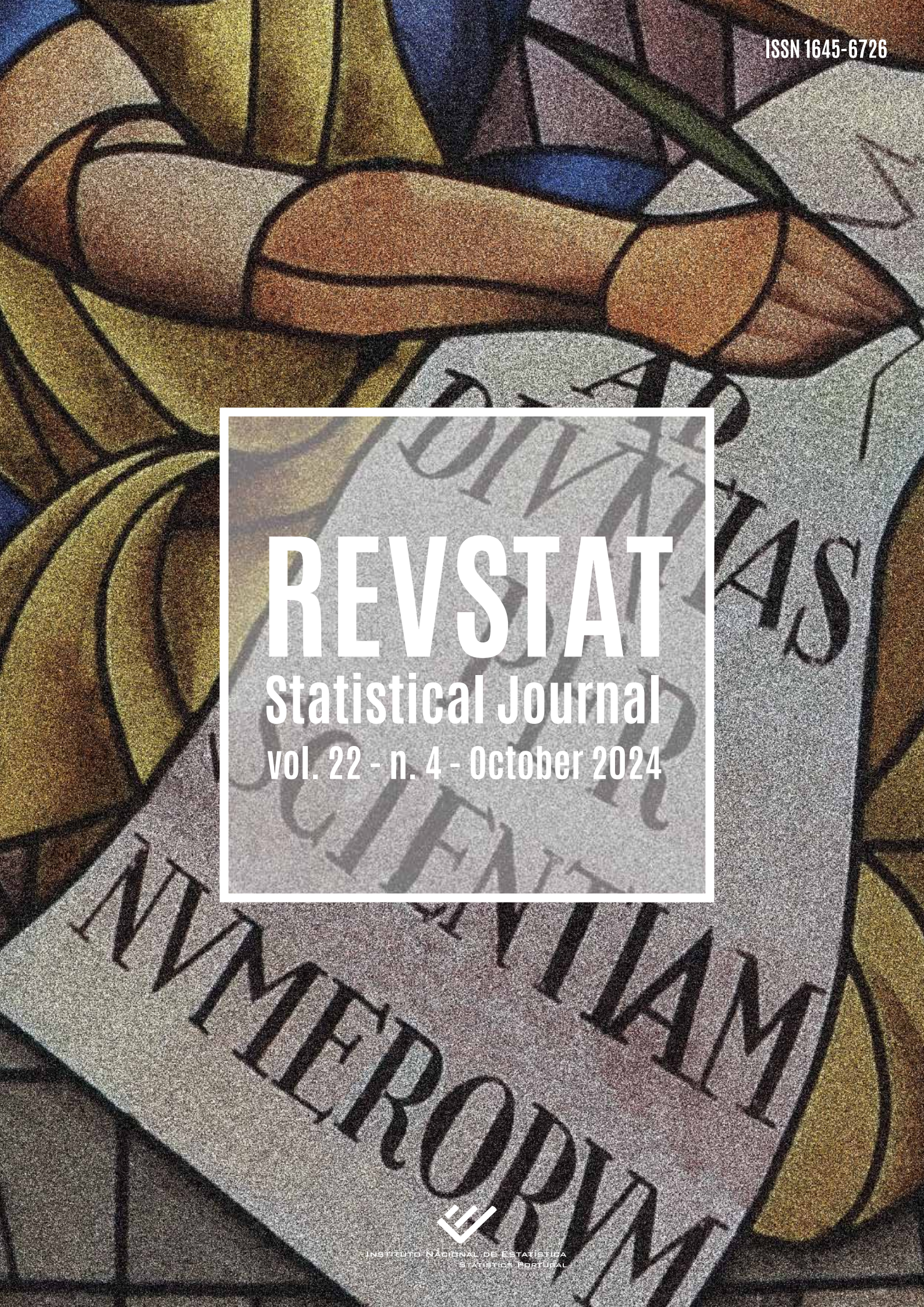
ISSN 1645-6726



REVSTAT

Statistical Journal

vol. 22 - n. 4 - October 2024

Instituto Nacional de Estatística
Statistics Portugal

REVSTAT — Statistical Journal, vol.22, n. 4 (October 2024)

vol.1, 2003- . - Lisbon : Statistics Portugal, 2003- .

Continues: Revista de Estatística = ISSN 0873-4275.

ISSN 1645-6726 ; e-ISSN 2183-0371

Editorial Board (2024-2025)

Publisher – Statistics Portugal

Layout-Graphic Design – Carlos Perpétuo | **Cover Design*** – Helena Nogueira

Edition - 130 copies | **Legal Deposit Registration** - 191915/03 | **Price** [VAT included] - € 9,00

**image*: stain glass window by Abel Manta (1888-1982)

INDEX

Abstracted/indexed in: *Journal Citation Reports™ – JCR (Clarivate); DOAJ – Directory of Open Access Journals; Current Index to Statistics; Google Scholar; Mathematical Reviews® (MathSciNet®); Zentralblatt für Mathematic; SCImago Journal & Country Rank; Scopus®.*

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 411–432
https://doi.org/10.57805/revstat.v22i4.542

Bootstrap Resampling Method for Estimation of Fuzzy Regression Parameters and a Sample Application

Authors: Dervış Topuz ✉
– Niğde Zübeyde Hanim School of Health, Niğde Ömer Halisdemir University, Niğde, Türkiye
topuz@ohu.edu.tr

Volkan Özkaya
– Faculty of Health Sciences, Department of Nutrition and Dietetics, İstanbul Medipol University, İstanbul, Türkiye
volkan.ozkaya@medipol.edu.tr

Betül Çiçek
– Faculty of Health Sciences, Department of Nutrition and Dietetics, Erciyes University, Kayseri, Türkiye
bcicek@erciyes.edu.tr

Received: January 2022 Revised: October 2022 Accepted: October 2022

Abstract:

- In fuzzy regression modeling, the fuzzy least squares technique is based on minimizing the squares of the total difference between observed and predicted outcomes. When the sample size is small, bootstrap resampling method is appropriate and useful for improving model estimation. The bootstrap resampling technique relies on resampling observations and resampling errors for bootstrap regression. The aim of this study is to investigate the use of Bootstrap in fuzzy regression modeling to estimate mean prediction with smaller errors at a particular α-segment, and apply it on a clinical data set. The behavior and properties of the least-squares estimators are affected when deviations or fuzziness arise in the sample and/or by slight changes in the data set. Bootstrap technique, on the other hand, provide robust estimators of the parameters which avoid such adverse effects.

Keywords:

- *bootstrapping; fuzzines; fuzzy linear regression; mean prediction error; fuzzy level; fuzzy confidence interval.*

AMS Subject Classification:

- 49A05, 78B26.

✉ Corresponding author

1. INTRODUCTION

When examining any problem, more than one independent variable may be related to the dependent variable. In order to explain the relationship between such variables, linear regression analysis is the most frequently used technique in statistical models. On the basis of the technique; while evaluating an observed event, it is essential to investigate which events are affected. These events may be one or more, as well as indirectly or directly affected. The technique expresses to what extent the observation values and the affected events are related with the help of a function. Generally, what percentage of the total change in an observed event can be explained by the affected event is evaluated according to the coefficient of determination. However, due to the occurrence of diagnostic procedures related to events in health services (disease, birth, death, etc.) at several levels, complex clinical uncertainties may arise in the relationships due to the insufficiency and uncertainty of information due to the nature of the data and various measurement methods used. In addition, as the sources of uncertainties in clinical relationships, reasons such as the subjective nature of medical history information, objectivity in the examination method, the fact that patient information may contain falsehood, measurement errors in the results of laboratory and other diagnostic tests, due to various restrictive factors, reasons such as the formation of sample sizes in the form of small data sets is shown [21]. In these situations involving uncertainties of medical applications, general facts are that the decisions made by experts can often cause contradictions since valid and reliable sampling and analysis techniques are not preferred. In this case, the calculations made have become questionable.

Various estimation techniques have been developed in order to solve these problems. These techniques are the bootstrap resampling method developed by Efron [21] and fuzzy least squares regression (FLSR) analysis technique by Tanaka *et al.* [59], [24]. The techniques can also be used as a correction method in cases where the assumptions about the error values of the regression model are not realized [55]. These can ensure that there is no difference between the actual observation value and the estimated values or that the difference is minimal. Fuzzy regression analysis is a valid and reliable technique for investigating and predicting data sets by measuring a concept that contains some degree of ambiguity or uncertainty [56]. In models where the data are insufficient or imperfect, caused by the imprecision or vagueness, it has been proven to be useful to use fuzzy models [9], [64]. The importance of using the bootstrap resampling technique and fuzzy regression technique in analysis for estimating model parameters has been increasingly recognized in recent years. Bootstrap resampling technique and fuzzy least squares technique are alternative techniques used in many areas such as time series and simulation in estimating linear and non-linear regression parameters.

This study aims to give illustration and application of bootstrap resampling technique in fuzzy least squares regression analysis in examining the clinical relationships between variables. Because classical least-squares technique is influenced by the outliers, therefore the presents of outliers may distort the estimates. Accordingly, bootstrap fuzzy regression methods have been created to modify least-squares methods so that the outliers have much less influence on the final estimates. We proposed the bootstrap fuzzy least squares regression technique as the estimation approach. The proposed model fitting approach is highly robust to the presence of outliers and properly determines outlier points and neutralizes the negative influence of outliers in the estimation procedure. In addition, we can provide more

general approaches that can consider different estimation scenarios. Some hierarchical algorithms concerning bootstrap technique in OLS and fuzzy least squares regression analysis are demonstrated. The basics of the bootstrap resampling technique and their applications to the clinic numerical example that can be described by fuzzy least squares regression model were discussed and compared the results with ordinary least squares regression technique results. It was also aimed to estimate the bias, standard error and confidence interval of the regression coefficients calculated by the techniques and to compare the performance of bootstrap ordinary least squares technique (BOLS) with the related estimates using some comparison criteria. The expectation for the future research topic on fuzzy regression is that many other new proposals and applications will appear in this context. The extension of the proposed procedure to the case of fuzzy input-fuzzy output observations, potential subjects for future researches. It turns out that traditional/common methods in the literature, as well as several other robust methods of fuzzy regression, can be formulated as special instances of bootstrap fuzzy regression.

2. BOOTSTRAP RESAMPLING TECHNIQUE

One of the most important purposes of statistical analysis is that the sample taken from the population must represent the population. The bootstrap resampling technique was developed by Efron [20] as a general technique for assessing the statistical accuracy of an estimator. The main purpose of the technique is to calculate the predictive θ value by choosing random samples with width n volumes, independent of a certain unknown distribution $f(x;\theta)$ and accept it as the predictor of the parameter θ. The bootstrap resampling technique is theoretically used to estimate values associated with the sampling distribution of estimators and test statistics.

The ordinary sampling techniques use some assumptions related to the form of the estimator distribution. These are the cases where standard assumptions are invalid, e.g. n volume is small, data contains uncertainty, data shows non-standard twist. In these situations, the use of these standard techniques may not give reliable and valid results. When these assumptions are doubtful or when the calculation of standard errors is necessary when parametric inference is impossible, the bootstrap resampling technique makes calculations without the need for these distributive assumptions because the sample population is considered [21]. The results calculated by the estimators can be used as an experimental distribution for statistics [11], [20]. The technique has been rarely used, although it is used to generate the estimation of the standard error of a statistic, confidence intervals and distributions by repeated use of the observed data [21], [23], [24], [53].

With the application of the in 4.1. bootstrap algorithm, the bias between population parameters and estimators will be reduced without increasing the sample size, and by obtaining the sampling distributions of the estimators, it will be provided to calculate the standard error of the estimators more accurately [10], [15], [16].

3. FUZZY LEAST SQUARES REGRESSION (FLSR) ANALYSIS

Fuzzy regression analysis is a fuzzy (or possibility) type of ordinary regression analysis. Fuzzy regression analysis studies the relationships between a response variable and a set of explanatory variables in complex systems involving imprecise data. The approach is one of the most widely used statistical techniques for evaluating the functional relationship between dependent and independent variables in uncertainty situations. In fuzzy regression analysis, the relationship between dependent variables and independent variables is not as precise as in ordinary regression analysis [64]. In these uncertain cases, fuzzy techniques can explain the effects of independent variables in a more realistic way. A commonly used technique of the parameter estimation of the fuzzy regression model is the least-squares method. The fuzzy least squares (FLS) technique, which is an extension of the least squares technique to fuzzy set theory, was used by to estimate fuzzy parameters [9], [17], [64]. These methods are very important because sometimes even a single observation can change the value of the parameter estimates, and omitting this observation from the data may lead to totally different estimates [12], [19], [31].

The approach is based on blurring the coefficients. Blurring can be done in two ways. It is possible by 1) blurring the model coefficients estimated by the ordinary least squares technique at a specified "h level", or 2) estimating the coefficients as fuzzy numbers [29], [45], [48]. However, in 1988, Diamond [18] concluded that "Tanaka *et al.* [61] used linear programming techniques to develop a model superficially resembling linear regression, but it is unclear what the relation is to a least squares concept, or that any measure of best fit by residuals is present". Most of the researches on fuzzy regression analysis focuses on the possibilistic regression [59], [61] and on the fuzzy Least-Squares (LS) regression [8], [18]. Recently, robust approaches to fuzzy regression have been considered as alternative approaches to fuzzy regression analysis [13], [15]. The bootstrap resampling technique using fuzzy data is developed in different approaches [46], [58] have considered the problem of hypothesis testing about the mean of a fuzzy random variable. Akbari and Rezaei [1] present a bootstrap fuzzy test for variance. Ferraro *et al.* (2010) [27] "International Journal of Approximate Reasoning 51 (2010) 759–770" proposed to use of a bootstrap procedure to evaluate the accuracy of the estimators in FLS regression. This idea is also investigated and proposed by many authors like "Akbari *et al.* (2012) [2], [24]. In this regard, Peters (1994) [51] considered outliers in Tanaka's possibilistic approach [59] with crisp input-output data which was later extended by Chen (2001) [10] to the model with fuzzy output-crisp input data. Hung and Yang (2006) [31] proposed an omission approach for Tanaka's approach [59] which had the ability to consider the effect of each observation while omitted on the value of the objective function of the model. Nasrabadi *et al.* (2007) [49] proposed an LP-based approach to outliers detection in fuzzy regression analysis. Varga (2007) [63] presented robust estimation approaches to fuzzy and non-fuzzy regression models. Nasrabadi and Hashemi (2008) [50] suggested a robust nonlinear fuzzy regression model using multilayered feedforward neural networks. Kula and Apaydin (2008) [36] proposed a robust fuzzy regression analysis based on the ranking of fuzzy sets. D'Urso and Massari (2013) [19] proposed weighted least-squares and least-median squares estimation for fuzzy linear regression analysis. Yang, Yin and Chen (2013) [66] present a robustified fuzzy varying coefficient model for fuzzy input-fuzzy output variables. Shakouri and Nadimi (2013) [57] investigated a method for outlier detection in fuzzy linear regression problems. Ferraro and Giordani (2013) [28] dealt with robustness in the

field of regression analysis for imprecise information managed in terms of fuzzy sets. Leski and Kotas (2015) [41], by introducing an objective function based on Huber's M-estimators and Yager's OWA operators, proposed a robust fuzzy-regression model. Chachi (2019) [12] introduced a weighted objective function to overcome the disadvantages of the LS fuzzy regression approaches in the presence of outliers. Arefi (2020) [5] investigated a quantile fuzzy regression based on fuzzy outputs and fuzzy parameters. Akbari and Hesamian (2019) [3] investigated a partial-robust-ridge-based regression model with fuzzy predictors-responses. Bootstrap fuzzy resampling technique tests for the mean and variance with Dp, q-distance [52]. They proposed bootstrap fuzzy linear regression model (BFLRM), a linear regression model with fuzzy dependent, crisp explanatory and fuzzy coefficients [59], [60]. Most of these developed fuzzy regression models are evaluated with fuzzy outputs and fuzzy parameters but non-fuzzy (net) inputs. Fuzzy least squares regression (FLSR) analysis technique, which is generally based on linear programming (LP), is proposed in order to minimize the fuzziness of the analyzed data and the total spread of the output (see, for example [12], [17] [29]). Hesamian and Akbari (2020) [32] proposed a robust varying coefficient approach to fuzzy multiple regression model. Hesamian and Akbari (2021) [33] adopted a two-stage robust procedure to propose and estimate the components of a robust multiple regression model with fuzzy intercepts and crisp coefficients. Khammar *et al.* (2020) [37], Khammar *et al.* (2021) [38], Khammar *et al.* (2021) [39] presented general approaches to fit fuzzy regression models crisp/fuzzy input and fuzzy output. Asadolahi *et al.* (2021) [6] proposed a robust support vector regression with exact predictors and fuzzy responses. Taheri and Chachi (2021) [62] investigated a robust variable-spread fuzzy regression model. Chachi and Chaji (2021) [14] considered quantile fuzzy regression using OWA operators. In the context of multi-attribute decision-making problems, Chachi *et al.* (2021) [13] developed a multi-objective two-stage optimization and decision technique for fuzzy regression modeling problems in order to handle both of the weak performances analysis of fuzzy regression models and their sensitivity to outliers. As mentioned above, during the last years, considerable attention was given to robust estimation problems in fuzzy environments, and several methodologies were developed in the literature [15].

According to the FLSR approach, it is assumed that the deviations between the observed values and the predicted values are caused by the uncertainty of the system structure or the blurring of the regression coefficients, not from measurement and observation errors, contrary to the OLSR analysis method [9]. That is, it assumes that the coefficients of the regression analysis model are related to its blur. For this purpose, the formula below is employed to estimate parameters of FLSR:

$$f = X \times \widetilde{\beta} \to \widetilde{Y}_i, \widetilde{Y}_i = f\left(\widetilde{\beta}, X\right) \tag{3.1}$$

It is given by the function 1. Here, $\widetilde{Y}_i$, denotes the fuzzy dependent variable in the symmetric triangular property structure estimated and is shown as $\widetilde{Y}_i = (\widetilde{y}_c,\ \widetilde{e}_s)$, $\widetilde{y}_c$ denotes the mean value (center), and $\widetilde{e}_s$ denotes the spread value.

In the case of fuzzy observations, consider a fuzzy linear regression for crisp explanatory and fuzzy response observations as follows:

$$\widetilde{Y}_i = f\left(\widetilde{\beta}, X\right) = \widetilde{\beta}_0 + \widetilde{\beta}_1 X_{i1} + ... + \widetilde{\beta}_{p-1} X_{i(p-1)} = \widetilde{\beta}_0 + \sum_{i=1}^{n} \widetilde{\beta}_i X_i \tag{3.2.a}$$

(3.2.b) $\widetilde{Y}_i = \{c_0,\ s_0\} + \{c_1,\ s_1\}X_{i1} + \{c_2,\ s_2\}X_{i2} + ... + \{c_{p-1},\ s_{p-1}\}X_{i(p-1)}$

in which $\widetilde{\beta}_j = \left[\widetilde{\beta}_0\ and\ \ \widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\beta}_3, ...\widetilde{\beta}_j ....., \widetilde{\beta}_{p-1}\right]^t$ are the coefficient values of the independent variables in the function and it is a set of dependent and independent variables formed in the form of $\{Y_i,\ X_{i1}, X_{i2},\ X_{i3}, ..., X_{(p-1)n}\ \} = \{Y_i, X_{ij}\ \}$, and each dependent variable observation is expressed as $x \epsilon X\ \ (i = 1, ..., n,\quad j = 1, 2, ..., p-1)$. That is, they are crisp values of the explanatory variables. It is defined by $(i = 1, 2, 3, ..., n)$. In the fuzzy least squares regression model, the data of the dependent $\widetilde{Y}_i$ variable can be real numbers or fuzzy numbers. It is generally assumed that the data for the dependent Y_i variable are symmetrical fuzzy numbers of interval type [35].

$\widetilde{\beta}_j = \left[\widetilde{\beta}_0\ ve\ \ \widetilde{\beta}_1, \widetilde{\beta}_2, \widetilde{\beta}_3, ...\widetilde{\beta}_j ....., \widetilde{\beta}_{p-1}\right]^t$ are fuzzy regression coefficients vectors with a symmetric triangular fuzzy number structure and they are fuzzy numbers in the form of $\widetilde{\beta}_j = (c_j,\ s_j)\widetilde{\beta}_j, (j : 0, 1, 2, 3, ..., p-1).c_j,$ is the $\mu_{\widetilde{\beta}_i}(c_j) = 1$ value representing the midpoint of the coefficients, that is, the center value, and has the form $c_j = [c_1, c_2 c_3, ..., c_n]^t$. s_j, shows the spread of the coefficients belonging to the fuzzy regression analysis model and is $s_j = [s_1, s_2, s_3, ..., s_n\ ]^t$ shaped [64].

Each coefficient value $\widetilde{\beta}_i = \{c_j,\ s_j\} = \left\{\widetilde{\beta}_i : c_j - s_j \leq \widetilde{\beta}_i \leq c_j + s_j\right\}$ has a symmetric triangular property structure and is $\widetilde{\beta}_i (j : 0, 1, 2, 3, ..., p-1)$ [59].

The $\widetilde{\beta}_i = \{c_j,\ s_j\}$ value of the fuzzy coefficients was estimated by the minimum blur method proposed by Tanaka. The method is given in the following equation. In least squares regression analysis proposed by Tanaka and Watada (1988) [60], the linear programming (LP) formulation considers triangular membership functions (not necessarily symmetric). The spreads of the calculated fuzzy coefficients are calculated with the help of equation. The LP formulation is as follows (3):

$$(3.3) \qquad \min\ \ Z(x)\ \ s^t|X_i| = \min_{c,\,s}\left[s_0 + \sum_{j=0}^{n} s_j\ |X_{ij}|\ \right]$$

$$\min_{c,\,s} J = c_1, c_2, .., c_n\ ,\ \ c_j \geq 0,\ \forall i,\ \ i = 1, 2, .., m \text{ and}$$

$$\min_{c,\,s} J = s_{\ 1}, s_2, ..., s_n,\ \ s_j \geq 0\ \ \forall i,\ \ j = 0, 1, 2, ..., n$$

$$\sum_{j=0}^{n} c_j X_{ij} + (1-h)\left[\sum_{j=0}^{n} s_j\ |X_{ij}|\ \right]\ \ \geq \widetilde{y}_c + (1-h)\widetilde{e}_s \quad \forall i,\ \ i = 1, 2, ..., n$$

$$\sum_{j=0}^{n} c_j X_{ij} - (1-h)\left[\sum_{j=0}^{n} s_j\ |X_{ij}|\ \right]\ \ \leq \widetilde{y}_c - (1-h)\widetilde{e}_s \quad \forall i,\ \ i = 1, 2, ..., n$$

Here $Z(x)$: function shows the total blur in the model. m: is the number of observations regarding the dependent variable. j : the number of independent variable x_{ij} : is the i-th observation value of the j-th independent variable. For each predicted $\widetilde{Y}_i$ observation value, the constraint number must be $2xn$ [43]. In order to minimize the total spread, the level h, $\widetilde{Y}_i$, the predictor of each observation value Y_i, is assumed to have a turbidity tolerance

$\mu_{\widetilde{Y}_i}(Y_i) \geq h \;\; i = 1, 2, .., m$ [30]. In Equation 3, the objective function is weighted with the absolute values of the measurements of the distributions of the independent variables. The application of bootstrap resampling technique in fuzzy least squares regression analysis is given below.

4. BOOTSTRAP FUZZY REGRESSION ANALYSIS

In this section, we introduce bootstrap resampling technique procedure. In general, regression technique for bootstrap is divided into two approaches: the first is based on the resampling observations approach and the second is based on the resampling errors. Bootstrap technique based on resampling errors is known as more suitable for the case of deterministic, whereas bootstrap resampling technique based on the drawing i.i.d. sample from the observations pairs is more appropriate for the case of random. However, bootstrap resampling technique pairs can also be used for deterministic [1]. The bootstrap is a "model-dependent" technique in terms of its implementation and performance although the bootstrap requires no theoretical formula for the quantity to be estimated and is less model-dependent than the traditional approach. In this paper, we use bootstrap technique based on the resampling errors. The bootstrap fuzzy regression analysis procedure is as follows:

Method: To describe the resampling methods we start with an n sized sample $w_i = (Y_i, X_{ji})'$ and assume that w_i's are drawn independently and identically from a distribution of F, where $Y_i = (y_1, y_2, .., y_n)'$ contains the responses, $X_{ji} = (x_{j1}, x_{j2}, ..., x_{jn})'$ is a matrix of dimension nxk, where $j = 1, 2, ...k, i = 1, 2, 3, ..., n$.

4.1. Bootstrapping Regression Algorithm

Here, two approaches for bootstrapping regression methods were given. The choice of either methods depends upon the regressors are fixed or random. If the regressors are fixed, the bootstrap uses resampling of the error term. If the regressors are random, the bootstrap uses resampling of observation sets w_i [55].

4.2. Bootstrap Based On The Resampling Errors

If the regressors are fixed, as in desing experiment, then the bootstrap resampling must preserve that structure. The bootstrap procedure based on the resampling errors as follows [55]:

$1^{(e)}$. Fit the full-sampling least-squares regression equation to estimate the regression coefficients of the model (6.a).

$2^{(e)}$. Calculate the e_i values $\left(e_i = Y_i - \hat{Y}_i\right)$

$3^{(e)}$. Draw an n sized bootstrap random sample with replacement $e_1^{(b)}, e_2^{(b)}, e_3^{(b)}, ..., e_n^{(b)}$ from the e_i values calculated in step $2^{(e)}$ giving $1/n$ probability each e_i values and Calculate the centered residual of $\bar{\hat{e}}_i^{(b)}$ [42] [53] [65]:

$4^{(e)}$. Compute the bootstrap $Y_i^{(b)}$ values by adding resampled residuals onto the ordinary least squares regression fit, holding the regression desing fixed [16] [55].

$$Y_i^{(b)} = X\hat{\beta} + \bar{\hat{e}}_i^{(b)} \tag{4.1}$$

$5^{(e)}$. Obtain least squares estimates from the 1-th bootstrap sample:

$$\widetilde{\beta}^{(b1)} = (X'X)^{-1}X'Y^{(b)} \quad \text{(we need } Y^*\text{)} \tag{4.2}$$

$$\widetilde{\beta}^{(b1)} = \hat{\beta}(X'X)^{-1}X'e^{(b)} \quad \text{(we do not need } Y^*\text{)} \tag{4.3}$$

$6^{(e)}$. Repeat steps $3^{(e)}, 4^{(e)}$ and $5^{(e)}$ for $r = 1, 2, ..., B$, and proceed as in resampling with random regressors $7^{(e)}$ and $8^{(e)}$.

An illustrative example that presents how the regression parameters are estimated from the bootstrap based on the resampling observations was given in Table 1.

By resampling residuals and randomly reattaching them to fitted values, the procedure implicitly assumes that the errors are identically distributed. Bootstrapping draws an analogy between the fitted value $\hat{Y}_i$ in the sample and Y in the population, and between the residual e_i in the sample and the error ε_i in the population [21]. In bootstrap resampling technique principle, the sample represents the population as the bootstrap samples. According to the weak law of large numbers, the empirical distribution function converges in probability to the true distribution function [42]. Note that define the bootstrap observation $Y_i^{(b)}$, by treating $\hat{\beta}$ as the "true" parameter and $e_i^{(b)}$ as the "population" of errors [54].

$7^{(e)}$. Obtain the probability distribution $\left(F\left(\widetilde{\beta}^{(b)}\right)\right)$ of bootstrap estimates $\widetilde{\beta}^{(b1)}$, $\widetilde{\beta}^{(b2)}, ..., \widetilde{\beta}^{(bB)}$ and use the $\left(F\left(\widetilde{\beta}^{(b)}\right)\right)$ to estimate regression coefficients, variances and confidence intervals as follows. The bootstrap estimate of regression coefficient is the mean of the distribution $\left(F\left(\widetilde{\beta}^{(b)}\right)\right)$ [25] [55].

$$\beta^{(b)} = \sum_{b=1}^{B} \beta^{(br)}/B = \beta^{(br)} \tag{4.4}$$

$8^{(e)}$. Thus, the bootstrap regression equation is

$$\min_{a_c, a_s} J1 = \begin{bmatrix} c0 + X_{i1} * c1 + X_{i2} * c2 + X_{i3} * c3 + X_{i4} * c4 - s0 - X_{i1} * s1 - X_{i2} * s2 - X_{i3} * s3 - X_{i4} * s4 <= Y_1 \\ c0 + X_{i1} * c1 + X_{i2} * c2 + X_{i3} * c3 + X_{i4} * c4 - s0 + X_{i1} * s1 + X_{i2} * s2 + X_{i3} * s3 + X_{i4} * s4 >= Y_1 \end{bmatrix}$$

$$\min_{a_c, a_s} Jn = \begin{bmatrix} c0 + X_{i150} * c1 + X_{i150} * c2 + X_{i150} * c3 + X_{i150} * c4 - s0 - X_{i150} * s1 - X_{i150} * s2 - X_{i150} * s3 - X_{i150} * s4 <= Y_1 \\ c0 + X_{i150} * c1 + X_{i150} * c2 + X_{i150} * c3 + X_{i150} * c4 - s0 + X_{i150} * s1 + X_{i150} * s2 + X_{i150} * s3 + X_{i150} * s4 >= Y_1 \end{bmatrix}$$

$$\widetilde{Y}_i = f\left(\widetilde{\beta}, x\right) = \widetilde{\beta}_0^{(b)} + \widetilde{\beta}_1^{(b)} x_{j1}^{(b)} + \widetilde{\beta}_2^{(b)} x_{j2}^{(b)} + ... + \widetilde{\beta}_s^{(b)} x_{jn}^{(b)} \tag{4.5}$$

where $\widetilde{\beta}_0^{(b)}$ is unbiased estimator of β [40].

4.3. The bootstrap bias, variance, confidence and percentile interval. The bootstrap bias equals,

$$\text{bias}_b = \widetilde{\beta}^{(b)} - \widetilde{\beta} \tag{4.6}$$

Further discussion are described in Efron and Tibshirani (1998) [23]. The bootstrap variance from the distribution $\left(F\left(\widetilde{\beta}^{(b)}\right)\right)$ are calculated by [53] [55].

$$\text{var}\left(\widetilde{\beta}^{(b)}\right) = \sum_{i=1}^{B}\left[\left(\widetilde{\beta}^{(br)} - \widetilde{\beta}^{(b)}\right)\left(\widetilde{\beta}^{(br)} - \widetilde{\beta}^{(b)}\right)'\right] \Big/ (B-1), \quad r = 1, 2, ..., B \tag{4.7}$$

The bootstrap confidence interval by normal approach is obtained by

$$\left(\widetilde{\beta}^{(b)} - t_{n-p,\ \frac{\alpha}{2}} * S_e\left(\widetilde{\beta}^{(b)}\right) < \beta < \widetilde{\beta}^{(b)} + t_{n-p,\ \frac{\alpha}{2}} * S_e(\widetilde{\beta}^{(b)})\right) = 1 - \alpha \tag{4.8}$$

where $t_{n-p,\ \frac{\alpha}{2}}$ is the critical value of t with probability $\alpha/2$ the right for $n-p$ degrees of freedom, and $S_e\left(\widetilde{\beta}^{(b)}\right)$ is the standard error of the $\widetilde{\beta}^{(b)}$. If the sample size is n $\geq$ 30, then Z distribution values are used instead of t in estimation of confidence intervals [22].

A non-parametric confidence interval named percentile Interval can be constructed from the quantiles of the bootstrap sampling distribution of $\widetilde{\beta}^{(b)}$. The $(\alpha/2)\%$ and $(1-\alpha/2)\%$ percentile interval is

$$\widetilde{\beta}^{(br)}_{(lower)} < \beta < \widetilde{\beta}^{(br)}_{(upper)} \tag{4.9}$$

where $\widetilde{\beta}^{(br)}$ is the ordered bootstrap estimates of regression coefficient from Equation 9 or 10, lower=$(\alpha/2)B$ and upper=$(1-\alpha/2)B$.

5. ILLUSTRATIVE EXAMPLE

Numerical examples are used to illustrate the fuzzy regression model that are summarized in previous sections. This example focuses on illustration and application of bootstrap technique in fuzzy regression analysis. Rapidly changing scientific and technological developments in recent years have negatively affected the health status of individuals by changing their nutritional habits. One of the main indicators of a healthy life is to have a stabile body composition. In recent years, along with the increasing prevalence of overweight and obesity worldwide, it has become even more crucial how to have a stabile body composition. In addition to overweight and obese individuals, it has become important to maintain the stability of body composition in the elderly, athletes and individuals with certain diseases. For such cases, anthropo-plyometric measurements can be used to evaluate the development-growth and nutritional status of individuals on body composition. In addition, the effects of dietary patterns of different diseases can be monitored and body composition can be determined.

In this study, in order to estimate total fat ($DEXATF$) calculated according to DEXA method (Y) values with minimum error, Triceps values of independent variables such as Body Mass Index (BMI)(kg/m^2)(X_1), age ($YEAR$) (X_2) waist circumference fat percentage

($WCFP$) (X_3) were used as material in the model. These values were used in the classical bootstrap regression analysis method ($BOLSR$) and fuzzy linear bootstrap regression analysis methods ($BFLSR$), and the results were compared by calculating the estimated values of the coefficients and statistical values. The sample size was determined as 50 participants in order to determine whether more reliable results can be calculated in a shorter time with small data sets in cases where the constraints of the classical bootstrap regression analysis method cannot be met and the uncertainties in the datasets are not minimized.

The data used in the current study was obtained by the permissions of Drug Researches Local Ethics Committee of Erciyes University Faculty of Medicine (Date: 02.12.2008, Number: 2008/613) and Human Researches Ethics Committee of Kocaeli University (Date: 10.03.2009, Number: 2009/48) and the support of Scientific Research Project Coordination Unit of Erciyes University (Project code: TSY-09-772). The study was conducted in accordance with the principles of the Declaration of Helsinki. The study sample was consisted of randomly selected 137 voluntary participants, aged between 18-65 years and admitted to Kocaeli University, Faculty of Medicine, Department of Nuclear Medicine from May to July 2009. Of the participants, 67 (50%) were females and 67 (50%) were males, respectively. Women with pregnancy/suspected pregnancy and in the menstruation period, participants with metabolic and endocrine diseases and with any systemic diseases (liver, kidney, heart) and participants prescribed with hormonal drugs and anti-oedematous drugs were excluded.

The data pairs $w_i = (Y_i, X_{ji})'$ of Table 1 population, $(i = 1, ..., 50)$ are used to demonstrate the proposed procedure in case where the crisp input X and crisp output Yi .

Table 1: $n = 50$ volume original data set.

No	DEXATF (Y)	BMI (X_1)	YEAR (X_2)	WCFP (X_3)
1	31.20	28.30	44.00	34.82
2	26.50	21.30	26.00	44861
3	34.80	28.40	54.00	39.12
.	.	.	.	.
48	53.40	40.30	54.00	58.16
49	37.00	36.60	37.00	35.07
50	24.50	44859	24.00	19.07

DEXATF: total fat (DEXATF) calculated according to DEXA method; BMI: Body Mass Index (kg/m^2); YEAR; WCFP: Waist Circumference Fat Percentage.

The bootstrap algorithm based on error terms has been applied to the data in Table 1 as follows:

$1^{(e)}$. First, the ordinary least squares regression (OLSR) model was fitted to data given in and the results of the ordinary least squares regression was summarized in Table 2.

All of the regressions in Table 2 are significant (p<0.01) and the determination of coefficients $R^2 = 0.933$, respectively. The regression of total fat calculated according to DEXA method on the Body Mass Index (kg/m^2), YEAR and waist circumference Fat

Percentage is significant as result of variance analysis (P<0.01**). According to the t-tests for significance of regression coefficients, all of the regression coefficients are significant (P<0.01). Therefore, BOLSR can be substituted as an alternative modelling approach. The illustration of the bootstrap ($B = 1000$ bootstrap samples, each of size $n = 30$) regression procedure, from the data given in Table 1, calculation the bootstrap estimates of the regression parameters for each sample are shown in Table 3.

Table 2: The summary statistics of regression coefficients for OLS regression.

Variables	$\hat{\beta}$	**S.E.**($\hat{\beta}$)	**t**	**Sig**	**95%Confidence Interval**
Constant	−6.973	3.031	−2.300	.026	(−13.074)–(−0.871)
BMI (X_1)	0.984	0.149	6.612	.000	(0.685)–(1.284)
YEAR (X_2)	−0.111	0.061	−1.809	.07	(−2.234)–(0.013)
WCFP (X_3)	0.399	0.125	3.177	.003	(0.146)–(0.651)
R^2=0.933, N = 50, SSE = 3.720, F = 103.853**					

DEXATF: total fat (DEXATF) calculated according to DEXA method; BMI: Body Mass Index BMI: Body Mass Index (kg/m^2); YEAR: WCFP: Waist Circumference Fat Percentage; SSE: sum of squares of error.

$2^{(e)}$**.** The values in Table 3 are obtained by calculating the values of e_i with $e_i = Y_i - \hat{Y}_i$.

Table 3: Bootstrap residual instances created by assigning the probability 1/n to each e_i value.

No	Y_i	$\hat{Y}_i$	e_i	**1/50**	**r**	$e_1^{(b)}$	$e_2^{(b)}$	$e_3^{(b)}$	$e_4^{(b)}$	**.**	$e_{48}^{(b)}$	$e_{49}^{(b)}$	$e_{50}^{(b)}$	$\bar{e}_i^{(b)}$
1	31.20	29.88	1.32	0.03	**1**	0.02	−0.11	−0.08	0.03	**.**	−0.05	−0.02	−0.11	−0.07
2	26.50	21.91	4.59	0.09	**2**	−0.06	−0.05	−0.10	−0.05	.	−0.05	−0.08	−0.05	−0.04
3	34.80	30.59	4.21	0.08	**3**	0.11	0.07	−0.11	0.10	**.**	−0.12	−0.05	−0.05	−0.02
4	31.30	28.92	2.38	0.05	**4**	−0.04	−0.08	−0.05	−0.05	**.**	−0.01	0.02	−0.10	−0.03
5	20.40	26.02	−5.62	−0.11	**5**	−0.02	0.01	0.10	−0.02	**.**	0.07	0.07	−0.05	−0.06
6	30.30	29.17	1.13	0.02	**6**	0.06	0.09	−0.10	−0.05	**.**	0.05	−0.02	0.02	−0.04
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
48	53.40	49.89	3.51	0.07	**48**	−0.06	−0.02	−0.04	0.09	**.**	−0.02	−0.04	0.01	−0.02
49	37.00	38.93	−1.93	−0.04	**49**	0.07	−0.01	−0.01	0.05	**.**	0.02	0.07	0.11	0.00
50	24.50	22.67	1.83	0.04	**50**	−0.05	−0.05	−0.01	0.03	**.**	−0.11	−0.01	−0.08	0.00
					.	.	.	.	.	.	.	.	.	.
					997	−0.10	0.18	0.09	0.05	**.**	−0.13	−0.11	0.38	−0.01
					998	0.05	−0.10	−0.22	0.09	**.**	−0.11	−0.08	−0.02	−0.04
					999	0.18	0.07	−0.07	0.09	**.**	−0.02	−0.03	−0.03	−0.05
					1000	−0.10	0.18	0.09	0.05	**.**	−0.13	−0.11	0.38	0.00
$\hat{\bar{e}}_i^{(b)} = \frac{\sum_{b=1}^{1000} e_i^{(b)}}{1000}$						−0.001	0.000	−0.004	0.001		−0.002	0.000	−0.001	

$3^{(e)}$**.** Draw an n sized bootstrap random sample with replacement $e_1^{(b)}, e_2^{(b)}, e_3^{(b)}, ..., e_n^{(b)}$ from the e_i values calculated in step $2^{(e)}$ giving 1/n probability each e_i values.

$4^{(e)}$**.** Calculated the bootstrap Y_i^* values by adding resampled residuals onto the ordinary least squares regression fit, holding the regression design fixed (Table 4).

Table 4: Bootstrap $Y^{(b)}$ values calculated with resampled residuals.

No	$Y^{(b)}$	$\hat{\beta}_0$	**B**MI	$\hat{\beta}_1$	YEAR	$\hat{\beta}_2$	**W**CFP	$\hat{\beta}_3$	$\bar{\hat{e}}_i^{(b)}$
1	29.88	−6.973	28.30	0.984	44.00	−0.111	34.82	0.399	−0.001
2	21.92	−6.973	21.30	0.984	26.00	−0.111	27.10	0.399	0.000
3	0.58	−6.973	28.40	0.984	54.00	−0.111	39.12	0.399	−0.004
.	.	.	.	.	.	.	.	.	.
48	9.89	−6.973	40.30	0.984	54.00	−0.111	58.16	0.399	−0.002
49	38.93	−6.973	36.60	0.984	37.00	−0.111	35.07	0.399	0.000
50	22.67	−6.973	25.10	0.984	24.00	−0.111	19.07	0.399	−0.001

$5^{(e)}$**.** Obtain least squares estimates from the 1^{th} bootstrap sample:

$$Y^{(b)} = -6.973 + 0.984 * BMI - 0.111 * YEAR + 0.399 * WCFP + \bar{\hat{e}}_i^{(b)} \tag{5.1}$$

$6^{(e)}$**.** Repeat steps $3^{(e)}$, $4^{(e)}$ and $5^{(e)}$ for r =1,2,...,B, and proceed as in resampling with random regressors $7^{(e)}$ and $8^{(e)}$ (Table 5).

Table 5: Some bootstrap descriptive statistics based on the resampling of the $(n = 50)$ error term of the data in Table 2.

Variables	Observed	$\hat{\beta}^*_{ort}$	$S_e(\hat{\beta}^*)$	$SS(\hat{\beta}^*)$	**Confidence intervals**	
					95% Confidence Interval	
Constant	−6.973	−69.590	31.809	0.0133	−12.977	−13.997
BMI(X_1)	0.984	0.9862	0.1246	−0.0020	0.802	12.145
YEAR(X_2)	−0.111	−0.1129	0.0641	−0.0020	−0.224	−0.0050
WCFP(X_3)	0.399	0.3993	0.0974	0.0005	0.250	0.6100

DEXATF: total fat (DEXATF) calculated according to DEXA method; BMI: Body Mass Index BMI: Body Mass Index (kg/m^2); YEAR; WCFP: Waist Circumference Fat Percentage.

$7^{(e)}$**.** By using the new observation points that have been formed, the parameters are estimated with the FLS regression analysis method:

MIN = 50*s0+1488.2*s1+1897*s2+1913*s3;

$$\min_{a_c, a_s} J1 = \begin{bmatrix} c0 + 28.30 * c1 + 44 * c2 + 34.82 * c3 * 0.5 * s0 + 28.30 * 0.5 * s1 + 44 * 0.5 * s2 + 34.82 * 0.5 * s3 \geq = 29.88; \\ c0 + 28.30 * c1 + 44 * c2 + 34.82 * c3 - 0.5 * s0 - 28.30 * 0.5 * s1 - 44 * 0.5 * s2 - 34.82 * 0.5 * s3 >= 29.88; \end{bmatrix}$$

(5.2)

$$\min_{a_c, a_s} J1 = \begin{bmatrix} c0 + 25.10 * c1 + 24 * c2 + 19.07 * c3 * 0.5 * s0 + 25.10 * 0.5 * s1 + 24 * 0.5 * s2 + 19.07 * 0.5 * s3 \geq 22.67; \\ c0 + 25.10 * c1 + 24 * c2 + 19.07 * c3 - 0.5 * s0 - 25.10 * 0.5 * s1 - 24 * 0.5 * s2 - 19.07 * 0.5 * s3 \leq 22.67; \end{bmatrix}$$

$FREE(c0);\ FREE(c1);\ FREE(c2);\ FREE(c3);\ END$

(5.3) $\widetilde{Y}_i = (24.461; 12.40) + (0.116; 0.0) * BMI + (-0.197; 0.0) * YEAR + (0.369; 0.0) * WCFP$

The data pairs $w_i = (\widetilde{Y}_i, X_{ji})'$ of Table 6, (i = 1,...,50) are used to demonstrate the proposed procedure in case where the crisp input X_{ji}, risp output $\widetilde{Y}_i$ and fuzzy regression coefficients.

Table 6: Some Bootstrap Fuzzy Descriptor Statistics Based on the Resampling of the Error Term Belonging to the Data in Table 1 ($n = 50$).

Variables	Observed	c_j	s_j	**Confidence intervals**	
				95%Confidence Interval	
Constant	−6.973	24.461	12.408	−6.988	−6.964
BMI(X_1)	0.984	0.116	0.00	0.983	0.984
YEAR(X_2)	−0.111	−0.197	0.00	−0.111	0.111
WCFP(X_3)	0.399	0.369	0.00	0.399	0.400
$R^2 = 1.0$, N = 50, SSE = 0,0075, F = 25142495,53**					

DEXATF: total fat (DEXATF) calculated according to DEXA method; BMI: Body Mass Index BMI: Body Mass Index (kg/m^2); YEAR; WCFP: Waist Circumference Fat Percentage.

Estimates of the bootstrap regression coefficients in the form of were calculated. Also, this model explains response variable using fewer variables although there is no procedure available in FLR which can be used as variable selection method.

6. DISCUSSION AND CONCLUSIONS

In this study, using the samples obtained by bootstrap resampling technique in ordinary least squares and fuzzy least squares regression analysis techniques, it has been tried to reveal which of them is more effective to estimate parameters.

Parameter estimates as well as their standard errors and confidence intervals statistics from bootstrapping ordinary least squares regression and bootstrapping fuzzy regression coefficients are presented in Table 7 for prediction of DEXATF (Y) (gr).

Table 7: BOLS and BFLS regression ($n = 50$, $B = 1000$) parameter estimations and the regression coefficients statistics for estimation of some DEXATF (gr).

	Variables	Observed	Average	S.E.	Bias	Confidence intervals	
						95% Confidence Interval	
BOLSR	Constant	−6.973	−6.959	31.809	0.0133	−12.977	−13.997
	BMI(X_1)	0.984	0.9862	0.1246	−0.0020	0.802	12.145
	YEAR(X_2)	−0.111	−0.1129	0.0641	−0.0020	−0.224	−0.005
	WCFP(X_3)	0.399	0.3993	0.0974	0.0005	0.250	0.610
BFLSR	Constant	−6.973	24.461		12.408	−6.988	−6.964
	BMI(X_1)	0.984	0.116	0.00	0.00	0.983	0.984
	YEAR(X_2)	−0.111	−0.197	0.00	0.00	−0.111	0.111
	WCFP(X_3)	0.399	0.369	0.00	0.00	0.399	0.400

DEXATF: total fat (DEXATF) calculated according to DEXA method; BMI: Body Mass Index (kg/m^2); YEAR; WCFP: Waist Circumference Fat Percentage

$B = 10000$ bootstrap samples are generated randomly to reflect the exact behavior of the bootstrap procedure and the distributions of bootstrap regression parameter estimations ($\widetilde{\beta}^{(b)}$) are graphed in Figure 2(a), 2(b), 2(c) (Figures 1). The histograms of the bootstrap estimates conform quite well to the limiting normal distribution for all regression coefficients. Hence, the confidence intervals should be based on that distribution, where B is sufficiently large ($B = 1000$). And, bootstrap fuzzy regressions are generated by putting each one of the observation sets in place in the model and the regression coefficients are estimated as $\widetilde{\beta}^{(b)}$. To reflect the exact behavior of the bootstrap sample procedure the distributions of fuzzy regression parameter estimations $\widetilde{\beta}^{(b)}$ are graphed in Figures 2(d) (Figures 1). The histograms of the bootstrap fuzzy estimates are no similar to the normal distribution for bootstrap OLS regression coefficients.

The fuzzy bootstrap regression standard errors of the BMI and $YEAR$ coefficients are substantially small than the estimated asymptotic OLS and bootstrap OLS standard errors, because of the inadequacy of the bootstrap in small samples. The confidence intervals based on the bootstrap fuzzy regression standard errors are very similar to the percentile intervals of the BMI and GS coefficients; however, the confidence intervals based on the OLS and bootstrap OLS standard errors are quite different from the percentile and confidence intervals based on the bootstrap standard errors. Comparing the bootstrap fuzzy coefficients averages $\widetilde{\beta}_0^{(br)}$, $\widetilde{\beta}_1^{(br)}$ and $\widetilde{\beta}_2^{(br)}$ with the corresponding OLS and bootstrap OLS estimates $\hat{\beta}_0^{(br)}$, $\hat{\beta}_1^{(br)}$, $\hat{\beta}_2^{(br)}$ and $\widetilde{\beta}_0^{(br)}$, $\widetilde{\beta}_1^{(br)}$, $\widetilde{\beta}_2^{(br)}$ and β shows that there is a little bias in the bootstrap coefficients.

The shape of these graphs show that a histogram of the replicates with an overlaid smooth density estimate and the skewness of the distribution of regression parameter estimate from the OLS bootstrap and fuzzy bootstrap replicate. The shape of these graphs show that a histogram of the replicates with an overlaid smooth density estimate and the skewness of the distribution of regression parameter estimate from the OLS bootstrap and fuzzy bootstrap replicate.

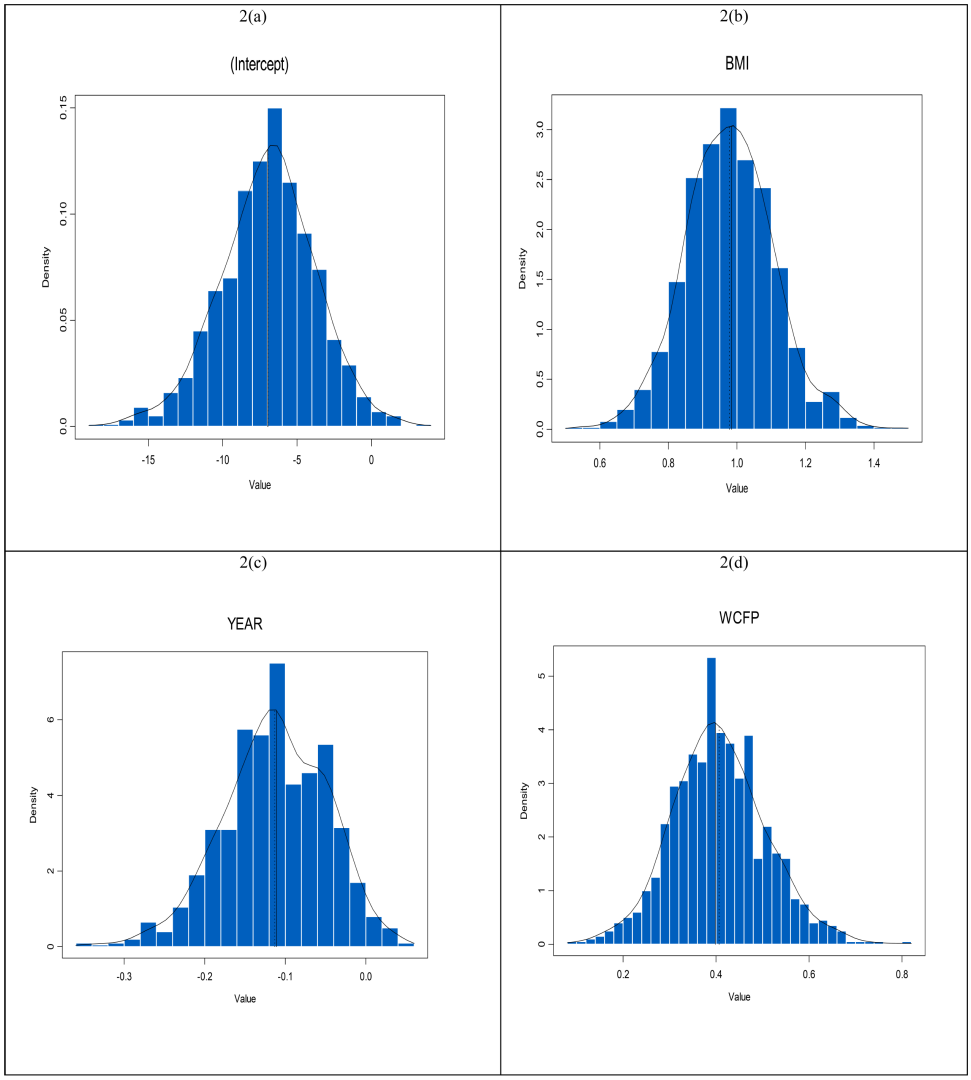


Figure 1: Histogram of bootstrap ($B = 1000$, (a), (b), (c), (d)) regression parameter estimates.

To examine the appearence of the distribution $\left(F\left(\beta^{(b)}\right)\right)$ of the replicates (B=10 000), the distribution plots of $\widetilde{\beta}^{(b)}$ from Equation 4.2 are given in Figures 1. The vertical lines of these plots give the mean of the B bootstrap parameter estimates ($\widetilde{\beta}^{(b)}$) and show the shape of distribution of bootstrap parameter estimates. Although, the larger bootstrap replicates (B) are used, the smoother distribution of $\widetilde{\beta}^{(b)}$ could usually be obtained in these plots (Fox, 1997) [25]. The number of bootstrap replications B depends on the application and size of sample. The bootstrap replications sufficient to be $B = 100$ for standard error estimates, for confidence interval estimates B≅1000, and for standard deviation estimate 50≤ B ≤ 100 were suggested by Leger *et al.* (1992) [40] and Efron (1990) [22]. In fact, it is known from the statistical theory of the bootstrap that a finite total of n^n possible bootstrap samples exist. If it was computed the parameter estimates for each of these n^n samples, it would obtain the true bootstrap estimates of parameters however, such extreme computation is wasteful and unnecessary [53].

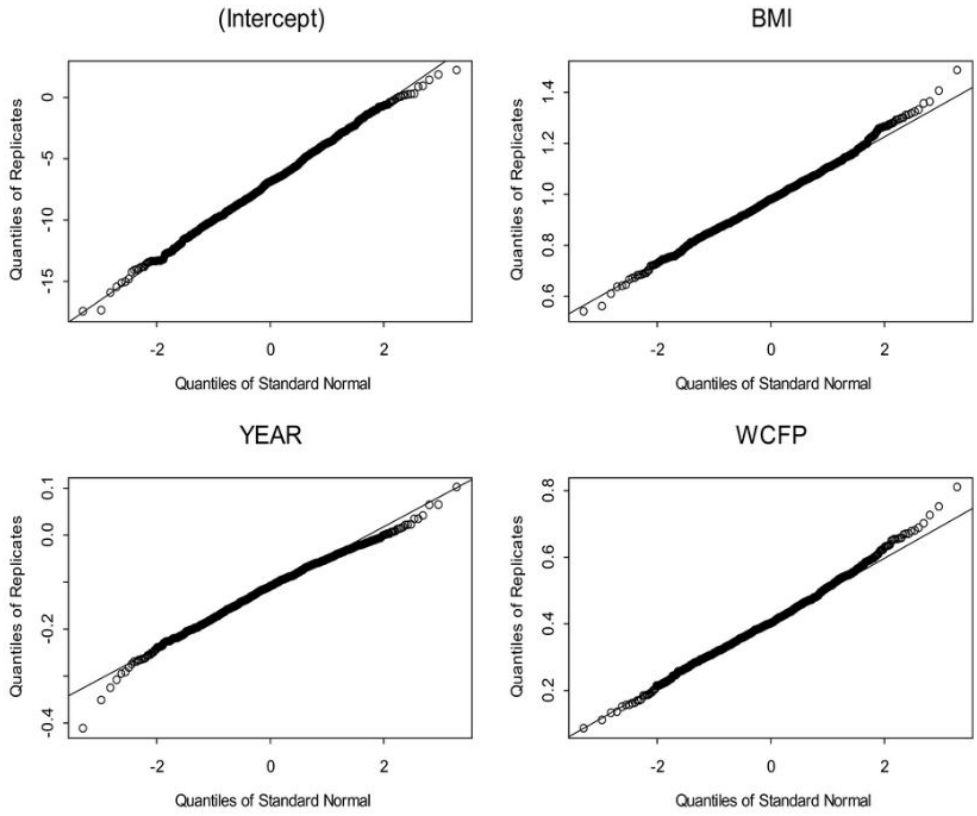


Figure 2: Normal Quantile – Quantile.

The bootstrap resampling technique is one of the most important concepts in statistics introduced. In classical techniques, the bootstrap resampling technique has become a very powerful tool used to estimate quantities associated with the sampling distribution of estimators and test statistics. In application of bootstrap resampling technique, there is often some uncertainty about the certain error structure, and a well-chosen resampling technique can give robust inferences to the certain error structure of the data. Indeed, it is harmful to pretend that mere calculation can replace thought about central issues such as the structure of a problem, the type of answer required, the sampling design and data quality. In these cases, for linear regression with normal random errors εj having constant variance, the least squares theory of regression estimation and inference provides clean, exact and optimal methods for analysis.

For generalizations to non-normal errors and non-constant variance, precise methods seldom exist, and we are faced with approximate techniques based on linear approximations to estimators and central limit theorems. Bootstrap resampling technique have the potential to provide more accurate and more valid analysis for modelling in complex problems. With ordinary least squares linear regression, in ideal conditions resampling essentially not only reproduces the exact theoretical analysis, but also offers the potential to deal with non-ideal circumstances such as non-constant variance. Despite its extent and usefulness, resampling technique should be carefully applied. Unless certain basic ideas are understood, it is all too easy to produce a wrong solution to the problem. Bootstrap resampling techniques are intended to help avoid tedious calculations based on questionable assumptions.

In conclusion, in this study it is aimed to describe: basic ideas, the standard error of bootstrap fuzzy regression technique, confidence intervals of the regression coefficients, application to bootstrap ordinary least squares technique and bootstrap fuzzy linear regression technique. The fuzzy regression technique is a new statistical technique that combines the classical regression technique with the theory of fuzzy logic. When functional relationship is not known in advance, fuzzy regression technique is introduced as an alternative technique which helps model crisp/crisp or crisp/fuzzy data. Also, the correct functional relationship between dependent variable and independent variable is not known. Bootstrap resampling technique is preferable in fuzzy least squares regression analysis and ordinary least squares regression techniques because of some theoretical properties like having any distributional assumptions on the residuals and hence, allows for inference even if the errors do not follow normal distribution.

The most important advantages of the bootstrap fuzzy least squares regression technique and bootstrap ordinary least squares technique are:

- they need smallers sample than ordinary least squares technique,
- they can be used when there are doubts about the distribution of the population,
- they can be used in cases of insufficient sample size and parametric assumptions are not realized,
- they can also be used in cases where the sample selection is not random,
- in cases of very large sample sizes, the methods can be applied by creating subgroups.

The bootstrap fuzzy least squares regression and bootstrap ordinary least squares techniques estimate the variation of a statistic from the variation of that statistic between subsamples, rather than from parametric assumptions and may yield similar results in many situations. However, it is a mistake to expect that bootstrap fuzzy least squares regression technique and bootstrap ordinary least squares regression technique always give valid and confident results. The confidence of results depend on the structure of the data and distribution function. Application of both regression techniques depend on development of statistical computer packages featured these analyses.

The estimations of the bootstrap standard error and confidence intervals of the regression coefficients are nearly equal to the standard error of regression coefficient estimates of the bootstrap fuzzy least squares regression technique. However, bootstrap fuzzy least squares regression technique gives regression coefficients, which generally have smaller standard errors and narrow confidence intervals than bootstrap regression technique. If the OLSR model did not satisfy the related model assumptions, the bootstrap regression technique and bootstrap fuzzy least squares regression techniques could be used for fitting the model and provide better estimates. Because the bootstrap resampling technique and bootstrap fuzzy least squares regression technique do not require above assumptions [7], [27]. Due to the computation of the standard errors and since confidence intervals are based on the distribution of bootstrap samples, not on assumptions about normal distributions. The assumption guessed behind bootstrap resampling technique is to treat the sample as if it were the entire population [4], [7].

In this research, we have presented model of fuzzy least squares regression for the literature. We have shown that the development of an adequate bootstrap resampling theory in the fuzzy context would be very profitable because in this context the asymptotic approximations are, in most cases, difficult to handle and hence, they are useless to make inferences. A real application to predict $DEXATF(Y)(g)$ in clinical data obtained was shown. $BOLS$ and $BFLS$ regression were obtained and also as can be seen from the statistical values calculated from a clinical numerical sample, the error of the $BFLS$ method regarding the estimates calculated according to the error criteria was detected to be lower than the errors calculated from the $BOLS$ method. Due to these results, we trust the results obtained with the $BFLSR$ method more than the results obtained with the $BOLSR$ method. It can be concluded that $BOLSR$ and $BFLSR$ methods have similar performance. Among these models, the $BFLSR$ method is proposed to be preferred. Although the bootstrap resampling technique is sometimes mentioned as a replacement for "standard statistics techniques", it is concluded that this thought is wrong, since the bootstrap resampling technique depends on the theoric elements of classic logic.

ACKNOWLEDGMENTS

The author is thankful to the referee and the editor for their constructive comments that led to a significant improvement of the paper.

ETHICS COMMITTEE APPROVAL

The study was conducted in accordance with the principles of the Declaration of Helsinki. The research permission was obtained from Erciyes University Ethics Committee (dated 12/02/2008 and number 2008/613) and Kocaeli University Ethics Committee (Date: 10.03.2009, Number: 2009/48). Informed consent was obtained from study participants.

DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

REFERENCES

[1] Akbari, M.G. and Rezaei, A. (2009). Bootstrap statistical inference for the variance based on fuzzy data, *Austrian Journal of Statistics*, **38**(2), 121–130.

[2] Akbari, H. and Matthews, H.D. (2012). Global cooling updates: Reflective roofs and pavements, *Energy and Buildings*, **55**, 2–6.

[3] Akbari, M. and Hesamian, H. (2019). A partial-robust-ridge-based regression model with fuzzy predictors-responses, *Journal of Computational and Applied Mathematics*, **351**, 290–301.

[4] Ariani, D.; Yuki, N.N. and Yuniarti, D. (2017). Perbandingan Metode Bootstrap Dan Jackknife Resampling Dalam Menentukan Nilai Estimasi Dan Interval Konfidensi Parameter Regresi., *EKSPONENSIAL*, **8**(1), 43–49.

[5] Arefi, M. (2020). Quantile fuzzy regression based on fuzzy outputs and fuzzy parameters, *Soft Computing*, **24**(1), 311–320.

[6] Asadolahi, M.; Akbari, M.; Hesamian, G. and Arefi, M. (2021). A robust support vector regression with exact predictors and fuzzy responses, *International Journal of Approximate Reasoning*, **132**, 206–225.

[7] Beran, R. (2003). The Impact of the bootstrap on statistical algorithms and theory, *Statistical Science*, **18**(2), 175–184

[8] Celmins, A. (1987). Least squares model fitting to fuzzy vector data, *Fuzzy Sets and Systems*, **22**(3), 245–269.

[9] Chang, Y.H.O. and Ayyub, B.M. (2001). Fuzyy regression methotds-a comparative assessment, *Fuzzy Sets and Systems*, **119**(2), 187–203.

[10] Chen, Y.S. (2001). Outliers detection and confidence interval modification in fuzzy regression, *Fuzzy Sets and Systems*, **119**(2), 259–272.

[11] Casella, G. (2003). Introduction to the silver anniversary of the bootstrap, *Statistical Science*, **18**(2), 133–134.

[12] Chachi, J. (2019). A weighted least squares fuzzy regression for crisp input-fuzzy output data, *IEEE Transactions on Fuzzy Systems*, **27**(4), 739–748.

[13] Chachi, J.; Kazemifard, A. and Jalalvand, M.J. (2021). A multi-attribute assessment of fuzzy regression models, *Iranian Journal of Fuzzy Systems*, **18**(4), 131–148.

[14] Chachi, J. and Chaji, A. (2021). An OWA-based approach to quantile fuzzy regression, *Computers & Industrial Engineering*, **159**, 107498.

[15] Chukhrova, N. and Johannssen, A. (2019). Fuzzy regression analysis: Systematic review and bibliography, *Applied Soft Computing*, **84**, 105708.

[16] Diciccio, T. and Tibshirani, R. (1987). Bootstrap confidence intervals and bootstrap approximations, *Journal of the American Statistical Association*, **82**(397), 163–170.

[17] Diamond, P. (1987). *Least squares fitting of several fuzzy variables*, Preprints of Second IFSA World Congress, Tokyo, Japan, pp. 329–331.

[18] Diamond, P. (1988). Fuzzy least squares, *Information Science*, **46**(3), 141–157.

[19] D'Urso, P. and Massari, R. (2013). Weighted least squares and least median squares estimation for the fuzzy linear regression analysis, *Metron*, **71**(3), 279–306.

[20] Efron, B. (1979). Bootstrap methods: Anoher look at the jackknife, *The Annals of Statistics*, **7**, 1–26.

[21] Efron, B. (1982). The jackknife, the bootstrap and other resampling plans, *Society For Industrial And Applied Mathematics*, **38**, 29–35.

[22] Efron, B. (1990). More efficient bootstrap computations, *Journal of the American Statistical Association*, **85**(409), 79–89.

[23] Efron, B. and Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*, CRC Press, Boca Raton, Florida, pp. 60–82.

[24] Efron, B. (2003). Second thought on bootstrapping, *Statistical Science*, **18**(2), 135–140.

[25] Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*, Sage Publications, Thousand Oaks, CA.

[26] Friedl, H. and Stampfer, E. (2002). *Resampling Methods*, Encyclopedia of Environmetrics, 3 (A. El-Shaarawi, W. Piegorsch, Eds.), Wiley, Chichester, 1768–1770.

[27] Ferraro, M.; Coppi, R.; Rodriguez, G.G. and Colubi, A. (2010). A linear regression model for imprecise response, *International Journal of Approximate Reasoning*, **51**(7), 759–770.

[28] Ferraro, M. and Giordani, P. (2013). A proposal of robust regression for random fuzzy sets, *Synergies of Soft Computing and Statistics for Intelligent Data Analysis*, Springer, Berlin, Heidelberg, 115–123.

[29] Hong, D.H.; Song, J.K. and Young, H. (2001). Fuzzy least-squares linear regression analysis using shape preserving operations, *Information Sciences*, **138**, 185–193.

[30] Hojati, M.; Bector, C.R. and Smimou, K.A. (2005). Simple method for computation of fuzzy linear regression, *European Journal of Operational Research*, **166**, 172–184.

[31] Hung, W.L. and Yang, M.S. (2019). n omission approach for detecting outliers in fuzzy regressions models, *Fuzzy Sets and Systems*, **157**, 3109–3122.

[32] Hesamian, G. and Akbari, M.G. (2021). A robust multiple regression model based on fuzzy random variables, *Journal of Computational and Applied Mathematics*, **388**, 1–13.

[33] Hesamian, G. and Akbari, M.G. (2020). A robust varying coefficient approach to fuzzy multiple regression model, *Journal of Computational and Applied Mathematics*, **375**, 113270.

[34] Jimenez-Gamero, M.D.; Pino-Mejias, R. and Rojas-Medar, M.A. (2005). A bootstrap test for the expectation of fuzzy random variables, *UNICAMP*.

[35] Kim, B. and Bishu, R.R. (1998). Evaluation of fuzzy linear regression models by comparing membership functions. *Fuzzy Sets and Systems*, **100**(1–3), 343–352.

[36] Kula, K. and Apaydin, A. (2008). Fuzzy robust regression analysis based on the ranking of fuzzy sets, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **22**(3), 245–269.

[37] Khammar, A.H.; Arefi, M. and Akbari, M.G. (2020). A robust least-squares fuzzy regression model based on kernel function, *Iranian Journal of Fuzzy Systems*, **17**(4), 105–119.

[38] Khammar, A.H.; Arefi, M. and Akbari, M.G. (2021). A general approach to fuzzy regression models based on different loss functions, *Soft Computing*, **25**(2), 835–849.

[39] Khammar, A.H.; Arefi, M. and Akbari, M.G. (2021). Quantile fuzzy varying coefficient regression based on kernel function, *Soft Computing*, **107**, 107313.

[40] Leger, C.; Politis, D.N. and Romano, J.P. (1992). Bootstrap technology and applications, *Technometrics*, **34**(4), 378–397.

[41] Leski, J.M. and Kotas, M. (2015). The statistical inferences of fuzzy regression based on bootstrap techniques, *Fuzzy Sets and Systems*, **279**, 112–129.

[42] Lee, W.J.; Jung, H.Y.; Yoon, J.H. and CHOI, S.H. (2015). A partial-robust-ridge-based regression model with fuzzy predictors-responses, *Soft Computing*, **19**, 883–890.

[43] Moskowitz, H. and Kim, K. (1993). On assesing the H value in fuzzy linear regression, *Fuzzy Sets and Systems*, **58**(3), 303–327.

[44] Mooney, C.Z. and Duval, R.D. (1993). Bootstrapping: A nonparametric approach to statistical inference, *SAGE Pub.*, London, 80.

[45] Ming, M.; Friedman, M. and Kandel, A. (1997). General fuzzy least squares, *Fuzzy Sets and Systems*, **88**, 107–118.

[46] Montenegro, M.; Colubi, A.; Rosa Casals, M. and Gil, M.A. (2004). Asymptotic and bootstrap techniques for testing the expected value of a fuzzy random variable, *Metrika*, **59**(1), 31–49.

[47] Mohammed, Z.R. (2008). Bootstrapping: A nonparametric approach to identify the effect of sparsity of data in the binary regression models, *Journal of Applied Sciences, Network for Scientific Information, Asian*, **17**(8), 2991–2997.

[48] Nasrabadi, M.M. and Nasrabadi, E.A. (2004). Mathematical-programming approach to fuzzy linear regression analysis, *Applied Mathematics and Computation*, **155**(3), 873–881.

[49] Nasrabadi, E.; Hashemi, S.M. and Ghatee, M. (2007). An LP-based approach to outliers detection in fuzzy regression analysis. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **15**, 441–456.

[50] Nasrabadi, E. and Hashemi, S.M. (2008). Robust fuzzy regression analysis using neural networks, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **16**, 597–598.

[51] Peters, G. (1994). Fuzzy linear regression with fuzzy intervals, *Fuzzy Sets and Systems*, **63**, 45–55.

[52] Sadeghpour Gildeh, B. and Rahimpour, S. (2013). Fuzzy bootstrap test for the mean and variance with Dp, q-distance, *International Journal of Machine Learning and Computing*, **3**, 21–23.

[53] Stine, R. (1990). Modern methods of data analysis, *SAGE Pub.*, Scotland, 325–373.

[54] Shao, J. (1996). Bootstrap model selection, *Journal of the American statistical Association*, **91**(434), 655–665.

[55] Sahinler, S. and Topuz, D. (2007). Bootstrap and Jackknife resampling algorithms for estimation of regression parameters, *Journal of Applied Quantitative Methods*, **2**, 188–199.

[56] Shakouri, H.; Nadimi, R. and Ghaderi, F. (2009). A hybrid TSK-FR model to study short-term variations of the electricity demand versus the temperature changes, *Expert Systems with Applications*, **36**, 1765–1772.

[57] Shakouri, H. and Nadimi, R. (2013). Outlier detection in fuzzy linear regression with crisp input-output by linguistic variable view, *Applied Soft Computing*, **13**, 734–742.

[58] Sungkono, J. (2015). Bootstrap resampling observasi pada estimasi parameter regresi menggunakan Software R, *MAGISTRA*, **27**, 92.

[59] Tanaka, H.; Uejima, S. and Asai, K. (1982). Linear regression analysis with fuzzy model, IEEE Transactions on Systems, *Man and Cyberneties*, **12**(6), 903–907.

[60] Tanaka, H. and Watada, J. (1988). Possibilistic linear systems and their application to the linear regression model, *Fuzzy Sets and Systems*, **27**(3), 275–289.

[61] Tanaka, H.; Hayashi, I. and Watada, J. (1989). Possibilistic linear regression analysis for fuzzy data *European Journal of Operational Research*, **40**, 389–396.

[62] Taheri, S.M. and Chachi, J. (2020). A robust variable-spread fuzzy regression model. In "Recent Developments and the New Direction in Soft-Computing Foundations and Applications", Springer, Cham. 309–320.

[63] Varga, S. (2007). Robust estimations in classical regression models versus robust estimations in fuzzy regression models, *Kybernetika*, **43**, 503–508.

[64] Wang, H.F. and Tsaur, R.C. (2000). Insight of a fuzzy regression model, *Fuzzy Sets and Systems*, **112**, 355–369.

[65] Wu, C.F.J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis, *Annals of Statistics*, **14**, 1261–1295.

[66] Yang, Z.; Yin, Y. and Chen, Y. (2013). Robust fuzzy varying coefficient regression analysis with crisp inputs and gaussian fuzzy output, *Journal of Computing Science and Engineering*, **7**, 263–271.

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 433–454
https://doi.org/10.57805/revstat.v22i4.500

Estimation and Diagnostic for a Partially Linear Regression Based on an Extension of the Rice Distribution

Authors: J. C. S. Vasconcelos
– Exact Sciences Department, University of São Paulo, Brazil
juliocezarvasconcelos@hotmail.com

E. M. M. Ortega ✉
– Exact Sciences Department, University of São Paulo, Brazil
edwin@usp.br

G. M. Cordeiro
– Statistics Department, Federal University of Pernambuco, Brazil
gausscordeiro@gmail.com

J. S. Vasconcelos
– Department of Rural Engineering, Paulista State University, Brazil
julianojsv@gmail.com

M. A. M. Biaggioni
– Department of Rural Engineering, Paulista State University, Brazil
m.biaggioni@unesp.br

Received: December 2020 Revised: November 2022 Accepted: November 2022

Abstract:

- We introduce an extension of the Rice distribution and estimate its parameters by maximum likelihood. We define two regressions based on this extended distribution to model volumetric shrinkage of the wood and milk production. The performance of the parameter estimators is investigated in finite samples using Monte Carlo simulations. Also, we propose the quantile residuals for the regression models whose empirical distribution is close to normality. The usefulness of the new regressions is proved empirically through two applications to agricultural data.

Keywords:

- *cubic splines; milk production; regression extensions; simulation study; wood data.*

AMS Subject Classification:

- 62.

✉ Corresponding author

1. INTRODUCTION

The Rice distribution [17] is generally observed when the global magnitude of a vector is related to its direction components, such as when wind speed is analyzed in two directions, i.e., a two-dimensional component vector. If the components are independent and normally distributed with equal variances, the general wind speed has a Rice distribution. It is also used to model dispersion (or variability) of line-of-sight transmission between two stations which applies to FM radio waves, microwaves, magnetic resonance images in the presence of noise and satellite transmissions. It is also employed to model Rician fading, which describes how the cancellation of signals affects the propagation of radio waves. In recent years, it has been utilized by various authors for different applications. For example, [8] introduced a generalized Rice distribution based on the linearity characteristics of a system to model situations where the maximum amplitude is close to the signal's mean amplitude; [12] presented a Bayesian approach to estimate its parameters, and [26] studied a new approach to analyze images based on the maximum likelihood method that permits obtaining simultaneous estimates of the image and signal noise.

The Rice distribution is relatively unknown in the area of applied statistics. One of the objectives of this paper is to generalize the Rice distribution to be applied in different research areas. We emphasize that the papers mentioned previously do not provide regressions which have been widely employed in many fields. A fundamental conjecture that should be examined with caution regarding a data set is that when the covariates express nonlinear effects on the response variable and adopting a parametric regression may not be a suitable alternative. To overcome this circumstance, generalized semiparametric models have been proposed. For example, [7] and [9] introduced the generalized additive model (GAM) which aggregates the properties of generalized linear models with additive models; and [18] demonstrated that nonparametric regression can be considered an interesting extension of the parametric regression, and the two can be combined to produce the semiparametric regression. Another model widely applied in recent years is the generalized additive model for location, scale, and shape (GAMLSS) [19]. Various authors have published papers involving partially linear regressions. [22] introduced the symmetric generalized partial linear model; [24] presented an extension of the log-normal distribution from two perspectives, one of which was the partially linear regression; and [11] proposed a new strategy to select Bayesian models and an efficient estimation method for partially linear model.

Based on these contexts, the objectives of this paper are described below. We define the *odd log-logistic Rice* (OLLRc) distribution that can be applied to model bimodal, trimodal and asymmetric data. Based on this distribution, we introduce a parametric regression with two systematic components and illustrate its flexibility using volumetric shrinkage wood data, see for example [23]. We propose a new partially linear regression and show its utility by analyzing milk production in the Northeast of Brazil. We prove that the empirical distribution of the quantile residuals (qrs) for both regressions has approximately the standard normal distribution. We provide two applications to real data (shrinkage volume of three wood species and milk production data) to illustrate the flexibility of the OLLRc partially linear regression model.

The remaining sections are as follows. Section 2 defines the OLLRc distribution, and provides some mathematical properties. Section 3 introduces a parametric regression based on the new distribution. Section 4 proposes the OLLRc partially linear regression and performs some simulations for the distribution of the penalized maximum likelihood estimators. Simulation results for the residuals are reported in Section 5. The usefulness of the new regressions is proved by means of two real data sets in Section 6. Section 7 ends with some concluding remarks.

2. THE ODD LOG-LOGISTIC RICE DISTRIBUTION

If G denotes a baseline distribution, the cumulative distribution function (cdf) of the the odd log-logistic-G (OLL-G) (OLL-G) generator [5] is defined by

$$F(y) = \int_0^{\frac{G(y)}{\bar{G}(y)}} \frac{\nu\, u^{\nu-1}}{(1+u^{\nu})^2} du = \frac{G(y)^{\nu}}{G(y)^{\nu} + [1-G(y)]^{\nu}}, \quad y > 0, \tag{2.1}$$

where $\nu > 0$ is the shape parameter

This class of generalized distributions has been deeply investigated in the last years; see, for example, the references in [16] and [24]. The probability density function (pdf) corresponding to (2.1) can be expressed as

$$f(y) = \frac{\nu\, g(y)\{G(y)[1-G(y)]\}^{\nu-1}}{\{G(y)^{\nu} + [1-G(y)]^{\nu}\}^2}, \tag{2.2}$$

where $g(y) = d\,G(y)/d\,y$ is the baseline density.

The Rice cdf with two parameters $\mu > 0$ and $\sigma > 0$ is (for $y > 0$)

$$G_{\mu,\sigma}(y) = 1 - Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right), \tag{2.3}$$

where $Q_1(a, b)$ is the Marcum Q-function, namely

$$Q_M(a,b) = \int_b^{\infty} x\left(\frac{x}{a}\right)^{M-1} \exp\left(-\frac{x^2+a^2}{2}\right) I_{M-1}(ax)dx,$$

I_{M-1} is the modified Bessel function of the first kind of order $M-1$ (for $\eta \in \mathbb{R}, \eta \neq 0$),

$$I_{\eta}(z) = \sum_{m=0}^{\infty} \frac{(-1)^m}{m!\,\Gamma(m+\eta+1)} \left(\frac{z}{\eta}\right)^{2m+\eta},$$

and $\Gamma(\cdot)$ is the gamma function. The Marcum Q-function is defined in the `VGAM` package of **R** software. For more details, see [27].

The pdf corresponding to (2.3) has the form

$$g_{\mu,\sigma}(y) = \frac{y}{\mu^2} \exp\left(-\frac{y^2+\sigma^2}{2\mu^2}\right) I_0\left(\frac{\sigma y}{\mu^2}\right), \tag{2.4}$$

where $I_0(z) = \sum_{m=0}^{\infty} z^{2m}/[4^m\,(m!)^2]$.

The Rice distribution can be obtained following a simple extension of the Rayleigh distribution. Let $X = \sqrt{T_1^2 + T_2^2}$, where $T_1 \sim \mathrm{N}(\delta_1, \mu^2)$ and $T_2 \sim \mathrm{N}(\delta_2, \mu^2)$ are independent random variables. Then, X has the Rice density (2.4), where $\sigma = \sqrt{\delta_1^2 + \delta_2^2}$. If $\sigma = 0$, then (2.4) is just the Rayleigh density. So, the parameter μ in the Rice distribution is the standard deviation of two Gaussian contributions and σ represents a distance term. The Rice distributions tends to the $\mathrm{N}(\sigma, \mu^2)$ distribution if $\sigma y/\mu^2$ goes to ∞.

The OLLRc cdf (for $y > 0$) is defined by taking $G(x)$ in (2.1) as the Rice cdf (2.3)

$$F(y) = \frac{\left[1 - Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)\right]^{\nu}}{\left[1 - Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)\right]^{\nu} + Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)^{\nu}} \tag{2.5}$$

The OLLRc density function follows by inserting (2.3) and (2.4) in Equation (2.2)

$$\begin{aligned} f(y) = {} & \frac{\nu\, y}{\mu^2} \exp\left\{\frac{-(y^2+\sigma^2)}{2\mu^2}\right\} I_0\left(\frac{y\,\sigma}{\mu^2}\right) \times \left\{\left[1 - Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)\right] Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)\right\}^{\nu-1} \\ & \times \left\{\left[1 - Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)\right]^{\nu} + Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu}\right)^{\nu}\right\}^{-2}, \end{aligned} \tag{2.6}$$

where all parameters are positive. The OLLRc density function can be expressed as a combination of exponentiated Rice densities (see Appendix A).

Henceforth, let $Y \sim \mathrm{OLLRc}(\mu, \sigma, \nu)$ be a random variable with density (2.6). Some shapes of (2.6) reported in Figure 1 reveal that the density of Y is very flexible for bimodal and trimodal data.

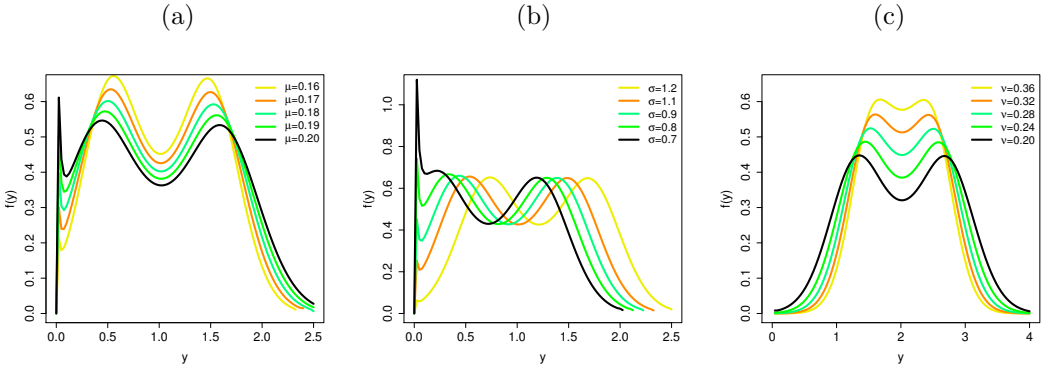


Figure 1: Plots of the OLLRc density. (a) For $\sigma = 1$ and $\nu = 0.18$ varying μ. (b) For $\mu = 0.16$ and $\nu = 0.17$ varying σ. (c) For $\mu = 0.25$ and $\sigma = 2$ varying ν.

By inverting Equation (2.5), the quantile function (qf) of Y, say $y = H(u) = F^{-1}(u)$, is

$$y = H(u) = H_{\mathrm{Rice}}\left\{\frac{u^{1/\nu}}{u^{1/\nu} + (1-u)^{1/\nu}}\right\}, \qquad u \in (0,1), \tag{2.7}$$

where $H_{\mathrm{Rice}}(u) = G^{-1}_{\mu,\sigma}(u)$ is the Rice qf.

Figure 2 displays plots of the density of Y and histograms from two simulated data sets with $100,000$ replications. They show that the simulated values are consistent with the OLLRc distribution, where we note trimodal and bimodal shapes.

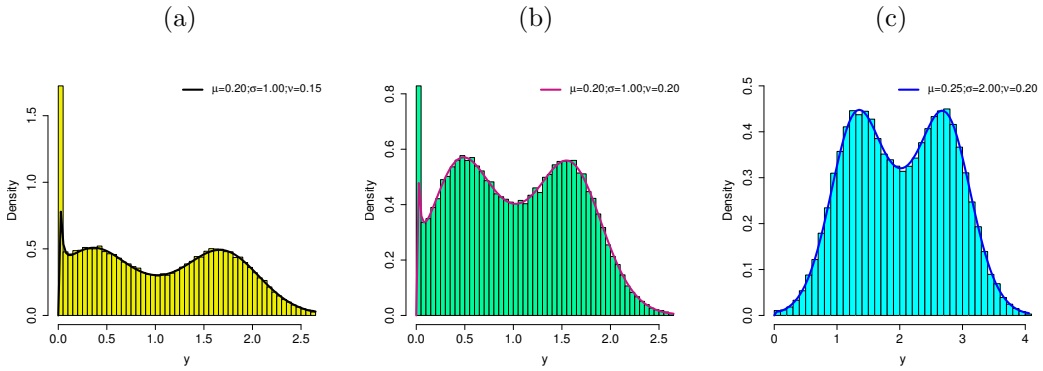


Figure 2: Histograms and plots of the OLLRc density.

The influence of the shape parameter ν on the skewness and kurtosis of Y can be easily investigated based on quantile measures. Figure 3(a) displays the Bowley skewness

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

whereas Figure 3(b) provides the Moors kurtosis

$$M = \frac{H(3/8) - H(1/8) + H(7/8) - H(5/8)}{H(6/8) - H(2/8)}.$$

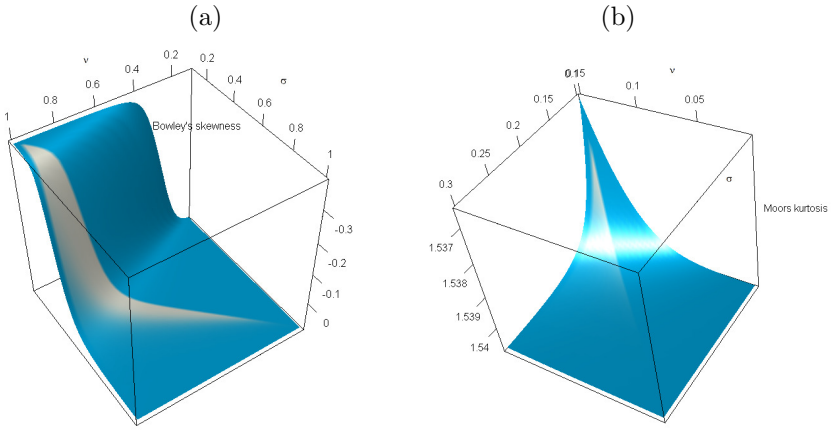


Figure 3: (a) Bowley's skewness. (b) Moors kurtosis.

By varying $\nu \in [0.1, 1]$, Figure 3(a) displays the Bowley skewness of Y for $\mu = 0.1$ and $\sigma \in [0.1, 1]$, whereas Figure 3(b) reports the Moors kurtosis of Y for $\mu = 0.1$ and $\sigma \in [0.1, 0.3]$.

3. THE OLLRC REGRESSION

The OLLRc regression is defined by two systematic components considering that the parameters μ_i and σ_i in the density (2.6) are given by (for $i = 1, ..., n$)

$$\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_1) \qquad \text{and} \qquad \sigma_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta}_2), \tag{3.1}$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \cdots, \beta_{1p})^\top$ and $\beta_2 = (\beta_{21}, \cdots, \beta_{2p})^\top$ are vectors of unknown coefficients and $\mathbf{x}_i^\top = (x_{i1}, \cdots, x_{ip})$ is a vector of covariates associated with the ith observation.

The OLLRc regression includes two special models: the Rice (for $\nu = 1$) and Rayleigh (for $\nu = 1$ and $\sigma_i = 0$) regressions.

The log-likelihood function for the vector $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \nu)^\top$ from a random sample $(y_1, \mathbf{x}_1), \cdots, (y_n, \mathbf{x}_n)$ has the form

$$\begin{aligned} l(\boldsymbol{\theta}) \;=\; & n\log(\nu) + \sum_{i=1}^{n} \log\left(\frac{y_i}{\mu_i}\right) - \sum_{i=1}^{n}\left(\frac{y_i^2 + \sigma_i^2}{2\mu_i^2}\right) + \sum_{i=1}^{n} \log\left[I_0\left(\frac{y_i\sigma_i}{\mu_i}\right)\right] + \\ & (\nu - 1)\sum_{i=1}^{n} \log\left\{\left[1 - Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)\right] Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)\right\} - \\ & 2\sum_{i=1}^{n} \log\left\{\left[1 - Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)\right]^{\nu} + Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)^{\nu}\right\}. \end{aligned} \tag{3.2}$$

The maximum likelihood estimate (MLE) $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be calculated by maximizing (3.2) using the **R** software and standard likelihood techniques can be adopted for inference purposes. Initial values for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ can be taken from the fitted Rice regression model (when $\nu = 1$).

Under conditions that are fulfilled for the parameter vector $\boldsymbol{\theta}$ in the interior of the parameter space but not on the boundary, the asymptotic distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is multivariate normal $N_{2p+1}(0, K(\boldsymbol{\theta})^{-1})$, where $K(\boldsymbol{\theta})$ is the information matrix. The asymptotic covariance matrix $K(\boldsymbol{\theta})^{-1}$ of $\widehat{\boldsymbol{\theta}}$ can be approximated by the inverse of the $(2p + 1) \times (2p + 1)$ observed information matrix $-\ddot{\mathbf{L}}(\boldsymbol{\theta})$, whose elements can be calculated numerically. The approximate multivariate normal distribution $N_{2p+1}(0, -\ddot{\mathbf{L}}(\boldsymbol{\theta})^{-1})$ for $\widehat{\boldsymbol{\theta}}$ can be used in the classical way to construct approximate confidence regions for some parameters in $\boldsymbol{\theta}$.

We can use likelihood ratio (LR) statistics for comparing some special models with the OLLRc regression model in the usual way. Further details are given by [14].

3.1. Simulation studies

Five thousands Monte Carlo simulations are carried out in the **R** software to examine the consistency of the MLEs under two scenarios: the OLLRc distribution and the OLLRc regression. By setting $n = 25$, 80, 160 and 320, a random sample is drawn from the OLLRc(μ, σ, ν) distribution, and the MLEs are calculated in each of these replications. For the regression scenario, we also consider 700.

The OLLRc distribution

We generate observations from the OLLRc distribution using (2.7) and $u \sim U(0,1)$ with $\mu = 0.15$, $\sigma = 1$ and $\nu = 0.2$. We calculate the MLEs in each of the 5,000 simulations and then the average estimates (AEs), biases, and means squared errors (MSEs). The results in Table 1 indicate that the estimates are accurate since their biases and MSEs converge to zero when n increases.

Table 1: Simulation findings from the OLLRc distribution.

Parameter	$n = 25$			$n = 80$		
	AE	Bias	MSE	AE	Bias	MSE
μ	0.177	0.027	0.017	0.160	0.010	0.004
σ	0.971	−0.029	0.024	0.994	−0.006	0.003
ν	0.269	0.069	0.073	0.227	0.027	0.018
Parameter	$n = 160$			$n = 320$		
	AE	Bias	MSE	AE	Bias	MSE
μ	0.154	0.004	0.001	0.152	0.002	0.000
σ	0.998	−0.002	0.001	0.998	−0.002	0.000
ν	0.211	0.011	0.006	0.208	0.008	0.002

The OLLRc regression

Consider the OLLRc regression with $\mu_i = \exp(\beta_{10} + \beta_{11}x_{i1} + \beta_{12}x_{i2})$ and $\sigma_i = \exp(\beta_{20} + \beta_{21}x_{i1} + \beta_{22}x_{i2})$ and fixed parameters $\beta_{10} = -2$, $\beta_{11} = 0.7$, $\beta_{12} = 1.8$, $\beta_{20} = 0.6$, $\beta_{21} = -0.8$ and $\beta_{22} = 0.4$.

For the generation process, we consider: $Y_i \sim \text{OLLRc}(\mu_i, \sigma_i, \nu)$, $x_{i1} \sim \text{Bernoulli}(0.5)$ and $x_{i2} \sim U(0,1)$. The simulation results from the fitted OLLRc regression in Table 2 indicate that the AEs go to the true parameters and that the biases and MSEs tend to vanish when n increases in agreement with first-order asymptotic theory.

4. THE OLLRC PARTIALLY LINEAR REGRESSION

The dependent variables can be influenced by explanatory variables with linear and non-linear effects in many areas. Recently, several works have been published related to regression models, for example, [2], [28], [13], [24], [22], [4], [25], among others.

In this context, considering the penalized smoothing based on the cubic-spline, we construct the partially linear regression based on the OLLRc distribution. The systematic component for the parameter μ_i in terms of the explanatory variables $\mathbf{x}_1, ..., \mathbf{x}_p$ (linear effects) and $\mathbf{t} = (t_i)$ (non-linear effect) has the form (for $i = 1, ..., n$)

$$\mu_i = \exp\Big\{\mathbf{x}_i^\top \boldsymbol{\beta}_1 + h(t_i)\Big\}, \tag{4.1}$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, ..., \beta_{1p})^\top$ is the unknown parameter vector and $h(\cdot)$ is an unknown smooth function of t_i.

Table 2: Simulation findings from the OLLRc regression with $\nu = 0.6$ and $\nu = 2$.

n	Parameter	$\nu = 0.6$			$\nu = 2$		
		AE	Bias	MSE	AE	Bias	MSE
25	β_{10}	−2.438	−0.438	0.530	−2.439	−0.439	0.632
	β_{11}	0.671	−0.029	0.141	0.736	0.036	0.159
	β_{12}	2.075	0.275	0.625	2.142	0.342	0.625
	β_{20}	0.580	−0.020	0.008	0.596	−0.004	0.002
	β_{21}	−0.681	0.119	0.046	−0.737	0.063	0.018
	β_{22}	0.434	1.234	0.041	0.406	1.206	0.012
	ν	0.546	−0.054	0.086	1.951	−0.049	0.387
80	β_{10}	−2.118	−0.118	0.088	−2.081	−0.081	0.113
	β_{11}	0.673	−0.027	0.032	0.700	<0.001	0.036
	β_{12}	1.836	0.036	0.092	1.859	0.059	0.099
	β_{20}	0.597	−0.003	0.002	0.600	<0.001	<0.001
	β_{21}	−0.747	0.053	0.020	−0.800	<0.001	0.006
	β_{22}	0.405	1.205	0.016	0.400	1.200	0.003
	ν	0.568	−0.032	0.023	2.023	0.023	0.161
160	β_{10}	−2.050	−0.050	0.035	−2.016	−0.016	0.051
	β_{11}	0.691	−0.009	0.015	0.699	−0.001	0.018
	β_{12}	1.819	0.019	0.041	1.811	0.011	0.049
	β_{20}	0.600	0.000	0.001	0.600	<0.001	<0.001
	β_{21}	−0.781	0.019	0.012	−0.807	−0.007	0.004
	β_{22}	0.399	1.199	0.008	0.398	1.198	0.001
	ν	0.590	−0.010	0.009	2.036	0.036	0.078
320	β_{10}	−2.021	−0.021	0.016	−2.007	−0.007	0.025
	β_{11}	0.697	−0.003	0.007	0.703	0.003	0.008
	β_{12}	1.809	0.009	0.019	1.803	0.003	0.023
	β_{20}	0.600	<0.001	0.001	0.600	<0.001	<0.001
	β_{21}	−0.796	0.004	0.007	−0.804	−0.004	0.002
	β_{22}	0.400	1.200	0.004	0.400	1.200	0.001
	ν	0.597	−0.003	0.004	2.020	0.020	0.038
700	β_{10}	−2.007	−0.007	0.007	−2.002	−0.002	0.011
	β_{11}	0.702	0.002	0.003	0.701	0.001	0.004
	β_{12}	1.803	0.003	0.009	1.802	0.002	0.011
	β_{20}	0.600	<0.001	<0.001	0.600	<0.001	<0.001
	β_{21}	−0.805	−0.005	0.004	−0.803	−0.003	0.001
	β_{22}	0.400	1.200	0.002	0.400	1.200	<0.001
	ν	0.601	0.001	0.002	2.011	0.011	0.017

For the partially linear regression (4.1), we consider the penalty based on the second order derivative of the function $h(\cdot)$ [15].

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^\top, \sigma, \nu)^\top$ be the parameter vector related to the parametric part and $\lambda > 0$ be the smoothing parameter that controls the smoothness of the curve. Consider a smooth function $h(t)$ (second order differentiable function in the interval $[a, b]$), such that it is a cubic smoothing splines where the nodes are the ordered values of $t_1, ..., t_n$, say $t_1^0 < t_2^0 < ... < t_q^0$, and q indicates the amount of distinct values for the explanatory variable t_i that is controlled in a non-parametric way.

The penalty described above can be expressed in matrix notation [6]. Let d_i be the distance between two subsequent and different control points called nodes i and $i+1$, i.e., $d_i = t_{i+1}^0 - t_i^0$ (for $i = 1, ..., q-1$). We define the elements q_{ij} (for $i = 1, ..., q$ and $j = 2, ..., q-1$)

of the $q \times (q-2)$ tridiagonal matrix $\mathbf{A}$ by

$$q_{j-1,j} = d_{j-1}^{-1}, \quad q_{jj} = -d_{j-1}^{-1} - d_j^{-1}, \quad q_{j+1,j} = d_j^{-1} \quad \text{and} \quad q_{ij} = 0 \quad \text{for} \quad |i-j| \geq 2.$$

The elements r_{ij} (for $i = 2, ..., q-1$ and $j = 2, ..., q-1$) of the $(q-2) \times (q-2)$ symmetric matrix $\mathbf{B}$ are

$$\begin{aligned} r_{ii} &= \frac{1}{3}(d_{i-1} + d_i) \quad \text{for} \quad i = 2, ..., q-1, \\ r_{i,i+1} &= r_{i+1,i} = \frac{1}{6} d_i \quad \text{for} \quad i = 2, ..., q-2 \quad \text{and} \\ r_{ij} &= 0 \quad \text{for} \quad |i-j| \geq 2. \end{aligned}$$

Further, let $\mathbf{K} = \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^{\mathbf{T}}$ be a $q \times q$ positive definite matrix. The parameters associated with the covariates in the linear and nonlinear effects ($\boldsymbol{\theta}$ and $\mathbf{h} = (h(t_1^0), ..., h(t_q^0))$, respectively) are determined by maximizing the penalized log-likelihood function

$$\begin{aligned} l_p(\boldsymbol{\theta}, \mathbf{h}) &= n \log(\nu) + \sum_{i=1}^{n} \log\left(\frac{y_i}{\mu_i}\right) - \sum_{i=1}^{n}\left(\frac{y_i^2 + \sigma^2}{2\mu_i^2}\right) + \sum_{i=1}^{n} \log\left[I_0\left(\frac{y_i \sigma}{\mu_i}\right)\right] + \\ &\quad (\nu - 1) \sum_{i=1}^{n} \log\left\{\left[1 - Q_1\left(\frac{\sigma}{\mu_i}, \frac{y_i}{\mu_i}\right)\right] Q_1\left(\frac{\sigma}{\mu_i}, \frac{y_i}{\mu_i}\right)\right\} - \\ &\quad 2 \sum_{i=1}^{n} \log\left\{\left[1 - Q_1\left(\frac{\sigma}{\mu_i}, \frac{y_i}{\mu_i}\right)\right]^{\nu} + Q_1\left(\frac{\sigma}{\mu_i}, \frac{y_i}{\mu_i}\right)^{\nu}\right\} - \frac{\lambda}{2} \mathbf{h}^T \mathbf{K}\, \mathbf{h}, \end{aligned} \tag{4.2}$$

where λ is the unknown smoothing parameter. The maximization of (4.2) is equivalent to the cubic smoothing spline. We use the gamlss(·) function from the `gamlss` [20] package to implement the OLLRc regression, and calculate the penalized maximum likelihood estimates (PMLEs). The cs(·) function is used to model the nonlinear effect based on cubic smoothing splines function [21].

4.1. Simulations for the OLLRc partially linear regression

We present a Monte Carlo study with a smooth function to verify the adequacy of the PMLEs in this regression. The covariates (linear and non-linear effects) and the response variable are generated as follows: $x_{i1} \sim U(0,1)$, $x_{i2} \sim$ Bernoulli(0.5), $t_{i3} \sim U(0, 0.045)$ (scenario 1), $t_{i3} \sim U(0, 0.03)$ (scenario 2) and $y_i \sim \text{OLLRc}(\mu_i, \sigma, \nu)$.

Further, the systematic component is $\mu_i = \exp\{0.5x_{i1} - 0.4x_{i2} + h(t_{i3})\}$, where $\beta_{11} = 0.5$, $\beta_{12} = -0.4$, $\sigma = 0.1$ and $\nu = 0.7$ (for $n = 60, 180,$ 400 and 700). In addition, 1,000 Monte Carlo samples are generated and, for each sample size, the PMLEs of the parameters are found for each replication, and then the AEs, biases and MSEs are calculated. The numbers in Table 3 indicate that the AEs tend to the true parameters and the biases and MSEs converge to zero when n increases. Thus, the sample distribution of the PMLEs is approximately normal.

Regarding the analysis of the nonlinear effect (t_{i3}), the true smooth curves $h(t_{i3}) = \sin(-50t_{i3}\pi) + \cos(30\pi t_{i3})$ and $h(t_{i3}) = \cos(100t_{i3}\pi) + \tan(15\pi t_{i3} - 1)$ and their respective estimated curves (based on 1,000 simulations) are displayed in Figure 4 for both scenario 1. We note that the estimated curves approach to the true curve for large sample sizes (as expected).

Table 3: Findings for the OLLRc partially linear regression.

n	Parameter	Scenario 1			Scenario 2		
		AE	Bias	MSE	AE	Bias	MSE
60	β_{11}	0.508	0.008	0.186	0.495	0.005	0.171
	β_{12}	−0.415	−0.015	0.064	−0.407	−0.007	0.055
	σ	0.186	0.086	0.016	0.134	0.034	0.004
	ν	0.636	−0.064	0.028	0.700	0.000	0.024
180	β_{11}	0.518	0.018	0.047	0.498	−0.002	0.048
	β_{12}	−0.409	−0.009	0.015	−0.405	−0.005	0.014
	σ	0.136	0.036	0.004	0.104	0.004	0.001
	ν	0.687	−0.013	0.005	0.730	0.030	0.007
400	β_{11}	0.510	0.010	0.018	0.504	0.004	0.019
	β_{12}	−0.404	−0.004	0.006	−0.403	−0.003	0.006
	σ	0.112	0.012	0.002	0.095	−0.005	0.001
	ν	0.703	0.003	0.002	0.737	0.037	0.004
700	β_{11}	0.505	0.005	0.011	0.494	−0.006	0.011
	β_{12}	−0.406	−0.006	0.004	−0.405	−0.005	0.003
	σ	0.100	0.000	0.001	0.090	−0.010	0.001
	ν	0.710	0.010	0.001	0.740	0.040	0.003

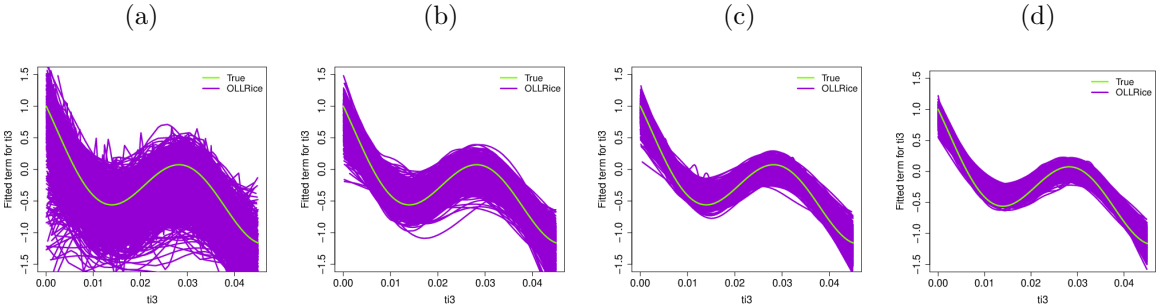


Figure 4: The generated and estimated curves for $h(t_{i3})$ (scenario 1): (a) $n = 60$, (b) $n = 180$. (c) $n = 400$. (d) $n = 700$.

5. RESIDUAL ANALYSIS AND SIMULATIONS

We define the quantile residuals (qrs) [3] for the OLLRc regression as

$$qr_i = \Phi^{-1}\left\{ \frac{\left[1 - Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)\right]^{\nu}}{\left[1 - Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)\right]^{\nu} + Q_1\left(\frac{\sigma_i}{\mu_i}, \frac{y_i}{\mu_i}\right)^{\nu}} \right\}, \tag{5.1}$$

where $\Phi(\cdot)^{-1}$ is the inverse of the standard normal cdf and μ_i and σ_i are given in Equation (3.1).

Simulations for the OLLRc regression

Some simulations of sizes 25, 80, 160, 320 and 700 are performed using the algorithm of Section 3.1 to examine the empirical distribution of these residuals. Figure 5 (for $\nu = 0.6$)

show that this distribution becomes closer to the standard normal distribution when n increases.

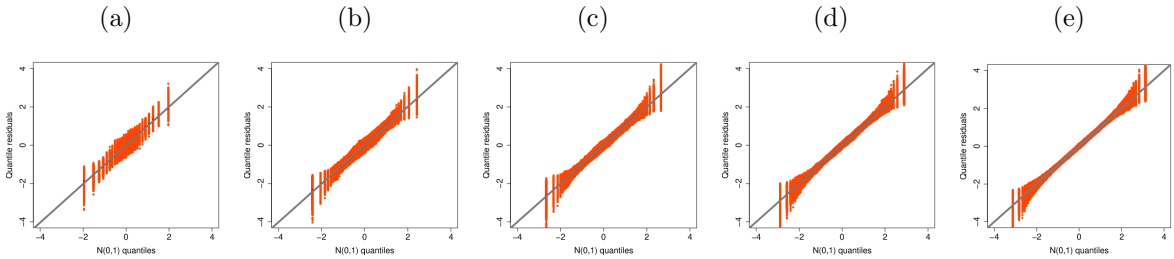


Figure 5: Normal probability plots of the qrs ($\nu = 0.6$). (a) $n = 25$. (b) $n = 80$. (c) $n = 160$. (d) $n = 320$. (e) $n = 700$.

Simulations for the OLLRc partially linear regression

Consider a simulation study to investigate the empirical distribution of the qrs for the OLLRc partially linear regression by generating 60, 180, 400 and 700 observations from Equation (4.1). Normal probability plots in Figure 6 reveal that the empirical distribution of the qrs is close to the standard normal for all samples, and the approximation becomes better when n increases.

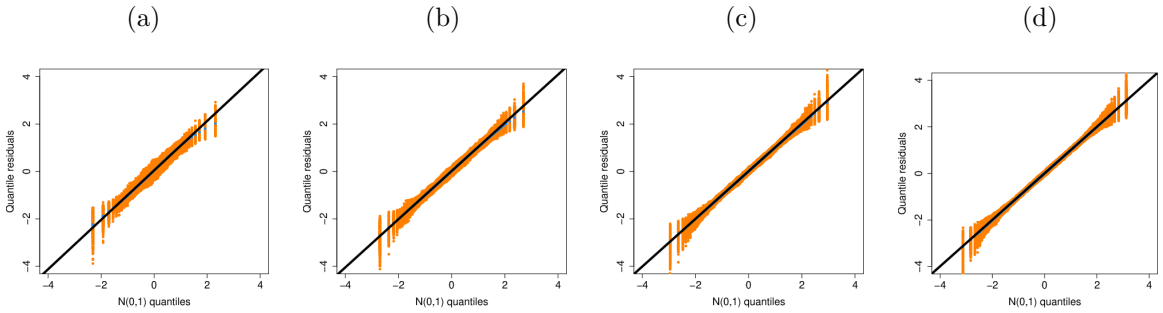


Figure 6: Normal probability plots of the qrs (scenario 1). (a) $n = 60$. (b) $n = 180$. (c) $n = 400$. (d) 700.

Thus, we use normal probability plots for the residuals (qr_i) with simulated envelopes for both models, as suggested by [1], as follows:

1. Fit the model and generate a sample of n independent observations using the fitted model as if it were the true model;
2. Fit the model to the generated sample using the data set ($\mathbf{x}_i$) and compute the values of the residuals;
3. Repeat steps (1) and (2) m times;
4. Obtain ordered values of the residuals, $qr^*_{(i)v}$, $i = 1, ..., n$ and $v = 1, ..., m$;
5. Consider n sets of the m ordered statistics and for each set compute the mean, minimum and maximum values;

6. Plot these values and the ordered residuals of the original sample against the normal scores. The minimum and maximum values of the m ordered statistics provide the envelope.

The residuals outside the limits provided by the simulated envelope require further investigation. Additionally, if a considerable proportion of points falls outside the envelope, we have evidence against the adequacy of the fitted model. Plots of the residuals against the fitted values can also be useful.

6. APPLICATIONS

We present two applications of the new regressions: the first for the OLLRc regression, and the second for the OLLRc partially linear regression.

6.1. Shrinkage volume data

We consider a data set referring to the shrinkage volume of three wood species: Cedrilho (Erismauncinatum Warm), Morototoni (ScheffleramorotoniAubl) and Pinus (Pinus spp). The shrinkage volume of wood is defined as the phenomenon related to the dimensional variation of wood due to moisture exchange with the surrounding environment until a condition of balance is attained, called the hygroscopic equilibrium moisture. The variations in the dimensions of wood specimens occur when they lose or gain moisture in relation to the saturation point of the fibers, which in general is in the range of 28% to 30% water. The dimensional variation involves either shrinkage or swelling. The shrinkage volume of wood varies widely among species, depending on the drying method and the behavior of the particular wood specimen, occasionally leading to alterations of shape and the formation of cracks and warping. Special precaution needs to be taken in situations that require wood stability. For structural framework, flooring, doors, door/window frames and furniture, cracking and warping can cause serious losses, requiring replacement. Thus correct drying methods to attain equilibrium moisture are essential. There are various explanations for the increase of contraction with higher temperature. One of them can be the reduction of the equilibrium moisture, but that factor has been experimentally found to cause an increase in contraction of less than 1%, when in reality the increase in contraction is much more than this. For these reasons, we study the effects of drying temperature and wood species on the shrinkage volume of wood specimens.

The experiment was carried out in the first half of 2020 at the School of Agronomic Sciences of Paulista State University (UNESP), located in the city of Botucatu, São Paulo, Brazil. The tests were carried out with wood specimens with volume of 20 cm^3 dried in a muffle furnace at final temperatures of 300 ^{o}C and 500 ^{o}C. A muffle furnace is type of oven that operates at high temperatures used in laboratories. The final temperature was applied for 10 minutes and the carbonization rate was 14.3 ^{o}C/min for each species. We used a pachymeter to measure the dimensions of each specimen to calculate the volume before and after carbonization in stable conditions. The variables involved are the following (for

$i = 1, ..., 36$): y_i: volumetric shrinkage (in cm^3); x_{i1}: temperature (0=300^{o}C, 1=500^{o}C) and x_{i2}: wood species (0=Cedrilho, 1=Morototoni, 2=Pinus) with two dummy variables (d_{i1}, d_{i2}).

First, we provide a marginal analysis of the response variable. Table 4 reports the MLEs (their standard errors in parentheses) of the parameters from the fitted OLLRc, Rice and Rayleigh distributions, and the statistics: Akaike Information Criterion (AIC) and Global Deviance (GD). These results indicate that the OLLRc distribution is the best model to the current data.

Table 4: Findings from the fitted distributions.

Distribution	$\log(\mu)$	$\log(\sigma)$	ν	AIC	GD
OLLRc	**−0.666**	**2.064**	**0.155**	**157.836**	**151.836**
	(0.109)	**(0.031)**	**(0.024)**		
Rice	0.804	2.021	1	162.263	158.263
	(0.124)	(0.052)	(—)		
Model	$\log(\mu)$	σ	ν		
Rayleigh	1.756	0	1	181.017	179.017
	(0.083)	(—)	(—)		

The likelihood ratio (LR) statistics in Table 8 indicate that the OLLRc distribution is the best model to these data among the three distributions. The estimated pdf of the fitted models in Figure 7 show that the OLLRc distribution gives the best fit to the shrinkage volume data.

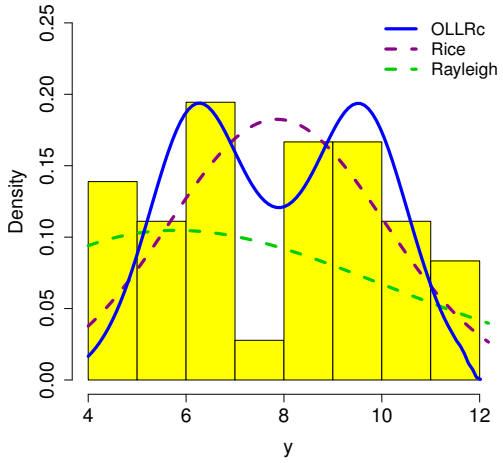


Figure 7: Estimated OLLRc, Rice and Rayleigh densities.

The OLLRc regression

The systematic components are given by

$$\mu_i = \exp(\beta_{10} + \beta_{11}x_{i1} + \beta_{12}d_{i1} + \beta_{13}d_{i2})$$

and

$$\sigma_i = \exp(\beta_{20} + \beta_{21}x_{i1} + \beta_{22}d_{i1} + \beta_{23}d_{i2}).$$

The MLEs, SEs and p-values from the fitted OLLRc regression to the current data are reported in Table 5. Some conclusions are addressed at the end of this application.

Table 5: Findings from the fitted OLLRc regression to the shrinkage volume data.

Parameter	Estimate	SE	p-value
β_{10}	−0.537	2.509	0.832
β_{11}	−0.346	0.191	0.082
β_{12}	−1.144	0.541	0.044
β_{13}	−0.786	0.219	0.001
β_{20}	1.687	0.135	<0.001
β_{21}	0.448	0.129	0.002
β_{22}	0.080	0.042	0.068
β_{23}	0.286	0.037	<0.001
ν	0.236	0.908	

The AIC and GD values in Table 6 confirm that the OLLRc regression is the best model to the shrinkage volume data.

Table 6: Adequacy statistics.

Regression	AIC	GD
OLLRc	**93.236**	**75.2356**
Rice	95.164	79.164
Rayleigh	178.373	170.373

Table 9 compares the new regression with two special models fitted to the wood volumetric retraction data, whose figures indicate that the OLLRc regression is the best model among the three.

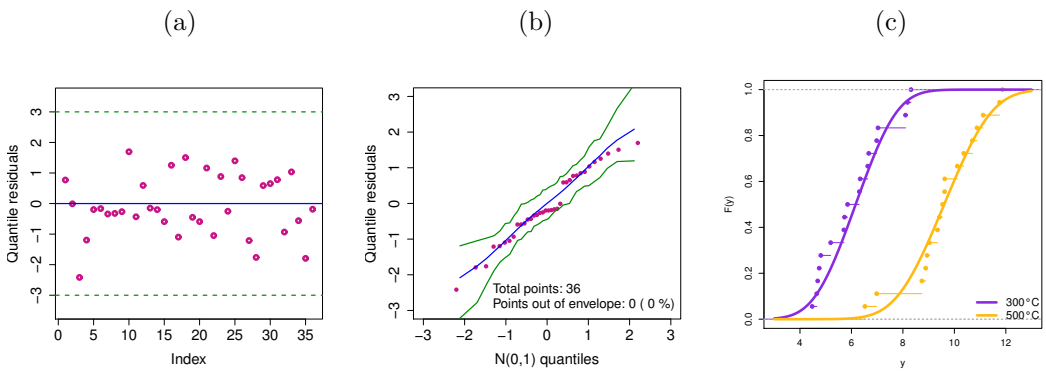


Figure 8: (a) Index plot of the qrs. (b) Normal probability plot for the qrs. (c) Empirical and estimated cdf for the temperature (300^{o}C, 500^{o}C).

Figures 8a and 8b display the index plot of the qrs and the normal probability plot with generated envelope for the OLLRc regression, respectively. These plots do not provide departure from the model assumptions. We report in Table 7 the presence of significant effects of the levels of wood from the fitted OLLRc regression to the shrinkage volume data.

Table 7: Findings for the three wood levels from the fitted OLLRc regression.

Test for link μ			
Hypotheses H_0	Estimate	SD	p-value
Cedrilho = Morototoni	−1.144	0.541	0.044
Cedrilho = Pinus	−0.786	0.219	0.001
Morototoni = Pinus	0.358	0.563	0.531

Test for link σ			
Hypotheses H_0	Estimate	SD	p-value
Cedrilho = Morototoni	0.080	0.042	0.068
Cedrilho = Pinus	0.286	0.037	<0.001
Morototoni = Pinus	0.206	0.023	<0.001

Assessing covariate effects on parameter μ

For a 5% significance level, we conclude:

The temperature levels are not statistically different for μ (see Table 5). The Cedrilho and Morototoni wood species and Cedrilho and Pinus wood species are statistically different for μ (see Table 5). The Pinus and Morototoni wood species are statistically different for μ (see Table 7).

Assessing covariate effects on parameter σ

The temperature levels are statistically different for σ (see Table 5 and Figure 8c). The Cedrilho and Morototoni wood species are not statistically different for σ (see Table 5). The Cedrilho and Pinus wood species are statistically different for σ (see Table 5). Morototoni and Pinus wood species are statistically different for σ (see Table 7).

Finally, the empirical and estimated cdf of the OLLRc regression are displayed in Figure 8c for different levels of temperatures, thus showing that this regression is suitable for the shrinkage volume data.

Table 8: LR tests (Application 1 without considering covariates).

Distributions	Hypotheses	LR statistic	p-value
OLLRc vs Rice	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	6.428	0.011
OLLRc vs Rayleigh	$H_0 : \sigma = 0$ and $\nu = 1$ vs $H_1 : H_0$ is false	27.182	<0.001

Table 9: LR statistics for three fitted regressions (Application 1).

Regressions	Hypotheses	LR statistic	p-value
OLLRc vs Rice	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	3.929	0.047
OLLRc vs Rayleigh	$H_0 : \sigma = 0$ and $\nu = 1$ vs $H_1 : H_0$ is false	95.137	<0.001

Table 10: LR tests (Application 2 without considering covariates).

Regressions	Hypotheses	LR statistic	p-value
OLLRc vs Rice	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	5.062	0.024
OLLRc vs Rayleigh	$H_0 : \nu = 0$ and $\nu = 1$ vs $H_1 : H_0$ is false	48.693	<0.001

Table 11: LR statistics for three fitted regressions (Application 2).

Models	Hypotheses	LR statistic	p-value
OLLRc vs Rice	$H_0 : \nu = 1$ vs $H_1 : H_0$ is false	96.007	<0.001
OLLRc vs Rayleigh	$H_0 : \sigma = 0$ and $\nu = 1$ vs $H_1 : H_0$ is false	148.996	<0.001

6.2. Milk production data

In the second application, the data referred to the quantity of cold milk, raw or homogenized, acquired (thousand liters) between the first quarter of 2005 until the fourth quarter of 2015 in Northeast Brazil. The Northeast is considered a new dairy farming region due to the expanded market for milk and dairy products in Brazil, including the Northeast itself, in recent years, driven by increased consumption, in turn related to rising purchasing power in the region, as well as stronger demand from other regions of Brazil and neighboring countries. To understand the advance of dairy farming in the Northeast, it is necessary to know something about the division of the region in terms of climate. Basically there are four sub-regions: the forest zone, sub-humid zone (agreste), mid-north and hinterland (sertão). Each of them has distinct physical characteristics that facilitate or hamper dairy farming. In this paper we study the states of Bahia and Pernambuco. The data set is obtained from the Brazilian Institute of Geography and Statistics [10]. The dependent variable is the production of milk produced, while the explanatory variables are: (i) state of production (Bahia and Pernambuco, two large producers of milk in the Northeast); and (ii) the quarter of production, between the first quarter of 2005 to the last quarter of 2015. This last covariable has a nonlinear effect on the quantity of cold milk. An option to analyze this data set is by means of the OLLRc partially linear regression. The variables under study are: y_i: production of cold milk (raw or homogenized) (thousand liters) (this variable was divided by 10,000); x_{i1}: states (Bahia and Pernambuco) and t_{i2}: quarter (from 1 to 44), for $i = 1, ..., 88$.

Table 12: Findings from the OLLRc partially linear regression.

Model	$\log(\mu)$	$\log(\sigma)$	ν	AIC	GD
OLLRc	**−0.146**	**1.905**	**0.326**	**372.189**	**366.189**
	(0.324)	**(0.028)**	**(0.158)**		
Rice	0.722	1.870	1	375.251	371.251
	(0.081)	(0.036)	(—)		
Model	$\log(\mu)$	σ	ν		
Rayleigh	1.616	0	1	416.881	414.881
	(0.053)	(—)	(—)		

We examine these data by studying the distribution of the response variable, that is, making a marginal analysis. Table 12 reports the results from three fitted distributions, which indicate that the OLLRc distribution can be chosen as the best model.

The numbers in Table 10 support the OLLRc distribution as the best fit model for these data. Figure 9 displays the estimated pdfs of the fitted models and shows that the wider distribution is the best for the current data.

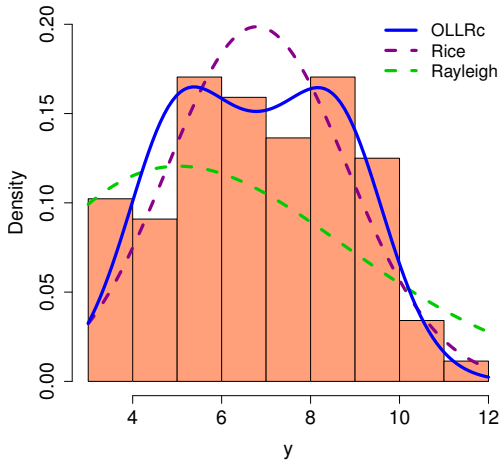


Figure 9: Estimated densities of the OLLRc, Rice and Rayleigh distributions.

Hence, the OLLRc distribution is a good candidate for modeling milk production data.

The OLLRc partially linear regression

Figure 10 displays the scatter plot between the response variable y_i and the covariate t_{i2}. So, there is a non-linear trend between these two variables, which requires the OLLRc partially linear regression model to analyze the current data.

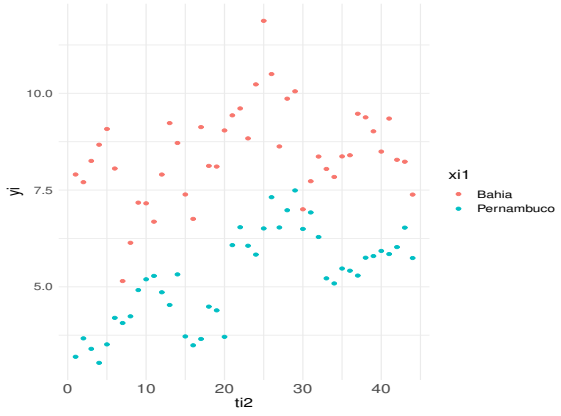


Figure 10: Scatter diagrams: production of cold milk versus quarter.

We now consider the systematic component:

$$\mu_i = \exp[\beta_{10} + \beta_{11} x_{i1} + h(t_{i2})],$$

where $h(\cdot)$ is an arbitrary smooth function associated with the explanatory variable t_{i2}; for more details, see Section 4.

Table 13 reports the generalized Akaike information criterion (GAIC), based on [19], and confirms that the OLLRc partially linear regression can be chosen as the best model.

Table 13: Model selection measures.

Model	GAIC
OLLRc	**256.802**
Rice	350.809
Rayleigh	401.797

Table 14 provides several quantities obtained from the fitted of this partially regression to the milk production data. There is a significant difference between the states of Bahia and Pernambuco in relation to milk production since the covariate x_{i1} is significant at a level of 5%.

Table 14: Findings from the fitted OLLRc partially linear regression.

Parameter	Estimate	SE	p-Value
β_{10}	1.882	0.036	<0.001
β_{11}	−0.476	0.031	<0.001
$\log(\sigma)$	−11.080	398.130	
ν	4.334	0.387	

The figures in Table 11 from two LR tests indicate that the wider regression is the best model for these data. Figure 11(a) provides the plots of the qrs against the observations index, whereas Figure 11(b) reports the normal probability plot with generated envelope. These plots support the wider linear regression for these data and that there are no observations falling outside the envelope.

Finally, Figure 11c provides the estimate of the non-linear effect. The vertical axis refers to the values of t_{i2} and the horizontal axis to the contribution of the estimated smooth curve to the values of t_i. We note from this plot that the amount of milk production is non-linear in relation to the quarter effect. In addition, a greater amount of milk production is achieved between quarters 20 to 35 (approximately).

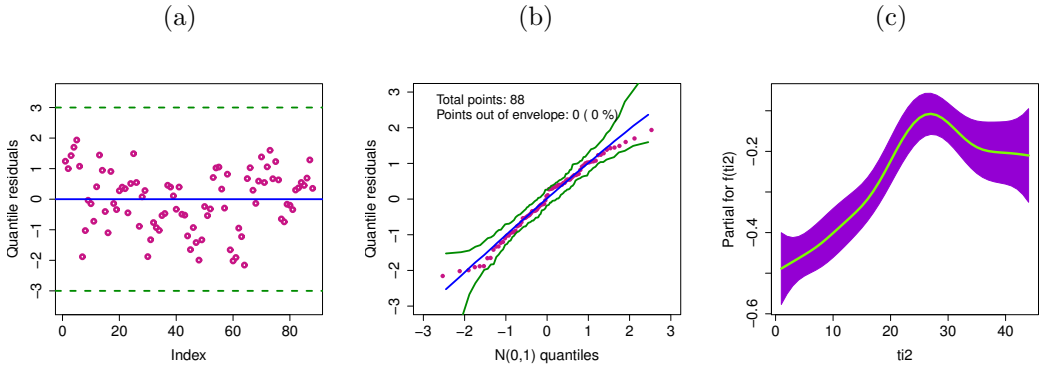


Figure 11: (a) Index plot of the qrs. (b) Normal probability plot for the qrs. (c) Smooth curve fitted from the OLLRc partially linear regression.

7. CONCLUDING REMARKS

The article presented the *odd log-logistic Rice* (OLLRc) distribution and proposed two regression models based on this distribution for the analysis of data that has no unimodal shape. We believe that this paper shows the first use of the Rice distribution in the context of a regression with two systematic components. We defined quantile residuals and provided some simulation studies. We proved the utility of the distribution and of the regressions by means of two data sets on the volumetric shrinkage of the wood and milk production. Future work can be developed using the new OLLRc model in different areas of research. Further, it may be of interest to propose a heteroscedastic semiparametric regression model based on the OLLRc distribution.

A. APPENDIX

Two power series follow for the numerator and denominator of (2.1) (for $\nu > 0$ real):

$$G(y)^{\nu} = \sum_{k=0}^{\infty} a_k \, G(y)^k \quad \text{and} \quad [1 - G(y)]^{\nu} = \sum_{k=0}^{\infty} (-1)^k \binom{\nu}{k} G(y)^k, \tag{A.1}$$

where

$$a_k = a_k(\nu) = \sum_{j=k}^{\infty} (-1)^{k+j} \binom{\nu}{j} \binom{j}{k}.$$

Inserting (A.1) in Equation (2.1) leads to

$$F(y) = \frac{\sum_{k=0}^{\infty} a_k \, G(y)^k}{\sum_{k=0}^{\infty} b_k \, G(y)^k} = \sum_{k=0}^{\infty} c_k \, G(y)^k \tag{A.2}$$

where $b_k = a_k + (-1)^k \binom{\nu}{k}$ (for $k \geq 0$), $c_0 = a_0/b_0$ and the coefficients c_k's (for $k \geq 1$) are calculated recursively as

$$c_k = b_0^{-1} \left(a_k - \sum_{r=1}^{k} b_r \, c_{k-r} \right).$$

By differentiating (A.2), the pdf of Y follows as

$$f(y) = \sum_{k=0}^{\infty} c_{k+1} \, h_{k+1}(y), \tag{A.3}$$

where $h_{k+1}(x) = (k+1) \, G(y)^k \, g(y)$ is the exponentiated-G (exp-G) density function with power parameter $k+1$.

Hence, the exp-Rice density can be expressed from (2.3) and (2.4) as

$$h_{k+1}(y) = \left[1 - Q_1\left(\frac{\sigma}{\mu}, \frac{y}{\mu} \right) \right]^k \frac{(k+1)y}{\mu^2} \exp\left(-\frac{y^2 + \sigma^2}{2\mu^2} \right) I_0\left(\frac{y\sigma}{\mu^2} \right) \tag{A.4}$$

The mathematical properties of the OLLRc distribution can be determined numerically by combining (A.3) and (A.4) for a given number of terms (say 10) in the linear combination.

ACKNOWLEDGMENTS

We would like to thank the Editor, and two anonymous reviewers for their time and valuable remarks. This work was supported by CNPq and CAPES, Brazil.

REFERENCES

[1] Atkinson, A.C. (1985). *Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Clarendon Press Oxford, Oxford.

[2] Cui, X.; Lu, Y. and Peng, H. (2017). Estimation of partially linear regression models under the partial consistency property, *Computational Statistics and Data Analysis*, **115**, 103–121.

[3] Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals, *Journal of Computational and Graphical Statistics*, **5**, 236–244.

[4] Ferreira, C.S.; Paula, G.A. and Lana, G.C. (2022). Estimation and diagnostic for partially linear models with first-order autoregressive skew-normal errors, *Computational Statistics*, **37**, 445–468.

[5] Gleaton, J.U. and Lynch, J.D. (2006). Properties of generalized log-logistic families of lifetime distributions, *Journal of Probability and Statistical Science*, **4**, 51–64.

[6] Green, P.J. and Silverman, B.W. (1993). Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach, CRC Press.

[7] Green, P.J. and Yandell, B.S. (1985). Semi-parametric generalized linear models, *In Generalized Linear Models*, Springer, New York, NY, 44–55.

[8] Hallbjomer, P. (2003). Modified Rice distribution for signals with limited available power, *IEEE antennas and wireless propagation letters*, **2**, 159–162.

[9] Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models, CRC Press, **43**.

[10] IBGE – INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. (2020). Pesquisa Trimestral do Leite: Período de 2005 até 2015, disponível em: `https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/21121-primeiros-resultadios-2leite.html?=&t=series-historicas`

[11] Jeong, S.; Park, T. and van Dyk, D.A. (2022). Bayesian model selection in additive partial linear models via locally adaptive splines, *Journal of Computational and Graphical Statistics*, **31**, 324–336.

[12] Lauwers, L.; Barbé, K.; Van Moer, W. and Pintelon, R. (2009). Estimating the parameters of a Rice distribution: A Bayesian approach. In: *2009 IEEE Instrumentation and Measurement Technology Conference*, IEEE, 114–117.

[13] Manghi, R.F.; Cysneiros, F.J.A. and Paula, G.A. (2019). Generalized additive partial linear models for analyzing correlated data, *Computational Statistics and Data Analysis*, **129**, 47–60.

[14] Sen, P.K. and Singer, J.M. (1993). *Large Sample Methods in Statistics: An Introduction with Applications*, New York, Chapman & Hall.

[15] O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems, *Statistical Science*, **1**, 502–518.

[16] Prataviera, F.; Ortega, E.M.M.; Cordeiro, G.M. and Braga, A.D.S. (2019). The heteroscedastic odd log-logistic generalized gamma regression model for censored data, *Communications in Statistics – Simulation and Computation*, **48**, 1815–1839.

[17] Rice, S.O. (1945). Mathematical analysis of random noise, *The Bell System Technical Journal*, **24**, 46–156.

[18] Ruppert, D.; Wand, M.P. and Carroll, R.J. (2003). Semiparametric Regression, Cambridge University Press, United Kingdom.

[19] Rigby, R.A. and Stasinopoulos, D.M. (2005). Generalized additive models for location, scale and shape, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**, 507–554.

[20] Stasinopoulos, D.M. and Rigby, R.A. (2007). Generalized additive models for location scale and shape (GAMLSS) in **R**. *Journal of Statistical Software*, **23**, 1–46.

[21] Stasinopoulos, M.D.; Rigby, R.A.; Heller, G.Z.; Voudouris, V. and De Bastiani, F. (2017). Flexible Regression and Smoothing: Using gamlss in R, CRC Press, New York.

[22] Vasconcelos, J.C.S. and Villegas, C. (2021a). Generalized symmetrical partial linear model, *Journal of Applied Statistics*, **48**, 557–572.

[23] Vasconcelos, J.C.S.; Cordeiro, G.M.; Ortega, E.M.M. and Rezende, E.M (2021b). A new regression model for bimodal data and applications in agriculture, *Journal of Applied Statistics*, **48**, 349–372.

[24] Vasconcelos, J.C.S.; Cordeiro, G.M.; Ortega, E.M.M. and Biaggioni, M.A. (2020). The parametric and additive partial linear regressions based on the generalized odd log-logistic log-normal distribution, *Communications in Statistics – Theory and Methods*, **51**, 3480–3507.

[25] Vasconcelos, J.C.S.; Cordeiro, G.M. and Ortega, E.M.M. (2022). The semiparametric regression model for bimodal data with different penalized smoothers applied to climatology, ethanol and air quality data, *Journal of Applied Statistics*, **49**, 248–267.

[26] Yakovleva, T.V. and Kulberg, N.S. (2013). Noise and signal estimation in MRI: two-parametric analysis of rice-distributed data by means of the maximum likelihood approach, *American Journal of Theoretical and Applied Statistics*, **2**, 67–79.

[27] Yee, T.W. (2008). The VGAM package, *R News*, **8**, 28–39.

[28] Zeng, Z. and Liu, X (2018). A difference-based approach in the partially linear model with dependent errors, *Journal of Inequalities and Applications*, **267**, 1–164.

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 455–477
https://doi.org/10.57805/revstat.v22i4.521

Applications of Composite lognormal Distributions

Authors: Jiahang Lyu
– Department of Mathematics, University of Manchester, UK
jiahang.lyu@manchester.ac.uk

Saralees Nadarajah ✉
– Department of Mathematics, University of Manchester, UK
mbbsssn2@manchester.ac.uk

Received: December 2021 Revised: November 2022 Accepted: November 2022

Abstract:

- The use of a power law distribution to model upper tails is common in many areas, most notably in physics. In this paper, we consider two data sets published in the physics literature. We show that composite lognormal distributions can provide better fits than the power law distribution even when the former are applied to the full data (as described in the data section) and the latter is applied just to the upper tail of the data.

Keywords:

- *Kolmogorov–Smirnov statistic; probability plot; quantile plot.*

AMS Subject Classification:

- Primary 62E15.

✉ Corresponding author

1. INTRODUCTION

Heavy tails are common in many areas of the sciences and engineering. These are commonly modeled by a power law distribution applied to the upper tail of the data, ignoring the body of the data. Ignoring the body of the data implies loss of information and loss of the power of the model.

The use of the power law distribution to model heavy tails is most common in the physics literature. Two recent papers published in the literature applying the power law distribution to model heavy tails are Campolieti [8] and Balthrop and Quan [3]. Campolieti [8] modeled the top 100 richest net wealth data from Canada Business Magazine. Models of this kind can be used to describe the economy of a country, accurate models of fitting wealth data can give a better prediction of the financial condition of a country. Balthrop and Quan [3] modeled the U.S. cumulative coal production data. Coal productivity is a significantly important factor to the economy of a country. Energy or electricity production relies mostly on coal production. Hence a good model of coal production data is essential for predicting the price of coal.

The aim of this paper is to show that composite lognormal distributions (Nadarajah and Bakar [14, 15]) can be used to model the entirety of the data sets in Campolieti [8] and Balthrop and Quan [3]. In addition, we show that these distributions provide better fits than the power law distribution even when the former are applied to the full data (as described in the data section) and the latter is applied just to the upper tail of the data.

The use of composite lognormal distributions to model data is not new. Cooray and Ananda [9] were the first to suggest the use of composite lognormal distributions. But Nadarajah and Bakar [14, 15] were the first to write an R package (R Core Team [16]) to implement the use of composite lognormal distributions. More recent papers on composite distributions include Calderín-Ojeda [4, 5], Calderín-Ojeda and Kwok [7], Calderín-Ojeda [6], Aminzadeh and Deng [2], Kim *et al.* [11] and Mutali and Vernic [13]. The distributions in these papers have been used to model among others city sizes.

The contents of this paper are organised as follows. Some details of the composite lognormal and power law distributions are given in Section 2. The two data sets and their summary statistics are given in Section 3. The fits of the distributions to the data sets are discussed in Section 4. Finally, some conclusions are noted in Section 5.

2. METHODS

2.1. Composite lognormal distributions

In this section, we discuss the composite lognormal distribution in Nadarajah and Bakar [14, 15]. The composite lognormal distribution is made up by joining together two distinct distributions: one for the body and the other for the tail. The body is described by the

lognormal distribution while the distribution for the tail can be arbitrary. The cumulative distribution function (cdf) of the composite lognormal distribution is

$$F(x) = \begin{cases} \dfrac{1}{1+\phi}\left[\Phi\left(\dfrac{\log\theta-\mu}{\sigma}\right)\right]^{-1}\Phi\left(\dfrac{\log x-\mu}{\sigma}\right), & \text{if } 0 < x \le \theta, \\ \dfrac{1}{1+\phi} + \dfrac{\phi}{1+\phi}\dfrac{F_0(x)-F_0(\theta)}{1-F_0(\theta)}, & \text{if } \theta < x < \infty, \end{cases} \tag{2.1}$$

where $\phi > 0$, θ denotes the point at which the two distributions are joined together, $\Phi(\cdot)$ denotes the cdf of the standard normal distribution, f_0 denotes the probability density function (pdf) of the tail, and F_0 denotes the cdf of the tail. The following conditions ensure that $F(x)$ is continuous and differentiable at θ:

$$\begin{aligned} \mu &= \log\theta + \sigma^2 + \theta\sigma^2\frac{f_0'(\theta)}{f_0(\theta)}, \\ \phi &= \left[\Phi\left(\frac{\log\theta-\mu}{\sigma}\right)\right]^{-1}\frac{1}{\theta\sigma}\psi\left(\frac{\log\theta-\mu}{\sigma}\right)\frac{1-F_0(\theta)}{f_0(\theta)}, \end{aligned} \tag{2.2}$$

where $\psi(\cdot)$ denotes the pdf of the standard normal distribution.

Different choices for f_0 and F_0 lead to different models for the composite lognormal distribution. In Section 4, we consider fourteen different models: the composite lognormal-Fréchet, composite lognormal-log logistic, composite lognormal-generalized Pareto, composite lognormal-Weibull, composite lognormal-inverse Weibull, composite lognormal-Pareto, composite lognormal-paralogistic, composite lognormal-inverse paralogistic, composite lognormal-Burr, composite lognormal-inverse Burr, composite lognormal-inverse Pareto, composite lognormal-inverse exponential, composite lognormal-exponential, composite lognormal-gamma, composite lognormal-inverse gamma, composite lognormal-transformed gamma and composite lognormal-inverse transformed gamma distributions. We fitted all of the distributions by the method of maximum likelihood. The best distribution was chosen according to the following information criteria:

- the Akaike Information Criterion (AIC) due to Akaike [1] defined by

  $$\mathrm{AIC} = 2k - 2\log\widehat{L},$$

  where k denotes the number of parameters and $\widehat{L}$ denotes the maximized likelihood;
- the Bayesian Information Criterion (BIC) due to Schwarz [17] defined by

  $$\mathrm{BIC} = k\log n - 2\log\widehat{L},$$

  where n denotes the number of data;
- the Hannan Quinn Criterion (HQC) due to Hannan and Quinn [10] defined by

  $$\mathrm{HQC} = -2\log\widehat{L} + 2k\log\log n.$$

The smaller the values of these criteria the better the fit. The goodness of fit of the distributions was assessed by the p-values of the Kolmogorov–Smirnov, Anderson Darling and Cramer von Mises statistics.

The fourteen different models considered include the following:

- the composite lognormal-inverse Burr distribution with

$$f_0(x) = \frac{\lambda_1\lambda_2\left(\frac{x}{\lambda_3}\right)^{\lambda_1\lambda_2}}{x\left[1+\left(\frac{x}{\lambda_3}\right)^{\lambda_2}\right]^{\lambda_1+1}}$$

and

$$F_0(x) = \left[1+\left(\frac{x}{\lambda_3}\right)^{-\lambda_1}\right]^{-\lambda_2},$$

where λ_1 and λ_2 are shape parameters while λ_3 is a scale parameter;

- the composite lognormal-generalised Pareto distribution with

$$f_0(x) = \frac{1}{\lambda_3}\left[1+\frac{\lambda_1(x-\lambda_2)}{\lambda_3}\right]^{-\frac{1}{\lambda_1}-1}$$

and

$$F_0(x) = 1-\left[1+\frac{\lambda_1(x-\lambda_2)}{\lambda_3}\right]^{-\frac{1}{\lambda_1}},$$

where λ_1 is a shape parameter, λ_2 is a location parameter and λ_3 is a scale parameter;

- the composite lognormal-inverse paralogistic distribution with

$$f_0(x) = \frac{\lambda_1^2\left(\frac{x}{\lambda_2}\right)^{\lambda_1^2}}{x\left[1+\left(\frac{x}{\lambda_2}\right)^{\lambda_1}\right]^{\lambda_1+1}},$$

where λ_1 is a shape parameter and λ_2 is a scale parameter.

2.2. Estimation

Suppose $x_1, x_2, ..., x_n$ is a random sample from (2.1). Let $\boldsymbol{\Lambda}$ denote the parameters specifying $f_0(\cdot)$ and $F_0(\cdot)$. The maximum likelihood estimates of θ, σ and $\boldsymbol{\Lambda}$, say $\widehat{\theta}$, $\widehat{\sigma}$ and $\widehat{\boldsymbol{\Lambda}}$, respectively, were obtained as follows:

i) Compute the likelihood function

$$L(\theta,\sigma,\boldsymbol{\Lambda}) = \frac{\phi^{n-m}}{(1+\phi)^n[1-F_0(\theta)]^{n-m}}\left[\prod_{x_i\le\theta}\frac{\psi\left(\frac{\log x_i-\mu}{\sigma}\right)}{\Phi\left(\frac{\log\theta-\mu}{\sigma}\right)}\right]\left[\prod_{x_i>\theta}f_0(x_i)\right],$$

where

$$m = \sum_{i=1}^{n} I\{x_i \le \theta\}$$

and $I\{\cdot\}$ denotes the indicator function. μ and ϕ are given by (2.2). Hence, they are functions of θ, σ and $\boldsymbol{\Lambda}$.

ii) Take its log as

$$\begin{aligned}\log L(\theta,\sigma,\boldsymbol{\Lambda}) \;=\; & (n-m)\log\phi - n\log(1+\phi)\\ & - m\log\left[\Phi\left(\frac{\log\theta-\mu}{\sigma}\right)\right] + (m-n)\log[1-F_0(\theta)]\\ & + \sum_{x_i\le\theta}\log\psi\left(\frac{\log x_i-\mu}{\sigma}\right)\\ & + \sum_{x_i>\theta}\log f_0(x_i).\end{aligned}$$

iii) Set initial values for θ, σ and $\boldsymbol{\Lambda}$.

iv) Maximize the log-likelihood function to obtain

$$\widehat{\theta},\widehat{\sigma},\widehat{\boldsymbol{\Lambda}} = \operatorname{argmax}_{\theta,\sigma,\boldsymbol{\Lambda}}\ \log L(\theta,\sigma,\boldsymbol{\Lambda}), \tag{2.3}$$

using the optim function in R.

v) Repeat steps iii) and iv) for a range of initial values to make sure that $\widehat{\theta}$, $\widehat{\sigma}$ and $\widehat{\boldsymbol{\Lambda}}$ are unique.

In Section 4, we compare the best of the composite lognormal distributions to the power law distribution given by the cdf:

$$F(x) = 1 - \left(\frac{K}{x}\right)^{\alpha} \tag{2.4}$$

for $x > K$ and $\alpha > 0$. For a given random sample $x_1, x_2, ..., x_n$ from (2.4), the maximum likelihood estimates of K and α are

$$\widehat{K} = \min(x_1, x_2, ..., x_n)$$

and

$$\widehat{\alpha} = n\left[\sum_{i=1}^{n}\log\left(\frac{x_i}{\widehat{K}}\right)\right]^{-1},$$

respectively.

Campolieti [8] and Balthrop and Quan [3] fitted the power law distribution to the upper tail of the data. They used the following procedure to estimate the parameters:

1. Order the data as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$.

2. Let $\widetilde{K} = x_{(i)}$ and estimate α by

$$\widetilde{\alpha} = \left[\sum_{i=1}^{n} I\left\{x_i \ge \widetilde{K}\right\}\right]\left[\sum_{i=1}^{n}\log\left(\frac{x_i}{\widetilde{K}}\right)\right]^{-1}.$$

3. Compute the Kolmogorov–Smirnov statistic

$$\sup_{x\ge\widetilde{K}}\left|\widetilde{F}(x) - 1 + \left(\frac{\widetilde{K}}{x}\right)^{\widetilde{\alpha}}\right|,$$

where $\widetilde{F}(\cdot)$ denotes the empirical cdf of the data.

4. Repeat steps 2 and 3 for $i = 1, 2, ..., n-1$.

5. Choose $\widetilde{K}$ and $\widetilde{\alpha}$ to correspond to the smallest value of the Kolmogorov–Smirnov statistic.

3. DATA SETS

The two data sets are described in Sections 3.1 and 3.2. We shall refer to the data as described in these sections as "full data" sets.

3.1. Canadian net wealth data

The data were collected from the rich 100 list "Canadian Business magazine". The rich list is published every year on line by the magazine. However, due to the webpages being updated, we were able to get the data only for the years 2014–2018, 2012 and 2009. Due to changes in policy from one year to another, 2014 had 101 data points while for 2018 had 98 data points. The remaining years had 100 data points each.

Table 1: Summary statistics of net wealth data in billions of nominal Canadian Dollars (the deflated figures by the Consumer Price Index are given in the second row for each year).

Year	Mean	Median	Standard deviation	Skewness	Kurtosis	Min	Max	CPI
2018	3.44 0.0258	2.11 0.0158	4.61 0.0346	6.02 5.9245	47.39 43.4279	1.07 0.008	41.14 0.308	133.4
2017	3.07 0.0235	2.03 0.0155	4.25 0.03259	6.47 6.3692	53.48 49.41976	0.875 0.0067	39.13 0.3	130.4
2016	2.88 0.0224	1.89 0.0147	4.00 0.0311	6.47 6.375	53.31 49.2505	0.835 0.0065	36.76 0.2863	128.4
2015	2.56 0.0202	1.80 0.0142	3.34 0.0264	6.40 6.3065	52.47 48.4212	0.782 0.0062	30.738 0.2428	126.6
2014	2.29 0.0183	1.46 0.0116	2.91 0.0232	5.91 5.8207	46.35 42.438	0.721 0.0058	26.075 0.2083	125.2
2012	2.02 0.0166	1.39 0.0114	2.33 0.0191	5.29 5.2145	38.88 35.1037	0.654 0.0054	20.129 0.1654	121.7
2009	1.73 0.0151	1.15 0.01005	2.39 0.0208	6.48 6.3829	53.72 49.6533	0.49 0.0043	21.99 0.1922	114.4

Table 1 gives the summary statistics of the data in terms of nominal and real figures. The Consumer Price Index was taken from `https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1810000501`. Since the rich people are getting richer with time in terms of both nominal and real figures, the values of mean, median, maximum and minimum increase with year. The skewness is positive every year. The kurtosis is much greater than 3 every year, meaning that the data are heavy tailed.

3.2. Cumulative coal production data

The data were collected from the U.S. Energy Information Administration (EIA) website. Balthrop and Quan [3] used the cumulative yearly production data from 1983 to 2016. Due to the website being updated, we use the data from 2001 to 2018. The data contained a large number of zeros (over 800 data points were zero), and these were removed before fitting of the distributions.

Table 2: Summary statistics of coal data (unit: short tons).

Mean	4413710
SD	38518025
Skewness	27.62912
Kurtosis	954.948
Min	43
Max	1528026392
Sample size	4180

Table 2 shows that the skewness is positive. The kurtosis is once again much larger than 3, which indicates the data has a large heavy tail.

4. RESULTS AND DISCUSSION

In this section, we illustrate the flexibility of the composite lognormal distributions using the two real data sets. Fourteen of the composite lognormal distributions were fitted to both data sets in full. For comparison, the power law distribution is also fitted to the full data sets. The power law distribution is also fitted to the upper tail of the data sets.

In the discussion throughout Sections 4.1 and 4.2, "p-values" refer to p-values of the Kolmogorov–Smirnov statistic. But Tables 10 and 12 also report p-values of Anderson Darling and Cramer von Mises statistics. The conclusions based on these p-values are the same as those based on p-values of the Kolmogorov–Smirnov statistic.

In Section 4.3, we investigate finite sample performance of the maximum likelihood estimators of composite lognormal distributions to see if the conclusions reported in Sections 4.1 and 4.2 are reasonable.

4.1. Canadian net wealth data

Tables 3 to 9 give the best three distributions giving the smallest information criteria for each year. The power law distribution does not make the best three distributions for any of the years.

Table 3: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2009.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−261.42	−251.42	−240.99	−247.20
Composite lognormal-inverse paralogistic	−251.60	−243.60	−235.79	−240.44
Composite lognormal-generalized Pareto	−209.98	−199.98	−189.56	−195.76

Table 4: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2012.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−231.72	−221.72	−211.30	−217.50
Composite lognormal-inverse paralogistic	−222.22	−214.22	−206.41	−211.06
Composite lognormal-generalized Pareto	−221.91	−211.91	−201.49	−207.69

Table 5: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2014.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−244.64	−234.64	−224.18	−230.40
Composite lognormal-inverse paralogistic	−237.97	−229.97	−222.13	−226.80
Composite lognormal-generalized Pareto	−228.74	−218.74	−208.28	−214.51

Table 6: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2015.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−211.45	−201.45	−191.03	−197.23
Composite lognormal-inverse paralogistic	−208.74	−200.74	−192.92	−197.58
Composite lognormal-generalized Pareto	−236.70	−226.70	−216.28	−222.48

Table 7: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2016.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−239.28	−229.28	−218.86	−225.06
Composite lognormal-inverse paralogistic	−229.61	−221.61	−213.80	−218.45
Composite lognormal-generalized Pareto	−234.94	−224.94	−214.52	−220.72

Table 8: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2017.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−240.90	−230.90	−220.48	−226.69
Composite lognormal-inverse paralogistic	−240.07	−232.07	−224.25	−228.91
Composite lognormal-generalized Pareto	−215.20	−205.20	−194.78	−200.98

Table 9: The three best composite lognormal distributions according to information criteria for Canadian net wealth data in 2018.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-inverse Burr	−251.60	−241.60	−231.26	−237.42
Composite lognormal-inverse paralogistic	−244.83	−236.83	−229.08	−233.69
Composite lognormal-generalized Pareto	−233.40	−223.40	−213.06	−219.22

Table 10 lists the p-values for the power law distribution and the very best composite lognormal distributions chosen as the ones having the smallest information criteria values.

Table 10: Fitted models and p-values for the Canadian net wealth data (the first row of p-values for each year is for the Kolmogorov–Smirnov statistic, the second row of p-values for each year is for the Anderson Darling statistic, the third row of p-values for each year is for the Cramer von Mises statistic).

Year	n	Composite model		Power law fitted to full data	Power law fitted to upper tail		
		Best model	p-value	p-value	p-value	$\widetilde{K}$	no of data $> \widetilde{K}$
2018	98	Composite lognormal-inverse Burr	0.973 0.970 0.974	0.060 0.055 0.058	0.99 0.98 0.99	2.77	37
2017	100	Composite lognormal-inverse paralogistic	0.853 0.855 0.855	0.011 0.010 0.008	0.994 0.994 0.995	2.96	27
2016	100	Composite lognormal-inverse Burr	0.9996 0.999 0.998	0.098 0.095 0.097	0.99 0.95 0.96	2.35	38
2015	100	Composite lognormal-generalised Pareto	0.844 0.840 0.851	0.020 0.030 0.035	0.98 0.99 0.96	1.96	48
2014	101	Composite lognormal-inverse Burr	0.921 0.935 0.922	0.063 0.065 0.068	0.92 0.93 0.90	1.85	42
2012	100	Composite lognormal-inverse Burr	0.996 0.999 0.995	0.065 0.061 0.062	0.9 0.9 0.91	1.48	46
2009	100	Composite lognormal-inverse Burr	0.998 0.995 0.999	0.020 0.025 0.022	0.959 0.966 0.954	1.17	48

The p-values for the very best composite lognormal distributions range from 0.8 to 0.99. The largest of these p-values is 0.9996 (2016) and the smallest is 0.844 (2015). The composite lognormal inverse Burr distribution gives the largest p-values for five of the seven years.

The p-values for the power law distribution are always less than 0.1 when applied to the full data. When applied to the tail (containing a fraction of the full data), the p-values are much closer to 1. But for four of the seven years the p-values for the very best composite lognormal distributions are still greater. For the year 2018, the p-value for the very best composite lognormal distribution is slightly smaller (0.973 compared to 0.99), but the power law tail models only 37 of the 98 observations. For the year 2017, the p-value for the very best composite lognormal distribution is again slightly smaller (0.853 compared to 0.994), but the power law tail models only 27 of the 100 observations. For the year 2015, the p-value for the very best composite lognormal distribution is again slightly smaller (0.844 compared to 0.98), but the power law tail models only 48 of the 100 observations.

The probability and quantile plots comparing the fits of the power law distribution and the very best composite lognormal distributions are shown in Figures 1 to 7.

Both the quantile and probability plots confirm that the composite lognormal distributions provide better fits than the power law distribution. Nearly all of the plotted points in the probability plots lie close to the 45 degree line for the composite lognormal distributions. The quantile plots show that the composite lognormal distributions provide good fits to the data except for a few extremely large observations. The power law distribution fitted to the full data gives poor fits. The power law distribution fitted to the tail gives much better fits but still not good as the composite lognormal distributions.

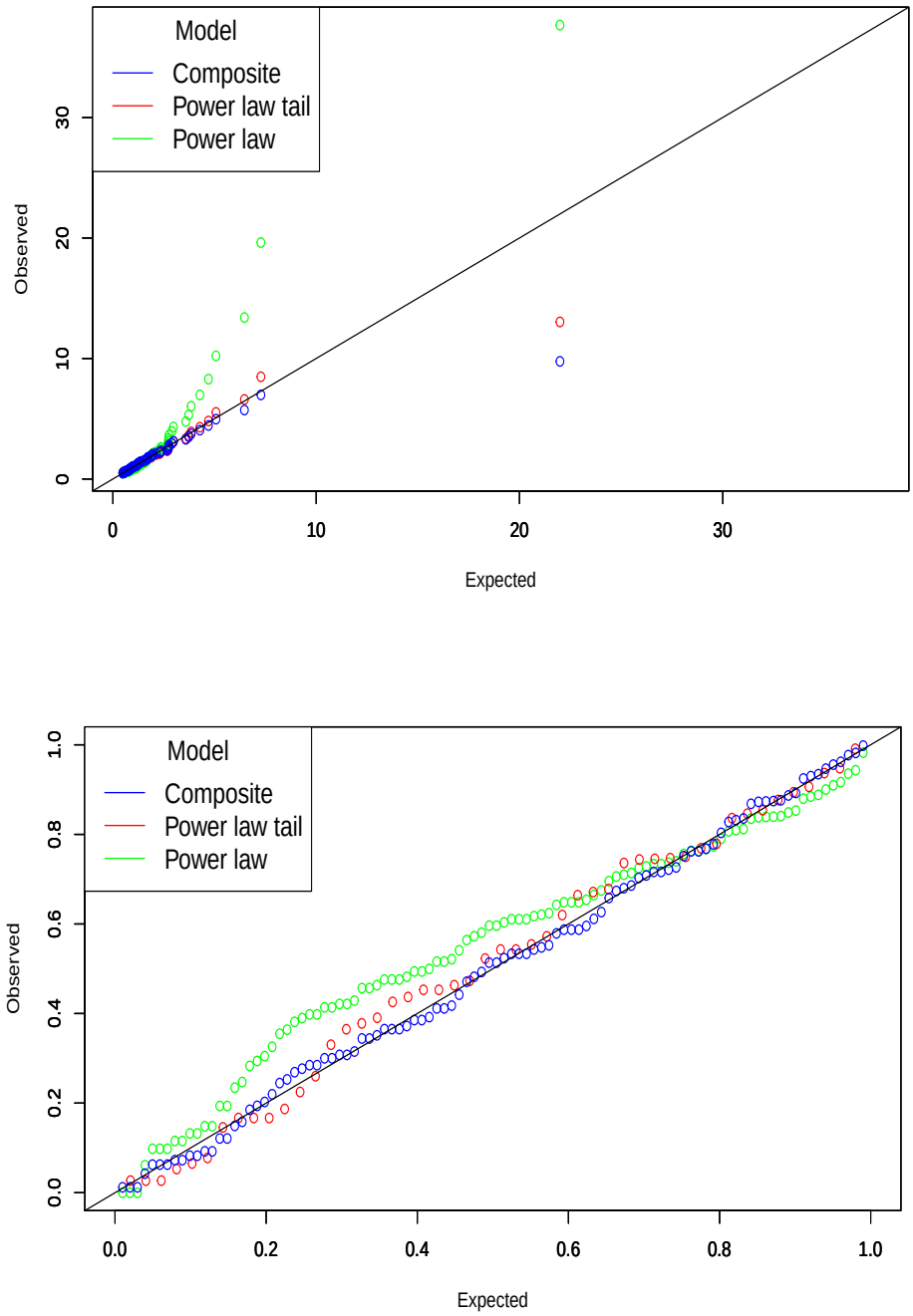


Figure 1: Quantile (left) and probability (right) plots for the fits of the composite lognormal-inverse Burr and power law distributions for Canadian net wealth data in 2009.

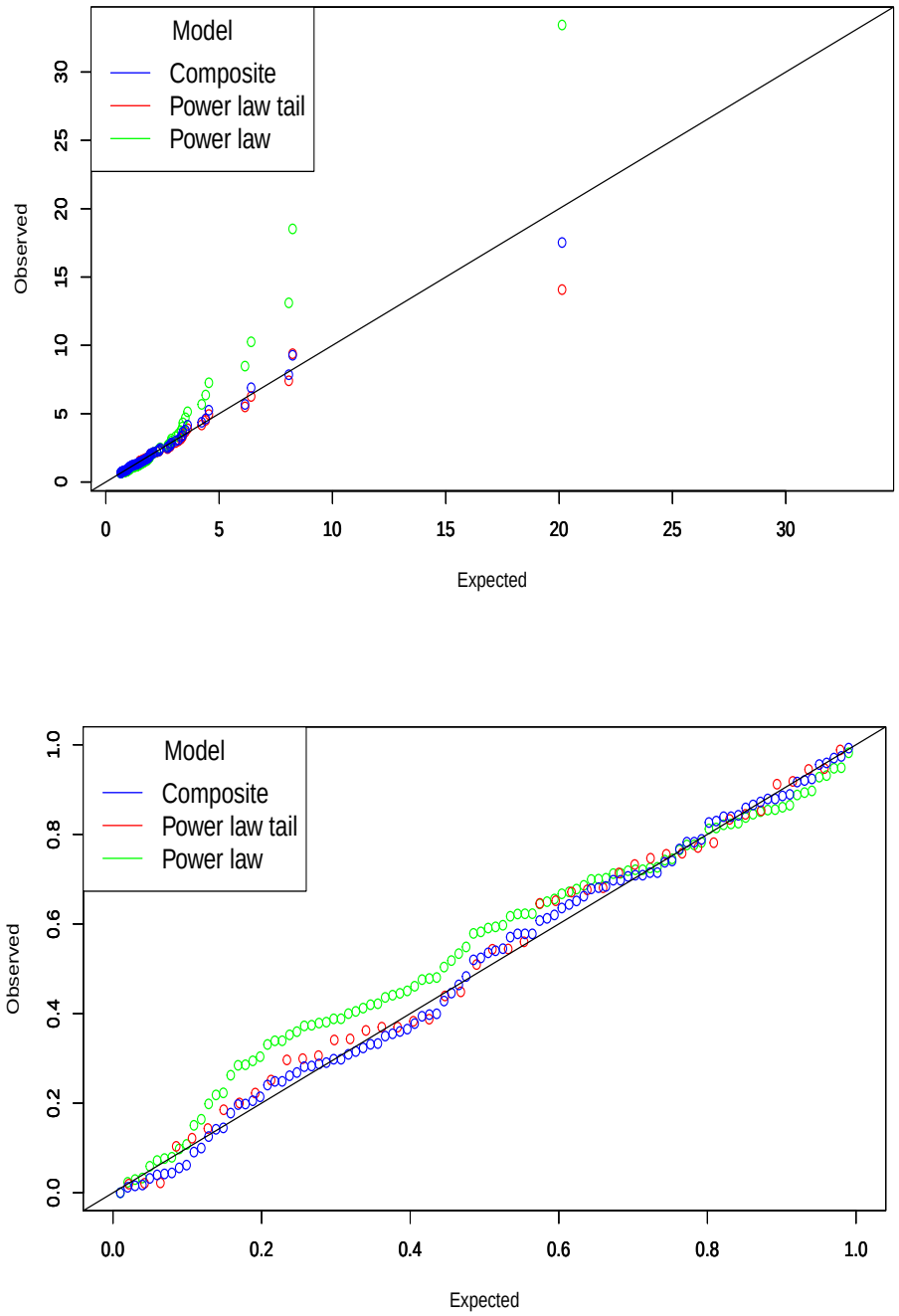


Figure 2: Quantile (left) and probability (right) plots for the fits of the composite lognormal-inverse Burr and power law distributions for Canadian net wealth data in 2012.

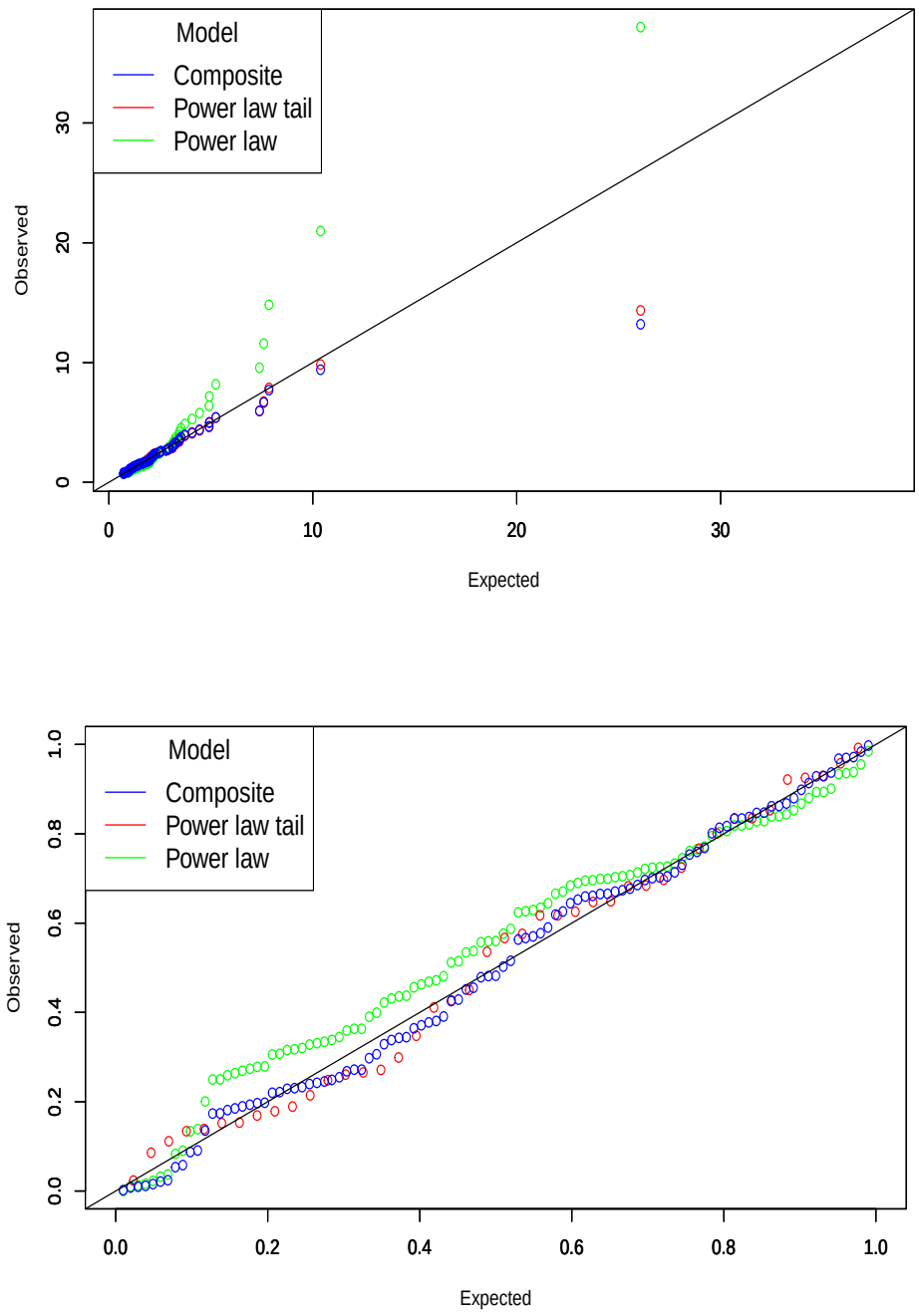


Figure 3: Quantile (left) and probability (right) plots for the fits of the composite lognormal-inverse Burr and power law distributions for Canadian net wealth data in 2014.

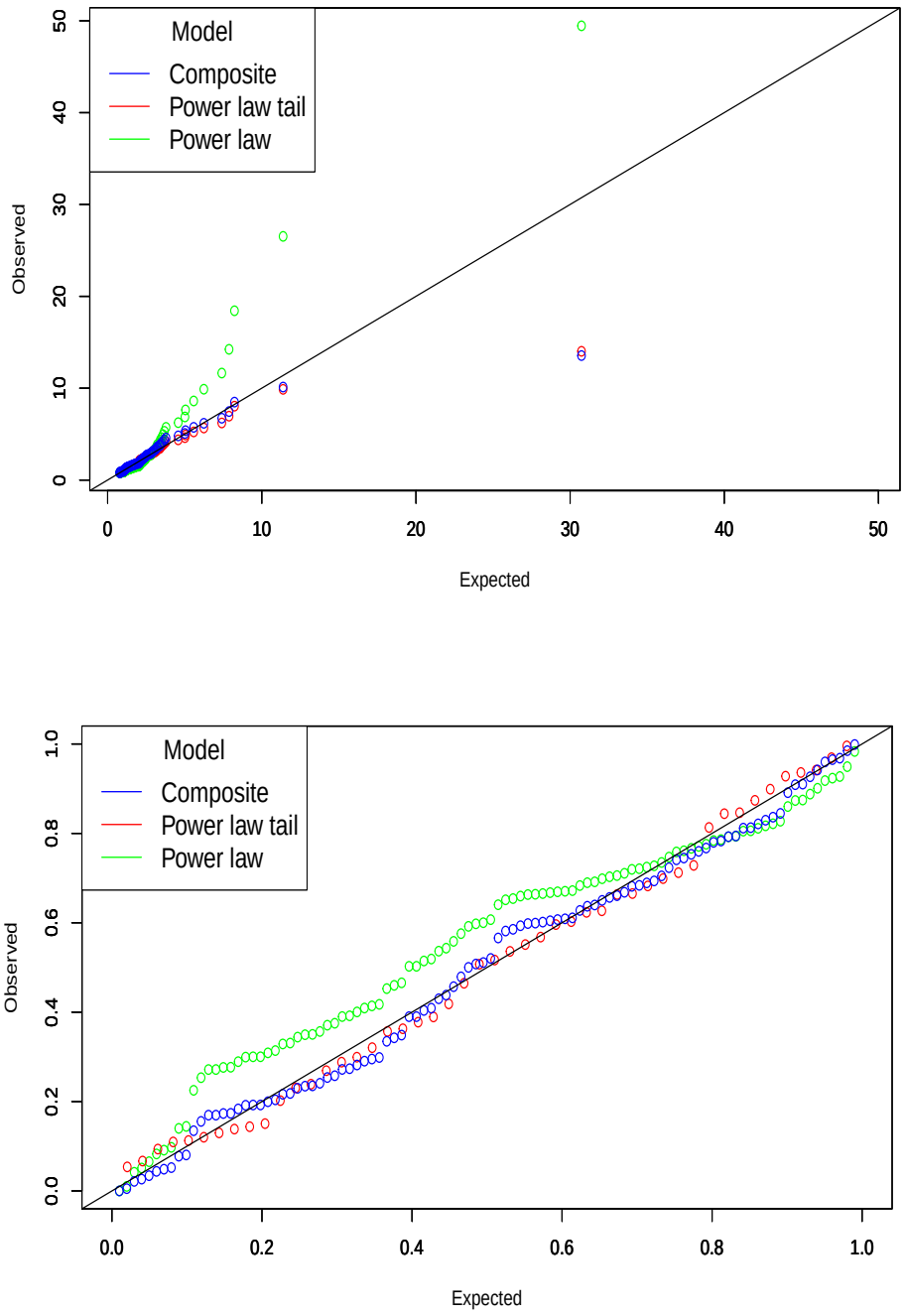


Figure 4: Quantile (left) and probability (right) plots for the fits of the composite lognormal-generalised Pareto and power law distributions for Canadian net wealth data in 2015.

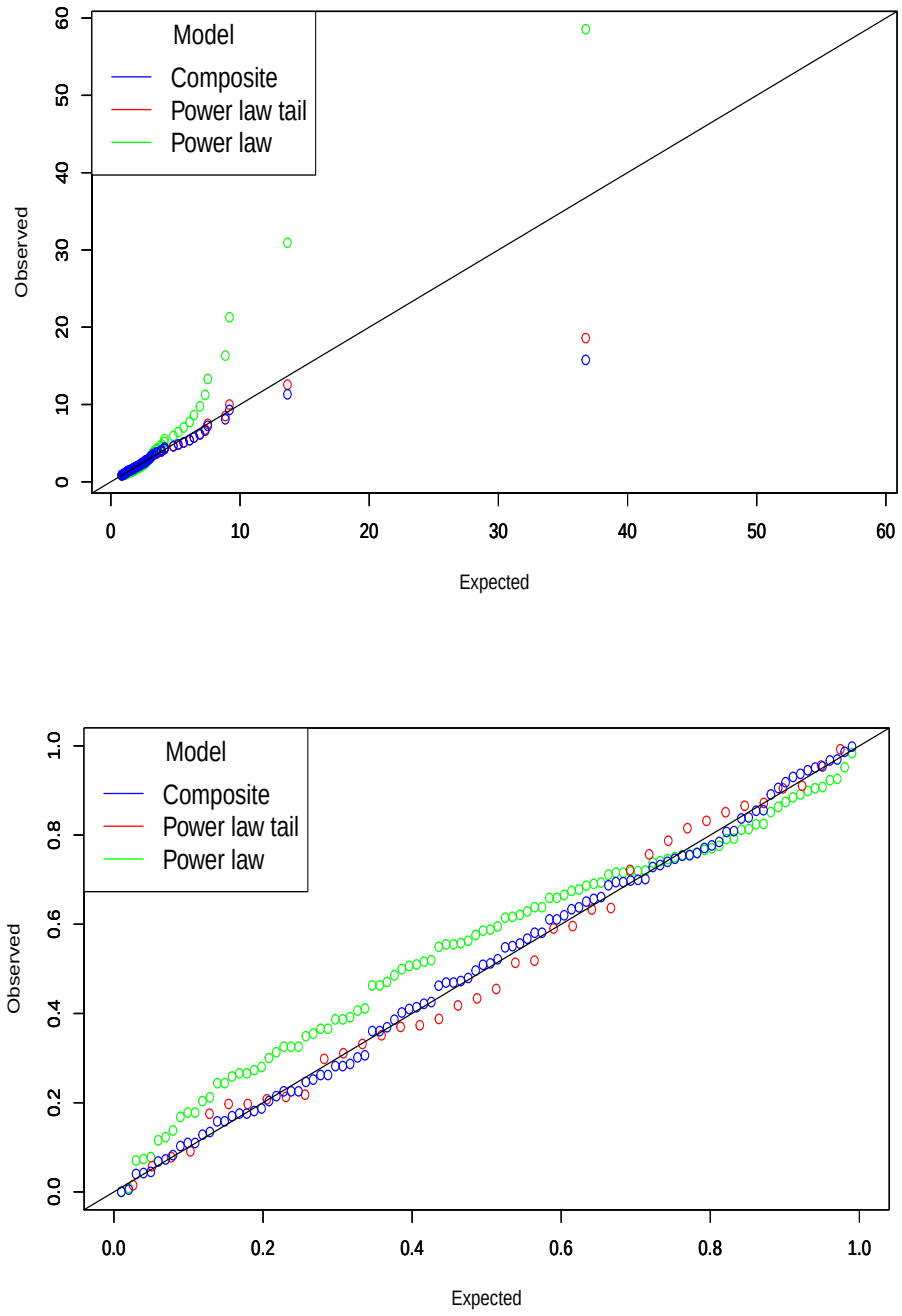


Figure 5: Quantile (left) and probability (right) plots for the fits of the composite lognormal-inverse Burr and power law distributions for Canadian net wealth data in 2016.

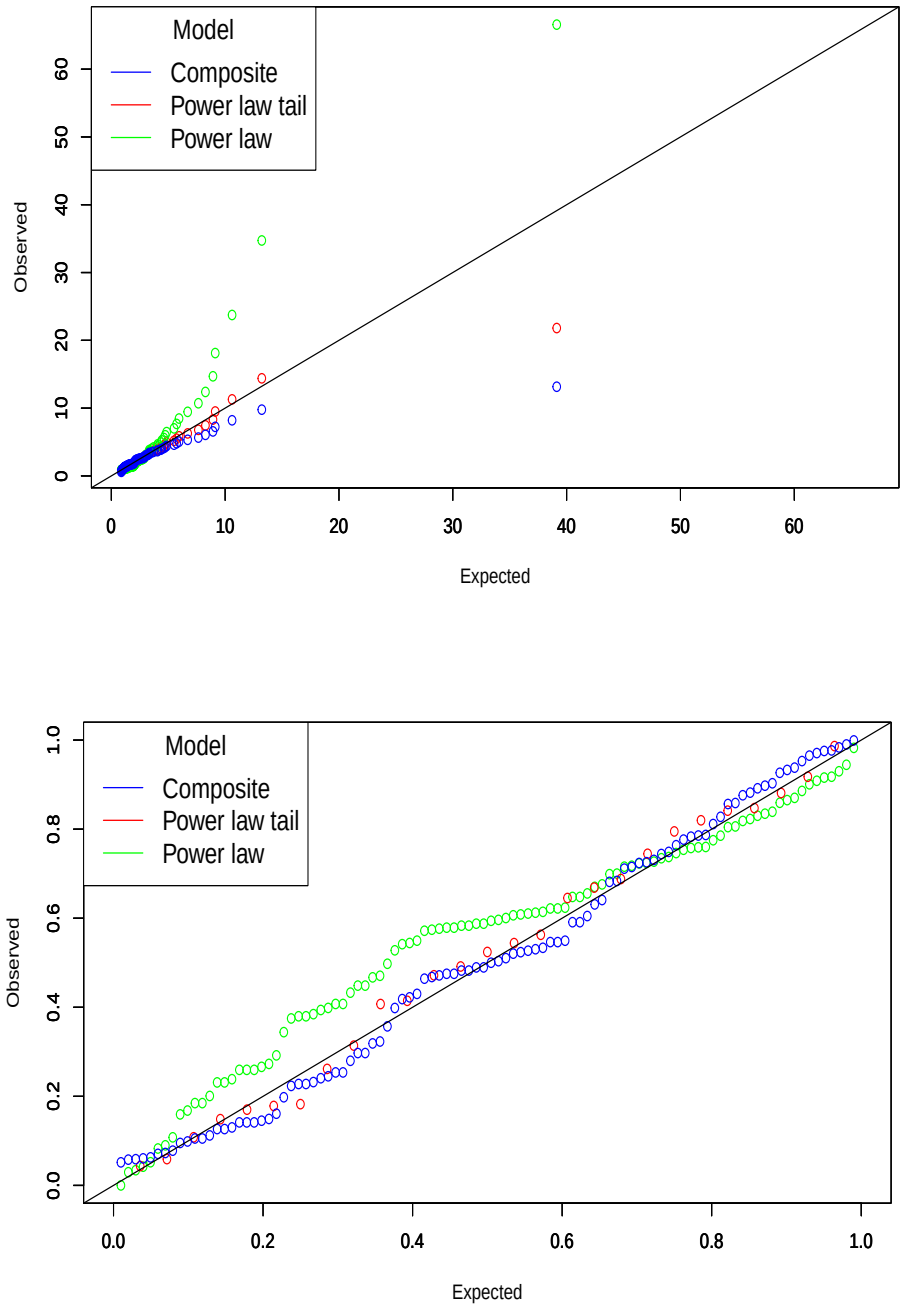


Figure 6: Quantile (left) and probability (right) plots for the fits of the composite lognormal-inverse paralogistic and power law distributions for Canadian net wealth data in 2017.

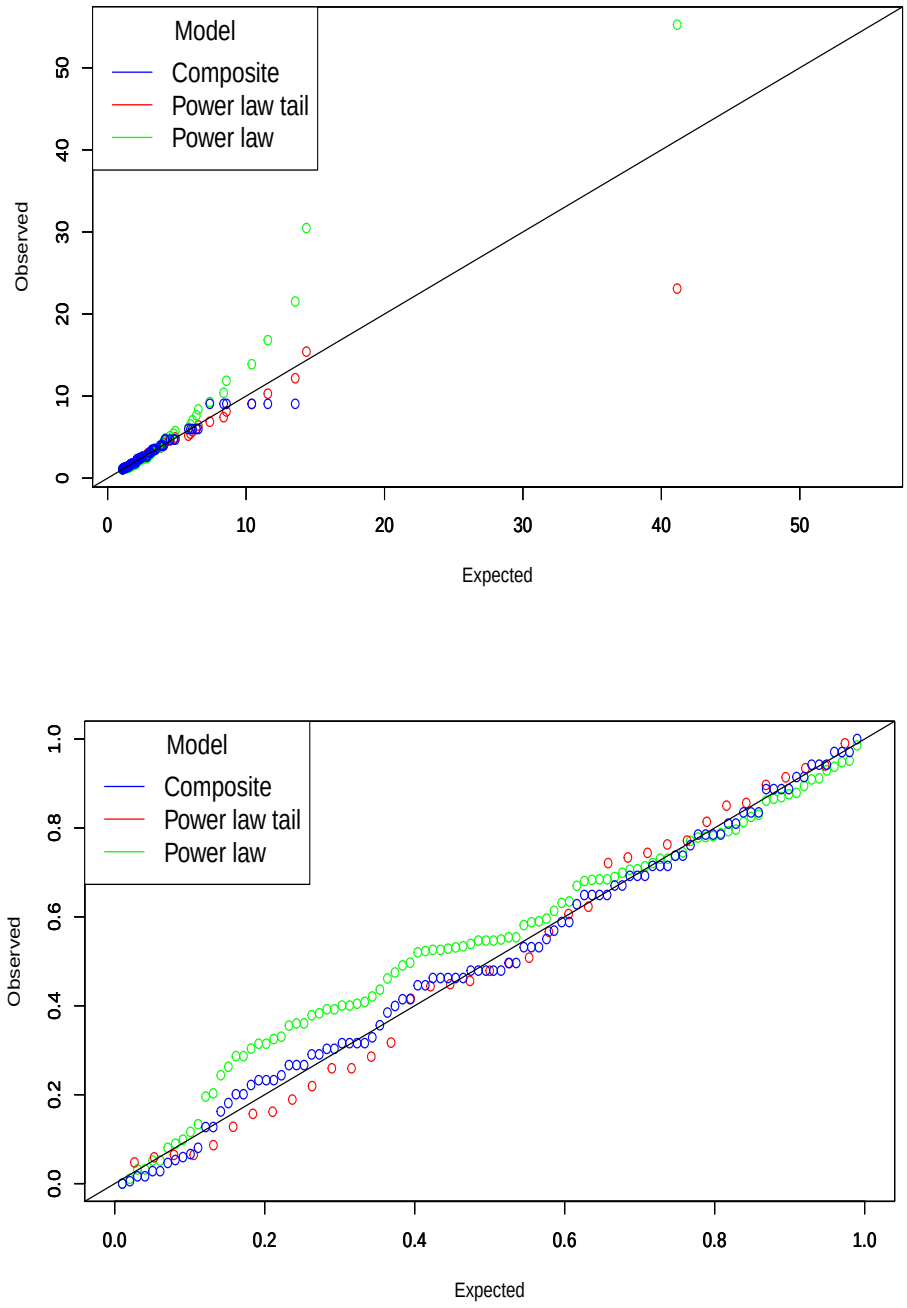


Figure 7: Quantile (left) and probability (right) plots for the fits of the composite lognormal-inverse Burr and power law distributions for Canadian net wealth data in 2018.

4.2. Cumulative coal production data

Table 11 gives the best five distributions giving the smallest information criteria. The power law distribution once again does not make the best five distributions. The p-values of the best five distributions given in Table 12 range from 0.94 to 0.97. The largest p-value of 0.9723 is given by the composite lognormal-generalised Pareto distribution. The smallest p-value of 0.9407 is given by the composite lognormal-Pareto distribution. The composite lognormal-generalised Pareto distribution also gives the smallest information criteria values.

The fit of the power law distribution to the full data ($n = 4180$) gave a p-value $< 2.210^{-16}$. The fit of the power law distribution to the tail containing 1095 observations of the full data gave $\widetilde{K} = 1027417$ and p-value $= 0.108$. All of the p-values in Table 12 are much greater than 0.108.

Table 11: The five best composite lognormal distributions according to information criteria for cumulative coal production data.

Model	$-2\log L$	AIC	BIC	HQC
Composite lognormal-generalised Pareto	−9464.96	−9454.96	−9429.60	−9445.99
Composite lognormal-log logistic	−9134.55	−9126.55	−9107.54	−9119.83
Composite lognormal-inverse paralogistic	−9270.34	−9262.34	−9243.33	−9255.61
Composite lognormal-paralogistic	−9219.03	−9211.03	−9192.01	−9204.30
Composite lognormal-Pareto	−9410.33	−9402.33	−9383.32	−9395.60

Table 12: p-values for the five best composite lognormal distributions for cumulative coal production data.

Model	p-values		
	Kolmogorov–Smirnov	Anderson Darling	Cramer von Mises
Composite lognormal-generalised Pareto	0.972	0.971	0.974
Composite lognormal-log logistic	0.965	0.964	0.966
Composite lognormal-inverse paralogistic	0.963	0.960	0.959
Composite lognormal-paralogistic	0.944	0.950	0.948
Composite lognormal-Pareto	0.941	0.940	0.938

The probability and quantile plots comparing the fits of the power law and composite lognormal-generalised Pareto distributions are shown in Figure 8.

The probability plot shows that the composite lognormal-generalised Pareto distribution provides a near perfect fit. The quantile plot shows that the composite lognormal-generalised Pareto distribution provides a good fit except for some extremely large observations. Neither of the two power law models provide as good a fit as the composite lognormal-generalised Pareto distribution.

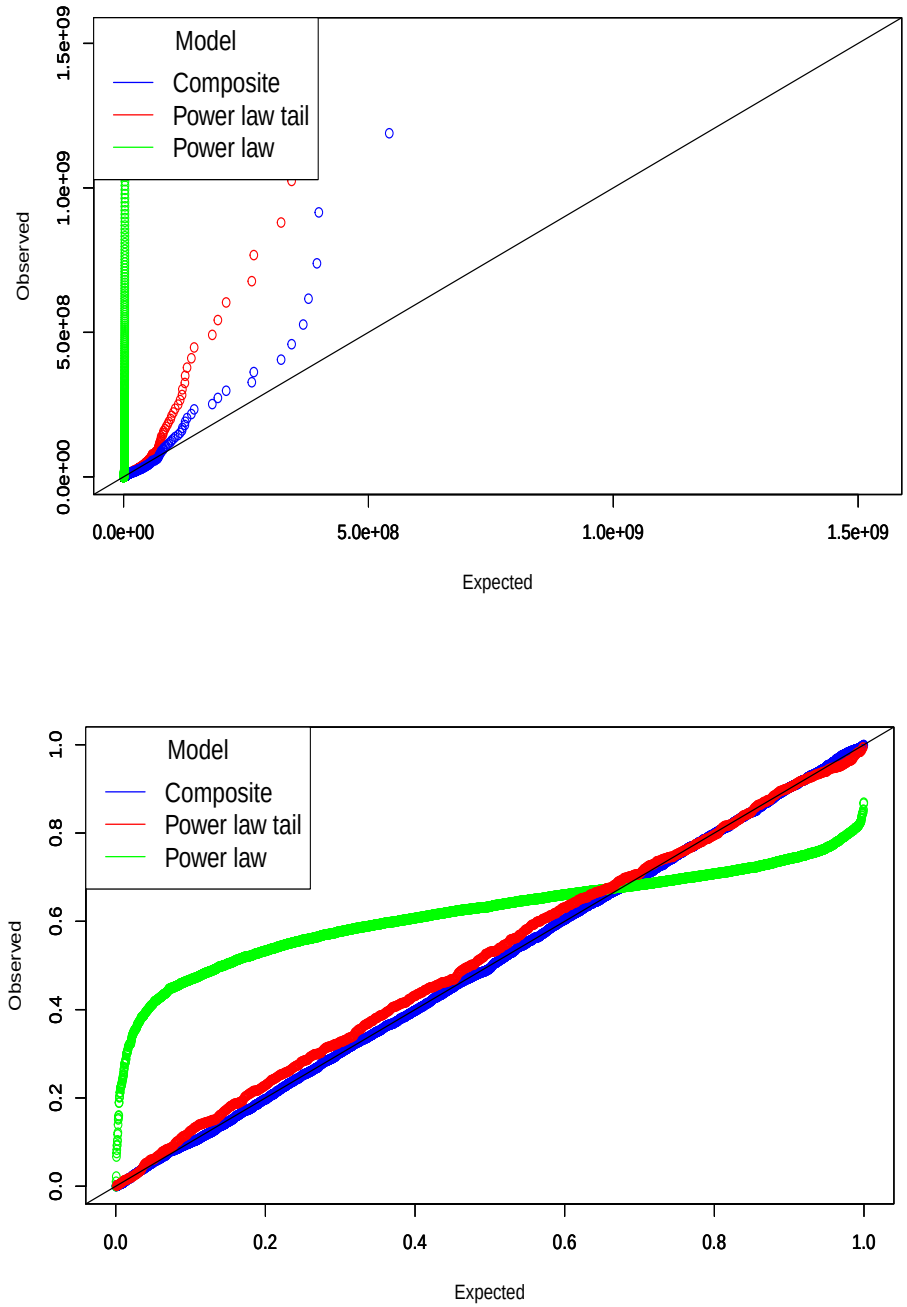


Figure 8: Quantile (left) and probability (right) plots for the fits of the composite lognormal-generalised Pareto and power law distributions for cumulative coal production data.

4.3. A simulation study

In this section, we assess the performance of the maximum likelihood estimates given by (2.3) with respect to sample size n. The assessment of the performance of the maximum likelihood estimates of $(\theta, \sigma, \mathbf{\Lambda})$ is based on a simulation study:

1. Generate ten thousand samples of size n from the composite lognormal distribution by inverting

$$F(x) = u_k$$

for $k = 1, 2, ..., n$, where $u_1, u_2, ..., u_n$ is a random sample from uniform$(0, 1)$ and F is given by (2.1);

2. Compute the maximum likelihood estimates for the ten thousand samples in step 1, say $\left(\widehat{\theta}_i, \widehat{\sigma}_i, \widehat{\mathbf{\Lambda}}_i\right)$ for $i = 1, 2, ..., 10000$.

3. Compute the biases and mean squared errors given by

$$\widehat{\text{bias}}_e(n) = \frac{1}{10000} \sum_{i=1}^{10000} (\widehat{e}_i - e)$$

and

$$\widehat{\text{MSE}}_e(n) = \frac{1}{10000} \sum_{i=1}^{10000} (\widehat{e}_i - e)^2$$

for $e = \theta, \sigma, \mathbf{\Lambda}$.

We repeated these steps for $n = 10, 11, ..., 500$ with $\theta = 1$, $\sigma = 1$ and $\mathbf{\Lambda}$ corresponding to the composite lognormal-inverse paralogistic distribution; so, computing $\widehat{\text{bias}}_\theta(n)$, $\widehat{\text{bias}}_\sigma(n)$, $\widehat{\text{bias}}_{\lambda_1}(n)$, $\widehat{\text{bias}}_{\lambda_2}(n)$, $\widehat{\text{MSE}}_\theta(n)$, $\widehat{\text{MSE}}_\sigma(n)$, $\widehat{\text{MSE}}_{\lambda_1}(n)$ and $\widehat{\text{MSE}}_{\lambda_2}(n)$ for $n = 10, 11, ..., 500$.

Figures 9 and 10 show how the biases and the mean squared errors vary with respect to n. The red line corresponds to the biases being zero. The following observations can be made:

1. The magnitude of the biases of the estimators generally decrease to zero;
2. The mean squared errors of the estimators generally decrease to zero;
3. The biases are generally negative for λ_1 and λ_2;
4. The biases appear largest in magnitude for λ_1 and λ_2;
5. The mean squared errors appear largest for θ and σ.

The results of the simulation study show that: the accuracy of the estimators of θ, σ, λ_1 and λ_2 as measured by bias is reasonable for all $n \geq 300$; the accuracy of the estimators of θ, σ, λ_1 and λ_2 as measured by mean squared error is reasonable for all $n \geq 300$. The sample size used in Section 4.2 is much greater than 300 but the sample sizes in Section 4.1 are not greater than 300. Hence, the conclusions in Section 4.2 should be reasonable but the conclusions in Section 4.1 should be treated conservatively.

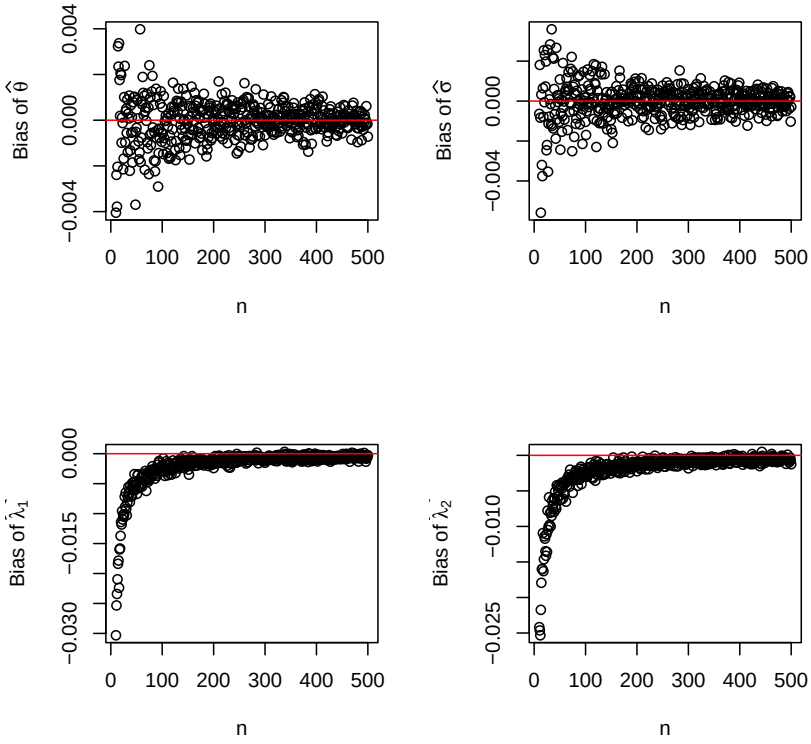


Figure 9: $\widehat{\text{bias}}_\theta(n)$, $\widehat{\text{bias}}_\sigma(n)$, $\widehat{\text{bias}}_{\lambda_1}(n)$ and $\widehat{\text{bias}}_{\lambda_2}(n)$ versus n.

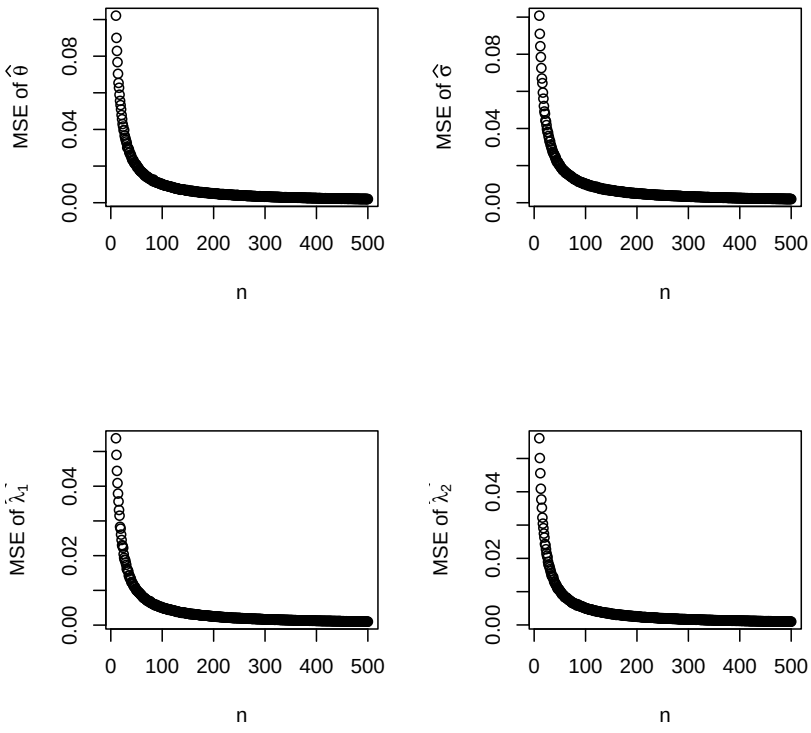


Figure 10: $\widehat{\text{MSE}}_\theta(n)$, $\widehat{\text{MSE}}_\sigma(n)$, $\widehat{\text{MSE}}_{\lambda_1}(n)$ and $\widehat{\text{MSE}}_{\lambda_2}(n)$ versus n.

We have presented results only for $\theta = 1$, $\sigma = 1$ and a particular composite lognormal distribution. But the results were similar for other choices for θ and σ and other composite lognormal distributions.

5. CONCLUSIONS

In this paper, we have illustrated the power of composite lognormal distributions for two real data sets recently published in the physics literature. These data sets (in full or in part) have been previously modeled by the power law distribution. All of the composite lognormal distributions provide much better fits than the power law distribution when both were fitted to the full data sets. For the first data set, several of the composite lognormal distributions (composite lognormal-inverse Burr, composite lognormal-inverse paralogistic and composite lognormal-generalised Pareto distributions) provide better fits than the power law distribution even when the former were fitted to the full data and the latter was fitted only to the upper tail. For the second data set, all of the composite lognormal distributions provide much better fits than the power law distribution even when the former were fitted to the full data and the latter was fitted only to the upper tail. The goodness of fit was assessed by probability plots, quantile plots and p-values of the Kolmogorov–Smirnov, Anderson Darling and Cramer von Mises statistics. Software for fitting composite lognormal distributions is freely available from Nadarajah and Bakar [14].

Finally, we like to point out that the use of the power law distribution to model the two real data sets was motivated by a theoretical framework. Lee *et al.* [12] describe the rationale for the composite lognormal distributions in (2.1) as "the lognormal distribution models a large portion of the data well, but quickly fades away to zero. Thus it fits poorly a portion of the tail. On the other hand, F_0 fits the tail portion well, but fits the other portion poorly. By combining two distributions with one fitting the portion below a given threshold and the other fitting the portion larger than the threshold, the composite distribution (2.1) was proposed". But to the best of our knowledge there is no theoretical motivation yet for the composite lognormal distributions. Finding a theoretical motivation for the composite lognormal distributions is a possible future work.

ACKNOWLEDGMENTS

The authors would like to thank the Editor and the two referees for careful reading and comments which greatly improved the paper.

REFERENCES

[1] Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, **19**, 716–723.

[2] Aminzadeh, M.S. and Deng, M. (2019). Bayesian predictive modeling for inverse gamma-Pareto composite distribution, *Communications in Statistics – Theory and Methods*, **48**, 1938–1954.

[3] Balthrop, A. and Quan, S. (2019). The power-law distribution of cumulative coal production, *Physica A: Statistical Mechanics and Its Applications*, **530**, 121573.

[4] Calderín-Ojeda, E. (2015). On the composite Weibull – Burr model to describe claim data, *Communications in Statistics: Case Studies, Data Analysis and Applications*, **1**, 59–69.

[5] Calderín-Ojeda, E. (2016). The distribution of all French communes: A composite parametric approach, *Physica A – Statistical Mechanics and Its Applications*, **450**, 385–394.

[6] Calderín-Ojeda, E. (2018). A note on parameter estimation in the composite Weibull–Pareto distribution, *Risks*, **6**, doi: 10.3390/risks6010011

[7] Calderín-Ojeda, E. and Kwok, C.F. (2016). Modeling claims data with composite Stoppa models, *Scandinavian Actuarial Journal*, **9**, 817–836.

[8] Campolieti, M. (2018). Heavy-tailed distributions and the distribution of wealth: Evidence from rich lists in Canada, 1999–2017, *Physica A: Statistical Mechanics and Its Applications*, **503**, 263–272.

[9] Cooray, K. and Ananda, M.M.A. (2005). Modeling actuarial data with a composite lognormal-Pareto model, *Scandinavian Actuarial Journal*, 321–334.

[10] Hannan, E.J. and Quinn, B.G. (1979). The determination of the order of an autoregression, *Journal of the Royal Statistical Society, B*, **41**, 190–195.

[11] Kim, Y.; Kim, H.; Lee, G. and Min, K.-H. (2019). A modified hybrid gamma and generalized Pareto distribution for precipitation data, *Asia-Pacific Journal of Atmospheric Sciences*, **55**, 609–616.

[12] Lee, C.; Famoye, F. and Alzaatreh, A.Y. (2013). Methods for generating families of univariate continuous distributions in the recent decades, *Wiley Interdisciplinary Reviews: Computational Statistics*, **5**, 219–238.

[13] Mutali, S. and Vernic, R. (2020). On the composite lognormal – Pareto distribution with uncertain threshold, *Communications in Statistics – Simulation and Computation*, doi: 10.1080/03610918.2020.1743860

[14] Nadarajah, S. and Bakar, S.A.A. (2013). CompLognormal: An R package for composite lognormal distributions, *R Journal*, **5**, 97–103.

[15] Nadarajah, S. and Bakar, S.A.A. (2014). New composite models for the Danish fire insurance data, *Scandinavian Actuarial Journal*, 180–187.

[16] R Core Team (2022). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[17] Schwarz, G.E. (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, 461–464.

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 479–501
https://doi.org/10.57805/revstat.v22i4.448

Estimations of Confidence Sets for the Unit Generalized Rayleigh Parameters Using Records Data

Authors: Xuanjia Zuo, Liang Wang ✉
– School of Mathematics, Yunnan Normal University, Kunming, PR China
zo_syunka@163.com, liang610112@163.com

Yuhlong Lio
– Department of Mathematical Sciences, University of South Dakota, Vermillion, USA
Yuhlong.Lio@usd.edu

Yogesh Mani Tripathi
– Department of Mathematics, Indian Institute of Technology Patna, Patna, India
yogesh@iitp.ac.in

Received: April 2022 Revised: December 2022 Accepted: December 2022

Abstract:

- Estimations of confidence sets are explored for the unit generalized Rayleigh distribution parameters with records data. Using the proposed pivotal quantities, equal-tailed confidence regions of the parameters are constructed. The associated optimal confidence sets with minimum-size are developed by non-linear optimization technique, and various numerical approaches are established for complex computations. For comparison and complementary, the likelihood asymptotic confidence sets of parameters are derived. Extensive simulation studies are conducted to evaluate the performance of all methods, and two practical examples are given for illustrations. Some alternative extensions are provided for pursuing potential higher accuracy confidence sets.

Keywords:

- *unit generalized Rayleigh distribution; records data; confidence set estimation; pivotal quantity; nonlinear optimization.*

AMS Subject Classification:

- 62F25, 62N01.

✉ Corresponding author

1. INTRODUCTION

Due to practical limitations such as time and(or) budget constraint, it is not easy to obtain complete sample in practice; expecially, when the test units feature character of high reliable and expensive. Therefore, censored data frequently appear during the data collection, where only a portion of exact failure times are observed under such limitated situations, and various censoring schemes are implemented in experimental procedures simulataneously. Common censoring schemes used in experiments include Type-I censoring, Type-II censoring, progressive censoring, as well as hybrid censoring. Interested readers may refer to, for example, the monographes of Balakrishnan and Cramer [6] and Lawless [21] for a comprehensive review. However, besides conventional censored data appeared from aforementioned data collection schemes, there are many other incomplete data types occurred in field and experiment situations such as reliability engineering, survival analysis, hydrology, economics, mining and meteorology among others, and records data is one of popular observation among them. For example, Guo et al.*et al.* [14] gave an example regarding a kind of the rock crushing machine, where the size of the rock being crushed is also obtained when the crush strength is larger than the previously one appearing as record data. Soliman *et al.* [24] investigated a reliability experiment, where the exact measurements of failure under operating stress are observed sequentially and the record-breaking values are only collected in this case due to the practical operating mechanism. The initial conception of records is introduced by Chandler [7] that could be described as follows. Let T_n, $n = 1, 2, ...$ be a series of independent and identically distributed (i.i.d.) random variables. Then an observation T_j is called an upper record, if $T_j > T_i$ for every $j > i$. Due to its wide application in practical fields, records have received wide attention and are discussed by many authors. See, for example, some recent contributions of Asgharzadeh *et al.* [3], Dey *et al.* [9], Singh *et al.* [25], Wang and Ye [30] among others. For more details, one could refer to monographes of Ahsanullah [1] and Nevzorov [23] for a comprehensive review.

In practical lifetime studies, various distributions like exponential, Weibull, gamma, normal, etc., have been proposed for data analysis from various perspectives. One of characteristics of these aforementioned traditional models is that these distributions all feature infinity support $(-\infty, \infty)$ or $(0, \infty)$. However, there are many situations, where observations collected from practical situation are bounded within a specified range, and in turn distributions with finite support may provide better modelling performance than those with infinity support from goodness-of-fitting perspectives. For example, Zhang and Xie [32] used an upper-truncated Weibull distribution to fit the pit depth data of a water pipe where the upper bound of pit depths is the thickness of the water pipe. The same model is further implemented to describe the wind speed data by Kantar and Usta [17]. Vicari *et al.* [27] proposed a generalized Topp–Leone distribution for fitting the V-I indices data of globular clusters with bounded support. Under such aforementioned studies, all of the authors mentioned that the implemented bounded models have better data fitting accuracy than traditional distributions with infinity support in their practical discussions. Therefore, distributions with bounded domain have potential theoretical and practical applications where such models may provide higher weight to the bounded data and give better fitting effect in data analysis, and has been extensively studied by many authors from various perspectives (e.g., [5], [8], [20], [26]).

Among different bounded models, distributions with unit support have attracted considerable attention in practice, where the associated observation within $(0, 1)$ is an important

and common occurred data type in reality such as birth rate, mortality data, as well as indices data from fields of energy, reliability and economic among others. There are various distributions with unit bound like beta, Kumaraswamy, Topp–Lenoe models among others. Some discussions and applications for such unit models could be found in the works of Genc [11], Ghitany *et al.* [13], Makouei *et al.* [22] and Wang [31]. Recently, Jha *et al.* [16] proposed another unit generalized Rayleigh distribution (UGRD) as follows. Let T be an UGRD random variable, the associated cumulative distribution function (CDF), probability density function (PDF) and hazard rate function (HRF) of T are respectively given by

$$F(t) = 1 - \left(1 - e^{-(\lambda \ln t)^2}\right)^{\theta}, \quad 0 < t < 1, \tag{1.1}$$

$$f(t) = -\frac{2\theta\lambda^2 \ln t}{t} e^{-(\lambda \ln t)^2} \left(1 - e^{-(\lambda \ln t)^2}\right)^{\theta-1}, \tag{1.2}$$

and

$$H(t) = \frac{-\frac{2\theta\lambda^2 \ln t}{t} e^{-(\lambda \ln t)^2}}{1 - e^{-(\lambda \ln t)^2}}, \tag{1.3}$$

where $\theta > 0$ and $\lambda > 0$ are shape and scale parameters, respectively. It is noted that the shape parameter θ affects the geometric shape of density curve and the scale parameter λ not only determines the steepness of density curve but also specifically exhibits the value of random variable. Hereafter, the UGRD with parameters θ and λ is denoted by UGRD(θ, λ) for concision. Further, plots of CDF, PDF and HRF of the UGRD are presented in Figure 1 for illustration, and it is noted visually that the UGRD has very flexible fitting ability and may be used as an alternative bound model to traditional Beta, Kumaraswamy and Topp–Leone distributions.

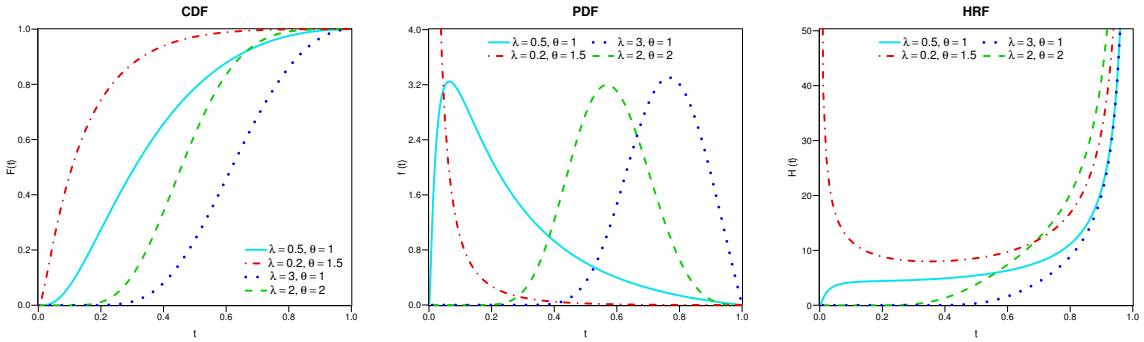


Figure 1: CDF, PDF and HRF of UGRD with different parameters.

In both theoretical and practical studies, point estimation is one of the most used approach in statistical inference. However, point estimation sometimes could not produce robust results, especially when estimates heavily depend on sample size. Since sample size appears frequently as moderate or small due to practical limitations, estimations of confidence sets are proposed alternatively in consequence, and have been discussed by many authors from different perspectives. For instance, Asgharzadeh *et al.* [2] provided the exact confidence intervals and regions when a bathtub-shaped distribution is used, and similar results are also obtained by Kinaci *et al.* [18] for the parameters of the generalized inverted exponential distribution.

Wu [28] constructed the confidence sets for the Weibull parameters under progressively censored data. Based on a modified progressively hybrid censored data, Zhu [33] proposed an adaptive Newton–Raphson algorithm based exact confidence region for a bathtub-shaped distribution. Motivated by such reasons as mentioned above and due to the flexibility and wide applications of the UGRD, the current investigation explores estimations of confidence sets for the UGRD parameters when records data is available, and various approaches are presented for constructing confidence intervals and confidence regions for the UGRD parameters in consequence.

The rest parts of this paper are arranged as follows. In Section 2, various estimates of confidence sets with equal-tailed and minimum-size are established for the UGRD parameters. Extensive numerical simulations are carried out in Section 3 to investigate the performance of different results, and two real life examples are also presented for illustrations. In Section 4, some extended results are further provided for exploring some more potential confidence sets for UGRD parameters with better performance. Finally, some concluding remarks are given in Section 5.

2. ESTIMATIONS OF CONFIDENCE SETS

Based on records data, different confidence sets of the UGRD parameters are established in this section. The equal-tailed confidence intervals and confidence regions are constructed respectively based on the proposed pivotal quantities, and the associated minimum-size confidence sets are also established in consequence. Moreover, conventional asymptotic confidence sets are provided for comparison.

2.1. Equal-tailed confidence sets

The equal-tailed confidence sets (ECSs) are discussed here for UGRD parameters λ and θ including the equal-tailed confidence intervals (ECI) and equal-tailed confidence region (ECR), respectively.

To construct the ECSs, two useful results are provided as follows.

Lemma 2.1. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper records from UGRD(λ, θ). Denote pivotal quantities*

$$\Psi(\lambda) = (n-1)\left[\frac{\ln(1-\exp(-(\lambda \ln T_n)^2))}{\ln(1-\exp(-(\lambda \ln T_1)^2))} - 1\right]^{-1} \tag{2.1}$$

and

$$\Upsilon(\lambda, \theta) = -2\theta \ln(1-\exp(-(\lambda \ln T_n)^2)). \tag{2.2}$$

Then $\Psi(\lambda)$ follows the F distribution with 2 and $2(n-1)$ degrees of freedom, $\Upsilon(\lambda, \theta)$ has a chi-square distribution with $2n$ degree of freedom, and $\Psi(\lambda)$ and $\Upsilon(\lambda, \theta)$ are statistically independent.

Proof: The proof is provided in part A of the Supplementary file. □

Lemma 2.2. *For arbitrary numbers a and b with $0 < b < a < 1$, let*

$$h(\lambda) = \frac{\ln(1 - \exp(-(\lambda \ln a)^2))}{\ln(1 - \exp(-(\lambda \ln b)^2))}, \lambda > 0, \tag{2.3}$$

then function $h(\lambda)$ increases in λ with $\lim_{\lambda \to 0} h(\lambda) = 1$ and $\lim_{\lambda \to \infty} h(\lambda) = \infty$.

Proof: The proof is provided in part B of the Supplementary file. □

Corollary 2.1. *According to Lemma 2.2, function $\Psi(\lambda)$ decreases in λ with range $(0, \infty)$.*

In the following, the ECIs of parameters λ and θ as well as the ECR of parameter vector (λ, θ) are established, respectively.

Theorem 2.1. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper record from UGRD(λ, θ). For arbitrary $0 < \gamma < 1$, a $100(1-\gamma)\%$ ECI of λ is given by*

$$\left[\psi\left(F_{\gamma/2}^{2,2(n-1)}\right), \psi\left(F_{1-\gamma/2}^{2,2(n-1)}\right)\right], \tag{2.4}$$

where $\psi(x)$ is the solution of equation $\Psi(\lambda) = x$ w.r.t. λ, and $F_p^{m_1,m_2}$ is the upper $100p\%$ percentile of F distribution with m_1 and m_2 degrees of freedom.

Proof: The proof is provided in part C of the Supplementary file. □

Theorem 2.2. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper record from UGRD(λ, θ). For given λ and arbitrary $0 < \gamma < 1$, a $100(1-\gamma)\%$ ECI of θ can be constructed as*

$$\left[\frac{\chi_{1-\gamma/2}^{2n}}{B(\lambda)}, \frac{\chi_{\gamma/2}^{2n}}{B(\lambda)}\right] \quad with \quad B(\lambda) = -2\ln(1 - \exp(-(\lambda \ln T_n)^2)), \tag{2.5}$$

where χ_p^m denotes the upper $100p\%$ percentile of chi-square distirbution with m degrees of freedom.

Proof: Using the distribution property of the pivotal quantity $\Upsilon(\lambda, \theta)$ given in Lemma 2.1, the result could be be established directly by following similar line as Theorem 2.1, and the details are omitted here for concision. □

Remark 2.1. It is noted from Theorem 2.2 that the ECI of θ is available with known λ. However, parameter λ is unknown in practical applications. To overcome this drawback, following alternative way is proposed to establish the ECI of parameter θ when parameter λ is unknown.

Let $\psi(Y)$ be the unique solution of λ from equation $\Psi(\lambda) = Y$, where Y is a random sample generating from F distribution with 2 and $2(n-1)$ degrees of freedom. Using the substitution method of Weerahandi [29], a generalized pivotal quantity of θ can be constructed as

$$S = \frac{\Upsilon(\psi(Y), \theta)}{B(\psi(Y))}. \tag{2.6}$$

Correspondingly, an approach termed as Algorithm 1 is provided to obtain the ECI of θ under unknown λ situation.

Algorithm 1: ECI of θ with unknown λ

STEP 1. Generate a random sample Y from the F distribution with 2 and $2(n-1)$ degrees of freedom, then $\psi(Y)$ can be solved through equation $\Psi(\lambda) = Y$.

STEP 2. Generate a random value of $\Upsilon(\psi(Y), \theta)$ from chi-square distirbution with $2n$ degrees of freedom and calculate S in (2.6).

STEP 3. Repeat Steps 1 and 2 M times and obtain a group values of S arranged in the ascending order, $S_1, S_2, ..., S_M$.

STEP 4. For $0 < \gamma < 1$, an ECI of θ with unknown λ can be constructed by

$$\left[S_{\lceil M\frac{\gamma}{2}\rceil}, S_{\lceil M(1-\frac{\gamma}{2})\rceil}\right], \tag{2.7}$$

where '$\lceil\cdot\rceil$' refers to the ceiling function.

Further, an ECR of parameter vector (λ, θ) is established as follows.

Theorem 2.3. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper record from UGRD(λ, θ). For arbitrary $0 < \gamma < 1$, a $100(1-\gamma)\%$ ECR of (λ,θ) can be written as*

$$\left\{(\lambda,\theta)\middle|\psi\left(F^{2,2(n-1)}_{\frac{1-\sqrt{1-\gamma}}{2}}\right) < \lambda < \psi\left(F^{2,2(n-1)}_{\frac{1+\sqrt{1-\gamma}}{2}}\right), \frac{\chi^{2n}_{\frac{1+\sqrt{1-\gamma}}{2}}}{B(\lambda)} < \theta < \frac{\chi^{2n}_{\frac{1-\sqrt{1-\gamma}}{2}}}{B(\lambda)}\right\}, \tag{2.8}$$

where associated notations are defined in Theorems 2.1 and 2.2, respectively.

Proof: The proof is provided in part D of the Supplementary file. □

2.2. Minimum-size confidence sets

It is noted from Subsection 2.1 that the proposed ECSs are obtained under equal-tailed approach, and such results someitmes may not have minimum sizes. Alternatively, optimal confidence sets for parameters λ, θ and (λ, θ) are proposed here. Specifically, the minimum-size confidence sets (MCSs) including the minimum-length confidence intervals (MCIs) of λ and θ as well as the minimum-area confidence region (MCR) of (λ, θ) are constructed respectively, and the associated numercial algorithms are also proposed for optimization computation.

Theorem 2.4. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper record from UGRD(λ, θ). For arbitrary $0 < \gamma < 1$, a $100(1-\gamma)\%$ MCI of λ is given by*

$$[\psi(x_2^*), \psi(x_1^*)], \tag{2.9}$$

where x_1^* and x_2^* are the solutions of the following non-linear system

$$\begin{cases} \frac{\psi'(x_1)}{\psi'(x_2)} = \frac{P^F_{2,2(n-1)}(x_1)}{P^F_{2,2(n-1)}(x_2)}, \\ F_{2,2(n-1)}(x_2) - F_{2,2(n-1)}(x_1) = 1-\gamma, \end{cases}$$

where $\psi'(x)$ is the derivative of $\psi(x)$ with respect to x, $F_{m_1,m_2}(x)$ is the CDF of F distribution with m_1 and m_2 degrees of freedom and $P^F_{m_1,m_2}(x)$ is the corresponding density function of $F_{m_1,m_2}(x)$.

Proof: The proof is provided in part E of the Supplementary file. □

Clearly, there is no closed form solution (x_1^*, x_2^*) for the MCI of λ in Theorem 2.4, and a numerical approach entitled Algorithm 2 is provided with pre-fixed accuracy level $\sigma > 0$.

Algorithm 2: MCI of λ in Theorem 2.4

STEP 1. Let $\dot{x}_1 = F_\gamma^{2,2(n-1)}$ be the upper bound of x_1 and $N = \lfloor \dot{x}_1/\sigma \rfloor$, where '$\lfloor\cdot\rfloor$' is the floor function.

STEP 2. Obtain a value of $\dot{x}_2$ by computing the equation $F_{2,2(n-1)}(\dot{x}_2) - F_{2,2(n-1)}(\dot{x}_1) = 1 - \gamma$.

STEP 3. Let $\dot{x}_1 = \dot{x}_1 - \sigma$.

STEP 4. Repeat Step 2 and Step 3 until $\dot{x}_1 < 0$ and obtain N groups $(\dot{x}_1^{[i]}, \dot{x}_2^{[i]})$, $i = 1, ..., N$.

STEP 5. The numerical MCI of λ can be constructed as $\left[\psi\left(\dot{x}_2^{[k]}\right), \psi\left(\dot{x}_1^{[k]}\right)\right]$, where k satisfies the equation $\psi(\dot{x}_1^{[k]}) - \psi(\dot{x}_2^{[k]}) = \min_{i=1}^{N}[\psi(\dot{x}_1^{[i]}) - \psi(\dot{x}_2^{[i]})]$.

Theorem 2.5. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper record from UGRD(λ, θ). For given λ and arbitrary $0 < \gamma < 1$, a $100(1-\gamma)\%$ MCI of θ can be constructed as*

$$\left[\frac{y_1^*}{B(\lambda)}, \frac{y_2^*}{B(\lambda)}\right], \tag{2.10}$$

where y_1^* and y_2^* are the solutions of the following non-linear system

$$\begin{cases} P_{2n}^{\chi}(y_2) = P_{2n}^{\chi}(y_1), \\ \chi_{2n}(y_2) - \chi_{2n}(y_1) = 1 - \gamma \end{cases}$$

and both $\chi_m(y)$ and $P_m^{\chi}(y)$ are the CDF and PDF of chi-square distribution with m degree of freedom, respectively.

Proof: For given λ, using pivotal quantity $\Upsilon(\lambda, \theta)$ in Lemma 2.1, the $100(1-\gamma)\%$ MCI of θ can be obtained similarly as in Theorem 2.4 and the details are omitted for concision. □

In addition, a numerical approach called Algorithm 3 are presented for obtaining the MCI of θ.

Remark 2.2. It is noted that the MCI of θ with unknown λ can be still constructed under the alternative approach as

$$\left[S_{\lceil j^* \rceil}, S_{\lceil j^*+M-(M\gamma+1)\rceil}\right], \tag{2.11}$$

where notation $S_{\lceil . \rceil}$ is defined in Algorithm 1 and j^* is an integer satisfying

$$S_{\lceil j^*+M-(M\gamma+1)\rceil} - S_{\lceil j^* \rceil} = \min_{j=1}^{\lceil M\gamma \rceil} \left[S_{\lceil j+M-(M\gamma+1)\rceil} - S_{\lceil j \rceil}\right].$$

Algorithm 3: MCI of θ in Theorem 2.5

STEP 1. Let $p = \sigma$ be the initial value.

STEP 2. Obtain the solutions $\dot{y}_1$ and $\dot{y}_2$ from euqation $P_{2n}^{\chi}(y) = p$, where $0 < \dot{y}_1 < \dot{y}_2$.

STEP 3. Calculate $C = \chi_{2n}(\dot{y}_2) - \chi_{2n}(\dot{y}_1)$, then let $\sigma^* = C - (1-\gamma)$.

STEP 4. If $\sigma^* > \sigma$, then let $p = p + \sigma$, otherwise if $\sigma^* < -\sigma$, let $p = p - \sigma$.

STEP 5. Repeat Steps 2–4 until $|\sigma^*| \leqslant \sigma$, for known λ, the numerical MCI of θ can be given by $[\dot{y}_1/B(\lambda), \dot{y}_2/B(\lambda)]$.

Similarly, the MCR of (λ, θ) is also established as follows.

Theorem 2.6. *Let $T = \{T_1, T_2, ..., T_n\}$ be upper record from UGRD(λ, θ). For an arbitrary $0 < \gamma < 1$, the $100(1-\gamma)\%$ MCR of (λ, θ) is given by*

$$\left\{(\lambda, \theta) \middle| \psi(x_2^*) < \lambda < \psi(x_1^*), \frac{y_1^*}{B(\lambda)} < \theta < \frac{y_2^*}{B(\lambda)}\right\}, \tag{2.12}$$

where $(x_1^*, x_2^*, y_1^*, y_2^*)$ are the solutions of the following non-linear system

$$\begin{cases} P_{2n}^{\chi}(y_1) = P_{2n}^{\chi}(y_2), \\ \left[F_{2,2(n-1)}(x_2) - F_{2,2(n-1)}(x_1)\right]\left[\chi_{2n}(y_2) - \chi_{2n}(y_1)\right] = 1-\gamma, \\ \dfrac{\psi'(x_2)B(\psi(x_1))}{\psi'(x_1)B(\psi(x_2))} = -\dfrac{P_{2,2(n-1)}^{F}(x_2)}{P_{2,2(n-1)}^{F}(x_1)}, \\ \dfrac{\psi'(x_1)\left[F_{2,2(n-1)}(x_2) - F_{2,2(n-1)}(x_1)\right]P_{2n}^{\chi}(y_1)}{\left[\chi_{2n}(y_2) - \chi_{2n}(y_1)\right]P_{2,2(n-1)}^{F}(x_1)\int_{\psi(x_2)}^{\psi(x_1)} \frac{1}{B(\lambda)} d\lambda} = -\dfrac{B(\psi(x_1))}{y_2 - y_1}. \end{cases}$$

Proof: The proof is provided in part F of the Supplementary file. □

For finding solution of MCR, the associated numerical approach termed as Algorithm 4 is provided in consequence.

Algorithm 4: MCR of (λ, θ) in Theorem 2.6

STEP 1. Set $\dot{y}_1 = \sigma$ and obtain $\tilde{y}_1$ and $\tilde{y}_2$ from the following equations

$$\begin{cases} P_{2n}^{\chi}(\tilde{y}_1) = P_{2n}^{\chi}(\tilde{y}_2) \\ \chi_{2n}(\tilde{y}_2) - \chi_{2n}(\tilde{y}_1) = (1-\gamma) \end{cases}, 0 < \tilde{y}_1 < \tilde{y}_2.$$

Then make $M = \lfloor \tilde{y}_1/\sigma \rfloor$.

STEP 2. Obtain a value of $\dot{y}_2 (> \dot{y}_1)$ by equation $P_{2n}^{\chi}(\dot{y}_1) = P_{2n}^{\chi}(\dot{y}_2)$, and calculate $\gamma^* = 1 - (1-\gamma)[\chi_{2n}(\dot{y}_2) - \chi_{2n}(\dot{y}_1)]^{-1}$.

STEP 3. For $i = 1, 2, ..., M$, obtain N_i groups $(\dot{x}_1^{[i_j]}, \dot{x}_2^{[i_j]})$, $j = 1, 2, ..., N_i$ by substituting γ^* for γ in Algorithm 1.

STEP 4. Let $\dot{y}_1 = \dot{y}_1 + \sigma$.

STEP 5. Repeat Step 2–Step 4 until $\dot{y}_1 > \tilde{y}_1$ and obtain $\sum_{i=1}^{M} N_i$ groups solutions $(\dot{x}_1^{[i_j]}, \dot{x}_2^{[i_j]}, \dot{y}_1^{[i]}, \dot{y}_2^{[i]})$, $i = 1, 2, ..., M$, $j = 1, 2, ..., N_i$.

STEP 6. The numercial MCR of (λ, θ) can be constructed as

$$\left\{ (\lambda, \theta) \middle| \psi\left(\dot{x}_2^{[i_{j^*}^*]}\right) < \lambda < \psi\left(\dot{x}_1^{[i_{j^*}^*]}\right), \frac{\dot{y}_1^{[i^*]}}{B(\lambda)} < \theta < \frac{\dot{y}_2^{[i^*]}}{B(\lambda)} \right\},$$

where $(\dot{x}_1^{[i_{j^*}^*]}, \dot{x}_2^{[i_{j^*}^*]}, \dot{y}_1^{[i^*]}, \dot{y}_2^{[i^*]})$ conform to the following equation

$$\int_{\psi\left(\dot{x}_2^{[i_{j^*}^*]}\right)}^{\psi\left(\dot{x}_1^{[i_{j^*}^*]}\right)} \frac{\dot{y}_2^{[i^*]} - \dot{y}_1^{[i^*]}}{B(\lambda)} d\lambda = \min_{(i,j)=(1,1)}^{(M,N_i)} \left[\int_{\psi\left(\dot{x}_2^{[i_j]}\right)}^{\psi\left(\dot{x}_1^{[i_j]}\right)} \frac{\dot{y}_2^{[i]} - \dot{y}_1^{[i]}}{B(\lambda)} d\lambda \right].$$

2.3. Asymptotic confidence sets

For comparison, traditional asymptotic confidence sets (ACSs) of UGRD parameters are also constructed based on asymptotic theory, where asymptotic confidence intervals (ACIs) of of λ and θ as well as asymptotic confidence region (ACR) of of (λ, θ) are obtained, respectively.

Let $T_1, T_2, ..., T_n$ be upper records from UGRD(λ, θ), and $t_1, t_2, ..., t_n$ be the associated observations. Therefore, log-likelihood function $\ell(\lambda, \theta)$ of λ and θ can be expressed from Ahsanullah [1] as

$$\begin{aligned} \ell(\lambda, \theta) \;=\; & n \ln(2\theta\lambda^2) + \theta \ln\left(1 - e^{-(\lambda \ln t_n)^2}\right) \\ & - \sum_{i=1}^{n} \ln\left[-\left(1 - e^{-(\lambda \ln t_i)^2}\right) t_i^{-1} \ln t_i \right] + (\lambda \ln t_i)^2. \end{aligned} \tag{2.13}$$

By taking derivatives, MLE $\hat{\lambda}$ of λ can be obtained from equation

$$\frac{n}{\lambda^2} - \frac{n(\ln t_n)^2 e^{-(\lambda \ln t_n)^2}}{\left(1 - e^{-(\lambda \ln t_n)^2}\right) \ln\left(1 - e^{-(\lambda \ln t_n)^2}\right)} - \sum_{i=1}^{n} \frac{(\ln t_i)^2}{1 - e^{-(\lambda \ln t_i)^2}} = 0, \tag{2.14}$$

whereas the MLE $\hat{\theta}$ of θ is given by

$$\hat{\theta} = -\frac{n}{\ln\left(1 - e^{-\left(\hat{\lambda} \ln t_n\right)^2}\right)}.$$

Remark 2.3. It is worth mentioning that the UGRD MLEs $\hat{\lambda}$ and $\hat{\theta}$ uniquely exist under records situation, and the associated existence and uniqueness are provided in part G of the Supplementary file. Therefore, altough there is no closed form for MLE $\hat{\lambda}$ in equation (2.14), the associated estimate could be obtained in a simple way by using numerical approaches like biasection or fixed-point iteration methods.

Further, let $\boldsymbol{\beta} = (\lambda, \theta)' = (\beta_1, \beta_2)'$ with $\beta_1 = \lambda, \beta_2 = \theta$, the observed information matrix of $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)'$ is given by

$$\mathbf{I}(\hat{\boldsymbol{\beta}}) = \left(\begin{array}{cc} -\frac{\partial^2 \ell(\lambda,\theta)}{\partial \lambda^2} & -\frac{\partial^2 \ell(\lambda,\theta)}{\partial \lambda \partial \theta} \\ -\frac{\partial^2 \ell(\lambda,\theta)}{\partial \lambda \partial \theta} & -\frac{\partial^2 \ell(\lambda,\theta)}{\partial \theta^2} \end{array} \right) \Bigg|_{\lambda=\hat{\lambda}, \theta=\hat{\theta}}, \tag{2.15}$$

where

$$\frac{\partial^2 \ell(\lambda,\theta)}{\partial^2 \lambda} = -\frac{2n}{\lambda^2} - \omega(t_n) + \theta \sum_{i=1}^{n} \omega(t_i), \quad \frac{\partial^2 \ell(\lambda,\theta)}{\partial^2 \theta} = -\frac{n}{\theta^2},$$
$$\frac{\partial^2 \ell(\lambda,\theta)}{\partial \lambda \partial \theta} = \frac{2\lambda (\ln t_n)^2 e^{-(\lambda \ln t_n)^2}}{1 - e^{-(\lambda \ln t_n)^2}}$$

and

$$\omega(t) = \frac{2(\ln t)^2 e^{-(\lambda \ln t)^2} \left[1 - 2\lambda^2 (\ln t)^2 - e^{-(\lambda \ln t)^2}\right]}{\left[1 - e^{-(\lambda \ln t)^2}\right]^2}.$$

Therefore, the variance-covariance matrix of $(\hat{\lambda}, \hat{\theta})$ can be constructed as

$$\mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \mathrm{Var}(\hat{\lambda}) & \mathrm{Cov}(\hat{\lambda}, \hat{\theta}) \\ \mathrm{Cov}(\hat{\lambda}, \hat{\theta}) & \mathrm{Var}(\hat{\theta}) \end{pmatrix}.$$

Consequently, the asymptotic distribution of $\hat{\boldsymbol{\beta}}$ can be obtained under some mild regularity conditions as $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} \to N(0, \mathbf{I}^{-1}(\hat{\boldsymbol{\beta}}))$.

For arbitrary $0 < \gamma < 1$, the $100(1-\gamma)\%$ ACI of β_i can be constructed by

$$\left[\hat{\beta}_i + u_{1-\gamma/2}\sqrt{\mathrm{Var}(\hat{\beta}_i)}, \hat{\beta}_i + u_{\gamma/2}\sqrt{\mathrm{Var}(\hat{\beta}_i)}\right], i = 1, 2, \tag{2.16}$$

where u_γ is the upper $100\gamma\%$ percentile of standard normal distribution. Moreover, the $100(1-\gamma)\%$ ACR of (λ, θ) can be obtained as follows

$$\left\{(\lambda, \theta) \middle| (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mathbf{I}(\hat{\boldsymbol{\beta}}) (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) < \chi^2_\gamma \right\}. \tag{2.17}$$

Remark 2.4. In some cases, the lower confidence bounds of the ACIs (2.16) sometimes may be negative. To overcome this drawback, one could use logarithmic transformation and delta method to obtain the asymptotic normality distribution of $\ln\hat{\beta}_i, i = 1, 2$ as $\ln\hat{\beta}_i - \ln\beta_i \to N\Big(0, \mathrm{Var}\Big(\ln\hat{\beta}_i\Big)\Big)$, with $\mathrm{Var}(\ln\hat{\beta}_i) = \mathrm{Var}(\hat{\beta}_i)/\hat{\beta}_i^2$. Therefore, the $100(1-\gamma)\%$ modified ACI of β_i can be constructed in this manner as

$$\left[\frac{\hat{\beta}_i}{\exp\left(u_{\gamma/2}\sqrt{\mathrm{Var}\Big(\ln\hat{\beta}_i\Big)}\right)}, \hat{\beta}_i \exp\left(u_{\gamma/2}\sqrt{\mathrm{Var}\Big(\ln\hat{\beta}_i\Big)}\right)\right], i = 1, 2. \tag{2.18}$$

3. NUMERICAL ILLUSTRATION

Extensive simulation studies are carried out to investigate the performance of the proposed results. In addition, two real-life examples are also presented to show the applicability of our methods.

3.1. Simulation studies

In simulation studies, performance of ECSs, MCSs and ACSs are compared in terms of criteria quantities including average width (AW) for confidence intervals, average area (AA) for confidence regions and coverage probability (CP) for all confidence sets.

For generating records data, another sampling approach termed as Algorithm 5 is provided as follows.

Algorithm 5: Upper record values from UGRD(λ, θ)

STEP 1. Generate n i.i.d. samples $u_1, u_2, ..., u_n$ from uniform distribution with range $(0, 1)$.

STEP 2. Calculate $v_i = -\ln(1-u_i),\ i = 1, 2, ..., n$.

STEP 3. Take $w_i = 1 - \exp(-\sum_{j=1}^{i} v_j)$, then $w_1, w_2, ..., w_n$ are upper records standard uniform distribution.

STEP 4. Implementing inverse transformation

$$t_i = \exp[-(-\ln(1-(1-w_i)^{1/\theta}))^{1/2}/\lambda],\ i = 1, 2, ..., n$$

then $t_1, t_2, ..., t_n$ are the upper record values from UGRD(λ, θ).

In this simulation, parameter values (λ, θ) are randomly chosen as $(0.5, 1)$, $(0.5, 0.5)$ and $(3, 1)$, sample sizes $n = 3$, 4, 5, 6, 7 and 8 are considered and the significance level is $\gamma = 0.05$.

For all numerical computation in minimum-size confidence sets, the accuracy level is taken to be $\sigma = 0.001$, and the simulations are conducted based 10,000 times of repetitions, where the ECI and MCI of θ are obtained under unknown λ cases by using the strategies provided in Remarks 2.1 and 2.2. The simulated associated criteria quantities AW, AA and CP are tabulated in Tables 1–3. In addition, for complementary and comparison, performance of ECI and MCI for θ given in Theorems 2.2 and 2.5 are also investigated with known λ, the associated criteria quantities are obtained by using the true values of λ in simulation and the associted results are tabulated in Table 4.

Table 1: AWs, AAs and CPs (within brackets) for UGRD confidence sets with $\lambda = 0.5,\ \theta = 1$.

		$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
λ	ACI	1.9864 (0.8226)	1.7601 (0.8835)	1.6818 (0.9043)	1.6299 (0.9244)	1.6118 (0.9370)	1.5961 (0.9434)
	ECI	1.8250 (0.9523)	1.6381 (0.9479)	1.5820 (0.9486)	1.5363 (0.9483)	1.5169 (0.9518)	1.5036 (0.9484)
	MCI	1.6773 (0.9279)	1.5208 (0.9286)	1.4777 (0.9342)	1.4434 (0.9322)	1.4294 (0.9395)	1.4208 (0.9346)
θ	ACI	8.3863 (0.9791)	5.0168 (0.9748)	3.3495 (0.9765)	2.5414 (0.9773)	2.0836 (0.9728)	1.8349 (0.9753)
	ECI	9.0399 (0.9334)	5.2259 (0.9361)	3.1536 (0.9410)	2.3884 (0.9454)	1.9646 (0.9436)	1.7430 (0.9446)
	MCI	5.7770 (0.9496)	4.0595 (0.9443)	2.7762 (0.9508)	2.1864 (0.9557)	1.8336 (0.9509)	1.6346 (0.9489)
(λ,θ)	ACR	12.5066 (0.7982)	8.0762 (0.9010)	5.5025 (0.9369)	4.1949 (0.9559)	3.4725 (0.9629)	3.0587 (0.9626)
	ECR	14.0787 (0.9417)	8.5107 (0.9523)	5.2939 (0.9536)	4.1231 (0.9478)	3.3691 (0.9537)	2.9745 (0.9461)
	MCR	9.0862 (0.9451)	6.3987 (0.9469)	4.5192 (0.9527)	3.5098 (0.9523)	2.9310 (0.9531)	2.6252 (0.9475)

Table 2: AWs, AAs and CPs (within brackets) for UGRD confidence sets with $\lambda = 0.5,\ \theta = 0.5$.

		$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
λ	ACI	5.4513 (0.8705)	5.1275 (0.9194)	5.2425 (0.9465)	4.688 (0.9608)	4.5967 (0.9703)	4.383 (0.9744)
	ECI	4.4116 (0.9489)	4.1500 (0.9490)	4.3133 (0.9492)	3.8481 (0.9476)	3.7679 (0.9491)	3.6012 (0.9500)
	MCI	3.9530 (0.9404)	3.7088 (0.9426)	3.8517 (0.9388)	3.4429 (0.9412)	3.3775 (0.9456)	3.2236 (0.9465)
θ	ACI	4.9927 (0.9815)	2.1162 (0.9783)	1.3507 (0.9752)	1.1324 (0.9807)	0.9712 (0.9835)	0.8753 (0.9908)
	ECI	4.7218 (0.9391)	1.9903 (0.9416)	1.2216 (0.9451)	1.0477 (0.9486)	0.9117 (0.9539)	0.8309 (0.9569)
	MCI	3.4154 (0.9449)	1.6898 (0.9535)	1.1217 (0.9584)	0.9786 (0.9624)	0.8604 (0.9615)	0.7893 (0.9568)
(λ,θ)	ACR	20.0447 (0.8857)	7.6137 (0.9481)	5.7514 (0.9661)	4.4766 (0.9747)	3.8706 (0.9838)	3.3478 (0.9881)
	ECR	22.3566 (0.9517)	6.6294 (0.9519)	4.9081 (0.9524)	3.7994 (0.9514)	3.2875 (0.9622)	2.8493 (0.9587)
	MCR	14.3081 (0.9530)	5.3074 (0.9532)	4.0968 (0.9516)	3.2366 (0.9537)	2.8335 (0.9616)	2.4759 (0.9641)

Table 3: AWs, AAs and CPs (within brackets) for UGRD confidence sets with $\lambda = 3,\ \theta = 1$.

		$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
λ	ACI	11.7857 (0.8416)	10.6667 (0.8856)	10.0701 (0.9086)	9.7815 (0.9252)	9.5784 (0.9352)	9.4684 (0.9422)
	ECI	10.9231 (0.9431)	9.9324 (0.9497)	9.4579 (0.9498)	9.179 (0.9511)	9.0211 (0.9501)	8.9084 (0.953)
	MCI	10.0412 (0.9382)	9.2208 (0.9331)	8.8449 (0.9341)	8.6182 (0.9359)	8.5062 (0.9365)	8.4214 (0.9412)
θ	ACI	11.8480 (0.9799)	5.0491 (0.9799)	3.3021 (0.9748)	2.5118 (0.9734)	2.103 (0.975)	1.8341 (0.9742)
	ECI	12.7022 (0.9425)	4.9763 (0.947)	3.0889 (0.9391)	2.3403 (0.9411)	1.9792 (0.9434)	1.7371 (0.9447)
	MCI	8.2348 (0.9479)	4.0257 (0.9529)	2.7213 (0.9526)	2.1438 (0.9504)	1.8473 (0.9512)	1.639 (0.9499)
(λ,θ)	ACR	96.2407 (0.8087)	49.7994 (0.9003)	32.6615 (0.9376)	24.8758 (0.9530)	20.9596 (0.9606)	18.2307 (0.9645)
	ECR	98.9104 (0.9537)	50.5229 (0.9502)	32.5485 (0.9517)	24.2006 (0.9499)	20.4082 (0.95)	17.6974 (0.9505)
	MCR	73.1492 (0.9467)	38.7621 (0.9504)	26.6496 (0.9522)	20.6069 (0.9508)	17.7536 (0.9523)	15.6203 (0.9513)

Table 4: AWs and CPs (within brackets) of ECIs and MCIs of θ with known λ.

$\lambda = 0.5,\ \theta = 1$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
ECI	3.3429 (0.9499)	2.5739 (0.9487)	2.1547 (0.9516)	1.8892 (0.9492)	1.7095 (0.9518)	1.5629 (0.9525)
MCI	3.0857 (0.9496)	2.4258 (0.9486)	2.0559 (0.9514)	1.8173 (0.9491)	1.6540 (0.9519)	1.5186 (0.9534)

$\lambda = 0.5,\ \theta = 0.5$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
ECI	1.6397 (0.9530)	1.2780 (0.9521)	1.0774 (0.9515)	0.9416 (0.9570)	0.8589 (0.9607)	0.7948 (0.9741)
MCI	1.5136 (0.9526)	1.2045 (0.9518)	1.0281 (0.9524)	0.9058 (0.9534)	0.8309 (0.9615)	0.7722 (0.9728)

$\lambda = 3,\ \theta = 1$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
ECI	3.3284 (0.9440)	2.5828 (0.9463)	2.1577 (0.9533)	1.8856 (0.9507)	1.7096 (0.9498)	1.5634 (0.9515)
MCI	3.0723 (0.9461)	2.4342 (0.9476)	2.0589 (0.9522)	1.8139 (0.9484)	1.6540 (0.9480)	1.5190 (0.9521)

From Tables 1–4, it is noted that

- AWs and AAs of all confidence sets decrease with increase of sample size n. Such phenomenon indicate the consistence property of the proposed results when sample size increases.
- Under fixed sample size n, MCIs of λ have the best performance than ECIs and ACIs in terms of AWs, whereas the ACI estimates of λ feature the largest AWs. In addition, the ACIs for λ have lowest CPs than those of ECIs and MCIs.

- For parameter θ, MCIs of θ are superior to ECIs and ACIs in terms of AWs, whereas although the CPs of ACI are highest, AWs of ACIs are larger than the other two interval estimates in general.
- For confidence region of (λ, θ), MCRs have the smallest AAs, whereas the CPs of all three confidence regions of (λ, θ) are close to the nominal significance level in most cases.
- From Table 4, for given parameter λ, AWs of both ECIs and MCIs of θ are smaller than those of θ with unknown λ shown in Tables 1–3, and the CPs under this case are still close to the nominal significance level.

To sum up, the simulation results indicate that the ECSs and MCSs of different parameters λ, θ and (λ, θ) perform better than traditional likelihood based ACRs in general, and the proposed MCSs are recommended as superior choices in practice.

3.2. Real data illustration

In this subsection, two real life examples are presented to demonstrate the practicality of the proposed methods. For comparison, another three unit bounded distributions namely Beta distribution (BeD), Kumaraswamy distribution (KuD) and Topp–Leone distribution (TLD) are considered as competitors of UGRD in this illustration. The corresponding PDFs of BeD, KuD and TLD are given respectively by:

$$\begin{aligned} BeD:&\ f_1(t) = t^{\alpha-1}(1-x)^{\beta-1}[B(\alpha,\beta)]^{-1},\ \alpha>0, \beta>0,\ 0<t<1,\\ KuD:&\ f_2(t) = \alpha\beta t^{\alpha-1}(1-t^{\alpha})^{\beta-1},\ \alpha>0,\ \beta>0,\ 0<t<1,\\ TLD:&\ f_3(t) = 2\alpha(1-t)t^{\alpha-1}(2-t)^{\alpha-1},\ 0<\alpha<1,\ 0<t<1. \end{aligned}$$

It is noted that above three distributions are also common used unit models that are widely implemented in practical data analysis. (e.g., Arora *et al.* [4], Gupta and Nadarajah [10] and Kohansal [19]).

Example 1. (Reservoir capacity ratio data) In this real life example, a data set from `http://cdec.water.ca.gov/dynamicapp/QueryMonthly?s=SHA` is considered for illustration. The data set is about the water capacities of Shasta reservoir in California, USA for the month of October from 2008 to 2019. Since the maximum capacity of Shasta reservoir is 4552000 acre-foot, the original data were converted to the $(0, 1)$ interval by dividing 4552000 acre-foot. The transformed data are shown as follows:

$$\begin{aligned} &0.28180,\ 0.37520,\ 0.71891,\ 0.70887,\ 0.54177,\ 0.38317\\ &0.24360,\ 0.31113,\ 0.60748,\ 0.69789,\ 0.48134,\ 0.71859. \end{aligned}$$

To check whether the UGRD could be used to fit the real life data, Kolmogorov–Smirnov (K-S) test is carried out for UGRD, BeD, KuD and TLD respectively under origin complete data, the associated results are tabulated in Table 5. It is noted from Table 5 that comparing with BeD, KuD and TLD, the UGRD seems more proper to fit the reservoir capacity ratio data.

Table 5: MLEs and K-S test of fitted distributions for reservoir capacity ratio data.

	MLE of model parameters	K-S distance	p-value
UGRD	$(\hat{\lambda}, \hat{\theta}) = (1.2369, 1.1284)$	0.1893	0.7166
BeD	$(\hat{\alpha}, \hat{\beta}) = (3.9505, 3.8693)$	0.1949	0.6834
KuD	$(\hat{\alpha}, \hat{\beta}) = (2.9870, 4.7499)$	0.1959	0.6777
TLD	$\hat{\alpha} = 2.8190$	0.2071	0.6111

In addition, the corresponding quantile-quantile (Q-Q) plot, probability-probability (P-P) plot and empirical cumulative distribution (ECD) plot of UGRD are also provided in Figure 2, which also indicates that the UGRD is a reasonable model used here.

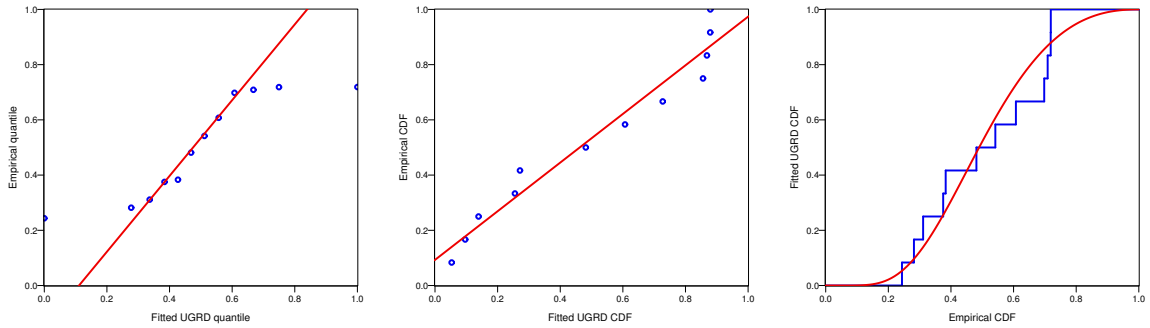


Figure 2: Q-Q, P-P and ECD plots of UGRD under the reservoir capacity ratio data.

Based on the reservoir capacity ratio data, a group of records data of size 3 is obtained as follows:

$$0.28180,\ 0.37520,\ 0.71891.$$

Then different confidence sets are estimated with $\gamma = 0.05$ and $\sigma = 10^{-4}$, and the associated results are listed in Tables 6 and 7 respectively, where the widths of confidence intervals and the areas of confidence regions are provided in parentheses. It is observed that the MCSs outperform the other competitors according to their criteria widths and areas. In Table 6, it is also noted that the lower confidence bounds of ACIs are negative. Using the alternative results in Remark 2.4, the modified ACIs and associated interval widths (within parentheses) of λ and θ are $[0.3264, 2.2162](1.8898)$ and $[0.2652, 5.0985](4.8333)$ respectively.

Table 6: Confidence intervals for λ and θ under reservoir capacity ratio records data.

Parameter	ACI	ECI	MCI
λ	[-0.0326,1.7337] (1.7663)	[0.0007,1.5843] (1.5836)	[0.0001,1.4140] (1.4139)
θ	[-0.4311,2.7566] (3.1877)	[0.1042,3.0771] (2.9729)	[0.0279,2.4935] (2.4656)

Table 7: Confidence regions for (λ, θ) under reservoir capacity ratio records data.

	Confidence regions	Areas
ACR	$\left\{ \begin{pmatrix} 0.8506-\lambda \\ 1.1628-\theta \end{pmatrix}' \begin{pmatrix} 0.2030 & 0.2068 \\ 0.2068 & 0.6613 \end{pmatrix} \begin{pmatrix} 0.8506-\lambda \\ 1.1628-\theta \end{pmatrix} < 5.9915 \right\}$	(5.6939)
ECR	$\left\{ 0.0001 < \lambda < 1.7333, \frac{0.9528}{B(\lambda)} < \theta < \frac{16.2120}{B(\lambda)} \right\}$	(5.4465)
MCR	$\left\{ 0.0002 < \lambda < 1.5389, \frac{0.3790}{B(\lambda)} < \theta < \frac{15.1281}{B(\lambda)} \right\}$	(4.2208)

Moreover, for further illustration, the confidence regions and contour of log-likelihood function (2.13) are plotted in Figure 3 to show the superority of MCR and the uniqueness of MLEs in this real data example.

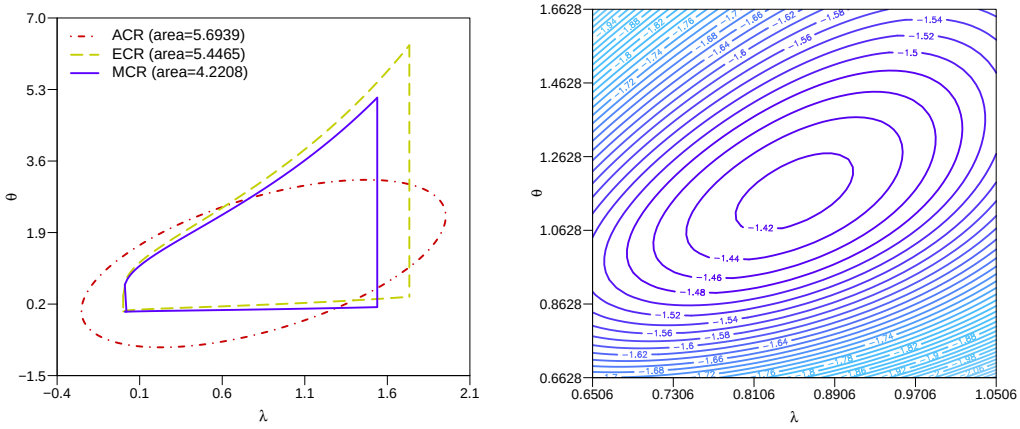


Figure 3: Contour of log-likelihood function (left) and confidence regions (right) under reservoir capacity ratio records data.

Example 2. (Electricity supply rate data) Another real life data set drawn out from `https://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?end=2018&locations=KE&start=2009` is used for illustration. The data set is about the electricity supply rate in Kenya from 2009 to 2018. The original data are shown as follows:

$$0.23000,\ 0.19200,\ 0.38581,\ 0.40793,\ 0.43049,$$
$$0.36000,\ 0.41600,\ 0.65400,\ 0.63589,\ 0.75000.$$

By computation, MLEs and corresponding K-S test results for UGRD, BeD, KuD and TLD are listed in Table 8 under these data. It is also noted that UGRD have best performance among these models to fit the electricity supply rate data. Meanwhile, the associated Q-Q, P-P and ECD plots of UGRD are also shown in Figure 4 for illustration.

Similarly, records from the original observations are given as follows:

$$0.23000,\ 0.38581,\ 0.40793,\ 0.43049,\ 0.65400,\ 0.75000.$$

Table 8: MLEs and K-S test results of fitted distributions for electricity supply rate data.

	MLE of model parameters	K-S distance	p-value
UGRD	$(\hat{\lambda}, \hat{\theta}) = (1.1002, 1.2703)$	0.1971	0.7636
BeD	$(\hat{\alpha}, \hat{\beta}) = (3.4950, 4.3081)$	0.2277	0.6008
KuD	$(\hat{\alpha}, \hat{\beta}) = (2.5636, 5.0465)$	0.2391	0.5404
TLD	$\hat{\alpha} = 2.2064$	0.2790	0.3505

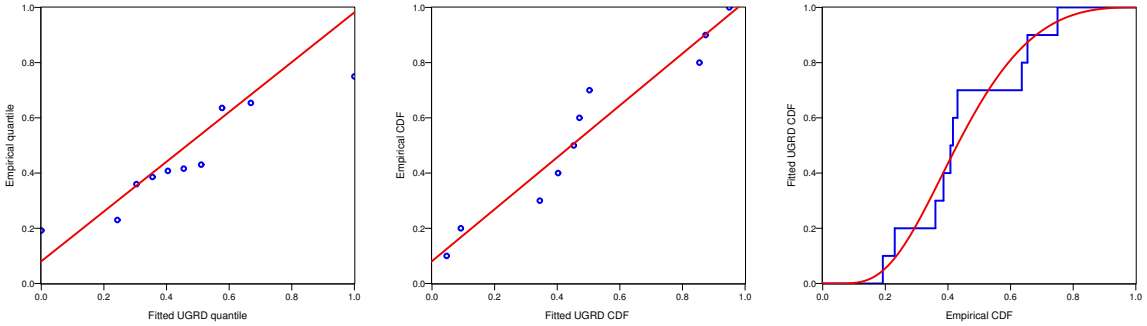


Figure 4: Q-Q plot, P-P plot and ECD plot for the electricity supply rate data of UGRD.

Different ACSs, ECSs and MCSs are shown in Tables 9 and 10 under same setting as Example 1. From the results in Tables 9 and 10, the MCSs still perform best among all estimates.

Table 9: Confidence intervals for λ and θ under electricity supply rate records data.

Parameter	ACI	ECI	MCI
λ	$[0.2086, 1.4980]$ (1.2894)	$[0.1165, 1.4757]$ (1.3592)	$[0.0700, 1.3917]$ (1.3217)
θ	$[0.1008, 4.1260]$ (4.0253)	$[0.5652, 4.5006]$ (3.9351)	$[0.3726, 3.9809]$ (3.6083)

Table 10: Confidence regions for (λ, θ) under electricity supply rate records data.

	Confidence Regions	Areas
ACR	$\left\{ \begin{pmatrix} 0.8533-\lambda \\ 2.1134-\theta \end{pmatrix}' \begin{pmatrix} 13.0909 & -2.2738 \\ -2.2738 & 1.3433 \end{pmatrix} \begin{pmatrix} 0.8533-\lambda \\ 2.1134-\theta \end{pmatrix} < 5.9915 \right\}$	(5.3421)
ECR	$\left\{ 0.0703 < \lambda < 1.5907, \frac{3.7632}{B(\lambda)} < \theta < \frac{25.4910}{B(\lambda)} \right\}$	(5.8436)
MCR	$\left\{ 0.0001 < \lambda < 1.4603, \frac{2.8890}{B(\lambda)} < \theta < \frac{24.1103}{B(\lambda)} \right\}$	(4.9916)

In addition, the plots of confidence regions and contour curve of log-likelihood function are also presented in Figure 5 for illustration and comparison.

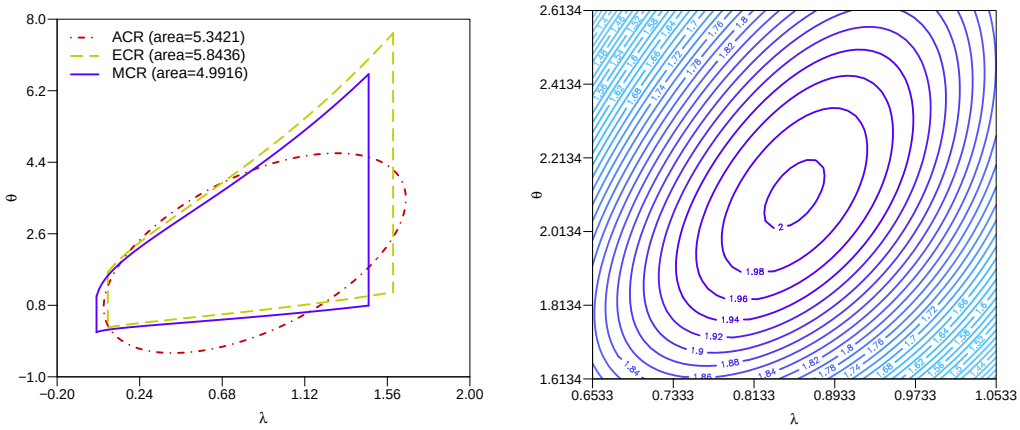


Figure 5: Contour of log-likelihood function (left) and confidence regions (right) under electricity supply rate records data.

4. EXTENSION WORK

In statistical inference, accuracy of confidence sets is one of main concerns in analysis which in turn affects the practical performance of applications. Following similar approach of previous inferential procedure, some extended results are proposed in this section for complementary, where a series of pivotal quantities is constructed, and then alternative generalized confidence sets are provided in consequence.

Using notations $W_1, W_2, ..., W_n$ and associated distribution properties given in Lemma 2.1, let $\xi_k = \sum_{i=1}^{k} W_i$ and $\eta_k = \sum_{i=k+1}^{n} W_i$, $k = 1, 2, ..., n-1$, one has

$$\Psi_k(\lambda) = \frac{2\xi_k/2k}{2\eta_k/2(n-k)} = \frac{(n-k)}{k}\left[\frac{\ln\left(1-\exp\left(-(\lambda \ln T_n)^2\right)\right)}{\ln\left(1-\exp\left(-(\lambda \ln T_k)^2\right)\right)} - 1\right]^{-1} \tag{4.1}$$

and

$$\Upsilon_k(\lambda, \theta) = 2(\xi_k + \eta_k) = -2\theta \ln(1 - \exp(-(\lambda \ln T_n)^2)), k = 1, ..., n-1 \tag{4.2}$$

follow F and chi-square distributions with $(2k, 2(n-k))$ and $2n$ degrees of freedom, respectively. Meanwhile, quantities $\Psi_k(\lambda)$ and $\Upsilon_k(\lambda, \theta)$ are statistically independent. Moreover, it is also noted from Lemma 2.2 that $\Psi_k(\lambda)$ decreases in λ with range $(0, \infty)$.

Using quantities $\Psi_k(\lambda), \Upsilon_k(\lambda, \theta)$ and following similar way as Theorems 2.1 and 2.3, for arbitrary $0 < \gamma < 1$, a series of $100(1-\gamma)\%$ ECIs of λ can be constructed as

$$\left[\psi_k\left(F_{\gamma/2}^{2k,2(n-k)}\right), \psi_k\left(F_{1-\gamma/2k}^{2,2(n-k)}\right)\right], k = 1, 2, ..., n-1, \tag{4.3}$$

where $\psi_k(x)$ refers to the solution of $\Psi_k(\lambda) = x$, and correspondingly a group of $100(1-\gamma)\%$ ECRs of (λ, θ) can be written as

$$\left\{(\lambda,\theta)\middle|\psi_k\left(F^{2k,2(n-k)}_{\frac{1-\sqrt{1-\gamma}}{2}}\right) < \lambda < \psi_k\left(F^{2k,2(n-k)}_{\frac{1+\sqrt{1-\gamma}}{2}}\right), \frac{\chi^{2n}_{\frac{1+\sqrt{1-\gamma}}{2}}}{B(\lambda)} < \theta < \frac{\chi^{2n}_{\frac{1-\sqrt{1-\gamma}}{2}}}{B(\lambda)}\right\} \tag{4.4}$$

with $k = 1, 2, ..., n-1$.

It is observed that there are $n-1$ confidence intervals and regions obtained under this manner, and their sizes may be different with different k. To find the optimal confidence sets among the proposed results, the following criterions are provided.

Criterion 1. The best ECI of λ is obtained as k^*-th one among proposed ECIs, where k^* satisfies

$$\psi_{k^*}\left(F^{2k^*,2(n-k^*)}_{1-\gamma/2}\right) - \psi_{k^*}\left(F^{2k^*,2(n-k^*)}_{\gamma/2}\right) = \min_{k=1}^{n-1}\left[\psi_k\left(F^{2k,2(n-k)}_{1-\gamma/2}\right) - \psi_k\left(F^{2k,2(n-k)}_{\gamma/2}\right)\right].$$

Criterion 2. The best ECR of (λ,θ) is obtained as k^*-th one among all ECRs, where k^* satisfies

$$\int_{\psi_{k^*}\left(F^{2k^*,2(n-k^*)}_{\frac{1-\sqrt{1-\gamma}}{2}}\right)}^{\psi_{k^*}\left(F^{2k^*,2(n-k^*)}_{\frac{1+\sqrt{1-\gamma}}{2}}\right)} \frac{\chi^{2n}_{\frac{1-\sqrt{1-\gamma}}{2}} - \chi^{2n}_{\frac{1+\sqrt{1-\gamma}}{2}}}{B(\lambda)} d\lambda = \min_{k=1}^{n-1}\left[\int_{\psi_k\left(F^{2k,2(n-k)}_{\frac{1-\sqrt{1-\gamma}}{2}}\right)}^{\psi_k\left(F^{2k,2(n-k)}_{\frac{1+\sqrt{1-\gamma}}{2}}\right)} \frac{\chi^{2n}_{\frac{1-\sqrt{1-\gamma}}{2}} - \chi^{2n}_{\frac{1+\sqrt{1-\gamma}}{2}}}{B(\lambda)} d\lambda\right].$$

Note from (4.2) that, since $\Upsilon_k(\lambda,\theta) = \Upsilon(\lambda,\theta)$ does not change with k and the associated ECI of θ coincides with the results obtained in Theorem 2.2. Further, following the similar approach of Subsection 2.2, a series of MCSs for parameters λ and (λ,θ) could be also obtained based on pivotal quantities $\Psi_k(\lambda)$ and $\Upsilon_k(\lambda,\theta)$, $k = 1, 2, ..., n-1$, the detailed results are omitted for concision and saving space. In addition, the optimal confidence sets for such MCSs could be also selected by using similar method as shown in Criterions 1 and 2.

For illustration, the confidence sets for parameters λ and (λ,θ) are reconstructed by using the records data in Example 2 with $k = 1, 2, 3, 4, 5$, where the significance level is $\gamma = 0.05$ as same as previous. The associated results are tabulated in Tables 11 and 12. From the results, it is seen that for ECSs estimation, the optimal ECI of λ and ECR of (λ,θ) are obtained at $k = 4$, whereas the associated optimal MCI of λ and MCR of (λ,θ) are also obtained at $k = 4$. In addition, one also note that the MCSs perform better than the associated ECSs at given k and that all the confidence sets obtained by using the proposed pivotal quantities $\Psi_k(\lambda), \Upsilon_k(\lambda,\theta), k = 4$ have smaller sizes than the ACSs in Tables 9 and 10. Further, plots of extended ECR, MCR of (λ,θ) with $k = 4$ and traditional ACR are also presented in Figure 6, which indicate that the proposed extended confidence regions have better performance in this manner.

Table 11: ECIs for λ and ECRs for (λ, θ) with different k under electricity supply rate records data.

k	ECIs	ECRs
1	$[0.1165, 1.4757]$ (1.3592)	$\left\{0.0703 < \lambda < 1.5907, \frac{3.7632}{B(\lambda)} < \theta < \frac{25.4910}{B(\lambda)}\right\}$ (5.8436)
2	$[0.0512, 1.6687]$ (1.6175)	$\left\{0.0220 < \lambda < 1.8063, \frac{3.7632}{B(\lambda)} < \theta < \frac{25.4910}{B(\lambda)}\right\}$ (7.4175)
3	$[0.0015, 1.2970]$ (1.2955)	$\left\{0.0002 < \lambda < 1.4326, \frac{3.7632}{B(\lambda)} < \theta < \frac{25.4910}{B(\lambda)}\right\}$ (4.9472)
4	$[0.0001, 0.9519]$ (0.9518)	$\left\{0.0001 < \lambda < 1.0930, \frac{3.7632}{B(\lambda)} < \theta < \frac{25.4910}{B(\lambda)}\right\}$ (3.1794)
5	$[0.0001, 2.2521]$ (2.2520)	$\left\{0.0001 < \lambda < 2.5783, \frac{3.7632}{B(\lambda)} < \theta < \frac{25.4910}{B(\lambda)}\right\}$ (15.0503)

Note: the interval widths and region areas are listed in the parentheses.

Table 12: MCIs for λ and MCRs for (λ, θ) with different k under electricity supply rate records data.

k	MCIs	MCRs
1	$[0.0700, 1.3917]$ (1.3217)	$\left\{0.0001 < \lambda < 1.4603, \frac{2.8890}{B(\lambda)} < \theta < \frac{24.1103}{B(\lambda)}\right\}$ (4.9916)
2	$[0.0101, 1.5144]$ (1.5043)	$\left\{0.0001 < \lambda < 1.6370, \frac{2.8346}{B(\lambda)} < \theta < \frac{24.3403}{B(\lambda)}\right\}$ (6.1677)
3	$[0.0001, 1.1420]$ (1.1419)	$\left\{0.0002 < \lambda < 1.2459, \frac{2.7388}{B(\lambda)} < \theta < \frac{24.7603}{B(\lambda)}\right\}$ (3.9759)
4	$[0.0001, 0.7925]$ (0.7924)	$\left\{0.0001 < \lambda < 0.8702, \frac{2.5663}{B(\lambda)} < \theta < \frac{25.5503}{B(\lambda)}\right\}$ (2.3655)
5	$[0.0001, 1.8748]$ (1.8747)	$\left\{0.0001 < \lambda < 2.0143, \frac{2.4165}{B(\lambda)} < \theta < \frac{26.2903}{B(\lambda)}\right\}$ (10.0198)

Note: the interval widths and region areas are listed in the parentheses.

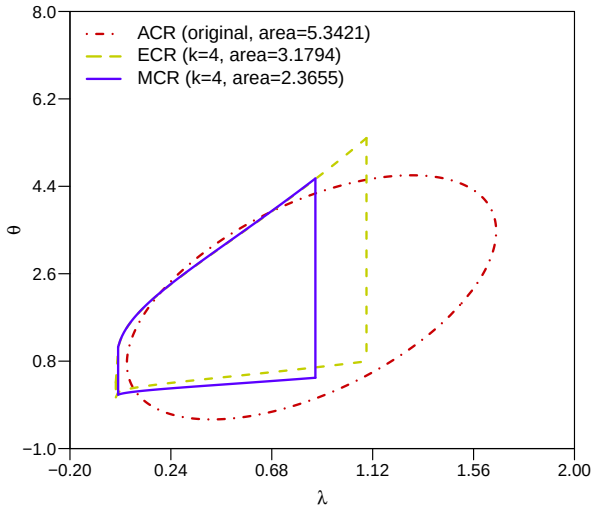


Figure 6: Plots of ECR and MCR with $k = 4$ and traditional ACR.

5. CONCLUSION

In this paper, different confidence sets of parameters from the unit generalized Rayleigh distribution are explored under records data. By constructing pivotal quantities, equal-tailed confidence sets are established for model parameters. Further, the associated minimum-size confidence sets are constructed based on optimization techniques, and the algorithms along with Lagrange multiplier method are also provided for computation. In addition, conventional likelihood based asymptotic confidence sets are also constructed for comparison. Extensive simulation studies and two real life examples are carried out to investigate the performance of different methods, and the results indicate that the proposed pivotal quantities based ECSs and MCSs perform better than common likelihood based confidence sets. Furthermore, a series of confidence sets are also proposed as extension based on constructed alternative pivotal quantities which sometimes may further provide potential better estimates.

ACKNOWLEDGMENTS

We express our sincere thanks to the Editor and anonymous reviewers for their useful comments and suggestions on an earlier version of this manuscript which led to this improved version. This work of Liang Wang is supported by the National Natural Science Foundation of China (No. 12061091, 12201395), the Yunnan Fundamental Research Projects (No. 202401AT070116) and the Yunnan Key Laboratory of Modern Analytical Mathematics and Applications (No. 202302AN360007). This work of Xuanjia Zuo was supported by the Postgraduate Innovation Research Funding of Yunnan Normal University (No. YJSJJ21-B36).

REFERENCES

[1] Ahsanullah, M. (2004). *Record Values – Theory and Applications*, University Press of America, New York.

[2] Asgharzadeh, A.; Abdi, M. and Wu, S.J. (2015). Interval estimation for the two-parameter bathtub-shaped lifetime distribution based on records, *Hacettepe Journal of Mathematics and Statistics*, **44**(2), 399–416.

[3] Asgharzadeh, A.; Fallah, A.; Raqab, M.Z. and Valiollahi, R. (2018). Statistical inference based on Lindley record data, *Statistical Papers*, **59**(2), 759–779.

[4] Arora, S.; Mahajan, K.K. and Kumari, R. (2021). Bayes estimators for the reliability and hazard rate functions of Topp–Leone distribution using Type-II censored data, *Communications in Statistics – Simulation and Computation*, **50**(8), 2327–2344.

[5] Bakouch, H.S.; Hussain, T.; Chesneau, C. and Jonas, T. (2022). A notable bounded probability distribution for environmental and lifetime data, *Earth Science Informatics*, **15**, 1607–1620.

[6] Balakrishnan, N. and Cramer, E. (2014). *Art of Progressive Censoring*, Springer, New York.

[7] Chandler, K.N. (1952). The distribution and frequency of record values, *Journal of the Royal Statistical Society: Series B*, **14**(2), 220–228.

[8] Condino, F. and Domma, F. (2017). A new distribution function with bounded support: the reflected generalized Topp–Leone power series distribution, *METRON*, **75**, 51–68.

[9] Dey, S.; Dey, T. and Luckett, D.J. (2016). Statistical inference for the generalized inverted exponential distribution based on upper record values, *Mathematics and Computers in Simulation*, **120**, 64–78.

[10] Gupta A.K. and Nadarajah S. (2004). *Handbook of Beta Distribution and Its Applications*, CRC Press, Boca Raton.

[11] Genc, A.I. (2013). Estimation of P(X>Y) with Topp–Leone distribution, *Journal of Statistical Computation and Simulation*, **83**(2), 326–339.

[12] Ghitany, M.E.; Tuan, V.K. and Balakrishnan, N. (2014). Likelihood estimation for a general class of inverse exponentiated distributions based on complete and progressively censored data, *Journal of Statistical Computation and Simulation*, **84**(1), 96–106.

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

[14] Guo, B.; Zhu, N.; Wang, W. and Wang, H. (2020). Constructing exact tolerance intervals for the exponential distribution based on record values, *Quality and Reliability Engineering International*, **36**(7), 2398–2410.

[15] Johnson, N.L.; Kotz, S. and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, Wiley, New York.

[16] Jha, M.K.; Tripathi, Y.M. and Dey, S. (2020). Multicomponent stress-strength reliability estimation based on unit generalized Rayleigh distribution, *International Journal of Quality & Reliability Management*, **38**(10), 2048–2079.

[17] Kantar, Y.M. and Usta, I. (2015). Analysis of the upper-truncated Weibull distribution for wind speed, *Energy Conversion and Management*, **96**, 81–88.

[18] Kinaci, I.; Wu, S.J. and Kud, C. (2019). Confidence intervals and regions for the generalized inverted exponential distribution based on progressively censored and upper records data, *REVSTAT – Statistical Journal*, **17**(4), 429–448.

[19] Kohansal, A. (2022). Inference on stress-strength model for a Kumaraswamy distribution based on hybrid progressive censored sample, *REVSTAT – Statistical Journal*, **20**(1), 51–83.

[20] Korkmaz, M.C. (2020). A new heavy-tailed distribution defined on the bounded interval: the logit slash distribution and its application, *Journal of Applied Statistics*, **47**(12), 2097–2119.

[21] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, Wiley, New York.

[22] Makouei, R.; Khamnei, H.J. and Salehi, M. (2021). Moments of order statistics and k-record values arising from the complementary beta distribution with application, *Journal of Computational and Applied Mathematics*, **390**, 113386.

[23] Nevzorov, V.B. (2001). *Records: Mathematical Theory*, American Mathematical Society, Rhode Island.

[24] Soliman, A.A.; Abd Ellah, A.H. and Sultan, K.S. (2006). Comparison of estimates using record statistics from Weibull model: Bayesian and non-Bayesian approaches, *Computational Statistics & Data Analysis*, **51**(3), 2065–2077.

[25] Singh, S.; Dey, S. and Kumar, D. (2020). Statistical inference based on generalized Lindley record values, *Journal of Applied Statistics*, **47**(9), 1543–1561.

[26] Teamah, A.E.A.; El-Hadidy, M.A.A. and ElGhoul, M.M. (2022). On bounded range distribution of a Wiener process, *Communications in Statistics – Theory and Methods*, **51**(4), 919–942.

[27] Vicari, D.; Van Dorp, J.R. and Kotz, S. (2008). Two-sided generalized Topp and Leone (TS-GTL) distributions, *Journal of Applied Statistics*, **35**(10), 1115–1129.

[28] Wu, S.J. (2002). Estimations of the parameters of the Weibull distribution with progressively censored data, *The Japan Statistical Society*, **32**(2), 155–163.

[29] Weerahandi, S. (2004). *Generalized Inference in Repeated Measures: Exact Methods in MANOVA and Mixed Models*, Wiley, New York.

[30] Wang, B.X. and Ye, Z. (2015). Inference on the Weibull distribution based on record values, *Computational Statistics & Data Analysis*, **83**, 26–36.

[31] Wang, L. (2018). Inference of progressively censored competing risks data from Kumaraswamy distributions, *Journal of Computational and Applied Mathematics*, **343**, 719–736.

[32] Zhang, T. and Xie, M. (2011). On the upper truncated Weibull distribution and its reliability implications, *Reliability Engineering & System Safety*, **96**(1), 194–200.

[33] Zhu, T. (2021). A new approach for estimating parameters of two-parameter bathtub-shaped lifetime distribution under modified progressive hybrid censoring, *Quality and Reliability Engineering International*, **37**(5), 2288–2304.

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 503–526
https://doi.org/10.57805/revstat.v22i4.524

A Simple Mean-Parameterized Maxwell Regression Model for Positive Response Variables

Author: Artur J. Lemonte
– Departamento de Estatística, Universidade Federal do Rio Grande do Norte, Brazil
arturlemonte@gmail.com

Received: May 2021 Revised: December 2022 Accepted: December 2022

Abstract:

- We study a quite simple parametric regression model that may be very useful to model positive response variables in practice. The frequentist approach is considered to perform inferences, and the traditional maximum likelihood method is employed to estimate the unknown parameters. Monte Carlo simulation results indicate that the maximum likelihood approach is quite effective to estimate the model parameters. We also derive a closed-form expression for the second-order bias of the maximum likelihood estimator, which is easily computed as an ordinary linear regression and is then used to define bias-corrected maximum likelihood estimates. We consider the normalized quantile residuals for the new parametric regression model to assess departures from model assumptions, and global and local influence methods are also discussed. Applications to real data are considered to illustrate the new regression model in practice, and comparisons with two of the most popular existing regression models are made.

Keywords:

- *maximum likelihood estimation; Maxwell distribution; Maxwell-Boltzmann distribution; parametric inference.*

AMS Subject Classification:

- 62F10, 62J05, 62J20.

✉ Corresponding author

1. INTRODUCTION

The Maxwell distribution, also known in the statistic and physic literatures as Maxwell–Boltzmann distribution, has probability density function (PDF) in the form

$$f(y;\alpha)=\sqrt{\frac{2}{\pi}}\frac{y^2\mathrm{e}^{-y^2/(2\alpha^2)}}{\alpha^3},\quad y>0,$$

where $\alpha>0$ is the scale parameter. The mean and variance of the Maxwell distribution reduce to

$$\mathbb{E}(Y)=2\alpha\sqrt{\frac{2}{\pi}},\quad \text{and}\quad \mathbb{VAR}(Y)=\frac{\alpha^2(3\pi-8)}{\pi}.$$

There are, of course, some works related specifically to the one-parameter Maxwell distribution in the statistic literature. The reader is referred to Tyagi and Bhattacharya [31], Bekker and Roux [3], Dey and Maiti [13], Dey *et al.* [12], Al-Baldawi [1], Li [24], Fan [15], Dar *et al.* [11] and Hossain *et al.* [20], among others. It is evident that the one-parameter Maxwell distribution has noticeable scientific importance and, of course, it leaves open quite a number of new directions of research. In this paper, we provide a complete study regarding this one-parameter family of distributions in a parametric regression setup on the basis of a mean-parameterized Maxwell distribution.

In a parametric regression framework, it is typically more useful to model directly the mean (mode or median) of the response variable. In the last few years, several works have been published and so contributed to the regression literature on parameterizations based on the mean, mode, or median. To mention a few, but not limited to, we refer the reader to Yao and Li [33], Lemonte and Bazan [23], Chen *et al.* [6], Castellares *et al.* [5], Bourguignon *et al.* [4], Gallardo *et al.* [17], Gómez *et al.* [19], Leão *et al.* [22] and Menezes *et al.* [26]. In this paper, in order to obtain a regression structure for the mean of the Maxwell distribution, we shall work with a different parameterization of the Maxwell PDF. Let $\mu=2\alpha(2/\pi)^{1/2}$ and, hence, $\alpha=(1/2)\mu(2/\pi)^{-1/2}$. In this case, substituting this expression in the Maxwell PDF, a reparameterization for the PDF is obtained; that is, the mean-parameterized Maxwell PDF is given by

$$f(y;\mu)=\left(\frac{2}{\pi}\right)^2\frac{8y^2}{\mu^3}\exp\left(-\frac{4y^2}{\pi\mu^2}\right),\quad y>0, \tag{1.1}$$

so that $\mathbb{E}(Y)=\mu>0$ is the mean of the Maxwell distribution. Additionally, we have that $\mathbb{VAR}(Y)=0.178\mu^2\propto\mu^2$. The cumulative distribution function (CDF) of the mean-parameterized Maxwell takes the form

$$F(y,\mu)=\frac{2\gamma(3/2,4y^2/(\pi\mu^2))}{\sqrt{\pi}},\quad y>0,$$

where $\gamma(a,x)=\int_0^x t^{a-1}\mathrm{e}^{-t}dt$ is the lower incomplete gamma function. We shall use the notation $\mathrm{Mw}(\mu)$ to refer to this distribution. We have that $\lim_{y\to0}f(y)=\lim_{y\to\infty}f(y)=0$ and, in addition, the mode is simply given by $\mu\sqrt{\pi}/2$. The Maxwell failure rate function is given by

$$r(y)=\left(\frac{2}{\pi}\right)^2\frac{8y^2}{\mu^3}\left[1-\frac{2\gamma(3/2,4y^2/(\pi\mu^2))}{\sqrt{\pi}}\right]^{-1}\exp\left(-\frac{4y^2}{\pi\mu^2}\right),\quad y>0.$$

Figure 1 displays some plots of the PDF and failure rate function for some values of μ.

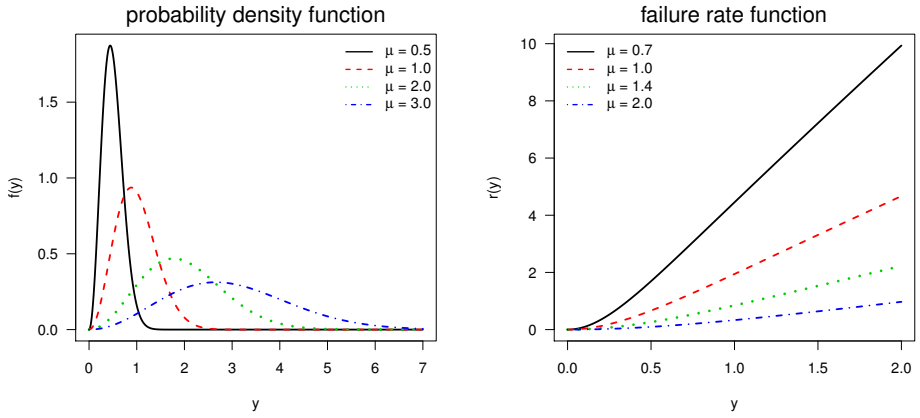


Figure 1: Density and failure rate functions.

We have the following propositions.

Proposition 1. The Maxwell PDF is log-concave for all values of $\mu > 0$.

Proof: The result follows by noting that the second derivative of $\log(f(y;\mu))$ is given by

$$\frac{\mathrm{d}^2 \log(f(y;\mu))}{\mathrm{d}y^2} = -\left(\frac{2}{y^2} + \frac{8}{\pi\mu^2}\right) < 0. \qquad \square$$

Proposition 2. For any $\mu > 0$, the Maxwell failure rate function is monotone increasing.

Proof: The result holds by using the log-concavity of the Maxwell PDF. $\square$

Remark 1. It is rather easy to generate random variates from the mean-parameterized Maxwell distribution. If U follows a gamma distribution with shape parameter $3/2$ and scale parameter 1, then $Y = \mu\sqrt{\pi U/4} \sim \mathrm{Mw}(\mu)$.

In this paper, we shall provide a parametric regression structure for the Maxwell distribution parameter, which involves covariates (explanatory variables) and unknown regression parameters. Furthermore, some quantities (e.g., score function, Fisher information matrix, etc.) related to the mean-parameterized Maxwell regression model are simple and compact, which makes the frequentist approach very easy to implement. Obviously the Bayesian approach has its merits and could also be considered and, in addition, these methodologies could be compared and contrasted. However, the comparison of these two methodologies is beyond the scope of this paper and hence can be considered in a future work. Also, it is quite common in practice, after modeling the real data at hand, to check the regression model assumptions and conduct diagnostic studies in order to detect possible atypical observations that may distort the results of the analysis. A first way to perform sensitivity

analysis is by means of global influence starting from the case deletion proposed by Cook [7]. In addition, Cook [8] introduced a general framework to detect atypical observations under small perturbations on the data or in the model. In this paper, global and local influence are also considered to detect atypical observations in the class of Maxwell regression models. Throughout this paper, an *atypical* observation means that it can be an outlier[1], or observation with a large residual in absolute value, or an influential observation in the sense of global or local influences. Finally, it is well-known the residuals carry important information concerning the appropriateness of assumptions that underlie statistical models, and thereby play an important role in checking model adequacy identifying discrepancies between models and data. Hence, we propose the normalized quantile residual introduced by Dunn and Smyth [14] for the Maxwell regression model to study discrepancies between the model and data. In summary, the main contributions of this paper are as follows:

- We propose a Maxwell distribution parameterized in terms of its mean, allowing easy interpretation of the distribution parameter.
- Based on the mean-parameterized Maxwell distribution, we propose a novel parametric regression model for positive response variables, which is quite simple and may be very useful in practice, allowing for parameter interpretation in terms of the response in the original scale; that is, the regression parameters are interpretable in terms of the mean of the variable of interest.
- The direct modeling of the mean parameter in the mean-parameterized Maxwell regression model will promote its wider use in practice, putting it on the same level of interpretability and parsimony of some well-known regression models for positive response variables.
- The simulation and data analysis examples in this article reinforce that the proposed framework is a quite simple yet flexible way to model positive response variables.

The rest of this paper is organized as follows. The mean-parameterized Maxwell regression model is introduced in Section 2, and likelihood-based inference, as well as Monte Carlo simulation experiments are also performed. In Section 3, we propose diagnostic measures (i.e., global and local influence) for the mean-parameterized Maxwell regression model and, in particular, the normal curvature of local influence is derived under a specific perturbation scheme, namely: case weighting perturbation. Additionally, we also consider the normalized quantile residual to assess departures from the underlying distribution. Section 4 contains real data applications of the mean-parameterized Maxwell regression model for illustrative purposes. The paper ends up with some concluding remarks in Section 5.

2. THE MAXWELL REGRESSION MODEL

The model. Let $Y_1, ..., Y_n$ be n independent random variables, where each Y_i $(i = 1, ..., n)$ is Maxwell distributed and has PDF (1.1) with mean parameter μ_i; that is, $Y_i \sim \mathrm{Mw}(\mu_i)$ for $i = 1, ..., n$. In this work, we assume the following functional relation:

$$\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}, \tag{2.1}$$

[1] An outlying observation, or "outlier," is one that appears to deviate markedly from other members of the sample in which it occurs.

where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is a vector of unknown regression coefficients, $\boldsymbol{\beta} \in \mathbb{R}^p$ with $p < n$, and $\boldsymbol{x}_i^\top = (x_{i1}, ..., x_{ip})$ are observations on p known covariates (or independent variables, or regressors). Generally, we have $x_{i1} = 1$ (for $i = 1, ..., n$) in practice and, hence, β_1 corresponds to the intercept parameter. It is worth emphasizing that other other links for the mean parameter in (2.1) could be considered, namely: identity ($\mu_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$), and square root ($\sqrt{\mu_i} = \boldsymbol{x}_i^\top \boldsymbol{\beta}$). However, the logarithm function is the most common and useful in such a case; that is, the main advantage of the exponential form $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ is that the requirement $\mu_i > 0$ is automatically satisfied for all $i = 1, ..., n$, whereas the identity and square root do not ensure such a requirement for all $i = 1, ..., n$. Note that the variance $\mathbb{VAR}(Y_i) = 0.178\mu_i^2 \propto \mu_i^2$ is a function of μ_i and, as a consequence, of the covariate values. Hence, non-constant response variances are naturally accommodated into the regression model. Moreover, we assume that the model matrix $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]^\top$ has column rank p.

Remark 2. Let ξ be the mode of the mean-parameterized Maxwell distribution, and so we have that $\xi = \mu\sqrt{\pi}/2$. The mean-parameterized Maxwell regression model is defined by the link function $\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\beta}$, for $i = 1, ..., n$, where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$, and $\boldsymbol{x}_i^\top = (x_{i1}, ..., x_{ip})$. Let $x_{i1} = 1$ (for $i = 1, ..., n$) and, hence, β_1 corresponds to the intercept parameter. In this case, $\log(\mu_i) = \beta_1 + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip}$. Note that

$$\log(\xi_i) = \beta_1^* + \beta_2\, x_{i2} + \cdots + \beta_p\, x_{ip}, \quad i = 1, ..., n,$$

where $\beta_1^* = \beta_1 + \log(\sqrt{\pi}/2)$ corresponds to the 'adjusted' intercept. Therefore, we can easily obtain the Maxwell modal regression model from the mean-parameterized Maxwell regression model.

Parameter estimation. Let $\boldsymbol{y} = (y_1, ..., y_n)^\top$ be the n-vector of the observed responses. We have that the parameter vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ represents the effects of the covariates on the mean parameter of the Maxwell regression model and, hence, we are interested in estimating this regression parameter vector. To do so, we shall consider the traditional maximum likelihood (ML) method. The log-likelihood function for this class of regression models, except for an unimportant constant term, has the form

$$\ell(\boldsymbol{\beta}) = -3\sum_{i=1}^{n} \log(\mu_i) - \frac{4}{\pi}\sum_{i=1}^{n} \frac{y_i^2}{\mu_i^2},$$

where $\mu_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ for $i = 1, ..., n$. The ML estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^\top$ of $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$ is obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. The maximization can be performed, for example, in the R software [28] by using the `optim(...)` function. The score function, obtained by differentiating the log-likelihood function $\ell(\boldsymbol{\beta})$ with respect to the unknown parameters, is given by the p-vector $\boldsymbol{U}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{s}$, where $\boldsymbol{s} = (s_1, ..., s_n)^\top$ with $s_i = 8y_i^2/(\pi\mu_i^2) - 3$. After some algebra, the expected (Fisher) information matrix for $\boldsymbol{\beta}$ takes the form $\boldsymbol{K} = 6\boldsymbol{X}^\top \boldsymbol{X}$.

The ML estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^\top$ can also be obtained by solving the nonlinear system of equations $\boldsymbol{U}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{0}_p$, where $\boldsymbol{0}_p$ denotes a p-dimensional vector of zeros. There is no closed-form expression for the ML estimate $\widehat{\boldsymbol{\beta}}$ and its computation has to be performed numerically using a nonlinear optimization algorithm. For example, the Newton–Raphson iterative technique (or the Gauss–Newton and Quasi-Newton methods) could be applied to

solve these equations and obtain $\widehat{\boldsymbol{\beta}}$ numerically. On the other hand, one can use the Fisher scoring method to estimate $\boldsymbol{\beta}$ by iteratively solving the equation

$$\boldsymbol{\beta}^{(m+1)} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{z}^{(m)}, \tag{2.2}$$

where $\boldsymbol{z} = (z_1, ..., z_n)^\top = \boldsymbol{X}\boldsymbol{\beta} + (1/6)\boldsymbol{s}$ acts as an adjusted dependent variable, and $m = 0, 1, ...$ is the iteration counter. The cycles through the scheme (2.2) consists of an iterative ordinary least squares algorithm to optimize the log-likelihood function, and the iterations go on until convergence is achieved (a stopping criterion must be defined). Equation (2.2) reveals that the calculation of the ML estimate $\widehat{\boldsymbol{\beta}}$ can be carried out using any software with a matrix algebra library as, for example, the R software. The optimization algorithms require the specification of initial values to be used in the iterative scheme.

In the following, we make some assumptions on the behavior of $\ell(\boldsymbol{\beta})$ as the sample size n approaches infinity, such as the regularity of the first three derivatives of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, and the existence and uniqueness of the ML estimate of $\boldsymbol{\beta}$; see, for example, Cox and Hinkley [9]. When n is large and under standard regularity conditions, the ML estimators of the Maxwell regression parameters are asymptotically normal, asymptotically unbiased and have asymptotic variance-covariance matrix given by the inverse of the expected Fisher information matrix: $\widehat{\boldsymbol{\beta}} \overset{a}{\sim} N_p(\boldsymbol{\beta}, \boldsymbol{K}^{-1})$. This asymptotic normal distribution can be used to construct approximate confidence intervals for the Maxwell regression parameters. Let β_r $(r = 1, ..., p)$ be r-th component of $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$. The asymptotic confidence interval for β_r is simply given by $\widehat{\beta}_r \pm \Phi^{-1}(1 - \vartheta/2)\,\mathrm{se}(\widehat{\beta}_r)$, for $r = 1, ..., p$, with asymptotic coverage of $100(1 - \vartheta)\%$. Here, $\mathrm{se}(\cdot)$ is the square root of the diagonal element of $\boldsymbol{K}(\widehat{\boldsymbol{\beta}})^{-1}$ corresponding to each parameter (i.e., the asymptotic standard error), and $\Phi^{-1}(\cdot)$ is the standard normal quantile function.

Finite sample bias of the ML estimator. It is well-known that ML estimators are asymptotically unbiased and efficient, but for small samples, the ML estimators may not be unbiased. Here, we shall provide a general closed-form expression for the second-order biases of the ML estimators of the Maxwell regression parameters. To that end, we shall use the general expression given by Cox and Snell [10, Eq. (20)]. The closed-form expression will, in turn, allow us to obtain bias-corrected estimates of the unknown parameters. We shall use the following notation: $\kappa_{rs} = \mathbb{E}(\partial^2 \ell(\boldsymbol{\beta})/\partial\beta_r \partial\beta_s)$, $\kappa_{rst} = \mathbb{E}(\partial^3 \ell(\boldsymbol{\beta})/\partial\beta_r \partial\beta_s \partial\beta_t)$ and $\kappa_{rs}^{(t)} = \partial\kappa_{rs}/\partial\beta_t$, for $r, s, t = 1, ..., p$. After some algebra, we obtain

$$\kappa_{rs} = -6\sum_{i=1}^{n} x_{ir}x_{is}, \quad \kappa_{rst} = 12\sum_{i=1}^{n} x_{ir}x_{is}x_{it}, \quad \text{and} \quad \kappa_{rs}^{(t)} = 0.$$

Let B_a denote the second-order bias of $\widehat{\beta}_a$ $(a = 1, ..., p)$. From Cox and Snell [10], we can express B_a in the form

$$B_a = \sum_{s,t,u}{}' \kappa^{a,s}\kappa^{t,u}\left(\kappa_{st}^{(u)} - \frac{1}{2}\kappa_{stu}\right),$$

where $\kappa^{r,s}$ is the (r, s)-th element of $\boldsymbol{K}^{-1}$, and $\sum'$ denotes the summation over all combinations of parameters $\beta_1, ..., \beta_p$. Plugging the cumulants given before into this expression, we can obtain the bias of $\widehat{\boldsymbol{\beta}}$, say $\boldsymbol{B}$, in matrix form. We can show after some algebra that the $p \times 1$ bias vector $\boldsymbol{B}$ reduces to

$$\boldsymbol{B} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{\delta}, \tag{2.3}$$

where $\boldsymbol{\delta}$ is the n-vector containing the elements of the main diagonal of the matrix $-(6\pi)^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$. Note that the second-order bias vector $\boldsymbol{B}$ is simply the set coefficients from a simple ordinary least squares regression of $\boldsymbol{\delta}$ on the columns of the model matrix $\boldsymbol{X}$. As expression (2.3) makes clear, it is possible to express the bias vector of $\widehat{\boldsymbol{\beta}}$ as the solution of an ordinary least squares regression. Additionally, the bias vector $\boldsymbol{B}$ involves simple operations on matrices and vectors, and we can calculate it numerically via software with numerical linear algebra facilities such as R with minimal effort. It is worth emphasizing that the bias vector $\boldsymbol{B}$ will be small when $\boldsymbol{\delta}$ is orthogonal to the columns of $\boldsymbol{X}$. However, the second-order bias vector $\boldsymbol{B}$ may be large in small and moderate sized samples. From (2.3), we define the bias-corrected ML estimate $\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{B}$. We say that $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, ..., \widetilde{\beta}_p)^\top$ is bias-adjusted ML estimate to order n^{-1}, since its bias is of order n^{-2}. It is expected that $\widetilde{\boldsymbol{\beta}}$ has superior finite-sample behavior relative to $\widehat{\boldsymbol{\beta}}$, whose bias is of order n^{-1}. It is not difficult to show that $\widetilde{\boldsymbol{\beta}} \overset{a}{\sim} N_p(\boldsymbol{\beta}, \boldsymbol{K}^{-1})$.

Simulation study. In what follows, we report Monte Carlo simulation experiments for the mean-parameterized Maxwell regression model. To explore the performance of the ML method in estimating the regression parameter vector $\boldsymbol{\beta}$, we report the results of simulations designed to evaluate the accuracy of the ML estimators of $\boldsymbol{\beta}$. The bias-adjusted ML estimate is also considered in the Monte Carlo simulations. The Monte Carlo experiments were carried out using $\log(\mu_i) = \beta_1\, x_{i1} + \beta_2\, x_{i2} + \beta_3\, x_{i3}$, where $x_{i1} = 1$ $(i = 1, ..., n)$, and $n = 10, 20, 30, 50, 80$ and 150. The true values of the regression parameters were taken as $\beta_1 = 1.0$, $\beta_2 = 0.5$ and $\beta_3 = 1.5$. The values of x_{i2} were obtained as random draws of the standard normal distribution, and the values of x_{i3} were obtained as random draws of the exponential distribution with mean equals 1. The covariate values were held constant throughout the simulations. We evaluate the point estimates by considering the following quantities: the mean, the relative bias[2] (RB), and the mean square error (MSE). These quantities are computed from 15,000 Monte Carlo replications. The numerical results are presented in Table 1. Note that the performance of the ML estimator of $\boldsymbol{\beta}$ is good, exhibiting small bias in all cases considered. It is noteworthy that the bias-adjusted estimator is better than the usual ML estimator for estimating the Maxwell regression parameters, mainly in very small sample sizes. However, for large sample sizes, the bias-corrected ML estimator becomes less justifiable. As expected, the MSE decreases as the sample size increases. In short, the numerical results reveal that the ML method can be used quite effectively to estimate the Maxwell regression parameters, and the bias-corrected ML estimator becomes a good alternative when the sample size is very small.

We now consider a Monte Carlo simulation study in the following way. First, we simulate data from the mean-parameterized Maxwell regression model and analyse the simulated data using the following models: mean-parameterized Maxwell, gamma, and inverse-Gaussian regression models. Next, we simulate data from a gamma model and analyse the simulated data using all three models (mean-parameterized Maxwell, gamma, and inverse-Gaussian regression models). Finally, we simulate data from an inverse Gaussian model and analyse the simulated data using all three models (mean-parameterized Maxwell, gamma, and inverse-Gaussian regression models). The gamma and inverse Gaussian regression models are very useful models for continuous positive response variables [see, for example, 25]. The Monte Carlo experiments were carried out using $\log(\mu_i) = \beta_1 + \beta_2\, x_i$, for $i = 1, ..., n$, and $n = 50, 90$ and 150.

[2]The relative bias of an estimate $\widehat{\theta}$, defined as $[\mathbb{E}(\widehat{\theta}) - \theta]/\theta$, is obtained by estimating $\mathbb{E}(\widehat{\theta})$ by Monte Carlo.

Table 1: Simulation results regarding the point estimates of the Maxwell model parameters.

		ML estimator			Bias-corrected ML estimator		
		$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widetilde{\beta}_1$	$\widetilde{\beta}_2$	$\widetilde{\beta}_3$
$n = 10$	Mean	0.913	0.550	1.516	0.939	0.536	1.512
	RB	−0.087	0.101	0.010	−0.061	0.072	0.008
	MSE	0.297	0.533	0.665	0.293	0.546	0.661
$n = 20$	Mean	0.971	0.510	1.499	0.981	0.506	1.498
	RB	−0.029	0.019	0.000	−0.019	0.012	−0.002
	MSE	0.245	0.500	0.517	0.244	0.503	0.515
$n = 30$	Mean	0.988	0.495	1.496	0.991	0.497	1.498
	RB	−0.012	−0.010	−0.003	−0.009	−0.006	−0.002
	MSE	0.199	0.489	0.524	0.199	0.487	0.526
$n = 50$	Mean	0.991	0.494	1.502	0.994	0.495	1.503
	RB	−0.009	−0.013	0.002	−0.006	−0.009	0.002
	MSE	0.187	0.470	0.455	0.187	0.468	0.455
$n = 80$	Mean	0.994	0.500	1.500	0.996	0.500	1.500
	RB	−0.006	0.001	0.000	−0.004	0.000	0.000
	MSE	0.182	0.439	0.444	0.182	0.439	0.444
$n = 150$	Mean	0.996	0.501	1.500	0.997	0.501	1.500
	RB	−0.004	0.002	0.000	−0.003	0.002	0.000
	MSE	0.175	0.430	0.432	0.175	0.430	0.432

The values of covariate x_i were obtained as random draws of the uniform distribution on the unit interval $(0, 1)$, and the covariate values were held constant throughout the simulations. We set $\beta_1 = 1.0$ and $\beta_2 = 0.8$. For the gamma and inverse Gaussian models, we consider the precision parameter, say ϕ, equals to $\phi = 4$ and $\phi = 5$, respectively. Tables 2 and 3 list the simulation results based on 10,000 Monte Carlo replications for the true data generating process (DGP) under three different scenarios: the Maxwell model as the true DGP, the gamma model as the true DGP, and the inverse Gaussian model as the true DGP. In Table 2 we present the point estimates, standard deviation (SD) between parentheses, and the values of Akaike information criterion (AIC) and Bayesian information criterion (BIC), whereas in Table 3 we present the coverage probability (CP) of the confidence intervals for β_1 and β_2 at the nominal levels 90% and 95%.

From Table 2, as expected, note that the Maxwell model yields the best fit under the Maxwell DGP, as well as the gamma and inverse models when these models correspond to the true DGPs; see the AIC and BIC values for the fitted models. It is also interesting to note that under the gamma DGP, the Maxwell model outperforms the inverse Gaussian model based on the AIC and BIC values. It is worth mentioning that under the inverse Gaussian DGP, the SDs of the ML estimates of the model parameters become larger than in the other two DGPs (Maxwell and gamma models). On the other hand, the ML estimates are close to the true values of the regression parameters, which indicates the 'robustness' of each model when estimating the regression parameters under model misspecification. From the numerical results in Table 3, we have that under the Maxwell DGP, the coverage rates of the confidence intervals are close to the nominal significance levels for all regression models, being the Maxwell regression model with the best performance, as expected. However, it

is noteworthy that the coverage rates of the confidence intervals of the Maxwell regression parameters under the gamma and inverse Gaussian DGPs are not near the nominal levels, mainly under the gamma DGP. Finally, it should be mentioned that much more numerical work is needed to come to any general conclusion about the 'robustness' of the Maxwell regression model under model misspecification and, hence, future research regarding this issue can be conducted in a separate paper elsewhere.

Table 2: Simulation results considering three different data generating process.

n	Model	Maxwell DGP			
		β_1	β_2	AIC	BIC
50	Maxwell	1.005(0.098)	0.780(0.179)	233.618	237.807
	Gamma	1.007(0.101)	0.780(0.191)	236.195	244.478
	Inverse Gaussian	1.005(0.104)	0.783(0.198)	245.476	253.759
90	Maxwell	1.002(0.083)	0.785(0.151)	349.574	354.574
	Gamma	1.004(0.086)	0.783(0.160)	352.967	362.466
	Inverse Gaussian	1.003(0.090)	0.785(0.169)	367.312	376.812
150	Maxwell	1.000(0.057)	0.790(0.108)	582.682	588.703
	Gamma	1.002(0.059)	0.788(0.113)	587.642	598.674
	Inverse Gaussian	1.001(0.062)	0.790(0.120)	611.891	622.923

n	Model	Gamma DGP			
		β_1	β_2	AIC	BIC
50	Maxwell	1.033(0.145)	0.770(0.238)	249.305	253.494
	Gamma	1.008(0.137)	0.767(0.227)	247.040	255.324
	Inverse Gaussian	1.003(0.141)	0.778(0.239)	252.414	260.697
90	Maxwell	1.038(0.112)	0.777(0.183)	375.505	380.505
	Gamma	1.012(0.104)	0.773(0.169)	371.079	380.578
	Inverse Gaussian	1.010(0.105)	0.778(0.172)	379.219	388.719
150	Maxwell	1.029(0.092)	0.799(0.148)	625.968	631.989
	Gamma	1.001(0.086)	0.796(0.138)	617.266	628.297
	Inverse Gaussian	1.001(0.085)	0.796(0.139)	631.973	643.005

n	Model	Inverse Gaussian DGP			
		β_1	β_2	AIC	BIC
50	Maxwell	1.121(0.225)	0.938(0.483)	351.457	355.646
	Gamma	1.006(0.172)	0.778(0.358)	281.375	289.658
	Inverse Gaussian	1.002(0.176)	0.788(0.368)	272.734	281.017
90	Maxwell	1.128(0.203)	0.905(0.445)	525.594	530.594
	Gamma	1.009(0.168)	0.748(0.350)	417.941	427.441
	Inverse Gaussian	1.010(0.161)	0.744(0.332)	405.532	415.032
150	Maxwell	1.133(0.164)	0.950(0.362)	891.968	897.989
	Gamma	1.012(0.129)	0.784(0.269)	702.632	713.664
	Inverse Gaussian	1.009(0.127)	0.789(0.265)	681.196	692.228

Table 3: Coverage rates of confidence intervals considering three different data generating process.

n	Model	Maxwell DGP			
		CP(90%)		CP(95%)	
		β_1	β_2	β_1	β_2
50	Maxwell	0.892	0.899	0.955	0.941
	Gamma	0.890	0.900	0.941	0.945
	Inverse Gaussian	0.855	0.903	0.917	0.941
90	Maxwell	0.917	0.897	0.952	0.948
	Gamma	0.900	0.862	0.941	0.945
	Inverse Gaussian	0.848	0.879	0.921	0.945
150	Maxwell	0.893	0.896	0.952	0.941
	Gamma	0.897	0.876	0.962	0.945
	Inverse Gaussian	0.855	0.866	0.928	0.928

n	Model	Gamma DGP			
		CP(90%)		CP(95%)	
		β_1	β_2	β_1	β_2
50	Maxwell	0.776	0.800	0.841	0.883
	Gamma	0.883	0.893	0.934	0.948
	Inverse Gaussian	0.824	0.869	0.893	0.931
90	Maxwell	0.759	0.814	0.841	0.879
	Gamma	0.890	0.890	0.938	0.948
	Inverse Gaussian	0.828	0.900	0.917	0.945
150	Maxwell	0.759	0.790	0.828	0.855
	Gamma	0.879	0.883	0.934	0.945
	Inverse Gaussian	0.852	0.883	0.900	0.934

n	Model	Inverse Gaussian DGP			
		CP(90%)		CP(95%)	
		β_1	β_2	β_1	β_2
50	Maxwell	0.886	0.828	0.955	0.897
	Gamma	0.941	0.921	0.969	0.948
	Inverse Gaussian	0.907	0.903	0.948	0.955
90	Maxwell	0.872	0.790	0.938	0.886
	Gamma	0.948	0.900	0.983	0.972
	Inverse Gaussian	0.910	0.897	0.966	0.962
150	Maxwell	0.834	0.731	0.872	0.831
	Gamma	0.914	0.872	0.962	0.934
	Inverse Gaussian	0.879	0.872	0.928	0.941

3. DIAGNOSTIC MEASURES

It is well-known that regression models are sensitive to the underlying model assumptions and hence a sensitivity analysis is strongly advisable after fitting regression models to a dataset. In order to assess the sensitivity of the ML estimates of the mean-parameterized Maxwell model parameters in the presence of atypical observations, we shall consider the global and local influence methods [7, 8]. Additionally, the normalized quantile residual will be considered to assess departures from the underlying distribution.

Global influence. A first way to perform sensitivity analysis is by means of global influence starting from the case deletion proposed by Cook [7], which is a common approach to study the effect of dropping the i-th case from the dataset. Let $\widehat{\boldsymbol{\beta}}_{(-i)}$ be the ML estimate of $\boldsymbol{\beta}$ without the i-th observation in the sample. To assess the influence of the i-th case on the ML estimate $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_1, ..., \widehat{\beta}_p)^\top$, the basic idea is to compare the difference between $\widehat{\boldsymbol{\beta}}_{(-i)}$ and $\widehat{\boldsymbol{\beta}}$. If the deletion of an observation seriously influences an estimate, more attention should be paid to that particular observation. Hence, if $\widehat{\boldsymbol{\beta}}_{(-i)}$ is far from $\widehat{\boldsymbol{\beta}}$, then this case is regarded as an influential observation. To measure the global influence, the generalized Cook distance is defined as the standardized norm of $\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}}$ in the form $GD_i = (\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}})^\top \boldsymbol{J}_n(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}})$ for $i = 1, ..., n$, where $\boldsymbol{J}_n(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X}$ is the observed (Fisher) information matrix, and $\boldsymbol{W} = \text{diag}\{w_1, ..., w_n\}$ with $w_i = 16y_i^2/(\pi\mu_i^2)$. Note that we have to compute $\widehat{\boldsymbol{\beta}}_{(-i)}$ for all $i = 1, ..., n$. To avoid employing the direct model estimation for all observations, we can use the following one-step approximation to reduce the number of models to be fitted: $\widehat{\boldsymbol{\beta}}_{(-i)} \doteq \widehat{\boldsymbol{\beta}} - \boldsymbol{J}_n(\widehat{\boldsymbol{\beta}})^{-1}\dot{\boldsymbol{L}}_i(\widehat{\boldsymbol{\beta}})$, where $\dot{\boldsymbol{L}}_i(\boldsymbol{\beta}) = \partial\ell_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}$, and $\ell_i(\boldsymbol{\beta}) = -3\log(\mu_i) - 4y_i^2/(\pi\mu_i^2)$. It follows that $\widehat{\boldsymbol{\beta}}_{(-i)} - \widehat{\boldsymbol{\beta}} \doteq -\boldsymbol{J}_n(\widehat{\boldsymbol{\beta}})^{-1}\boldsymbol{x}_i\,\widehat{s}_i$, where $\widehat{s}_i := s_i(\widehat{\boldsymbol{\beta}})$. Hence, the generalized Cook distance reduces to $GD_i = \widehat{s}_i^2\,\boldsymbol{x}_i^\top(\boldsymbol{X}^\top\widehat{\boldsymbol{W}}\boldsymbol{X})^{-1}\boldsymbol{x}_i$, for $i = 1, ..., n$, where $\widehat{\boldsymbol{W}} := \boldsymbol{W}(\widehat{\boldsymbol{\beta}})$. The index plot of GD_i may reveal those influential observations on the ML estimates of the Maxwell regression parameters.

Local influence. In the following, the local influence method under a specific perturbation scheme (case weighting perturbation) is carried out in order to assess the sensitivity of the ML estimates of the Maxwell regression parameters. Let $\boldsymbol{\omega} \in \boldsymbol{\Omega}$ be a k-dimensional vector of perturbations, where $\boldsymbol{\Omega} \subset \mathbb{R}^k$ is an open set. The perturbed log-likelihood function is denoted by $\ell(\boldsymbol{\beta}|\boldsymbol{\omega})$. The vector of no perturbation is $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ such that $\ell(\boldsymbol{\beta}|\boldsymbol{\omega}_0) = \ell(\boldsymbol{\beta})$. The Cook's idea for assessing local influence is essentially analyzing the local behavior of the log-likelihood displacement $LD_{\boldsymbol{\omega}} = 2[\ell(\widehat{\boldsymbol{\beta}}) - \ell(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}})]$, where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\omega}}$ denotes the ML estimate under $\ell(\boldsymbol{\beta}|\boldsymbol{\omega})$, around $\boldsymbol{\omega}_0$ by evaluating the curvature of the plot of $LD_{\boldsymbol{\omega}_0 + a\boldsymbol{d}}$ against a, where $a \in \mathbb{R}$ and $\boldsymbol{d}$ is a unit norm direction. One of the measures of particular interest is the direction $\boldsymbol{d}_{\max}$ corresponding to the largest curvature $C_{\boldsymbol{d}_{\max}}$. Cook [8] proved that the normal curvature at the direction $\boldsymbol{d}$ is given by $C_{\boldsymbol{d}}(\boldsymbol{\beta}) = 2|\boldsymbol{d}^\top\boldsymbol{\Delta}^\top\boldsymbol{J}_n(\boldsymbol{\beta})^{-1}\boldsymbol{\Delta}\boldsymbol{d}|$, where $\boldsymbol{\Delta} = \partial^2\ell(\boldsymbol{\beta}|\boldsymbol{\omega})/\partial\boldsymbol{\beta}\partial\boldsymbol{\omega}^\top$ and $\boldsymbol{J}_n(\boldsymbol{\beta})$ are evaluated at $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\omega}_0$. We have that $\boldsymbol{J}_n(\boldsymbol{\beta}) = \boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}$ and, after some algebra, we can show that $\boldsymbol{\Delta} = \boldsymbol{X}^\top\boldsymbol{S}$, where $\boldsymbol{S} = \text{diag}\{s_1, ..., s_n\}$. Let $(1/2)C_{\boldsymbol{d}_{\max}}$ be the largest eigenvalue of $\boldsymbol{L} = -\boldsymbol{\Delta}^\top\boldsymbol{J}_n(\boldsymbol{\beta})^{-1}\boldsymbol{\Delta}$, and $\boldsymbol{d}_{\max}$ be the corresponding unit norm eigenvector ($||\boldsymbol{d}_{\max}|| = 1$). The index plot of the largest eigenvector ($\boldsymbol{d}_{\max}$) of $\boldsymbol{L}$ may reveal those influential observations on the ML estimate $\widehat{\boldsymbol{\beta}}$.

Residuals. Usually, the residuals are defined in order to study departures from the response distribution assumptions. More precisely, the residuals carry important information concerning the appropriateness of assumptions that underlie statistical models, and thereby play an important role in checking model adequacy. The use of residuals for assessing the adequacy of fitted regression models is nowadays commonplace due to the widespread availability of statistical software, many of which are capable of displaying residuals and diagnostic plots, at least for the more commonly used models. We shall consider the normalized quantile residuals proposed in Dunn and Smyth [14] to check the adequacy of the Maxwell regression model fitted to a dataset, which is simply defined as

$$R_i = \Phi^{-1}\left(\frac{2\gamma(3/2, 4y_i^2/(\pi\widehat{\mu}_i^2))}{\sqrt{\pi}}\right), \quad i = 1, ..., n, \tag{3.1}$$

where $\widehat{\mu}_i = \exp(\boldsymbol{x}_i^\top \widehat{\boldsymbol{\beta}})$. The normalized quantile residuals in (3.1) have a standard normal distribution asymptotically [14, 16]. Since the exact distribution of the above residual is not known, it is usual to add envelopes as suggested by Atkinson [2, § 4.2] into the normal quantile-quantile plot (QQ-plot) for R_i to decide whether the observed residuals are consistent with the fitted regression model. Thus, observations corresponding to absolute residuals outside the limits provided by the simulated envelope are worthy of further investigation. Additionally, if a considerable proportion of points falls outside the envelope, then one has evidence against the adequacy of the fitted model.

Remark 3. The simple closed-form expression for the bias vector of the ML estimators of the Maxwell regression parameters in (2.3) can be used to define improved Pearson residuals [see, for example, 10] for the mean-parameterized Maxwell regression model. Hence, future research can be done to compare through Monte Carlo simulations the improved Pearson residuals and the normalized quantile residuals.

4. ILLUSTRATIVE EXAMPLES

In what follows, we shall consider real data examples to illustrate the Maxwell regression model in practice. All computations regarding the mean-parameterized Maxwell regression model were carried out using the R program. The R code to compute the ML estimates of the mean-parameterized Maxwell regression model parameters is provided in the Appendix.

Life of metal pieces data. Here, we consider the biaxial fatigue data on the life (in cycles to failure) of metal pieces reported by Rieck and Nedelman [29]. The response variable (Y) is the life (in number of cycles to failure) of $n = 46$ metal pieces, and the explanatory variable (x) is the work per cycle ($\mathrm{mJ/m^3}$). We assume that $Y_i \sim \mathrm{Mw}(\mu_i)$, for $i = 1, ..., 46$, where $\log(\mu_i) = \beta_1 + \beta_2 \log(x_i)$. The ML estimates, asymptotic standard errors (SE) and the 95% asymptotic confidence intervals (CI) of the Maxwell regression parameters are listed in Table 4. Figure 2 displays the normalized quantile residuals for the Maxwell regression model. We have in this figure the quantile residuals against the index, and the normal QQ-plot (with generated envelopes), respectively. Note that the residuals appear satisfactory (random) and, more important, there is no observation falling outside the envelope. Therefore, the mean-parameterized Maxwell regression model provides a good fit to the biaxial fatigue data. Figure 2 shows the index plot of the generalized Cook distance, as well as the index plot of $|\boldsymbol{d}_{\max}|$. The generalized Cook distance identifies the cases #4 and #46 as possible influential observations on the ML estimates of the Maxwell regression parameters. We remove each of these observations individually from the dataset and, after that, we fit the mean-parameterized Maxwell regression model. We observe that there is no inferential change regarding the regression parameters when removing the cases #4 and #46 from the dataset and, hence, these observations have no influence on the ML estimates of the Maxwell regression parameters. The estimated Maxwell regression model is

$$\log(\widehat{\mu}_i) = 12.4733 - 1.706 \log(x_i), \quad i = 1, ..., 46.$$

The coefficients of the mean-parameterized Maxwell regression model can be interpreted as follows. The expected life (in cycles to failure) of a metal piece should decrease approximately 81.84% $[(1 - \mathrm{e}^{-1.7060}) \times 100\%]$ as the logarithm of work per cycle increases one unity;

that is, there is a decrease in the expected rate of life (in cycles to failure) by a factor of (approximately) 0.1816 [$\exp(-1.7060) = 0.1816$].

Advertising media data. Next, we shall consider data corresponding to the impact of newspapers on sales. These data are the advertising budget (in thousands of dollars) along with sales. The advertising experiment has $n = 200$ observations, and they are available in the R package `datarium` [21]. The response variable (Y) corresponds to the sales (in thousands of dollars), while the covariate (x) corresponds to the advertising budget on newspapers (in thousands of dollars). We assume that $Y_i \sim \text{Mw}(\mu_i)$, for $i = 1, ..., 200$, where $\log(\mu_i) = \beta_1 + \beta_2\, x_i$. The mean-parameterized Maxwell regression estimates are provided in Table 4. In addition, Figure 3 confirms that the Maxwell regression model is suitable to model the data, since there are no observations falling outside the envelope. The index plots of GD_i (generalized Cook distance) and $|\boldsymbol{d}_{\max}|$ (local influence) are presented in Figure 3. It is identified the cases #37 and #129 as possible influential observations on the ML estimates of the mean-parameterized Maxwell regression parameters. We remove each of these observations individually from the dataset and, after that, we fit the Maxwell regression model. There is no inferential change regarding the regression parameters when removing these cases from the dataset, so these observations have no influence on the ML estimates of the Maxwell regression parameters. The estimated Maxwell regression model is

$$\log(\widehat{\mu}_i) = 2.6879 + 0.003\, x_i, \quad i = 1, ..., 200,$$

and the ML estimates of the mean-parameterized Maxwell regression parameters deliver the following interpretation. The expected sale (in thousands of dollars) should increase (approximately) 0.301% [$(\text{e}^{0.003} - 1) \times 100\%$] as the advertising budget on newspapers increases one thousand dollars; that is, there is an increase in the expected sale by a factor of (approximately) 1.003 [$\exp(0.003) = 1.003$].

Radioimmunoassay data. Now, we consider the radioimmunoassay data, reported in Tiede and Pagano [30]. These data were obtained from the Nuclear Medicine Department of the Veteran's Administration Hospital, Buffalo, New York. The variable of interest (Y) is the radioactivity count rate, and the covariate (x) corresponds to the dose concentration (measured in micro-international units per milliliter). We assume that $Y_i \sim \text{Mw}(\mu_i)$, for $i = 1, ..., 14$, where $\log(\mu_i) = \beta_1 + \beta_2\, x_i$. Table 4 lists the ML estimates, asymptotic SEs and the 95% asymptotic CIs of the Maxwell regression parameters. Residuals plots are displayed in Figure 4, which confirms that the mean-parameterized Maxwell regression model is suitable to model the data, since there are no observations falling outside the envelope. Figure 4 displays the index plot of the generalized Cook distance, as well as the index plot of $|\boldsymbol{d}_{\max}|$. It is identified the cases #1, #2 and #14 as possible influential observations on the ML estimates of the Maxwell regression parameters. We remove each of these observations individually from the dataset and, after that, we fit the mean-parameterized Maxwell regression model. It is noteworthy that there is no inferential change regarding the regression parameters when removing the cases #1, #2 and #14 from the dataset, revealing that these observations have no influence on the ML estimates of the Maxwell regression parameters. The estimated Maxwell regression model is

$$\log(\widehat{\mu}_i) = 8.6091 - 0.0190\, x_i, \quad i = 1, ..., 14.$$

The mean-parameterized Maxwell parameter estimates deliver interesting interpretation. The expected radioactivity count rate should decrease approximately 1.88% $[(1-\mathrm{e}^{-0.0190})\times 100\%]$ as the dose concentration increases one unity; that is, there is a decrease in the expected radioactivity count rate by a factor of (approximately) 0.98 $[\exp(-0.0190)=0.98]$.

Table 4: Parameter estimates.

Parameter	Life of metal pieces data		
	Estimate	SE	95% CI
β_1	12.4733	0.4007	(11.688; 13.259)
β_2	−1.7060	0.1114	(−1.924; −1.488)

Parameter	Advertising media data		
	Estimate	SE	95% CI
β_1	2.6879	0.0498	(2.590; 2.785)
β_2	0.0030	0.0011	(0.001; 0.005)

Parameter	Radioimmunoassay data		
	Estimate	SE	95% CI
β_1	8.6091	0.1390	(8.337; 8.881)
β_2	−0.0190	0.0032	(−0.025; −0.013)

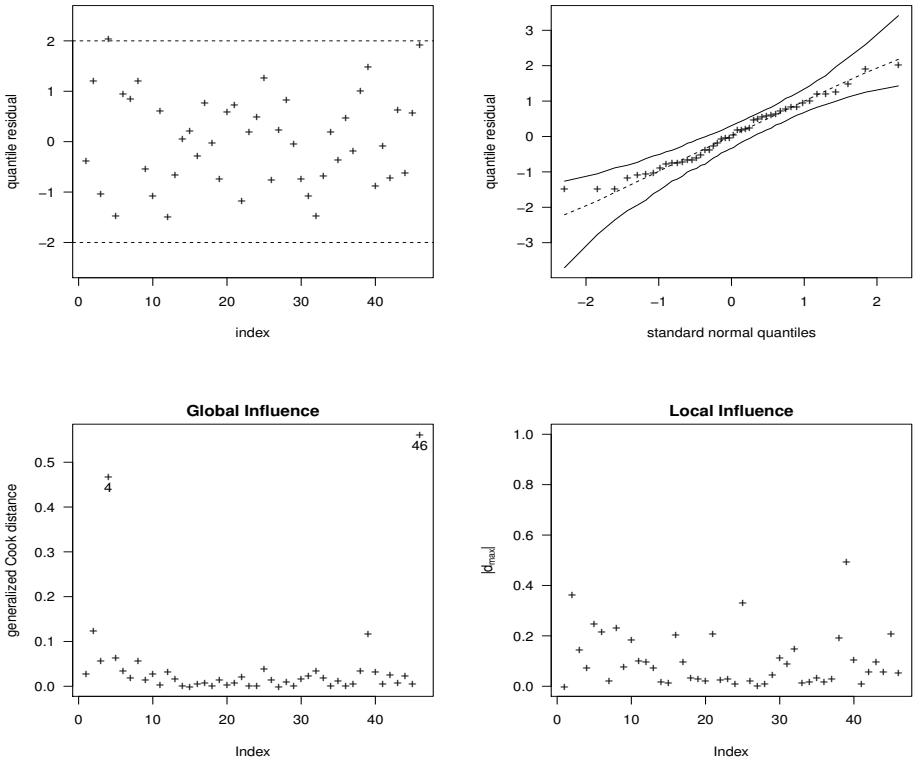


Figure 2: Residuals plots (top), and influence plots (bottom); life of metal pieces data.

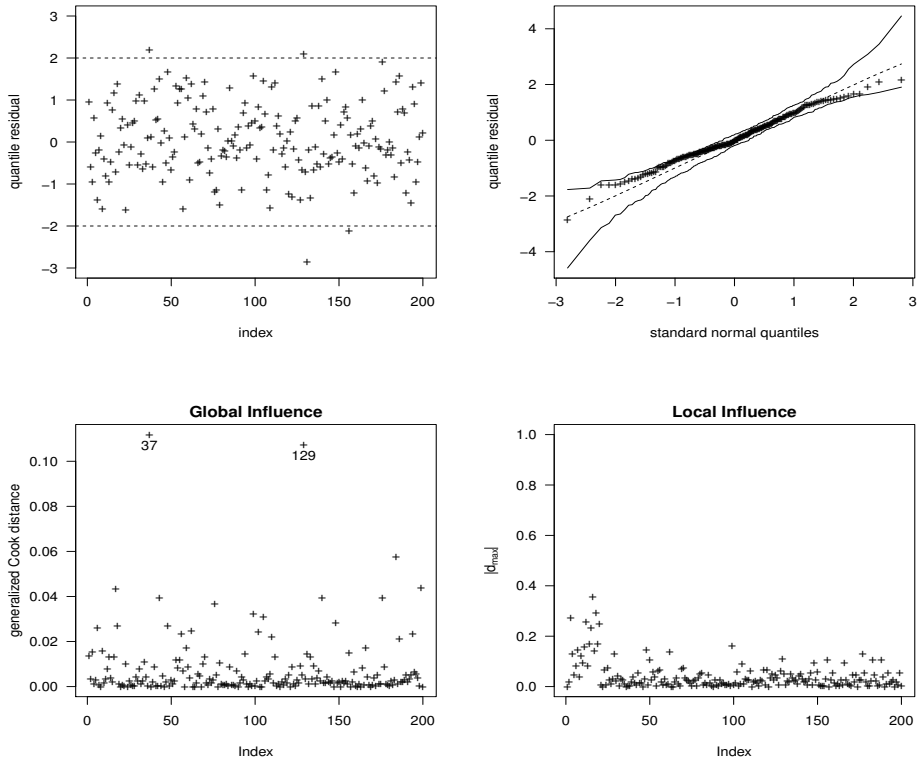


Figure 3: Residuals plots (top), and influence plots (bottom); advertising media data.

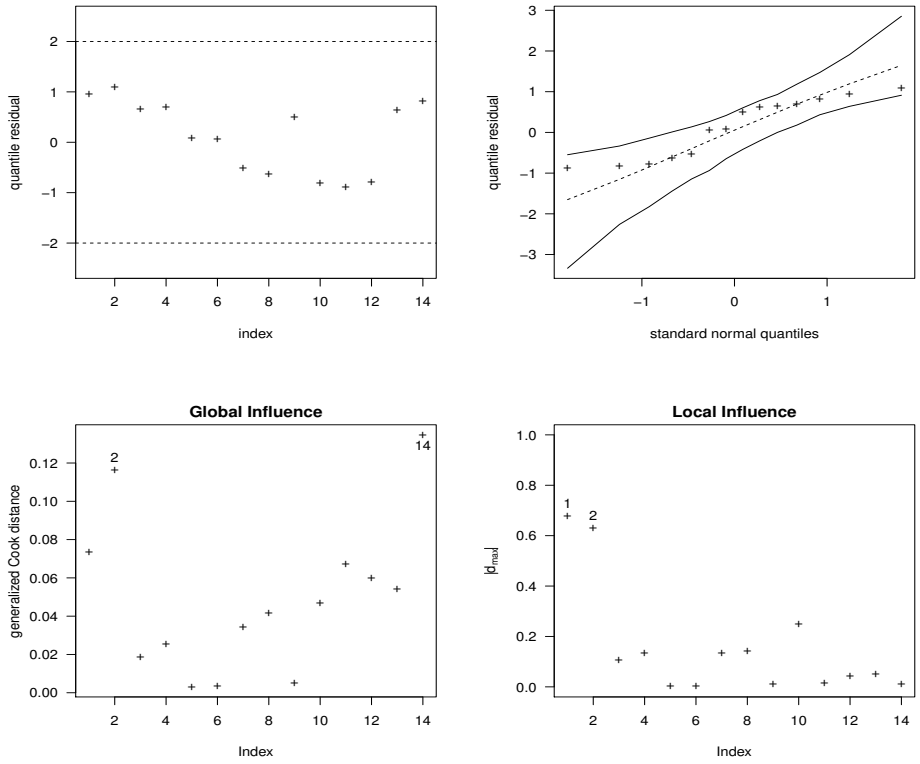


Figure 4: Residuals plots (top), and influence plots (bottom); radioimmunoassay data.

4.1. Competing models

Obviously, there are plenty of regression models in the statistic literature that can be used to model continuous positive response variables. Perhaps the most useful (and simple as well) regression models to deal with positive response variables are the gamma and inverse Gaussian generalized linear models [25]. Beyond the very simple form of these models, the gamma and inverse Gaussian regression models have been quite used in practice mainly because of the well-developed R function `glm()`. The gamma PDF is

$$f(y) = \frac{\phi^{\phi} y^{\phi-1}}{\Gamma(\phi)\mu^{\phi}} \exp\left(-\frac{\phi y}{\mu}\right), \quad y > 0,$$

where $\mu > 0$ is the mean, and $\phi > 0$ is the precision parameter. In the generalized linear model terminology, $\phi^{-1} > 0$ corresponds to the dispersion parameter. We shall use the notation $\mathrm{Ga}(\mu, \phi)$ to refer to this distribution. The gamma distribution reduces to the exponential distribution when $\phi = 1$. If $Y \sim \mathrm{Ga}(\mu, \phi)$, the variance is $\mathbb{VAR}(Y) = \phi^{-1}\mu^2 \propto \mu^2$. Remembering that the variance of the mean-parameterized Maxwell distribution is $0.178\mu^2 \propto \mu^2$ and, hence, the heteroscedastic form based on the gamma and Maxwell distributions are similar, thus modeling the variance of the response variable in a quadratic form. The inverse Gaussian PDF is

$$f(y) = \left(\frac{\phi}{2\pi y^3}\right)^{1/2} \exp\left(-\frac{\phi(y-\mu)^2}{2\mu^2 y}\right), \quad y > 0,$$

where $\mu > 0$ is the mean, and $\phi > 0$ is the precision parameter. We shall use the notation $\mathrm{IG}(\mu, \phi)$ to refer to this distribution. If $Y \sim \mathrm{IG}(\mu, \phi)$, the variance is $\mathbb{VAR}(Y) = \phi^{-1}\mu^3 \propto \mu^3$.

In the following, we fit the gamma and inverse Gaussian regression models to the data previously analyzed using the mean-parameterized Maxwell regression model. For each of the three datasets previously analyzed, we consider the same regression structures for the mean parameter of the gamma and inverse Gaussian models that were considered for the mean of the Maxwell model. The parameter estimates of the gamma and inverse Gaussian parameters are listed in Tables 5 and 6, respectively. Residuals plots are displayed in Figures 5 and 6 for the gamma and inverse Gaussian regression models, respectively. We consider the deviance residual for these regression models, which appear to be a very good choice in the generalized linear model framework [27]. Similar to the mean-parameterized Maxwell regression model, Figure 5 also reveals that the gamma regression model seems to be appropriate to fit these real datasets, once none observation is outside the envelope. On the other hand, the inverse Gaussian regression model appears not suitable to model the advertising media data (some observations are outside the envelope), but it appears suitable to model the other datasets (see Figure 6). At this moment, the natural question is which one is the best in modeling these datasets. The next section addresses this question.

Table 5: Parameter estimates; gamma regression.

Parameter	Life of metal pieces data: $\widehat{\phi}^{-1} = 0.1554$		
	Estimate	SE	95% CI
β_1	12.4449	0.3869	(11.686; 13.203)
β_2	−1.6945	0.1076	(−1.905; −1.484)

Parameter	Advertising media data: $\widehat{\phi}^{-1} = 0.1318$		
	Estimate	SE	95% CI
β_1	2.7047	0.0443	(2.618; 2.791)
β_2	0.0031	0.0010	(0.001; 0.005)

Parameter	Radioimmunoassay data: $\widehat{\phi}^{-1} = 0.0990$		
	Estimate	SE	95% CI
β_1	8.6514	0.1071	(8.441; 8.861)
β_2	−0.0191	0.0025	(−0.024; −0.014)

Table 6: Parameter estimates; inverse Gaussian regression.

Parameter	Life of metal pieces data: $\widehat{\phi}^{-1} = 0.00034$		
	Estimate	SE	95% CI
β_1	11.7484	0.5068	(10.755; 12.742)
β_2	−1.5103	0.1280	(−1.761; −1.260)

Parameter	Advertising media data: $\widehat{\phi}^{-1} = 0.0079$		
	Estimate	SE	95% CI
β_1	2.7057	0.0438	(2.620; 2.792)
β_2	0.0031	0.0010	(0.001; 0.005)

Parameter	Radioimmunoassay data: $\widehat{\phi}^{-1} = 3.15\text{e-}05$		
	Estimate	SE	95% CI
β_1	8.5721	0.1259	(8.325; 8.819)
β_2	−0.0170	0.0019	(−0.021; −0.013)

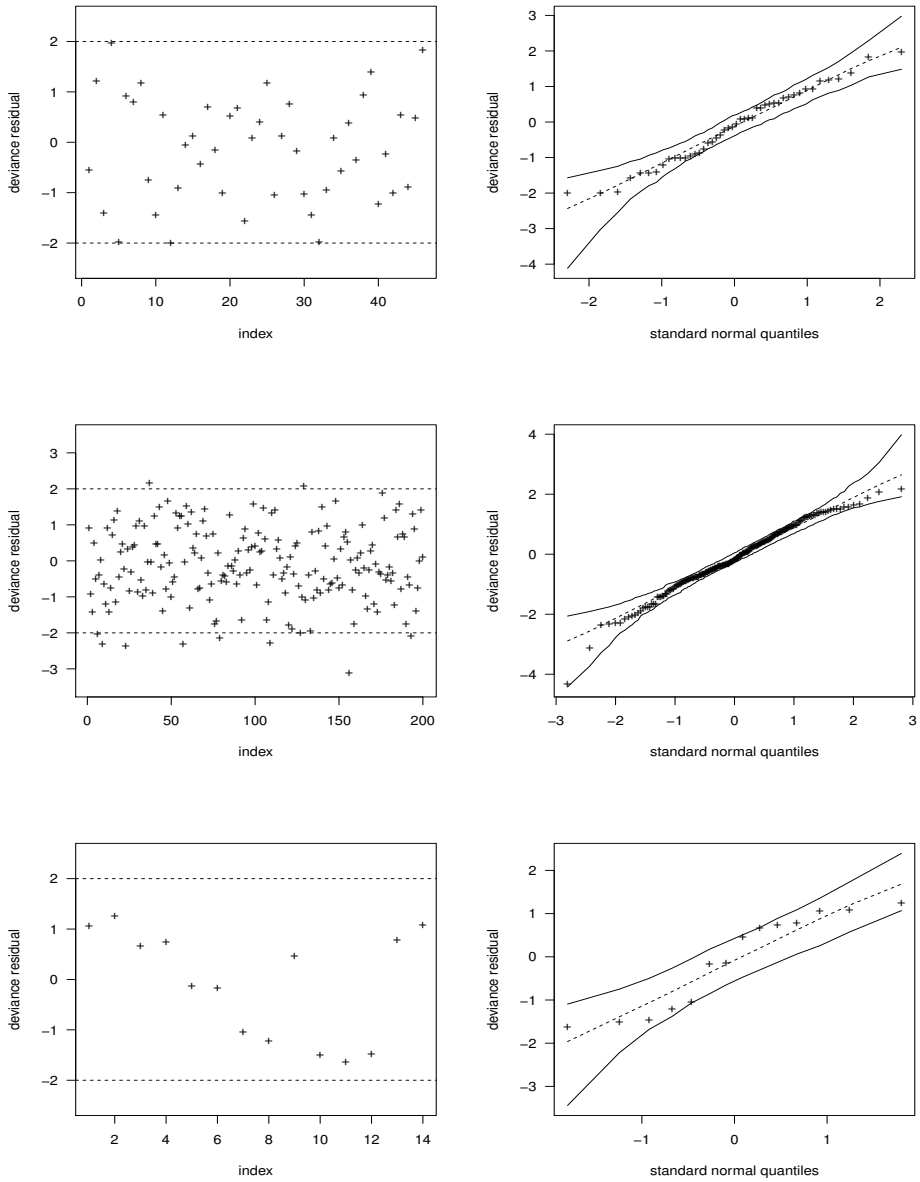


Figure 5: Residuals plots for the gamma regression: life of metal pieces data (top), advertising media data (middle), and radioimmunoassay data (bottom).

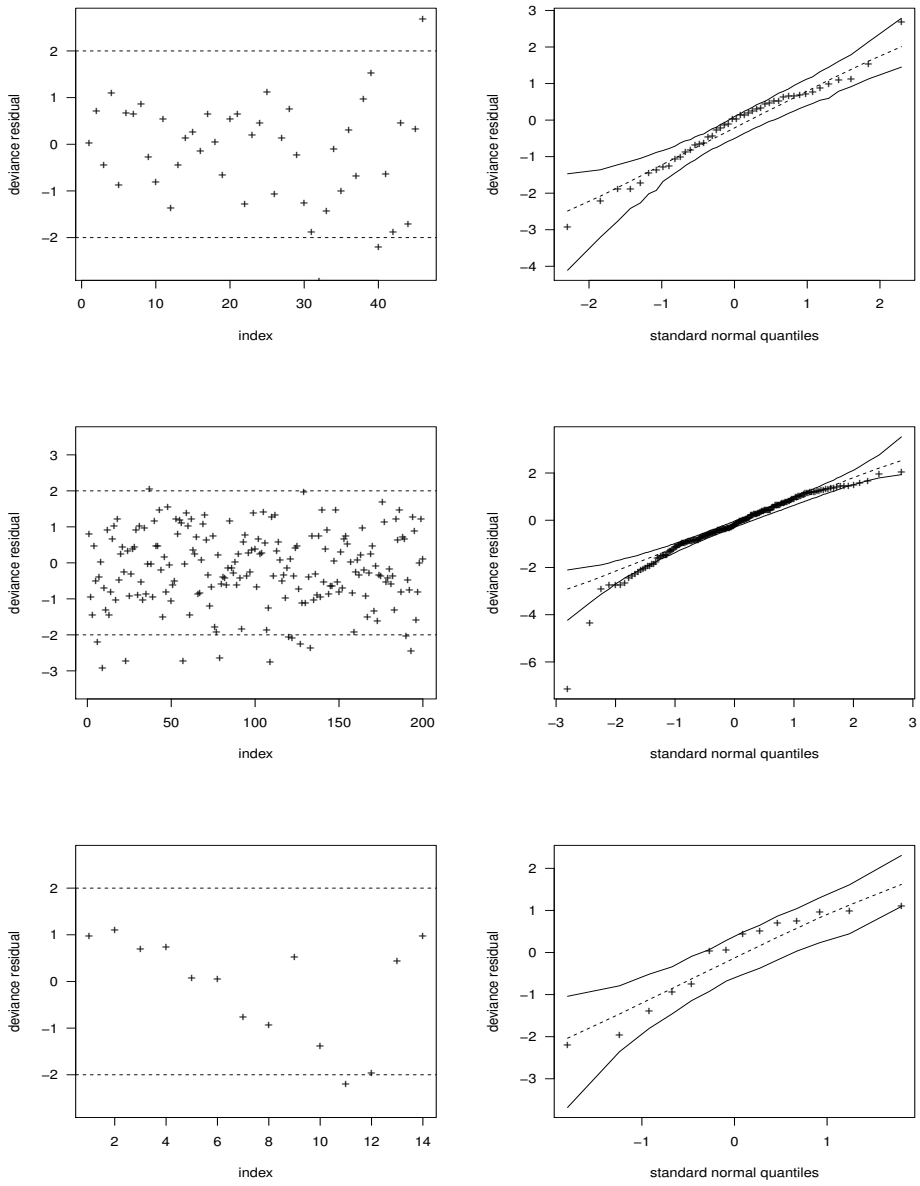


Figure 6: Residuals plots for the inverse Gaussian regression: life of metal pieces (top), advertising media data (middle), and radioimmunoassay data (bottom).

4.2. Choosing the best model

Here, we try to put some light on the following natural question: what is the best regression model to fit the data in the previous sections among the mean-parameterized Maxwell, gamma and inverse Gaussian regression models. It is worth stressing that this question is not easy to be answered in generality. The values of AIC and BIC of all fitted regression models are listed in Table 7. The Maxwell and gamma regressions outperform the inverse Gaussian regression to model the life of metal pieces data as well as the advertising media data, while these three regression models can be considered equivalent to model the radioimmunoassay data. On the basis of AIC and BIC values, it seems that the Maxwell and gamma regression models should be chosen as the best regression models to fit these three datasets. In terms of parsimony, the mean-parameterized Maxwell regression model should be preferable, once it has the advantage of having fewer parameters to be estimated than the gamma regression model. Remembering that in the gamma regression model is necessary to estimate a precision parameter, while in the Maxwell regression model is not.

Table 7: AIC and BIC values.

Model	Life of metal pieces		Advertising media		Radioimmunoassay	
	AIC	BIC	AIC	BIC	AIC	BIC
Maxwell	635.14	638.80	1295.17	1301.77	240.21	241.50
Gamma	635.73	641.22	1292.80	1302.70	239.49	241.41
Inverse Gaussian	647.68	653.17	1322.40	1332.30	241.38	243.30

From now on, we only consider the mean-parameterized Maxwell and gamma regression models. By following with the analysis in order to select the best regression model, we shall consider the generalized likelihood ratio test statistic ($V_{\rm LR}$) proposed by Vuong [32]. The statistic $V_{\rm LR}$ measures the distance between two models in terms of the Kullback–Leibler information criterion. The test statistic can be expressed as $V_{\rm LR} = \Lambda\Psi^{-1/2}$, and

$$\Lambda = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\log\left(\frac{\mathrm{Mw}(\widehat{\mu}_i)}{\mathrm{Ga}(\widehat{\mu}_i,\widehat{\phi})}\right),$$

$$\Psi = \frac{1}{n}\sum_{i=1}^{n}\left[\log\left(\frac{\mathrm{Mw}(\widehat{\mu}_i)}{\mathrm{Ga}(\widehat{\mu}_i,\widehat{\phi})}\right)\right]^2 - \left[\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{\mathrm{Mw}(\widehat{\mu}_i)}{\mathrm{Ga}(\widehat{\mu}_i,\widehat{\phi})}\right)\right]^2.$$

The statistic $V_{\rm LR}$ converges in distribution to a standard normal distribution under the null hypothesis of equivalence of the models. The null hypothesis is not rejected if $|V_{\rm LR}| \leq \Phi^{-1}(1-\alpha/2)$, where $\Phi^{-1}(\cdot)$ is the standard normal quantile function, and α is the significance level. On the other hand, we reject at significance level α the null hypothesis in favor of the Maxwell model being better (worse) than the gamma model if $V_{\rm LR} > \Phi^{-1}(1-\alpha)$ ($V_{\rm LR} < -\Phi^{-1}(1-\alpha)$). Table 8 lists the observed values of $V_{\rm LR}$ (and the corresponding p-values), indicating that the mean-parameterized Maxwell and gamma regression models are equivalent to fit these datasets. However, in terms of parsimony, the mean-parameterized Maxwell regression model should be preferable as mentioned early. In summary, the results in this section reveal that the mean-parameterized Maxwell regression model can be a good (and simple as well) alternative to the well-developed gamma regression model in practice.

Table 8: Generalized likelihood ratio statistic.

Data	V_{LR}	p-value
Life of metal pieces	−0.9320	0.3513
Advertising media	−0.5243	0.6001
Radioimmunoassay	−1.2102	0.2262

5. CONCLUDING REMARKS

In this paper, based on the mean-parameterized Maxwell distribution, a parametric class of regression models to deal with positive response variables was studied. By employing the frequentist approach, the estimation of the Maxwell regression parameters is conducted by the maximum likelihood method. We also provide a closed-form expression for the expected Fisher information matrix. Monte Carlo simulation experiments reveal that the maximum likelihood method is quite effective to estimate the Maxwell model parameters, and that the initial guesses we recommend for the Maxwell regression parameters worked perfectly well in the Monte Carlo simulations as well as real data applications. We also give a simple formula for calculating bias-corrected maximum likelihood estimates of the mean-parameterized Maxwell regression parameters. We discuss diagnostic techniques (global and local influence, and residuals analysis) for the mean-parameterized Maxwell regression model. Diagnostic methods have been an important tool in regression analysis to detect anomalies with the fitted model, such as departures from the model assumptions, presence of outliers and presence of influential observations. In particular, an appropriate matrix for assessing local influence on the Maxwell parameter estimates under a specific perturbation scheme is obtained. Additionally, we illustrate the methodology developed in this paper by means of applications to real data. We verify through the real data applications that the mean-parameterized Maxwell regression model was superior to the well-known inverse Gaussian regression model, and was very similar to the gamma regression model, which is, probably, the most used regression model to deal with positive response variables in practice. Finally, it is worth stressing that the formulas related with the mean-parameterized Maxwell regression model are manageable (such as log-likelihood function, score function, expected Fisher information matrix, etc.) and with the use of modern computer resources and its numerical capabilities, this regression model may prove to be an useful addition to the arsenal of applied statisticians.

The previous developments regarding the mean-parameterized Maxwell regression model indicate that this model can be indeed very useful in practice. Therefore, we would like to point out that the current work opens new possibilities for future works. In particular, an interesting extension of the mean-parameterized Maxwell regression model which allows for explanatory variables to be measured with error may be developed. Also, one may study the mean-parameterized Maxwell regression model under random effects. Additionally, due to recent advances in computational technology, one may explore other estimation methods for the mean-parameterized Maxwell regression model such as the Bayesian approach. In addition, Bayesian influence diagnostics can also be treated via the Kullback–Leibler divergence and, hence, atypical observations can also be identified in a Bayesian context. A very interesting extension of the developments considered in this paper would be to study the mean-parameterized Maxwell regression model in a semiparametric context. Obviously an in-depth investigation of such studies is beyond the scope of the current paper, but certainly are very interesting topics for future works.

APPENDIX. The R code

```
## R function to estimate the mean-parameterized
## Maxwell parameters (link function = "log")
Maxwell.reg <- function(formula, data){
    cl <- match.call()
    if (missing(data))
      data <- environment(formula)
    mf <- match.call(expand.dots = FALSE)
    m <- match(c("formula", "data"), names(mf), 0L)
    mf <- mf[c(1L, m)]
    mf$drop.unused.levels <- TRUE
    oformula <- as.formula(formula)
    mf$formula <- formula
    mf[[1L]] <- as.name("model.frame")
    mf <- eval(mf, parent.frame())
    mt <- terms(formula, data = data)
    Y <- model.response(mf, "numeric")
    X <- model.matrix(mf)
    if (length(Y) < 1)
      stop("empty model")
    if (!(min(Y) >= 0))
      stop("invalid dependent variable")
    floglikMax <- function(vPar){
      veta <- X%*%vPar
      vmu <- exp(veta)
      loglik <- sum( -3*log(vmu) - (4/pi)*(Y^2/vmu^2) )
      loglik
    }
    fscoreMax <- function(vPar){
      veta <- X%*%vPar
      vmu <- exp(veta)
      vt <- (8/pi)*(Y^2/vmu^2) - 3
      vt <- as.vector(vt)
      score <- t(X)%*%vt
      score
    }
    fFisherMax <- function(){
      6*(t(X)%*%X)
    }
    start <- c( solve(t(X)%*%X)%*%t(X)%*%log(Y+0.1) )
    opt <- optim(start, fn = floglikMax, gr = fscoreMax, method = "BFGS",
                 control=list(fnscale=-1), hessian=FALSE)
    if (opt$conv != 0)
      stop("algorithm did not converge")
    beta <- opt$par
    se <- sqrt(diag(solve(fFisherMax())))
    z.value <- beta/se
    p.value <- 2*(1 - pnorm(abs(z.value)))
    names(beta) <- colnames(X)
    rval <- cbind( round(beta, 6), round(se, 6),
                   round(z.value, 6), round(p.value, 6) )
    colnames(rval) <- c("Estimate", "Std. Error",
                          "z value", "Pr(>|z|)")
    return(rval)
}

## Example: Life of metal pieces data
## y = "number of cycles to failure" and x = "work  per cycle"
data(Biaxial, package="ssym")
attach(Biaxial)
y <- Life
x <- log(Work)
Maxwell.fit <- Maxwell.reg(y ~ x)
Maxwell.fit
             Estimate Std. Error   z value Pr(>|z|)
(Intercept) 12.473253   0.400698  31.12885        0
x           -1.706004   0.111374 -15.31775        0
```

ACKNOWLEDGMENTS

The author would like to thank two anonymous reviewers for their insightful comments and suggestions. The author also acknowledges the financial support of the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq: grant 304776/2019–0).

REFERENCES

[1] Al-Baldawi, T.H.K. (2015). Some Bayes estimators for Maxwell distribution with conjugate informative priors, *Al-Mustansiriyah Journal of Science*, **26**, 64–69.

[2] Atkinson, A.C. (1985). *Plots, Transformation and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford University Press, New York.

[3] Bekker, A. and Roux, J.J.J. (2005). Reliability characteristics of the Maxwell distribution: A Bayes estimation study, *Communications in Statistics – Theory and Methods*, **34**, 2169–2178.

[4] Bourguignon, M.; Leão, J. and Gallardo, D.I. (2019). Parametric modal regression with varying precision, *Biometrical Journal*, **61**, 1–19.

[5] Castellares, F.; Ferrari, S.L.P. and Lemonte, A.J. (2018). On the Bell distribution and its associated regression model for count data, *Applied Mathematical Modelling*, **6**, 172–185.

[6] Chen, Y.; Genovese, C.R.; Tibshirani, R.J. and Wasserman, L. (2016). Nonparametric modal regression, *The Annals of Statistics*, **44**, 489–514.

[7] Cook, R.D. (1977). Detection of influential observation in linear regression, *Technometrics*, **19**, 15–18.

[8] Cook, R.D. (1986). Assessment of local influence, *Journal of the Royal Statistical Society B*, **48**, 133–169.

[9] Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

[10] Cox, D.R. and Snell, E.J. (1968). A general definition of residuals (with discussion), *Journal of the Royal Statistical Society B*, **30**, 248–275.

[11] Dar, A.A.; Ahmed, A. and Reshi, J.A. (2017). Bayesian analysis of Maxwell–Boltzmann distribution under different loss functions and prior distributions, *Pakistan Journal of Statistics*, **33**, 419–440.

[12] Dey, S.; Dey, T. and Maiti, S.S. (2013). Bayesian inference for Maxwell distribution under conjugate prior, *Model Assisted Statistics and Applications*, **8**, 193–203.

[13] Dey, S. and Maiti, S.S. (2013). Estimation of the parameter of Maxwell distribution under different loss functions, *Journal of Statistical Theory and Practice*, **4**, 279–287.

[14] Dunn, P.K. and Smyth, G.K. (1996). Randomised quantile residuals, *Journal of Computational and Graphical Statistics*, **5**, 236–244.

[15] Fan, G. (2016). Estimation of the loss and risk functions of parameter of Maxwell distribution, *Science Journal of Applied Mathematics and Statistics*, **4**, 129–133.

[16] Feng, C.; Sadeghpour, A. and Li, L. (2017). Randomized quantile residuals: an omnibus model diagnostic tool with unified reference distribution, *arXiv* preprint arXiv:1708.08527.

[17] Gallardo, D.I.; Gómez-Déniz, E.; Leão, J. and Gómez, H.W. (2020). Estimation and diagnostic tools in reparameterized slashed Rayleigh regression model. An application to chemical data, *Chemometrics and Intelligent Laboratory Systems*, **207**, 104–189.

[18] Green, P.J. and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman and Hall, Boca Raton.

[19] Gómez, Y.M.; Gallardo, D.I.; Leão, J. and Gómez, H.W. (2020). Extended exponential regression model: diagnostics and application to mineral data, *Symmetry*, **12**, 2042.

[20] Hossain, M.P.; Sanusi, R.A.; Omar, M.H. and Riaz, M. (2019). On designing Maxwell CUSUM control chart: an efficient way to monitor failure rates in boring processes, *The International Journal of Advanced Manufacturing Technology*, **100**, 1923–1930.

[21] Kassambara, A. (2019). *datarium: Data Bank for Statistical Analysis and Visualization*, R package version 0.1.0.

[22] Leão, J.; Bourguignon, M.; Saulo, H.; Santos-Neto, M. and Calsavara, V. (2021). The negative binomial beta prime regression model with cure rate: application with a aelanoma dataset, *Journal of Statistical Theory and Practice*, **15**, p63.

[23] Lemonte, A.J. and Bazan, J. (2016). New class of Johnson S_B distributions and its associated regression model for rates and proportions, *Biometrical Journal*, **58**, 727–746.

[24] Li, L. (2016). Minimaxestimation of the parameter of Maxwell distribution under different loss functions, *American Journal of Theoretical and Applied Statistics*, **5**, 202–207.

[25] McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

[26] Menezes, A.F.B.; Mazucheli, J. and Chakraborty, S. (2021). A collection of parametric modal regression models for bounded data, *Journal of Biopharmaceutical Statistics*, **31**, 490–506.

[27] Pierce, D.A. and Schafer, D.W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association*, **81**, 977–986.

[28] R Core Team (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

[29] Rieck, J.R. and Nedelman, J.R. (1991). A log-linear model for the Birnbaum–Saunders distribution, *Technometrics*, **33**, 51–60.

[30] Tiede, J.J. and Pagano, M. (1979). Application of robust calibration to radioimmunoassay, *Biometrics*, **35**, 567–574.

[31] Tyagi, R.K. and Bhattacharya, S.K. (1989). Bayes estimator of the Maxwell's velocity distribution function, *Statistica*, **49**, 563–566.

[32] Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica*, **57**, 307–333.

[33] Yao, W. and Li, L. (2014). A new regression model: modal linear regression, *Scandinavian Journal of Statistics*, **41**, 656–671.

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 527–544
https://doi.org/10.57805/revstat.v22i4.541

Kernel Estimation of the Dynamic Cumulative Past Inaccuracy Measure for Right Censored Dependent Data

Authors: K. V. Viswakala
– Department of Statistics, University of Kerala, Thiruvananthapuram, Kerala, India
viswakalakv@gmail.com

E. I. Abdul Sathar ✉
– Department of Statistics, University of Kerala, Thiruvananthapuram, Kerala, India
sathare@gmail.com

Received: March 2022 Revised: December 2022 Accepted: December 2022

Abstract:

- This paper proposes a nonparametric estimator for the lifetime distribution's dynamic cumulative past inaccuracy measure based on censored dependent data. The asymptotic properties of the estimator are discussed under suitable regularity conditions. We use Monte-Carlo simulations to compare the estimator's performance to that of an empirical estimator using mean squared errors to test the estimator's properties numerically. The methods are demonstrated using two different real data sets.

Keywords:

- *dynamic cumulative past inaccuracy measure; alpha-mixing; mean squared error (MSE); mean integrated squared error (MISE).*

AMS Subject Classification:

- 62B10, 62G20.

✉ Corresponding author

1. INTRODUCTION

Let $f(x)$ and $g(x)$ be the probability density functions (pdfs) of the failure times of two systems X and Y, with distribution functions $F(x) = P(X \leq x)$ and $G(x) = P(Y \leq x)$ respectively. Kerridge's [12] measure of inaccuracy between X and Y is given by

$$I(X,Y) = -\int_0^\infty f(x) \log g(x) dx. \tag{1.1}$$

It has been known for a long time as a helpful tool for determining the degree of error in experimental results. It can also be interpreted as an error that occurred when an experiment's true density function, $f(x)$, was assigned to $g(x)$ by the experimenter. Kerridge [12] discussed the application of inaccuracy measures in statistical inference. This measure is also applied in the field of economics. International demand or cross-country demand analysis estimates the demand for goods or services for a group of countries. James and Anita [10] address the outlier problem in the international demand analysis, which can be remedied using inaccuracy measures. Kayal and Sunoj [11] introduced a generalized dynamic conditional Kerridge's inaccuracy measure, which can be represented as the sum of conditional Renyi's divergence and Renyi's entropy. Rajesh *et al.* [17] and Sathar *et al.* [20] suggested nonparametric estimator for inaccuracy measure in the reliability context, such as residual life distributions and past life distributions, respectively, and found their properties under some regularity conditions.

Hooda and Tuteja [9] defined some nonadditive measures of relative information and inaccuracy. Using reversible symmetry, Bhatia and Taneja [2] defined the quantitative-qualitative measure of inaccuracy. Straightforwardly, Gur Dial [8] established the noiseless coding theorems for subjective probability codes for nonadditive measures of inaccuracy. Goel *et al.* [7] introduce and discuss a measure of inaccuracy between the distributions of n^{th} record value. Although this measure is inapplicable when the random variables' pdfs are void, Kundu *et al.* [16] proposed an alternative measure of inaccuracy called dynamic cumulative past inaccuracy between random variables X and Y, which is represented as

$$\bar{C}I(X,Y) = -\int_0^\infty F(x) \log G(x) dx. \tag{1.2}$$

Kundu *et al.* [16] investigated general results for this measure. Relying on various applications of stochastic classes in reliability and information theory fields, Khorashadizadeh [13] studied new classes of the lifetime in terms of cumulative inaccuracy along with their relations with other famous aging classes. Also, some characterization results are obtained under the proportional reversed hazard rate model. Di Crescenzo and Longobardi [4] defined the empirical expression of cumulative inaccuracy in connection with empirical cumulative entropy.

In many realistic situations, if a system is found to be down at time t, the random variable $[t - X|X \leq x]$ describes the time elapsed between the failure of a system and the time. Based on this idea, the cumulative inaccuracy measure between two past lifetimes, analogous to the measure, (1.2), is defined by Kumar and Taneja [15] and Kundu *et al.* [16] independently as

$$\bar{C}I(X,Y,t) = -\int_0^t \frac{F(x)}{F(t)} \log\left[\frac{G(x)}{G(t)}\right] dx, \tag{1.3}$$

and so called dynamic cumulative past inaccuracy measure. Clearly when $t = 0$, (1.3) becomes (1.2). (1.3) equivalently can be written as

$$\begin{aligned} \bar{CI}(X,Y,t) &= -\frac{1}{F(t)}\int_0^t F(x)\log G(x)dx + \frac{\log G(t)}{F(t)}\int_0^t F(x)dx \\ &= \bar{\mathcal{A}}_t + \bar{\mathcal{B}}_t, \end{aligned} \tag{1.4}$$

where

$$\bar{\mathcal{A}}_t = -\frac{1}{F(t)}\int_0^t F(x)\log G(x)dx \quad \text{and} \quad \bar{\mathcal{B}}_t = \frac{\log G(t)}{F(t)}\int_0^t F(x)dx.$$

Ghosh and Kundu [6] introduced the notion of cumulative past inaccuracy of order α and study the proposed measure for conditionally specified models of two components failed at different time instants, called generalized conditional cumulative past inaccuracy, and their properties are discussed.

Example 1.1. Let the random variables X and Y have the following distribution functions $F(x) = 2x - x^2$, and $G(x) = x^\lambda$ respectively, $x \in [0,1]$. Then for $t \in [0,1]$, the dynamic cumulative past inaccuracy measure, $\bar{CI}_n(t)$ is obtained as

$$\bar{CI}_n(t) = \frac{\lambda t(2t-9)}{18(t-2)}.$$

Figure 1 depicts the dynamic cumulative past inaccuracy measure for for $t \in [0,1]$ and for $\lambda \in \{4,\ 6,\ ...,\ 12\}$. According to the graph, the dynamic cumulative past inaccuracy measure is an increasing function in λ and t.

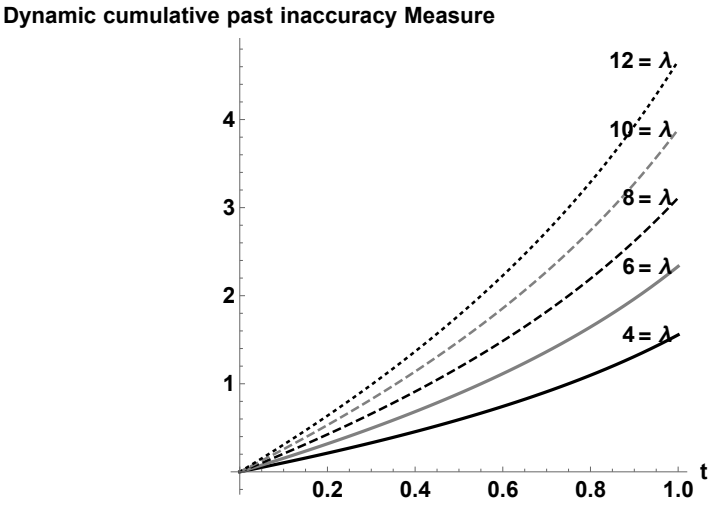


Figure 1: Plot of $\bar{CI}(X,Y,t)$ against $t \in [0,1]$ for different parameter λ.

From a practical standpoint, it appears more reasonable to forego independence in favour of some dependency. For example, if a family's income is exclusively dependent on the

salary of one of its members, an accident or the death of that individual will have a negative influence on the family's performance. However, this will not be the case when examined from the perspective of society as a whole. Random variables are derived from specific types of mixing conditions that have already been defined in the literature. Alpha-mixing is a strong mixing condition with many practical applications among the various mixing conditions used in the literature.

Censorship is either desirable or unavoidable in life testing and can take several forms. Withdrawals from a clinical trial, death unrelated to the condition under study, and a person still alive at the end of the follow-up period are all examples of random censoring. Right censorship is one of the most common types of censorship. Right censoring is appropriate in studies of electrical equipment failure, the occurrence of a specific disease, and so on.

Motivated by the emerging work and the importance of (1.3), we intend to develop a kernel function-based estimation technique for this measure in practical situations. This paper considers the nonparametric estimation of (1.3) under right censoring and discusses some of its properties. Throughout this paper, we assume that the random variables are alpha-mixing (Rosenblatt [19]).

This paper's outline is as follows: In Section 2, we present a nonparametric estimator for (1.4) in censored samples. Section 3 looks into the asymptotic properties of the estimator. Section 4 contains a simulation study to demonstrate the estimator's behaviour and a comparison to an empirical estimator. Furthermore, they are compared to two different real-world data sets.

2. KERNEL ESTIMATION

In this section, we propose a nonparametric estimator for the cumulative past inaccuracy measure for right censored data sets. Consider $\{X_i\}$, $\{Y_i\}$, $i = 1, 2, ..., n$ be identically distributed random samples have distribution functions be $F(x) = Pr(X_i \leq x)$ and $G(x) = Pr(Y_i \leq x)$ respectively. We use independent and identically distributed random variable R_{1i} and R_{2i} with corresponding distribution functions $P_1(x)$ and $P_2(x)$ for creating right-censored data from X_i and Y_i respectively. Note that R_{1i} and R_{2i} are independent of X_i and Y_i respectively. Let $C_i = \min(X_i, R_{1i})$, $C_i^* = \min(Y_i, R_{2i})$, $\delta_i = I(X_i \leq R_{1i})$ and $\delta_i^* = I(Y_i \leq R_{2i})$. Then the kernel density estimator of (1.4) under right censoring is as follows:

$$\begin{aligned} \bar{CI}_n(t) &= \bar{\mathcal{A}}_{nt} + \bar{\mathcal{B}}_{nt}, \\ &= -\frac{1}{F_n(t)} \int_0^t F_n(x) \log G_n(x) dx + \frac{\log G_n(t)}{F_n(t)} \int_0^t F_n(x) dx, \end{aligned} \tag{2.1}$$

where

$$F_n(t) = \int_0^t \frac{1}{nh} \sum_{i=1}^{n} \frac{K\left(\frac{x-C_i}{h}\right)\delta_i dx}{1 - P_1(C_i)} \quad \text{and} \quad G_n(t) = \int_0^t \frac{1}{nh} \sum_{i=1}^{n} \frac{K\left(\frac{x-C_i^*}{h}\right)\delta_i^* dx}{1 - P_2(C_i^*)},$$

respectively are the nonparametric density estimator for $F(t)$ and $G(t)$ under censoring and $K(\cdot)$ be the kernel function. For the positive integers, i and j $h \to 0$, $nh \to \infty$ and the following assumptions hold:

(i) $f^{(k)}(x), 1 \leq k \leq 2j$, exists and $f^{(2j)}(x)$ is bounded

and if $K(\cdot)$ satisfies,

(ii) $K(s) \geq 0,\ -\infty < s < \infty\ ,\ \int_{\mathbb{R}^+} K(s)ds = 1,\ \int_{\mathbb{R}^+} s^a K^i(s)ds = 0$ for positive odd integer a, $\int_{\mathbb{R}^+} s^b K^i(s)ds < \infty$ for positive even integer b.

Denote $Q^*(t) = P(C_1 \leq t, \delta_1 = 1)$ the sub distribution function for the uncensored observations, and $q^*(t) = [1 - P_1(t)]f(t)$ the corresponding density, then a reasonable estimate of $f(t)$ can be obtained from Cai [3] as $q^*(t)/[1 - P_1(t)]$. Consider the transformation $\alpha = \frac{x - C_i}{h}$, then we get

$$\begin{aligned} E\left[\frac{1}{h}\frac{K\left(\frac{x-C_i}{h}\right)\delta_i}{1-P_1(C_i)}\right] &= \frac{1}{h}\int_{\mathbb{R}^+}\frac{K\left(\frac{x-C_i}{h}\right)}{1-P_1(C_i)}q^*(C_i)dC_i, \\ &= \int_{\mathbb{R}^+} K(\alpha)f(x-\alpha h)d\alpha, \\ &= \int_{\mathbb{R}^+} K(\alpha)\left[f(x) - f^{(1)}(x)\alpha h + \frac{f^{(2)}(x)}{2!}\alpha^2 h^2 - ...\right], \\ &= f(x) + \frac{h^2}{2}\int_{\mathbb{R}^+}\alpha^2 K(\alpha)d\alpha f^{(2)}(x) + O(h_n^2), \end{aligned} \tag{2.2}$$

and using Lemma 2 in Elias Masry [5], we get

$$E\left[\frac{1}{h}\frac{K\left(\frac{x-C_i}{h}\right)\delta_i}{1-P_1(C_i)}\right]^2 = \frac{C_k}{h}\frac{f(x)}{1-P_1(x)}, \tag{2.3}$$

where $C_k = \int_{\mathbb{R}^+} K^2(\alpha)d\alpha$. Let $K_1 = \frac{K\left(\frac{x-C_i}{h}\right)\delta_i}{1-P_1(C_i)}$, then using (2.2) and (2.3), we get

$$\begin{aligned} \text{Bias}[F_n(t)] &= \int_0^t E\left(\frac{K_1}{h}\right)dx - F(x), \\ &= \frac{h^2}{2}\int_{\mathbb{R}^+}\alpha^2 K(\alpha)d\alpha\int_0^t f^{(2)}(x)dx + O(h^4), \\ \text{Var}[F_n(t)] &\approx \frac{1}{n}\left\{\int_0^t E\left(\frac{K_1}{h}\right)^2 dx - \int_0^t\left[E\left(\frac{K_1}{h}\right)\right]^2 dx\right\} \\ &\quad + \left\{\int_0^t E\left(\frac{K_1}{h}\right)dx - F(x)\right\}^2, \\ &= \frac{C_k}{nh}\int_0^t\frac{f(x)}{1-P_1(x)}dx. \end{aligned}$$

2.1. Estimation of $\bar{\mathcal{A}}_t$ and $\bar{\mathcal{B}}_t$

Using Taylor's series expansion, we have

$$\log G_n(x) = \log G(x) + \frac{G_n(x) - G(x)}{G(x)} + R_n,$$

where

$$R_n = \int_0^1 \frac{2(1-\tau)}{\{G(x)+\tau[G_n(x)-G(x)]\}^2}[G_n(x)-G(x)]^2 d\tau.$$

Hence,

$$\begin{aligned} F_n(x)\log G_n(x) - F(x)\log G(x) \; &= \\ &= \log G(x)[F_n(x)-F(x)] + \frac{1}{G(x)}[F_n(x)-F(x)][G_n(x)-G(x)] \\ &\quad + \frac{F(x)}{G(x)}[G_n(x)-G(x)] + R_n[F_n(x)-F(x)] + F(x)R_n. \end{aligned} \tag{2.4}$$

Next, we need to find $E|R_n|^j$, for any positive integer j. For this, consider $V_1 = \left\{x : |G_n(x)-G(x)| \le \frac{G(x)}{2}\right\}$ and V_1^c is the compliment of V_1. Clearly, for $\theta \in V_1$ and for every $0 \le \epsilon \le 1$, we have

$$0 < G(x)\Big(1-\frac{\epsilon}{2}\Big) \le G(x) + \epsilon[G_n(x)-G(x)] < G(x)\Big(1+\frac{\epsilon}{2}\Big).$$

Equivalently, we get

$$0 < \frac{(1-\epsilon)[G_n(x)-G(x)]^2}{\{G(x)+\epsilon[G_n(x)-G(x)]\}^2} \le \frac{(1-\epsilon)[G_n(x)-G(x)]^2}{\big[(1-\frac{\epsilon}{2})G(x)\big]^2}.$$

Let $I(\cdot)$ denotes the indicator function, since $\int_0^1 \frac{(1-\epsilon)}{\big(1-\frac{\epsilon}{2}\big)^2} d\epsilon < 1$, then for every positive integer j we get

$$E|R_n|^j I(V_1) \; \le \; \frac{1}{[G(x)]^{2j}} E[G_n^*(x)-G(x)]^{2j}.$$

Also, we have

$$E|R_n|^j I(V_1^c) \; \le \; E\left[\left|\frac{1}{G_n(x)} - \frac{1}{G(x)} - \frac{G_n(x)-G(x)}{G(x)}\right|^j I(V_1^c)\right].$$

For $1 \le i \le n$, we have $K(x-Y_i) \ne 0$ and $m < K(\alpha) < N$ so that $G_n(x) \ge \frac{m}{nh}$, or equivalently, $\frac{1}{G_n(x)} \le \frac{nh}{m}$. Also, $G_n(x) \le \frac{N}{h}$ and $nh^2 \to \infty$ implies for sufficiently large n,

$$\begin{aligned} E|R_n|^j I(V_1^c) \; &\le \; \left|\frac{nh^2}{m} + h - \frac{h}{G(x)} - \frac{N}{G(x)}\right|^j \frac{1}{h^j} E[I(V_1^c)], \\ &= \; O(n^j h^j) P\left[|G_n^*(x)-G(x)| \ge \frac{G(x)}{2}\right], \\ &\le \; O(n^j h^j)\left\{P\left[|G_n^*(x)-E[G_n^*(x)]| \ge \frac{G(x)}{4}\right]\right. \\ &\quad \left. + \; P\left[|E[G_n^*(x)]-G(x)| \ge \frac{G(x)}{4}\right]\right\}. \end{aligned}$$

For sufficiently large n,

$$P\left[|E[G_n^*(x)] - G(x)| \geq \frac{G(x)}{4}\right] = 0,$$

and

$$P\left[|G_n(x) - E[G_n(x)]| \geq \frac{G(x)}{4}\right] \leq 2\exp\{-Cnh\},$$

for some constant C, (see Rao [18]), we obtain for sufficiently large n

$$E|R_n|^j I(V_1^c) \leq 2\exp\{-Cnh\}.$$

Also, we have

$$\begin{aligned} E|R_n|^j &= E|R_n|^j I(V_1) + E|R_n|^j I(V_1^c), \\ &\leq \frac{1}{[G(x)]^{2j}} E[G_n(x) - G(x)]^{2j} + O(n^j h^j)\exp\{-Cnh\}. \end{aligned}$$

In particular for $j = 1, 2$ in the above inequality, we get

$$\begin{aligned} E|R_n| &\leq \frac{1}{[G(x)]^2} E[G_n(x) - G(x)]^2 + O(nh)\exp\{-Cnh\}, \\ &= O\left(\frac{1}{nh}\right) + O(h^4) + O(nh), \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} E|R_n|^2 &\leq \frac{1}{[G(x)]^4} E[G_n(x) - G(x)]^4 + O(n^2h^2)\exp\{-Cnh\}, \\ &= O\left(\frac{h^3}{n}\right) + O\left(\frac{1}{n^2h^2}\right) + O(h^8) + O(n^2h^2), \end{aligned} \tag{2.6}$$

since nh goes to infinity, $O(nh)\exp\{-Cnh\}$ and $O(n^2h^2)\exp\{-Cnh\}$ have smaller orders than that of $E[G_n(x) - G(x)]$ and $E[G_n(x) - G(x)]^2$ respectively.

In order to simplify the notation define $\mathfrak{h}_n(t) = \int_0^t F_n(x)\log G_n(x)dx$, $\mathfrak{g}_n(t) = \int_0^t F_n(x)dx$, $\mathfrak{h}(t) = \int_0^t F(x)\log G(x)dx$ and $\mathfrak{g}(t) = \int_0^t F(x)dx$ so that we can easily prove that

$$\frac{\mathfrak{h}_n(t)}{F_n(t)} - \frac{\mathfrak{h}(t)}{F(t)} \approx \frac{\mathfrak{h}_n(t) - \frac{\mathfrak{h}(t)}{F(t)}F_n(t)}{F(t)}, \tag{2.7}$$

and

$$\frac{\log G_n(t)\mathfrak{g}_n(t)}{F_n(t)} - \frac{\log G(t)\mathfrak{g}(t)}{F(t)} \approx \frac{\log G_n(t)\mathfrak{g}_n(t) - \frac{\log G(t)\mathfrak{g}(t)}{F(t)}F_n(t)}{F(t)}. \tag{2.8}$$

Hence using (2.4)–(2.8), we get the following:

$$\begin{aligned} \text{Bias}[\bar{\mathcal{A}}_{nt}] &= -\text{Bias}\left[\frac{\mathfrak{h}_n(t)}{F_n(t)} - \frac{\mathfrak{h}(t)}{F(t)}\right], \\ &= \frac{-h^2}{2}\int_{\mathbb{R}^+}\alpha^2 K(\alpha)d\alpha\left\{\frac{1}{F(t)}\int_0^t\left[\log G(x)\int_0^x f^{(2)}(y)dy\right.\right. \\ &\quad \left.\left.+\frac{F(x)}{G(x)}\int_0^x g^{(2)}(y)dy\right]dx - \frac{\mathfrak{h}(t)}{F^2(t)}\int_0^t f^{(2)}(x)dx\right\} + O(h^4), \end{aligned} \tag{2.9}$$

and

$$\begin{aligned}\text{Bias}[\bar{\mathcal{B}}_{nt}] &= \frac{1}{F(t)}\,\text{Bias}[\mathfrak{g}_{\mathfrak{n}}(t)\log G_n(t)] - \frac{\log G(t)\mathfrak{g}(t)}{F^2(t)}\,\text{Bias}[F_n(t)],\\ &= \frac{h^2}{2}\int_{\mathbb{R}^+}\alpha^2K(\alpha)d\alpha\Bigg\{\frac{\log G(t)}{F(t)}\int_0^t\int_0^x f^{(2)}(y)dydx + \frac{\mathfrak{g}(t)}{F(t)G(t)}\int_0^t g^{(2)}(x)dx\\ &\quad - \frac{\log G(t)\mathfrak{g}(t)}{F^2(t)}\int_0^t f^{(2)}(x)dx\Bigg\} + O(h^4). \qquad (2.10)\end{aligned}$$

Moreover,

$$\begin{aligned}\text{Var}[\bar{\mathcal{A}}_{nt}] &= \text{Var}\left[\frac{\mathfrak{h}_{\mathfrak{n}}(t)}{F_n(t)} - \frac{\mathfrak{h}(t)}{F(t)}\right],\\ &\approx \frac{C_k}{nh}\frac{1}{F^2(t)}\Bigg\{\int_0^t \log^2 G(x)\int_0^x \frac{f(y)}{1-P_1(y)}dydx\\ &\quad + \int_0^t\left[\frac{F(x)}{G(x)}\right]^2\int_0^x\frac{g(y)}{1-P_2(y)}dydx + \frac{\mathfrak{h}^2(t)}{F^2(t)}\int_0^t\frac{f(x)}{1-P_1(x)}dx\Bigg\}, \qquad (2.11)\end{aligned}$$

and

$$\begin{aligned}\text{Var}[\bar{\mathcal{B}}_{nt}] &= \frac{1}{F^2(t)}\,\text{Var}[\mathfrak{g}_{\mathfrak{n}}(t)\log G_n(t)] + \left[\frac{\log G(t)\mathfrak{g}(t)}{F^2(t)}\right]^2\text{Var}[F_n(t)],\\ &\approx \frac{C_k}{nh}\frac{\log^2 G(t)}{F^2(t)}\Bigg\{\int_0^t\int_0^x\frac{f(y)}{1-P_1(y)}dydx + \frac{\mathfrak{g}^2(t)}{F^2(t)}\int_0^t\frac{f(x)}{1-P_1(x)}dx\Bigg\}\\ &\quad + \frac{C_k}{nh}\frac{\mathfrak{g}^2(t)}{G^2(t)F^2(t)}\int_0^t\frac{g(x)}{1-P_2(x)}dx. \qquad (2.12)\end{aligned}$$

The following theorem gives bias and variance of the proposed estimator.

Theorem 2.1. *Under the assumptions given in Section 2, bias and variance of* $\bar{C}I_n(t)$ *is given as*

$$\begin{aligned}\text{Bias}\left[\bar{C}I_n(t)\right] &= \frac{h^2}{2}\int_{\mathbb{R}^+}\alpha^2K(\alpha)d\alpha\Bigg\{\frac{\log G(t)}{F(t)}\int_0^t\int_0^x f^{(2)}(y)dydx + \frac{\mathfrak{g}(t)}{F(t)G(t)}\int_0^t g^{(2)}(x)dx\\ &\quad - \frac{\log G(t)\mathfrak{g}(t)}{F^2(t)}\int_0^t f^{(2)}(x)dx - \frac{1}{F(t)}\int_0^t\log G(x)\int_0^x f^{(2)}(y)dydx\\ &\quad - \frac{1}{F(t)}\int_0^t\frac{F(x)}{G(x)}\int_0^x g^{(2)}(y)dydx + \frac{\mathfrak{h}(t)}{F^2(t)}\int_0^t f^{(2)}(x)dx\Bigg\},\end{aligned}$$

and

$$\begin{aligned}\text{Var}\left[\bar{C}I_n(t)\right] &\approx \frac{C_k}{nh}\frac{1}{F^2(t)}\Bigg\{\int_0^t\log^2 G(x)\int_0^x\frac{f(y)}{1-P_1(y)}dydx + \int_0^t\left[\frac{F(x)}{G(x)}\right]^2\int_0^x\frac{g(y)}{1-P_2(y)}dydx\\ &\quad + \frac{\mathfrak{h}^2(t)}{F^2(t)}\int_0^t\frac{f(x)}{1-P_1(x)}dx + \log^2 G(t)\int_0^t\int_0^x\frac{f(y)}{1-P_1(y)}dydx\\ &\quad + \frac{\log^2 G(t)\mathfrak{g}^2(t)}{F^2(t)}\int_0^t\frac{f(x)}{1-P_1(x)}dx + \frac{\mathfrak{g}^2(t)}{G^2(t)}\int_0^t\frac{g(x)}{1-P_2(x)}dx\Bigg\}.\end{aligned}$$

Proof: Using the equations (2.9), (2.10), (2.11) and (2.12), the result follows. □

The following example shows the application of Theorem 2.1.

Example 2.1. Consider the two non-negative random variables X and Y have the pdfs $f(x)$ and $g(x)$ respectively, so that for $x \in (0,1)$

$$\begin{aligned} f(x) &= 2x \text{ and } F(x) = P(X \leq x) = x^2, \\ g(x) &= 3x^2 \text{ and } G(x) = P(Y \leq x) = x^3. \end{aligned}$$

Let the random variables X and Y be right censored by uniform random variables with parameters $(0, 0.5)$ and $(0.5, 1)$, respectively. Then we get

$$\text{Bias}\left[\bar{CI}_n(t)\right] = \frac{-2h^2}{t}\int_{\mathbb{R}^+} \alpha^2 K(\alpha)d\alpha,$$

and

$$\begin{aligned} \text{Var}\left[\bar{CI}_n(t)\right] \approx \frac{C_k}{nh}\frac{1}{t^4}\Bigg\{ & \int_0^t \frac{9\log^2 x}{2}[\log(1-2x) - 2x]dx - \frac{3(t-2)t + 6(t-1)\log(1-t)}{4t} \\ & - \frac{t^2}{18}[(3\log t - 1)^2 + 9\log^2 t][2t + \log(1-2t)] - \frac{1}{12}[t(2+t) + 2\log(1-t)] \\ & + \frac{9\log^2 t}{2}[t - t^2 - (t - \frac{1}{2})\log(1-2t)]\Bigg\}. \end{aligned}$$

3. ASYMPTOTIC PROPERTIES

In this section, we discuss some asymptotic properties of (1.4). The following theorem reveals the consistency property of the estimator.

Theorem 3.1. *Under the assumptions given in Section 2, $\bar{CI}_n(t)$ is a consistent estimator of $\bar{CI}(X, Y, t)$.*

Proof: We have

$$\bar{CI}_n(t) = \frac{\log G_n(t)\mathfrak{g}_\mathfrak{n}(t)}{F_n(t)} - \frac{\mathfrak{h}_\mathfrak{n}(t)}{F_n(t)}.$$

$\text{MSE}[\mathfrak{h}_\mathfrak{n}(t)] \to 0$, $\text{MSE}[\mathfrak{g}_\mathfrak{n}(t)] \to 0$, $\text{MSE}[F_n(t)] \to 0$, $\text{MSE}[\log G_n(t)] \to 0$, when $n \to \infty$, and using Slutsky's theorem we obtain desired result. □

In the following theorem, we check the asymptotic nature of the estimator's mean integrated squared error (MISE).

Theorem 3.2. *Under the assumptions given in Section 2, the MISE of $\bar{CI}_n(t)$ tends to zero as $n \to \infty$.*

Proof:

$$\begin{aligned} \text{MISE}[\bar{CI}_n(t)] &= E\int\left[\bar{CI}_n(t) - \bar{CI}(X,Y,t)\right]^2 dt, \\ &= \text{MISE}[\bar{\mathcal{A}}_{nt}] + \text{MISE}[\bar{\mathcal{B}}_{nt}] + 2E\int\left[\bar{\mathcal{A}}_{nt} - \bar{\mathcal{A}}_t\right]\left[\bar{\mathcal{B}}_{nt} - \bar{\mathcal{B}}_t\right]dt. \end{aligned}$$

Also

$$\text{MISE}[\bar{\mathcal{A}}_{nt}] \leq \int \frac{1}{F^2(t)}\Bigg\{\text{MSE}[\mathfrak{h}_{\mathfrak{n}}(t)] + \left[\frac{\mathfrak{h}(t)}{F(t)}\right]^2 \text{MSE}[F_n(t)] - 2\frac{\mathfrak{h}(t)}{F(t)}\text{MSE}^{\frac{1}{2}}[\mathfrak{h}_{\mathfrak{n}}(t)]\text{MSE}^{\frac{1}{2}}[F_n(t)]\Bigg\}dt \to 0,$$

as $n \to \infty$. Using similar steps and applying Holder's inequality, we get the proof. □

The following theorem states the asymptotic normal distribution of the proposed estimator.

Theorem 3.3. *Let $\bar{C}I_n(t)$ be nonparametric estimator of $\bar{C}I(X,Y,t)$, $K(x)$ be a kernel and h satisfying the conditions for bandwidth. Then for fixed t*

$$(nh)^{\frac{1}{2}}\left[\frac{\bar{C}I_n(t) - \bar{C}I(X,Y,t)}{\sigma_{\bar{C}I_n}}\right]$$

follows normal distribution with mean zero and variance 2, as $n \to \infty$ with

$$\begin{aligned}\sigma^2_{\bar{C}I_n} &= \frac{C_k}{F^2(t)}\Bigg\{\int_0^t \log^2 G(x)\int_0^x \frac{f(y)}{1-P_1(y)}dydx + \int_0^t \left[\frac{F(x)}{G(x)}\right]^2 \int_0^x \frac{g(y)}{1-P_2(y)}dydx \\ &+ \frac{\mathfrak{h}^2(t)}{F^2(t)}\int_0^t \frac{f(x)}{1-P_1(x)}dx + \log^2 G(t)\int_0^t\int_0^x \frac{f(y)}{1-P_1(y)}dydx \\ &+ \frac{\log^2 G(t)\mathfrak{g}^2(t)}{F^2(t)}\int_0^t \frac{f(x)}{1-P_1(x)}dx + \frac{\mathfrak{g}^2(t)}{G^2(t)}\int_0^t \frac{g(x)}{1-P_2(x)}dx\Bigg\}.\end{aligned}$$

Proof: We have

$$\begin{aligned}\bar{\mathcal{A}}_{nt} - \bar{\mathcal{A}}_t &= -\frac{\mathfrak{h}_{\mathfrak{n}}(t)}{F_n(t)} + \frac{\mathfrak{h}(t)}{F(t)}, \\ &= -\frac{[\mathfrak{h}_{\mathfrak{n}}(t) - \mathfrak{h}(t)]}{F_n(t)} + \frac{\mathfrak{h}(t)[F_n(t) - F(t)]}{F(t)F_n(t)}, \\ &= -\frac{1}{F_n(t)}\int_0^t \Bigg\{\log G(x)\Big[F_n(x) - F(x)\Big] + \frac{F(x)}{G(x)}\Big[G_n(x) - G(x)\Big]\Bigg\}dx \\ &\quad + \frac{\mathfrak{h}(t)}{F(t)F_n(t)}\Big[F_n(x) - F(x)\Big].\end{aligned}$$

Using asymptotic normality and almost sure convergence properties of of $F_n(t)$ given in Cai [3], we get

$$(nh)^{\frac{1}{2}}\left[\frac{\bar{\mathcal{A}}_{nt} - \bar{\mathcal{A}}_t}{\sigma_{\bar{\mathcal{A}}}}\right]$$

asymptotically follows standard normal distribution with

$$\begin{aligned}\sigma^2_{\bar{\mathcal{A}}} &= \frac{C_k}{F^2(t)}\Bigg\{\int_0^t \log^2 G(x)\int_0^x \frac{f(y)}{1-P_1(y)}dydx + \int_0^t \left[\frac{F(x)}{G(x)}\right]^2 \int_0^x \frac{g(y)}{1-P_2(y)}dydx \\ &+ \frac{\mathfrak{h}^2(t)}{F^2(t)}\int_0^t \frac{f(x)}{1-P_1(x)}dx\Bigg\}.\end{aligned}$$

Similarly, we get

$$(nh)^{\frac{1}{2}}\left[\frac{\bar{\mathcal{B}}_{nt}-\bar{\mathcal{B}}_t}{\sigma_{\bar{\mathcal{B}}}}\right]$$

asymptotically follows standard normal distribution with

$$\sigma^2_{\bar{\mathcal{B}}} = \frac{C_k}{F^2(t)}\left\{\log^2 G(t)\left[\int_0^t\int_0^x \frac{f(y)}{1-P_1(y)}dydx + \frac{\mathfrak{g}^2(t)}{F^2(t)}\int_0^t \frac{f(x)}{1-P_1(x)}dx\right] + \frac{\mathfrak{g}^2(t)}{G^2(t)}\int_0^t \frac{g(x)}{1-P_2(x)}dx\right\}.$$

Hence the proof. □

In the following theorem, we check the almost sure convergence property of the suggested estimator.

Theorem 3.4. *Let $\bar{C}I_n(t)$ be a nonparametric estimator of $\bar{C}I(X,Y,t)$, suppose that $F(\cdot), G(\cdot), f(\cdot)$ and $g(\cdot)$ satisfy the Lipschitz conditions and the kernel $K(\cdot)$ satisfies the requirements and for $0<\tau<\infty$, the marginal distribution function of R satisfies $L(\tau)<1$ (see Cai [3]) then*

$$\sup_{0\le t\le\tau}\left|\bar{C}I_n(t)-\bar{C}I(X,Y,t)\right|\to 0 \quad \textit{almost surely.}$$

Proof: We have

$$\left|\bar{C}I_n(t)-\bar{C}I(X,Y,t)\right| \le \left|\bar{\mathcal{A}}_{nt}-\bar{\mathcal{A}}_t\right| + \left|\bar{\mathcal{B}}_{nt}-\bar{\mathcal{B}}_t\right|.$$

Also,

$$\begin{aligned}\left|\bar{\mathcal{A}}_{nt}-\bar{\mathcal{A}}_t\right| &= \left|\frac{\mathfrak{h}_{\mathfrak{n}}(t)}{F_n(t)}-\frac{\mathfrak{h}(t)}{F(t)}\right| \\ &\le \left|\frac{\mathfrak{h}_{\mathfrak{n}}(t)-\mathfrak{h}(t)}{F_n(t)}\right| + \left|\frac{\mathfrak{h}(t)}{F_n(t)F(t)}\right|\left|F_n(t)-F(t)\right| \\ &\le \frac{1}{F_n(t)}\int_0^t\left\{|\log G(x)||F_n(x)-F(x)| + \left|\frac{F(x)}{G(x)}\right||G_n(x)-G(x)|\right\}dx \\ &\quad + \left|\frac{\mathfrak{h}(t)}{F_n(t)F(t)}\right|\left|F_n(t)-F(t)\right|.\end{aligned}$$

Similarly, we get

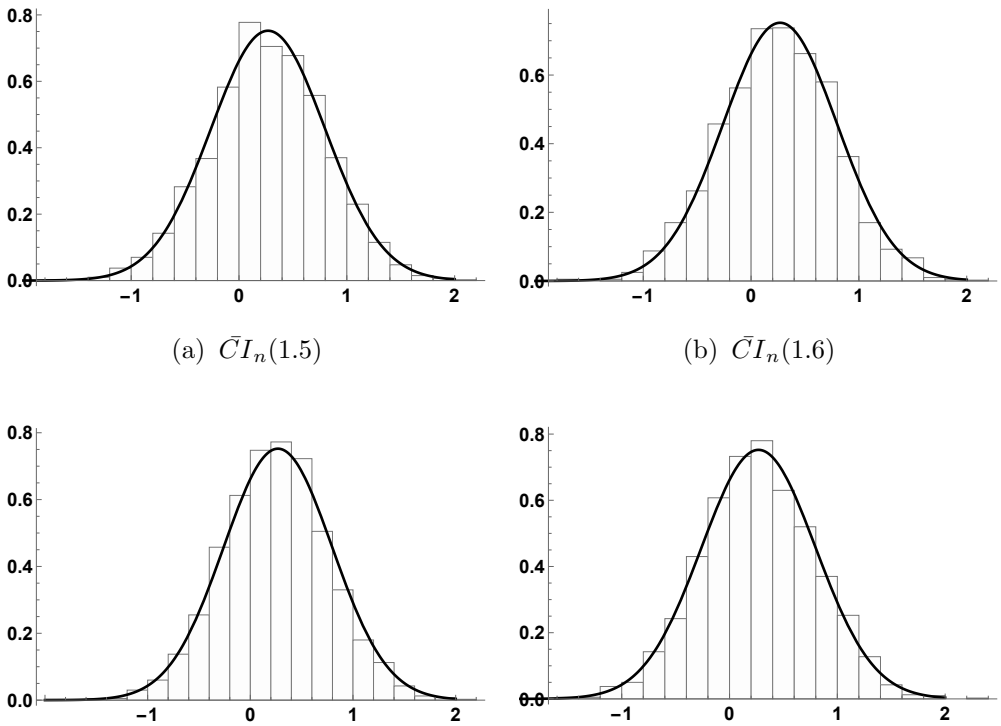
$$\begin{aligned}|\bar{\mathcal{B}}_{nt}-\bar{\mathcal{B}}_t| \le{}& \frac{\log G(t)}{F_n(t)}\left\{\int_0^t|F_n(x)-F(x)|dx + \left|\frac{\mathfrak{g}(t)}{F(t)}\right|\left|F_n(t)-F(t)\right|\right\} \\ &+ \frac{\mathfrak{g}(t)}{G(t)F_n(t)}|G_n(t)-G(t)|.\end{aligned}$$

By using the almost sure convergence of $G_n(t)$, $F_n(t)$ given in Cai [3], the proof immediately follows. □

4. NUMERICAL ANALYSIS

In this section, we simulate to evaluate the performance of the proposed estimators. We are interested in random variables of the form $U = \sqrt{(1-\rho)}|V|$, where V are generated from the AR(1) model to obtain dependent samples.

We generate two sets of 2000 simulated samples with white noise distributed as normal density and parameters of $(0, 1)$ and $(0, 2)$, respectively, to test the proposed estimator's asymptotic normality. Exponential distributions with parameters 1 and 2 are used for right censoring observations. The kernel function, in this case, is Epanechnikov, and it has the form $0.75(1-u^2)I(|u| < 1)$. The process is repeated 250 times, and the estimator's histogram with a normal curve is shown in Figure 2 for $t = 1.5$, 1.6, 1.7, and 1.8. We passed the AIC and BIC tests, indicating that the estimator has asymptotic normality.



(a) $\bar{C}I_n(1.5)$ (b) $\bar{C}I_n(1.6)$

(c) $\bar{C}I_n(1.7)$ (d) $\bar{C}I_n(1.8)$

Figure 2: Histogram of $\bar{C}I_n(t)$ with normal density curve from sample of size 250.

Table 1 shows the MISE of the estimator for varying parameter ρ, and the table values show that as sample size increases, the MISE approaches zero. In Table 1, we also compute the estimator's 95% confidence interval when $t = 0.6$ and ρ varies. We can conclude from the table values that the confidence interval width for these data sets decreases as the sample size increases.

Table 1: Comparison of MISE of $\bar{CI}_n(x)$ and 95% confidence interval of $\bar{CI}_n(0.6)$.

n	50		100	
ρ	MISE	95% CI	MISE	95% CI
−0.9	0.27742	(0.26248, 0.89153)	0.18867	(0.45588, 0.54162)
−0.6	0.17139	(0.41373, 0.60759)	0.12089	(0.48320, 0.57749)
−0.3	0.25134	(0.41558, 0.71311)	0.17854	(0.43049, 0.64481)
0	0.24360	(0.37707, 0.73257)	0.19241	(0.47052, 0.59493)
0.3	0.24152	(0.44217, 0.66990)	0.19816	(0.46348, 0.55320)
0.6	0.19935	(0.47671, 0.54945)	0.17724	(0.48481, 0.53651)
0.9	0.24180	(0.48576, 0.62959)	0.20584	(0.49064, 0.56981)

n	200		300	
ρ	MISE	95% CI	MISE	95% CI
−0.9	0.09656	(0.48291, 0.53108)	0.00513	(0.48801, 0.51252)
−0.6	0.08262	(0.49566, 0.55799)	0.00476	(0.49782, 0.53547)
−0.3	0.08787	(0.45501, 0.55358)	0.00449	(0.48234, 0.54069)
0	0.09106	(0.48313, 0.54717)	0.00346	(0.48922, 0.53103)
0.3	0.10085	(0.49187, 0.53199)	0.00132	(0.49807, 0.52636)
0.6	0.07573	(0.48991, 0.52993)	0.00648	(0.49629, 0.51122)
0.9	0.10222	(0.49436, 0.55430)	0.00670	(0.49809, 0.54510)

Figure 3 shows the proposed estimator's value and its upper and lower confidence bounds when $t = 1.5$ and $\rho = 0.5$. It is observed that $\bar{CI}_n(1.5)$ is not monotone for n for fixed values $t = 1.5$ and $\rho = 0.5$. Also, as sample size n increases, the width of the confidence intervals narrows, and both bounds approach the kernel estimator, which means the kernel estimator is more precise when the sample size becomes large.

Figure 3: $\bar{CI}_n(1.5)$ and confidence bounds when $\rho = 0.5$.

4.1. Comparison with an empirical estimator

Consider $\{X_i\}$, $\{Y_i\}$, $i = 1, 2, ..., n$ be identically distributed random samples have survival functions be $F(x) = Pr(X_i \geq x)$ and $G(x) = Pr(Y_i \geq x)$ respectively. We use independent and identically distributed random variable R_{1i} and R_{2i} with corresponding distribution functions $P_1(x)$ and $P_2(x)$ for creating right censored data from X_i and Y_i respectively. Note that R_{1i} and R_{2i} are independent of X_i and Y_i respectively. Let $C_i = \min(X_i, R_{1i})$ and $C_i^* = \min(Y_i, R_{2i})$. In this censoring scheme one can observe (C_i, δ_i) and (C_i^*, δ_i^*), where $\delta_i = I(X_i \leq R_{1i})$ and $\delta_i^* = I(Y_i \leq R_{2i})$. Denote $\{C_{i:n}\}_{1\leq i\leq n}$ and $\{C_{i:n}^*\}_{1\leq i\leq n}$, be the sample order statistics, where ties within lifetimes or within censoring times are ordered arbitrarily and ties among lifetimes and censoring times are treated as if the former precedes the latter. $\{\delta_{i:n}\}_{1\leq i\leq n}$ and $\{\delta_{i:n}^*\}_{1\leq i\leq n}$ are the concomitant of i^{th} order statistic of each sample. Let

$$Z_j = \sum_{r=1}^{n} I(C_i \leq C_{j:n}^*), \quad i = 1, 2, ..., n, \tag{4.1}$$

the number of random variables of the first censored sample that are less than or equal to j^{th} order statistics of the second censored sample. Also we rename by $C_{(j,1)} < C_{(j,2)} < ...$ the random sample of the first censored sample belonging to $(C_{j:n}^*, C_{(j+1):n}^*]$, if any. Then in the context of right censoring, we get the estimator of cumulative inaccuracy measure owing to Di Crescenzo and Longobardi (2013), as follows:

$$\begin{aligned} \bar{CI}_n^{\text{cen}}(t) &= -\frac{1}{n}\sum_{j=1}^{n-1}\left[\frac{Z_{j+1}C_{(j+1):n}^* - Z_jC_{j:n}^* - \sum_{k=1}^{Z_{j+1}-Z_j} C_{(j,k)}}{F_*(t)\sum_{r=1}^{n} I(R_{1r} > C_{j:n}^*)}\right] \\ &\quad \times \ln\left(\frac{j}{G_*(t)\sum_{r=1}^{n} I(R_{2r} > Y_{2j:n})}\right) I(Y_{1j:n} \leq t), \end{aligned} \tag{4.2}$$

where $F_*(t)$ and $G_*(t)$ are the Kaplan–Meier estimators of distribution functions $F(t)$ and $G(t)$ respectively defined as

$$F_*(x) = 1 - \prod_{1\leq i\leq n}\left(1 - \frac{\delta_{i:n}}{n-i+1}\right)^{I(C_{i:n}\leq t)} \text{ and } G_*(x) = 1 - \prod_{1\leq i\leq n}\left(1 - \frac{\delta_{i:n}^*}{n-i+1}\right)^{I(C_{i:n}^*\leq t)}. \tag{4.3}$$

Table 2 shows the results of comparing the proposed estimator with the empirical estimator using bias and MSE for varying t. We can conclude from these data sets that bias and MSE decrease with increasing sample size n and are inversely proportional to t.

Table 2: Comparison of |Bias| and MSE of the estimators $\bar{CI}_n(t)$ and $\bar{CI}_n^{\text{cen}}(t)$.

|Bias|						
n	100		200		300	
t	$\bar{CI}_n(t)$	$\bar{CI}_n^{\text{cen}}(t)$	$\bar{CI}_n(t)$	$\bar{CI}_n^{\text{cen}}(t)$	$\bar{CI}_n(t)$	$\bar{CI}_n^{\text{cen}}(t)$
1.1	0.17026	0.33860	0.09752	0.28616	0.02315	0.10148
1.3	0.19966	0.56255	0.11951	0.34930	0.03831	0.13926
1.5	0.22769	0.79647	0.14579	0.41002	0.05833	0.16947
1.7	0.28286	1.03660	0.16794	0.46752	0.08318	0.19762
1.9	0.29928	1.27998	0.18381	0.52152	0.11240	0.21901

MSE						
n	100		200		300	
t	$\bar{CI}_n(t)$	$\bar{CI}_n^{\text{cen}}(t)$	$\bar{CI}_n(t)$	$\bar{CI}_n^{\text{cen}}(t)$	$\bar{CI}_n(t)$	$\bar{CI}_n^{\text{cen}}(t)$
1.1	0.02933	0.11467	0.01043	0.09070	0.00157	0.01826
1.3	0.04012	0.31647	0.01442	0.13500	0.00350	0.05177
1.5	0.05189	0.63436	0.02134	0.18629	0.00552	0.08249
1.7	0.08305	1.07454	0.02827	0.24301	0.00761	0.11512
1.9	0.09261	1.63838	0.03387	0.30381	0.00973	0.13043

4.1.1. Real data Analysis

Example 4.1. We use 101 data points from Andrews and Herzberg [1], representing the stress-rupture life of kevlar 49/epoxy strands subjected to constant sustained pressure at 90% stress until all fail, giving us complete data with exact failure times. The fitting details for this data set are given in Table 3. We generated 100 bootstrap samples from the data sets, which were right censored by exponential models with parameters 0.09 and 0.03, respectively. Under the assumption that $F(x)$ is distributed as an Extreme value and $G(x)$ is distributed as a Weibull model, we plotted the mean values of the kernel and empirical estimators in Figure 4 using the Epanechnikov kernel. We also find the kernel estimator's relative efficiencies compared to the empirical estimator, which is plotted in Figure 6(a).

Example 4.2. We consider the 20 failure times of equal-load share samples from Table 1's sample 1, as investigated by Kim and Kvam [14]. We discovered better models for the data sets using AIC and BIC, shown in Table 3. We assumed that $F(x)$ is an exponential distribution and $G(x)$ is a Weibull distribution. We generated 100 bootstrap samples from the data sets, right censored by exponential models with parameters of 0.2 and 0.1, respectively. The Epanechnikov kernel is used as the kernel form. The mean values of empirical and kernel estimators are calculated and plotted in Figure 5. The relative efficiencies of the kernel estimator to the empirical estimator are also determined and plotted in Figure 6(b).

In Figures 4 and 5, we plot $\bar{CI}(X,Y,t)$, $\bar{CI}_n(t)$ and $\bar{CI}_n^{\text{cen}}(t)$ with respect to Examples 4.1 and 4.2 respectively. We can observe from Figure 4, $\bar{CI}(X,Y,t)$ and $\bar{CI}_n(t)$ are non decreasing and monotonic, while $\bar{CI}_n^{\text{cen}}(t)$ is non increasing in t. In Figure 5, $\bar{CI}(X,Y,t)$ and $\bar{CI}_n(t)$ are monotonic. Furthermore, in both cases, the kernel estimator outperforms the empirical estimator. We can conclude from Figures 6 that the relative efficiencies of kernel

estimators in comparison to empirical estimators are decreasing functions in t. Furthermore, the graph shows that relative efficiencies are greater than one, indicating that the kernel estimator outperforms the empirical estimator in the Examples 4.1 and 4.2.

Table 3: Fitting details of real data.

	Distribution	Parameters	AIC	BIC
Example 4.1:	Extreme value	(0.60869, 0.70011)	−2.48095	−2.49513
	Weibull	(0.94876, 1.12481, −0.07909)	−2.23536	−2.25600
	Pareto	(7, 5.61164,1.19461,−0.07909)	−2.13630	−2.16296
	Frechet	(2.50547, 1.36598, −0.89583)	−2.05503	−2.07566
Example 4.2:	Exponential	0.19053	−5.25114	−5.19792
	Weibull	(0.74337, 4.36222, 0.13999)	−4.86134	−4.66002
	Pareto	(7, 1.86990, 2.96583, 0.14000)	−4.402009	−4.11923
	Gamma	(0.83604, 6.27797)	−4.39639	−4.27688

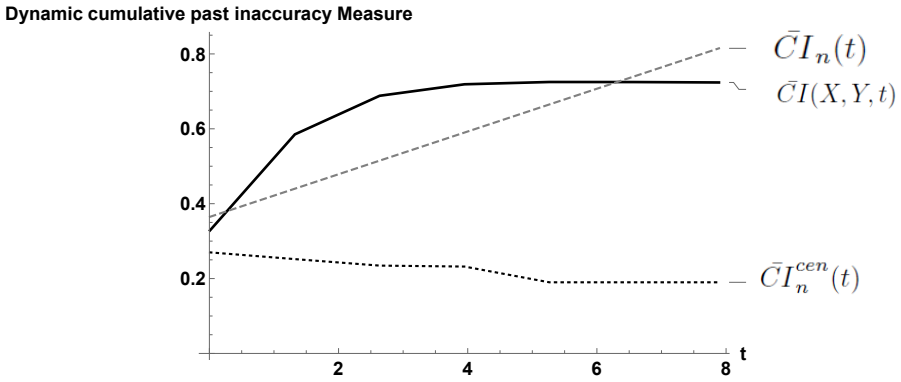


Figure 4: Comparison of $\bar{C}I(X,Y,t)$, $\bar{C}I_n(t)$ and $\bar{C}I_n^{\text{cen}}(t)$ for the stress-rupture lives of kevlar 49/epoxy strands.

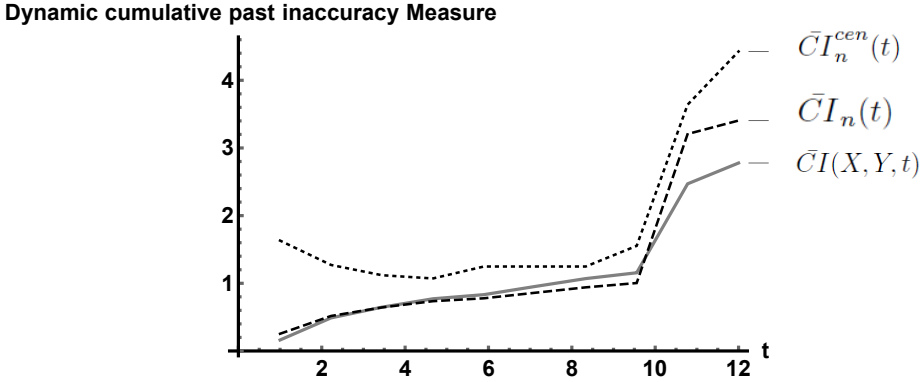
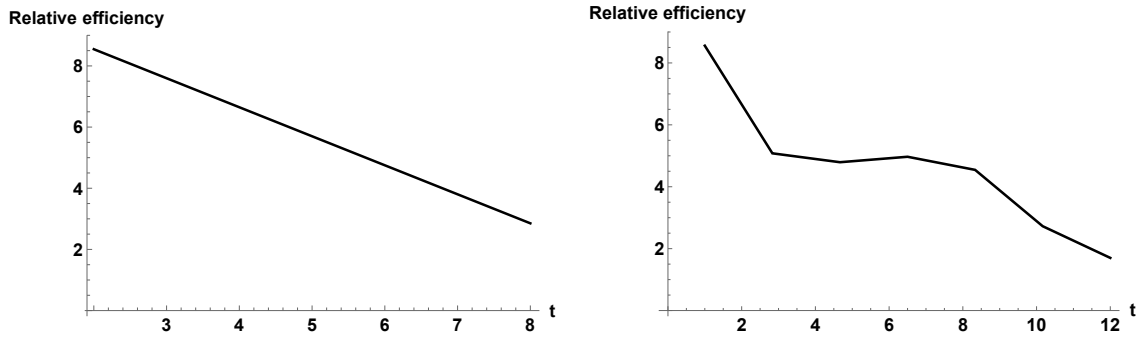


Figure 5: Comparison of $\bar{C}I(X,Y,t)$, $\bar{C}I_n(t)$ and $\bar{C}I_n^{\text{cen}}(t)$ for the 20 failure times of equal-load share samples.



(a) Example 4.1 (b) Example 4.2

Figure 6: Comparison of $\bar{CI}_n(t)$'s relative efficiency with respect to $\bar{CI}_n^{\text{cen}}(t)$.

ACKNOWLEDGMENTS

We thank the referees for their attentive reading of the paper and helpful suggestions, which enhanced the manuscript's presentation.

REFERENCES

[1] Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, Springer, New York.

[2] Bhatia, P.K. and Taneja, H.C. (1993). On quantitative-qualitative measure of inaccuracy reversible symmetry, *Information Sciences*, **67**, 277–282.

[3] Cai, Z. (1998). Kernel density and hazard rate estimation for censored dependent data, *Journal of Multivariate Analysis*, **67**, 23–34.

[4] Di Crescenzo, A. and Longobardi, M. (2013). *Stochastic Comparisons of Cumulative Entropies*. In "Stochastic Orders in Reliability and Risk" (H. Li and X. Li, Eds.), Springer, New York, 168–182.

[5] Elias Masry (1986). Recursive Probability Density Estimation for Weakly Dependent Stationary Processes, *IEE Transactions on Information Theory*, **32**(2), 254–267.

[6] Ghosh, A. and Kundu, C. (2018). On generalized conditional cumulative past inaccuracy measure, *Applications of Mathematics*, **63**(2), 167–193.

[7] Goel, R.; Taneja, H.C. and Kumar, V. (2018). Kerridge measure of inaccuracy for record statistics, *Journal of Information and Optimization Sciences*, **39**, 1149–1161.

[8] Gur Dial (1987). On non-additive measures of inaccuracy and a coding theorem, *Journal of Information and Optimization Sciences*, **8**(1), 113–118.

[9] Hooda, D.S. and Tuteja, R.K. (1985). On characterization of non-additive measures of relative information and inaccuracy, *Bulletin Calcutta Mathematical Society*, **77**, 363–369.

[10] James, L.S. Jr. and Anita, R. (2006). Modeling international consumption patterns, *Review of Income and Wealth*, **52**, 603–624.

[11] Kayal, S. and Sunoj, S.M. (2017). Generalized kerridge's inaccuracy measure for conditionally specified models, *Communications in Statistics - Theory and Methods*, **46**, 8257–8268.

[12] Kerridge, D.F. (1961). Inaccuracy and inference, *Journal of the Royal Statistical Society, Series B*, **23**, 184–194.

[13] Khorashadizadeh, M. (2018). More Results on Dynamic Cumulative Inaccuracy Measure, *Journal of Iranian Statistical Society*, **17**(1), 89–108.

[14] Kim, H. and Kvam P.H. (2004). Reliability Estimation Based on System Data with an Unknown Load Share Rule, *Lifetime Data Analysis*, **10**, 83–94.

[15] Kumar, V. and Taneja, H.C. (2015). Dynamic Cumulative Residual and Past Inaccuracy Measures, *Journal of Statistical Theory and Applications*, **14**(4), 399–412.

[16] Kundu, C.; Di Crescenzo, A. and Longobardi, M. (2016). On cumulative residual (past) inaccuracy for truncated random variables, *Metrika*, **79**, 335–356.

[17] Rajesh, G.; Sathar, A.E.I. and Viswakala, K.V. (2017). Estimation of inaccuracy measure for censored dependent data, *Communications in Statistics – Theory and Methods*, **46**(20), 10058–10070.

[18] Rao, B.L.S.P. (1983). *Nonparametric Functional Estimation*, Academic Press, New York.

[19] Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition, *Proceedings of the National Academy of Sciences of the United States of America*, **42**(1), 43–47.

[20] Sathar, E.I.A.; Viswakala, K.V. and Rajesh, G. (2021). Estimation of past inaccuracy measure for the right censored dependent data, *Communications in Statistics – Theory and Methods*, **50**(6), 1446–1455.

REVSTAT – Statistical Journal
Volume 22, Number 4, October 2024, 545–570
https://doi.org/10.57805/revstat.v22i4.543

Bootstrapping Order Statistics with Variable Rank

Authors: M. E. Sobh
– Department of Mathematics, Faculty of Science, Mansoura University, Egypt
mohamedebrahim2014@mans.edu.eg
m.ebraheem160@yahoo.com

H. M. Barakat ✉
– Department of Mathematics, Faculty of Science, Zagazig University, Egypt
hmbarakat@hotmail.com

Received: June 2021 Revised: December 2022 Accepted: December 2022

Abstract:

- This work investigates the strong consistency of bootstrapping central and intermediate order statistics for an appropriate choice of re-sample size for known and unknown normalizing constants. We show that when the normalizing constants are estimated from the data, the bootstrap distribution for central and intermediate order statistics may be weakly or strongly consistent. A simulation study is conducted to show numerically how to choose the bootstrap sample size to give the best approximation of the bootstrapping distribution for the central and intermediate quantiles.

Keywords:

- *bootstrap technique; central order statistics; intermediate order statistics; weak consistency; strong consistency.*

AMS Subject Classification:

- 62G30, 62F40.

✉ Corresponding author

1. INTRODUCTION

Let $X_1, X_2, ..., X_n$ be iid random variables (RVs) with a common distribution function (DF) $F(x)$ and let $X_{1:n}, X_{2:n}, ..., X_{n:n}$ be the corresponding order statistics. The DF of the k-th order statistic $X_{k:n}$, $1 \leq k \leq n$, is given by

$$F_{k:n}(x) = P(X_{k:n} \leq x) = B_{F(x)}(k, n-k+1), \tag{1.1}$$

where $B_x(a,b)$ is the usual incomplete beta function with the shape parameters $a, b > 0$ (cf. David and Nagaraja [19]). A sequence $\{X_{k:n}\}$ is called a sequence of order statistics with variable rank (cf. [3]) if $1 < k_n < n$ and $\min\{k_n, n-k_n\} \to \infty$, as $n \to \infty$ (denoted by $\min\{k_n, n-k_n\} \underset{n}{\longrightarrow} \infty$), where we have the following two cases:

1. If $\frac{k_n}{n} \underset{n}{\longrightarrow} 0$ (or $\frac{k_n}{n} \underset{n}{\longrightarrow} 1$), then $X_{k_n:n}$ is called the lower intermediate order statistic (or the upper intermediate order statistics);
2. If $\frac{k_n}{n} \underset{n}{\longrightarrow} p$ $(0 < p < 1)$, then $X_{k_n:n}$ is called the central order statistic.

A prominent example for the central order statistics is the p-th sample quantile (including the median, quartiles, percentiles etc.), where $k_n = [np] + 1$ and $[\cdot]$ is the greatest integer function (see David and Nagaraja [19]). On the other hand, the intermediate order statistics also have many applications, e.g., they can be used to estimate the probabilities of the future extremes and tail quantiles of the underlying distribution that are extremely relative to the available sample size, cf. [33]. Moreover, many authors, e.g., Mason [30] and Teugels [36] have also found estimates that are based, in part, on intermediate order statistics.

The literature abounds with many different results for intermediate and central order statistics and their applications. Interested readers may refer to Balkema and de Haan [3, 4], Barakat [5, 6], Barakat and El-Shandidy [7], Barakat and Omar [8, 9], Chibisov [18], Falk [21], Falk and Wisheckel [22], Frey and Zhang [23], Ho and Lee [27], Nagaraja and Nagaraja [31], Peng and Yang [32], Smirnov [35], and Wu [37]. The bootstrap method introduced in Efron [20] is a general procedure for approximating the sampling distributions of statistics based on re-sampling from the data at hand. There are several forms of the bootstrap and additionally several other re-sampling methods that are related to it, such as jackknifing, cross-validation, randomization tests, and permutation tests. The bootstrap method is shown to be successful in many situations and is accepted as an alternative to the asymptotic methods (for more details, see [14] and [31]). Let $X_n = (X_1, X_2, ..., X_n)$ be a random sample from an unknown DF $F(x)$. For $m = m(n) \underset{n}{\longrightarrow} \infty$, assume that $Y_i, i = 1, 2, ..., m$, are conditionally iid RVs with distribution

$$P(Y_i = X_j | X_n) = \frac{1}{n}, j = 1, 2, ..., n, \ i \in \{1, 2, ..., m\},$$

then $(Y_1, Y_2, ..., Y_m)$ is a random re-sample of size m from the empirical DF

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x)}(X_i) = \frac{1}{n} Q_n(x),$$

where $I_A(x)$ is the indicator function and $Q_n(x)$ is an RV distributed as a binomial distribution with parameters n and $F(x)$, denoted by $Q_n(x) \sim \mathrm{B}(n, F(x))$. Furthermore, let the

extreme value theory (see [14]) be applicable to the extreme order statistic $X_{k:n}$, which means that there exist normalizing constants $a_n > 0$ and b_n such that $F_{k:n}(a_n x + b_n) = B_{F(a_n x+b_n)}(k, n-k+1)$ weakly converges, as $n \to \infty$ (denoted by $\xrightarrow[n]{\text{w}}$) to a non-degenerate DF $G(x)$, where $G(x)$ is one of the extreme value distributions. Now, let $Y_{1:m}, Y_{2:m}, ..., Y_{m:m}$ be the corresponding order statistics of $Y_1, Y_2, ..., Y_m$, and define

$$H_{n,m}(a_m x + b_m) = P(Y_{k:m} \leq a_m x + b_m | X_n) = B_{F_n(a_m x+b_m)}(k, m-k+1).$$

$H_{n,m}(a_m x + b_m)$ is called the bootstrap distribution of $a_n^{-1}(X_{k:n} - b_n)$, where n and m are the sample size and re-sample size, respectively. A full-sample bootstrap is the case when $m = n$. In contrast, m out of n bootstrap technique is the case when $m < n$. One of the bootstrap's desired properties is consistency; namely, the bootstrap's limit distribution is the same as the original statistic distribution. For a long time, it has been known that a full-sample bootstrap does not work for order statistics. This seminal result was apparently first revealed for extremes by Athreya and Fukuchi [1] and Fukuchi [24]. Moreover, it was proved for intermediate order statistics by Geluk and de Haan [25] and Barakat *et al.* [16]. Finally, for central order statistics, this result was proved by Barakat *et al.* [16]. Athreya and Fukuchi [1] and Fukuchi [24] (see also Athreya and Fukuchi [2]) studied the consistency of bootstrapping extremes for known and unknown normalizing constants and they showed that the bootstrap DF fails to be consistent in the full-sample bootstrap case. Moreover, they showed that the bootstrap DF is a weakly consistent estimate if $m = o(n)$ and it is strongly consistent if $m = o(\frac{n}{\log n})$. Barakat *et al.* [11] extended this result to the extreme generalized order statistics. Later, Barakat *et al.* [16] have got some similar results for the order statistics with variable ranks. Namely, they showed that the bootstrapping central and intermediate quantiles fail to be consistent in the full-sample bootstrap case. Moreover, they also showed that when the normalizing constants are known, the bootstrap DFs for central and intermediate order statistics are weakly consistent when $m = o(n)$ (see, Theorems 4.1 and 4.2 in [16]). Barakat *et al.* [13] extended this result to the case where we use the bootstrap for estimating a central, or an intermediate quantile under power normalization.

The main aim of the present work is to extend the results of [16] by investigating the strong consistency of bootstrapping central and intermediate order statistics for an appropriate choice of re-sample size for known and unknown normalizing constants. A simulation study is conducted to illustrate how to choose the re-sample's size. Sections 2 and 3 are devoted respectively to the intermediate and central order statistics, while the simulation study is conducted in Section 4. Finally, we conclude the paper in Section 5. The rest of this introductory section will be devoted to review some basic results pertaining to the asymptotic behaviour of the central and intermediate order statistics, which are the essential pillars of our study.

1.1. Some important aspects of the asymptotic theory of order statistics with variable rank

The following lemma (Lemma 1.1 in Barakat [6]) is a cornerstone of the asymptotic theory of order statistics with variable rank.

Lemma 1.1 (cf. [6], see also [28]). *For any sequence of variable ranks $\{k_n\}$, let $\{u_n, n \geq 1\}$ be a sequence of real numbers and let $-\infty \leq \tau \leq \infty$. Then,*

$$F_{k_n:n}(u_n) = P(X_{k_n:n} \leq u_n) \underset{n}{\longrightarrow} \mathcal{N}(\tau), \tag{1.2}$$

if and only if

$$\frac{nF(u_n) - k_n}{\sqrt{k_n(1 - \frac{k_n}{n})}} \underset{n}{\longrightarrow} \tau, \tag{1.3}$$

where $\mathcal{N}(\cdot)$ is the standard normal DF and $F_{k_n:n}$ is defined in (1.1).

Since the variable ranks are classified into central and intermediate ranks, we will consider each of the two cases separately.

1.1.1. Asymptotic theory of the intermediate order statistics

If $\frac{k_n}{n} \underset{n}{\longrightarrow} 0$ (i.e., the lower intermediate case), then by using the linear parametrization's transformation $u_n = a_n x + b_n$ and $\tau = U(x)$, (1.3) will be reduced to

$$\frac{nF(a_n x + b_n) - k_n}{\sqrt{k_n}} \underset{n}{\longrightarrow} U(x), \tag{1.4}$$

(cf. [18]). A sequence of intermediate rank $\{k_n\}$ is said to satisfy the Chibisov's condition ([18]), if

$$\sqrt{k_{n+z_n(\nu)}} - \sqrt{k_n} \underset{n}{\longrightarrow} \frac{\alpha \nu l}{2}, \tag{1.5}$$

for any sequence of integer values $z_n(\nu)$, with $\frac{z_n(\nu)}{n^{1-\frac{\alpha}{2}}} \underset{n}{\longrightarrow} \nu$, where $0 < \alpha < 1$, $l > 0$, and ν is any real number. Chibisov [18] showed that, whenever $\{k_n\}$ satisfies the condition (1.5), the only possible non-degenerate forms for $\mathcal{N}(U(x))$ in (1.2) are $\mathcal{N}(U_{i;\beta}(x))$, $i = 1, 2, 3$, where $U_{3;\beta}(x) = U_3(x) = x$, $\forall x$,

$$U_{2;\beta}(x) = \begin{cases} -\beta \log |x|, & x \leq 0, \\ \infty, & x > 0, \end{cases} \qquad U_{1;\beta}(x) = \begin{cases} -\infty, & x \leq 0, \\ \beta \log x, & x > 0, \end{cases}$$

and β is a positive constant depending only on α, l and the type of the DF $F(x)$. Chibisov [18] noted that, the condition (1.5) implies $\frac{k_n}{n^\alpha} \underset{n}{\longrightarrow} l^2$. On the other hand, Barakat and Omar [8] showed that the last condition implies the Chibisov's condition, which means that the Chibisov rank sequences are widely-used and the Chibisov's limit types are vastly applicable. Recently, Barakat *et al.* [12] characterized the asymptotic behaviour of the scale normalizing constant a_n.

Lemma 1.2 ([12]). *Let $L(n) = \exp(\sqrt{n})$. Furthermore, let $F(a_n x + b_n) \in D^{(l,\alpha)}(\mathcal{N}(U(x)))$ mean that (1.4) is satisfied for $k_n \sim l^2 n^\alpha$. Then, for any $\varepsilon > 0$,*

1. $a_n L^{\frac{1}{\beta}+\varepsilon}(k_n) \underset{n}{\longrightarrow} \infty$ *and* $a_n L^{\frac{1}{\beta}-\varepsilon}(k_n) \underset{n}{\longrightarrow} 0$, *if* $F(a_n x + b_n) \in D^{(l,\alpha)}(\mathcal{N}(U_{1;\beta}(x)))$;
2. $a_n L^{\frac{-1}{\beta}+\varepsilon}(k_n) \underset{n}{\longrightarrow} \infty$ *and* $a_n L^{\frac{-1}{\beta}-\varepsilon}(k_n) \underset{n}{\longrightarrow} 0$, *if* $F(a_n x + b_n) \in D^{(l,\alpha)}(\mathcal{N}(U_{2;\beta}(x)))$;
3. $a_n L^{+\varepsilon}(k_n) \underset{n}{\longrightarrow} \infty$ *and* $a_n L^{-\varepsilon}(k_n) \underset{n}{\longrightarrow} 0$, *if* $F(a_n x + b_n) \in D^{(l,\alpha)}(\mathcal{N}(U_3(x)))$.

1.1.2. Asymptotic theory of the central order statistics

Smirnov [35] revealed that it is possible to find two rank sequences $\{k_n\}$ and $\{k_n^\star\}$ with $\frac{k_n}{n} \sim \frac{k_n^\star}{n} \sim p$, $0 < p < 1$, to lead to different non-degenerate limiting DFs for $X_{k_n:n}$ and $X_{k_n^\star:n}$. However, this is not possible if $k_n \sim k_n^\star \sim pn + o(\sqrt{n})$. Under this condition, Smirnov [35] showed that with $u_n = a_n x + b_n$ and $\tau = V(x)$, (1.3) will be reduced to

$$\sqrt{n}\frac{F(a_n x + b_n) - p}{C_p} \xrightarrow[n]{} V(x), \tag{1.6}$$

where $C_p = \sqrt{p(1-p)}$. Smirnov [35] showed that, whenever $\{k_n\}$ satisfies the condition $k_n \sim pn + o(\sqrt{n})$, the only possible non-degenerate forms for $\mathcal{N}(V(x))$ in (1.2) are $\mathcal{N}(V_{i;\beta}(x))$, $i = 1, 2, 3, 4$, where

$$V_{1;\beta}(x) = \begin{cases} -\infty, & x \le 0, \\ cx^\beta, & x > 0, \end{cases} \qquad V_{2;\beta}(x) = \begin{cases} -c\,|\,x\,|^\beta, & x \le 0, \\ \infty, & x > 0, \end{cases}$$

$$V_{3;\beta}(x) = \begin{cases} -c_1\,|\,x\,|^\beta, & x \le 0, \\ c_2 x^\beta, & x > 0, \end{cases} \qquad V_{4;\beta}(x) = W_4(x) = \begin{cases} -\infty, & x \le -1, \\ 0, & -1 < x \le 1, \\ \infty, & x > 1, \end{cases}$$

$c = c_1 = \frac{1}{\sqrt{p(1-p)}}$, $c_2 = \frac{c_1}{A}$, and $A > 0$. In this case, we say that the DF F belongs to the domain of normal p-attraction of the limit type $V_{i;\beta}(x)$, $i \in \{1, 2, 3, 4\}$, and we write $F \in D^{(p)}(V_{i;\beta}(x))$.

2. BOOTSTRAPPING INTERMEDIATE ORDER STATISTICS

In this section, we investigate the strong consistency of the bootstrap distribution $H_{n,m}(a_m x + b_m) = P(X_{k_m:m} \le a_m x + b_m | X_n)$, where k_n is the Chibisov rank sequence, which satisfies the condition (1.5), and the condition (1.4) is satisfied with $U(x) = U_{i;\beta}(x)$, $i \in \{1, 2, 3\}$, for some suitable normalizing constants $a_n > 0$ and b_n.

2.1. Almost sure consistency of bootstrapping intermediate for known normalizing constants

Barakat *et al.* [16] proved the weak limit relation $\sup_{x\in\mathbb{R}} |H_{n,m}(a_m x + b_m) - \mathcal{N}(U_{i;\beta}(x))| \xrightarrow[n]{\text{p}} 0$, if $m = o(n)$, where "$\xrightarrow[n]{\text{p}}$" stands for convergence in probability, as $n \to \infty$. The following theorem extends this result.

Theorem 2.1. *Let m be chosen such that $\sum_{n=1}^{\infty} \lambda^{\sqrt{\frac{n}{m}}} < \infty$, for each $\lambda \in (0, 1)$. Then,*

$$\sup_{x\in\mathbb{R}} |H_{n,m}(a_m x + b_m) - \mathcal{N}(U_{i;\beta}(x))| \xrightarrow[n]{\text{w.p.1}} 0,$$

where the symbol "$\xrightarrow[n]{\text{w.p.1}}$" denotes the convergence with probability one (almost surely convergence) as $n \to \infty$.

Proof: Let $\bar{k}_n = \frac{k_n}{n}$. Then, we have

$$\frac{mF_n(a_mx+b_m)-k_m}{\sqrt{k_m}} = \sqrt{m}\frac{F_n(a_mx+b_m)-\bar{k}_m}{\sqrt{\bar{k}_m}} = \sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-n\bar{k}_m}{\sqrt{n\bar{k}_m}}\right)$$
$$= \sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right)+\frac{mF(a_mx+b_m)-k_m}{\sqrt{k_m}}.$$

On the other hand, from the assumptions of the theorem, we get

$$\frac{mF(a_mx+b_m)-k_m}{\sqrt{k_m}} \xrightarrow[n]{} U_{i;\beta}(x),\ i\in\{1,2,3\}.$$

Thus, to prove the theorem, we only need to show that

$$\sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right) \xrightarrow[n]{\text{w.p.1}} 0.$$

By Borel–Cantelli lemma, it is enough to prove that

$$\sum_{n=1}^{\infty} P\left(\sqrt{\frac{m}{n}}\left|\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right|>\epsilon\right)<\infty,$$

for every $\epsilon>0$. Now for each $\theta>0$ we get

$$\sqrt{\frac{m}{n}}\log P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right)>\epsilon\right)=$$

$$\sqrt{\frac{m}{n}}\log P\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}>\sqrt{\frac{n}{m}}\,\epsilon\right)=\sqrt{\frac{m}{n}}\log P\Big(e^{\theta T_{n,m}}>e^{\theta\sqrt{\frac{n}{m}}\,\epsilon}\Big),$$

where

$$T_{n,m}=\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}.$$

By using the Markov inequality, we get

$$\sqrt{\frac{m}{n}}\log P\Big(e^{\theta T_{n,m}}>e^{\theta\sqrt{\frac{n}{m}}\,\epsilon}\Big)\le\sqrt{\frac{m}{n}}\log\left(e^{-\theta\sqrt{\frac{n}{m}}\,\epsilon}E\Big(e^{\theta T_{n,m}}\Big)\right)=$$

$$\sqrt{\frac{m}{n}}\left(-\sqrt{\frac{n}{m}}\theta\epsilon+\log\varphi_m(\theta)\right)=-\theta\epsilon+\sqrt{\frac{m}{n}}\log\varphi_m(\theta)\xrightarrow[n]{}-\theta\epsilon,$$

where $\varphi_m(\theta)$ is the moment generating function for the standard normal DF. Therefore, for sufficiently large n, we get the following relation:

$$\sum_{n=1}^{\infty} P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right)>\epsilon\right)=$$

$$\sum_{n=1}^{\infty}\exp\left\{\log P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right)>\epsilon\right)\right\}\le\sum_{n=1}^{\infty}e^{-\theta\epsilon\sqrt{\frac{n}{m}}}<\infty,$$

for every $\epsilon>0$, since the condition $\sum_{n=1}^{\infty}\lambda^{\sqrt{\frac{n}{m}}}<\infty$, for each $\lambda\in(0,1)$, guarantees the convergence of the infinite series $\sum_{n=1}^{\infty}\exp\{-\theta\epsilon\sqrt{\frac{n}{m}}\}$, for every $\epsilon>0$. By similar reasoning we can show that

$$\sum_{n=1}^{\infty} P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(a_mx+b_m)-nF(a_mx+b_m)}{\sqrt{n\bar{k}_m}}\right)<-\epsilon\right)<\infty,$$

for every $\epsilon>0$. The theorem is proved. □

2.2. Almost sure consistency of bootstrapping intermediate for unknown normalizing constants

If the DF F is unknown, the normalizing constants a_m and b_m need to be estimated from the sample data for $H_{n,m}(\cdot)$ to be of use. Let $\hat{a}_m$ and $\hat{b}_m$ be estimators of a_m and b_m based on $X_n = (X_1, X_2, ..., X_n)$. Define the bootstrap distribution for the normalized intermediate order statistic $a_n^{-1}(X_{k_n:n} - b_n)$ with the estimated normalizing constants by

$$\hat{H}_{n,m}(\hat{a}_m x + \hat{b}_m) = P\Big(Y_{k_m:m} \le \hat{a}_m x + \hat{b}_m | X_n\Big).$$

Fukuchi [24] provided some sufficient conditions for the bootstrap distribution of the maximum order statistics to be consistent. The following theorem extends the Fukuchi's result by providing sufficient conditions for $\hat{H}_{n,m}(\hat{a}_m x + \hat{b}_m)$ to be consistent.

Theorem 2.2. *Let $m = m(n)$. Then,*

$$\sup_{x\in\mathbb{R}} |\hat{H}_{n,m}(\hat{a}_m x + \hat{b}_m) - \mathcal{N}(U_{i;\beta}(x))| \xrightarrow[n]{\text{w.p.1}} 0, \ \ i = 1, 2, 3,$$

if (**i**) $H_{n,m}(x) \xrightarrow[n]{\text{w.p.1}} \mathcal{N}(U_{i;\beta}(x))$,

(**ii**) $\dfrac{\hat{a}_m}{a_m} \xrightarrow[n]{\text{w.p.1}} 1$,

and

(**iii**) $\dfrac{\hat{b}_m - b_m}{a_m} \xrightarrow[n]{\text{w.p.1}} 0.$

Moreover, this theorem holds if "$\xrightarrow[n]{\text{w.p.1}}$" is replaced by "$\xrightarrow[n]{\text{p}}$".

Proof: First, we note that (i) is equivalent to

$$\sqrt{m}\frac{F_n(a_m x + b_m) - \bar{k}_m}{\sqrt{\bar{k}_m}} \xrightarrow[n]{\text{w.p.1}} U_{i;\beta}(x).$$

Moreover, for every $\epsilon > 0$, the relations (ii) and (iii) imply

$$(1-\epsilon)a_m < \hat{a}_m < (1+\epsilon)a_m \tag{2.1}$$

and

$$b_m - \epsilon a_m < \hat{b}_m < b_m + \epsilon\, a_m, \tag{2.2}$$

respectively. Now, fix $x > 0$, the relations (2.1) and (2.2) yield

$$\begin{aligned}\limsup_{n\to\infty} \sqrt{m}\frac{F_n(\hat{a}_m x + \hat{b}_m) - \bar{k}_m}{\sqrt{\bar{k}_m}} &\le \limsup_{n\to\infty} \sqrt{m}\frac{F_n((1+\epsilon)x + \epsilon)a_m + b_m) - \bar{k}_m}{\sqrt{\bar{k}_m}} \\ &\le U_{i;\beta}((1+\epsilon)x + \epsilon).\end{aligned}$$

By a similar way we can prove that

$$\liminf_{n\to\infty} \sqrt{m}\frac{F_n(\hat{a}_m x + \hat{b}_m) - \bar{k}_m}{\sqrt{\bar{k}_m}} \ge U_{i;\beta}((1-\epsilon)x - \epsilon).$$

Since $U_{i;\beta}(x)$ is continuous, we get

$$\lim_{n\to\infty}\sqrt{m}\frac{F_n(\hat{a}_m x+\hat{b}_m)-\bar{k}_m}{\sqrt{\bar{k}_m}}=U_{i;\beta}(x).$$

By a similar argument, the same limit relation can easily be proved for $x<0$. Thus, $H_{n,m}(\hat{a}_m x+\hat{b}_m)\xrightarrow[n]{\text{w.p.1}}\mathcal{N}(U_{i;\beta}(x))$. Now, suppose that the conditions (i), (ii), and (iii) hold in probability. Then, for any subsequence $\{n_i\}_{i=1}^{\infty}$ of $\{n\}_{n=1}^{\infty}$, there exists a further subsequence $\{n_i^{\star}\}_{i=1}^{\infty}$ such that (i), (ii), and (iii) hold w.p.1. Then, by applying the first part of the theorem, we get

$$\sup_{x\in\mathbb{R}}|\hat{H}_{n_i^{\star},m(n_i^{\star})}(\hat{a}_{m(n_i^{\star})}x+\hat{b}_{m(n_i^{\star})})-\mathcal{N}(U_{i;\beta}(x))|\xrightarrow[n]{\text{w.p.1}}0.$$

The theorem is established. □

Now, for the bootstrap distribution $\hat{H}_{n,m}(\hat{a}_m x+\hat{b}_m)$ to be consistent, we need to choose $\hat{a}_m$ and $\hat{b}_m$ satisfying the conditions (ii) and (iii) in Theorem 2.2. Since a_m and b_m are functionals of F then, the natural choices of $\hat{a}_m$ and $\hat{b}_m$ are the empirical counter parts of a_m and b_m. In the next theorem, we give appropriate choices for $\hat{a}_m$ and $\hat{b}_m$ for each domain of attraction of $\mathcal{N}(U_{i;\beta}(x)), i=1,2,3$.

Theorem 2.3. *Let $k_n^{'}=\frac{n}{m}k_m$, $k_n^{''}=\frac{n}{m}(k_m+\sqrt{k_m})$, and x_0 be the left endpoint of F (i.e., $x_0=\inf\{x:F(x)>0)$. Then,*

- **(i)** *if $F(a_n x+b_n)\in D^{(l,\alpha)}(\mathcal{N}(U_{1;\beta}(x)))$, $\hat{a}_m=F_n^{-1}\left(\frac{k_m}{m}\right)-\hat{x}_0=X_{k_n^{'}:n}-X_{k_n:n}$, and $\hat{b}_m=X_{k_n:n}$, where $\hat{x}_0=X_{k_n:n}$ is an estimator for x_0;*
- **(ii)** *if $F(a_n x+b_n)\in D^{(l,\alpha)}(\mathcal{N}(U_{2;\beta}(x)))$, $\hat{a}_m=-F_n^{-1}\left(\frac{k_m}{m}\right)=-X_{k_n^{'}:n}$, and $\hat{b}_m=0$;*
- **(iii)** *if $F(a_n x+b_n)\in D^{(l,\alpha)}(\mathcal{N}(U_3(x)))$, $\hat{a}_m=F_n^{-1}\left(\frac{k_m+\sqrt{k_m}}{m}-\frac{k_m}{m}\right)=X_{k_n^{''}:n}-X_{k_n^{'}:n}$, and $\hat{b}_m=F_n^{-1}\left(\frac{k_m}{m}\right)=X_{k_n^{'}:n}$.*

If $m=o(n)$, then

$$\sup_{x\in\mathbb{R}}\left|\hat{H}_{n,m}(\hat{a}_m x+\hat{b}_m)-\mathcal{N}(U_{i;\beta}(x))\right|\xrightarrow[n]{\text{p}}0. \tag{2.3}$$

Moreover, if $\sum_{n=1}^{\infty}\lambda^{\sqrt{\frac{n}{m}}}<\infty$, for each $\lambda\in(0,1)$, then (2.3) holds w.p.1.

Proof: First, let $F(a_n x+b_n)\in D^{(l,\alpha)}(\mathcal{N}(U_{1;\beta}(x)))$. Therefore, in view of the result of Chibisov [18], we have $b_n=b_m=x_0>-\infty$. In order to apply Parts (ii) and (iii) in Theorem 2.2, it suffices to show that

$$\frac{\hat{a}_m}{a_m}=\frac{X_{k_n^{'}:n}-X_{k_n:n}}{a_m}\xrightarrow[n]{}1 \tag{2.4}$$

and

$$\frac{\hat{b}_m-b_m}{a_m}=\frac{X_{k_n:n}-x_0}{a_m}\xrightarrow[n]{}0, \tag{2.5}$$

both in probability or w.p.1. First, let us focus on the case of convergence in probability.

Now, we have

$$\frac{\hat{a}_m}{a_m} = \frac{X_{k_n':n} - x_0}{a_m} - \frac{a_n}{a_m} \times \frac{X_{k_n:n} - x_0}{a_n}, \quad \frac{\hat{b}_m - b_m}{a_m} = \frac{a_n}{a_m} \times \frac{X_{k_n:n} - x_0}{a_n},$$

and $P(\frac{X_{k_n:n}-x_0}{a_n} \leq x) \xrightarrow[n]{\text{w}} \mathcal{N}(U_{1;\beta}(x))$, where $\mathcal{N}(U_{1;\beta}(x))$ is a non-degenerate DF. Therefore, to prove (2.4) and (2.5), it is sufficient to show that

$$\frac{X_{k_n':n} - x_0}{a_m} \xrightarrow[n]{\text{p}} 1 \tag{2.6}$$

and

$$\frac{a_n}{a_m} \xrightarrow[n]{} 0. \tag{2.7}$$

First we prove (2.6). Clearly,

$$\frac{nF(a_m x + b_m) - k_n'}{\sqrt{k_n'}} = \sqrt{\frac{n}{m}} \left(\sqrt{m} \frac{F(a_m x + b_m) - \bar{k}_m}{\sqrt{\bar{k}_m}} \right) \xrightarrow[n]{} \begin{cases} \infty, & \text{if } x > 1, \\ -\infty, & \text{if } x < 1. \end{cases} \tag{2.8}$$

Thus, from (2.8), we get

$$P\left(\frac{X_{k_n':n} - x_0}{a_m} < \epsilon + 1 \right) \xrightarrow[n]{} \mathcal{N}(\infty) = 1,$$

which in turn implies

$$P\left(\frac{X_{k_n':n} - x_0}{a_m} > \epsilon + 1 \right) \xrightarrow[n]{} 0. \tag{2.9}$$

Similarly we have

$$P\left(\frac{X_{k_n':n} - x_0}{a_m} < -\epsilon + 1 \right) \xrightarrow[n]{} \mathcal{N}(-\infty) = 0. \tag{2.10}$$

From (2.9) and (2.10), we get

$$P\left(\left| \frac{X_{k_n':n} - x_0}{a_m} - 1 \right| > \epsilon \right) \xrightarrow[n]{} 0.$$

Hence (2.6) is proved. Turning now to prove (2.7). By using Lemma 1.2 and the condition $m = o(n)$, we get

$$\frac{a_n}{a_m} \sim \frac{L^{\frac{-l}{\beta}}(k_n)}{L^{\frac{-l}{\beta}}(k_m)} = \frac{e^{\frac{-l}{\beta} n^{\frac{\alpha}{2}}}}{e^{\frac{-l}{\beta} m^{\frac{\alpha}{2}}}} = e^{\frac{-l}{\beta} n^{\frac{\alpha}{2}} \left(1 - \left(\frac{m}{n}\right)^{\frac{\alpha}{2}}\right)} \xrightarrow[n]{} 0,$$

which proves (2.7). Finally, in order to switch from convergence in probability to convergence w.p.1, we argue by the same way as in the end of the proof of Theorem 2.1. This completes the proof of Part (i). Now, assume that $F(a_n x + b_n) \in D^{(l,\alpha)}(\mathcal{N}(U_{2;\beta}(x)))$. Therefore, in view of the result of Chibisov [18], we have $x_0 = -\infty$ and $b_n = b_m = 0$ (this legitimates the choice $\hat{b}_m = 0$). On the other hand, by Theorem 2.3 in order to prove Part (ii) of the theorem, it suffices to show that

$$\frac{\hat{a}_m}{a_m} = \frac{-X_{k_n':n}}{a_m} \xrightarrow[n]{} 1 \tag{2.11}$$

and

$$\frac{\hat{b}_m - b_m}{a_m} \xrightarrow[n]{} 0, \tag{2.12}$$

both in probability or w.p.1. First, let us focus on the case of convergence in probability.

It is clear that (2.12) is satisfied (actually $\frac{\hat{b}_m - b_m}{a_m} = 0$, for all m). Therefore, we have only to prove (2.11). Clearly, we have

$$\frac{nF(a_m x + b_m) - k_n'}{\sqrt{k_n'}} = \sqrt{\frac{n}{m}}\left(\sqrt{m}\frac{F(a_m x) - \bar{k}_m}{\sqrt{\bar{k}_m}}\right) \underset{n}{\longrightarrow} \begin{cases} -\infty, & \text{if } x < -1, \\ \infty, & \text{if } x > -1. \end{cases} \tag{2.13}$$

Thus, from (2.13), we get

$$P\left(\frac{X_{k_n':n}}{a_m} < -(\epsilon + 1)\right) \underset{n}{\longrightarrow} \mathcal{N}(-\infty) = 0,$$

which implies to

$$P\left(\frac{-X_{k_n':n}}{a_m} > \epsilon + 1\right) \underset{n}{\longrightarrow} 0. \tag{2.14}$$

Similarly we have

$$P\left(\frac{X_{k_n':n}}{a_m} < -(1 - \epsilon)\right) \underset{n}{\longrightarrow} \mathcal{N}(\infty) = 1,$$

which in turn is equivalent to

$$P\left(\frac{-X_{k_n':n}}{a_m} < 1 - \epsilon\right) \underset{n}{\longrightarrow} 0. \tag{2.15}$$

From (2.14) and (2.15), we get $P\left(\left|\frac{-X_{k_n':n}}{a_m} - 1\right| > \epsilon\right) \underset{n}{\longrightarrow} 0$, which proves (2.11), as well as Part (ii), when the convergence in the probability. In order to switch to the convergence w.p.1, we again argue by the same way as in the end of the proof of Theorem 2.1. Finally, assume that $F(a_n x + b_n) \in D^{(l,\alpha)}(\mathcal{N}(U_3(x)))$. By Theorem 2.2, in order to prove Part (iii), it suffices to show that

$$\frac{\hat{a}_m}{a_m} = \frac{X_{k_n'':n} - X_{k_n':n}}{a_m} \underset{n}{\longrightarrow} 1 \tag{2.16}$$

and

$$\frac{\hat{b}_m - b_m}{a_m} = \frac{X_{k_n':n} - b_m}{a_m} \underset{n}{\longrightarrow} 0, \tag{2.17}$$

both in probability or w.p.1. First, let us again focus on the case of convergence in probability and write

$$\frac{\hat{a}_m}{a_m} = \frac{X_{k_n'':n} - X_{k_n':n}}{a_m} = \frac{X_{k_n'':n} - b_m}{a_m} - \frac{X_{k_n'} - b_m}{a_m}.$$

Hence, to prove (2.16) and (2.17), it is sufficient to show that

$$\frac{X_{k_n'':n} - b_m}{a_m} \underset{n}{\xrightarrow{\text{p}}} 1 \tag{2.18}$$

and

$$\frac{X_{k_n':n} - b_m}{a_m} \underset{n}{\xrightarrow{\text{p}}} 0. \tag{2.19}$$

First, we prove (2.18). One can write

$$\frac{nF(a_m x + b_m) - k_n''}{\sqrt{k_n''}} = \frac{nF(a_m x + b_m) - \frac{n}{m}(k_m + \sqrt{k_m})}{\sqrt{\frac{n}{m}(k_m + \sqrt{k_m})}}$$

$$= \sqrt{\frac{n}{m}}\left(\frac{mF(a_m x+b_m)-(k_m+\sqrt{k_m})}{\sqrt{(k_m+\sqrt{k_m})}}\right) = \sqrt{\frac{n}{m}}\left(\frac{mF(a_m x+b_m)-(k_m+\sqrt{k_m})}{\sqrt{k_m(1+\frac{1}{\sqrt{k_m}})}}\right)$$

$$= \sqrt{\frac{n}{m}}\left(\frac{mF(a_m x+b_m)-(k_m+\sqrt{k_m})}{\sqrt{k_m(1+o(1))}}\right) = \sqrt{\frac{n}{m}}\left(\frac{mF(a_m x+b_m)-k_m}{\sqrt{k_m(1+o(1))}} - \frac{\sqrt{k_m}}{\sqrt{k_m(1+o(1))}}\right).$$

On the other hand, the assumption of the theorem yields

$$\frac{mF(a_m x+b_m)-k_m}{\sqrt{k_m(1+o(1))}} \underset{n}{\longrightarrow} x.$$

Therefore, we get

$$\frac{nF(a_m x+b_m)-k_n''}{\sqrt{k_n''}} \underset{n}{\longrightarrow} \begin{cases} \infty, & \text{if } x>1, \\ -\infty, & \text{if } x<1. \end{cases} \tag{2.20}$$

Thus, for every $\epsilon > 0$, we get

$$P\left(\frac{X_{k_n'':n}-b_m}{a_m} < \epsilon+1\right) \underset{n}{\longrightarrow} \mathcal{N}(\infty)=1,$$

which implies

$$P\left(\frac{X_{k_n'':n}-b_m}{a_m} > \epsilon+1\right) \underset{n}{\longrightarrow} 0. \tag{2.21}$$

Moreover, by the same way we get

$$P\left(\frac{X_{k_n'':n}-b_m}{a_m} < 1-\epsilon\right) \underset{n}{\longrightarrow} \mathcal{N}(-\infty)=0. \tag{2.22}$$

Thus, (2.21) and (2.22) lead to

$$P\left(\left|\frac{X_{k_n'':n}-b_m}{a_m}-1\right| > \epsilon\right) \underset{n}{\longrightarrow} 0,$$

which proves (2.18). Next, we prove (2.19). We have

$$\frac{nF(a_m x+b_m)-k_n'}{\sqrt{k_n'}} = \frac{nF(a_m x+b_m)-\frac{n}{m}k_m}{\sqrt{\frac{n}{m}k_m}}$$

$$= \sqrt{\frac{n}{m}}\left(\sqrt{m}\frac{F(a_m x+b_m)-\bar{k}_m}{\sqrt{\bar{k}_m}}\right) \underset{n}{\longrightarrow} \begin{cases} \infty, & \text{if } x>0, \\ -\infty, & \text{if } x<0. \end{cases} \tag{2.23}$$

Thus, (2.23) yields $P\left(\frac{X_{k_n':n}-b_m}{a_m} < \epsilon\right) \underset{n}{\longrightarrow} \mathcal{N}(\infty)=1$, which implies that

$$P\left(\frac{X_{k_n':n}-b_m}{a_m} > \epsilon\right) \underset{n}{\longrightarrow} 0. \tag{2.24}$$

Similarly we have

$$P\left(\frac{X_{k_n':n}-b_m}{a_m} < -\epsilon\right) \underset{n}{\longrightarrow} \mathcal{N}(-\infty)=0. \tag{2.25}$$

Therefore, the relations (2.24) and (2.25) imply

$$P\left(\left|\frac{X_{k'_n:n}-b_m}{a_m}\right|>\epsilon\right)\underset{n}{\longrightarrow}0,$$

and this proves (2.19). In order to switch to the convergence w.p.1, we argue by the same way as in the end of the proof of Theorem 2.1. This completes the proof of the theorem. □

3. BOOTSTRAPPING CENTRAL ORDER STATISTICS

In this section, we discuss the strong consistency of the bootstrap distribution $H^{\star}_{n,m}(c_m x+d_m)=P(X_{k_m:m}\le c_m x+d_m|X_n)$, where k_n is the central rank sequence, which satisfies the condition $k_n\sim pn+o(\sqrt{n})$, and (1.6) is satisfied with $V(x)=V_{i;\beta}(x),\ i=1,2,3,4$, for some suitable normalizing constants $c_n>0$ and d_n.

3.1. Almost sure consistency of bootstrapping central for known normalizing constants

Barakat *et al.* [16] proved the weak limit relations $\sup_{x\in\mathbb{R}}|H^{\star}_{n,m}(c_m x+d_m)-\mathcal{N}(V_{i;\beta}(x))|\xrightarrow[n]{\text{p}}0,\ \ i=1,2,3,4$, if $m=o(n)$. The following theorem extends this result.

Theorem 3.1. *If* $\sum_{n=1}^{\infty}\lambda^{\sqrt{\frac{n}{m}}}<\infty$, *for each* $\lambda\in(0,1)$, *then*

$$\sup_{x\in\mathbb{R}}|H^{\star}_{n,m}(c_m x+d_m)-\mathcal{N}(V_{i;\beta}(x))|\xrightarrow[n]{\text{w.p.1}}0,\ \ i=1,2,3,4.$$

Proof: On one hand, we have

$$\sqrt{m}\frac{F_n(c_m x+d_m)-p}{C_p}=\sqrt{\frac{m}{n}}\left(\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{np(1-p)}}\right)+\sqrt{m}\frac{F(c_m x+d_m)-p}{C_p}.$$

On the other hand, the assumption of the theorem guarantees that $F(c_m x+d_m)\sim p$, as $n\to\infty$, and

$$\sqrt{m}\frac{F(c_m x+d_m)-C_p}{C_p}\underset{n}{\longrightarrow}V_{i;\beta}(x).$$

Thus, to prove

$$\sqrt{m}\frac{F_n(c_m x+d_m)-C_p}{C_p}\xrightarrow[n]{\text{w.p.1}}V_{i;\beta}(x),$$

we need only to show that

$$\sqrt{\frac{m}{n}}\left(\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}\right)\xrightarrow[n]{\text{w.p.1}}0.$$

By Borel–Cantelli lemma, it is enough to prove that

$$\sum_{n=1}^{\infty} P\left(\sqrt{\frac{m}{n}}\left|\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}\right|>\epsilon\right)<\infty,$$

for every $\epsilon>0$. Now, for each $\theta>0$ we have

$$\sqrt{\frac{m}{n}}\log P\left(\sqrt{\frac{m}{n}}\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}>\epsilon\right)=$$

$$\sqrt{\frac{m}{n}}\log P\left(\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}>\sqrt{\frac{n}{m}}\ \epsilon\right)=\sqrt{\frac{m}{n}}\log P\Big(e^{\theta T_{n,m}}>e^{\theta\sqrt{\frac{n}{m}}\ \epsilon}\Big),$$

where $T_{n,m}$ is defined as

$$T_{n,m}=\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}.$$

By using Markov inequality we get

$$\begin{aligned}\sqrt{\frac{m}{n}}\log P\Big(e^{\theta T_{n,m}}>e^{\theta\sqrt{\frac{n}{m}}\ \epsilon}\Big)\ &\le\ \sqrt{\frac{m}{n}}\log\ \Big(e^{-\theta\sqrt{\frac{n}{m}}\ \epsilon}E\Big(e^{\theta T_{n,m}}\Big)\Big)\\ &=\ -\theta\epsilon+\sqrt{\frac{m}{n}}\log\varphi_m(\theta)\ \underset{n}{\longrightarrow}\ -\theta\epsilon.\end{aligned}$$

Therefore, for sufficiently large n, we get

$$\sum_{n=1}^{\infty} P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}\right)>\epsilon\right)=$$

$$\sum_{i=1}^{\infty}\exp\left\{\log P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}\right)>\epsilon\right)\right\}\le\sum_{i=1}^{\infty}e^{-\theta\epsilon\sqrt{\frac{n}{m}}}<\infty,$$

for every $\epsilon>0$, since the condition $\sum_{n=1}^{\infty}\lambda^{\sqrt{\frac{n}{m}}}<\infty$, for each $\lambda\in(0,1)$, guarantees the convergence of the infinite series $\sum_{n=1}^{\infty}\exp\{-\theta\epsilon\sqrt{\frac{n}{m}}\}$, for every $\epsilon>0$. By similar reasoning we can show that

$$\sum_{n=1}^{\infty} P\left(\sqrt{\frac{m}{n}}\left(\frac{nF_n(c_m x+d_m)-nF(c_m x+d_m)}{\sqrt{nF(c_m x+d_m)(1-F(c_m x+d_m))}}\right)<-\epsilon\right)<\infty,$$

for every $\epsilon>0$. The theorem is proved. □

3.2. Limits of bootstrap distribution for central order statistics when normalizing constants are unknown

Let $\hat{c}_m$ and $\hat{d}_m$ be estimators of c_m and d_m based on $X_n=(X_1,X_2,...,X_n)$, respectively. Define the bootstrap distribution for the normalized central order statistic $c_n^{-1}(X_{k_n:n}-d_n)$ with the estimated normalizing constants by $\hat{H}^{\star}_{n,m}(\hat{c}_m x+\hat{d}_m)=P(Y_{k_m:m}\le\hat{c}_m x+\hat{d}_m|X_n)$. In order to study the limit of bootstrap distribution for central order statistics when normalizing constants are unknown, we start with the following essential theorem.

Theorem 3.2. *Let $m = m(n)$. Then, for all the continuity points of $V_{i;\beta}(x)$, $i = 1, 2, 3$ (see Remark 3.1), we have*

$$\sup_{x\in\mathbb{R}}\left|\hat{H}^{\star}_{n,m}(\hat{c}_m x + \hat{d}_m) - \mathcal{N}(V_{i;\beta}(x))\right| \xrightarrow[n]{\text{w.p.1}} 0,$$

if (**i**) $H^{\star}_{n,m}(x) \xrightarrow[n]{\text{w.p.1}} \mathcal{N}(V_{i;\beta}(x))$,

(**ii**) $\dfrac{\hat{c}_m}{c_m} \xrightarrow[n]{\text{w.p.1}} 1$,

and

(**iii**) $\dfrac{\hat{d}_m - d_m}{c_m} \xrightarrow[n]{\text{w.p.1}} 0.$

Moreover, this theorem holds if "$\xrightarrow[n]{\text{w.p.1}}$" is replaced by "$\xrightarrow[n]{\text{p}}$".

Proof: The proof of the theorem is similar to the proof of Theorem 2.2. □

Remark 3.1. A quick look at the possible non-degenerate limit laws $\mathcal{N}(V_{i;\beta}(x))$, $i = 1, 2, 3$, reveals that each of these limit laws has at most one discontinuity point.

For the bootstrap distribution $\hat{H}^{\star}_{n,m}(\hat{c}_m x + \hat{d}_m)$ to be consistent, we need to choose $\hat{c}_m$ and $\hat{d}_m$ satisfying the conditions (ii) and (iii) in Theorem 3.2. In the next theorem, we suggest choices for $\hat{c}_m$ and $\hat{d}_m$ as a functional of the empirical distribution for the domains of attraction $F(c_n x + d_n) \in D^{(p)}\mathcal{N}(V_{i;\beta}(x)), i = 1, 2, 3$.

Theorem 3.3. *Let $k_n^{'} = [pn] + 1$, $k_n^{''} = [\frac{n}{\sqrt{m}} + pn] + 1$, and $k_n^{'''} = [pn - \frac{n}{\sqrt{m}}] + 1$. Then,*

1. *if $F(c_n x + d_n) \in D^{(p)}(\mathcal{N}(V_{1;\beta}(x)))$, $\hat{c}_m = F_n^{-1}\left(p + \frac{1}{\sqrt{m}}\right) - F_n^{-1}(p) = X_{k_n^{''}:n} - X_{k_n^{'}:n}$, and $\hat{d}_m = F_n^{-1}(p) = X_{k_n^{'}:n}$;*
2. *if $F(c_n x + d_n) \in D^{(p)}(\mathcal{N}(V_{2;\beta}(x)))$, $\hat{c}_m = F_n^{-1}(p) - F_n^{-1}\left(p - \frac{1}{\sqrt{m}}\right) = X_{k_n^{'}:n} - X_{k_n^{'''}:n}$, and $\hat{d}_m = F_n^{-1}(p) = X_{k_n^{'}:n}$;*
3. *if $F(c_n x + d_n) \in D^{(p)}(\mathcal{N}(V_{3;\beta}(x)))$, $\hat{c}_m = F_n^{-1}\left(p + \frac{1}{\sqrt{m}}\right) - F_n^{-1}(p) = X_{k_n^{''}:n} - X_{k_n^{'}:n}$, and $\hat{d}_m = F_n^{-1}(p) = X_{k_n^{'}:n}$.*

If $m = o(n)$, then

$$\sup_{x\in\mathbb{R}}\left|\hat{H}^{\star}_{n,m}(\hat{c}_m x + \hat{d}_m) - \mathcal{N}(V_{i;\beta}(x))\right| \xrightarrow[n]{\text{p}} 0. \tag{3.1}$$

Moreover, if $\sum_{n=1}^{\infty} \lambda^{\sqrt{\frac{n}{m}}} < \infty$ for each $\lambda \in (0, 1)$ then (3.1) holds w.p.1.

Proof: Let $F(c_n x + d_n) \in D^{(p)}(\mathcal{N}(V_{1;\beta}(x)))$. From Theorem 3.2, it suffices to show that

$$\frac{\hat{c}_m}{c_m} = \frac{X_{k_n^{''}:n} - X_{k_n^{'}:n}}{c_m} \xrightarrow[n]{} 1 \tag{3.2}$$

and

$$\frac{\hat{d}_m - d_m}{c_m} = \frac{X_{k_n^{'}:n} - d_m}{c_m} \xrightarrow[n]{} 0, \tag{3.3}$$

both in probability or w.p.1. First, let us focus on the case of convergence in probability.

We start with

$$\frac{\hat{c}_m}{c_m} = \frac{X_{k_n'':n} - X_{k_n':n}}{c_m} = \frac{X_{k_n'':n} - d_m}{c_m} - \frac{X_{k_n':n} - d_m}{c_m}.$$

Thus, to prove (3.2) and (3.3), it is sufficient to show that

$$\frac{X_{k_n'':n} - d_m}{c_m} \xrightarrow[n]{\mathrm{p}} 1 \tag{3.4}$$

and

$$\frac{X_{k_n':n} - d_m}{c_m} \xrightarrow[n]{\mathrm{p}} 0. \tag{3.5}$$

We start with the proof of (3.4). By using the relations $[\frac{n}{\sqrt{m}} + pn] = \frac{n}{\sqrt{m}} + pn - \delta$, where $0 \le \delta < 1$, and $\frac{1}{\sqrt{m}} + p + \frac{1-\delta}{n} \sim p$, as $n \to \infty$, we get

$$\frac{nF(c_m x + d_m) - k_n''}{\sqrt{k_n''(1 - \frac{k_n''}{n})}} = \sqrt{n}\frac{F(c_m x + d_m) - \left(\frac{1}{\sqrt{m}} + p + \frac{1-\delta}{n}\right)}{\sqrt{\left(\frac{1}{\sqrt{m}} + p + \frac{1-\delta}{n}\right)\left(1 - \left(\frac{1}{\sqrt{m}} + p + \frac{1-\delta}{n}\right)\right)}}$$

$$\sim \sqrt{\frac{n}{m}}\left(\sqrt{m}\frac{F(c_m x + d_m) - p}{C_p} - \frac{1 + \frac{\sqrt{m}}{n}(1-\delta)}{C_p}\right) \xrightarrow[n]{} \begin{cases} \infty, & \text{if } x > 1, \\ -\infty, & \text{if } x < 1. \end{cases} \tag{3.6}$$

Relation (3.6) follows from the two obvious relations

$$\frac{1 + \frac{\sqrt{m}}{n}(1-\delta)}{C_p} \xrightarrow[n]{} \frac{1}{C_p} = c \quad \text{and} \quad \sqrt{m}\frac{F(c_m x + d_m) - p}{C_p} \xrightarrow[n]{} cx^{\beta}.$$

The relation (3.6) yields $P(\frac{X_{k_n'':n} - d_m}{c_m} < \epsilon + 1) \xrightarrow[n]{} \mathcal{N}(\infty) = 1$, which is equivalent to

$$P\left(\frac{X_{k_n'':n} - d_m}{c_m} > \epsilon + 1\right) \xrightarrow[n]{} 0. \tag{3.7}$$

Similarly we have

$$P\left(\frac{X_{k_n'':n} - d_m}{c_m} < -\epsilon + 1\right) \xrightarrow[n]{} \mathcal{N}(-\infty) = 0. \tag{3.8}$$

From (3.7) and (3.8) we get $P(|\frac{X_{k_n'':n} - d_m}{c_m} - 1| > \epsilon) \xrightarrow[n]{} 0$, which proves (3.4). Now, we prove (3.5). One can easily deduce that

$$\sqrt{n}\frac{F(c_m x + d_m) - p}{C_p} = \sqrt{\frac{n}{m}}\left(\sqrt{m}\frac{F(c_m x + d_m) - p}{C_p}\right) \xrightarrow[n]{} \begin{cases} \infty, & \text{if } x > 0, \\ -\infty, & \text{if } x < 0. \end{cases} \tag{3.9}$$

Thus, from (3.9), we have $P(\frac{X_{k_n':n} - d_m}{c_m} < \epsilon) \xrightarrow[n]{} \mathcal{N}(\infty) = 1$, which is equivalent to

$$P\left(\frac{X_{k_n':n} - d_m}{c_m} > \epsilon\right) \xrightarrow[n]{} 0. \tag{3.10}$$

Similarly we obtain

$$P\left(\frac{X_{k'_n:n}-d_m}{c_m}<-\epsilon\right)\underset{n}{\longrightarrow}0. \tag{3.11}$$

Therefore, by combining the relations (3.10) and (3.11), we get $P(|\frac{X_{k'_n:n}-d_m}{c_m}|>\epsilon)\underset{n}{\longrightarrow}0$, which proves (3.5). In order to switch to convergence w.p.1, we proceed as in the end of the proof of Theorem 2.1. This completes the proof of Part (i).

Now, let $F(c_nx+d_n)\in D^{(p)}(\mathcal{N}(V_{2;\beta}(x)))$. From Theorem 3.2, it suffices to show that

$$\frac{\hat{c}_m}{c_m}=\frac{X_{k'_n:n}-X_{k'''_n:n}}{c_m}\underset{n}{\longrightarrow}1 \tag{3.12}$$

and

$$\frac{\hat{d}_m-d_m}{c_m}=\frac{X_{k'_n:n}-d_m}{c_m}\underset{n}{\longrightarrow}0, \tag{3.13}$$

both in probability or w.p.1. Again, we first focus on the case of the convergence in probability and we start with

$$\frac{\hat{c}_m}{c_m}=\frac{X_{k'_n:n}-X_{k'''_n:n}}{c_m}=\frac{X_{k'_n:n}-d_m}{c_m}-\frac{X_{k'''_n:n}-d_m}{c_m}.$$

Hence, to prove (3.12) and (3.13), it is sufficient to show that

$$\frac{X_{k'''_n:n}-d_m}{c_m}\underset{n}{\overset{\mathrm{p}}{\longrightarrow}}-1 \tag{3.14}$$

and

$$\frac{X_{k'_n:n}-d_m}{c_m}\underset{n}{\overset{\mathrm{p}}{\longrightarrow}}0. \tag{3.15}$$

We prove (3.14). By applying the relations $[pn-\frac{n}{\sqrt{m}}]=pn-\frac{n}{\sqrt{m}}-\delta$, $0\le\delta<1$, and $p-\frac{1}{\sqrt{m}}+\frac{1-\delta}{n}\sim p$, as $n\to\infty$, we can deduce that

$$\begin{aligned}\frac{nF(c_mx+d_m)-k'''_n}{\sqrt{k'''_n(1-\frac{k'''_n}{n})}}&=\sqrt{n}\frac{F(c_mx+d_m)-\left(p-\frac{1}{\sqrt{m}}+\frac{1-\delta}{n}\right)}{\sqrt{\left(p-\frac{1}{\sqrt{m}}+\frac{1-\delta}{n}\right)\left(1-\left(p-\frac{1}{\sqrt{m}}+\frac{1-\delta}{n}\right)\right)}}\\ &\sim\sqrt{\frac{n}{m}}\left(\sqrt{m}\frac{F(c_mx+d_m)-p}{C_p}-\frac{-1+\frac{\sqrt{m}}{n}(1-\delta)}{C_p}\right)\underset{n}{\longrightarrow}\begin{cases}-\infty, & \text{if } |x|>1,\\ \infty, & \text{if } |x|<1.\end{cases}\end{aligned} \tag{3.16}$$

Thus, on account (3.16), we get $P(\frac{X_{k'''_n:n}-d_m}{c_m}<\epsilon-1)\underset{n}{\longrightarrow}\mathcal{N}(\infty)=1$, which is equivalent to

$$P\left(\frac{X_{k'''_n:n}-d_m}{c_m}>\epsilon-1\right)\underset{n}{\longrightarrow}0. \tag{3.17}$$

In the same manner, we have

$$P\left(\frac{X_{k'''_n:n}-d_m}{c_m}<-\epsilon-1\right)\underset{n}{\longrightarrow}\mathcal{N}(-\infty)=0. \tag{3.18}$$

From (3.17) and (3.18), we get $P(|\frac{X_{k'''_n:n}-d_m}{c_m}+1|>\epsilon)\underset{n}{\longrightarrow}0$. Hence (3.14) is proved.

We turn now to prove (3.15). We start with the obvious limit relation

$$\sqrt{n}\frac{F(c_m x + d_m) - p}{C_p} = \sqrt{\frac{n}{m}}\left(\sqrt{m}\frac{F(c_m x + d_m) - p}{C_p}\right) \underset{n}{\longrightarrow} \begin{cases} \infty, & \text{if } x > 0, \\ -\infty, & \text{if } x < 0, \end{cases} \tag{3.19}$$

which in turn implies that $P(\frac{X_{k'_n:n} - d_m}{c_m} < \epsilon) \underset{n}{\longrightarrow} \mathcal{N}(\infty) = 1$ and hence

$$P\left(\frac{X_{k'_n:n} - d_m}{c_m} > \epsilon\right) \underset{n}{\longrightarrow} 0. \tag{3.20}$$

Moreover, the limit relation (3.19) yields

$$P\left(\frac{X_{k'_n:n} - d_m}{c_m} < -\epsilon\right) \underset{n}{\longrightarrow} 0. \tag{3.21}$$

By combining (3.20) and (3.21), we get $P(|\frac{X_{k'_n:n} - d_m}{c_m}| > \epsilon) \underset{n}{\longrightarrow} 0$, which proves (3.15). Finally, the fact that the convergence in (3.14) and (3.15) is w.p.1 can be easily proved by the same way as in the end of the proof of Theorem 2.1. This completes the proof of Part (ii).

Finally, consider the case $F(c_n x + d_n) \in D^{(p)}(\mathcal{N}(V_{3;\beta}(x)))$. From Theorem 3.2, it suffices to show that

$$\frac{\hat{c}_m}{c_m} = \frac{X_{k''_n:n} - X_{k'_n:n}}{c_m} \underset{n}{\longrightarrow} 1 \tag{3.22}$$

and

$$\frac{\hat{d}_m - d_m}{c_m} = \frac{X_{k'_n:n} - d_m}{c_m} \underset{n}{\longrightarrow} 0, \tag{3.23}$$

both in probability or w.p.1. We first focus on the case of the convergence in probability and we start with

$$\frac{\hat{c}_m}{c_m} = \frac{X_{k''_n:n} - X_{k'_n:n}}{c_m} = \frac{X_{k''_n:n} - d_m}{c_m} - \frac{X_{k'_n:n} - d_m}{c_m}.$$

Therefore, to prove (3.22) and (3.23), it is sufficient to show that

$$\frac{X_{k'_n:n} - d_m}{c_m} \underset{n}{\xrightarrow{\text{p}}} 1 \tag{3.24}$$

and

$$\frac{X_{k''_n:n} - d_m}{c_m} \underset{n}{\xrightarrow{\text{p}}} 0. \tag{3.25}$$

By proceeding as we did in Parts (i) and (ii), we can easily show that

$$\frac{nF(c_m x + d_m) - k''_n}{\sqrt{k''_n(1 - \frac{k''_n}{n})}} \underset{n}{\longrightarrow} \begin{cases} \infty, & \text{if } x > 1, \\ -\infty, & \text{if } x < 1. \end{cases} \tag{3.26}$$

Again, by proceeding as we did in Parts (i) and (ii), the relation (3.26) yields $P(|\frac{X_{k''_n:n} - d_m}{c_m} - 1| > \epsilon) \underset{n}{\longrightarrow} 0$, which in turn proves (3.24). On the other hand, the proof of the relation (3.25) follows also by proceeding as we did in Parts (i) and (ii). Finally, we can prove that the convergence in both the relations (3.24) and (3.25) is w.p.1, by the same way as in the end of the proof of Theorem 2.1. The proof is complete. □

3.3. Bootstrapping sample quantiles when the DFs of these quantiles weakly converge to $\mathcal{N}(x)$ and F is unknown

It has been known for a long time that the DF of the sample quantile $X_{k_n:n} = X_{[pn]+1:n}$, $0 < p < 1$, based on a continuous DF $F(x)$ with positive probability density (PDF) $f(x)$ in a neighborhood of the p-th population quantile $x_o = F^{-1}(p)$, weakly converges to the standard normal DF (e.g., see [35]). In the present subsection, we will study the limit bootstrapping sample quantiles when the PDF f is unknown. We start with a classical result; its proof can be found in many known references among them [35].

Lemma 3.1. *Let $X_{k_n:n} = X_{[pn]+1:n}$, $0 < p < 1$, be a sample quantile, which is based on a continuous DF $F(x)$ with a positive PDF $f(x)$ in a neighborhood of the p-th population quantile $x_0 = F^{-1}(p)$. Then,*

$$P(X_{k_n:n} < c_n x + d_n) = P\left(\sqrt{n} f(F^{-1}(p)) \frac{X_{k_n:n} - F^{-1}(p)}{\sqrt{p(1-p)}} \leq x\right) \xrightarrow[n]{\text{w}} \mathcal{N}(x). \tag{3.27}$$

where $c_n = \frac{\sqrt{p(1-p)}}{\sqrt{n} f(x_0)}$ and $d_n = x_0 = F^{-1}(p)$.

It is known (cf. [34]) that $X_{k_n:n}$ is a consistent estimator of $F^{-1}(p)$. Moreover, the relation (3.27) can be used to construct an approximate confidence interval for $F^{-1}(p)$, if either the form of f is completely specified around $F^{-1}(p)$ or a good estimator for $f(F^{-1}(p))$ is available. Siddiqui [34] proposed an estimator for $\frac{1}{f(x_0)} = \frac{1}{f(F^{-1}(p))}$ in the form $S_{rn} = \frac{n}{2r}(X_{[np]+r:n} - X_{[np]-r+1:n})$. Moreover, Siddiqui [34] showed that this estimator is asymptotically normal DF, when r is chosen to be of order $n^{\frac{1}{2}}$. Bloch and Gastwirth [17] showed that, if $r = o(n)$ and $r \xrightarrow[n]{} \infty$ then, S_{rn} is a consistent estimator for $\frac{1}{f(F^{-1}(p))}$. Now, we study the bootstrap distribution of $X_{k_m:m}$, $k_m = [mp] + 1$, which defined for unknown normalizing constants by $H^*_{n,m}(\hat{c}_m x + \hat{d}_m) = P(X_{k_m:m} < \hat{c}_m x + \hat{d}_m \,|\, X_n)$, where $\hat{c}_m$ and $\hat{d}_m$ are some estimators of c_m and d_m, respectively.

Theorem 3.4. *Let $\hat{c}_m = \frac{\sqrt{p(1-p)}}{\sqrt{m}} S_{rm}$, $\hat{d}_m = X_{[np]+1:n}$, where $r = o(m)$. Then,*

$$\sup_{x \in \mathbb{R}} |H^*_{n,m}(\hat{c}_m x + \hat{d}_m) - \mathcal{N}(x))| \xrightarrow[n]{\text{p}} 0, \text{ if } m = o(n).$$

Moreover, if there exist $\lambda \in (0,1)$ such that $\sum_{n=1}^{\infty} \lambda^{\sqrt{\frac{n}{m}}} < \infty$ then

$$\sup_{x \in \mathbb{R}} |H^*_{n,m}(\hat{c}_m x + \hat{d}_m) - \mathcal{N}(x))| \xrightarrow[n]{\text{w.p.1}} 0.$$

Proof: In order to the bootstrap distribution $H^*_{n,m}(\hat{c}_m x + \hat{d}_m)$ to be consistent, we have to prove that $\hat{c}_m$ and $\hat{d}_m$ satisfy the conditions (ii) and (iii) in Theorem 3.2, respectively. Since S_{rm} is a consistent estimator for $\frac{1}{f(x_0)}$ (cf. [17]), we get

$$\frac{\hat{c}_m}{c_m} = \frac{\sqrt{p(1-p)}/\sqrt{m} S_{rm}}{\sqrt{p(1-p)}/\sqrt{m} f(x_o)} = S_{rm} f(x_0) \xrightarrow[n]{\text{p}} 1.$$

On the other hand, we have $\frac{c_n}{c_m} = \frac{\sqrt{p(1-p)}/\sqrt{n}f(x_0)}{\sqrt{p(1-p)}/\sqrt{m}f(x_0)} = \sqrt{\frac{m}{n}} \underset{n}{\longrightarrow} 0$. Thus, on account of Lemma 3.1, we get

$$\frac{\hat{d}_m - d_m}{c_m} = \frac{X_{[np]+1:n} - F^{-1}(p)}{c_n} \frac{c_n}{c_m} \underset{n}{\xrightarrow{\text{p}}} 0.$$

Therefore, the conditions (ii) and (iii) in Theorem 3.2 are proved when the convergence is in probability. The proof of these conditions when the convergence is w.p.1 is achieved by the same way as in the end of the proof of Theorem 2.1. The proof is complete. □

4. SIMULATION STUDY

In this section, we address two applications of the earlier theoretical findings. Firstly, we provide a p-value-based method for choosing m. We present a simulation study in Example 4.1 that is carried out using Mathematica 11 to explain how we choose numerically the values of m to give the best approximation of the bootstrapping DFs for the central and intermediate quantiles. In Example 4.1, we choose normality to highlight the key issue that pertains to the selection of m. On the other hand, under typical circumstances, the majority of the practical issues that any researcher faces result in the asymptotic normality of the quantiles (e.g., see Lemma 3.1). Consequently, based on the Kolmogorov–Smirnov test of normality and the corresponding p-values, the best value of m (that corresponds to the largest p-value) should be chosen such that $\sum_{n=1}^{\infty} \lambda^{\sqrt{\frac{n}{m}}} < \infty$, for each $\lambda \in (0,1)$ (see Remark 4.1). Although this method is applied when the quantiles being bootstrapped are asymptotically normal, other possible asymptotically laws given in Theorems 2.1, 2.2, 3.1, and 3.2 can be considered by applying a similar algorithm. Secondly, in Example 4.2, based on several large samples from a logistic distribution, we construct confidence intervals for the median using the bootstrapping methodology and the approach provided in Example 4.1. Additionally, predicted coverage probabilities are included with each computed confidence interval.

4.1. Examples

Example 4.1. This example relies on the fact that the sample median $\mathcal{S}_{1;n} = X_{[\frac{n}{2}+1]:n}$, and the sample intermediate quantiles $\mathcal{S}_{2;n} = X_{[2\sqrt{n}]:n}$, $\mathcal{S}_{3;n} = X_{[\sqrt{n}]:n}$, $\mathcal{S}_{4;n} = X_{2[\sqrt[3]{n}]:n}$, and $\mathcal{S}_{5;n} = X_{[\sqrt[3]{n}]:n}$ based on the standard normal DF weakly converge to the normal DF. Let $\hat{\mathcal{S}}_{i;m}$, $i = 1, 2, ..., 5$, be the corresponding bootstrapping statistics of $\mathcal{S}_{i;n}$, $i = 1, 2, ..., 5$, respectively, where each of these bootstrapping statistics is based on a sub-sample with replacement of size m (a bootstrap sample of size m). According, to the results of Sections 2 and 3, we expect that the bootstrapping DFs of the statistics $\hat{\mathcal{S}}_{i;m}$, $i = 1, 2, ..., 5$, converge to the normal DF provided that $m \ll n$ (i.e., $m = o(n)$).

This study, shown in Table 1, is achieved via the following algorithm:

- (**i**) Generate a random sample (parent sample) of size $n = 100,000$ from $\mathcal{N}(\cdot)$;
- (**ii**) Determine a value of m $(100, 200, ..., 5000$, as shown in Table 1) and generate a sub-sample with a replacement of size m (a bootstrap sample) from the parent sample;

(**iii**) Determine each of the sample bootstrapping statistics $\hat{\mathcal{S}}_{i;m}$, $i = 1, 2, ..., 5$;

(**iv**) Repeat the steps (ii) and (iii) 1000 times to obtain the observed sample bootstrapping statistics $\hat{\mathcal{S}}_{ij;m}$, $i = 1, 2, ..., 5$; $j = 1, 2, ..., 1000$;

(**v**) By using the Kolmogorov–Smirnov test, check the normality of the data sets $\{\hat{\mathcal{S}}_{ij;m},\ i = 1, 2, ..., 5;\ j = 1, 2, ..., 1000\}$ and determine the corresponding p-values (see Remark 4.2);

(**vi**) Repeat the steps (ii)–(v) 100 times for each chosen m and compute the average p-values (denoted by $\overline{p}$) for each chosen m and each of the five statistics. These averages, $\overline{p}$, are written as entries in Table 1, where the best $\overline{p}$ is distinguished by an asterisk.

It is noted that for $n = 100,000$, the best choice of m falls in the interval $[200, 400]$, i.e., the values 200 to 400 are 0.2–0.4% of the value of n (see Remark 4.1). Moreover, the $\overline{p}$ for the central case are higher than those for the intermediate case.

Table 1: $\overline{p}$ corresponding to the checking normality of different bootstrap central and intermediate quantiles for various values of m.

$k_n \rightarrow$	Central	Intermediate			
$m \downarrow$	$k_n = [\frac{n}{2}] + 1$	$k_n = [\sqrt{n}]$	$k_n = 2[\sqrt{n}]$	$k_n = [\sqrt[3]{n}]$	$k_n = 2[\sqrt[3]{n}]$
100	0.424534	0.241040	0.374819	0.0471849	0.221942
200	0.445344	0.35888*	0.354039	0.0759685	0.136431
300	0.47536*	0.326927	0.422355	0.135734	0.28982*
400	0.461186	0.294167	0.43096*	0.18197*	0.213884
500	0.415695	0.254239	0.396061	0.160607	0.229961
600	0.413815	0.145734	0.231875	0.171207	0.206003
700	0.423271	0.165254	0.275231	0.141310	0.206607
800	0.447738	0.095997	0.249245	0.140825	0.154863
900	0.396874	0.104246	0.248103	0.082727	0.137661
1000	0.416074	0.136514	0.134745	0.094409	0.099289
2000	0.388266	0.104154	0.135125	0.020539	0.149453
3000	0.389145	0.002338	0.207131	0.001534	0.009307
4000	0.356465	0.003308	0.113068	0.000084	0.050480
5000	0.338578	0.024744	0.014881	0.000383	0.049058

Remark 4.1. According, to the results of Sections 2 and 3, the best performance of the bootstrapping DFs of the central and intermediate order statistics occurs at the values of m for which $\sum_{n=1}^{\infty} \lambda^{\sqrt{\frac{n}{m}}} < \infty$, for each $\lambda \in (0, 1)$. On the other hand, according, to [1] the condition $\sqrt{m} = o(\frac{2\sqrt{n}}{\log n})$ is a sufficient condition for $\sum_{n=1}^{\infty} \lambda^{\sqrt{\frac{n}{m}}} < \infty$, which implies that the best performance of the bootstrapping DFs of the central and intermediate order statistics occurs when $m \ll 3000$. Therefore, the simulation output endorses this anticipated result.

Remark 4.2. In the earlier version of this paper, in order to implement Part (v) of the given algorithm, we fitted the data sets $\{\hat{\mathcal{S}}_{ij;m},\ i = 1, 2, ..., 5;\ j = 1, 2, ..., 1000\}$ to the normal DF by using the Kolmogorov–Smirnov test after calculating the sample mean and standard deviation. However, one referee point out to an important issue that the Kolmogorov–Smirnov test can be used to fit the normal DF only when parameters are not estimated from the data

(cf. [29]). Since our focus here is only on checking the normality of the bootstrap samples, we apply the Kolmogorov–Smirnov test to check the normality of the given sample bootstrapping statistics without estimating any parameters. Namely, in Mathematica 11, there are two ways to fit any data to the normal DF. The first is to provide the mean and variance values; if not, estimate them based on the data. The second choice is to examine the data's normality without figuring out what the fitted normal distribution's parameter values should be. The second choice was adopted.

Example 4.2. In this example, we generate three samples of sizes $n = 100,000$, $n = 50,000$, and $n = 30,000$, from the logistic distribution with location and scale parameters $0 = \text{mean} = \text{median}$ and 1, respectively. We construct a confidence interval for each median, which pertains to the three samples, using the bootstrapping technique. We first apply the p-value-based method for choosing m, which is given in Example 4.1, where 100 bootstrap runs are taken into consideration. In Table 2, the three best values of m and the corresponding best average values of p-values are given. We currently have a sample of 100 observed medians for each of the three initial samples. These median samples follow a normal DF with unknown parameters. Use these median samples to estimate these unknown parameters. Finally, construct a 99% confidence interval of each median pertaining to the three original samples (of sizes 100,000, 50,000, and 30,000). For each of these three samples, we get one constructed confidence interval. Therefore, according to the algorithm given below, we have 10000 confidence intervals to be checked whether each of them contains zero. Due to the use of the bootstrapping technique and also estimating the unknown parameters, we anticipate that the significant levels (SL) of these confidence intervals are smaller than 99%. We estimate the average lower limit ($\overline{L}$), average upper limit ($\overline{U}$), and coverage probability (CP) of the estimated confidence intervals. By doing this, we can estimate the quality of these confidence intervals and subsequently the quality of the suggested approaches. These findings are presented in Table 2, which demonstrates that the SL is not less than 96%, which endorses the results of Theorems 3.3 and 3.4. Moreover, the results presented in Table 2 is achieved via the following algorithm:

(**i**) Generate a random sample (parent sample) of size n ($n = 100,000$; $n = 50,0000$, and $n = 30,000$) from the standard logistic distribution;

(**ii**) Apply the p-value-based method which is given in Example 4.1 and choose the best m corresponding to the largest p-value (e.g., for the $n = 100,000$ case, we have $m = 300$);

(**iii**) Generate M ($M = 100$) sub-samples of size m with replacement from the parent sample and calculate the median for each sample;

(**iv**) Calculate the mean μ_B and standard deviation σ_B for the set of the sample medians (100 medians) in step (iii). In addition, calculate a 99% confidence interval for the population median according to the usual law $\mu_B \pm z_{\alpha/2}\frac{\sigma_B}{\sqrt{M}}$, where $\alpha = 0.01$, and z_p is the p-th quantile of the standard normal DF;

(**v**) Repeat steps (iii) and (iv) 100 times. Determine how many times (say $0 \le n_1 \le 100$), the population median (i.e., zero) falls within the constructed confidence intervals.

(**vi**) Repeat steps (i)–(v) 100 times. In each of those times, we get in step (v), $0 \le n_i \le 100, i = 1, 2, ..., 100$;

(**vii**) Compute $\overline{L}$, $\overline{U}$, and CP= $\frac{\sum_{i=1}^{100} n_i}{10000}$.

Table 2: The average p-values, $\overline{L}$, $\overline{U}$, and CP for median for three different samples from logistic distribution.

$n \downarrow$	$m \downarrow$	p-value	$\overline{L}$	$\overline{U}$	CP
$n = 100,000$	300	0.480606	−0.03	0.03	97.21
$n = 50,000$	200	0.435146	−0.04	0.03	96.69
$n = 30,000$	150	0.420365	−0.04	0.04	96.39

4.2. Discussion

In the light of the preceding simulation study given in Examples 4.1 and 4.2, we consider a virtual case study to show how the developed bootstrap technique in this paper saves time and cost. Suppose our purpose is modeling (i.e., to detect its asymptotic distribution) the sample median of some random phenomenon that is governed by a DF that satisfies the conditions given in Lemma 3.1 (i.e., the sample median from this DF weakly converges to the normal DF).

The usual way to achieve this purpose is to get a large number N of independent random samples, and from each of them, we determine the median. By finding a suitable DF (a normal DF with specified mean and variance) that fits this median-data set (the set of the collected sample medians) we can achieve our aim. As an example, if $N = 1000$ and each sample has a size of 200, we will need 200,000 observations. On the other hand, if we had one large sample of size 100,000 (say) and apply the bootstrap technique, we can achieve our aim by choosing $m \in [200, 400]$ (as the simulation study shows). In this case, bearing in mind that obtaining a large number of independent samples, even of moderated sample sizes, is more difficult and costly than obtaining one sample of a large size, we find that the bootstrap technique is very beneficial. Moreover, regarding the natural question that which of the usual way and bootstrap technique allows us to make better inference on the population median, the theoretical results concerning the bootstrap technique, and especially the result of this paper, guarantees both ways are asymptotically the same. Therefore, one of the most important advantages of the given bootstrap technique is that it enables us to model the different quantiles via one large sample instead of a large number of independent samples.

Undoubtedly the cornerstone of the bootstrap technique given in this section is determining the best value of m. The theoretical result of the paper stipulates that m is small concerning n. One reviewer of this paper provided an elegant intuition about why one wants m to be small, namely, "it is because of discreteness. When m is big, the bootstrap distribution will have big chunks of probability, which can make the distribution less normal than when m is small". The given algorithm to determine m depends on four determinants, which are the parent DF of the given large sample, the sample size n, the number of replications of the p-value, and the number of bootstrap runs. Of course, when any researcher applies the given algorithm he should consider his determinants. However, we repeated the preceding simulation study with different determinants to shed some light on the influence of these determinants on the choice of m.

1. Figures 1–4 suggest that one essentially wants to make m as small as possible, with respect to n, as long as the sufficient condition given in Remark 4.1 is satisfied (and of course, we preserve the necessary requirements that $m \underset{n}{\longrightarrow} \infty$ and $\frac{m}{n} \underset{n}{\longrightarrow} 0$).
2. When the sample size n becomes smaller than 100,000 (with fixing the other determinants), the range of m (the ratio of the best value of m to n) changes by a small amount as shown in Figures 1 and 2. Namely, at $n = 50,000$, the best value of m is about 200 with $\overline{p}$=0.45, while at $n = 30,000$, the best value of m is about 150 with $\overline{p}$=0.43.
3. The change in the number of bootstrap runs (with fixing the other determinants) does not influence the range of the choice of m ($0.2 - 0.4\%$ of the value of n). On the other hand, increasing this number makes $\overline{p}$ decrease and become more stable, see Figure 3.
4. The change in the number of replications of the p-value (with fixing the other determinants) has no influence on the choice of m. On the other hand, increasing this number makes $\overline{p}$ more stable, see Figure 4.

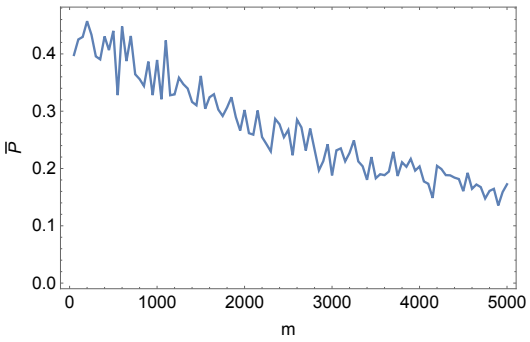


Figure 1: $\overline{p}$ vs. m at $n = 50,000$.

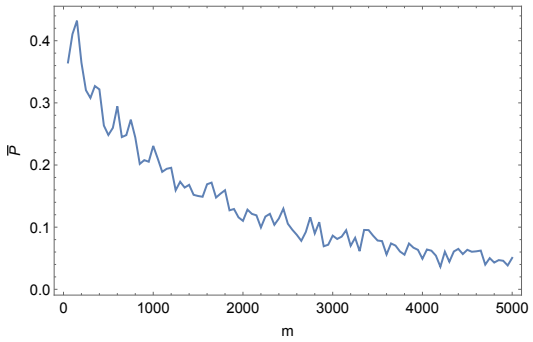


Figure 2: $\overline{p}$ vs. m at $n = 30,000$.

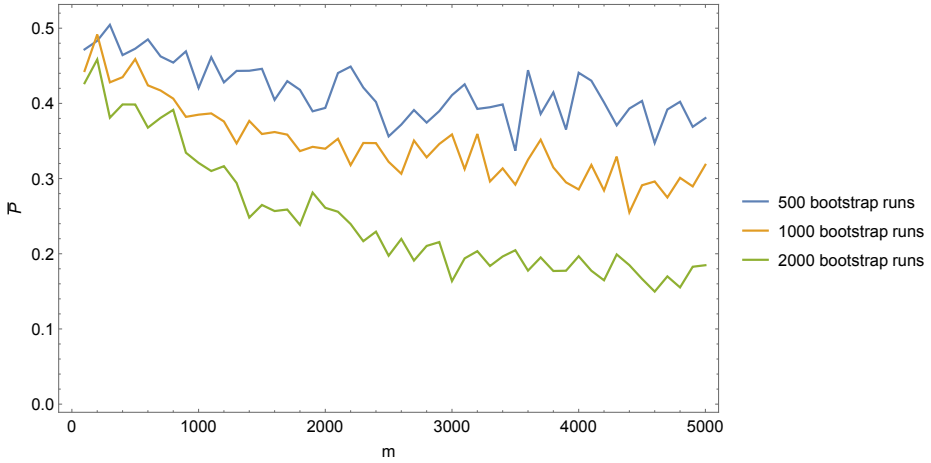


Figure 3: $\overline{p}$ vs. m at different values of the bootstrap runs.

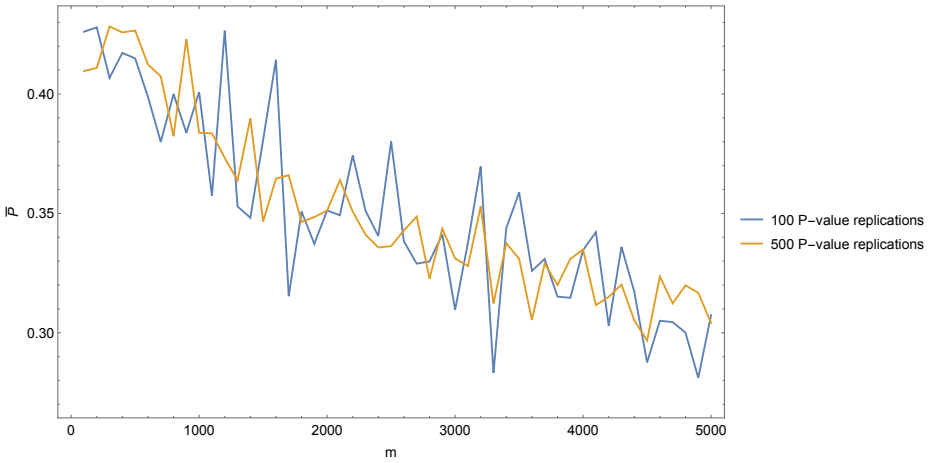


Figure 4: $\overline{p}$ vis m at different values of p-value replications.

5. CONCLUDING REMARKS

The bootstrap is an extremely flexible technique that can be applied to a wide variety of problems. One of the desired properties of the bootstrapping method is consistency, which guarantees that the limit of the bootstrap distribution is the same as that of the distribution of the given statistic.

In this paper, we investigated the strong consistency of bootstrapping central and intermediate order statistics for an appropriate choice of re-sample size for known and unknown normalizing constants. Consequently, inference concerning quartiles can now be performed by applying the bootstrap technique. For central order statistics, one can use the bootstrap to obtain a confidence interval for the p-th population quantile. On the other hand, it is well known that the asymptotic behavior of intermediate quantiles is one of the pillar factors in choosing a suitable value of threshold in the peak over threshold (POT) approach and the constructing related estimators (the Hill estimators) of the tail index (cf. [10, 14, 15, 26]. Therefore, the study of bootstrapping intermediate order statistics will pave the way to use and improve the modeling of extreme values via the POT approach. This potential application of the bootstrapping intermediate order statistics will be the subject of future studies.

The implemented simulation study in this paper aims to show how we choose numerically the values of m to give the best approximation (performance) of the bootstrapping DF for the central and intermediate quantiles. To our best knowledge, there is no such study was done in the literature even for extreme order statistics.

ACKNOWLEDGMENTS

We are grateful to the editor and anonymous referees for their valuable comments and helpful suggestions, which improved the presentation of this paper. We also thank El-Adll, M. E., for the participation in constructive discussions during the preparation of this work.

REFERENCES

[1] Athreya, K.B. and Fukuchi, J. (1994). *Bootstrapping extremes of i.i.d random variables.* In: "Proceedings of the Conference on Extreme Value Theory and Applications", held at NIST, Maryland, 23–29.

[2] Athreya, K.B. and Fukuchi, J. (1997). Confidence interval for end point of a c.d.f, via bootstrap, *Journal of Statistical Planning and Inference*, **58**(2), 299–320.

[3] Balkema, A.A. and de Haan, L. (1978). Limit distributions for order statistics I, *Theory of Probability & Its Applications*, **23**(1), 77–92.

[4] Balkema, A.A. and de Haan, L. (1978). Limit distributions for order statistics II, *Theory of Probability & Its Applications*, **23**(2), 341–358.

[5] Barakat, H.M. (1997). On the continuation of the limit distribution of the extreme and central terms of a sample, *Test*, **6**(23), 51–68.

[6] Barakat, H.M. (2003). On the restricted convergence of intermediate order statistics, *Probability and Mathematical Statistics – Wroclaw Univesity*, **23**(2), 229–240.

[7] Barakat, H.M. and El-Shandidy, M.A. (1990). Some limit theorems of intermediate term of a random number of independent random variables, *Commentationes Mathematicae Universitatis Carolinae*, **31**(2), 323–336.

[8] Barakat, H.M. and Omar, A.R. (2011). On limit distributions for intermediate order statistics under power normalization, *Mathematical Methods of Statistics*, **20**(4), 365–377.

[9] Barakat, H.M. and Omar, A.R. (2016). A note on domains of attraction of the limit laws of intermediate order statistics under power normalization, *Statistical Methodology*, **31**, 1–7.

[10] Barakat, H.M.; Nigm, E.M. and Alaswed, A.M. (2017). The Hill estimators under power normalization, *Applied Mathematical Modelling*, **45**, 813–822.

[11] Barakat, H.M.; Nigm, E.M. and El-Adll, M.E. (2011). Bootstrap for extreme generalized order statistics, *Arabian Journal for Science and Engineering*, **36**(6), 1083–1090.

[12] Barakat, H.M.; Nigm, E.M. and Harpy, M.H. (2020). Limit theorems for univariate and bivariate order statistics with variable ranks, *Statistics*, **54**(4), 737–755.

[13] Barakat, H.M.; Nigm, E.M. and Khaled, O.M. (2015). Bootstrap method for central and intermediate order statistics under power normalization, *Kybernetika*, **51**(6), 923–932.

[14] Barakat, H.M.; Nigm, E.M. and Khaled, O.M. (2019). *Statistical Techniques for Modelling Extreme Value Data and Related Applications*, Cambridge Scholars Publishing, London.

[15] Barakat, H.M.; Nigm, E.M.; Khaled, O.M. and Alaswed, A.M. (2018). The estimations under power normalization for the tail index, with comparison, *AStA Advances in Statistical Analysis*, **102**(3), 431–454.

[16] Barakat, H.M.; Nigm, E.M.; Khaled, O.M. and Momenkhan, F.A. (2015). Bootstrap method for order statistics and modeling study of the air pollution, *Communications in Statistics – Simulation and Computation*, **44**(6), 1477–1491.

[17] Bloch, D.A. and Gastwirth, J.L. (1968). On a simple estimate of the reciprocal of the density function, *The Annals of Mathematical Statistics*, **39**(3), 1083–1085.

[18] Chibisov, D.M. (1964). On limit distributions for order statistics, *Theory of Probability & Its Applications*, **9**(1), 142–148.

[19] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics*, Wiley, Hoboken.

[20] Eforn, B. (1979). Bootstrap methods: another look at the jackknife, *The Annals of Statistics*, **7**, 1–26.

[21] Falk, M. (1989). A note of uniform asymptotic normality of intermediate order statistics, *Annals of the Institute of Statistical Mathematics*, **41**(1), 19–29.

[22] Falk, M. and Wisheckel, F. (2018). Multivariate order statistics: the intermediate case, *Sankhya A*, **80**(1), 110–120.

[23] Frey, J. and Zhang, Y. (2017). What do interpolated nonparametric confidence intervals for population quantilesguarantee?, *The American Statistician*, **71**(4), 305–309.

[24] Fukuchi, J. (1994). *Bootstrapping Extremes of Random Variables*, Doctoral dissertation, Iowa State University.

[25] Geluk, J. and Haan, L. de (2002). On bootstrap sample size in extreme value theory, *Publications de l'Institut Mathematique*, **71**(85), 21–25.

[26] Haan, L. de and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*, Springer Series in Operations Research, New York.

[27] Ho, Y.H.S. and Lee, S.M.S. (2005). Iterated smoothed bootstrap confidence intervals for population quantiles, *The Annals of Statistics*, **33**(1), 437–462.

[28] Leadbetter, M.R.; Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*, Springer-Verlag, New York, Heidelberg.

[29] Lilliefors, H.W. (1967). On the Kolmogorov–Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association*, **62**(318), 399–402.

[30] Mason, D.M. (1982). Laws of large numbers for sums of extreme values, *The Annals of Probability*, **10**, 750–764.

[31] Nagaraja, C.H. and Nagaraja, H.N. (2019). Distribution-free approximate methods for constructing confidence intervals for quantiles, *International Statistical Review*, **88**(1), 5–100.

[32] Peng, L. and Yang, J. (2009). Jackknife method for intermediate quantiles, *Journal of Statistical Planning and Inference*, **139**(7), 2373–2381.

[33] Pickands, J. (1975). Statistical inference using extreme order statistics, *The Annals of Statistics*, **3**(1), 119–131.

[34] Siddiqui, M.M. (1960). Distribution of quantiles in samples from a bivariate population, *Journal of Research of the National Bureau of Standards*, **64**, 145–150.

[35] Smirnov, N.V. (1952). Limit distribution for terms of a variational series, *American Mathematical Society*, **11**, 82–143.

[36] Teugels, J.L. (1981). Limit theorems on order statistics, *The Annals of Probability*, **9**, 868–880.

[37] Wu, C.Y. (1966). The types of limit distributions for some terms of variational series, *Statistica Sinica*, **15**, 749–762.

REVSTAT — Statistical journal

Aims and Scope

The aim of REVSTAT — Statistical Journal is to publish articles of high scientific content, developing Statistical Science focused on innovative theory, methods, and applications in different areas of knowledge. Important survey/review contributing to Probability and Statistics advancement is also welcome.

Background

Statistics Portugal started in 1996 the publication of the scientific statistical journal *Revista de Estatística*, in Portuguese, a quarterly publication whose goal was the publication of papers containing original research results, and application studies, namely in the economic, social, and demographic fields. Statistics Portugal was aware of how vital statistical culture is in understanding most phenomena in the present-day world, and of its responsibilities in disseminating statistical knowledge.

In 1998 it was decided to publish papers in English. This step has been taken to achieve a larger diffusion, and to encourage foreign contributors to submit their work. At the time, the editorial board was mainly composed by Portuguese university professors, and this has been the first step aimed at changing the character of *Revista de Estatística* from a national to an international scientific journal. In 2001, the *Revista de Estatística* published a three volumes special issue containing extended abstracts of the invited and contributed papers presented at the 23rd European Meeting of Statisticians (EMS). During the EMS 2001, its editor-in-chief invited several international participants to join the editorial staff.

In 2003 the name changed to REVSTAT — Statistical Journal, published in English, with a prestigious international editorial board, hoping to become one more place where scientists may feel proud of publishing their research results.

Editorial policy

REVSTAT — Statistical Journal is an open access peer-reviewed journal published quarterly, in English, by Statistics Portugal.

The editorial policy of REVSTAT is mainly placed on the originality and importance of the research. The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage revstat.ine.pt based in Open Journal System (OJS). The only working language allowed is English. Authors intending to submit any work must register, login and follow the guidelines.

There are no fees for publishing accepted manuscripts that will be made available in open access.

All articles consistent with REVSTAT aims and scope will undergo scientific evaluation by at least two reviewers, one from the Editorial Board and another external. Authors can suggest

an editor or reviewer who is expert on the paper subject providing her/his complete information, namely: name, affiliation, email and, if possible, personal URL or ORCID number.

 Also, in the context of archiving policy, REVSTAT is a *blue* journal welcoming authors to deposit their works in other scientific repositories regarding the use of the published edition and providing its source.

Journal prints may be ordered at expenses of the author(s), and prior to publication.

Abstract and Indexing services

REVSTAT — *Journal Citation Reports - JCR (Clarivate); DOAJ-Directory of Open Access Journals; Current Index to Statistics; Google Scholar; Mathematical Reviews® (MathSciNet®); Zentralblatt für Mathematic; Scimago Journal & Country Rank; Scopus*

Author guidelines

The whole submission and review processes for REVSTAT are conducted exclusively online on the journal's webpage https://revstat.ine.pt/ based in Open Journal System (OJS). Authors intending to submit any work must ***register***, ***login*** and follow the indications choosing ***Submissions***.

REVSTAT — Statistical Journal adopts the COPE guidelines on publication ethics.

Work presentation

- the only working language is English;
- the first page should include the name, ORCID iD (optional), Institution, country, and email-address of the author(s);
- a summary of fewer than one hundred words, followed by a maximum of six keywords and the MSC 2020 subject classification should be included also in the first page;
- manuscripts should be typed only in black, in double-spacing, with a left margin of at least 3 cm, with numbered lines, and a maximum of 25 pages, in .pdf format;
- the title should be with no more than 120 characters (with spaces);
- figures must be a minimum of 300dpi and will be reproduced online as in the original work, however, authors should take into account that the printed version is always in black and grey tones;
- authors are encouraged to submit articles using LaTeX which macros are available at *REVSTAT style*;
- citations in text should be included in the text by name and year in parentheses, as in the following examples: § article title in lowercase (Author 1980); § This theorem was proved later by AuthorB and AuthorC (1990); § This subject has been widely addressed (AuthorA 1990; AuthorB et al. 1995; AuthorA and AuthorB 1998);
- references should be listed in alphabetical order of the author's scientific surname at the end of the article;

- acknowledgments of people, grants or funds should be placed in a short section before the References title page. Note that religious beliefs, ethnic background, citizenship and political orientations of the author(s) are not allowed in the text;
- authors are welcome to suggest one of the Editors or Associate Editors or yet other reviewer expert on the subject providing a complete information, namely: name, affiliation, email, and personal URL or ORCID number in the Comments for the Editor (submission form).

Accepted papers

After final revision and acceptance of an article for publication, authors are requested to provide the corresponding LaTeX file, as in REVSTAT style.

Supplementary files may be included and submitted separately in .tiff, .gif, .jpg, .png .eps, .ps or .pdf format. These supplementary files may be published online along with an article, containing data, programming code, extra figures, or extra proofs, etc; however, REVSTAT is not responsible for any supporting information supplied by the author(s).

Copyright Notice

Upon acceptance of an article, the author(s) will be asked to transfer copyright of the article to the publisher, Statistics Portugal, to ensure the widest possible dissemination of information.

According to REVSTAT's archiving policy, after assigning the copyright form, authors may cite and use limited excerpts (figures, tables, etc.) of their works accepted/published in REVSTAT in other publications and may deposit only the published edition in scientific repositories providing its source as REVSTAT while the original place of publication. The Executive Editor of the Journal must be notified in writing in advance.