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

Amir Masoud Malekfar¹, Farzad Eskandari²*

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


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.

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 \( \alpha \) for fitting the models on interval-valued data based on the MCR structure and some particular distributions of \( \varepsilon \). 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 \( (n \to \infty) \), cause each asymptotic distribution depends on n. According to some particular distributions of \( \varepsilon \), Section 6 assesses a better estima-
tor 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, $(\inf A, \sup A) \in \mathbb{R}^2$ with $\inf A \leq \sup A$. Equivalently, the point $(\text{mid } A, \text{spr } A) \in \mathbb{R} \times \mathbb{R}^+$, 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 $A = [\text{mid } A \pm \text{spr } A]$ or $A = [\inf A \pm \sup A]$ is used in each case.

Remark 2.1. The $(\inf, \sup)$- and the $(\text{mid, spr})$-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 $(\text{mid, spr})$-parametrization, since it only involves a non-negativity constraint on the second component, which is more operative. Additionally, the meaning of the $(\text{mid, spr})$-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 $(\text{mid, spr})$-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 $\text{spr } X$ 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

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

where $\alpha$ and $\beta$ are the coefficients, $\gamma$ is an intercept term affecting the mid component
of \( Y \), and \( \epsilon \) is an interval-valued random error variable such that \( E(\epsilon | X) = [-\delta, \delta] \in k_c(\mathbb{R}) \), 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

\[
Y = \alpha X^M + \beta X^S + \gamma + \epsilon. \tag{2.2}
\]

Model 2.2 are generalized linear models. Two specific assumptions

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

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 \( i = 1, \ldots, k \),

Symbols and formulas are introduced based on each of Models 2.2 and Section 2 in this section. Given one random sample \( \{X^{ib}\}_i^{k} \) from \( X \). So we define

\[
X^{*b} = \{X^{ib}_1, \ldots, X^{ib}_k\}, \tag{3.3}
\]

where \( X^{ib}_i \) is the \( b \)th observed sample of the \( i \)th interval of \( X \). Hence \( X^{ib}_i \epsilon (\inf(X_i), \sup(X_i)) \).

\[
spr(X^{*b}) = X^{rb} = X^{Sib} = \{X^{Sb}_1, \ldots, X^{Sb}_k\}, \tag{3.4}
\]

where \( X^{Sb}_i \) is the \( b \)th calculated value of \( X^{S}_i \), the spread of the \( i \)th interval of \( X \).

\[
\text{mid}(X^{*b}) = X^{Mib} = \{X^{Mb}_1, \ldots, X^{Mb}_k\}, \tag{3.5}
\]

where \( X^{Mb}_i \) is the \( b \)th sample of the \( i \)th value of \( X^M \).

\[
Y^{*b} = \{Y^{ib}_1, \ldots, Y^{ib}_k\}, \tag{3.6}
\]
where $Y^b_i$ is obtained from $Y^b_i = \alpha X^M_i + \beta X^S_i + \gamma + \epsilon^b_i$.

$$Y^{M;b} = \{Y^{M}_1, \ldots, Y^{M}_k\},$$

where $Y^{M}_i = \alpha X^M_i + \gamma + \epsilon^M_i$.

$$Y^{I;b} = \{Y^{I}_1, \ldots, Y^{I}_k\},$$

where each sample $Y^{I}_i \in [Y^b_i - 0.0005, Y^b_i + 0.0005]$.

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

$$X^* = \{X^{*;b}\}^n_{b=1} = \{X^{*;1}, \ldots, X^{*;n}\} = \{X^1, \ldots, X^k, \ldots, X^1, \ldots, X^n\}. \quad (3.7)$$

$$Y^* = \{Y^{*;b}\}^n_{b=1} = \{Y^{*;1}, \ldots, Y^{*;n}\} = \{Y^1, \ldots, Y^k, \ldots, Y^1, \ldots, Y^n\}. \quad (3.8)$$

$$\text{mid} (X^*) = X^{M*} = \{X^{M;b}\}^n_{b=1} = \{X^1, \ldots, X^k, \ldots, X^1, \ldots, X^n\}. \quad (3.9)$$

$$\text{spr}(X^*) = X^{S*} = \{X^{S;b}\}^n_{b=1} = \{X^1, \ldots, X^k, \ldots, X^1, \ldots, X^n\}. \quad (3.10)$$

$$Y^{M*} = \{Y^{M;b}\}^n_{b=1} = \{Y^1, \ldots, Y^k, \ldots, Y^1, \ldots, Y^n\}. \quad (3.11)$$

$$Y^{I*} = \{Y^{I;b}\}^n_{b=1} = \{Y^1, \ldots, Y^k, \ldots, Y^1, \ldots, Y^n\}. \quad (3.11)$$

### 3.1 The Process of Producing Random Sample

Use the uniform distribution, randomly produce the $b^{th}$ single-valued data point $X^b_i$ from $X_i \in [a_i, c_i]$. Here $b$ is the index of the $b^{th}$ sample and $i$ is the index of the $i^{th}$ interval. $X_i$ is the $i^{th}$ interval of the only predictor $X$. So $X^b_i \in [a_i, c_i]$ for $i = 1, \ldots, k$, $k$ is the number of the intervals, and $b = 1, \ldots, n$ 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 $b^{th}$ single-valued data point $X^{M;b}_i$ from $X^{M}\in [d_i, d_i]$ for $i = 1, \ldots, k$ and $b= 1, \ldots, n$. $X^M_i$ is the $i^{th}$ interval of $X^M$. So $X^{M;b}_i \in [d_i, d_i]$ for $i = 1, \ldots, k$ and $b = 1, \ldots, n$. We generate each random sample $X^{M;b}$ according to Equation 3.9. 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 $b^{th}$ single-valued data point $Y^b_i$ is generated, in each of Models 2.2, according to the $b^{th}$ single-valued data point $X^b_i$. 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 $b^{th}$ single-valued data point $Y^{I;b}_i \in [Y^b_i - 0.0005, Y^b_i + 0.0005]$ for $i = 1, \ldots, k$ and $b = 1, \ldots, n$. Hence we generate each random sample $Y^{I;b} = \{Y^{I}_1, \ldots, Y^{I}_k\}$. We produce the random sample $Y^{I*} = \{Y^1, \ldots, Y^k, \ldots, Y^1, \ldots, Y^n\}$, by using
all of the samples $Y^{i,b}$ for $b = 1, \ldots, n$. Difference of $Y$ and $Y^I$ value is very small. This paper will show that the difference is useful.

Produce the $b$th random sample:

$$
Y^{\ast b} = \{Y_{b1}^{1}, \ldots, Y_{bk}^{1}\}, \quad Y^{I \ast b} = \{Y_{ib}^{1}, \ldots, Y_{ik}^{1}\}, \quad X^{\ast b} = \{X_{b1}^{1}, \ldots, X_{bk}^{1}\}, \quad IX^{\ast b} = \{1, X_{b1}^{1}, \ldots, X_{bk}^{1}\}.
$$

Difference of $Y$ and $Y^I$ value is very small. This paper will show that the difference is useful.

So we produce:

$$
Y^\ast = \{Y_{1}^{1}, \ldots, Y_{k}^{1}, \ldots, Y_{1}^{n}, \ldots, Y_{k}^{n}\},
$$

$$
Y^{I \ast} = \{Y_{1}^{I1}, \ldots, Y_{k}^{I1}, \ldots, Y_{1}^{In}, \ldots, Y_{k}^{In}\},
$$

$$
X^{\ast} = \{X_{1}^{1}, \ldots, X_{k}^{1}, \ldots, X_{1}^{n}, \ldots, X_{k}^{n}\},
$$

$$
IX^{\ast} = \{1, X_{1}^{1}, \ldots, X_{k}^{1}, \ldots, X_{1}^{n}, \ldots, X_{k}^{n}\}.
$$

4. A Proposal Method

Based on the new structure of MCR in Section 3 and the estimator of $\omega_1$ in the simple linear regression model $Y = \omega_0 + \omega_1 X + \varepsilon$, Subsection 4.2 proposes a new estimator of $\alpha$, 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 $\alpha$. This is the best estimator of $\alpha$ 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 $\varepsilon$ in Model 2.2?

The effectiveness of $\varepsilon$ 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 $Y^\ast b$ and $X^{M \ast b}$ (see Sections 2-3) for $b = 1, \ldots, n$. Also, we re-introduce Estimator 4.13 based on $Y^\ast$ and $X^{M \ast}$ (see Sections 2-3). In Section 6, Table 1 shows some particular distributions of $\varepsilon$ that Estimator 4.12 is better than Estimator 4.14. We re-introduce:

$$
\hat{\alpha} = \frac{1}{n} \sum_{b=1}^{n} \hat{\alpha}^b = \frac{1}{n} \sum_{b=1}^{n} \left( \frac{\text{Cov} \left( Y^{\ast b}, X^{M \ast b} \right)}{\text{Var} \left( X^{M \ast b} \right)} \right) 
$$

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

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

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

### 4.2 A New Estimator

This Subsection introduces Estimator 4.14 as a new estimator of $\alpha$ 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 $Y^{I+s}$ and $X^{M+s}$ (see Sections 2-3). In Section 6, Table 2 shows some particular distributions of $\epsilon$ that Estimator 4.14 is better than Estimator 4.12. We propose:

$$\overline{\alpha} = \frac{1}{n} \sum_{b=1}^{n} \overline{\alpha^b} = \frac{1}{n} \sum_{b=1}^{n} \frac{\text{Cov}(Y^{I+b}, X^{M+b})}{\text{Var}(X^{M+b})}.$$  

(4.14)
Assessment, Estimation and Modeling of the . . . 85

\[
\begin{bmatrix}
\gamma \\
\alpha
\end{bmatrix}
= \frac{1}{n} \sum_{b=1}^{n} \begin{bmatrix}
\gamma^b \\
\alpha^b
\end{bmatrix} = \frac{1}{n} \sum_{b=1}^{n} \begin{bmatrix}
1 \ X^b_1 \\
1 \ X^b_k
\end{bmatrix} \begin{bmatrix}
1 \ X^b_1 \\
1 \ X^b_k
\end{bmatrix}^T \begin{bmatrix}
1 \ X^b_1 \\
1 \ X^b_k
\end{bmatrix}^{-1} \begin{bmatrix}
1 \ X^b_1 \\
1 \ X^b_k
\end{bmatrix}^T \begin{bmatrix}
Y^b_1 \\
Y^b_k
\end{bmatrix}
\]

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

\[
\text{MSE}(\hat{\alpha}) = \left( \frac{1}{n} \sum_{b=1}^{n} \frac{\text{Cov}(Y^b_1, X^b_1)}{\text{Var}(X^b_1)} \right)^2 + \text{Var} \left( \frac{1}{n} \sum_{b=1}^{n} \frac{\text{Cov}(Y^b_1, X^b_1)}{\text{Var}(X^b_1)} \right).
\]

Two tables in Section 6, Tables 1-2, show that MSE values of the estimators depend on distributions of \( \varepsilon \) in Model 2.2. Also the mentioned section represents that under particular distributions of \( \varepsilon \), 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 \( \tilde{\alpha} \) 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 \( \varepsilon \).
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 \( \hat{\alpha} = \frac{\hat{\sigma}_Y}{\hat{\sigma}_M} \), so \( \hat{\sigma}_Y = \hat{\sigma}_M \hat{\alpha} \).

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

\[
\text{var}(\hat{\alpha}) = \frac{\text{var}(Y^*)}{n \cdot \text{var}(X^M)}
\] (5.16)

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

\[
\sqrt{n} (\hat{\alpha} - \alpha) \overset{L}{\rightarrow} N\left(0, \frac{\sigma^2_M}{\sigma^2_{M*}} \right)
\] (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 εM* variable is:

\[
\varepsilon^M = ((\varepsilon^{M_1}), \ldots, (\varepsilon^{M_n}))^T
\] (5.18)
Where
\[ \varepsilon^{M*b} = \text{mean}(Y^b) - \left( \frac{\text{cov}(Y^b, X^{M*b})}{\text{var}(X^{M*b})} \right) \left( (X^{M*b}) + \hat{\gamma} \right) \]
for \( b = 1, \ldots, n \).

\( \ldots \hat{\gamma} \) is offered in Equation 4.13.

**Definition 1.**
We know \( \hat{\alpha} = (\hat{\alpha}^1, \ldots, \hat{\alpha}^n) \), where \( \hat{\alpha}^b \) is the \( b \)th value of \( \hat{\alpha} \) according to the \( b \)th 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:

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

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

**Proof.** Let us write according to Theorem 5.1

\[ \frac{\hat{\alpha} - \alpha}{\text{Cov}(\varepsilon^{M*}, X^{M*})} \]

\[ E\left( \sqrt{n} (\hat{\alpha} - \alpha) \right) = \sqrt{n} E(\hat{\alpha} - \alpha) = 0 \text{ as } n \to \infty, \]

\[ \text{Var}\left( \sqrt{n} (\hat{\alpha} - \alpha) \right) = n \left( \frac{\sigma_{\varepsilon^{M*}}^2}{n \sigma_{X^{M*}}^2} \right) = \frac{\sigma_{\varepsilon^{M*}}^2}{\sigma_{X^{M*}}^2} \text{ as } n \to \infty. \]

\[ \Box \]

### 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, \( n \to \infty \).

**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 \( e^{M*} \) and \( e^{S*} \) (see Section 2) are the center and the spread of \( e^* \) 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^*}, e^{S^*}, e^{M^*} and e^* . We know \( \bar{\alpha} = \frac{\hat{\sigma}_{YI}^2}{\hat{\sigma}_{XM}^2} \), so
\[
\bar{\alpha} - \alpha = \frac{\hat{\gamma}}{\hat{\sigma}_{XM}^2} - \alpha.
\]
Hence:
\[
\text{Cov} (Y^{I^*}, X^{M^*}) = \text{Cov} (Y^{M^*} + Y^{S^*} + e^{M^*} + e^{S^*}, X^{M^*}) = \text{Cov} (Y^{M^*}, X^{M^*})
\]
\[
= \text{Cov} (\alpha X^{M*} + e^{M*}, X^{M*}) = \alpha \hat{\sigma}_{XM}^2 + \text{Cov} (e^{M*}, X^{M*}).
\]

We have \( \bar{\alpha} - \alpha = \frac{\text{Cov}(e^{M*}, X^{M*})}{\hat{\sigma}_{XM}^2} \), therefore \( E(\bar{\alpha} - \alpha) = E\left( \frac{\text{Cov}(e^{M*}, X^{M*})}{\hat{\sigma}_{XM}^2} \right) = 0 \), since \( X^{M^*} \) is independent of \( e^{M^*} \). Also \( \text{Var}(\bar{\alpha}) = \frac{1}{n} \text{Var}\left( \frac{\text{Cov}(e^{M*}, X^{M*})}{\hat{\sigma}_{XM}^2} \right) = \frac{1}{n} \left( \frac{\sigma_{M^*}^2}{\hat{\sigma}_{XM}^2} \right) \to 0 \) as \( n \to \infty \). So \( \bar{\alpha} \), namely Estimator 4.14, is a consistent estimator of \( \alpha \).

\[\square\]

Based on Sections 2-3, the variance of the consistent estimator is:
\[
\text{var}(\bar{\alpha}) = \frac{\text{var}(Y^{I^*})}{n \cdot \text{var}(X^{M^*})} \quad (5.20)
\]

**Theorem 5.4.** Under conditions of Model 2.2, if \( E\left( ||Y||^4 \right) < \infty \) (where \( ||.|| \) denotes the Euclidean norm) and \( 0 < \frac{\sigma_{I^*}^2}{\sigma_{XM}^2} < \infty \), the
\[
\sqrt{n} \left( \bar{\alpha} - \alpha \right) \xrightarrow{L} N \left( 0, \frac{\sigma_{M^*}^2}{\sigma_{XM}^2} \right) \quad (5.21)
\]

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 \( e^{M^*} \) variable is:
\[
e^{M^*} = (e^{M^*}_1, \ldots, e^{M^*_n})^T \quad (5.22)
\]

Where
\[
e^{M^*_b} = \text{mean}(Y^{I^*_b}) - \left( \frac{\text{Cov}(Y^{I^*_b}, X^{M^*_b})}{\text{var}(X^{M^*_b})} \cdot (X^{M^*_b}) + \bar{\gamma} \right) \quad \text{for } b = 1, \ldots, n.
\]

\( \bar{\gamma} \) 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:
\[
\left[ E\left( \sum_{b=1}^{n} \left| \alpha^{M^*_b} - \alpha \right|^3 \right) \right] \leq o\left( \sum_{b=1}^{n} \left( \frac{\sigma_{Y^{I^*_b}}}{\sigma_{XM^*_b}}^2 \right)^3 \right) \quad (5.23)
\]

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

\[ \tilde{\alpha} - \alpha = \frac{\text{Cov} (\varepsilon^M, X^M)}{\sigma^2_{X^M}}, \]

\[ E \left( \sqrt{n} \left( \tilde{\alpha} - \alpha \right) \right) = \sqrt{n} \cdot E \left( \tilde{\alpha} - \alpha \right) = 0 \text{ as } n \to \infty, \]

\[ \text{Var} \left( k_n \left( \tilde{\alpha} - \alpha \right) \right) = n \left( \frac{\sigma^2_{\varepsilon^M}}{\text{Var}(X^M)} \right) = \frac{\sigma^2_{\varepsilon^M}}{\sigma^2_{X^M}} \text{ as } n \to \infty. \]

\[ \square \]

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 (\(\alpha, \beta, \gamma\) = (0.6, 0.2, 0) in each of Models 2.2. We assess the accuracy of the estimators of \(\alpha\) in each model. \(X^8\) (see artificial 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 \(X^b_i\) from the \(i^{th}\) interval of the predictor \(X = (X^8 - 0.0005, X^8 + 0.0005)\) for \(i = 1, \ldots, 50\) and \(b = 1, \ldots, 1000\). 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 \(\varepsilon\) distribution in Models 2.2.

Which of the consist estimators of \(\alpha\) is better than the other one according to all distributions of \(\varepsilon\)?

The MSE column of Table 1 shows Estimator 4.12 is better than the other one according to some particular distributions of \(\varepsilon\), 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 \(\varepsilon\), 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 ε

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

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 $\varepsilon$ (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 $\varepsilon$ (see Table 2). Estimator 4.14 is better than the other one in Figure 2. So none of two consistent estimators of $\alpha$ is better than the other one based on all distributions of $\varepsilon$.

7. Conclusions

In the last decade, interval-valued variables have been modeled by LRM. 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 $\varepsilon$

<table>
<thead>
<tr>
<th>Estimator</th>
<th>$\hat{\alpha}$</th>
<th>Bias($\hat{\alpha}$)</th>
<th>Var ($\hat{\alpha}$)</th>
<th>MSE ($\hat{\alpha}$)</th>
<th>$\varepsilon$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(4.12)</td>
<td>0.600734</td>
<td>0.000734</td>
<td>0.000704</td>
<td>0.000704</td>
<td>N(0,1)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5909</td>
<td>-0.0090</td>
<td>5e-11</td>
<td>8e-05</td>
<td>N(0,1)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6001</td>
<td>0.0001</td>
<td>2e-05</td>
<td>2e-05</td>
<td>Beta(3,0.6)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5964</td>
<td>-0.0035</td>
<td>5e-11</td>
<td>1e-05</td>
<td>Beta(3,0.6)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.4563</td>
<td>-0.1436</td>
<td>105.6827</td>
<td>105.7034</td>
<td>C (0,1)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>5.8006</td>
<td>5.2006</td>
<td>5e-11</td>
<td>27.0464</td>
<td>C (0,1)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6001</td>
<td>0.0001</td>
<td>2e-05</td>
<td>2e-05</td>
<td>F(100,200,10)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.6007</td>
<td>0.00075</td>
<td>5e-11</td>
<td>5e-07</td>
<td>F(100,200,10)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.5988</td>
<td>-0.0011</td>
<td>0.0004</td>
<td>0.0004</td>
<td>F(10,200,10)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5795</td>
<td>-0.0204</td>
<td>5e-11</td>
<td>0.0004</td>
<td>F(10,200,10)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.5996</td>
<td>-0.0003</td>
<td>0.0003</td>
<td>0.0003</td>
<td>F(200,10,10)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5933</td>
<td>-0.0066</td>
<td>5e-11</td>
<td>4e-05</td>
<td>F(200,10,10)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.60003</td>
<td>3e-05</td>
<td>1e-06</td>
<td>1e-06</td>
<td>Ln(0.046,0.046)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5995</td>
<td>-0.0004</td>
<td>5e-11</td>
<td>1e-07</td>
<td>Ln(0.046,0.046)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.5999</td>
<td>-7e-05</td>
<td>0.0010</td>
<td>0.001</td>
<td>$\chi^2 (0.25,0.25)$</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.6149</td>
<td>0.0149</td>
<td>5e-11</td>
<td>0.0002</td>
<td>$\chi^2 (0.25,0.25)$</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6002</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0001</td>
<td>Beta(0.01,0.01)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.6113</td>
<td>0.0113</td>
<td>5e-11</td>
<td>0.0001</td>
<td>Beta(0.01,0.01)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6000</td>
<td>7e-05</td>
<td>7e-06</td>
<td>7e-06</td>
<td>N(0,0.1)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5990</td>
<td>-0.0009</td>
<td>5e-11</td>
<td>8e-07</td>
<td>N(0,0.1)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6000</td>
<td>1e-05</td>
<td>2e-06</td>
<td>2e-06</td>
<td>Gamma(0.1,5)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5999</td>
<td>-1e-05</td>
<td>5e-11</td>
<td>1e-10</td>
<td>Gamma(0.1,5)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6000</td>
<td>4e-05</td>
<td>6e-07</td>
<td>6e-07</td>
<td>Exp(30)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5996</td>
<td>-0.0003</td>
<td>5e-11</td>
<td>1e-07</td>
<td>Exp(30)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>810.1115</td>
<td>809.5115</td>
<td>582010775</td>
<td>582666084</td>
<td>F(1,1)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>-1.8195</td>
<td>-2.4195</td>
<td>5e-11</td>
<td>5.85444</td>
<td>F(1,1)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6028</td>
<td>0.0028</td>
<td>0.0150</td>
<td>0.0150</td>
<td>Bin(500,0.95)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5915</td>
<td>-0.0084</td>
<td>5e-11</td>
<td>7e-05</td>
<td>Bin(500,0.95)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.5998</td>
<td>-0.0001</td>
<td>0.0001</td>
<td>0.0001</td>
<td>Geo(0.85)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.5972</td>
<td>-0.0027</td>
<td>4e-11</td>
<td>7e-06</td>
<td>Geo(0.85)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.6002785</td>
<td>0.0002</td>
<td>0.0001</td>
<td>0.0001</td>
<td>Beta(0.2,0.2)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.6060</td>
<td>0.0060</td>
<td>5e-11</td>
<td>3e-05</td>
<td>Beta(0.2,0.2)</td>
</tr>
<tr>
<td>(4.12)</td>
<td>0.5994</td>
<td>-0.0005</td>
<td>7e-05</td>
<td>7e-05</td>
<td>Beta(0.7,0.7)</td>
</tr>
<tr>
<td>(4.14)</td>
<td>0.6000</td>
<td>2e-05</td>
<td>5e-11</td>
<td>6e-10</td>
<td>Beta(0.7,0.7)</td>
</tr>
</tbody>
</table>
Figure 1: Assessment of the MSE of two estimators according to each distribution of $\varepsilon$ from Table 1
Figure 2: Assessment of the MSE of two estimators according to each distribution of $\varepsilon$ from Table 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 $\varepsilon$ distribution in Models 2.2. Assessments of Models 2.2 on the interval-valued data show that none of the two consistent estimators of $\alpha$ is better than the other one based on all distributions of $\varepsilon$.

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 ($n \to \infty$). 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 $\varepsilon$ distribution in choosing a better estimator of $\alpha$ in each of Models 2.2. Second, the proposed method fully makes use of the variability of the interval-valued data of $X$ variable because the estimated coefficients of $\alpha$ 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 $\alpha$, 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 $\varepsilon$ distributions.
Acknowledgements

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

References


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


