Journal of Data Science and Modeling, Vol. 1, No. 1, 77-97, December 2022

Research Manuscript

Assessment, Estimation and Modeling of the Midpoint Coefficient for Imprecise Data

Amir Masoud Malekfar [1] , Farzad Eskandari[2,*]

[1]. PhD candidate, Department of Statistics, Allameh Tabataba'i University, Tehran, Iran.

2. Professor of statistics, Allameh Tabataba'i University, Tehran, Iran.

Received: 12/10/2018 Accepted: 28/10/2018

Abstract: Imprecise measurement tools produce imprecise data. Interval-valued data is usually used to deal with such imprecisions. Therefore interval-valued variables are used in estimation methods. Linear regression models have recently modeled them. If the response variable has any statistical distributions, interval-valued variables are modeled under the generalized linear models framework. In this article, we propose a new consistent estimator of a parameter in the generalized linear model with regard to distribution of the response variable in the exponential family. A simulation study shows that the new estimator is better than others based on particular distribution of the response variable. We present the optimal properties of the estimators in this research.

Keywords: Interval-valued data, Generalized linear models, Consistent estimator, Simulation, Optimal properties.

Mathematics Subject Classification (2010): 57R19, 57N65, 05E45, 62Gxx.

[*]Corresponding author:ffeskandari@yahoo.com

1. Introduction

Imprecise measurement tools are imprecise data producers. Imprecise-valued data and a desire for increased precision of data are two motives for statisticians to use interval-valued data rather than single-valued data. Diday (1987) revealed that singular-valued data, variably termed as classical or imprecise-valued data, will result in loss of information because singular-valued data depends on units of imprecise measurement. Therefore, Diday (1987, 1989, 1995), Émilion (1997), Bertrand and Goupil (2000), Billard and Diday (2002, 2003) and Billard and Diday (2006) proposed some methods for estimating, modeling and analyzing imprecise-valued data. Diday and Emilion (1996, 1996, 1998) used interval-valued data for collecting of imprecise-valued data. In recent years, Calle and Gómez (2001), Rivero and Valdes (2008), Trutschnig et al. (2009), Huber et al. (2009) and Billard (2011) have shown that interval-valued data is the best collecting method of big data and grouped data. Xu (2010) and Noirhomme - Fraiture and Brito (2011) showed that interval-valued data is difficult to analyze with classical methods.

Different analyzing and modeling methods of interval-valued data have recently been proposed as follows:

Billard and Diday (2000) introduced the Center Method (CM), which is built a linear regression model on centers of intervals. After obtaining parameter estimates, they applied the fitted model to both lower and upper bounds of a new observation to achieve an interval predicted response. This method concerns center points of interval-valued data only. Hence, high volumes of information are lost by the CM. Neto et al. (2004), de Carvalho et al. (2004) and Neto and de Carvalho (2008) introduced the Center and Range Method (CRM). The method utilizes not only centers but also ranges of intervals to fit regression models. Centers and ranges are separately used to do the fitting. Neto et al. (2005) and Neto and de Carvalho (2010) proposed the Constrained Center Method (CONCM). The regression model is the same as the CM under a restrictive condition. Neto et al. (2005) and Neto and de Carvalho (2010) proposed the Constrained Center and Range Method (CONCRM). The regression model is the same as the CRM under two restrictive conditions. The CRM assumes that centers and ranges are independent and fit models on them separately. In order to break assumption, Billard (2006) fitted centers and ranges simultaneously as a bivariate model, either with (BCRMI) or without (BCRMO) interaction terms between center and range variables (see also Neto et al., 2009). Blanco-Fernández et al. (2011) presented and estimated a more flexible simple linear model, the M model, between random interval variables. Wang et al. (2012) introduced the Complete Information Method (CIM). The CIM defines the inner product of interval-valued variables, and transforms the regression modeling into the computation of some inner products.

Diday and Emilion (1998) and Bertrand and Goupil (2000) introduced sample mean and variance of interval-valued data. Billard and Diday (2002, 2003) developed the method of observations with histogram-valued data. Gil et al. (2001, 2007), Billard (2007) and Neto et al. (2007) analyzed dependence versus independence on interval-valued data. A sample covariance was obtained by Billard (2007, 2008). Billard (2011) and Le-Rademacher and Billard (2012) completed the method.

In recent articles, interval-valued variables have been modeled by Linear Regression Models (LRMs). If the response variable has any statistical distributions in the exponential family, interval-valued variables will be modeled in Generalized Linear Models (GLMs) framework. This study aims to find a new consistent estimator of a parameter, with regard to distributions of the response variable, in GLMs. A new estimator of a parameter is proposed for fitting GLMs on the interval-valued data using a new structure of Monte Carlo Resampling (MCR).

Using this structure, we generate a large number of samples by randomly selecting a single-valued point within each observed interval of the interval-valued independent variables. We put each of the single-valued data in each of GLMs. After that, we generate single-valued points of response variable using the samples in each model. Then, we generate samples. We calculate the new and the old known estimator (Blanco-Fernández et al., 2011) based on each sample. Following that, we calculate the mean of the obtained estimated parameters from n, the sample size, repetitions.

Based on particular distributions of the response variables in the exponential family, the new estimator is considered to be more effective than the old known estimator (Blanco-Fernández et al., 2011) for fitting some GLMs on interval-valued data by using Monte Carlo Simulation (MCS). The Mean Square Error (MSE) is used as the method of evaluating the accuracy of the estimators. We introduce an asymptotic distribution of each estimator according to the Liapunov Theorem (LT) because each asymptotic distribution depends on n, the sample size. So we can not use the Central Limit Theorem (CLT).

Section 2 introduces the interval models of $Y = \alpha X^M + \beta X^S + \gamma + \varepsilon$. These models are GLMs. Definitions of symbols are proposed in Section 3. We propose a new structure of MCR in this section. Subsections 4.1-4.2 introduce two, one new and one old, estimators of α for fitting the models on interval-valued data based on the MCR structure and some particular distributions of ε. Following this, the MSE of each estimator will be presented. Section 5 studies two optimal properties, consistency, and asymptotic distribution, of the estimators in large-sample conditions. Subsections 5.1-5.2 introduce an asymptotic distribution of each estimator. Limited ranges of intervals and n samples taken from each interval, in large-sample conditions ($n \rightarrow \infty$), cause each asymptotic distribution depends on n. According to some particular distributions of ε, Section 6 assesses a better estimator from Section 4 for fitting the models on interval-valued data. The paper ends with the conclusion in Section 7.

2. Interval Models

From now on we will consider interval-valued experimental data belonging to the space $k_c(\mathbb{R}) = \{[a, b] : a, b \in \mathbb{R}, a \leq b\}$. Each interval $A \in k_c(\mathbb{R})$ is parametrized by means of a two-dimensional value, defined in terms of its endpoints, $(\inf A, \sup A) \in R^2$ with $\inf A \leq \sup A$. Equivalently, the point $(\text{mid } A, \text{ spr } A) \in R \times R^+$, where $\text{mid } A = \frac{\inf A + \sup A}{2}$ is the midpoint of the interval, and $\text{spr } A = \frac{\sup A - \inf A}{2}$ denotes the spread or radius, also characterizes the interval A. The notation $A = [\inf A, \sup A]$ or $A = [\text{mid } A \pm \text{spr } A]$ is used in each case.

Remark 2.1. The (inf, sup)- and the (mid, spr)-characterization for real intervals are the usual ones chosen for the treatment of interval data. Any other pair of points allowing the extremes of the interval to be computed, as the infimum and the amplitude, could be used. One useful characterization when the intervals are imprecise observations of a real-valued variable and it is possible to assume a distribution on that interval is the first and third quartile. Nevertheless, this is not the case considered in this work. The (inf , sup)-parametrization is not easy to use for statistical purposes given the order restriction that it involves. In this sense, it is more advisable to use the (mid, spr)-parametrization, since it only involves a non-negativity constraint on the second component, which is more operative. Additionally, the meaning of the (mid, spr)-description for interval data is very intuitive: the first component is related to the location of the interval and the second one to the imprecision (in the sense of the distance to a precise quantity of $\mathbb{R}$).

The formalization of the linear model for random intervals presented in Blanco-Fernández et al. (2011) is based on the (mid, spr)-representation of the intervals. The notation $A = [\text{mid } A \pm \text{ spr } A]$ can be split into two terms depending on the midpoint and spread values of A by means of the canonical decomposition $A = \text{mid } A\,[1 \pm 0] + \text{spr } A[0 \pm 1]$. This expression allows us to work separately with the mid and spr components of the interval, but keeping the interval arithmetic. The intervals $[1 \pm 0]$ and $[0 \pm 1]$ can be equivalently expressed in their (inf, sup)-representation as $[1, 1]$ and $[-1, 1]$, respectively.

Let X and Y be two random intervals with finite second-order moments, and spr X non-degenerated (so X is not reduced to a real random element). Based on the canonical decomposition, the generalized linear models between X and Y, the interval-valued random independent and response variable, is formalized as

$$Y = \alpha \text{ mid } X\,[1 \pm 0] + \beta \text{ spr } X\,[0 \pm 1] + \gamma\,[1 \pm 0] + \varepsilon, \tag{2.1}$$

where α and β are the coefficients, γ is an intercept term affecting the mid component of Y, and ε is an interval-valued random error variable such that $E(\varepsilon\,|X) = [-\delta,\delta]\in k_c(\mathbb{R})$, so that $\delta\geq 0$. For simpler notation, if we define $B = [\gamma-\delta,\gamma+\delta] \in k_c(\mathbb{R})$, the generalized linear function associated with the models 2.1 will be denoted by

$$E(Y\,|X\,) = \alpha\, X^M + \beta\, X^S + B,$$

where $X^M = X\,[1 \pm 0]$ and $X^S = X\,[0 \pm 1]$.

So based on the canonical decomposition, the linear models between X and Y is formalized as

$$Y = \alpha X^M + \beta X^S + \gamma + \varepsilon. \tag{2.2}$$

Model 2.2 are generalized linear models. Two specific assumptions

$$\begin{cases} H_0\text{: The consistency of } \overline{\overline{\alpha}} \\ H_1\text{: o.w} \end{cases}$$

are investigated according to Modes 2.2.

3. A New Structure of MCR

In this section, we propose a new structure of Monte Carlo Resampling (MCR). The structure is introduced to fit each of Models 2.2 on interval-valued data. One possible drawback of the approach is that it is computationally intensive. By the nature of MCS, a larger number of repetitions is always desired. The structure of MCR is implemented as follows: For i = 1, ..., k,

Symbols and formulas are introduced based on each of Models2.2 and Section 2 in this section.

Given one random sample $\{X_i^b\}_{i=1}^k$ from X. So we define

$$X^{*b} = \{X_1^b, \ldots, X_k^b\}, \tag{3.3}$$

where X_i^b is the bth observed sample of the ith interval of X. Hence $X_i^b \epsilon\ (\inf(X_i), \sup(X_i))$.

$$\operatorname{spr}\left(X^{*b}\right) = X^{r*b} = X^{S*b} = \{X_1^{Sb}, \ldots, X_k^{Sb}\}, \tag{3.4}$$

where X_i^{Sb} is the bth calculated value of X_i^S, the spread of the ith interval of X.

$$\operatorname{mid}(X^{*b}) = X^{M*b} = \{X_1^{Mb}, \ldots, X_k^{Mb}\}, \tag{3.5}$$

where X_i^{Mb} is the bth sample of the ith value of X^M.

$$Y^{*b} = \{Y_1^b, \ldots, Y_k^b\}, \tag{3.6}$$

where Y_i^b is obtained from $Y_i^b = \alpha X_i^{Mb}+\beta X_i^{Sb}+\gamma+\varepsilon_i^b$.

$$Y^{M*b}=\{Y_1^{Mb},\ldots,Y_k^{Mb}\},$$

where $Y_i^{Mb}= \alpha X_i^{Mb}+\gamma+\varepsilon_i^{Mb}$.

$$Y^{I*b}=\{Y_1^{Ib},\ \ldots,\ Y_k^{Ib}\},$$

where each sample $Y_i^{Ib}\epsilon[Y_i^b-0.0005,\ Y_i^b+0.0005]$.

When more than one sample is taken from each interval, we define:

$$X^*=\{X^{*b}\}_{b=1}^n=\left\{X^{*1},\ \ldots,\ X^{*n}\right\} = \{X_1^1,\ldots,X_k^1,\ldots,\ X_1^n,\ldots,X_k^n\}. \tag{3.7}$$

$$Y^*=\{Y^{*b}\}_{b=1}^n=\left\{Y^{*1},\ \ldots,\ Y^{*n}\right\} = \{Y_1^1,\ldots,Y_k^1,\ldots,\ Y_1^n,\ldots,Y_k^n\}. \tag{3.8}$$

$$\text{mid}\ (X^*) = X^{M*}=\{X^{M*b}\}_{b=1}^n=\{X_1^{M1},\ldots,X_k^{M1},\ldots,\ X_1^{Mn},\ldots,X_k^{Mn}\}. \tag{3.9}$$

$$\text{spr}(X^*) =X^{r*}=X^{S*}=\{X^{S*b}\}_{b=1}^n = \{X_1^{S1},\ldots,X_k^{S1}\ldots,\ X_1^{Sn},\ldots,X_k^{Sn}\}. \tag{3.10}$$

$$Y^{M*}=\{Y^{M*b}\}_{b=1}^n = \{Y_1^{M1},\ldots,Y_k^{M1},\ldots,Y_1^{Mn},\ldots,Y_k^{Mn}\}. \tag{3.11}$$

$$Y^{I*}=\{Y^{I*b}\}_{b=1}^n=\left\{Y_1^{I*1},\ \ldots,\ Y_k^{I*n}\right\}=\{Y_1^{I1},\ldots,Y_k^{I1},\ldots,\ Y_1^{In},\ldots,Y_k^{In}\}.$$

3.1 The Process of Producing Random Sample

Use the uniform distribution, randomly produce the bth single-valued data point X_i^b from $X_i=[a_i, c_i]$. Here b is the index of the bth sample and i is the index of the ith interval. X_i is the ith interval of the only predictor X. So $X_i^b\epsilon[a_i, c_i]$ for $i = 1, \ldots, k$, k is the number of the intervals, and $b = 1, \ldots, n$, n is the sample size. We generate each random sample X^{*b} according to Equation 3.3. We produce the random sample X^*, namely Equation 3.7, by using all of the samples X^{*b} for $b= 1, \ldots, n$. We randomly produce the bth single-valued data point X_i^{Mb} from $X_i^M=[d_i, d_i]$ for $i = 1, \ldots, k$ and $b= 1, \ldots, n$. X_i^M is the ith interval of X^M. So $X_i^{Mb}\epsilon[d_i, d_i]$ for $i = 1, \ldots, k$ and $b = 1, \ldots, n$. We generate each random sample X^{M*b} according to Equation 3.5. We produce the random sample X^{M*}, namely Equation 3.9, by using all of the random samples X^{M*b}. We produce all values of Y in each of Models 2.2. Also, the bth single-valued data point Y_i^b is generated, in each of Models 2.2, according to the bth single-valued data point X_i^b. We generate each random sample Y^{*b} according to Equation 3.6. We produce the random sample Y^* (see Equations 3.8, 3.10 and 3.11) by using all of the samples Y^{*b} for $b = 1, \ldots, n$.

We randomly generate the bth single-valued data point $Y_i^{Ib}\epsilon[Y_i^b-0.0005,\ Y_i^b+0.0005]$ for $i = 1, \ldots, k$ and $b = 1, \ldots, n$. Hence we generate each random sample $Y^{I*b}=\{Y_1^{Ib}, \ldots, Y_k^{Ib}\}$. We produce the random sample $Y^{I*}=\{Y_1^{I1},\ldots,Y_k^{I1},\ldots,\ Y_1^{In},\ldots,Y_k^{In}\}$, by using all of the samples Y^{I*b} for b = 1, . . . , n. Difference of Y and Y^{I} value is very small. This paper will show that the difference is useful.

Produce the bth random sample:

$$Y^{*b}=\{Y_1^b,\dots,Y_k^b\},\ Y^{I*b}=\{Y_1^{Ib},\dots,Y_k^{Ib}\}, X^{*b}=\{X_1^b,\dots,X_k^b\}, IX^{*b}=\{1,X_1^b,\dots,X_k^b\}$$

$$X^{M*b}=\{X_1^{Mb},\dots,X_k^{Mb}\},\ \ IX^{M*b}=\{1,X_1^{Mb},\dots,X_k^{Mb}\}, X^{S*b}=\{X_1^{Sb},\dots,X_k^{Sb}\}.$$

So we produce:

$$Y^{*}=\{Y_1^1,\dots,Y_k^1,\dots,Y_1^n,\dots,Y_k^n\},$$
$$Y^{I*}=\{Y_1^{I1},\dots,Y_k^{I1},\dots,Y_1^{In},\dots,Y_k^{In}\}, X^{*}=\{X_1^1,\dots,X_k^1,\dots,X_1^n,\dots,X_k^n\},$$
$$X^{M*}=\{X_1^{M1},\dots,X_k^{M1},\dots,X_1^{Mn},\dots,X_k^{Mn}\},\ X^{S*}=\{X_1^{S1},\dots,X_k^{S1},\dots,X_1^{Sn},\dots,X_k^{Sn}\}.$$

4. A Proposal Method

Based on the new structure of MCR in Section 3 and the estimator of ω_1 in the simple linear regression model Y $=\omega_0+\omega_1X+\varepsilon$, Subsection 4.2 proposes a new estimator of α, namely Estimator 4.14, to fit each of Models 2.2 on interval-valued data. Also Subsection 4.1 re-introduces Estimator 4.12 as the old known estimator of α. This is the best estimator of α which has ever been proposed by Blanco-Fernández et al. (2011) based on Y and X^M to fit each of the models on interval-valued data.

Is Estimator 4.12 better than Estimator 4.14 for all distributions of ε in Model 2.2?

The effectiveness of ε distributions in finding a better estimator is shown in Tables 1-2 and Figures 1-2. Subsections 4.1-4.2 introduce the estimation procedures to achieve two estimators.

4.1 The Old Known Estimator

This subsection re-introduces Estimator 4.12. We provide the estimator according to the random samples Y^{*b} and X^{M*b} (see Sections 2-3) for b=1, . . . , n. Also, we re-introduce Estimator 4.13 based on Y^{*} and X^{M*} (see Sections 2-3). In Section 6, Table 1 shows some particular distributions of ε that Estimator 4.12 is better than Estimator 4.14. We re-introduce:

$$\bar{\hat{\alpha}}=\frac{1}{n}\sum_{b=1}^{n}\hat{\alpha}^b=\frac{1}{n}\sum_{b=1}^{n}\left(\frac{\mathrm{Cov}\left(Y^{*b},\ X^{M*b}\right)}{\mathrm{Var}\,(X^{M*b})}\right) \tag{4.12}$$

$$\overline{\widehat{\begin{pmatrix}\gamma\\ \alpha\end{pmatrix}}}=\frac{1}{n}\sum_{b=1}^{n}\widehat{\begin{pmatrix}\gamma^{b}\\ \alpha^{b}\end{pmatrix}}=\frac{1}{n}\sum_{b=1}^{n}\left\{\begin{pmatrix}1 & X_1^{Mb}\\ & .\\ & .\\ & .\\ 1 & X_k^{Mb}\end{pmatrix}^{T}.\begin{pmatrix}1 & X_1^{Mb}\\ & .\\ & .\\ & .\\ 1 & X_k^{Mb}\end{pmatrix}\right\}^{-1}.\begin{pmatrix}1 & X_1^{Mb}\\ & .\\ & .\\ & .\\ 1 & X_k^{Mb}\end{pmatrix}^{T}.\begin{pmatrix}Y_1^{b}\\ .\\ .\\ .\\ Y_k^{b}\end{pmatrix}=$$

$$\left\{\begin{pmatrix}1 & X_1^{M1}\\ & .\\ & .\\ 1 & X_k^{M1}\\ & .\\ & .\\ 1 & X_1^{Mn}\\ & .\\ & .\\ 1 & X_k^{Mn}\end{pmatrix}^{T}.\begin{pmatrix}1 & X_1^{M1}\\ & .\\ & .\\ 1 & X_k^{M1}\\ & .\\ & .\\ 1 & X_1^{Mn}\\ & .\\ & .\\ 1 & X_k^{Mn}\end{pmatrix}\right\}^{-1}.\begin{pmatrix}1 & X_1^{M1}\\ & .\\ & .\\ 1 & X_k^{M1}\\ & .\\ & .\\ 1 & X_1^{Mn}\\ & .\\ & .\\ 1 & X_k^{Mn}\end{pmatrix}^{T}\begin{pmatrix}Y_1^{1}\\ .\\ .\\ Y_k^{1}\\ .\\ .\\ Y_1^{n}\\ .\\ .\\ Y_k^{n}\end{pmatrix}=\begin{pmatrix}\overline{\widehat{\gamma}}\\ \frac{\widehat{\sigma}_{Y*,X^{M*}}}{\widehat{\sigma}^{2}_{X^{M*}}}\end{pmatrix} \qquad (4.13)$$

Estimators 4.12 and 4.13 have a same answer of $\overline{\widehat{\alpha}}$. The MSE of Estimator 4.12 is:

$$\mathrm{MSE}\left(\overline{\widehat{\alpha}}\right)=\left(\frac{1}{n}\sum_{b=1}^{n}\left(\frac{\mathrm{Cov}\left(Y^{*b},X^{M*b}\right)}{\mathrm{Var}\left(X^{M*b}\right)}\right)-\alpha\right)^{2}+\mathrm{Var}\left(\frac{1}{n}\sum_{b=1}^{n}\left(\frac{\mathrm{Cov}\left(Y^{*b},X^{M*b}\right)}{\mathrm{Var}\left(X^{M*b}\right)}\right)\right).$$

It is introduced according to the MCR structure in Section 3.

4.2 A New Estimator

This Subsection introduces Estimator4.14 as a new estimator of α for fitting each of Model 2.2 on interval-valued data. The estimator is proposed based on the random samples X^{M*b} and Y^{I*b} (see Sections 2-3) for $b = 1, \ldots, n$. Also, we propose Estimator 4.15 based on Y^{I*} and X^{M*} (see Sections 2-3). In Section 6, Table 2 shows some particular distributions of ε that Estimator 4.14 is better than Estimator 4.12. We propose:

$$\overline{\widehat{\alpha}}=\frac{1}{n}\sum_{b=1}^{n}\widehat{\alpha}^{b}=\frac{1}{n}\sum_{b=1}^{n}\left(\frac{\mathrm{Cov}\left(Y^{I*b},X^{M*b}\right)}{\mathrm{Var}\left(X^{M*b}\right)}\right). \qquad (4.14)$$

$$\overline{\widehat{\begin{pmatrix}\gamma\\ \alpha\end{pmatrix}}}=\frac{1}{n}\sum_{b=1}^{n}\widehat{\begin{pmatrix}\gamma^{b}\\ \alpha^{b}\end{pmatrix}}=\frac{1}{n}\sum_{b=1}^{n}\left\{\left(\begin{pmatrix}1 & X_1^{Mb}\\ & .\\ & .\\ & .\\ 1 & X_k^{Mb}\end{pmatrix}^{T}\cdot\begin{pmatrix}1 & X_1^{Mb}\\ & .\\ & .\\ & .\\ 1 & X_k^{Mb}\end{pmatrix}\right)^{-1}\right\}\cdot\begin{pmatrix}1 & X_1^{Mb}\\ & .\\ & .\\ & .\\ 1 & X_k^{Mb}\end{pmatrix}^{T}\cdot\begin{pmatrix}Y_1^{Ib}\\ .\\ .\\ .\\ Y_k^{Ib}\end{pmatrix}=$$

$$\left\{\begin{pmatrix}1 & X_1^{M1}\\ & .\\ & .\\ 1 & X_k^{M1}\\ & .\\ & .\\ 1 & X_1^{Mn}\\ & .\\ & .\\ 1 & X_k^{Mn}\end{pmatrix}^{T}\cdot\begin{pmatrix}1 & X_1^{M1}\\ & .\\ & .\\ 1 & X_k^{M1}\\ & .\\ & .\\ 1 & X_1^{Mn}\\ & .\\ & .\\ 1 & X_k^{Mn}\end{pmatrix}\right\}^{-1}\cdot\begin{pmatrix}1 & X_1^{M1}\\ & .\\ & .\\ 1 & X_k^{M1}\\ & .\\ & .\\ 1 & X_1^{Mn}\\ & .\\ & .\\ 1 & X_k^{Mn}\end{pmatrix}^{T}\begin{pmatrix}Y_1^{I1}\\ .\\ .\\ Y_k^{I1}\\ .\\ .\\ Y_1^{In}\\ .\\ .\\ Y_k^{In}\end{pmatrix}=\begin{pmatrix}\overline{\widehat{\gamma}}\\ \frac{\widehat{\sigma}_{Y^{I*},X^{M*}}}{\widehat{\sigma}^{2}_{X^{M*}}}\end{pmatrix} \tag{4.15}$$

Estimators 4.14 and 4.15 have a same answer of $\overline{\widehat{\alpha}}$. Based on the MCR structure in Section 3, the MSE of Estimator 4.14 is:

$$\mathrm{MSE}\left(\overline{\widehat{\alpha}}\right)=\left(\frac{1}{n}\sum_{b=1}^{n}\left(\frac{\mathrm{Cov}\left(Y^{I*b},X^{M*b}\right)}{\mathrm{Var}\left(X^{M*b}\right)}\right)-\alpha\right)^{2}+\mathrm{Var}\left(\frac{1}{n}\sum_{b=1}^{n}\left(\frac{\mathrm{Cov}\left(Y^{I*b},X^{M*b}\right)}{\mathrm{Var}\left(X^{M*b}\right)}\right)\right).$$

Two tables in Section 6, Tables 1-2, show that MSE values of the estimators depend on distributions of ε in Model 2.2. Also the mentioned section represents that under particular distributions of ε, each of Estimators 4.12 and 4.14 is a better estimator with the least MSE value in comparison to the other one.

5. Optimal Properties of $\overline{\widehat{\alpha}}$ in Large-Sample Conditions

In this section, we utilize a resampling idea (see Section 3) to fit Model 2.2 on interval-valued data. It not only accounts for internal variations of intervals but also derives an approximate sampling distribution of each estimator via the new structure of MCR.

In this section, consistency of Estimators 4.12, 4.14 and asymptotic distribution of each estimator are studied in large-sample conditions, $n\to\infty$. Subsections 5.1-5.2 introduce an asymptotic distribution of each of the estimators according to Sections 2-3 and some particular distributions of ε.

This study uses the LT instead of the CLT. Limited ranges of intervals and n samples taken from each interval, in large-sample conditions (n→ ∞), cause each asymptotic distribution depends on n.

5.1 Optimal Properties of the Old Known Estimator

We study two optimal properties, consistency and asymptotic distribution, of Estimator 4.12 in large-sample conditions, n→∞, in this subsection.

Theorem 5.1. *Estimator 4.12 is a consistent estimator.*

Proof. We know, Y^{M*} and Y^{S*} (see Sections 2-3) are the center and the spread of Y^* variable. Also ε^{M*} and ε^{S*} (see Section 2) are the center and the spread of ε^* variable.

$X^{M*}$is independent of X^{S*}, Y^{S*}, ε^{S*} and ε^{M*}. We know $\overline{\widehat{\alpha}} = \frac{\widehat{\sigma}_{Y*,X^{M*}}}{\widehat{\sigma}^2_{X^{M*}}}$, so $\overline{\widehat{\alpha}}-\alpha=\frac{\widehat{\sigma}_{Y*,X^{M*}}}{\widehat{\sigma}^2_{X^{M*}}}-\alpha$. Hence:

$$\mathrm{Cov}\left(Y^*,X^{M*}\right)=\mathrm{Cov}\left(Y^{M*}+Y^{S*},X^{M*}\right)=\mathrm{Cov}\left(Y^{M*},X^{M*}\right)=\mathrm{Cov}\left(\alpha X^{M*}+\varepsilon^{M*},X^{M*}\right)$$

$$=\alpha\sigma^2_{X^{M*}}+\mathrm{Cov}\left(\varepsilon^{M*},X^{M*}\right).$$

We can write $\overline{\widehat{\alpha}}-\alpha=\frac{\mathrm{Cov}(\varepsilon^{M*},\quad X^{M*})}{\sigma^2_{X^{M*}}}$, therefore $E\left(\overline{\widehat{\alpha}}-\alpha\right)=E\left(\frac{\mathrm{Cov}(\varepsilon^{M*},\quad X^{M*})}{\sigma^2_{X^{M*}}}\right)=0$, since X^{M*} is independent of ε^{M*}. Also $\mathrm{Var}\left(\overline{\widehat{\alpha}}\right)=\frac{1}{n}\mathrm{Var}\left(\frac{\mathrm{Cov}(\varepsilon^{M*},X^{M*})}{\sigma^2_{X^{M*}}}\right)\rightarrow 0$ as $n\rightarrow\infty$. So $\overline{\widehat{\alpha}}$, namely Estimator 4.12, is a consistent estimator of α. □

Based on Sections 2-3, the variance of the consistent estimator is:

$$\mathrm{var}\left(\widehat{\alpha}\right)=\frac{\mathrm{var}\left(Y^*\right)}{n.\mathrm{var}\left(X^{M*}\right)} \tag{5.16}$$

Theorem 5.2. *Under conditions of Model 2.2, if E(*$\|\mathrm{mid}\ X, \mathrm{mid}\ Y\|^4$*)*$<\infty$ *(where* $\|.\|$ *denotes the Euclidean norm) and* $0<\frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}}<\infty$*, then*

$$\sqrt{n}\left(\overline{\widehat{\alpha}}-\alpha\right)\xrightarrow{L}N\left(0,\frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}}\right) \tag{5.17}$$

Distribution 5.17 is an asymptotic distribution of the old estimator. The distribution is rewritten according to Section 3 and Equation 5.18. In this subsection, a definition of ε^{M*} *variable is:*

$$\varepsilon^{M*}=((\varepsilon^{M*1})^T,\ \ldots,(\varepsilon^{M*n})^T)^T \tag{5.18}$$

Where

$$\varepsilon^{M*b} = \text{mean}\left(Y^{*b}\right) - \left(\frac{\text{cov}\left(Y^{*b}, X^{M*b}\right)}{\text{var}\,(X^{M*b})} . \left(\left(X^{M*b}\right)\right) + \bar{\hat{\gamma}}\right) \qquad \text{for } b = 1, \ldots, n.$$

. . . $\bar{\hat{\gamma}}$ *is offered in Equation 4.13.*

Definition 1.

We know $\hat{\alpha} = \left(\hat{\alpha}^1, \ldots, \hat{\alpha}^n\right)$, where $\hat{\alpha}^b$ is the bth value of $\hat{\alpha}$ according to the bth set of the data source (see Estimator 4.12).

We provide the LT condition according to Definition 1. The LT condition, namely Equation 5.19, for establishment of Distribution 5.17 is as follows:

$$\left[E(\sum_{b=1}^{n} \left|\hat{\alpha}^b - -\alpha\right|^3)\right]^2 = o(\sum_{b=1}^{n} (\frac{\sigma_{Y*b}}{\sqrt{\sigma^2_{X^{M*b}}}})^2)^3 \tag{5.19}$$

We propose two sides of the distribution by establishing the condition.

Proof. Let us write according to Theorem 5.1

$$\bar{\hat{\alpha}} - \alpha = \frac{\text{Cov}\left(\varepsilon^{M*}, \quad X^{M*}\right)}{\sigma^2_{X^{M*}}},$$

$$E\left(\sqrt{n}\ \left(\bar{\hat{\alpha}} - \alpha\right)\right) = \sqrt{n}\ .E\left(\bar{\hat{\alpha}} - \alpha\right) = 0 \ \ \textit{as} \ \ n \to \infty,$$

$$\text{Var}\left(\sqrt{n}\ \left(\bar{\hat{\alpha}} - \alpha\right)\right) = n\left(\frac{\sigma^2_{\varepsilon^{M*}}}{n.\sigma^2_{X^{M*}}}\right) = \frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}} \ \ \textit{as} \ \ n \to \infty.$$

□

5.2 Optimal Properties of the New Estimator

Let's study two optimal properties, consistency and asymptotic distribution, of Estimator 4.14 in large-sample conditions, n→∞.

Theorem 5.3. *Estimator 4.14 is a consistent estimator.*

Proof. We know, Y^{M*} and Y^{S*} (see Sections 2-3) are the center and the spread of Y^* variable. Also ε^{M*} and ε^{S*} (see Section 2) are the center and the spread of ε^* variable. Difference of Y^* and Y^{I*} are defined as e^*. We can say, e^* is an imposed error variable. So e^{M*} and e^{S*} are the center and the spread of e^*.

$X^{M*}$is independent of X^{S*}, Y^{S*}, e^{S*}, e^{M*}, ε^{S*}, ε^{M*} and ε^{*}. We know $\overline{\widehat{\alpha}} = \frac{\widehat{\sigma}_{Y^{I*},X^{M*}}}{\widehat{\sigma}^2_{X^{M*}}}$, so $\overline{\widehat{\alpha}}{-}\alpha{=}\ \frac{\widehat{\sigma}_{Y^{I*},X^{M*}}}{\widehat{\sigma}^2_{X^{M*}}}{-}\alpha$. Hence:

$$\begin{aligned}\operatorname{Cov}\left(Y^{I*},X^{M*}\right) &= \operatorname{Cov}\left(Y^{M*}{+}Y^{S*}+e^{M*}+e^{S*},X^{M*}\right) = \operatorname{Cov}\left(Y^{M*},X^{M*}\right)\\ &= \operatorname{Cov}\left(\alpha X^{M*}{+}\varepsilon^{M*},X^{M*}\right){=}\alpha\sigma^2_{X^{M*}}{+}\operatorname{Cov}\left(\varepsilon^{M*},X^{M*}\right).\end{aligned}$$

We have $\overline{\widehat{\alpha}}{-}\alpha{=}\frac{\operatorname{Cov}(\varepsilon^{M*},\quad X^{M*})}{\sigma^2_{X^{M*}}}$, therefore $E\left(\overline{\widehat{\alpha}}{-}\alpha\right) = E\left(\frac{\operatorname{Cov}(\varepsilon^{M*},\quad X^{M*})}{\sigma^2_{X^{M*}}}\right) = 0$, since X^{M*} is independent of ε^{M*}. Also $\operatorname{Var}\left(\overline{\widehat{\alpha}}\right){=}\frac{1}{n}\operatorname{Var}\left(\frac{\operatorname{Cov}(\varepsilon^{M*},X^{M*})}{\sigma^2_{X^{M*}}}\right) = \frac{1}{n}(\frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}}) \to 0$ as $n \to \infty$. So $\overline{\widehat{\alpha}}$, namely Estimator 4.14, is a consistent estimator of α. □

Based on Sections 2-3, the variance of the consistent estimator is:

$$\operatorname{var}\left(\overline{\widehat{\alpha}}\right){=}\frac{\operatorname{var}\left(Y^{I*}\right)}{n.\operatorname{var}\left(X^{M*}\right)} \tag{5.20}$$

Theorem 5.4. *Under conditions of Model 2.2, if* $E(\left\|\text{mid } X, \text{mid } Y^{I}\right\|^{4}){<}\ \infty$ *(where* $\|.\|$ *denotes the Euclidean norm) and* $0 < \frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}} < \infty$*, the*

$$\sqrt{n}\ \left(\overline{\widehat{\alpha}}{-}\alpha\right)\xrightarrow{L}N\left(0\,,\frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}}\right) \tag{5.21}$$

Distribution 5.21 is an asymptotic distribution of the new estimator. Distribution 5.21 is written according to Section 3 and Equation 5.22. In this subsection, a definition of ε^{M*} variable is:

$$\varepsilon^{M*}{=}((\varepsilon^{M*1})^{T},\ \ldots{,}(\varepsilon^{M*n})^{T})^{T} \tag{5.22}$$

Where

$$\varepsilon^{M*b}{=}\ \operatorname{mean}\left(Y^{I*b}\right)-\left(\frac{\operatorname{cov}\left(Y^{I*b},X^{M*b}\right)}{\operatorname{var}\left(X^{M*b}\right)}.\left(\left(X^{M*b}\right)\right)+\overline{\widehat{\gamma}}\right) \qquad \text{for } b = 1,\ \ldots,\ n.$$

$\overline{\widehat{\gamma}}$ is offered in Equation 4.15. Based on Definition 1, the LT condition (see Equation 5.23) for establishment of the asymptotic distribution is as follows:

$$\left[E(\sum_{b=1}^{n}\left|\widehat{\alpha}^{b}--\alpha\right|^{3})\right]^{2} = o(\sum_{b=1}^{n}(\frac{\sigma_{Y^{I*b}}}{\sqrt{\sigma^2_{X^{M*b}}}})^{2})^{3} \tag{5.23}$$

By establishing this condition, two sides of the asymptotic distribution are created.

Proof. of theorem 5.4. Let us write according to Theorem 5.3

$$\overline{\widehat{\alpha}}-\alpha=\frac{\mathrm{Cov}\left(\varepsilon^{M*},X^{M*}\right)}{\sigma^2_{X^{M*}}},$$

$$E\left(\sqrt{n}\left(\overline{\widehat{\alpha}}-\alpha\right)\right)=\sqrt{n}\,.E\left(\overline{\widehat{\alpha}}-\alpha\right)=0 \;\; \mathit{as} \;\; n\rightarrow\infty,$$

$$\mathrm{Var}\left(kk_n\left(\overline{\widehat{\alpha}}-\alpha\right)\right) \;=n\left(\frac{\sigma^2_{\varepsilon^{M*}}}{n.\sigma^2_{X^{M*}}}\right)=\frac{\sigma^2_{\varepsilon^{M*}}}{\sigma^2_{X^{M*}}} \;\; \mathit{as} \;\; n\rightarrow\infty.$$

□

6. A Simulation Study

We compare the estimators for each of Model 2.2, via simulated data sets, which are generated in the new structure of MCR (see Section 3). Let X, X^M, X^S be the interval-valued independent variable, the center of X, the spread of X, and Y be the interval-valued response variable. We consider one set of coefficients (α, β, γ) = (0.6, 0.2, 0) in each of Models 2.2. We assess the accuracy of the estimators of α in each model. X_8 (see artificialeg data from mplot package in R software) is as the center of X. By using the uniform distributions, we randomly generate a single-valued data point X_i^b from the i^{th} interval of the predictor X = (X_8−0.0005, X_8+0.0005) for i = 1, . . . , 50 and b = 1, . . . , 1000. The MCR structure is implemented according to 50 separate intervals of X and 1000 are samples taken from each of the intervals. We generate each random sample X^{*b} for each value b according to the MCR structure in Section 3. The remaining random samples are created based on the MCR structure in Section 3.

A simulation study shows that the correlation between any two intervals is almost zero and the internal correlation of each interval is almost one. Since 1000 samples are taken from each interval, the ranges of the intervals are equal to 0.001, depend on each other.

We summarize some simulation results in Tables 1-2. In each table, we report the mean of the obtained midpoint coefficients from 1000 repetitions based on each ε distribution in Models2.2.

Which of the consist estimators of α is better than the other one according to all distributions of ε?

The MSE column of Table 1 shows Estimator 4.12 is better than the other one according to some particular distributions of ε, in the last column of the table, in Models 2.2.

The MSE column of Table 2 shows that Estimator (14) is better than the other one according to some particular distributions of ε, in the last column of the table, in Models 2.2.

Table 1: The superiority of Estimator 4.12 based on some particular distributions of $\boldsymbol{\varepsilon}$

Estimator	$\overline{\widehat{\boldsymbol{\alpha}}}$	**Bias**($\overline{\widehat{\boldsymbol{\alpha}}}$)	**Var** ($\overline{\widehat{\boldsymbol{\alpha}}}$)	**MSE** ($\overline{\widehat{\boldsymbol{\alpha}}}$)	$\boldsymbol{\varepsilon}$
(4.12)	0.6016	0.0016	0.0018	0.0018	F(1,25,0)
(4.14)	0.5265	-0.0734	5e-11	0.0054	F(1,25,0)
(4.12)	0.5985	-0.0014	0.0042	0.0042	F(50,5)
(4.14)	0.5338	-0.0661	5e-11	0.0043	F(50,5)
(4.12)	0.5996	-0.0003	0.0001	0.0001	Gamma(1,2)
(4.14)	0.5800	-0.0199	5e-11	0.0003	Gamma(1,2)
(4.12)	0.6027	0.0027	0.0333	0.033	Gamma(0.5,0.1)
(4.14)	0.2925	-0.3074	5e-11	0.0945	Gamma(0.5,0.1)
(4.12)	0.5996	-0.0003	1e-05	1e-05	Beta(5,3)
(4.14)	0.6077	0.00771	5e-11	6e-05	Beta(5,3)
(4.12)	0.5996	-0.0003	1e-05	1e-05	Beta(5,5)
(4.14)	0.5897	-0.0102	5e-11	0.0001	Beta(5,5)
(4.12)	0.6001	0.0001	1e-05	1e-05	Beta(5,1)
(4.14)	0.5954	-0.0045	5e-11	2e-05	Beta(5,1)

None of the above tables, Tables 1-2, will show the absolute superiority of one of the estimators over another one.

Figure 1 shows the behaviors of the MSE of two estimators according to some particular distributions of ε (see Table 1). Estimator 4.12 is better than the other one in Figure 1.

Figure 2 shows the behaviors of the MSE of two estimators based on some particular distributions of ε (see Table 2). Estimator 4.14 is better than the other one in Figure 2. So none of two consistent estimators of α is better than the other one based on all distributions of ε.

7. Conclusions

In the last decade, interval-valued variables have been modeled by LRMs. If the response variable has any statistical distributions, interval-valued variables will be modeled in the GLMs framework.

One possible drawback of the new MCR structure is that it is computationally intensive. By the nature of MCS, a larger number of repetitions is always desired. However, we believe that its good properties outweigh this disadvantage. First, it relieves the need

Table 2: The superiority of Estimator (4.14) based on some particular distributions of $\boldsymbol{\varepsilon}$

Estimator	$\overline{\overline{\alpha}}$	**Bias**$(\overline{\overline{\alpha}})$	**Var** $(\overline{\overline{\alpha}})$	**MSE** $(\overline{\overline{\alpha}})$	$\boldsymbol{\varepsilon}$
(4.12)	0.600734	0.000734	0.0007042	0.000704	N(0,1)
(4.14)	0.5909	-0.0090	5e-11	8e-05	N(0,1)
(4.12)	0.6001	0.0001	2e-05	2e-05	Beta(3,0.6)
(4.14)	0.5964	-0.0035	5e-11	1e-05	Beta(3,0.6)
(4.12)	0.4563	-0.1436	105.6827	105.7034	C (0,1)
(4.14)	5.8006	5.2006	5e-11	27.0464	C (0,1)
(4.12)	0.6001	0.0001	2e-05	2e-05	F(100,200,10)
(4.14)	0.6007	0.00075	5e-11	5e-07	F(100,200,10)
(4.12)	0.5988	-0.0011	0.0004	0.0004	F(10,200,10)
(4.14)	0.5795	-0.0204	5e-11	0.0004	F(10,200,10)
(4.12)	0.5996	-0.0003	0.0003	0.0003	F(200,10,10)
(4.14)	0.5933	-0.0066	5e-11	4e-05	F(200,10,10)
(4.12)	0.60003	3e-05	1e-06	1e-06	Ln(0.046,0.046)
(4.14)	0.5995	-0.0004	5e-11	1e-07	Ln(0.046,0.046)
(4.12)	0.5999	-7e-05	0.0010	0.001	χ^2 (0.25,0.25)
(4.14)	0.6149	0.0149	5e-11	0.0002	χ^2 (0.25,0.25)
(4.12)	0.6002	0.0002	0.0001	0.0001	Beta(0.01,0.01)
(4.14)	0.6113	0.0113	5e-11	0.0001	Beta(0.01,0.01)
(4.12)	0.6000	7e-05	7e-06	7e-06	N(0,0.1)
(4.14)	0.5990	-0.0009	5e-11	8e-07	N(0,0.1)
(4.12)	0.6000	1e-05	2e-06	2e-06	Gamma(0.1,5)
(4.14)	0.5999	-1e-05	5e-11	1e-10	Gamma(0.1,5)
(4.12)	0.6000	4e-05	6e-07	6e-07	Exp(30)
(4.14)	0.5996	-0.0003	5e-11	1e-07	Exp(30)
(4.12)	810.1115	809.5115	582010775	582666084	F(1,1)
(4.14)	-1.8195	-2.4195	5e-11	5.85444	F(1,1)
(4.12)	0.6028	0.0028	0.0150	0.0150	Bin(500,0.95)
(4.14)	0.5915	-0.0084	5e-11	7e-05	Bin(500,0.95)
(4.12)	0.5998	-0.0001	0.0001	0.0001	Geo(0.85)
(4.14)	0.5972	-0.0027	4e-11	7e-06	Geo(0.85)
(4.12)	0.6002785	0.0002	0.0001	0.0001	Beta(0.2,0.2)
(4.14)	0.6060	0.0060	5e-11	3e-05	Beta(0.2,0.2)
(4.12)	0.5994	-0.0005	7e-05	7e-05	Beta(0.7,0.7)
(4.14)	0.6000	2e-05	5e-11	6e-10	Beta(0.7,0.7)

Figure 1: Assessment of the MSE of two estimators according to each distribution of $\boldsymbol{\varepsilon}$ from Table 1

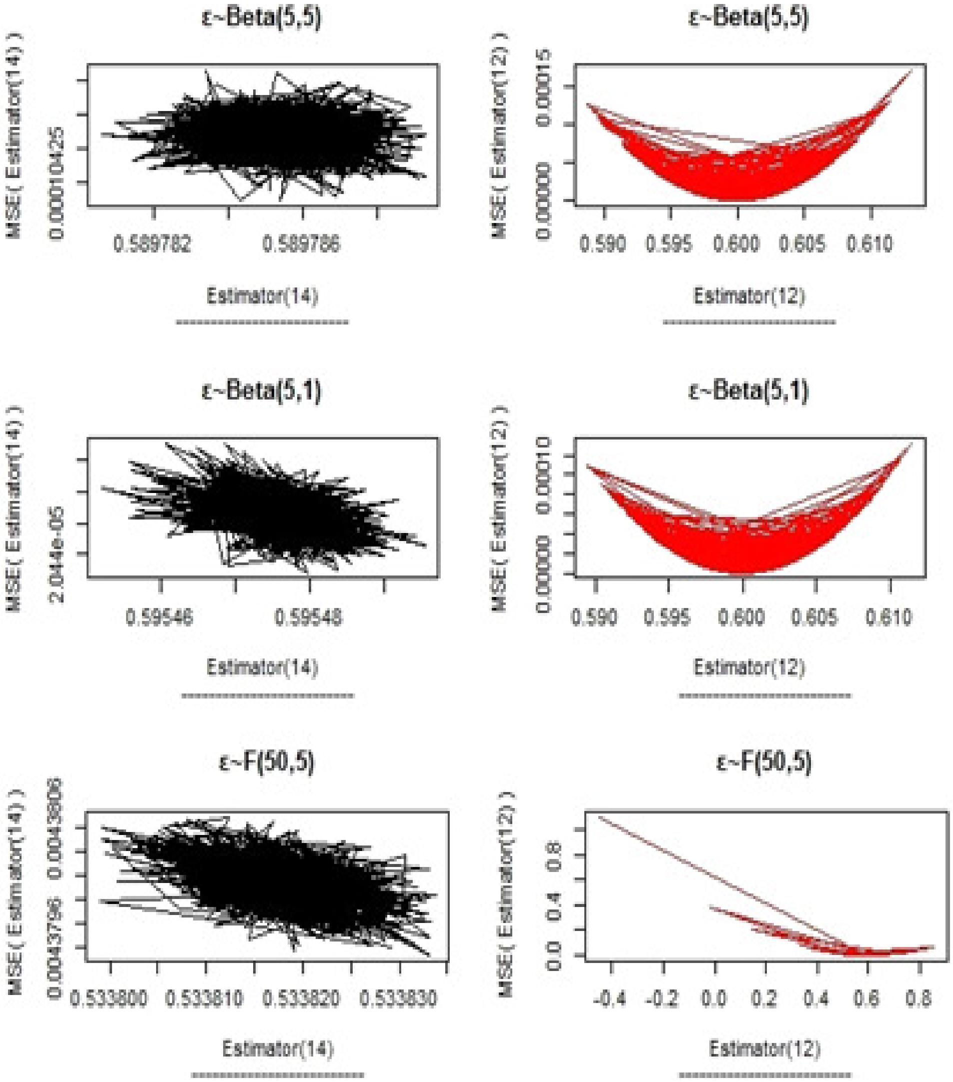
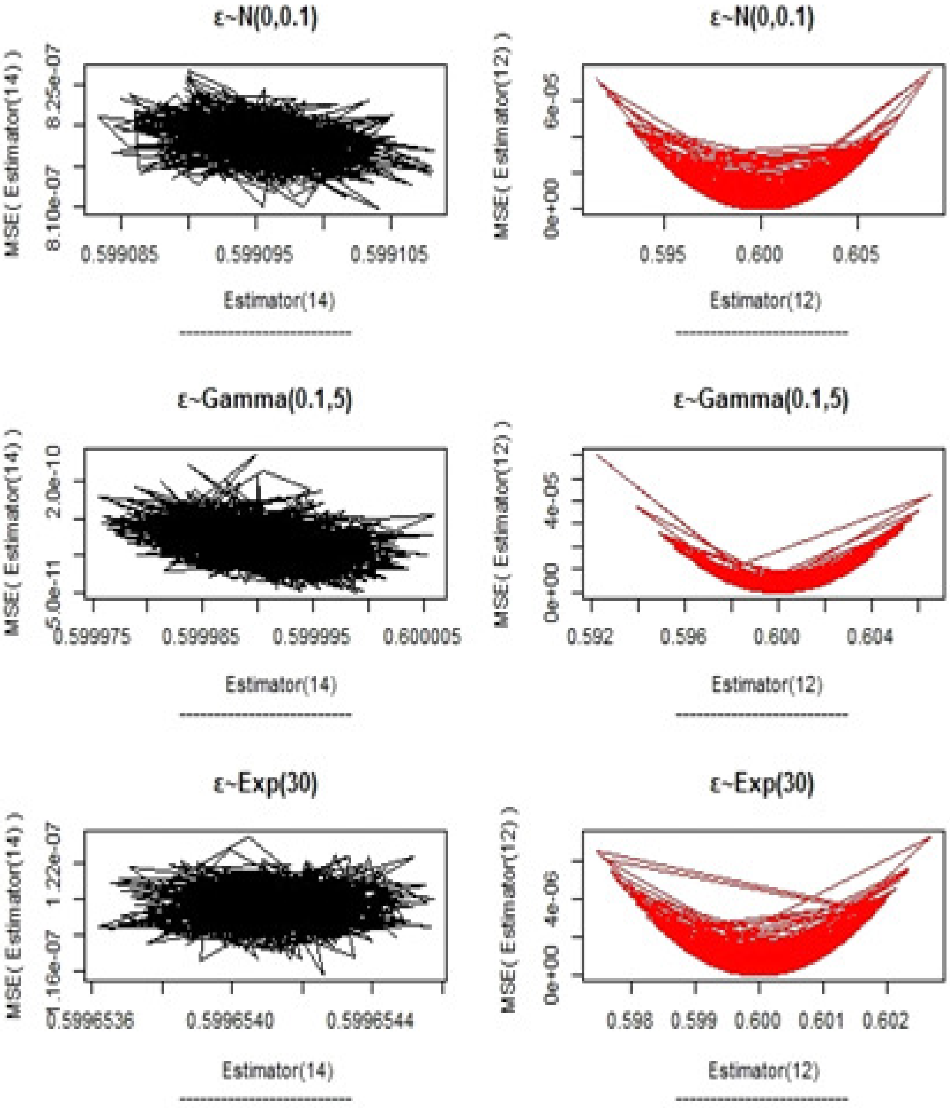


Figure 2: Assessment of the MSE of two estimators according to each distribution of **ε** fromTable 2



to develop complex methodologies for interval-valued data. Second, it is flexible in the sense that one can modify the structure of MCR depending on specific problems. For example, one may assume a non-uniform distribution within an interval such as truncated normal distribution.

In this paper, we introduce a new estimator, using the structure of MCR based on Y^{I} and X^{M}, alongside the old known estimator, based on Y and X^{M}. The structure of MCR randomly generates a large number of single-valued data sets; each of them consists of points randomly chosen within the observed intervals of X, X^{M} and Y^{I}. In the structure, internal variations in the interval-valued observations of the predictor variable X are fully utilized.

We summarize some simulation results in the tables. In each table, we report the mean of the obtained midpoint coefficients from n repetitions based on each ε distribution in Models 2.2. Assessments of Models 2.2 on the interval-valued data show that none of the two consistent estimators of α is better than the other one based on all distributions of ε.

By using the sampling distributions obtained from the random sampling process, a new asymptotic distribution of each estimator is introduced based on the LT. Since the ranges of the intervals are limited, so each asymptotic distribution depends on n in large-sample conditions ($n \rightarrow \infty$). Hence the condition of independence is not established. Also, the variance of each estimator depends on n. So the LT must be used instead of the CLT.

This research paper has certain theoretical, empirical, and methodological contributions: First, the study proposes a new method to investigate the effect of ε distribution in choosing a better estimator of α in each of Models2.2. Second, the proposed method fully makes use of the variability of the interval-valued data of X variable because the estimated coefficients of α from n repetitions, based on each estimator, are obtained via the new structure of MCR. Third, in the study, to increase the precision of the obtained estimated coefficients of α, we try to limit the sampling to intervals, single-valued data points, of the variable X based on the new structure of MCR. By entering the values listed in each of the models, we generate the values of the response variable. This will cause the imprecision imposed according to a random number from each of the X and Y intervals simultaneously to be limited to random numbers from each of the intervals of X in the parametric model. Last, this research presents the imprecision imposed to Y as Y^{I} according to the MCR structure in Section 3. The study shows that this imprecision is imposed on providing a consistent superior estimator for some specific ε distributions.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not for profit sectors.

References

Bertrand, P. and Goupil, F. (2000). *Descriptive statistics for symbolic data Analysis of symbolic data*. (pp. 106-124): Springer.

Billard, L. (2001). Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data, edited by H.-H. Bock and E. Diday. *Journal of Classification*. **18**(2), 291-294.

Billard, L. (2006). Symbolic data analysis: what is it? *Compstat 2006-Proceedings in Computational Statistics*. (pp. 261-269): Springer.

Billard, L. (2007).*Dependencies and variation components of symbolic interval-valued data Selected Contributions in Data Analysis and Classification*. (pp. 3-12): Springer.

Billard, L. (2008). Sample covariance functions for complex quantitative data. *Proceedings World Congress International Association Statistical Computing*.

Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining*. **4**(2), 149-156.

Billard, L., Diday, E. (2002). *Symbolic regression analysis Classification, Clustering, and Data Analysis*. (pp. 281-288): Springer.

Billard, L. and Diday, E. (2000). *Regression analysis for interval-valued data Data Analysis, Classification, and Related Methods*. (pp. 369-374): Springer.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: symbolic data analysis. *Journal of the American Statistical Association*. **98**(462), 470-487.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. John Wiley and Sons, Chichester.

Blanco-Fernández, A., Corral, N. and González-Rodríguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic. *Computational Statistics & Data Analysis*. **55**(9), 2568-2578.

Calle, M.L. and Gómez, G. (2001). Nonparametric Bayesian estimation from interval-censored data using Monte Carlo methods. *Journal of Statistical Planning and Inference*. **98**(1), 73-87.

De Carvalho, F., Lima Neto, E. and Tenorio, C. (2004). A new method to fit a linear regression model for interval-valued data. *Advances in Artificial Intelligence*, 295-306.

Diday, E. (1987). Introduction à l'analyse des données symboliques, Actes des journées "Symbolique-Numérique" pour l'apprentissage de connaissances partir d'observations, Université Paris 9 Dauphine. CEREMADE. Edités par E. Diday et Y. Kodratoff.

Diday, E. (1989). Introduction à l'approche symbolique en analyse des données. Revue française d'automatique, d'informatique et de recherche opérationnelle. *Recherche opérationnelle*. **23**(2), 193-236.

Diday, E. (1995). Probabilist, possibilist and belief objects for knowledge analysis. *Annals of Operations Research*. **55**(2), 225-276.

Diday, E. and Emilion, R. (1996a). Latices and capacities in analysis of probabilist object. *Paper presented at the Studies in Classication*.

Diday, E. and Emilion, R. (1996b). *Capacities and credibilities in analysis of probabilistic objects Ordinal and Symbolic Data Analysis*. (pp. 13-30): Springer.

Diday, E. and Emilion, R. (1998). *Capacities, Credibilities in Analysis of Probabilistic Objects by Histograms and Lattices Data Science, Classification, and Related Methods*. (pp. 353-357): Springer.

Emilion, R. (1997). Differentiation des capacites et des integrales de Choquet. *Comptes rendus de l'Academie des sciences. Serie 1, Mathematique*. **324**(4), 389-392.

Gil, M.Á., López-García, M.T.,Lubiano, M.A. and Montenegro, M. (2001). Regression and correlation analyses of a linear relation between random intervals. *test*. **10**(1), 183-201.

Gil, M.Á., González-Rodríguez, G., Colubi, A. and Montenegro, M. (2007). Testing linear independence in linear models with interval-valued data. *Computational Statistics & Data Analysis*. **51**(6), 3002-3015.

Huber, C., Solev, V. and Vonta, F. (2009). Interval censored and truncated data: Rate of convergence of NPMLE of the density. *Journal of Statistical Planning and Inference*. **13** , 1734-1749 ,(5)9.

Le-Rademacher, J. and Billard, L. (2012). Symbolic covariance principal component analysis and visualization for interval-valued data. *Journal of Computational and Graphical Statistics*. **21**(2), 413-432.

Neto, E.A.L., Cordeiro, G.M., de Carvalho, F.A.T., dos Anjos, U.U. and da Costa, A.G. (2009) Bivariate generalized linear model for interval-valued variables. *2009 International Joint Conference on Neural Networks*.

Neto, E.A.L., de Carvalho, F.A.T. and Tenorio, C.P. (2004). Univariate and multivariate linear regression methods to predict interval-valued features. *Australasian Joint Conference on Artificial Intelligence*.

Neto, E.A.L, de Carvalho, F.A.T. and Freire, E.S. (2005). Applying constrained linear regression models to predict interval-valued data. *Annual Conference on Artificial Intelligence*.

Neto, E.A.L. and de Carvalho, F.A.T. (2008). Centre and Range method for fitting a linear regression model to symbolic interval data. *Computational Statistics & Data Analysis*. **52**(3), 1500-1515.

Neto, E.A.L. and de Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*. **54**(2), 333-347.

Neto, E.A.L., de Carvalho, F.A.T. and Neto, J.F.C. (2007). Constrained linear regression models for interval-valued data with dependence. *IEEE International Conference on Systems, Man and Cybernetics*.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far beyond the classical data models: symbolic data analysis. *Statistical Analysis and Data Mining*. **4**(2), 157-170.

Rivero, C. and Valdes, T. (2008). An algorithm for robust linear estimation with grouped data. *Computational Statistics & Data Analysis*. **53**(2), 255-271.

Trutschnig, W., González-Rodríguez, G., Colubi, A., Gil, M.Á. (2009). A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread.*Information Sciences*. **179**(23), 3964-3972.

Wang, H., Guan, R. and Wu, J. (2012). Linear regression of interval-valued data based on complete information in hypercubes. *Journal of Systems Science and Systems Engineering*. **21**(4), 422-442.

Xu, W. (2010). *Symbolic data analysis: interval-valued data regression*. PhD thesis, University of Georgia.