mathbb{E}\left[t(y_i)\right]=-\frac{n}{\lambda},$$

resulting in

$$\begin{aligned}\kappa_{\lambda\lambda,\lambda} &= -\frac{n^2}{\lambda^3}-\frac{n}{\lambda^2}\mathbb{E}\left[\sum_{i=1}^{n}\left(1-y_i^{-1}\right)\right]\\ &= -\frac{n^2}{\lambda^3}-\frac{n}{\lambda^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(1-y_i^{-1}\right)\right].\\ &= -\frac{n^2}{\lambda^3}+\frac{n^2}{\lambda^3}\\ &= 0\end{aligned}$$

Thus, the bias of $\hat{\lambda}$ is

$$B(\hat{\lambda})=\left(\frac{\lambda^2}{n}\right)^2\left(\frac{2n}{2\lambda^3}\right)=\frac{\lambda}{n}.$$

Finally, it follows that the bias-corrected MLE $\hat{\lambda}$ is given by

$$\hat{\lambda}_{BC}=\hat{\lambda}-B(\hat{\lambda})=\hat{\lambda}\left(1-\frac{1}{n}\right).$$

The Pezeta regression model

Starting from the Pezeta distribution, in this section a new regression model will be introduced for the dependent variable with support at (0,1). This model has a regression structure on the median of the distribution. Thus, in the presence of outliers in the data, this new regression model has an advantage over regression models with a mean structure.

By taking $\mathrm{median}(Y)=\tau$ and isolating for λ, results in

$$\lambda=\frac{\ln(0.5)}{1-\tau^{-1}}.$$

Under this parameterization, the density function (1) becomes

$$f(y;\tau)=\frac{\ln(0.5)}{y^2\left(1-\tau^{-1}\right)}\exp\left[\frac{\ln(0.5)}{1-\tau^{-1}}\left(1-y^{-1}\right)\right],0<y<1, \quad (3)$$

and the corresponding cdf and quantile function are given by

$$F(y;\tau)=\exp\left[\frac{\ln(0.5)}{1-\tau^{-1}}\left(1-y^{-1}\right)\right],0<y<1 \quad (4)$$

and

$$Q(p;\tau)=\left[1-\frac{1-\tau^{-1}}{\ln(0.5)}\ln(p)\right]^{-1},0<p<1,$$

respectively, where $0<\tau<1$ denotes the median of Y.

The random variable Y with pdf (3) is denoted as $Y\sim\mathrm{pezeta}(\tau)$. Some plots of the pdf (3) are shown in Figure 2. These plots reveal that the pdf can be asymmetric to the left and asymmetric to the right.

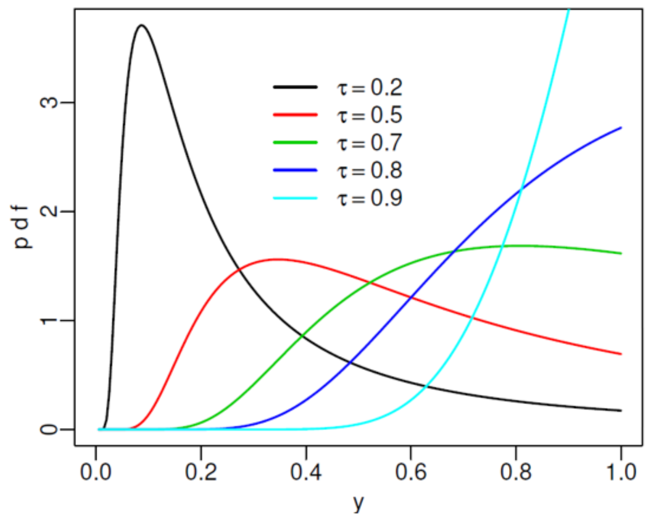


Figure 2 Some forms of the pdf (3), for special cases.

Here, the regression model for the median has the following regression structure

$$\eta_i=g(\tau_i)=X_i^T\beta.$$

Citation: Reis LDR. The Pezeta regression model: an alternative to unit Lindley regression model. *Biom Biostat Int J*. 2023;12(4):107‒112. DOI: 10.15406/bbij.2023.12.00393

where $\beta=\left(\beta_1,\cdots,\beta_k\right)^\top$ is k-vector of unknown parameters, $x_i=(x_{i1},\cdots,x_{ik})^\top$ is vector of k explanatory variables $(k<n)$, which are assumed fixed and known and η_i is the linear predictor. For model with intercept, it is assumed that $x_{i1}=1,\ \forall i$. The $g(\cdot)$ is a link function strictly monotonic and twice differentiable, such that $g:(0,1)\to\mathbb{R}$. Examples of some link functions can be: (i) standard logistic quantile function $g(\tau)=\ln\left[\tau/(1-\tau)\right]$; and (ii) standard Cauchy quantile function $g(\tau)=\tan\left(\pi(\tau-0.5)\right)$.

From Equation (3) the log-likelihood function for a random sample of size n is given by

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

where

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

Differentiating $\mathcal{L}_i(\tau_i)$ with respect to τ_i

$$\frac{\partial\mathcal{L}_i(\tau_i)}{\partial\tau_i}=\frac{1}{\tau_i-\tau_i^2}-\frac{\ln(0.5)}{\tau_i^2\left(1-\tau_i^{-1}\right)^2}\left(1-y_i^{-1}\right)$$

$$=a_i\left(1+\dot{y}_i\right), \tag{5}$$

where $a_i=1/\left(\tau_i-\tau_i^2\right)$ and $\dot{y}_i=\ln(0.5)\left(1-y_i^{-1}\right)/\left(1-\tau_i^{-1}\right)$. Since that $\mathbb{E}\left[\partial\mathcal{L}(\tau_i)/\partial\tau_i\right]=0$, then $\mathbb{E}\left[\dot{y}_i\right]=-1,\ \forall i$.

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

$$\frac{\partial\mathcal{L}(\beta)}{\partial\beta_j}=\sum_{i=1}^{n}\frac{\partial\mathcal{L}_i(\tau_i)}{\partial\tau_i}\frac{\mathrm{d}\tau_i}{\mathrm{d}\eta_i}\frac{\partial\eta_i}{\partial\beta_j}. \tag{6}$$

Note that, $\mathrm{d}\tau_i/\mathrm{d}\eta_i=1/g'(\tau_i)$ and $\partial\eta_i/\partial\beta_j=x_{ij}$, then the score vector of β_j is given by

$$\frac{\partial\mathcal{L}(\beta)}{\partial\beta_j}=\ \sum_{i=1}^{n}\frac{a_i\left(1+\dot{y}_i\right)}{g'(\tau_i)}x_{ij}.$$

The score vector in matrix form is $U(\beta)=X^\top Gv$, where X is a $n\times k$ matrix whose ith row is $x_i^\top, G=diag\{1/g'(\tau_1),\cdots 1/g'(\tau_n)\}$ (diagonal matrix) and $v=\left(a_1\left(1+\dot{y}_1\right),\cdots,a_n\left(1+\dot{y}_n\right)\right)^\top$.

The MLE of β, say $\hat{\beta}$, is the solution of $U(\beta)=0$. There is no analytical solution for this nonlinear system, and so the MLE of β must be obtained numerically, from iterative methods. However, these iterative methods require initial guesses for parameter values. As in Ribeiro-Reis,[3] the initial guess for $\hat{\beta}$ will be the ordinary least squares estimator of the regression $g(y)$ on X, which is $\hat{\beta}^{(0)}=\left(X^\top X\right)^{-1}X^\top g(y)$.

From Equation (6), the second derivative of $\mathcal{L}(\beta)$ with respect to β_l is

$$\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_j\partial\beta_l}=\sum_{i=1}^{n}\frac{\partial}{\partial\tau_i}\left(\frac{\partial\mathcal{L}_i(\tau_i)}{\partial\tau_i}\frac{1}{g'(\tau_i)}x_{ij}\right)\frac{\mathrm{d}\tau_i}{\mathrm{d}\eta_i}\frac{\partial\eta_i}{\partial\beta_l}$$

$$=\sum_{i=1}^{n}\left[\frac{\partial^2\mathcal{L}_i(\tau_i)}{\partial\tau_i^2}\frac{1}{g'(\tau_i)}x_{ij}+\frac{\partial\mathcal{L}_i(\tau_i)}{\partial\tau_i}\frac{\partial}{\partial\tau_i}\left(\frac{1}{g'(\tau_i)}\right)x_{ij}\right]\frac{1}{g'(\tau_i)}x_{il}.$$

Once that $\mathbb{E}\left[\partial\mathcal{L}_i(\tau_i)/\partial\tau_i\right]=0$, then

$$\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_j\partial\beta_l}\right]=\sum_{i=1}^{n}\mathbb{E}\left[\frac{\partial^2\mathcal{L}_i(\tau_i)}{\partial\tau_i^2}\right]\frac{1}{g'(\tau_i)^2}x_{ij}.$$

From Equation (5), follows that

$$\frac{\partial^2\mathcal{L}_i(\tau_i)}{\partial\tau_i^2}=\ a_i'\left(1+\dot{y}_i\right)+a_i\dot{y}_i',$$

where $a_i'=\partial a_i/\partial\tau_i$ and $\dot{y}_i'=\partial\dot{y}_i/\partial\tau_i$.

Since that $\mathbb{E}\left[\dot{y}_i\right]=-1$, then the expected value is

$$\mathbb{E}\left[\frac{\partial^2\mathcal{L}_i(\tau_i)}{\partial\tau_i^2}\right]=\ a_i'\left(1+\mathbb{E}\left[\dot{y}_i\right]\right)+a_i\mathbb{E}\left[\dot{y}_i'\right]=a_i\mathbb{E}\left[\dot{y}_i'\right].$$

We still have to

$$\dot{y}_i'=\ -\frac{\ln(0.5)}{\tau_i^2\left(1-\tau_i^{-1}\right)^2}\left(1-y_i^{-1}\right)=a_i\,\dot{y}_i,$$

resulting in $\mathbb{E}\left[\dot{y}_i'\right]=a_i\,\mathbb{E}\left[\dot{y}_i\right]=-a_i$ and hence $\mathbb{E}\left[\frac{\partial^2\mathcal{L}_i(\tau_i)}{\partial\tau_i^2}\right]=-a_i^2$.

Finally,

$$\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_j\partial\beta_l}\right]=-\sum_{i=1}^{n}\frac{a_i^2}{g'(\tau_i)^2}x_{ij}.$$

Let $P=\operatorname{diag}\left\{a_1^2,\cdots,a_n^2\right\}$, the expression in matrix form is

$$\mathbb{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_j\partial\beta_l}\right]=-X^\top PG^2X.$$

So, the Fisher expected information matrix is

$$\mathcal{K}(\beta)=X^\top PG^2X.$$

Under the usual regularity conditions for MLEs, when the sample size is large,

$$\hat{\beta}\overset{a}{\sim}\mathcal{N}_k\left(\beta,\mathcal{K}(\beta)^{-1}\right),$$

where $\overset{a}{\sim}$ denotes asymptotic distribution. So, confidence intervals and hypothesis testing can be performed using the normal distribution. Based on asymptotic distribution, the $100(1-\alpha)\%$ confidence intervals for β_j is given by

$$\hat{\beta}_j\pm z_{(1-\alpha/2)}\sqrt{\theta_{jj}},\ j=1,\cdots k,$$

where $z_{(1-\alpha/2)}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution and θ_{jj} denotes the jth diagonal element of the matrix $\mathcal{K}(\beta)^{-1}$.

Residuals

Residual analysis is a good indicator to tell if an estimated model is well-adjusted.[3] If the residuals do not show an adequate behavior, then the estimated model is poor. Here, the Dunn-Smyth[4] residuals will be addressed. The Dunn-Smyth residuals are defined as

$$\hat{r}_i=Q_N\left(F\left(y_i;\hat{\tau}_i\right)\right),$$

in which $Q_N(\cdot)$ denotes the quantile function of the standard normal distribution and $F\left(y_i;\hat{\tau}_i\right)$ is the cdf (4) evaluated in $\hat{\tau}$. If the model is well estimated, then the Dunn-Smyth residuals are expected to have a random behavior around zero, with approximately 95% of the values falling within the range $(-2,2)$.[5,6]

Simulation

To show the performance of the MLEs for the proposed regression model, a numerical study using Monte Carlo simulations, with 10000 repetitions, is performed. The simulated regression model is given by

$$\ln\left(\frac{\tau_i}{1-\tau_i}\right)=\beta_1+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4x_{i3},$$

in which all explanatory variables x's are generated from the standard normal distribution. Three sample sizes $n=\{50,100,300\}$ are considered, with the true values of the parameters being: $\beta_1=1.7$, $\beta_2=-2.4$, $\beta_3=0.9$ and $\beta_4=4.2$.

The performance measures analyzed in the simulations will be based on the average estimates (AEs), mean squared errors (MSEs) and the 95% coverage rates (CRs) for the parameters. The simulations were done in the matrix programming language Ox Console.[7]

The simulation results are shown in Table 1. As can be seen, as the sample size increases, the MLEs and CRs converge to their true values, and the MSEs decrease. Thus, we can see the good performance of the estimates for the regression model introduced here.

Table 1 Simulations results

n	Parameter	AE	MSE	CR (95%)
50	β_1	1.740808	0.022806	93.79
	β_2	− 2.401397	0.039183	93.51
	β_3	0.900474	0.018413	93.35
	β_4	4.19932	0.041534	94.14
150	β_1	1.713127	0.006826	94.91
	β_2	− 2.402275	0.008375	94.59
	β_3	0.900578	0.005515	94.65
	β_4	4.20142	0.007643	94.55
300	β_1	1.707051	0.003405	94.97
	β_2	− 2.400968	0.003944	94.90
	β_3	0.90126	0.003165	94.93
	β_4	4.200144	0.00353	94.64

Application

The Pezeta regression model is compared with the unit-Lindley (UL) regression model, which was introduced by Mazucheli et al.[8] The density function of the UL model is given by $f_{UL}(y;\tau)=\frac{(1-\tau)^2}{\tau(1-y)^3}\exp\left\{-\frac{y(1-\tau)}{\tau(1-y)}\right\}$, $0<y<1$, where $0<\tau<1$ denotes the mean of the distribution.

The data used here were analyzed by Smithson & Verkuilen.[9] The response variable (y) is the accuracy that presents scores on a test of reading accuracy taken by 44 children in Australian. The explanatory variables are dyslexia (x_2) and nonverbal intelligence quotient x_3. The variable x_2 is a categorical variable that takes value 1 if the child has dyslexia and value 0 if the child does not have dyslexia. The variable x_3 is converted into z scores. These data are available in the *betareg* package.[10]

The fitted model is given by

$$\ln\left(\frac{\tau_i}{1-\tau_i}\right)=\beta_1+\beta_2x_{i2}+\beta_3x_{i3}+\beta_4\left(x_{i2}\times x_{i3}\right), i=1,2,\ldots,44,$$

where τ_i refers to the median for the Pezeta regression model and to the mean for the UL regression model.

To discriminate between the two regression models, the usual statistics were used: Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Hannan–Quinn Information Criterion (HQIC). The model that presents the smallest values of these statistics is chosen as a superior model for the data in question. The formulas for the AIC, BIC and HQIC statistics can be consulted at Ribeiro-Reis.[6]

All calculations in this application were made using the language Ox Console.[7] The results of the estimates for the Pezeta and UL regression models are shown in Table 2. Note that the two models share the same sign for the parameter estimates. It is also noticed that all estimates of the coefficients for the Pezeta regression model are highly significant. In turn, in the UL regression model the β_4 estimate was not statistically significant.

Table 2 Summary estimates for Pezeta and UL regression models

Parameter	Estimate	Std error	z -value	p -value
		Pezeta		
β_1	1.98376	0.235234	8.433134	0.000000
β_2	− 1.248062	0.376479	− 3.315087	0.000916
β_3	1.204827	0.249365	4.831581	0.000001
β_4	− 1.256459	0.375852	− 3.34296	0.000829
		UL		
β_1	3.18122	0.166414	19.11631	0.000000
β_2	− 3.169079	0.27772	− 11.41107	0.000000
β_3	0.293898	0.176392	1.666164	0.095681
β_4	− 0.358143	0.275959	− 1.297811	0.194352

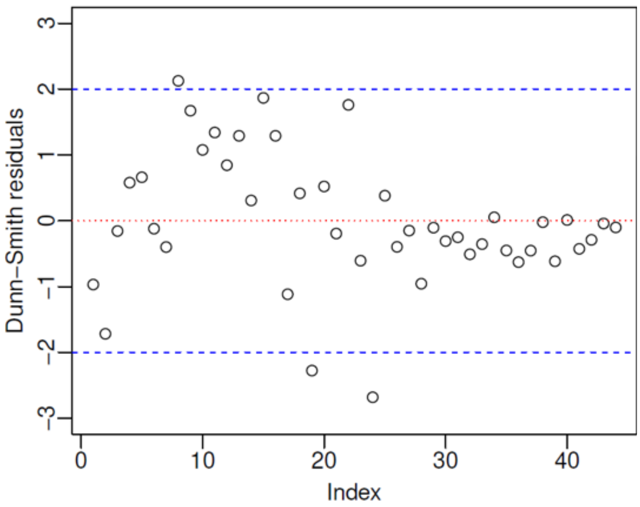
The statistics for the choices of the two models are in Table 3. It is noted that all three statistics have their lowest values for the Pezeta regression model, indicating that this model is more appropriate for the data in question.

Table 3 Information criteria

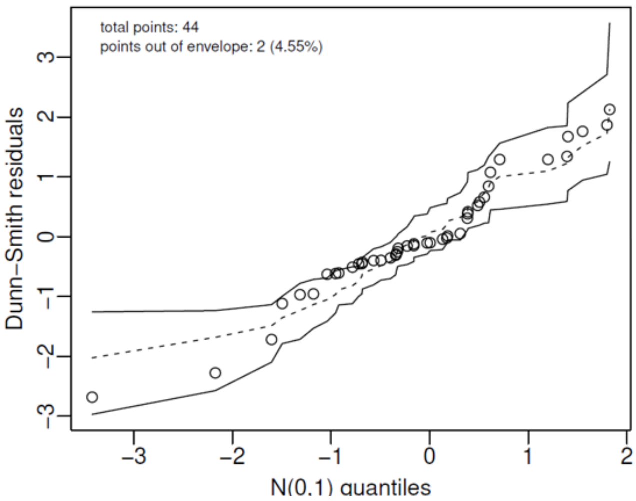
Model	AIC	BIC	HQIC
Pezeta	− 80.1893	− 73.0526	− 77.5427
UL	− 76.5169	− 69.3802	− 73.8703

The Dunn-Smyth residuals, with their respective simulated envelopes, for the Pezeta and UL regression models are shown in Figures 3 & 4, respectively. It is verified that the residuals for the Pezeta model presents a more random behavior around zero, than the UL model. The simulated envelope corroborates this, since in the Pezeta model there are only 4.55% of the observations outside the simulated envelope. In contrast, in the UL model, the number of observations outside the simulated envelope is 72.73%, indicating the poor fit of the UL model.

Citation: Reis LDR. The Pezeta regression model: an alternative to unit Lindley regression model. *Biom Biostat Int J.* 2023;12(4):107–112. DOI: 10.15406/bbij.2023.12.00393

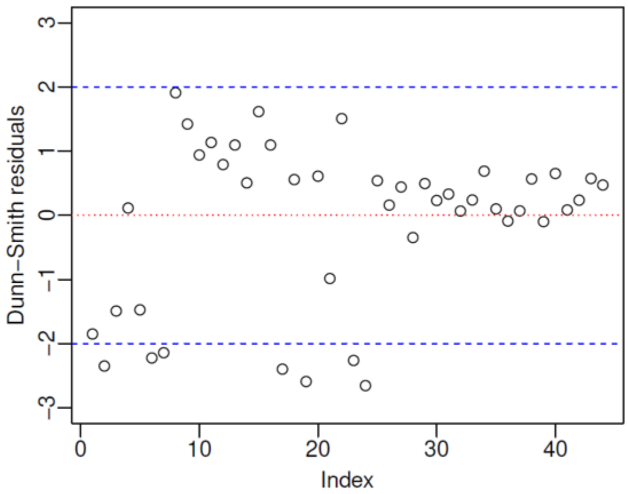


(a) residuals *versus* index

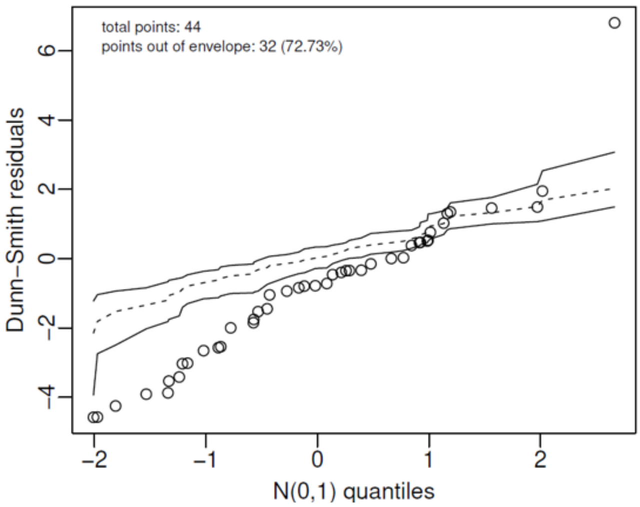


(b) simulated envelope

Figure 3 Dunn-Smyth residuals for Pezeta regression model.



(a) residuals *versus* index



(b) simulated envelope

Figure 4 Dunn-Smyth residuals for unit Lindley regression model.

Conclusion

In this paper, a new probability distribution with support on the interval (0,1) was proposed. This new distribution is obtained through a transformation of the random variable with exponential distribution. Several properties were discussed, such as mode, ordinary moments, quantile function, random number generation, exponential family and maximum likelihood estimation (with and without bias correction).

Subsequently, a regression model for the dependent variable in the unit interval was introduced. The regression is structured on the median of the distribution, which means that, in the presence of outliers in the data, the proposed regression model is more robust than the regression models with structure on the mean. The maximum likelihood method is considered for parameter estimation. Analytical expressions are obtained for the score vector and for the Fisher information matrix. Fisher's information matrix is very important to obtain the standard errors of the estimated coefficients.

A simulation study on finite samples showed that the maximum likelihood estimators are consistent, indicating that as the sample size increases, the estimators converge to their true parameters. An application to real data is also made, to show the usefulness of the model in practice. The proposed regression model is compared with the unit Lindley regression model. The results showed that the regression model proposed in this paper is superior to the unit Lindley regression model.

Suggestions for future research can be: (i) bias correction for the estimated coefficients of the regression model; (ii) introduce the version of the regression model for time series.

Acknowledgments

None.

Conflicts of interest

The author declare that there is no conflicts of interest.

Funding

None.

References

1. Abramowitz M, Stegun IA. H*andbook of mathematical functions with formulas, graphs and mathematical tables.* New York: Dover; 1965.
2. Cox DR, Snell EJ. A general definition of residuals. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1968;30(2):248–265.
3. 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.
4. Dunn PK, Smyth GK. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*. 1996;5(3):236–244.
5. Altun E. The log-weighted exponential regression model: alternative to the beta regression model. *Communications in Statistics-Theory and Methods*. 2021;50(10):2306–2321.
6. Ribeiro-Reis LD. The unit inverse Weibull distribution and its associated regression model. *Journal of Econometrics and Statistics*. 2023;3(1):55–68.
7. Doornik JA. *Ox: an object-oriented matrix programming language*. United Kingdom: Timberlake Consultants and Oxford; 2018.
8. Mazucheli J, Menezes AFB, Chakraborty S. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. *Journal of Applied Statistics*. 2019;46(4):700–714.
9. Smithson M, Verkuilen J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*. 2006;11(1):54.
10. Cribari Neto F, Zeileis A. Beta regression in R. *Journal of statistical software*. 2010;34:1–24.

Citation: Reis LDR. The Pezeta regression model: an alternative to unit Lindley regression model. *Biom Biostat Int J*. 2023;12(4):107–112.
DOI: 10.15406/bbij.2023.12.00393