



On the one parameter unit-Lindley distribution and its associated regression model for proportion data

Josmar Mazucheli, André Felipe Berdusco Menezes & Subrata Chakraborty

To cite this article: Josmar Mazucheli, André Felipe Berdusco Menezes & Subrata Chakraborty (2018): On the one parameter unit-Lindley distribution and its associated regression model for proportion data, Journal of Applied Statistics, DOI: [10.1080/02664763.2018.1511774](https://doi.org/10.1080/02664763.2018.1511774)

To link to this article: <https://doi.org/10.1080/02664763.2018.1511774>



Published online: 23 Aug 2018.



Submit your article to this journal [↗](#)




Article views: 5



View Crossmark data [↗](#)



On the one parameter unit-Lindley distribution and its associated regression model for proportion data

Josmar Mazucheli ^a, André Felipe Berdusco Menezes^a and Subrata Chakraborty^b

^aDepartment of Statistics, Universidade Estadual de Maringá, DEs, PR, Brazil; ^bDepartment of Statistics, Dibrugarh University, Assam, India

ABSTRACT

In this paper considering an appropriate transformation on the Lindley distribution, we propose the unit-Lindley distribution and investigate some of its statistical properties. An important fact associated with this new distribution is that it is possible to obtain the analytical expression for bias correction of the maximum likelihood estimator. Moreover, it belongs to the exponential family. This distribution allows us to incorporate covariates directly in the mean and consequently to quantify their influences on the average of the response variable. Finally, a practical application is presented to show that our model fits much better than the Beta regression.

ARTICLE HISTORY

Received 20 January 2018
Accepted 9 August 2018

KEYWORDS

Lindley distribution;
proportion data; maximum
likelihood estimation;
regression model

1. Introduction

In applied statistics, a common issue is to deal with the uncertainty phenomena observed in the bounded interval $(0, 1)$. Very often in real life we encounter measures like proportion or fraction of a certain characteristic, scores of some ability tests, different indices and rates, which lie in the interval $(0, 1)$ (see, for instance, [3,6,10,11,14,22,24], among others studies). In such cases continuous distributions with domain $(0, 1)$ are indispensable to probabilistic modeling of the phenomena. The two parameter Beta distribution (or the Pearson type IV distribution) is the most widely used model for such data in practice, mainly because its flexibility [13]. Though many distributions were proposed and studied as alternatives there is still no agreement on preference of a particular model.

In this paper we introduce a one parameter unit-Lindley distribution, [19], derived from a transformation on the Lindley distribution. As far as we know the only other one-parameter distribution in the unit interval is the Topp–Leone distribution [25]. Nevertheless, the Topp–Leone distribution does not possess important properties such as closed form expressions for the moments.

The main advantage of the unit-Lindley distribution lies on the fact that practitioners will have a new quite flexible, unimodal one-parameter distribution which possesses several vital properties that other distributions restricted to the interval $(0, 1)$ do not enjoy. For instance, the unit-Lindley distribution has only a single parameter and closed form expressions for cumulative distribution function (c.d.f), quantile function and simple expression

for moments unlike the well known Beta distribution (having two parameters, no closed form for c.d.f. and quantile function) and Kumaraswamy distribution (with two parameters, no closed form for moments). Moreover because of its simple formula for mean the unit-Lindley distribution allows us to directly incorporate the covariates in the mean in order to quantify their average influence on the response variable. This enabled us to present a new bounded regression model as a viable alternative to the widely used Beta regression model [2,8].

We provide a comprehensive account of statistical properties of the proposed distribution along with an application with data from the access of people in households with inadequate water supply and sewage in the cities of Brazil from the Southeast and Northeast regions, to demonstrate that the unit-Lindley regression yields a better fit than the Beta regression model.

The rest of this paper is structured as follows. In Section 2, we start with the model formulation and investigate several features such as moments, incomplete moments, behavior of the cumulative and probability density functions, Lorenz curve and quantile function. Parameter estimation by two different methods are discussed in Section 3. A simulation study to investigate the performance of the proposed estimators is presented in Section 4. A real life application related to the proportion of people with inadequate water supply and sewage is analyzed in Section 5. We conclude with some discussion in Section 6.

2. The unit-Lindley distribution

The Lindley distribution was introduced by Lindley [18] in the context of Bayesian inference. Its probability density function (p.d.f) is specified by

$$f(y | \theta) = \frac{\theta^2}{1 + \theta} (1 + y) \exp(-\theta y), \quad y > 0, \theta > 0.$$

The corresponding c.d.f. is

$$F(y | \theta) = 1 - \left(1 + \frac{\theta y}{1 + \theta}\right) \exp(-\theta y). \quad (1)$$

Ghitany *et al.* [9] studied the Lindley distribution and outlined that its mathematical properties are more flexible than those of the exponential distribution.

From (1) using the transformation $X = Y/(1 + Y)$ we propose a new distribution with support on the unit-interval. The c.d.f. and the p.d.f. of the resulting distribution are given, respectively, by

$$F(x | \theta) = 1 - \left(1 - \frac{\theta x}{(1 + \theta)(x - 1)}\right) \exp\left(-\frac{\theta x}{1 - x}\right), \quad 0 < x < 1, \quad \theta > 0. \quad (2)$$

$$f(x | \theta) = \frac{\theta^2}{1 + \theta} (1 - x)^{-3} \exp\left(-\frac{\theta x}{1 - x}\right), \quad 0 < x < 1, \quad \theta > 0. \quad (3)$$

The first derivative of $f(x | \theta)$ is

$$\frac{d}{dx} f(x | \theta) = \frac{\theta^2(\theta + 3x - 3)}{(1 + \theta)(x - 1)^5} \exp\left(-\frac{\theta x}{1 - x}\right)$$

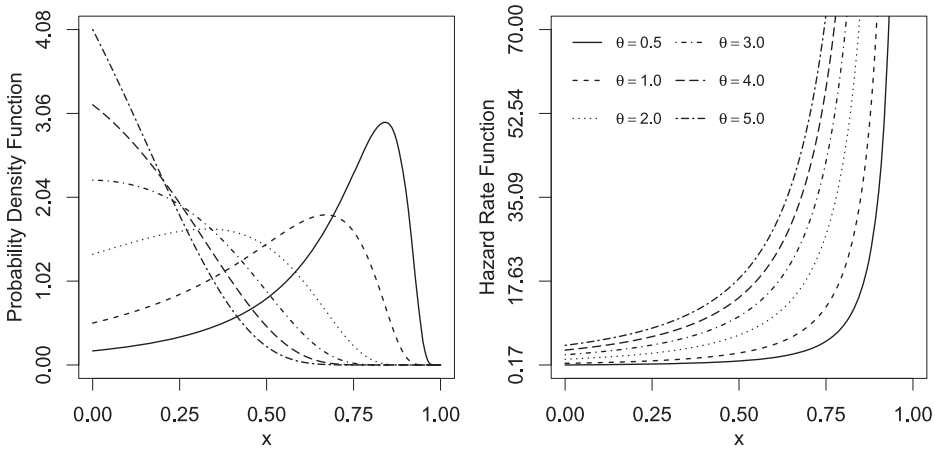


Figure 1. P.d.f and hazard rate function of unit-Lindley distribution for selected values of θ .

which implies that the p.d.f is unimodal with maximum at $X_{\max} = 1 - \theta/3$ for $\theta < 3$ and $X_{\max} = 0$ for $\theta \geq 3$. Figure 1 shows the p.d.f. of the unit-Lindley distribution for selected values of θ .

In what follows we shall discuss several important statistical properties of the unit-Lindley distribution.

2.1. Concavity

Proposition 2.1: *The c.d.f. of the unit-Lindley is concave for $\theta > 3$.*

Proof: The second derivative of $F(x | \theta)$ is

$$F''(x | \theta) = \frac{\theta^2(\theta + 3x - 3)}{(1 + \theta)(x - 1)^5} \exp\left(-\frac{\theta x}{1 - x}\right).$$

This implies for all x in $(0, 1)$, $F''(x | \theta) > 0$ only if $\theta < 0$ therefore it can never be convex and $F''(x | \theta) < 0$ if $\theta > 3$. Hence $F(x | \theta)$ is concave function of x for $\theta > 3$. ■

Proposition 2.2: *The c.d.f. of the unit-Lindley is bounded for $\theta > 3/2$ as follows.*

$$x^\theta \leq F(x | \theta) \leq 1 - \exp\left(-\frac{\theta x}{1 - x}\right).$$

Therefore the c.d.f. can be used to define new premium by distorting survival function [see 28].

Proposition 2.3: *The p.d.f. of the unit-Lindley is log-concave for all $0 < x < 1$ if $\theta \geq \frac{3}{2}$.*

Proof: The second derivative of $F(x | \theta)$ is

$$F''(x | \theta) = \frac{\theta^2(\theta + 3x - 3)}{(1 + \theta)(x - 1)^5} \exp\left(-\frac{\theta x}{1 - x}\right).$$

We know that $f(x | \theta)$ is log-concave (log-convex) function of x if for all x in $(0, 1)$ $\frac{d}{dx} \log f(x | \theta)$ is a non-increasing (non-decreasing) function of x . Note that

$$\frac{d^2}{dx^2} \log f(x | \theta) = \frac{d}{dx} \frac{f'(x | \theta)}{f(x | \theta)} = \frac{d}{dx} \frac{\theta + 3(x - 1)}{(x - 1)^2} = \frac{2\theta + 3(x - 1)}{(x - 1)^3}.$$

This is always < 0 for all x in $(0, 1)$ when $\theta \geq \frac{3}{2}$.

Hence $f(x | \theta)$ is log-concave for all $0 < x < 1$, if $\theta \geq \frac{3}{2}$. ■

As a consequence of the above proposition the following results hold for unit-Lindley distribution when $\theta \geq \frac{3}{2}$:

- $f(x | \theta)$ is log-concave for all $0 < x < 1$;
- $\int_0^x F(t) dt$ is log-concave for all $0 < x < 1$;
- $\bar{F}(x | \theta)$ is log-concave for all $0 < x < 1$;
- $\int_x^1 \bar{F}(t) dt$ is log-concave for all $0 < x < 1$;
- $\frac{f(x|\theta)}{\bar{F}(x|\theta)}$ is monotone increasing function in x for all $0 < x < 1$;
- Mean residual life (MRL) is a decreasing function of x ;
- The distribution is strongly unimodal;
- All moments exist;
- At most has an exponential tail.

2.2. Hazard rate function

The hazard rate function of the unit-Lindley distribution is given by

$$h(x | \theta) = \frac{f(x | \theta)}{1 - F(x | \theta)} = \frac{\theta^2}{(\theta - x + 1)(x - 1)^2}, \quad 0 < x < 1. \quad (4)$$

Since $d/dx h(x | \theta) = \theta^2 / ((x + 1)^3 (\theta - x + 1)^2) [2\theta - 3(x - 1)] > 0$ for all $\theta > 0$ the hazard rate function is increasing in x . Note that $\lim_{x \rightarrow 0} h(x | \theta) = \theta^2 / (1 + \theta)$ while $\lim_{x \rightarrow 1} h(x | \theta) = \infty$. The behavior of $h(x | \theta)$ considering different values of θ is illustrated on the right side of Figure 1.

2.3. Moments

The k th moment about origin of the unit-Lindley distribution is given by

$$\mu_{k'} = \mathbb{E}(X^k) = \frac{k}{(1 + \theta)} \int_0^1 \frac{x^{k-1}(1 - \theta + x)}{(1 - x)} \exp\left(-\frac{\theta x}{1 - x}\right) dx \quad k = 1, 2, \dots,$$

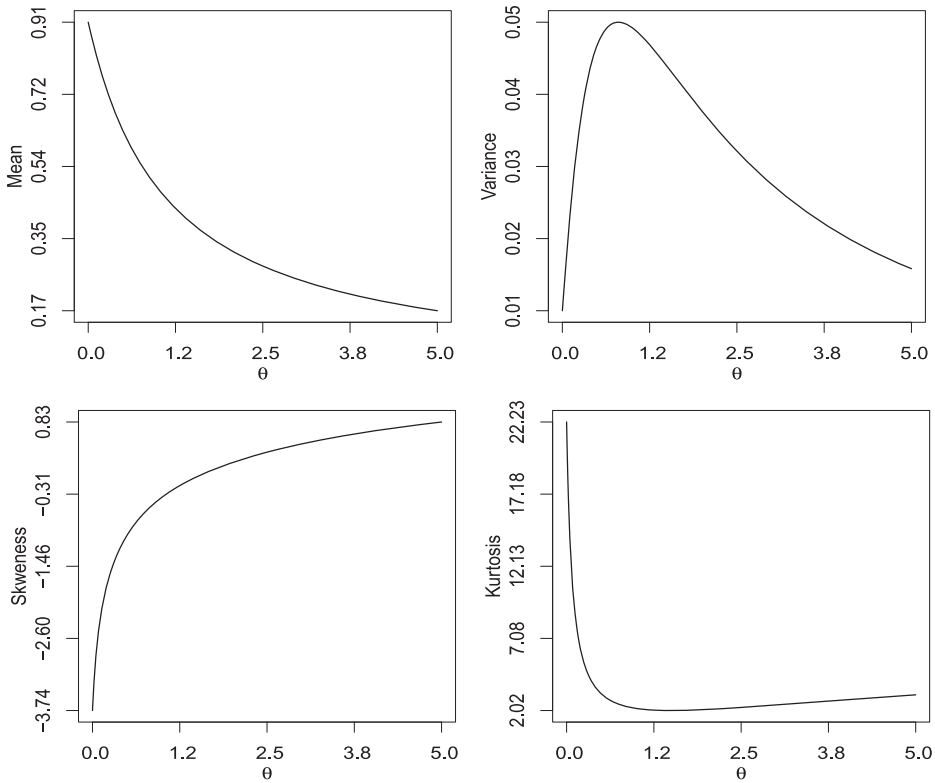


Figure 2. Mean, variance, skewness and kurtosis of unit-Lindley as a function of θ .

which can not be solved analytically. In particular, for $k = 1, 2, 3, 4$ we get

$$\begin{aligned} \mu'_1 &= \frac{1}{1 + \theta}, \\ \mu'_2 &= \frac{1}{1 + \theta} (\theta^2 e^\theta \text{Ei}(1, \theta) - \theta + 1), \\ \mu'_3 &= \frac{1}{1 + \theta} (e^\theta \text{Ei}(1, \theta) \theta^3 + 3 \theta^2 e^\theta \text{Ei}(1, \theta) - \theta^2 - 2 \theta + 1), \\ \mu'_4 &= \frac{1}{2(1 + \theta)} (e^\theta \text{Ei}(1, \theta) \theta^4 + 8 e^\theta \text{Ei}(1, \theta) \theta^3 - \theta^3 \\ &\quad + 12 \theta^2 e^\theta \text{Ei}(1, \theta) - 7 \theta^2 - 6 \theta + 2), \end{aligned}$$

where $\text{Ei}(a, z) = \int_1^\infty z^{-a} e^{-xz} dx$ is the exponential integral function [1].

From the plots of the mean, variance, skewness and kurtosis of the unit-Lindley distribution presented in Figure 2 it is observed that the mean decreases and skewness increases with the increase in θ whereas, the kurtosis initially decreases then increases with θ .

2.4. Incomplete moments

The k th incomplete moment of the unit-Lindley distribution is given by

$$\begin{aligned} T_k(t) &= \mathbb{E} \left(X^k \mid x < t \right) \\ &= \frac{k}{(1+\theta)} \int_0^t \frac{x^{k-1} (1-\theta+x)}{(1-x)} \exp \left(-\frac{\theta x}{1-x} \right) dx, \quad k = 1, 2, \dots \end{aligned}$$

which can not be solved analytically. In particular, for $k=1,2$ we have

$$\begin{aligned} T_1(t) &= \frac{[1 + (\theta - 1)t] e^{t-\theta/(t-1)}}{(t-1)(\theta+1)}, \\ T_2(t) &= \frac{e^{\frac{t-\theta}{t-1}} + (t-1) [\theta^2 e^\theta (\theta+3) \text{Ei}(1, \theta) - \theta^2 - 2\theta + 1] - \theta^2 e^\theta (\theta+3)(t-1) \text{Ei}(1, -\theta/(t-1))}{[(2t-1)\theta - t + 1]}. \end{aligned}$$

2.5. MRL function

For a non-negative continuous random variable X the MRL function is defined as $\mu(x \mid \theta) = \mathbb{E}(X - x \mid X > x)$ and is given by

$$\mu(x \mid \theta) = \frac{1}{S(x \mid \theta)} \int_x^\infty S(y \mid \theta) dy.$$

For the unit-Lindley distribution, we get $\mu(x \mid \theta) = (x-1)^2/(1+\theta+x)$. Note that $\lim_{x \rightarrow 0} \mu(x \mid \theta) = 1/(1+\theta)$ while $\lim_{x \rightarrow 1} \mu(x \mid \theta) = 0$.

2.6. Quantile function

Let X be a unit-Lindley random variable with c.d.f (2). The quantile function, $Q(p) = F^{-1}(p)$, can be written as

$$Q(p \mid \theta) = \frac{1 + \theta + W_{-1} \left((1 + \theta) (p - 1) e^{-(1+\theta)} \right)}{1 + W_{-1} \left((1 + \theta) (p - 1) e^{-(1+\theta)} \right)}, \quad (5)$$

such that $0 < p < 1$ and W_{-1} denotes the negative branch of the Lambert W function. The Lambert W function is a multivalued complex function defined as the solution of the equation $W(z) \exp[W(z)] = z$. For more on Lambert W function interested readers may refer to [4,12,26] and references cited therein.

2.7. Mean deviation

As pointed out, for example in [9], the amount of scatter in a population is measured to some extent by the totality of deviations from the mean and the median. These are known as the mean deviation about the mean and the mean deviation about the median and are

defined as

$$\delta(X) = \int_x^\infty |X - m|f(x | \theta) dx = 2 \left[mF(m) - \int_0^m xf(x | \theta) dx \right], \tag{6}$$

with $m = \mathbb{E}(X)$ or $m = \mathbb{M}edian(X)$, respectively . Considering (2) and (3) in (6) we get:

$$\delta(X) = \frac{2}{1 + \theta} \left(\left[e^{-\theta m/(1-m)}(1 - m) + m(1 + \theta) - 1 \right] \right).$$

For $m = \mathbb{E}(X)$ we get $\delta(X) = 2\theta e^{-1}/((1 + \theta)^2)$. Considering $m = Q(0.5 | \theta)$ we have the expression for the mean deviation about the median. The expression for $Q(\cdot | \theta)$ is given in Section 2.6 .

2.8. Lorenz curve

The Lorenz curve for a random variable X is defined as

$$L(F(q)) = \frac{1}{\mathbb{E}(X)} \mathbb{E}(X | X \leq q)F(q). \tag{7}$$

For the unit-Lindley distribution we have

$$\mathbb{E}(X | X \leq q)(q) = \frac{1}{(1 + \theta)(q - 1)} \left[e^{-\theta q/(1-q)}(1 - q + \theta q) + q - 1 \right].$$

Hence, from (7) we obtain the Lorenz function for the unit-Lindley distribution as

$$L(p) = \frac{1}{(1 + \theta)^2(p - 1)} [e^{-\theta p/(1-p)}(1 - p + \theta p) + p - 1],$$

where $q = F^{-1}(p)$ is given in Section 2.6 .

2.9. Stress strength reliability

Suppose that X and Y are two independent unit-Lindley random variables with parameters θ_1 and θ_2 , respectively, having p.d.f's $f_X(\cdot)$ and $f_Y(\cdot)$. Then the stress-strength reliability measure [15] is given by

$$\begin{aligned} R &= P(Y < X) = \int_0^1 f_X(x | \theta_1)F_Y(x | \theta_2) dx \\ &= \frac{\theta_2^2 (\theta_1 \theta_2^2 + 2\theta_1^2 \theta_2 + \theta_1^3 + \theta_2^2 + 4 \theta_1 \theta_2 + 3 \theta_1^2 + \theta_2 + 3 \theta_1)}{(\theta_1 + \theta_2)^3 (1 + \theta_2) (1 + \theta_1)}. \end{aligned} \tag{8}$$

2.10. Exponential family

A distribution belongs to the exponential family [7] if it is of the form

$$f(x | \theta) = \exp[Q(\theta), T(x | \theta) + D(\theta) + S(x | \theta)].$$

It can be easily seen that the proposed distribution belongs to the exponential family by rewriting the pdf in Equation (3) as

$$f(x | \theta) = \exp \left[-\frac{\theta x}{1-x} \right] \exp \left[\log \frac{\theta^2}{1+\theta} \right] \exp [\log(1-x)^{-3}],$$

where $Q(\theta) = \theta$, $T(x | \theta) = \frac{x}{1-x}$, $D(\theta) = \log \frac{\theta^2}{1+\theta}$, $S(x | \theta) = \log(1-x)^{-3}$.

Therefore, $T(\mathbf{x}) = \sum_{i=1}^n \frac{x_i}{1-x_i}$ is a complete sufficient estimator for θ based on a sample of size n from the proposed distribution. Beside that, since the distribution is exponential family a minimum-variance unbiased estimator can be obtained by bias corrected MLE.

3. Estimation

In this section, we shall consider the estimation of parameter θ of the unit-Lindley distribution by the maximum likelihood method and method of moments. For the maximum likelihood estimator (MLE) of θ we derive the closed-form expressions for the second-order bias-correction.

3.1. Maximum likelihood estimator

Let X_1, \dots, X_n be a random sample from the unit-Lindley distribution with p.d.f. (3). Then, for observed $\mathbf{x} = (x_1, \dots, x_n)$, the log-likelihood function of θ can be written as

$$\ell(\theta | \mathbf{x}) \propto 2n \log \theta - n \log(1 + \theta) - \theta t(\mathbf{x}), \quad (9)$$

where $t(\mathbf{x}) = \sum_{i=1}^n \frac{x_i}{1-x_i}$. The maximum likelihood estimate $\hat{\theta}$ of θ is obtained by solving the following linear equation

$$\frac{d}{d\theta} \ell(\theta | \mathbf{x}) = \frac{2n}{\theta} - \frac{n}{1+\theta} - t(\mathbf{x}) = 0,$$

which gives

$$\hat{\theta} = \frac{1}{2t(\mathbf{x})} [n - t(\mathbf{x}) + \sqrt{t(\mathbf{x})^2 + 6n t(\mathbf{x}) + n^2}]. \quad (10)$$

Next

$$\frac{d^2}{d\theta^2} \ell(\theta | \mathbf{x}) = \frac{n}{(1+\theta)^2} - \frac{2n}{\theta^2} < 0$$

for all θ , in particular for $\theta = \hat{\theta}$.

Since $d^2/d\theta^2 \ell(\theta | \mathbf{x})$ is data-independent, we have that $n \mathbb{E}[d^2/d\theta^2 \log f(X | \theta)] = d^2/d\theta^2 \ell(\theta | \mathbf{x})$. Thus, the expected Fisher information is $\mathbb{I}(\hat{\theta}) = 2n/\theta^2 - n/((1+\theta)^2)$.

From the large sample theory [see, 17, pp. 461–463], the asymptotic distribution of MLE $\widehat{\theta}$ of θ is such that

$$\sqrt{n} (\widehat{\theta} - \theta) \xrightarrow{D} N(0, \mathbb{V}(\widehat{\theta})),$$

where \xrightarrow{D} denotes convergence in distribution and $\mathbb{V}(\widehat{\theta})$ is just the inverse of the expected Fisher information written as $\mathbb{V}(\widehat{\theta}) = (\theta^2 (1 + \theta)^2) / (n (\theta^2 + 4\theta + 2))$. It is easy to see that for $\psi = g(\theta) = \mathbb{E}(X)$ $\widehat{\psi} = \widehat{\mathbb{E}}(X) = 1 / (1 + \widehat{\theta})$ and $\mathbb{V}(\widehat{\psi}) = \theta^2 / (n (\theta^2 + 4\theta + 2))$. Hence, the asymptotic 100 (1 - α)% confidence intervals for θ and ψ are given, respectively, by

$$\widehat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\widehat{\theta}^2 (1 + \widehat{\theta})^2}{n (\widehat{\theta}^2 + 4\widehat{\theta} + 2)}} \quad \text{and} \quad \frac{1}{1 + \widehat{\theta}} \pm z_{\alpha/2} \sqrt{\frac{\widehat{\theta}^2}{n (\widehat{\theta}^2 + 4\widehat{\theta} + 2)}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard Normal distribution.

It is important to note that for a Bayesian analysis we can use the Jeffreys invariant prior for θ , given by $\pi(\theta) \propto \sqrt{\mathbb{I}(\theta)}$. The Bayesian procedure is not considered in this paper.

Cox and Snell [5] provided a framework for estimating the bias, to $\mathcal{O}(n^{-1})$ for the MLEs of the parameters of regular densities. Hence, subtracting the estimated bias from the original MLE produces a bias-corrected estimator (BCE) that is unbiased to $\mathcal{O}(n^{-2})$. Following Cox and Snell [5] the analytical expression for bias-correction of a scalar $\widehat{\theta}$, given by

$$\mathcal{B}(\widehat{\theta}) = (\kappa^{11})^2 [0.5 \kappa_{111} + \kappa_{11,1}] + \mathcal{O}(n^{-2}), \tag{11}$$

where

$$\begin{aligned} \kappa^{11} &= \mathbb{E}\left[-\frac{d^2}{d\theta^2} \ell(\theta | \mathbf{x})\right]^{-1} = \frac{\theta^2 (1 + \theta)^2}{n(\theta^2 + 4\theta + 2)}, \\ \kappa_{11,1} &= \mathbb{E}\left[-\frac{d^2}{d\theta^2} \ell(\theta | \mathbf{x}) \times \frac{d}{d\theta} \ell(\theta | \mathbf{x})\right] = 0 \quad \text{and} \\ \kappa_{111} &= \mathbb{E}\left[-\frac{d^3}{d\theta^3} \ell(\theta | \mathbf{x})\right] = \frac{2 n(\theta^3 + 6\theta^2 + 6\theta + 2)}{\theta^3 (1 + \theta)^3}. \end{aligned}$$

Thus, the bias-corrected MLE $\widetilde{\theta}$ is

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\theta}^5 + 7\widehat{\theta}^4 + 12\widehat{\theta}^3 + 8\widehat{\theta}^2 + 2\widehat{\theta}}{(\widehat{\theta}^2 + 4\widehat{\theta} + 2)^2 n}, \tag{12}$$

where the right-hand side is $\widehat{\mathcal{B}}(\widehat{\theta})$.

Re-parameterizing (3) in terms of the mean $\mu = \frac{1}{1+\theta}$, the maximum likelihood of μ is obtained as

$$\widehat{\mu} = -\frac{1}{2 t(\mathbf{x})} \left[n + t(\mathbf{x}) - \sqrt{t(\mathbf{x})^2 + 6 n t(\mathbf{x}) + n^2} \right],$$

and the corresponding bias-corrected MLE $\widetilde{\mu}$ of μ as

$$\widetilde{\mu} = \widehat{\mu} - \frac{2 \widehat{\mu}^2 (2 \widehat{\mu} - 2)}{n (\widehat{\mu}^2 - 2 \widehat{\mu} - 1)^2}.$$

3.2. Method of moment estimator

Let X_1, \dots, X_n be a random sample from the unit-Lindley distribution with p.d.f (3). Then the method of moment estimator (MME) $\hat{\theta}_{\text{MME}}$ of θ is given by

$$\hat{\theta}_{\text{MME}} = \frac{1 - \bar{X}}{\bar{X}} = \frac{1}{\bar{X}} - 1, \quad (13)$$

which is positively biased, i.e. $E(\hat{\theta}) - \theta > 0$.

Proof: Let $\hat{\theta}_{\text{MME}} = g(\bar{X})$ and $g(t) = 1/t - 1$ for $t > 0$. Since $g''(t) = 2/t^3 > 0$, $g(t)$ is strictly convex. Thus, by Jensen's inequality, we have $E(g(\bar{X})) > g(E(\bar{X}))$. Since $g(E(\bar{X})) = g(1/(1 + \theta)) = \theta$ we get $E(\hat{\theta}) > \theta$. ■

Using the delta method the asymptotic variance of $\hat{\theta}_{\text{MME}}$ is given by

$$\mathbb{V}(\hat{\theta}_{\text{MME}}) = 1/\bar{X}^2 \mathbb{V}(\bar{X}), \quad (14)$$

where

$$\mathbb{V}(\bar{X}) = \frac{\theta^2 e^\theta \text{Ei}(1, \theta) - \theta + 1}{n^2(\theta + 1)}.$$

4. Simulation study

In this section, we conduct a Monte Carlo simulation in order to evaluate and compare the finite-sample behavior of the MLEs, its bias-corrected counterpart obtained by the Cox–Snell methodology (BCE) and the moment estimators (MME) of the parameter θ of the unit-Lindley distribution.

We have generated samples of size $n = 10, 20, 40, 60$ and 80 by considering $\mathbb{E}(X) = 0.1, 0.2, \dots, 0.7$, which implies that $\theta = 9.00, 4.00, \dots, 0.43$. To simulate observations from the unit-Lindley distribution we generated Y from Lindley distribution (see, `rlindley` function in `LindleyR` library) and then used the transformation $X = Y/(1 + Y)$. The simulation experiment was repeated $M = 10,000$ times. The performance evaluation was done based on the estimated bias and root mean-squared error (RMSE).

Table 1 shows that MLE and MME of θ are positive biased, while the BCE estimator achieve substantial bias reduction, especially for small and moderate sample sizes. It is also observed that the RMSE decreases as n increases, as expected. Additionally, the RMSE of the corrected estimates are smaller than those of the uncorrected estimates.

5. Real data analysis

In this section, our interest lies in imposing a regression structure for the variable of interest using the unit-Lindley distribution. In regression analysis it is very common to model the mean of the response. Since the unit-Lindley distribution has closed form expression for mean it can be used in this context. It is noteworthy that the re-parametrized p.d.f of the

Table 1. Estimated bias (root mean-squared error) of θ .

| θ | n | MLE | MME | BCE |
|----------|-----|-----------------|-----------------|------------------|
| 9.00 | 10 | 0.9005 (3.3515) | 0.7898 (3.3314) | 0.0026 (2.9083) |
| | 20 | 0.4240 (2.0844) | 0.3688 (2.0877) | -0.0011 (1.9398) |
| | 40 | 0.2077 (1.3864) | 0.1803 (1.3962) | 0.0005 (1.3369) |
| | 60 | 0.1364 (1.1046) | 0.1187 (1.1152) | -0.0005 (1.0781) |
| | 80 | 0.1036 (0.9488) | 0.0904 (0.9588) | 0.0013 (0.9315) |
| 4.00 | 10 | 0.3634 (1.3720) | 0.2918 (1.3718) | 0.0058 (1.1965) |
| | 20 | 0.1719 (0.8586) | 0.1380 (0.8742) | 0.0026 (0.8013) |
| | 40 | 0.0847 (0.5764) | 0.0682 (0.5926) | 0.0022 (0.5566) |
| | 60 | 0.0553 (0.4625) | 0.0438 (0.4770) | 0.0008 (0.4519) |
| | 80 | 0.0400 (0.3953) | 0.0315 (0.4083) | -0.0007 (0.3886) |
| 2.33 | 10 | 0.1903 (0.7468) | 0.1400 (0.7602) | 0.0027 (0.6571) |
| | 20 | 0.0898 (0.4729) | 0.0658 (0.4931) | 0.0007 (0.4437) |
| | 40 | 0.0444 (0.3162) | 0.0328 (0.3347) | 0.0010 (0.3061) |
| | 60 | 0.0291 (0.2533) | 0.0210 (0.2690) | 0.0004 (0.2480) |
| | 80 | 0.0220 (0.2180) | 0.0160 (0.2320) | 0.0006 (0.2145) |
| 1.50 | 10 | 0.1120 (0.4502) | 0.0770 (0.4722) | 0.0025 (0.3998) |
| | 20 | 0.0536 (0.2877) | 0.0365 (0.3105) | 0.0014 (0.2712) |
| | 40 | 0.0263 (0.1943) | 0.0177 (0.2123) | 0.0009 (0.1886) |
| | 60 | 0.0176 (0.1561) | 0.0120 (0.1716) | 0.0008 (0.1530) |
| | 80 | 0.0134 (0.1338) | 0.0092 (0.1477) | 0.0008 (0.1318) |
| 1.00 | 10 | 0.0648 (0.2847) | 0.0395 (0.3097) | -0.0018 (0.2562) |
| | 20 | 0.0310 (0.1838) | 0.0189 (0.2063) | -0.0008 (0.1744) |
| | 40 | 0.0154 (0.1244) | 0.0095 (0.1422) | -0.0002 (0.1212) |
| | 60 | 0.0102 (0.1000) | 0.0063 (0.1152) | -0.0001 (0.0983) |
| | 80 | 0.0075 (0.0861) | 0.0044 (0.0993) | -0.0002 (0.0850) |
| 0.67 | 10 | 0.0418 (0.1835) | 0.0238 (0.2106) | 0.0005 (0.1665) |
| | 20 | 0.0199 (0.1189) | 0.0115 (0.1412) | 0.0001 (0.1133) |
| | 40 | 0.0097 (0.0808) | 0.0056 (0.0975) | 0.0000 (0.0789) |
| | 60 | 0.0063 (0.0650) | 0.0037 (0.0789) | -0.0001 (0.0640) |
| | 80 | 0.0047 (0.0561) | 0.0028 (0.0682) | -0.0001 (0.0554) |
| 0.43 | 10 | 0.0248 (0.1136) | 0.0122 (0.1388) | 0.0000 (0.1040) |
| | 20 | 0.0116 (0.0743) | 0.0054 (0.0941) | -0.0004 (0.0711) |
| | 40 | 0.0057 (0.0507) | 0.0028 (0.0655) | -0.0002 (0.0497) |
| | 60 | 0.0039 (0.0410) | 0.0019 (0.0532) | 0.0000 (0.0404) |
| | 80 | 0.0029 (0.0354) | 0.0014 (0.0459) | 0.0000 (0.0350) |

unit-Lindley (UL) in terms of the mean can be written as

$$f(y | \mu) = \frac{(1 - \mu)^2}{\mu (1 - y)^3} \exp\left(-\frac{y(1 - \mu)}{\mu(1 - y)}\right), \tag{15}$$

where $0 < y < 1$ and $0 < \mu < 1$.

Let Y_1, \dots, Y_n be n independent random variables, where $Y_i \sim \text{UL}(\mu_i), i = 1, \dots, n$. The regression model is defined assuming that the mean of Y_i satisfies the following functional relation

$$g(\mu_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \tag{16}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ is a p -dimensional vector of regression coefficients ($p < n$) and $\mathbf{x}_i^\top = (x_{i1}, \dots, x_{ip})$ denotes the observations on p known covariates. Note that the variance of Y_i is a function of μ_i and, as a consequence, of the covariate values, which implies that non-constant response variances are naturally accommodated into the model. We shall

assume that the mean link function $g(\cdot)$ is a strictly monotonic and twice differentiable function that maps $(0, 1)$ into \mathbb{R} . Possible candidates for such function are the c.d.f.'s of the normal or the logistic distribution, among others [see 21].

Under a classical approach, the unknown parameter vector $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^\top$ are estimated by maximizing the log-likelihood function, which can be expressed as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^n \ell_i(\mu_i), \quad (17)$$

where

$$\ell_i(\mu_i) = 2 \log(1 - \mu_i) - \log(\mu_i) - 3 \log(1 - y_i) - \frac{y_i(1 - \mu_i)}{\mu_i(1 - y_i)}. \quad (18)$$

The data set used in this section is about the access of people in households with inadequate water supply and sewage in the cities of Brazil from the Southeast and Northeast regions. We are interested in analyzing the association between proportion of households with inadequate water supply and sewage and some sociodemographic variables of these cities. This data set are available in <http://atlasbrasil.org.br/2013/>, consist of 3197 cities and all variables were measured during the census in 2010.

Specifically, we consider the following covariates: region (REG = 0 for Southeast, REG = 1 for Northeast), life expectancy (LIFE), income per capita (INCPC) and human development index (HDI). We also consider the logit link function which ensures that the predicted mean stays within bounds $(0,1)$. Hence the regression structure for μ_i is given by

$$\logit(\mu_i) = \beta_0 + \beta_1 \text{HDI}_i + \beta_2 \text{REG}_i + \beta_3 \text{INCPC}_i + \beta_4 \text{LIFE}_i. \quad (19)$$

For the sake of comparison we also fit the Beta regression model [2,8]. The procedure NLMIXED [23] is used to perform the required computations.

In Table 2 the estimates, the standard errors and the 95% confidence intervals for the parameters of both models are presented. Although the models under investigation provide the same effect of the covariates under the response variable, it can be seen that the estimates of β_1 and β_2 are quite different. Moreover, looking at the 95% confidence intervals, we can see that all covariates are significant to explain the mean of the response variable. For instance, cities with greater values for HDI tend to have less proportion of households with inadequate water supply and sewage.

Table 2. Summary of the fitted models.

| Parameter | Beta | | | unit-Lindley | | |
|-----------|----------|--------|--------------------|--------------|--------|--------------------|
| | Estimate | S. E. | 95% C. I. | Estimate | S. E. | 95% C. I. |
| β_0 | 2.0806 | 0.6230 | (0.8595; 3.3017) | 5.7060 | 0.7831 | (4.1712; 7.2409) |
| β_1 | -2.8030 | 0.5875 | (-3.9545; -1.6515) | -7.8670 | 0.6239 | (-9.0899; -6.6441) |
| β_2 | 0.8228 | 0.0475 | (0.7297; 0.9160) | 0.9736 | 0.0510 | (0.8736; 1.0736) |
| β_3 | -0.0014 | 0.0002 | (-0.0018; -0.0010) | -0.0012 | 0.0001 | (-0.0014; -0.0009) |
| β_4 | -0.0349 | 0.0098 | (-0.0541; -0.0158) | -0.0471 | 0.0127 | (-0.0719; -0.0223) |
| ϕ | 12.7788 | 0.3515 | (12.0898; 13.4678) | — | — | — |

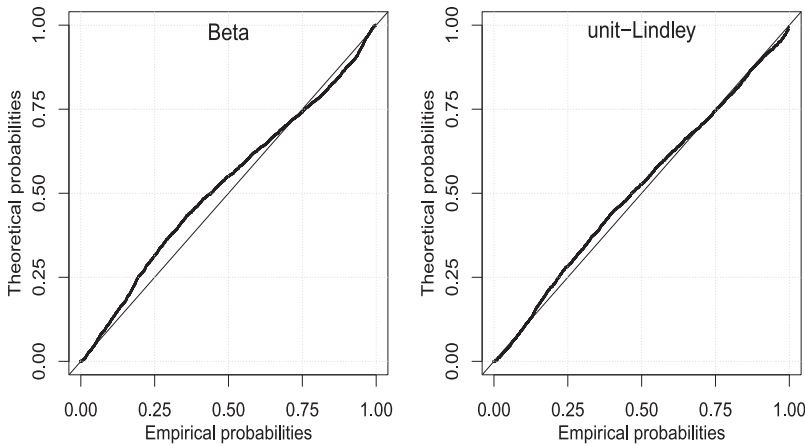


Figure 3. Theoretical and empirical probabilities of the Cox–Snell residuals.

Table 3. Likelihood-based statistics.

| Model | AIC | BIC | HQIC | Young (<i>p</i> -value) |
|--------------|--------------|--------------|--------------|--------------------------|
| unit-Lindley | −11,470.6765 | −11,440.3266 | −11,459.7950 | 5.5654 (< 0.0001) |
| Beta | −11,038.7573 | −11,002.3375 | −11,025.6996 | |

In order to evaluate the fitted models, we have calculated the residuals introduced by Cox and Snell [5]. This residuals are defined as:

$$\hat{e}_i = -\log [1 - \hat{F}(y_i)], \quad i = 1, \dots, n,$$

where $\hat{F}(\cdot)$ is an estimated of the c.d.f.

According to Lawless [16] if the model is appropriate, then the \hat{e}_i should behave approximately like a sample from the standard exponential distribution. Figure 3 shows the probability–probability plots, where the empirical probabilities of \hat{e}_i are compared with those of the standard exponential distribution. It is observed that the plotted points for the unit-Lindley regression are closer to the diagonal line than those of the Beta regression.

To discriminate between the unit-Lindley and the Beta regression models, we compute the likelihood-based statistics (Akaike’s Information Criterion (AIC), Bayesian Information criterion (BIC) and Hannan-Quinn Information Criterion (HQIC)). Finally, we consider the generalized likelihood statistic introduced by Vuong [27] for comparison of non-nested models, in an attempt to choose the better regression model. Based on the results presented in Table 3 we can conclude that the unit-Lindley regression provides the better fit.

6. Concluding remarks

In many fields of applied science certain indicators, percentages, proportions, ratios and rates measured in (0, 1) scale are treated as study variables for characterization of distinct phenomena. The current statistical literature provide very few choices of models to

deal with such variables. Two main such models are the Beta and Kumaraswamy distributions. The present paper has contributed a new one parameter probability distribution with bounded domain constructed by an simple intuitive variable transformation in the Lindley distribution. Random sample from the distribution can be easily simulated by simple transformation of sample generated from Lindley distribution. Several statistical properties of the proposed distribution are studied. Method of moments and maximum likelihood estimation are discussed and analytical expression for the bias correction of the MLE is derived. The fact that the unit-Lindley distribution allows us to incorporate a regression structure in the mean of the response variables, admit it to be seen as an alternative which is more parsimonious compared to the Beta regression model. Application of the proposed model to a real data set yielded a better fit than the Beta regression model. As such we envisage that our new distribution will be highly utilized across all relevant fields of science.

Acknowledgements

We would like to thank the Editor-in-Chief, Associate Editor and two referees for careful reading and for comments which greatly improved the paper.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Josmar Mazucheli  <http://orcid.org/0000-0001-6740-0445>

References

- [1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions with Formulas, graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series, Dover Publications, Incorporated, New York, 1974.
- [2] E. Cepeda-Cuervo, *Variability modeling in generalized linear models*, Ph.D. diss., Mathematics Institute, Universidade Federal do Rio de Janeiro, 2001.
- [3] D.O. Cook, R. Kieschnick, and B. McCullough, *Regression analysis of proportions in finance with self selection*, *J. Empir. Financ.* 15 (2008), pp. 860–867.
- [4] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth, *On the Lambert W function*, *Adv. Comput. Math.* 5 (1996), pp. 329–359.
- [5] D.R. Cox and E.J. Snell, *A general definition of residuals*, *J. R. Stat. Soc. Ser. B* 30 (1968), pp. 248–275.
- [6] F. Cribari-Neto and T.C. Souza, *Religious belief and intelligence: Worldwide evidence*, *Intelligence* 41 (2013), pp. 482–489.
- [7] A.J. Dobson, *An Introduction to Generalized Linear Models*, 2nd ed., Chapman and Hall/CRC, London, 2001.
- [8] S. Ferrari and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, *J. Appl. Stat.* 31 (2004), pp. 799–815.
- [9] M.E. Ghitany, B. Atieh, and S. Nadarajah, *Lindley distribution and its application*, *Math. Comput. Simul.* 78 (2008), pp. 493–506.
- [10] A.K. Gupta and S. Nadarajah, *Handbook of Beta Distribution and its Applications*, CRC Press, New York, 2004.
- [11] M. Hunger, J. Baumert, and R. Holle, *Analysis of sf-6d index data: Is beta regression appropriate?* *Value Health* 14 (2011), pp. 759–767.

- [12] P. Jodrá, *Computer generation of random variables with Lindley or Poisson–Lindley distribution via the Lambert W function*, *Math. Comput. Simul.* 81 (2010), pp. 851–859.
- [13] N.L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions*, 2nd ed., Vol. 2, John Wiley & Sons Inc., New York: 1995.
- [14] R. Kieschnick and B.D. McCullough, *Regression analysis of variates observed on (0, 1): Percentages, proportions and fractions*, *Stat. Model.* 3 (2003), pp. 193–213.
- [15] S. Kotz and L.Y. M. Pensky, *The Stress-Strength Model and its Generalizations: Theory and Applications*, World Scientific, London, 2003.
- [16] J.F. Lawless, *Statistical Models and Methods for Lifetime Data*, 2nd ed: John Wiley and Sons, Hoboken, NJ, 2003.
- [17] E.J. Lehmann and G. Casella, *Theory of Point Estimation*, Springer Verlag, New York, 1998.
- [18] D.V. Lindley, *Fiducial distributions and Bayes' theorem*, *J. R. Stat. Soc. Ser. B (Methodol.)* 20 (1958), pp. 102–107.
- [19] J. Mazucheli, A.F.B. Menezes, and S. Chakraborty, *On the one parameter unit-Lindley distribution and its associated regression model for proportion data* (2018). Available at arXiv preprint arXiv:1801.02512
- [20] J. Mazucheli, L.B. Fernandes, and R.P. Oliveira, *LindleyR: The Lindley Distribution and Its Modifications*, (2016). R package version 1.1.0.
- [21] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. Chapman and Hall, London, 1989.
- [22] L.E. Papke and J.M. Wooldridge, *Econometric methods for fractional response variables with an application to 401(k) plan participation rates*, *J. Appl. Econometrics* 11 (1996), pp. 619–632.
- [23] SAS, *The NLMIXED Procedure, SAS/STAT[®] User's Guide, Version 9.22*, SAS Institute Inc., Cary, NC, 2010.
- [24] T.C. Souza and F. Cribari-Neto, *Intelligence, religiosity and homosexuality non-acceptance: Empirical evidence*, *Intelligence* 52 (2015), pp. 63–70.
- [25] C.W. Topp and F.C. Leone, *A family of J-Shaped frequency functions*, *J. Amer. Statist. Assoc.* 50 (1955), pp. 209–219.
- [26] D. Veberič, *Lambert W function for applications in physics*, *Comput. Phys. Commun.* 183 (2012), pp. 2622–2628.
- [27] Q.H. Vuong, *Likelihood ratio tests for model selection and non-nested hypotheses*, *Econometrica* 57 (1989), pp. 307–333.
- [28] S. Wang, *Premium calculation by transforming the layer premium density*, *ASTIN Bull.* 26 (1996), pp. 71–92.