Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 51 (3) (2022), 882 – 899
DOI : 10.15672/hujms.932811

Research Article

Testing distributional assumption of unit-Lindley regression model

Deniz Ozonur

[1] *Gazi University, Faculty of Science, Department of Statistics, Teknikokullar, Ankara*

Abstract

This paper proposes smooth goodness of fit test statistic and its components to test the distributional assumption of the unit-Lindley regression model, which is useful for describing data measured between zero and one. Orthonormal polynomials on the unit-Lindley distribution, score functions and Fisher's information matrix are provided for the smooth test. Deviance and Pearson's chi-square tests are also adapted to the unit-Lindley regression model. A parametric bootstrap simulation study is conducted to compare type I errors and powers of the tests under different scenarios. Empirical findings demonstrate that the first smooth component, deviance, and chi-square tests have undesirable behavior for the unit-Lindley regression model. A real data set is analyzed by using the developed tests to show the adequacy of the unit-Lindley regression model. Model selection criteria and residual analysis prove that the unit-Lindley regression model provides a better fit than the Beta and simplex regression models for the real data set.

Mathematics Subject Classification (2020). 62G10, 62F40

Keywords. Chi-square test, deviance test, power of test, smooth test, unit-Lindley distribution, parametric bootstrap

1. Introduction

Regression models describe the relationship between response variable and independent variables. A linear regression model assumes that the error term follows a normal distribution with homogeneous variance and the response variable is a linear function of a set of independent variables.

In applied sciences, some data types can be expressed as proportions, percentages, rates, or fractions. Point rates of football teams [27], body fat percentage of a human body [40], biomass percentages of plant organs [31], cover proportion of a specific plant type [11] can be given as examples of such data. The linear regression model is not appropriate for data with response variables bounded on the unit interval since it may produce fitted values outside the unit interval.

In the literature, there are many regression models to model data on the unit interval. The Beta regression model is the most widely used model for modeling proportions in economics, actuarial, ecology, and environmental sciences [14]. The simplex regression model is used in several fields of science [8, 19]. The ToppLeone and Kumaraswamy

Email address: denizozonur@gazi.edu.tr
Received: 10.05.2021; Accepted: 17.03.2022

models are known models when a response variable is measured continuously on the unit interval [32, 42]. In recent years, alternative regression models have been provided for the bounded response variable. For instance, unit-gamma [26], unit-inverse Gaussian [16], unit-improved second-degree Lindley [3], log-Bilal [4], log-weighted exponential [2], quantile log exponential-power [20] regression models. Further, Mazucheli et al. [24] introduced the unit-Lindley (UL) regression model by applying a transformation to the original Lindley distribution [23]. Although several versions of the Lindley distribution such as generalized Lindley [44], extended Lindley [5], exponential Poisson Lindley [6], power Lindley [15], generalized weighted Lindley [33], inverse weighted Lindley [34] have been proposed, the UL distribution [24] has gained popularity due to its flexible properties. The probability density function (pdf) and cumulative distribution function (cdf) of the UL distribution are, respectively, given by

$$f\left(y;\theta\right)=\frac{\theta^2}{1+\theta}(1-y)^{-3}\exp\left(-\frac{\theta y}{1-y}\right)$$

and

$$F\left(y;\theta\right)=1-\left(1-\frac{\theta y}{\left(1+\theta\right)\left(y-1\right)}\right)\exp\left(-\frac{\theta y}{1-y}\right),$$

where $0<y<1$, $\theta>0$. The corresponding UL regression model is proposed as an alternative to the Beta regression model which is frequently used for modeling unit bounded data [24]. An important difference between the Beta and UL regression models is that while the UL regression model has only mean parameter, the Beta regression model has mean and precision parameters and includes separate submodels for each parameter [39]. Although the UL regression model is restricted in this respect, the UL regression model has some advantages over the Beta regression model. The main advantage of the UL distribution is that its cdf and quantile function can be expressed in closed form [24]. However, the cdf and quantile function of the Beta distribution are not available in closed form. Moreover, the UL regression model can provide a better fit than the Beta regression model for a data set on the unit interval. Therefore, the UL regression model draws attention as an alternative to the Beta regression model.

In the UL regression model, the response variable y_j is assumed to follow the density function

$$f(y_j;\mu_j)=\frac{(1-\mu_j)^2}{\mu_j(1-y_j)^3}\exp\left(-\frac{y_j(1-\mu_j)}{\mu_j(1-y_j)}\right),$$

where $0<y_j<1$ and μ_j is the mean of the response y_j. The moments about the origin of the response variable are presented in Appendix A.

The UL regression model specifies the relationship between the mean of y_j and linear predictor such that

$$g\left(\mu_j\right)=x_j^T\beta,$$

where $\beta=(\beta_1,\beta_2,\ldots,\beta_p)^T$ denotes a $p\times 1$ vector of parameters and $x_j=(x_{j1},x_{j2},\ldots,x_{jp})^T$ is a $p\times 1$ vector of covariates.

It is assumed that the link function $g(.)$ is a strictly monotonic and twice differentiable link function that maps $(0,1)$ into $\mathbb{R}$. Several link functions are avaliable in the literature such as logit, probit, complementary log-log and log-log link functions which ensures that the estimated mean stays within bounds (0,1). Specifically, we consider the logit link function $\text{logit}\left(\mu_j\right)=\log\left(\frac{\mu_j}{1-\mu_j}\right)$, since it provides interpretable regression parameters. It is noteworthy that the UL regression model with link function is similar to generalized linear models [12, 25]. However, the UL distribution of the response variable is not a member of canonical exponential family; therefore, the UL regression model is not a generalized linear model.

Although many regression models have been proposed for bounded response on the unit interval, none of them has proposed new goodness of fit test to test suitability of a model. Before making statistical inferences about model parameters, goodness of fit tests should be used to test the hypothesis assuming that an examined model is suitable or a distribution of a response variable is correct for a data set. Therefore, goodness of fit tests are crucial as they determine suitability of a model.

Our main goal in this study is to propose new tests based on the smooth goodness of fit test for the distributional assumption of the UL regression model. In order to test the distributional assumption of the UL regression model, null and alternative hypotheses are expressed as below:

$$H_0: \text{Response variable follows the UL distribution}$$
$$H_1: \text{Response variable does not follow the UL distribution.}$$

The smooth test was introduced by [28] for uniform distribution and extended to test composite hypothesis for location-scale families by several authors [18,21,22,35]. In order to derive an optimal test in a large sample size, Rayner et al. [36] developed the smooth test as a score test statistic, and this form of the smooth test has been applied to many distributions such as Gamma [10], Logistic [36], zero-inflated poisson [41], Nakagami [30] and Lindley [7] distributions. Rippon [37] adapted the smooth test to generalized linear models. Ozonur et al. [29] compared some goodness of fit tests with the smooth test for Poisson regression model. In this study, the smooth test statistic and its components are derived for the UL regression model. It is known that the deviance and Pearson's chi-square tests are two well-known goodness of fit tests for generalized linear models and the well-known tests are applied to the UL regression model in this study.

The motivations of this study can be given as follows: (i) to construct tests for a different regression model that is not a member of the generalized linear models, (ii) to empirically investigate the applicability of the proposed tests for the UL regression model, (iii) to provide better fits than alternative regression models with responses on the unit interval.

The remainder of the paper is organized as follows. In Section 2, the methodology of the smooth tests is presented for composite null hypothesis. In Section 3, the smooth goodness of fit tests are introduced and the well-known deviance and chi-square tests are adapted to the UL regression model. In Section 4, a parametric bootstrap simulation study is conducted to evaluate the performances of the goodness of fit tests. In Section 5, a real data set is analyzed to test the distributional assumption of a fitted UL regression model and the UL regression model is compared with the Beta and simplex regression models using model selection criteria and residual analysis. In Section 6, some conclusions are offered. Some details are given in three appendices.

2. Methodology

In this section, the basic process of the smooth tests is given for composite null hypothesis. A more detailed description of the methodology can be found in [36].

Let $f(y;\theta)$ be a probability density function, where $\theta = (\theta_1, \theta_2, ..., \theta_p)^T$ is a vector of nuisance parameters. The smooth goodness of test can be derived by embedding the null probability density function in an order k alternative probability density function,

$$f_k(y;\tau,\theta) = C(\tau,\theta)\exp\left\{\sum_{i=1}^{k}\tau_i h_i(y;\theta)\right\} f(y;\theta),$$

where $\tau = (\tau_1, \tau_2, ..., \tau_k)^T$ is a vector of real parameters, $C(\tau,\theta)$ is a normalisation constant guarantees that the integral of alternative density functions is 1 and $h_i(y;\theta)$ is a set of orthonormal functions on $f(y;\theta)$ with $h_0(y;\theta) = 1$. The inner product of the functions h_i and h_j is defined as follows:

$$\langle h_r, h_s \rangle = \int_{-\infty}^{\infty} h_r(y;\theta)\, h_s(y;\theta)\, f(y;\theta)\, dy = \delta_{rs}, \quad (r,s = 0,1,2,3,\ldots)$$

where $\delta_{rs} = 1$ if $r = s$ and 0 otherwise. Let E_0 be expectation with respect to the $f(y;\theta)$. The orthonormality implies that $E_0[h_r(y;\theta)\, h_s(y;\theta)] = \delta_{rs}$. The assumption here is that all expectations exist. Assume that $Y_1, \ldots, Y_n$ is a random sample from the distribution with probability density function $f_k(y;\tau,\theta)$. Testing for $f(y;\theta)$ is equivalent to testing $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$. The log-likelihood function using the alternative density function is given by

$$\log L = n \log C(\tau,\theta) + \sum_{i=1}^{k}\sum_{j=1}^{n} \tau_i h_i(y_j;\theta) + \sum_{j=1}^{n} \log f(y_j;\theta).$$

The partial derivatives of the log-likelihood function with respect to τ and θ are assumed to exist up to second order. The score statistic for testing $H_0 : \tau = 0$ against $H_1 : \tau \neq 0$ is $S = U_\tau^T \Sigma^{-1} U_\tau$. Using the partial derivative of the log-likelihood function with respect to τ, the score vector $U_\tau = U_\tau(\theta)$ has rth element $\{h_r(y_1;\theta) + h_r(y_2;\theta) + \cdots + h_r(y_n;\theta)\}$. The asymptotic covariance matrix Σ of U_τ is given by

$$\Sigma = I_{\tau\tau} - I_{\tau\theta} I_{\theta\theta}^{-1} I_{\theta\tau} = nM,$$

where

$$M = I_k - Cov_0\left[h_r, \frac{\partial \log f}{\partial\theta}\right]\left\{Var_0\left(\frac{\partial \log f}{\partial\theta}\right)\right\}^{-1} Cov_0\left[\frac{\partial \log f}{\partial\theta}, h_r\right],$$

in which I_k is the $k \times k$ identity matrix and zero subscripts indicate evaluations under the null hypothesis [36]. Let $\hat{\theta}$ be the maximum likelihood estimator of θ under the null hypothesis. Therefore, the score statistic is given by

$$S = \left\{\sum_{j=1}^{n} h_r\left(y_j;\hat{\theta}\right)/\sqrt{n}\right\}^T M^{-1}\left(\hat{\theta}\right)\left\{\sum_{j=1}^{n} h_r\left(y_j;\hat{\theta}\right)/\sqrt{n}\right\}.$$

The matrix M is required to be non-singular. When the matrix M is reduced to the $k \times k$ identity matrix I_k, the score statistic takes the form

$$S = \sum_{r=1}^{k} V_r^2\left(\hat{\theta}\right),$$

where $V_r\left(\hat{\theta}\right) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_r\left(y_j, \hat{\theta}\right)$. Under the null hypothesis, the components V_r are asymptotically mutually independent and asymptotically standard normal. Therefore, the score test statistic S is asymptotically χ_k^2 distributed [36].

3. Goodness of fit tests

In this section, the smooth goodness of fit test statistic of order three and its components are firstly provided for the UL regression model. Secondly, the most widely used deviance and chi-square goodness of fit tests are adapted to the UL regression model.

3.1. Smooth test

To construct a smooth test, the UL probability density function $f(y_j;\mu_j)$ is nested in the following smooth density function

$$f_k(y_j;\tau,\mu_j) = C(\tau,\mu_j) \exp\left\{\sum_{i=1}^{k} \tau_i h_i(y_j;\mu_j)\right\} f(y_j;\mu_j), \tag{3.1}$$

where $\tau = (\tau_1, \tau_2, ..., \tau_k)^T$ is a $k \times 1$ vector of parameters, $C(\tau, \mu_j)$ is a normalisation function so that the smooth density function integrates to one, $h_i(y_j; \mu_j)$ are orthonormal polynomials up to order k on the UL distribution and f_k is the smooth alternative of order k. When $\tau = 0$, f_k collapses to original response distribution f in Equation (3.1). Therefore, testing the $H_0 : \tau = 0$ against the $H_1 : \tau \neq 0$ is equivalent to testing goodness of fit of the UL distribution.

Let $\hat{\gamma}_0 = (\tau_0^T = 0^T, \hat{\beta}^T)^T$ denotes the maximum likelihood estimate (mle) of the full parameter vector $\gamma = (\tau^T, \beta^T)^T$ under the H_0 where β is the vector of regression parameters.

Following the smooth test structure for generalized linear models in [37], the smooth test statistic is derived as a score test statistic

$$S = U_\tau^T(\hat{\gamma}_0) B^{-1}(\hat{\gamma}_0) U_\tau(\hat{\gamma}_0), \tag{3.2}$$

where $U_\tau(\hat{\gamma}_0) = \sqrt{n}(V_1, \ldots, V_k)^T$ is the score vector with $V_r = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_r(y_j, \hat{\mu}_j)$ and $B(\hat{\gamma}_0) = \left[I_{\tau\tau} - I_{\tau\beta} I_{\beta\beta}^{-1} I_{\beta\tau} \right]_{\hat{\gamma}_0}$ is the asymptotic covariance matrix of $U_\tau(\hat{\gamma}_0)$ under the H_0. The derivation of the score vector and elements of the asymptotic covariance matrix are given in Appendix B.

As seen from Equation (3.2), the convenient sums of squares structure of the smooth test statistic has been lost. However, it is possible to recover the structure by decomposing the B matrix. By using singular value decomposition, an orthogonal matrix Q and a diagonal matrix $D = \text{diag}(d_1^2, ..., d_k^2)$ can be obtained and the B matrix is partitioned as $B = \left[I_{\tau\tau} - I_{\tau\beta} I_{\beta\beta}^{-1} I_{\beta\tau} \right]_{\hat{\gamma}_0} = QDQ^T$. Therefore, the smooth test statistic is converted into its convenient sums of squares structure as below:

$$\begin{aligned} S &= \sqrt{n}(V_1, V_2, \ldots, V_k) QD^{-1}Q^T (V_1, V_2, \ldots, V_k)^T \sqrt{n} \\ &= \frac{U_1^2}{d_1^2} + \frac{U_2^2}{d_2^2} + \cdots + \frac{U_k^2}{d_k^2}, \end{aligned}$$

where $(U_1, U_2, \ldots, U_k) = \sqrt{n}(V_1, V_2, \ldots, V_k) Q$ and the $\frac{U_r^2}{d_r^2}$ $(r = 1, 2, ..., k)$ is the rth squared component of the smooth test statistic. For simplicity, the component $\frac{U_r^2}{d_r^2}$ is denoted as C_r^2.

Since the smooth test statistic is developed as a score test statistic, it has the asymptotic χ_k^2 distribution under the null hypothesis [36]. The components $C_1^2, C_2^2, ..., C_k^2$ are asymptotically independent and asymptotically follow the χ_1^2 distribution under the null hypothesis [36]. The squared components of the smooth test statistic can be used as goodness of fit test statistics [36].

Note that in order to calculate the smooth test statistic, the orthonormal polynomials up to order k on the UL distribution are required. In this study, we consider $k = 3$, since the sixth central moment of the response variable is needed in the even third orthonormal polynomial [7]. The orthonormal polynomials up to order three on the UL distribution are obtained by applying the Gram-Schmidt orthogonalization process and presented in Appendix C.

3.2. Deviance test

As a type of likelihood ratio test statistic, McCullagh and Nelder [25] defined the deviance test statistic (D) to compare saturated model with interested model for a generalized linear model (GLM).

The deviance statistic is calculated as follows:

$$D = 2\left(l_S\left(\tilde{\beta}\right) - l\left(\hat{\beta}\right)\right),$$

where $l_S\left(\tilde{\beta}\right)$ and $l\left(\hat{\beta}\right)$ are the maximum value of the log-likelihood functions for the saturated and interested models, respectively.

For the saturated UL regression model, the maximum likelihood estimates $\tilde{\mu}_j$ are obtained by differentiating the log-likelihood function $l(\beta)$ with respect to each μ_j and solving the estimating equations. Therefore, the maximum value of the log-likelihood function for the saturated model is $l_S\left(\tilde{\beta}\right) = \sum_{j=1}^{n} l_j(\tilde{\mu}_j)$.

For the interested UL regression model, the fitted values $\hat{\mu}_j$ are calculated by using inverse logit link function $\frac{\exp\left(x_j^T\hat{\beta}\right)}{1+\exp\left(x_j^T\hat{\beta}\right)}$ where $\hat{\beta}$ is the mle of the regression parameter vector β. Thus, the maximum value of the log-likelihood function for the interested model is $l\left(\hat{\beta}\right) = \sum_{j=1}^{n} l_j(\hat{\mu}_j)$. The mle of β is computed by Fisher scoring algorithm using the score function U_β and the information matrix $I_{\beta\beta}$ in Appendix B.

3.3. Chi-square test

McCullagh and Nelder [25] proposed Pearsons chi-square test statistic for generalized linear models and we applied the statistic to UL regression model as follows:

$$\chi^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{\mu}_j)^2}{\hat{\sigma}_j^2},$$

where $\hat{\mu}_j$ and $\hat{\sigma}_j^2 = (1-\hat{\mu}_j)^2\left[\frac{1}{\mu_j}\exp\left((1-\hat{\mu}_j)/\hat{\mu}_j\right)E_1\left((1-\hat{\mu}_j)/\hat{\mu}_j\right) - 1\right]$ are the estimated mean and variance of the response y_j, respectively.

4. Simulation study

In this section, we compare the smooth goodness of fit test and its squared components with the deviance and Pearson chi-square tests according to their simulated type I errors and powers by using R 3.5.1 software.

In the simulation study, we consider two different models. As the first model, we use

$$\text{logit}(\mu_j) = -0.4 + 0.3x_{1j} + 0.8x_{2j}$$

model and as the second model, we examine

$$\text{logit}(\mu_j) = 1.8 - 1.2x_{1j} + 0.6x_{2j}$$

model. In each model, sample sizes are taken as $n = 15, 30, 60$ and the covariate values are obtained as random draws from the uniform distribution $\mathcal{U}(0,1)$.

Although the smooth test statistic and its components have asymptotic distributions under the null hypothesis, unpresented simulation study shows that empirical levels based on their asymptotic null distributions are much smaller than the nominal level for small sample sizes. The simulation code used in the analysis of the asymptotic distributions of smooth test statistics is available from the author upon request. Furthermore, in the generalized linear model context, the deviance and chi-square test statistics asymptotically follow the χ^2_{n-p} distribution under null hypothesis where p is the number of estimated parameters. However, the UL regression model is not a generalized linear model as mentioned in Section 1 and the asymptotic distributions of the deviance and chi-square statistics are not clear for the UL regression model. Therefore, parametric bootstrap procedure is recommended to obtain type I errors and powers of the tests. The procedure is conducted by the following steps:

(i) Calculate the means μ_j of the considered regression model by using the inverse link function.
(ii) Simulate response variables y_j using the UL or alternative distribution with the calculated means μ_j and any other required parameters.
(iii) Fit the UL regression model to the simulated response variables and calculate the test statistic T_m based on the fitted UL regression model.
(iv) Simulate response variables ${y_j}^*$ from the UL distribution using the estimated means in step (iii). Fit the UL regression model to the simulated response variables ${y_j}^*$ and calculate the test statistic T^* from the fitted model.
(v) Repeat step (iv) B number of times and obtain B test statistics ${T_b}^*$ for $b = 1, 2, .., , B$.
(vi) Calculate the bootstrap p-value as $p = \frac{\#({T_b}^* \geq T_m)}{B}$ and reject the null hypothesis if the p-value is smaller than the nominal level.
(vii) Repeat steps (ii)-(vi) M number of times and obtain rejection rate of the null hypothesis.

The parametric bootstrap procedure is performed with M=B=2000 replications and the response variables are generated from the UL distribution with means μ_j in step (ii) to obtain the type I errors of the tests. In order to generate response variables from the UL distribution, random numbers z_j are generated from the Lindley distribution via LindleyR package in R software and the transformation $z_j/(1+z_j)$ is used. Moreover, the Fisher scoring algorithm is implemented to obtain parameter estimates of the UL regression model. The estimates of regression parameters of the beta regression model are used as initial values for the UL regression model and fifty iterations are allowed in the algorithm. Empirical type I errors of the tests are presented in Table 1 for nominal levels $\alpha = 0.10, 0.05, 0.01$.

Table 1. Empirical type I errors of the tests for two models.

Model	n	Level	S	C_1^2	C_2^2	C_3^2	χ^2	D
$\text{logit}(\mu_j) = -0.4 + 0.3x_{1j} + 0.8x_{2j}$	15	0.10	0.096	0.083	0.098	0.099	0.097	0.096
		0.05	0.054	0.046	0.047	0.053	0.049	0.044
		0.01	0.009	0.013	0.011	0.014	0.007	0.006
	30	0.10	0.098	0.078	0.097	0.091	0.099	0.101
		0.05	0.055	0.058	0.051	0.050	0.044	0.046
		0.01	0.011	0.008	0.016	0.010	0.007	0.010
	60	0.10	0.101	0.106	0.109	0.106	0.097	0.103
		0.05	0.057	0.055	0.064	0.053	0.046	0.046
		0.01	0.015	0.015	0.012	0.016	0.008	0.009
$\text{logit}(\mu_j) = 1.8 - 1.2x_{1j} + 0.6x_{2j}$	15	0.10	0.088	0.081	0.099	0.099	0.092	0.093
		0.05	0.051	0.057	0.038	0.060	0.056	0.042
		0.01	0.010	0.010	0.013	0.011	0.006	0.015
	30	0.10	0.093	0.095	0.095	0.096	0.091	0.099
		0.05	0.043	0.047	0.046	0.043	0.049	0.049
		0.01	0.008	0.011	0.006	0.007	0.010	0.007
	60	0.10	0.096	0.091	0.107	0.100	0.086	0.094
		0.05	0.045	0.040	0.066	0.044	0.040	0.048
		0.01	0.013	0.011	0.011	0.008	0.006	0.011

As seen from Table 1, the empirical type I errors of the tests are close to the nominal levels irrespective of the models and sample sizes.

In order to evaluate performances of the tests in terms of power, we consider Beta distribution $Beta(\mu_j, \phi)$ with means μ_j and precision parameters $\phi = 3, 5, 8, 10, 15, 20, 50, 100$ and simplex distribution $Simplex(\mu_j, \sigma)$ with means μ_j and dispersion parameters $\sigma = 0.05, 1, 1.5, 3, 3.5, 4, 5, 10$ as alternative distributions with the probability density functions in Table 2. For the empirical powers of the tests, the nominal level is considered as 0.05 and the power results are summarized in Table 3 and Table 4. The simulation codes based on the parametric bootstrap method are available from the author on request.

Table 2. The probability density functions of the alternative distributions in the simulation study.

Alternative Distribution	Probability Density Function	Parameters
Beta	$f(y_j;\mu_j,\phi) = \frac{\Gamma(\phi) y_j^{\mu_j\phi-1}(1-y_j)^{(1-\mu_j)\phi-1}}{\Gamma(\mu_j\phi)\Gamma((1-\mu_j)\phi)}$	$0<\mu_j<1, \phi>0$
Simplex	$f(y_j;\mu_j,\sigma) = \frac{\exp\left(\frac{-1}{2\sigma^2}\frac{(y_j-\mu_j)^2}{y_j(1-y_j)\mu_j^2\left(1-\mu_j^2\right)}\right)}{\left(2\pi\sigma^2(y_j(1-y_j))^3\right)^{1/2}}$	$0<\mu_j<1, \sigma>0$

Table 3 presents the empirical powers of the tests under the $Beta(\mu_j, \phi)$ distribution with the same means μ_j as in the UL distribution. The power results are obtained for nominal level $\alpha = 0.05$, sample sizes $n = 15, 30, 60$ and precision parameters $\phi = 3, 5, 8, 10, 15, 20, 50, 100$. As the value of ϕ increases, the powers of C_1^2, χ^2, and D tests dramatically decrease and converge to 0.000 regardless of the models and sample sizes. This situation shows the problems of the C_1^2, χ^2, and D tests when ϕ is large. Therefore, the C_1^2, χ^2, and D tests should not be used as goodness of fit tests for the UL regression model. When S, C_2^2, and C_3^2 tests are compared, their powers first decrease and then increase as the value of ϕ increases. Although the power of the S test is higher than that of the C_2^2 and C_3^2 tests for $\phi = 3, 5$, the C_2^2 test can show better performance among them for larger ϕ values and small sample size (see Table 3, $n = 15, 30$, $\phi = 15, 20, 50$). The C_3^2 test can outperform the S and C_2^2 tests (see Table 3, Model 2, $\phi = 8, 10$). As the value of ϕ increases, the powers of S and C_2^2 tests converge to 1.000 and the C_3^2 test becomes the least powerful test among the S, C_2^2, and C_3^2 tests.

Table 4 presents empirical powers of the tests under $Simplex(\mu_j, \sigma)$ distribution with the same means μ_j as in the UL distribution. The power results are obtained for $\alpha = 0.05, n = 15, 30, 60$ and $\sigma = 0.05, 1, 1.5, 3, 3.5, 4, 5, 10$. As expected, the powers of all tests increase as the sample size increases. The powers of C_1^2, χ^2, and D tests are close to 0.0000 for small σ values regardless of models and sample sizes. Although their powers increase as the value of σ and sample size increase, these tests give no rejection for small σ values. It means that these tests are problematic; therefore, the C_1^2, χ^2, and D tests are not applicable for the UL regression model. As the value of σ increases, the powers of S, C_2^2, and C_3^2 tests first decrease and then increase up to 1.000. The C_3^2 test is the least powerful test among the three tests. Although the C_2^2 test provides better results than the S test for some small σ values (see Table 4, $\sigma = 1, 1.5$), the S test becomes the most powerful test among S, C_2^2, and C_3^2 tests as the value of σ increases.

Table 3. Empirical powers of the tests under the $Beta(\mu_j, \phi)$ distribution for $\alpha = 0.05$, $n = 15, 30, 60$ and $\phi = 3$, 5, 8, 10, 15, 20, 50, 100.

Model	n	ϕ	S	C_1^2	C_2^2	C_3^2	χ^2	D
$\text{logit}(\mu_j) = -0.4 + 0.3x_{1j} + 0.8x_{2j}$	15	3	0.4940	0.1610	0.3490	0.4690	0.4650	0.4840
		5	0.1840	0.0870	0.0980	0.1780	0.0790	0.0900
		8	0.2200	0.0700	0.2540	0.1260	0.0020	0.0040
		10	0.3140	0.0270	0.4100	0.0930	0.0000	0.0000
		15	0.6540	0.0010	0.8180	0.0650	0.0000	0.0000
		20	0.9010	0.0010	0.9510	0.1042	0.0000	0.0000
		50	0.9990	0.0000	1.0000	0.2780	0.0000	0.0000
		100	1.0000	0.0000	1.0000	0.7150	0.0000	0.0000
	30	3	0.8030	0.1830	0.6790	0.7440	0.7130	0.8120
		5	0.4510	0.3170	0.1640	0.3750	0.0800	0.1690
		8	0.4300	0.2740	0.3080	0.1620	0.0000	0.0030
		10	0.6680	0.1020	0.6350	0.1030	0.0000	0.0000
		15	0.9240	0.0360	0.9590	0.0740	0.0000	0.0000
		20	0.9780	0.0040	0.9990	0.1490	0.0000	0.0000
		50	1.0000	0.0000	1.0000	0.7480	0.0000	0.0000
		100	1.0000	0.0000	1.0000	0.9840	0.0000	0.0000
	60	3	0.9730	0.2400	0.9470	0.9330	0.8670	0.9540
		5	0.7990	0.7160	0.2730	0.6730	0.0800	0.2690
		8	0.8820	0.7530	0.4800	0.3240	0.0000	0.0030
		10	0.9900	0.6390	0.8700	0.2320	0.0000	0.0000
		15	1.0000	0.1920	1.0000	0.1720	0.0000	0.0000
		20	1.0000	0.0330	1.0000	0.2830	0.0000	0.0000
		50	1.0000	0.0000	1.0000	0.9960	0.0000	0.0000
		100	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000
$\text{logit}(\mu_j) = 1.8 - 1.2x_{1j} + 0.6x_{2j}$	15	3	0.9390	0.4950	0.8950	0.8260	0.9170	0.9670
		5	0.7680	0.1520	0.5750	0.7290	0.6160	0.7960
		8	0.4220	0.0520	0.1850	0.4800	0.2120	0.4280
		10	0.2450	0.0110	0.1580	0.3220	0.0790	0.2080
		15	0.1290	0.0050	0.3380	0.1370	0.0060	0.0500
		20	0.1320	0.0000	0.6260	0.0530	0.0010	0.0100
		50	0.6060	0.0000	0.9900	0.1260	0.0000	0.0000
		100	0.9900	0.0000	1.0000	0.2290	0.0000	0.0000
	30	3	0.9990	0.6980	0.9890	0.9520	0.9970	0.9990
		5	0.9770	0.2330	0.8300	0.9700	0.8320	0.9730
		8	0.7910	0.1500	0.2300	0.8300	0.2550	0.6910
		10	0.6310	0.0400	0.1720	0.6940	0.0970	0.4320
		15	0.4300	0.0020	0.5890	0.3210	0.0090	0.0790
		20	0.4240	0.0000	0.9080	0.1160	0.0010	0.0150
		50	0.9890	0.0000	1.0000	0.2930	0.0000	0.0000
		100	1.0000	0.0000	1.0000	0.8650	0.0000	0.0000
	60	3	1.0000	0.8860	0.9975	0.9830	1.0000	1.0000
		5	1.0000	0.5070	0.9760	1.0000	0.9650	0.9995
		8	0.9825	0.6130	0.3315	0.9885	0.3760	0.8990
		10	0.9480	0.2170	0.1660	0.9400	0.0795	0.6370
		15	0.9205	0.0050	0.8650	0.6060	0.0015	0.0925
		20	0.9800	0.0010	0.9970	0.2020	0.0000	0.0090
		50	1.0000	0.0000	1.0000	0.7543	0.0000	0.0000
		100	1.0000	0.0000	1.0000	0.9990	0.0000	0.0000

Table 4. Empirical powers of the tests under the $Simplex(\mu_j, \sigma)$ distribution for $\alpha = 0.05$, $n = 15, 30, 60$ and $\sigma = 0.05$, 1, 1.5, 3, 3.5, 4, 5, 10.

Model	n	σ	S	C_1^2	C_2^2	C_3^2	χ^2	D
$\text{logit}(\mu_j) = -0.4 + 0.3x_{1j} + 0.8x_{2j}$	15	0.05	1.0000	0.0000	1.0000	0.9870	0.0000	0.0000
		1	0.8240	0.0000	0.9230	0.5237	0.0000	0.0000
		1.5	0.2730	0.0280	0.3420	0.1032	0.0000	0.0000
		3	0.3010	0.0680	0.2700	0.1780	0.3290	0.3100
		3.5	0.5520	0.1620	0.4800	0.2640	0.6280	0.5980
		4	0.7000	0.2900	0.6120	0.3770	0.7720	0.7360
		5	0.8960	0.6170	0.7900	0.5640	0.9360	0.9250
		10	0.9960	0.9820	0.9566	0.9150	0.9990	0.9990
	30	0.05	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000
		1	0.9970	0.0050	0.9980	0.8240	0.0000	0.0000
		1.5	0.4030	0.1470	0.4400	0.2431	0.0000	0.0000
		3	0.6450	0.0880	0.6290	0.3340	0.5560	0.6110
		3.5	0.8630	0.1360	0.8640	0.4850	0.8720	0.8930
		4	0.9460	0.4350	0.9380	0.6320	0.9710	0.9740
		5	0.9990	0.8540	0.9920	0.8820	1.0000	0.9980
		10	1.0000	1.0000	0.9917	1.0000	1.0000	1.0000
	60	0.05	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000
		1	1.0000	0.0480	1.0000	0.9890	0.0000	0.0000
		1.5	0.9230	0.5770	0.7230	0.4570	0.0000	0.0000
		3	0.9470	0.1330	0.9300	0.5450	0.7760	0.8770
		3.5	0.9960	0.1750	0.9950	0.7610	0.9920	0.9960
		4	0.9990	0.6190	0.9990	0.9140	0.9990	0.9990
		5	1.0000	0.9820	1.0000	0.9930	1.0000	1.0000
		10	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
$\text{logit}(\mu_j) = 1.8 - 1.2x_{1j} + 0.6x_{2j}$	15	0.05	1.0000	0.0000	1.0000	0.8610	0.0000	0.0000
		1	0.6660	0.0000	0.9610	0.2150	0.0000	0.0000
		1.5	0.0990	0.0000	0.5400	0.0690	0.0000	0.0000
		3	0.1090	0.0010	0.0900	0.0890	0.0350	0.1270
		3.5	0.2070	0.0070	0.1950	0.1490	0.1210	0.3160
		4	0.3630	0.0300	0.3600	0.2140	0.2950	0.5170
		5	0.6580	0.1480	0.6330	0.2910	0.6200	0.7880
		10	0.9750	0.8420	0.9661	0.6600	0.9720	0.9900
	30	0.05	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000
		1	0.9990	0.0000	1.0000	0.6810	0.0000	0.0000
		1.5	0.5210	0.0000	0.9050	0.1100	0.0000	0.0000
		3	0.3130	0.0090	0.2570	0.1850	0.0240	0.2670
		3.5	0.4220	0.0100	0.3240	0.3160	0.1590	0.5790
		4	0.6780	0.0380	0.6340	0.4490	0.4640	0.8370
		5	0.9300	0.2290	0.9020	0.5960	0.8830	0.9780
		10	1.0000	0.9910	0.9990	0.8510	1.0000	1.0000
	60	0.05	1.0000	0.0000	1.0000	1.0000	0.0000	0.0000
		1	1.0000	0.0000	1.0000	0.9940	0.0000	0.0000
		1.5	0.9570	0.0000	0.9980	0.3450	0.0000	0.0000
		3	0.4260	0.0540	0.3580	0.2580	0.0140	0.4530
		3.5	0.7790	0.0700	0.7140	0.5880	0.2350	0.8740
		4	0.9510	0.0770	0.8860	0.8190	0.7030	0.9840
		5	0.9990	0.3650	0.9950	0.9160	0.9940	1.0000
		10	1.0000	1.0000	1.0000	0.9610	1.0000	1.0000

5. Real data analysis

In this section, we consider the data set on the access of people living in households with inadequate water supply and sewage in the state of Maranhão in Brazil. This data set is the part of the original data analyzed by [24]. We are interested in modeling the proportion of households with inadequate water supply and sewage (y) as a function of the logarithm of the income (x) in the state. The data set and computer code used in this section are available at https://avesis.gazi.edu.tr/denizozonur/dokumanlar. We consider the following regression model:

$$\text{logit}\left(\mu_j\right) = \beta_0 + \beta_1 \log\left(x_j\right), \quad j = 1, \ldots, 55.$$

Since the C_1^2, χ^2, and D tests are problematic for the UL regression model, the S, C_2^2, and C_3^2 tests are applied to examine the adequacy of the distributional assumption in the fitted UL regression model. We obtain the parametric bootstrap p-values of the test statistics for the real data set and the parametric bootstrap p-values are calculated by the following steps:

(i) Estimate the means μ_j and calculate the observed test statistic value, T_0, from the fitted UL regression model.
(ii) Generate response variables y_j^* from the UL distribution with estimated means $\hat{\mu}_j$.
(iii) For the generated response variables, fit the UL regression model using the covariates of the original data and recalculate T_0 statistic from the fitted model and call it T_0^*.
(iv) Repeat steps (ii) and (iii) a large number of R times and obtain R test statistics T_{0r}^* for $r = 1, 2, ..., R$.
(v) The bootstrap p value is calculated as $p = \frac{\#\left(T_{0r}^* \geq T_0\right)}{R}$.
(vi) Reject the null hypothesis if the bootstrap p-value is smaller than the nominal level of tests.

In the real data application, we considered the nominal level $\alpha = 0.05$ and the replication number $R = 5000$. According to the parametric bootstrap procedure, we obtain p-values of the test statistics S, C_2^2, and C_3^2 as 0.3504, 0.1192, and 0.7042, respectively. All the considered tests suggest that the unit-Lindley response distribution adequately fits the real data set.

For comparison purposes, Beta and simplex regression models are also fitted to the real data set, and Table 5 presents the maximum likelihood estimates, standard errors and $p-$values for the fitted UL, Beta, and simplex regression models. In order to estimate the parameters of the Beta and simplex regression models, the betareg [45] and VGAM [43] packages in R software are used, respectively. In Table 5, the parameter δ represents the precision parameter of the Beta regression model and the logarithm of the σ parameter of the simplex regression model, respectively.

Table 5. Summary of the fitted regression models.

	Unit-Lindley			Beta			Simplex		
Parameter	MLE	SE	$p-$value	MLE	SE	$p-$value	MLE	SE	$p-$value
β_0	5.3678	1.8467	0.004	3.8487	1.7687	0.0295	4.9151	1.8643	0.0083
β_1	-1.1588	0.3432	<0.001	-0.8636	0.3291	0.0086	-1.0878	0.3443	0.0015
δ				6.3710	1.1510	< 0.001	0.9312	0.0953	<0.001
AIC	-47.1328			-42.0342			-33.8521		
BIC	-43.1181			-36.0122			-27.8301		
$HQIC$	-45.5803			-39.7055			-31.5233		

As seen from Table 5, the parameter β_1 is statistically significant at 5 % level for all considered regression models. It is concluded that there is a negative relationship between the mean response (proportion of households with inadequate water supply and sewage) and the logarithm of the income in the state. In order to determine the best-fitted regression model among the three regression models, we also calculate Akaikes Information Criterion (AIC) [1]), Bayesian Information criterion (BIC) [38] and Hannan-Quinn Information criterion ($HQIC$) [17]. These criteria are calculated as follows:

$$AIC = 2p - 2\log\hat{L}, \quad BIC = p\log(n) - 2\log\hat{L}, \quad HQIC = 2p\log(\log(n)) - 2\log\hat{L},$$

where n is the number of observations, p is the number of parameters, and $\hat{L}$ is the maximum value of the likelihood function for a fitted model. The model achieving the lowest value of the selected criterion is chosen as the best model. Table 5 shows that the UL regression model provides the best fit for the data set, since it has the lowest values of the $AIC, BIC, HQIC$ statistics.

Moreover, residuals are widely used to assess the suitability of fitted statistical models. In order to evaluate the fitted regression models, three commonly used residuals such as randomized quantile [13], Cox-Snell [9], and Pearson residuals are examined. The randomized quantile residuals are given by

$$r_j = \Phi^{-1}(F(y_j)), \quad j = 1, \ldots, n$$

where $F(.)$ is the cdf of a response distribution and Φ is the cdf of the standard normal distribution. The randomized quantile residuals follow a standard normal distribution if the fitted model is valid.

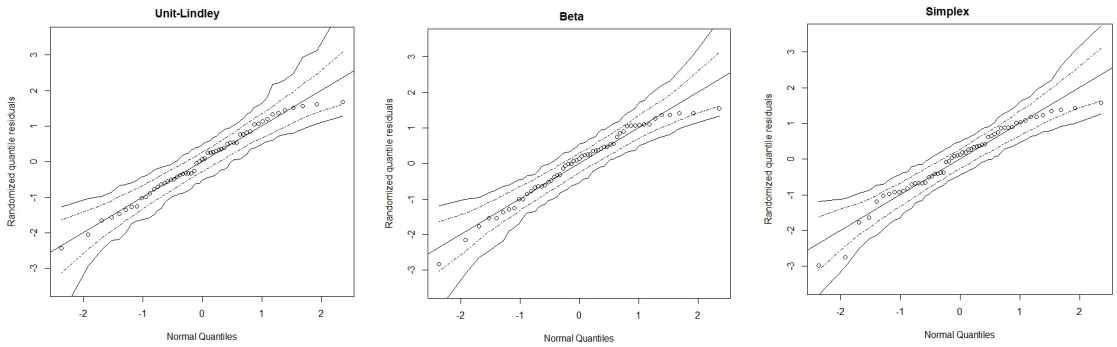

Figure 1. The quantile-quantile plots of the randomized quantile residuals for the considered models.

Figure 1 presents the quantile-quantile plots of the randomized quantile residuals with simulated envelopes for the UL, Beta, and simplex regression models. As seen from Figure 1, the UL regression model provides a better fit than the Beta and simplex regression models, since the residuals are much closer to the diagonal line in the UL regression model.

For the UL regression model, the Cox-Snell and Pearson residuals are also evaluated. The Cox-Snell residuals are calculated as follows:

$$e_j = -\log(1 - F(y_j)), \quad j = 1, \ldots, n$$

where $F(.)$ is the cdf of the UL distribution. The Cox-Snell residuals follow a standard exponential distribution distribution if the fitted model is appropriate.

The Pearson residuals are defined as

$$r_j^* = \frac{y_j - \hat{\mu}_j}{\sqrt{\hat{\sigma}_j^2}},$$

where $\hat{\sigma}_j^2 = (1-\hat{\mu}_j)^2 \left[\frac{1}{\mu_j} \exp\left((1-\hat{\mu}_j)/\hat{\mu}_j\right) E_1\left((1-\hat{\mu}_j)/\hat{\mu}_j\right) - 1\right]$ is the estimated variance of the UL response variable. The scatter plot of the Pearson residuals against the index of the observations should not show a detectable pattern and the residuals outside the interval $(-2, 2)$ are detected as potential outliers.

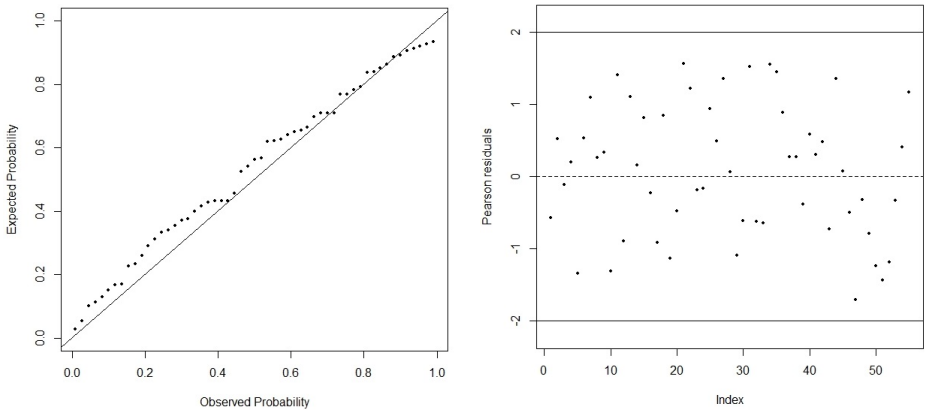

Figure 2. Two diagnostic plots for the UL regression model.

Figure 2 displays the Cox-Snell (left) and Pearson (right) residuals for the UL regression model and verifies that the UL regression model is suitable for the real data set.

6. Conclusions

This paper focuses on the unit-Lindley regression model assuming that the response variable follows the unit-Lindley distribution. The smooth goodness of fit test and its components are developed to test distributional assumption of the UL regression model. The smooth test statistic is converted into its convenient sums of squares structure by decomposing the variance-covariance matrix. In order to obtain the smooth test statistic, the moments about the origin of the response variable, the orthonormal polynomials on the unit-Lindley distribution, the score functions and Fisher's information matrix are derived. The most popular deviance and chi-square goodness of tests are adapted to unit-Lindley regression model. The maximum likelihood estimates of the regression parameters are obtained by using Fisher scoring algorithm. A parametric bootstrap simulation study is performed to compare proposed tests in terms of their type I errors and powers. The simulation study shows that empirical type I errors of all the tests are always close to the nominal levels. Empirical power results are obtained for the Beta and simplex alternative distributions. The powers of C_1^2, χ^2, and D tests are 0.0000 when the precision of the Beta distribution is large or the dispersion of the simplex distribution is small. These results demonstrate that C_1^2, χ^2, and D tests have undesirable behaviors. Therefore, they should not be used as the goodness of fit tests for the unit-Lindley regression model. For the Beta distribution, the smooth test S is the most powerful test among the S, C_2^2, and C_3^2 tests for small ϕ values. As the value of ϕ increases, the power of C_2^2 can provide better power results than that of S and C_3^2 tests for small sample sizes. The powers of S and C_2^2 tests converge to 1.0000 for large ϕ values regardless of the sample sizes and models. For the simplex distribution, the C_3^2 test is the least powerful test among the S, C_2^2, and C_3^2 tests and C_2^2 test can outperform the S test for small σ values. However, the S test gives higher power results than C_2^2 test as the value of σ increases. New goodness of fit tests demonstrate that unit-Lindley regression model adequately fits a real data set. The superiority of the unit-Lindley regression model over the Beta and simplex regression models is indicated by model selection criteria and residual analysis.

References

[1] H. Akaike, *A new look at the statistical model identification*, IEEE Trans. Automat. Contr. **19** (6), 716-723, 1974.

[2] E. Altun, *The log-weighted exponential regression model: alternative to the beta regression model*, Comm. Statist. Theory Methods **50** (10), 2306-2321, 2021.

[3] E. Altun and G.M. Cordeiro, *The unit-improved second-degree Lindley distribution: inference and regression modeling*, Comput. Statist. **35** (1), 259-279, 2020.

[4] E. Altun, M. El-Morshedy and M.S. Eliwa, *A new regression model for bounded response variable: an alternative to the beta and unit-Lindley regression models*, PloS one **16** (1), e0245627, 2021.

[5] H.S. Bakouch, B.M. Al-Zahrani, A.A. Al-Shomrani, V.A. Marchi and F. Louzada, *An extended Lindley distribution*, J. Korean Statist. Soc. **41** (1), 75-85, 2012.

[6] W. Barreto-Souza and H.S. Bakouch, *A new lifetime model with decreasing failure rate*, Stats. **47** (2), 465-476, 2013.

[7] D.J. Best and J.C.W. Rayner, *Smooth tests of fit for the Lindley distribution*, Stats **1** (1), 92-97, 2018.

[8] J.M. Carrasco and N. Reid, *Simplex regression models with measurement error*, Comm. Statist. Simulation Comput. **50** (11), 3420-3435, 2021.

[9] D.R. Cox and E.J. Snell, *A general definition of residuals*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **30** (2), 248-275, 1968.

[10] B. De Boeck, O. Thas, J.C.W. Rayner and D.J. Best, *Smooth tests for the gamma distribution*, J. Stat. Comput. Simul. **81** (7), 843-855, 2011.

[11] R.S. Defries, M.C. Hansen, J.R. Townshend, A.C. Janetos and T.R. Loveland, *A new global 1km dataset of percentage tree cover derived from remote sensing*, Glob. Change Biol. **6** (2), 247-254, 2000.

[12] A.J. Dobson and A.G. Barnett, *An Introduction to Generalized Linear Models*, Chapman and Hall/CRC, 2008.

[13] P.K. Dunn and G.K. Smyth, *Randomized quantile residuals*, J. Comput. Graph. Statist. **5** (3), 236-244, 1996.

[14] S. Ferrari and F. Cribari-Neto, *Beta regression for modelling rates and proportions*, J. Appl. Stat. **31** (7), 799-815, 2004.

[15] M.E. Ghitany, D.K. Al-Mutairi, N. Balakrishnan and L.J. Al-Enezi, *Power Lindley distribution and associated inference*, Comput. Statist. Data Anal. **64**, 20-33, 2013.

[16] M.E. Ghitany, J. Mazucheli, A.F.B. Menezes and F. Alqallaf, *The unit-inverse Gaussian distribution: a new alternative to two-parameter distributions on the unit interval*, Comm. Statist. Theory Methods **48** (14), 3423-3438, 2019.

[17] E.J. Hannan and B.G. Quinn, *The determination of the order of an autoregression*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **41** (2), 190-195, 1979.

[18] C.M. Jarque and A.K. Bera, *A test for normality of observations and regression residuals*, Int. Stat. Rev. **55** (2), 163-177, 1987.

[19] R. Kieschnick and B.D. McCullough, *Regression analysis of variates observed on (0, 1): percentages, proportions and fractions*, Stat. Model. **3** (3), 193-213, 2003.

[20] M.Ç. Korkmaz, E. Altun, M. Alizadeh and M. El-Morshedy, *The log exponential-power distribution: properties, estimations and quantile regression model*, Mathematics **9** (21), 1-19, 2021.

[21] J.A. Koziol, *Assessing multivariate normality: a compendium*, Comm. Statist. Theory Methods **15** (9), 2763-2783, 1986.

[22] J.A. Koziol, *An alternative formulation of Neyman's smooth goodness of fit tests under composite alternatives*, Metrika **34** (1), 17-24, 1987.

[23] D.V. Lindley, *Fiducial distributions and Bayes' theorem*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **20** (1), 102-107, 1958.

[24] J. Mazucheli, A.F.B. Menezes and S. Chakraborty, *On the one parameter unit-Lindley distribution and its associated regression model for proportion data*, J. Appl. Stat. **46** (4), 700-714, 2019.
[25] P. McCullagh and J.A. Nelder, *Generalized Linear Models*, Chapman and Hall, 1989.
[26] A.M. Mousa, A.A. El-Sheikh and M.A. Abdel-Fattah, *A gamma regression for bounded continuous variables*, Adv. Appl. Stat. **49** (4), 305-326, 2016.
[27] L.R. Nakamura, P.H. Cerqueira, T.G. Ramires, R.R. Pescim, R.A. Rigby and D.M. Stasinopoulos, *A new continuous distribution on the unit interval applied to modelling the points ratio of football teams*, J. Appl. Stat. **46** (3), 416-431, 2019.
[28] J. Neyman, *Smooth test for goodness of fit*, Scand. Actuar. J. **1937** (3-4), 149-199, 1937.
[29] D. Ozonur, H.T.K. Akdur and H. Bayrak, *Comparisons of tests of distributional assumption in Poisson regression model*, Comm. Statist. Simulation Comput. **46** (8), 6197-6207, 2017.
[30] D. Özonur, F. Gökpnar, E. Gökpnar and H. Bayrak, *Goodness of fit tests for Nakagami distribution based on smooth tests*, Comm. Statist. Theory Methods **45** (7), 1876-1886, 2016.
[31] H. Poorter and L. Sack, *Pitfalls and possibilities in the analysis of biomass allocation patterns in plants*, Front. Plant Sci. **3** (259), 1-10, 2012.
[32] G. Pumi, C. Rauber and F.M. Bayer, *Kumaraswamy regression model with Aranda-Ordaz link function*, Test **29**, 1051-1071, 2020.
[33] P.L. Ramos and F. Louzada, *The generalized weighted Lindley distribution: properties, estimation, and applications*, Cogent Math. **3** (1), 1-18, 2016.
[34] P.L. Ramos, F. Louzada, T.K. Shimizu and A.O. Luiz, *The inverse weighted Lindley distribution: properties, estimation and an application on a failure time data*, Comm. Statist. Theory Methods **48** (10), 2372-2389, 2019.
[35] J.C.W. Rayner and D.J. Best, *Neyman-type smooth tests for location-scale families*, Biometrika **73** (2), 437-446, 1986.
[36] J.C.W. Rayner, O. Thas and D.J. Best, *Smooth Tests of Goodness of Fit: Using R*, John Wiley and Sons, 2009.
[37] P. Rippon, *Application of smooth tests of goodness of fit to generalized linear models*, PhD thesis, University of Newcastle, 2013.
[38] G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (2), 461-464, 1978.
[39] M. Smithson and J. Verkuilen, *A better lemon-squeezer? Maximum likelihood regression with beta-distribuited dependent variables*, Psycholog. Meth. **11** (1), 54-71, 2006.
[40] M.G. Swainson, A.M. Batterham, C. Tsakirides, Z.H. Rutherford and K. Hind, *Prediction of whole-body fat percentage and visceral adipose tissue mass from five anthropometric variables*, PloS one **12** (5), 1-12, 2017.
[41] O. Thas and J.C.W. Rayner, *Smooth tests for the zero-inflated Poisson distribution*, Biometrics **61** (3), 808-815, 2005.
[42] C.W. Topp and F.C. Leone, *A family of J-shaped frequency functions*, J. Amer. Statist. Assoc. **50** (269), 209-219, 1955.
[43] T.W. Yee, *The VGAM package for categorical data analysis*, J. Stat. Softw. **32** (10), 1-34, 2010.
[44] H. Zakerzadeh and A. Dolati, *Generalized Lindley distribution*, J. Math. Ext. **3** (2), 1-17, 2009.
[45] A. Zeileis, F. Cribari-Neto, B. Grün and I. Kos-midis, *Beta regression in R*, J. Statist. Softw. **34** (2), 1-24, 2010.

Appendix A. Moments about the origin of the response variable

Let $\mu'_r = E(y^r)$ denotes rth moment about the origin of the response variable following the UL distribution. The moments are derived for $r = 2, 3, 4, 5, 6$ as follows:

$$\mu'_2 = 2\mu - 1 + \left\{(1-\mu)^2/\mu\ \right\}\exp\left((1-\mu)/\mu\ \right)E_1\left((1-\mu)/\mu\ \right)$$

$$\mu'_3 = \left(2\mu^2 - 1\right)/\mu\ + \left\{3(1-\mu)^2/\mu + (1-\mu)^3/\mu^2\ \right\}\exp\left((1-\mu)/\mu\ \right)E_1\left((1-\mu)/\mu\ \right)$$

$$\mu'_4 = (1-\mu)/2\ \left[\left(-(1-\mu)^2/\mu^{\ 2} - 7(1-\mu)/\mu + 2\mu/(1-\mu) - 6\ \right)\right.$$
$$\left. + \left\{\left((1-\mu)^3/\mu^3 + 8(1-\mu)^2/\mu^2 + 12(1-\mu)/\mu\ \right)\right\}\exp\left((1-\mu)/\mu\ \right)E_1\left((1-\mu)/\mu\ \right)\right]$$

$$\mu'_5 = (1-\mu)^2/\mu\left[\mu/(1-\mu)^2 - \ 5\mu/(1-\mu) - 7.5 - 2.5\mu/(1-\mu) - \ 1/3\right.$$
$$+ \left(-2\mu^2 + 3\mu - 1\right)/6\ + \left\{(1-\mu)^3/6\mu^3\ + 2.5(1-\mu)^2/\mu^2\right.$$
$$\left.\left. +\ 10(1-\mu)/\mu + 10\ \right\}\exp\left((1-\mu)/\mu\ \right)E_1\left((1-\mu)/\mu\ \right)\right]$$

$$\mu'_6 = (1-\mu)^2/\mu\left[\mu/(1-\mu)^2 - \ 6\mu/(1-\mu) - 14.25 - 12.5(1-\mu)/\mu - 23(1-\mu)^2/24\mu^2\right.$$
$$+ 11\,(1-\mu)/12\mu\ - (1-\mu)^3/24\mu^3\ + \left\{(1-\mu)^4/24\mu^4\ +\ (1-\mu)^3/\mu^3\right.$$
$$\left.\left. +7.5(1-\mu)^2/\mu^2 + 20(1-\mu)/\mu + 15\right\}\exp\left((1-\mu)/\mu\ \right)E_1\left((1-\mu)/\mu\ \right)\right],$$

where μ is the mean of the response variable and $E_n(z) = \int_1^\infty t^{-n}\exp(-zt)dt$ is the exponential integral function.

Appendix B. Score functions and information matrix

In this appendix, we obtain the score functions and information matrix for the smooth test statistic. Let $l(\beta)$ be log-likelihood function of the UL model

$$l(\beta) = \sum_{j=1}^{n} l_j(\mu_j),$$

where $l_j(\mu_j) = 2\log(1-\mu_j) - \log(\mu_j) - 3\log(1-y_j) - \frac{y_j(1-\mu_j)}{(1-y_j)\mu_j}$. The regression parameter vector β is estimated by solving the following score function

$$U_\beta = \frac{\partial l(\beta)}{\partial \beta_u} = \sum_{j=1}^{n}\frac{\partial l_j(\mu_j)}{\partial \mu_j}\frac{\partial \mu_j}{\partial \tau_j}\frac{\partial \tau_j}{\partial \beta_u}$$
$$= \sum_{j=1}^{n}\left(-\mu_j - 1 + \frac{y_j(1-\mu_j)}{(1-y_j)\mu_j}\right)x_{ju},$$

where $\tau_j = \log\left(\frac{\mu_j}{1-\mu_j}\right) = x_j^T\beta$ is the logit link function. In the study, we used Fisher's scoring algorithm to estimate the regression model parameters.

For the UL regression model using smooth alternative distribution f_k in Equation (3.1), the corresponding log-likelihood function is

$$\log L\ = \sum_{j=1}^{n}\log C(\tau,\mu_j) + \sum_{i=1}^{k}\tau_i\sum_{j=1}^{n}h_i(y_j;\mu_j) + \sum_{j=1}^{n}\log f(y_j;\mu_j).$$

Since h_r $(r = 1, 2, .., k)$ is the rth order orthonormal polynomial with $h_0 = 1$, the expected value of the h_r is $E_0(h_r) = 0$ for $r \geq 1$ under H_0 [36] and the score function with respect to τ is given by

$$U_\tau = \frac{\partial \log L}{\partial \tau_r}\bigg|_{\hat{\gamma}_0} = \sum_{j=1}^{n} h_r(y_j, \hat{\mu}_j) = \sqrt{n} V_r,$$

where $V_r = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} h_r(y_j; \hat{\mu}_j)$.

The Fisher information matrix is partitioned as

$$I = \begin{bmatrix} I_{\tau\tau} & I_{\tau\beta} \\ I_{\beta\tau} & I_{\beta\beta} \end{bmatrix}$$

and the sub-matrices of the information matrix are derived as follows:

$$(I_{\tau\tau})(\hat{\gamma}_0) = -E_0\left[\frac{\partial^2 \log L}{\partial \tau_r \partial \tau_s}\right] = \begin{bmatrix} n & 0 & \cdots & 0 \\ 0 & n & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & n \end{bmatrix} = nI_k,$$

where I_k is the $k \times k$ identity matrix.

We assume that h_r is the rth order orthonormal polynomial on the UL distribution; therefore, we can write the polynomial as $h_r = \sum_{i=0}^{r} a_{ir} y^i$. The coefficients of each orthonormal polynomial can be extracted from the orthonormal polynomials on the UL distribution in Appendix C and we have

$$\begin{aligned}
(I_{\tau\beta})(\hat{\gamma}_0) &= -E_0\left[\frac{\partial^2 \log L}{\partial \tau_r \partial \beta_u}\right] \\
&= \sum_{j=1}^{n} Cov\left(h_r, \frac{\partial \log f}{\partial \beta_u}\right) \\
&= \sum_{j=1}^{n} \mu_j x_{ju} \sum_{i=0}^{r} a_{ir} \Gamma(2+i) U\left(-1+i, -2, \frac{1-\mu_j}{\mu_j}\right) \\
&= \sum_{j=1}^{n} \mu_j x_{ju} A_{rj} \\
&= \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{k1} & \cdots & A_{kn} \end{bmatrix} \begin{bmatrix} \mu_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \mu_n \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \\
&= ADX,
\end{aligned}$$

where X is the $n \times p$ matrix of covariates, $D = diag(\mu_1, ..., \mu_n)$ is the $n \times n$ diagonal matrix and A is the $k \times n$ matrix with rows $A_1^T, A_2^T, ..., A_k^T$ where $U(a, b, z)$ is the confluent hypergeometric function.

$$
\begin{aligned}
(I_{\beta\beta})(\hat{\gamma}_0) &= -E_0\left[\frac{\partial^2 \log L}{\partial\beta_u \partial\beta_v}\right] \\
&= \sum_{j=1}^{n} Cov\left(\frac{\partial \log f}{\partial\beta_u}, \frac{\partial \log f}{\partial\beta_v}\right) \\
&= \begin{bmatrix}
\sum\limits_{j=1}^{n} x_{j1}^2 w_j & \sum\limits_{j=1}^{n} x_{j1}x_{j2} w_j & \cdots & \sum\limits_{j=1}^{n} x_{j1}x_{jp} w_j \\
\sum\limits_{j=1}^{n} x_{j2}x_{j1} w_j & \sum\limits_{j=1}^{n} x_{j2}^2 w_j & \cdots & \sum\limits_{j=1}^{n} x_{j2}x_{jp} w_j \\
\vdots & \vdots & \ddots & \vdots \\
\sum\limits_{j=1}^{n} x_{jp}x_{j1} w_j & \sum\limits_{j=1}^{n} x_{jp}x_{j2} w_j & \cdots & \sum\limits_{j=1}^{n} x_{jp}^2 w_j
\end{bmatrix} \\
&= \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ \vdots & \ddots & \vdots \\ x_{1p} & \cdots & x_{np} \end{bmatrix}
\begin{bmatrix}
-\mu_1^2 + 2\mu_1 + 1 & 0 & \cdots & 0 \\
0 & -\mu_2^2 + 2\mu_2 + 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & -\mu_n^2 + 2\mu_n + 1
\end{bmatrix} \\
&\quad \times \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \\
&= X^T W X,
\end{aligned}
$$

where W is the $n \times n$ diagonal weight matrix. Note that the $I_{\beta\beta}$ is the information matrix that has been used to estimate the regression parameters β in the Fisher scoring algorithm.

Appendix C. Orthonormal polynomials

The orthonormal polynomials of the UL distribution are generated by GramSchmidt orthogonalization process using the basis $\{1, y, y^2, y^3\}$. Let μ_2^* be the second central moment of the response variable. The first four orthonormal polynomials on the UL distribution are given by $h_0(y;\mu) = 1$ and $h_r(y;\mu)$ $(r = 1, 2, 3)$:

$$
\begin{aligned}
h_1(y;\mu) &= (y - \mu)/\sqrt{\mu_2^*}, \\
h_2(y;\mu) &= \Big(y^2 + ay + b\Big)/\sqrt{d}, \\
h_3(y;\mu) &= \Big(y^3 + cy^2 + py + t\Big)/\sqrt{e},
\end{aligned}
$$

where $a = -(\mu_3' - \mu\mu_2')/\mu_2^*$, $b = -\mu_2' - a\mu$, $d = \mu_4' + a^2\mu_2' + b^2 + 2a\mu_3' + 2b\mu_2' + 2ab\mu$, $c = -(\mu_5' + a\mu_4' + b\mu_3')/d$, $p = ac - (\mu_4' - \mu\mu_3')/\mu_2^*$, $t = -\mu_3' + bc + \mu(\mu_4' - \mu\mu_3')/\mu_2^*$, $e = \mu_6' + c^2\mu_4' + p^2\mu_2' + t^2 + 2c\mu_5' + 2p\mu_4' + 2t\mu_3' + 2cp\mu_3 + 2ct\mu_2' + 2pt\mu$.