Computational Statistics (2020) 35:259–279
https://doi.org/10.1007/s00180-019-00921-y

ORIGINAL PAPER

The unit-improved second-degree Lindley distribution: inference and regression modeling

Emrah Altun[1,2] · Gauss M. Cordeiro[3]

Received: 10 April 2019 / Accepted: 17 September 2019 / Published online: 26 September 2019

Abstract

We define a new one-parameter model on the unit interval, called the *unit-improved second-degree Lindley* distribution, and obtain some of its structural properties. The methods of maximum likelihood, bias-corrected maximum likelihood, moments, least squares and weighted least squares are used to estimate the unknown parameter. The finite sample performance of these methods are investigated by means of Monte Carlo simulations. Moreover, we introduce a new regression model as an alternative to the beta, unit-Lindley and simplex regression models and present a residual analysis based on Pearson and Cox–Snell residuals. The new models are proved empirically to be competitive to the beta, Kumaraswamy, simplex, unit-Lindley, unit-Gamma and Topp–Leone models by means of two real data sets. Empirical findings indicate that the proposed models can provide better fits than other competitive models when the data are close to the boundaries of the unit interval.

Keywords Beta regression · Bias-correction · Method of moments · Residual analysis

1 Introduction

The bounded distributions are essential for modeling proportions, percentages, especially observed in economic variables such as proportion of income spent, industry market shares, etc. The beta (Ferrari and Cribari-Neto 2004) and simplex (Kieschnick and McCullough 2003) regression models are commonly used approaches for modeling proportions in economics, actuarial, ecology and environmental sciences. Although

✉ Emrah Altun
emrahaltun123@gmail.com

Gauss M. Cordeiro
gausscordeiro@gmail.com

1 Department of Statistics, Bartin University, 74100 Bartin, Turkey

2 Department of Mathematics, Bartin University, 74100 Bartin, Turkey

3 Department of Statistics, Federal University of Pernambuco, Recife, Brazil

the beta distribution is widely used in several areas of sciences, it has some shortcomings. One of them is that its cumulative distribution function (cdf) is a special function. These type of special functions increase the computational time and complexity in statistical inference. The Topp–Leone model (Topp and Leone 1955) can be viewed as other widely used unit distribution. It has increased its popularity after the work of Nadarajah and Kotz (2003). One of its most important property is that its cdf has a simple form. The Kumaraswamy distribution (Kumaraswamy 1980) is another distribution defined on the unit interval which has increased its popularity after the article of Cordeiro and de Castro (2011). The unit-Gamma distribution (Grassia 1977) is also a widely used distribution in modeling the bounded data sets. Mazucheli et al. (2018a) studied the parameter estimation of its model parameters based on the bias-corrected maximum likelihood estimation method.

In recent years, several researchers have shown a great interest to define new distributions on bounded supports. Mazucheli et al. (2018b) proposed the unit-Birnbaum-Saunders distribution and demonstrated its performance in modeling the monthly water capacity. Recently, Mazucheli et al. (2019) introduced the unit-Lindley distribution and analyzed the access of people in households with inadequate water supply and sewage in the cities of Brazil using the unit-Lindley and beta regression models. Gómez-Déniz et al. (2014) studied the log-Lindley distribution with application to insurance data. Altun and Hamedani (2018) introduced the log-xgamma distribution and studied its statistical properties. Altun (2019) introduced the log-weighted exponential distributions and its associated regression model. The goal of this paper is to propose a new distribution defined on the unit interval which has tractable statistical properties. It arises from the *improved second-degree Lindley* (ISDL) distribution (Karuppusamy et al. 2017) and has several advantages over well-known models such as the beta, Kumaraswamy and Topp–Leone distributions. Some of its statistical properties are determined in closed-form, such as probability density function (pdf), ordinary and incomplete moments. More importantly, it can provide better fits than other well-known distributions defined on the unit interval. Moreover, a new regression model based on the proposed distribution is investigated as a useful alternative to the beta regression model.

The rest of the paper is organized as follows. In Sect. 2, we define the new distribution and obtain some of its mathematical properties. In Sect. 3, we estimate the model parameters using five different methods: maximum likelihood, bias-corrected maximum likelihood, moments, least squares and weighted least squares. In Sect. 4, a Monte-Carlo simulation study is conducted to evaluate the finite sample performance of the estimation methods. In Sect. 5, we introduce a new regression model based on the proposed distribution as an alternative to the beta regression model. In Sect. 6, two real data sets are analyzed to prove empirically the flexibility of the new models. Some conclusions are offered in Sect. 7.

2 The unit-ISDL distribution

The Lindley density is

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

where $\theta > 0$ is the scale parameter. The Lindley distribution is a mixture of the Exponential (θ) and Gamma ($2, \theta$) distributions. Its cdf is

$$F(x;\theta) = 1 - \frac{\theta+1+\theta x}{\theta+1}\exp(-\theta x), \quad x \geq 0.$$

Most of its statistical properties such as moments, stochastic ordering and entropies were obtained by Ghitany et al. (2008). Recently, Karuppusamy et al. (2017) introduced the ISDL density

$$f(x;\lambda) = \frac{\lambda^3}{\lambda^2+2\lambda+2}(1+x)^2\exp(-\lambda x), \quad x > 0, \tag{1}$$

where $\lambda > 0$ is the shape parameter. The cdf corresponding to (1) is

$$F(x;\lambda) = 1 - \left[1 + \frac{\lambda^2 x^2 + 2\left(\lambda^2+\lambda\right)x}{\lambda^2+2\lambda+2}\right]\exp(-\lambda x), \quad x \geq 0.$$

The ISDL density can be expressed as a mixture of the Gamma$(1, \lambda)$, Gamma$(2, \lambda)$ and Gamma$(3, \lambda)$ densities. So, Eq. (1) can be rewritten as follows

$$f(x;\lambda) = p_1 f_1(x) + p_2 f_2(x) + p_3 f_3(x), \tag{2}$$

where

$$p_1 = \frac{\lambda^2}{\lambda^2+2\lambda+2}, \quad p_2 = \frac{2\lambda}{\lambda^2+2\lambda+2}, \quad p_3 = \frac{2}{\lambda^2+2\lambda+2}, \tag{3}$$

and

$$f_1(x) = \lambda\exp(-\lambda x), \quad f_2(x) = \lambda^2 x\exp(-\lambda x), \quad f_3(x) = \lambda^3 x^2\exp(-\lambda x)/2. \tag{4}$$

Substituting (3) and (4) in (2), we find the ISDL density (1). Some statistical properties of the ISDL distribution can be easily obtained from those of the Gamma distribution. Here, we introduce a new distribution on the unit-interval by taking the ISDL distribution as the baseline model.

Proposition 1 *Let X be a random variable with pdf (1) and define the random variable $Y = X/(X+1)$. Then, the pdf of Y is*

$$f(y;\lambda) = \frac{\lambda^3(1-y)^{-2}}{\lambda^2+2\lambda+2}\left(1+\frac{y}{1-y}\right)^2\exp\left(-\frac{y\lambda}{1-y}\right), \quad 0 < y < 1, \tag{5}$$

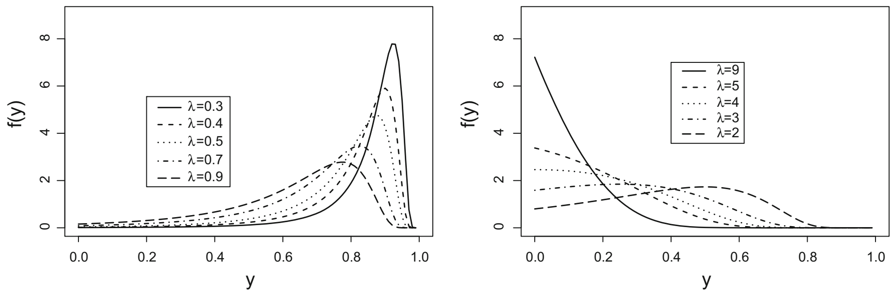

Fig. 1 Plots of the unit-ISDL density for some parameter values

where $\lambda > 0$ *is the shape parameter. Hereafter, the random variable* Y *having density* (5) *is denoted by* $Y \sim$ *unit-ISDL*(λ).

Note that the density given in (5) could be introduced for a three-component general mixture of Gamma densities using the same transformation. However, it increases the model complexity and its statistical properties cannot be obtained in closed-form. Therefore, particular case of a three-component general mixture of Gamma is used. The cdf corresponding to (5) (for $0 \leq y \leq 1$) is

$$F(y;\lambda) = 1 - \left[1 + \frac{\lambda^2(y/(1-y))^2 + 2(\lambda^2+\lambda)y/(1-y}{\lambda^2+2\lambda+2}\right]\exp\left(-\frac{y\lambda}{1-y}\right). \quad (6)$$

Figure 1 displays some possible unit-ISDL density shapes. It can be a good choice for modeling extremely left or right skewed data sets.

Random values from $Y \sim$ unit-ISDL(λ) are generated by solving $F(y;\lambda) = u$, where $u \sim uniform(0, 1)$. The **uniroot** function of R software can be used to solve this non-linear equation.

2.1 Moments

The kth ordinary moment of Y takes the form

$$E(Y^k) = \int_0^1 y^k \frac{\lambda^3(1-y)^{-2}}{\lambda^2+2\lambda+2}\left(1+\frac{y}{1-y}\right)^2 \exp\left(-\frac{y\lambda}{1-y}\right)dy.$$

The above integration can not be carried out analytically but only numerically. In particular,

$$E(Y) = \frac{\lambda+2}{\lambda^2+2\lambda+2} \quad \text{and} \quad E(Y^2) = \frac{2}{\lambda^2+2\lambda+2}.$$

The variance of Y follows easily from these results as

$$Var(Y) = \frac{2}{\lambda^2+2\lambda+2} - \frac{(\lambda+2)^2}{(\lambda^2+2\lambda+2)^2}.$$

Clearly, the mean and variance of the unit-ISDL distribution decrease when λ increases.

2.2 Incomplete moments

The rth incomplete moment of Y is

$$m_r(t) = E(Y^r \,|\, y < t) = \frac{\lambda^3}{\lambda^2 + 2\lambda + 2} \int_0^t \frac{y^r}{(1-y)^2} \left(1 + \frac{y}{1-y}\right)^2 \exp\left(-\frac{\lambda y}{1-y}\right) dy.$$

The above integration can not be carried out analytically. In particular, for $r = 1$, we have

$$m_1(t) = \frac{\lambda + 2}{\lambda^2 + 2\lambda + 2} - \frac{\exp\left(\frac{\lambda t}{t-1}\right)(\lambda^2 t - \lambda t^2 + \lambda + 2t^2 - 4t + 2)}{(t-1)^2 (\lambda^2 + 2\lambda + 2)}.$$

The first incomplete moment can be used to obtain the mean deviations and Lorenz and Bonferroni curves that are fundamental tools for analyzing data in economics and reliability.

2.3 Exponential family

A distribution belongs to the exponential family if it can be expressed as

$$f(y; \lambda) = \exp[Q(\lambda)T(y) + D(\lambda) + S(y)].$$

It is clear that the unit-ISDL distribution belongs to the exponential family by rewriting (5) as

$$\begin{aligned} f(y; \lambda) = \exp\left[-\frac{y\lambda}{1-y}\right] \exp\left[\ln\left(\frac{\lambda^3}{\lambda^2 + 2\lambda + 2}\right)\right] \\ \exp\left[-2\ln(1-y) + 2\ln\left(1 + \frac{y}{1-y}\right)\right], \end{aligned}$$

where $Q(\lambda) = \lambda$, $T(y) = y/(1-y)$,

$$S(y) = -2\ln(1-y) + 2\ln\left(1 + \frac{y}{1-y}\right) \quad \text{and} \quad D(\lambda) = \ln\left(\frac{\lambda^3}{\lambda^2 + 2\lambda + 2}\right).$$

Here, $T(\mathbf{y}) = \sum_{i=1}^n y_i/(1-y_i)$ is the sufficient statistic for λ.

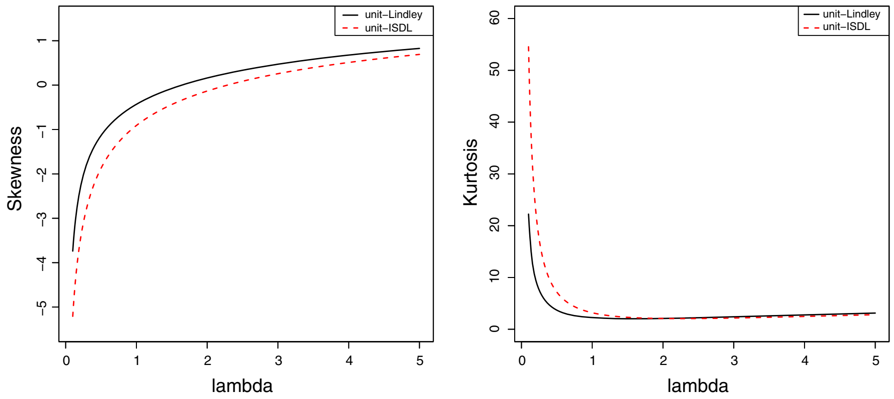

Fig. 2 Plots of the skewness and kurtosis measures of the unit-ISDL and unit-Lindley distributions

Table 1 The tail probabilities of the unit-Lindley and unit-ISDL distributions for selected parameter values

Parameter	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$
$P(X < 0.5)$					
Unit-Lindley	0.0882	0.1382	0.1913	0.2454	0.2989
Unit-ISDL	0.0196	0.0398	0.0669	0.0997	0.1370
$P(X < 0.4)$					
Unit-Lindley	0.0553	0.0882	0.1242	0.1621	0.2008
Unit-ISDL	0.0108	0.0225	0.0385	0.0585	0.0820
$P(X < 0.3)$					
Unit-Lindley	0.0337	0.0544	0.0776	0.1025	0.1284
Unit-ISDL	0.0060	0.0126	0.0218	0.0336	0.0478

2.4 Comparison of the unit-ISDL and unit-Lindley distributions

The unit-ISDL density has similar shapes of the unit-Lindley density. It is needed to clarify the differences between these two distributions. Therefore, the skewness and kurtosis measures of these two distributions are compared. Moreover, the tail-properties are also discussed. Figure 2 displays the numerical results for the skewness and kurtosis measures of the unit-ISDL and unit-Lindley distributions. The plots in Fig. 2 reveal that the skewness and kurtosis values of the unit-ISDL distribution have wider ranges than those of the unit-Lindley distribution. It is clear that the unit-ISDL distribution is a better choice than the unit-Lindley distribution for modeling left-skewed and leptokurtic data.

Table 1 gives the tail probabilities of the unit-Lindley and unit-ISDL distributions for selected parameter values. These probabilities indicate that the left tail of the unit-ISDL distribution is thinner than that one of the unit-Lindley distribution. Therefore, the new distribution can be preferable than the unit-Lindley for modeling left-skewed and thin-tailed data.

3 Estimation

In this section, we consider the methods of maximum likelihood, moments, least squares and weighted least squares to estimate the unknown parameter of the unit-ISDL distribution.

3.1 Maximum likelihood

Let $y_1, \ldots, y_n$ be a random sample from the unit-ISDL distribution. The log-likelihood function for λ is

$$\ell(\lambda) \propto 3n \ln(\lambda) - n \ln(\lambda^2 + 2\lambda + 2) - \lambda t(\mathbf{y}), \tag{7}$$

where $t(\mathbf{y}) = \sum_{i=1}^{n} y_i/(1 - y_i)$. By differentiating (7) with respect to λ gives

$$\frac{\partial \ell}{\partial \lambda} = \frac{3n}{\lambda} - \frac{n(2\lambda + 2)}{\lambda^2 + 2\lambda + 2} - t(\mathbf{y}) \tag{8}$$

Solving (8) for zero, the maximum likelihood estimate (MLE) of λ, say $\hat{\lambda}$, is

$$\begin{aligned}\hat{\lambda} = \frac{1}{3z} & \left[\left(10z^3 + n^3 + 12zn^2 + 3\sqrt{6}\sqrt{2z^6 + z^2n^4 + 14z^3n^3 + 54z^4n^2 + 16z^5n} + 48z^2n \right)^{\frac{1}{3}} \right] \\ & - (2z^2 - n^2 - 8zn) \\ & \times \left[3z \left(10z^3 + n^3 + 12zn^2 + 3\sqrt{6}\sqrt{2z^6 + z^2n^4 + 14z^3n^3 + 54z^4n^2 + 16z^5n} + 48z^2n \right)^{\frac{1}{3}} \right]^{-1} \\ & - \frac{2z - n}{3z},\end{aligned}$$

where $z = t(\mathbf{y})$. The second-order derivative of (7) with respect to λ is

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{n(\lambda^4 + 8\lambda^3 + 24\lambda^2 + 24\lambda + 12)}{\lambda^2(\lambda^2 + 2\lambda + 2)^2}.$$

Hence, the expected information is

$$I(\lambda) = E\left(-\frac{\partial^2 \ell}{\partial \lambda^2}\right) = \frac{n(\lambda^4 + 8\lambda^3 + 24\lambda^2 + 24\lambda + 12)}{\lambda^2(\lambda^2 + 2\lambda + 2)^2}. \tag{9}$$

The asymptotic variance of $\hat{\lambda}$ is easily obtained by inverting (9)

$$Var(\hat{\lambda}) = \frac{\lambda^2(\lambda^2 + 2\lambda + 2)^2}{n(\lambda^4 + 8\lambda^3 + 24\lambda^2 + 24\lambda + 12)}.$$

The equi-tailed $100(1 - p)\%$ confidence interval (CI) for λ is

$$\widehat{\lambda} \pm z_{p/2}\sqrt{Var(\widehat{\lambda})},$$

where $z_{p/2}$ is the upper $p/2$ quantile of the standard normal distribution.

3.2 Bias-corrected maximum likelihood

We adopt a "corrective" approach to reduce the bias of $\hat{\lambda}$ to order $O(n^{-2})$ since the MLE is biased to order $O(n^{-1})$ in finite samples. Following Cox and Snell (1968), when the observations are independent but not necessarily identically distributed, the bias-correction of $\hat{\lambda}$ can be expressed as

$$\mathrm{B}(\hat{\lambda}) = (\kappa^{11})^2 \left(\frac{1}{2}\kappa_{111} + \kappa_{11,1}\right) + O(n^{-2}),$$

where $\kappa^{11} = E\left(-\frac{\partial^2 \ell}{\partial \lambda^2}\right)^{-1}$, $\kappa_{111} = E\left(-\frac{\partial^3 \ell}{\partial \lambda^3}\right)$ and $\kappa_{11,1} = E\left(-\frac{\partial^2 \ell}{\partial \lambda^2}\frac{\partial \ell}{\partial \lambda}\right)$. These expressions for the unit-ISDL distribution are

$$\kappa^{11} = \frac{\lambda^2(\lambda^2+2\lambda+2)^2}{n(\lambda^4+8\lambda^3+24\lambda^2+24\lambda+12)},$$
$$\kappa_{111} = \frac{6n}{\lambda^3} - n\left[\frac{2(2\lambda+2)^3}{(\lambda^2+2\lambda+2)^3} - \frac{6(2\lambda+2)}{(\lambda^2+2\lambda+2)^2}\right],$$

and $\kappa_{11,1} = 0$.

Thus, the bias-corrected maximum likelihood estimate (BC-MLE) of λ is

$$\tilde{\lambda} = \hat{\lambda} - \frac{\lambda^4/(\lambda^2+2\lambda+2)^4}{n^2(\lambda^4+8\lambda^3+24\lambda^2+24\lambda+12)^2}\left[\frac{3n}{\lambda^3} - \frac{n}{2}\left(\frac{2(2\lambda+2)^3}{(\lambda^2+2\lambda+2)^3} - \frac{6(2\lambda+2)}{(\lambda^2+2\lambda+2)^2}\right)\right].$$

3.3 Method of moments

The method of moments (MOM) to estimate λ follows by equating the first theoretical moment of the unit-ISDL distribution to the sample mean, namely

$$\hat{\lambda}_{MOM} = \frac{\sqrt{-4\bar{y}^2+4\bar{y}+1} - 2\bar{y}+1}{2\bar{y}},$$

where $\bar{y} = \sum_{i=1}^{n} y_i \big/ n$.

3.4 Least squares

Let $y_{(1)}, \ldots, y_{(n)}$ denote the ordered sample of size n from the unit-ISDL distribution. The least squares estimator (LSE) of λ is found by minimizing

$$\sum_{i=1}^{n}\left[F(y_{(i)};\lambda) - \frac{i}{n+1}\right]^2, \qquad (10)$$

where $F(y_{(i)}; \lambda)$ is the cdf of the unit-ISDL distribution. Inserting (6) in Eq. (10) gives

$$\sum_{i=1}^{n}\left\{1-\left[1+\frac{\lambda^2\left(y_{(i)}/\left(1-y_{(i)}\right)\right)^2+2(\lambda^2+\lambda)y_{(i)}/\left(1-y_{(i)}\right)}{\lambda^2+2\lambda+2}\right]\exp\left(-\frac{y_{(i)}\lambda}{1-y_{(i)}}\right)-\frac{i}{n+1}\right\}^2.$$

3.5 Weighted least squares

The weighted least square estimate (WLSE) of λ follows by minimizing

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left(\begin{array}{l}1-\left[1+\frac{\lambda^2\left(y_{(i)}/\left(1-y_{(i)}\right)\right)^2+2(\lambda^2+\lambda)y_{(i)}/\left(1-y_{(i)}\right)}{\lambda^2+2\lambda+2}\right]\\ \times\exp\left(-\frac{y_{(i)}\lambda}{1-y_{(i)}}\right)-\frac{i}{n+1}\end{array}\right)^2.$$

4 The unit-ISDL regression model

The beta regression model is widely used when the response variable is in the interval (0, 1). We define a new regression model as an alternative to the beta regression model by re-parameterizing the new distribution using a convenient systematic component for the mean.

Let $\lambda = \left(\sqrt{-4\mu^2+4\mu+1}-2\mu+1\right)(2\mu)^{-1}$. Then, the pdf of the re-parametrized unit-ISDL distribution in terms of $E(Y)=\mu$ takes the form

$$\begin{aligned} f(y;\mu) &= \frac{\left[\left(\sqrt{-4\mu^2+4\mu+1}-2\mu+1\right)(2\mu)^{-1}\right]^3(1-y)^{-2}}{\left[\left(\sqrt{-4\mu^2+4\mu+1}-2\mu+1\right)(2\mu)^{-1}\right]^2+2\left(\sqrt{-4\mu^2+4\mu+1}-2\mu+1\right)(2\mu)^{-1}+2}\\ &\times\left(1+\frac{y}{1-y}\right)^2\exp\left(-\frac{y\left(\sqrt{-4\mu^2+4\mu+1}-2\mu+1\right)(2\mu)^{-1}}{1-y}\right),\quad 0<y<1, \end{aligned}\tag{11}$$

where $0<\mu<1$. The covariates are linked to the response variable Y in the usual way through the logit-link function

$$\mu_i=\frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)},\quad i=1,\ldots,n,\tag{12}$$

where $\boldsymbol{x}_i^T=\left(x_{i1},\ldots,x_{ip}\right)$ is the vector of covariates and $\boldsymbol{\beta}=\left(\beta_1,\ldots,\beta_p\right)^T$ is the vector of unknown regression coefficients. Other links can be considered in the similar manner.

Inserting (12) in (11), the log-likelihood function can be expressed as

$$\ell(\boldsymbol{\beta}) = 3\sum_{i=1}^{n} \ln\left[\left(\sqrt{-4\mu_i^2 + 4\mu_i + 1} - 2\mu_i + 1\right)(2\mu_i)^{-1}\right] - 2\sum_{i=1}^{n}(1 - y_i)$$
$$- \sum_{i=1}^{n} \ln\left(\begin{array}{l}\left[\left(\sqrt{-4\mu_i^2 + 4\mu_i + 1} - 2\mu_i + 1\right)(2\mu_i)^{-1}\right]^2 \\ +2\left(\sqrt{-4\mu_i^2 + 4\mu_i + 1} - 2\mu_i + 1\right)(2\mu_i)^{-1} + 2\end{array}\right)$$
$$+2\sum_{i=1}^{n}\ln\left(1 + \frac{y_i}{1 - y_i}\right) - \sum_{i=1}^{n}\frac{y_i\left(\sqrt{-4\mu_i^2 + 4\mu_i + 1} - 2\mu_i + 1\right)(2\mu_i)^{-1}}{1 - y_i},$$

where μ_i is given by (12). Under standard regularity conditions, the asymptotic distribution of $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is multivariate normal $N_p(0, K(\boldsymbol{\beta})^{-1})$, where $K(\boldsymbol{\beta})$ is the expected information matrix. The asymptotic covariance matrix $K(\boldsymbol{\beta})^{-1}$ of $\widehat{\boldsymbol{\beta}}$ can be approximated by the inverse of the $p \times p$ observed information matrix $-\ddot{\mathrm{L}}(\boldsymbol{\beta})$, whose elements are evaluated numerically by most statistical packages. The approximate multivariate normal distribution $N_p(0, -\ddot{\mathrm{L}}(\boldsymbol{\beta})^{-1})$ for $\widehat{\boldsymbol{\beta}}$ can be used in the classical way to construct approximate confidence intervals for the parameters in $\boldsymbol{\beta}$.

4.1 Residuals analysis

Residual analysis has a critical role in checking the adequacy of the fitted model. In order to analyze departures from the model assumptions, three types of residuals are usually utilized: the residuals introduced by Cox and Snell (1968), the randomized quantile residuals defined by Dunn and Smyth (1996) and Pearson residuals.

4.1.1 Cox and Snell residuals

Cox and Snell (1968) residuals are defined by

$$\hat{e}_i = -\ln\left[1 - F(y_i; \hat{\boldsymbol{\beta}})\right], \quad i = 1, \ldots, n,$$

where $F(\cdot)$ is the unit-ISDL cdf. If the fitted model is correct, Cox and Snell's residuals are approximately distributed as standard exponential distribution.

4.1.2 Randomized quantile residuals

The randomized quantile residuals are defined as

$$\hat{r}_i = \Phi^{-1}\left(\hat{u}_i\right),$$

where $\hat{u}_i = F(y_i; \hat{\boldsymbol{\beta}})$ and $\Phi^{-1}(z)$ is the inverse of the standard normal cdf. If the fitted model is correct, the randomized quantile residuals are distributed as standard normal distribution.

4.1.3 Pearson residuals

The Pearson residual is widely used to detect possible outliers in the data. The Pearson residual is based on the idea of subtracting off the mean and dividing by the standard deviation. For the unit-ISDL regression model, the Pearson residuals can be expressed as

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{\mathrm{Var}}\,(y_i)}},$$

where

$$\widehat{\mathrm{Var}}\,(y_i) = \frac{2\,\hat{\mu}_{\mathrm{i}}^2\left(\sqrt{-4\,\hat{\mu}_{\mathrm{i}}^2 + 4\,\hat{\mu}_{\mathrm{i}} + 1} - 2\,\hat{\mu}_{\mathrm{i}}\sqrt{-4\,\hat{\mu}_{\mathrm{i}}^2 + 4\,\hat{\mu}_{\mathrm{i}} + 1} + 1\right)}{8\,\hat{\mu}_{\mathrm{i}} + 4\,\hat{\mu}_{\mathrm{i}}\sqrt{-4\,\hat{\mu}_{\mathrm{i}}^2 + 4\,\hat{\mu}_{\mathrm{i}} + 1} + 2\sqrt{-4\,\hat{\mu}_{\mathrm{i}}^2 + 4\,\hat{\mu}_{\mathrm{i}} + 1} + 2}.$$

The plot of these residuals against the index of the observations should reveal no detectable pattern. If the fitted model is correct, the Pearson residuals lie in the interval $(-2, 2)$. The residuals outside this range are associated with potential outlier observations.

5 Simulations

In this section, we perform two simulation studies to examine the estimation methods in the proposed models.

5.1 Simulation study for the unit-ISDL distribution

First, we obtain the MLE, BC-MLE, MOM, LSE and WLSE of the unknown parameter λ in the unit-ISDL distribution. We compare the estimation efficiency of these five estimates by means of Monte Carlo simulations. The following simulation procedure is implemented:

1. Set the sample size n and the parameter λ;
2. Generate n random observations from the unit-ISDL (λ) distribution;
3. Use the generated observations in Step 2, estimate λ by means of the MLE, BC-MLE, MOM, LSE and WLSE methods;
4. Repeat N times the steps 2 and 3;
5. Use $\hat{\lambda}$ and λ to calculate the mean biases, mean relative estimates (MREs) and mean square errors (MSEs) from the following equations:

$$Bias = \sum_{j=1}^{N} \frac{\hat{\lambda}_j - \lambda}{N}, \quad MRE = \sum_{j=1}^{N} \frac{\hat{\lambda}_j/\lambda}{N}, \quad MSE = \sum_{j=1}^{N} \frac{(\hat{\lambda}_j - \lambda)^2}{N}.$$

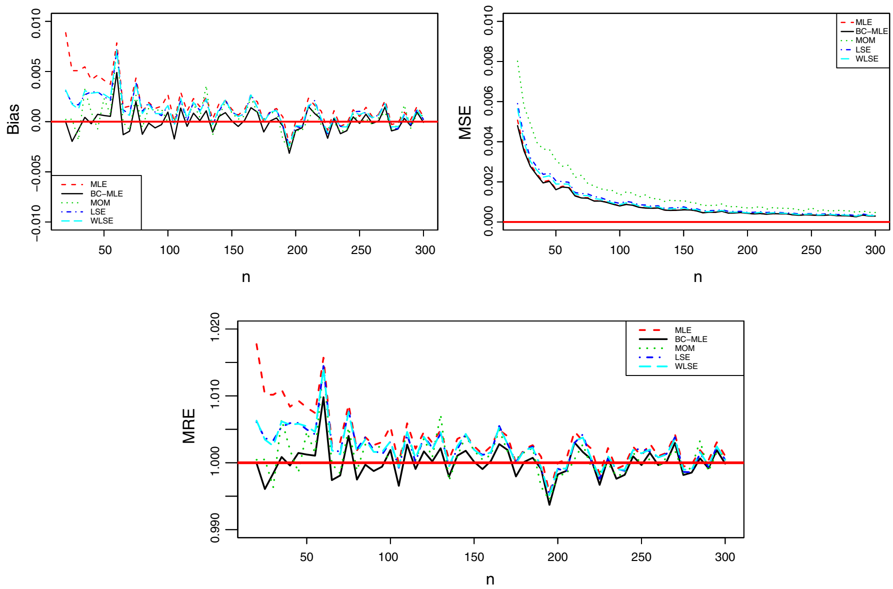

Fig. 3 Estimated biases, MSEs and MREs for the parameter λ of the unit-ISDL distribution

The simulation results are carried out using the **R** software. We take $\lambda = 0.5$, $N = 10{,}000$ and $n = 20, 25, 30, \ldots, 300$. We expect that the MREs are closer to one when the MSEs are near zero. The plots of the estimated mean biases, MSEs and MREs obtained by the MLE, BC-MLE, MOM, LSE and WLSE methods are displayed in Fig. 3. Based on these plots, we note that the mean biases and MSEs of all estimates tend to zero when n increases and, as expected, the MRE values tend to one. Further, they reveal that the BC-MLE method converges to the nominal value of λ faster than the MLE, MOM, LSE and WLSE methods. So, we can conclude that the BC-MLE method can be chosen as more reliable than the MLE, MOM, LSE and WLSE methods for the parameter λ of the new distribution.

5.2 Simulation study for the unit-ISDL regression

Second, we investigate the performance of the MLEs of the parameters in the unit-ISDL regression model by means of a simulation study. We generate $N = 10,000$ samples of sizes $n = 50, 250, 500$ and 1000 from model (12) under the systematic component:

$$\mu_i = \frac{\exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)}{1 + \exp\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}\right)},$$

where x_1 and x_1 are generated from a uniform $U(0, 1)$ distribution. We take parameter values $\beta_0 = 0.5$, $\beta_1 = 0.5$ and $\beta_2 = 2$, and the response variable y_i is generated from this systematic component. The precision of the MLEs is based on the averages of the estimates (AEs), mean biases and mean square errors (MSEe). Table 2 gives the simulation results. Based on the figures in this table, we note that the MSEs of the

Table 2 The AEs, biases and MSEs based on 10,000 simulations for the unit-ISDL regression with parameters: $\beta_0 = 0.5$, $\beta_1 = 0.5$ and $\beta_2 = 2$ for $n = 50, 250, 500$ and 1000

Sample size	Parameters	β_0	β_1	β_2
n = 50	AE	0.613679	0.617070	1.959483
	Bias	0.113679	0.117070	– 0.040517
	MSE	0.328236	0.074869	0.063743
n = 250	AE	0.552417	0.589977	1.962157
	Bias	0.052417	0.089977	– 0.037843
	MSE	0.205544	0.050730	0.047477
n = 500	AE	0.536856	0.552461	1.984456
	Bias	0.036856	0.052461	– 0.015544
	MSE	0.085162	0.031197	0.032380
n = 1000	AE	0.506274	0.515114	1.994224
	Bias	0.006274	0.015114	– 0.005776
	MSE	0.057508	0.014472	0.013230

MLEs of the parameters decay toward zero when the sample size increases as expected under first-order asymptotic theory. This fact reveals the consistency property of the MLEs.

6 Empirical studies

6.1 Univariate data modeling

In this section, we compare the unit-ISDL distribution with four alternative distributions by means of a real data set (Nadar et al. 2013). The data refer to the monthly water capacity from the Shasta reservoir in California, USA, taken for the month of February from 1991 to 2010. The information about the hazard shape can be helpful in selecting a suitable model. For this purpose, a device called the total time on test (TTT) plot (Aarset 1987) can be used. The TTT plot is obtained by plotting

$$G(r/n) = \left[\left(\sum_{i=1}^{r} y_{(i)}\right) + (n-r)y_{(r)}\right] \Big/ \sum_{i=1}^{n} y_{(i)},$$

against r/n, where $y_{(i)}$ $(i = 1, \ldots, n)$ are the order statistics of the sample (for $r = 1, \ldots, n$). The TTT plot given in Fig. 4 indicates that the hazard shape of the monthly water capacity data is increasing. As suggested by reviewer, we compare the unit-ISDL distribution with unit-three-component-gamma (unit-3CGamma) distribution. The pdf of 3CGamma distribution is

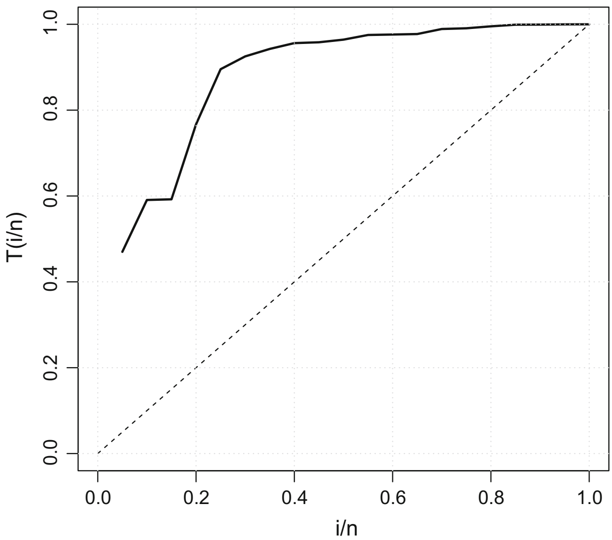

Fig. 4 The TTT plot of the monthly water capacity data

$$f(x;\alpha_1,\alpha_2,\alpha_3,\lambda_1,\lambda_2,\lambda_3) = p_1\frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)}x^{\alpha_1-1}\exp(-\lambda_1 x) + p_2\frac{\lambda_2^{\alpha_2}}{\Gamma(\alpha_2)}x^{\alpha_2-1}\exp(-\lambda_2 x) + p_3\frac{\lambda_3^{\alpha_3}}{\Gamma(\alpha_3)}x^{\alpha_3-1}\exp(-\lambda_3 x) \tag{13}$$

where

$$p_1 = \frac{\lambda_1^2}{\lambda_1^2+2\lambda_2+\lambda_3}, \quad p_2 = \frac{2\lambda_2}{\lambda_1^2+2\lambda_2+\lambda_3}, \quad p_3 = \frac{\lambda_3}{\lambda_1^2+2\lambda_2+\lambda_3} \tag{14}$$

Let X be a random variable with pdf (13). Using the transformation $Y = X/(X+1)$, the pdf of unit-3CGamma distribution $(0 < y < 1)$ is

$$\begin{aligned} f(y;\alpha_1,\alpha_2,\alpha_3,\lambda_1,\lambda_2,\lambda_3) = {} & p_1\frac{\lambda_1^{\alpha_1}}{\Gamma(\alpha_1)(1-y)^{\alpha_1+1}}y^{\alpha_1-1}\exp\left(-\lambda_1\frac{y}{1-y}\right) \\ & + p_2\frac{\lambda_2^{\alpha_2}}{\Gamma(\alpha_2)(1-y)^{\alpha_2+1}}y^{\alpha_2-1}\exp\left(-\lambda_2\frac{y}{1-y}\right) \\ & + p_3\frac{\lambda_3^{\alpha_3}}{\Gamma(\alpha_3)(1-y)^{\alpha_3+1}}y^{\alpha_3-1}\exp\left(-\lambda_3\frac{y}{1-y}\right) \end{aligned} \tag{15}$$

where p_1, p_2 and p_3 are defined in (14). The unit-ISDL distribution is compared with the following five distributions defined on the unit interval $(0 < y < 1)$ as well as unit-3CGamma distribution:

1. Beta distribution

$$f\left(y;\alpha,\beta\right)=\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}y^{\alpha-1}(1-y)^{\beta-1},\quad \alpha>0,\ \beta>0;$$

2. Kumaraswamy distribution

$$f\left(y;\alpha,\beta\right)=\alpha\beta y^{\alpha-1}\big(1-y^{\alpha}\big)^{\beta-1},\quad \alpha>0,\ \beta>0;$$

3. Topp–Leone distribution

$$f\left(y;\theta\right)=\theta(2-2y)\Big(2y-y^{2}\Big)^{\theta-1},\quad \theta>0;$$

4. Unit-Lindley distribution

$$f\left(y;\lambda\right)=\frac{\lambda^{2}}{1+\lambda}(1-y)^{-3}\exp\left(-\frac{\lambda y}{1-y}\right),\quad \lambda>0;$$

5. Unit-Gamma distribution

$$f\left(y;\alpha,\phi\right)=\frac{\alpha^{\phi}}{\Gamma\left(\phi\right)}y^{\alpha-1}\ \ln(-y)^{\phi-1},\quad \alpha>0,\ \phi>0.$$

We use the **R** software to estimate the model parameters. The estimated parameters of the unit-Gamma distribution are used as initial values of the shape and scale parameters of the unit-3CGamma distribution. The MLEs and corresponding standard errors (SEs), Kolmogorov–Smirnov (K–S) statistic and associated p value, Akaike Information Criteria (AIC), Consistent Akaikes Information Criteria (CAIC), Bayesian Information Criteria (BIC) and Hannan-Quinn Information Criteria (HQIC) for all fitted distributions are reported in Tables 3 and 4. The lower the values of these criteria, the better the fitted model to these data.

Table 3 lists the MLEs of the parameters for the fitted models to the monthly water capacity data, corresponding SEs and K-S test results with its p-values. A seen from K-S test results, all fitted models provides sufficient representation for the current data since all p-values are greater than 0.05. However, the unit-3CGamma distribution has the lowest value of K-S statistics.

To decide the best fitted distribution, the model selection criteria, AIC, CAIC, BIC and HQIC, are used and the results are reported in Table 4. The figures in this table reveal that unit-ISDL distribution has the lowest values of the model selection criteria, except AIC value. The unit-3CGamma distribution has the lowest value of AIC. However, there is no big difference between the AIC values of unit-ISDL and unit-3CGamma distributions. More importantly, the unit-ISDL distribution has fewer parameters than unit-3CGamma distribution. When considered the law of parsimony, the proposed distribution can be chosen as the best model for the current data.

Figure 5 displays the fitted densities to the monthly water capacity histogram and some estimated functions of the unit-ISDL distribution. The right panel of the Fig.

Table 3 MLEs and their SEs (on second line) of the fitted models and goodness-of-fit statistics for the monthly water capacity data

Models	Parameter estimations						K-S	p-value
Beta(α, β)	7.3154	2.9098					0.2359	0.1834
	2.3180	0.8754						
Kumaraswamy(α, β)	6.3476	4.4893					0.2209	0.2447
	1.5575	2.0409						
Topp–Leone (θ)	8.6664						0.2549	0.1241
	1.9379							
Unit-Lindley(λ)	0.4957						0.2421	0.1621
	0.0806							
Unit-Gamma(α, ϕ)	8.1070	2.8786					0.2360	0.1833
	2.6543	0.8628						
Unit-3CGamma ($\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$)	15.9181	3.1087	10.3250	3.9256	3.0633	2.6767	0.1183	0.9115
	6.0615	1.8927	16.4659	1.4779	2.2205	4.4009		
Unit-ISDL(λ)	0.7078						0.2050	0.3245
	0.0933							

Table 4 Model selection criteria of the fitted models for the monthly water capacity data

Models	AIC	CAIC	BIC	HQIC
Beta(α, β)	− 21.1239	− 20.4180	− 19.1324	− 20.7351
Kumaraswamy(α, β)	− 22.9494	− 22.2435	− 20.9580	− 22.5607
Topp–Leone (θ)	− 21.1753	− 20.9530	− 20.1795	− 20.9809
Unit-Lindley(λ)	− 25.6543	− 25.4321	− 24.6586	− 25.4600
Unit-Gamma(α, ϕ)	− 21.0561	− 20.3502	− 19.0646	− 20.6673
Unit-3CGamma($\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3$)	− 29.2270	− 22.7654	− 23.2526	− 28.0607
Unit-ISDL(λ)	− 28.8114	− 28.5892	− 27.8157	− 28.6170

5 reveals that the proposed distribution provides qualified fit to the monthly water capacity data.

6.2 Regression modeling

In this section, we compare the beta, simplex, unit-Lindley and unit-ISDL regression models. The **betareg** and **simplexreg** packages of the **R** software are used to estimate the parameters of the beta and simplex regression models, respectively. For details see https://cran.r-project.org/web/packages/betareg/betareg.pdf and https://cran.r-project.org/web/packages/simplexreg/simplexreg.pdf).

The **optim** function of the **R** software is used to obtain the estimated parameters of the unit-Lindley and unit-ISDL regressions.

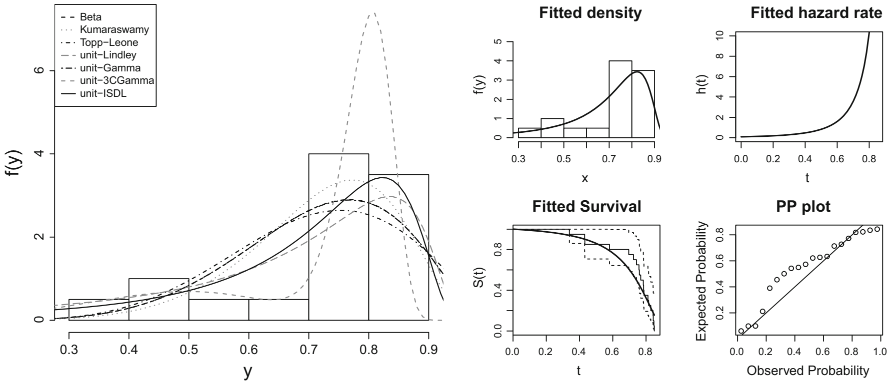

Fig. 5 The estimated pdfs of the fitted distributions to the monthly water capacity data (left-panel) and some estimated functions for the unit-ISDL distribution (right-panel)

The aim of the study is to relate the *long term interest* (LTI) rates of the Organisation for Economic Cooperation and Development (OECD) countries (y) with *foreign direct investment*. These variables are given in the "Appendix". The logit link function is used for the fitted regression models. Hence, the systematic component for μ_i is

$$\text{logit}\,(\mu_i) = \beta_0 + \beta_1 FDI_i.$$

Table 5 gives the MLEs, their standard errors (SEs) and corresponding p-values for the beta, simplex, unit-Lindley and unit-ISDL regressions fitted to the LTI rates. The parameter ψ is the dispersion parameter of the beta and simplex regression models. The values in Table 5 reveal that the parameter β_1 is statistically significant at 5% level for both regression models. We conclude that the FDI stocks explain the LTI rates. In other words, the LTI rates decrease when the FDI stocks increase.

The minimized $-\hat{\ell}$, AIC and BIC values are adopted to select the best fitted regression. Since the unit-ISDL regression has the lowest values of these statistics, it provides a better fit than the beta and simplex regressions for the current data. Moreover, Fig. 6 displays the randomized quantile residuals for the beta, simplex, unit-Lindley and unit-ISDL regressions. It is clear that the plotted points for the unit-ISDL regression are near to the diagonal line.

The plots of the Pearson and Cox–Snell residuals for the unit-ISDL regression are displayed in Fig. 7. They indicate that the Pearson residuals lie between $(-2, 2)$ and that none of the observations can be considered as possible outliers. The Probability–Probability (PP) plot of the Cox–Snell residuals reveals that the unit-ISDL regression provides an adequate fit to these data.

7 Conclusions

A new one-parameter distribution with bounded support is introduced. Some of its structural properties are obtained. The maximum likelihood, bias-corrected maximum

Table 5 MLEs, SEs and corresponding p-values for the fitted models

Parameters	Beta			Simplex			Unit-Lindley			Unit-ISDL		
	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value	Estimate	S.E.	p-value
β_0	-3.7591	0.1686	< 0.001	-3.7986	0.3286	< 0.001	-3.7171	0.2022	< 0.001	-3.7242	0.1998	< 0.001
β_1	-0.3716	0.1722	0.0309	-0.4421	0.1213	< 0.001	-0.5141	0.1750	0.0033	-0.5093	0.1737	0.0033
ψ	71.2000	18.8100	< 0.001	2.5375	0.1988	0.0262	–	–	–	–	–	–
$-\hat{\ell}$	-103.8000			-94.3895			-103.9809			-104.3926		
AIC	-201.6000			-182.7790			-203.9618			-204.7852		
BIC	-197.0209			-178.1199			-200.9091			-201.7325		

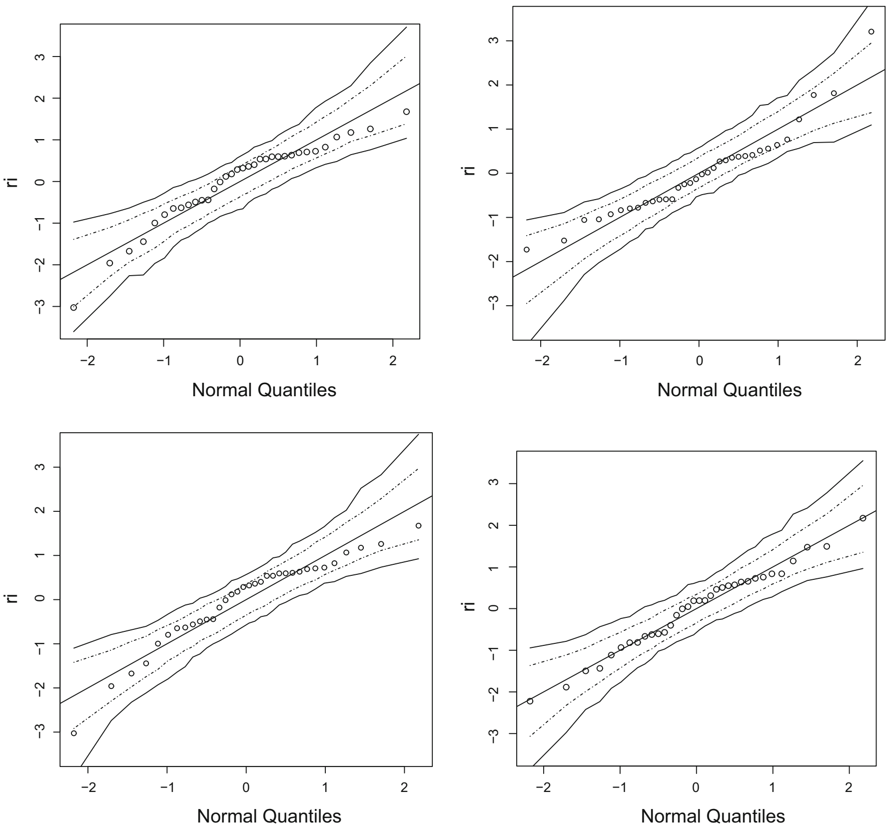

Fig. 6 The Quantile–Quantile (QQ) plot of the randomized quantile residuals for the beta (top-left), simplex (top-right), unit-Lindley (bottom-left) and unit-ISDL (bottom-right) regression models

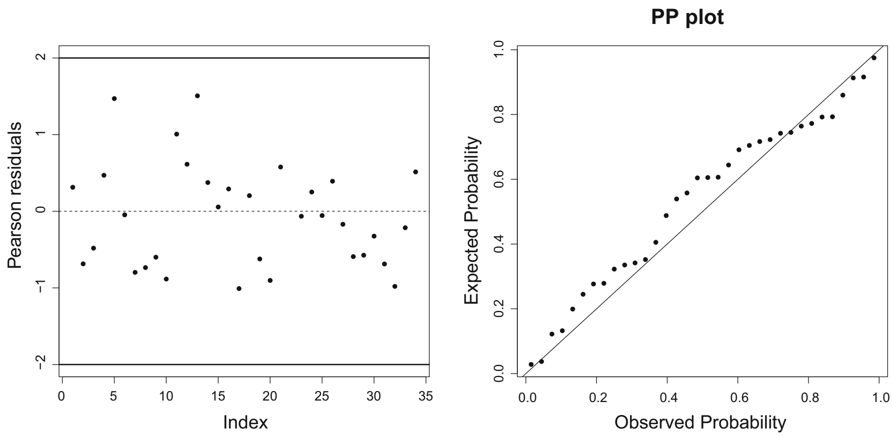

Fig. 7 The Pearson (left) and Cox–Snell residuals (right) plots for the unit-ISDL regression model

likelihood, moments, least squares and weighted least squares methods are discussed for estimating the unknown parameters of the *unit improved second-degree Lindley* (unit-ISDL) distribution via simulation study. A new regression model for the unit response variable is introduced and compared with the beta and simples regression models. Empirical findings reveal that the unit-ISDL regression model provides a better fit than the beta, unit-Lindley and simplex regression models when the response variable is close to the boundaries of the unit interval. An extensive study on residual analysis, leverage and outlier detection is planned as a future work of this study. We hope that the results given in this paper will be useful for practitioners in several areas.

Appendix

The data for Sect. 6.2 are given below:

1. Long term interest (LTI) rate (%): 2.640 0.596 0.680 2.190 4.560 2.140 0.410 0.530 0.750 0.280 4.390 3.390 5.190 0.800 2.160 2.640 0.060 2.549 0.930 0.310 0.540 7.750 0.470 2.810 1.760 3.170 1.760 1.010 0.990 1.318 0.550 0.040 1.374 2.890
2. Foreign Direct Investment (FDI) stocks (Outward) (% GDP): 30.78 57.87 121.52 90.17 45.39 11.08 55.92 51.54 56.31 43.34 11.64 20.85 21.99 276.22 28.81 27.56 30.6 21.02 5.93 7.24 380.1 15.76 305.44 8.94 48.05 5.41 23.68 3.56 14.53 41.9 71.7 162.75 61.86 40.43

References

Aarset AS (1987) How to identify a bathtub hazard rate. IEEE Trans Reliab 36:106–108

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

Altun E (2019) The log-weighted exponential regression model: alternative to the beta regression model. Commun Stat Theory Methods. Forthcoming

Cordeiro GM, de Castro M (2011) A new family of generalized distributions. J Stat Comput Simul 81:883–898

Cox DR, Snell EJ (1968) A general definition of residuals. J R Stat Soc Ser B (Methodol) 30:248–275

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

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

Ghitany ME, Atieh B, Nadarajah S (2008) Lindley distribution and its application. Math Comput Simul 78:493–506

Grassia A (1977) On a family of distributions with argument between 0 and 1 obtained by transformation of the gamma and derived compound distributions. Aust J Stat 19:108–114

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

Karuppusamy S, Balakrishnan V, Sadasivan K (2017) Improved second-degree Lindley distribution and its applications. IOSR J Math 13:1–10

Kieschnick R, McCullough BD (2003) Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. Stat Model 3:193–213

Kumaraswamy P (1980) A generalized probability density function for double-bounded random processes. J Hydrol 46:79–88

Mazucheli J, Menezes AFB, Dey S (2018a) Improved maximum-likelihood estimators for the parameters of the unit-gamma distribution. Commun Stat Theory Methods 47:3767–3778

Mazucheli J, Menezes AF, Dey S (2018b) The unit-Birnbaum-Saunders distribution with applications. Chil J Stat 9:47–57

Mazucheli J, Menezes AFB, Chakraborty S (2019) On the one parameter unit-Lindley distribution and its associated regression model for proportion data. J Appl Stat 46:700–714

Nadar M, Papadopoulos A, Kızılaslan F (2013) Statistical analysis for Kumaraswamy's distribution based on record data. Stat Pap 54:355–369

Nadarajah S, Kotz S (2003) Moments of some J-shaped distributions. J Appl Stat 30:311–317

Topp CW, Leone FC (1955) A family of J-shaped frequency functions. J Am Stat Assoc 50:209–219