

Constrained Interval-valued Linear Regression Model

Feng Li*, Shoumei Li*, Nana Tang*, Thierry Denceux*[†]

*College of Applied Science

Beijing University of Technology, Beijing 100124, China

Email: lishoumei2011@163.com

[†]Sorbonne Universités

Université de Technologie de Compiègne, CNRS

UMR 7253 Heudiasyc, France

Abstract—In current interval-valued linear regression models, meaningless predictions may be generated because the lower bounds of the predicted intervals may be greater than their upper bounds. To avoid this problem, we propose a constrained interval-valued linear regression model based on random set theory. However, due to the introduction of constraints in this model, the expectation of the errors is no longer zero, and estimation provided by traditional least square may produce systematic bias. To address this issue, we introduce a two-step procedure: in the first step, a dummy variable is defined and plugged into the regression model to ensure that the expectation of errors is zero; least square estimation is then used in the second step. To show the validity of proposed method, experiments on simulated and real data are presented.

I. INTRODUCTION

The classical linear regression model has been well developed in the pre-computer age of statistics, but even in today's computer era there are still good reasons to study and use it. In past research, real-valued data have been exclusively considered. However, in practice, set-valued observations, and particularly interval-valued data, may be obtained in many fields, such as medical diagnosis and weather forecasting. The analysis of such data plays an important role in various fields, and it is expected to have further extensive applications to other areas.

Interval-valued data are quite common in our daily life. For example, the highest and lowest values of daily temperature are interval data. Until recently, a single point representing an interval-valued observation has often been used to estimate the parameters of the traditional linear regression. For instance, when modeling stock prices, only the closing prices are typically considered in the model, in which case the information of stock price fluctuations during each day is lost. Alternatively, we can take into account the highest and lowest prices of each day, which reflect the variability of stock prices over this period.

The notion of set-valued random variable was first introduced by Robbins [1] in 1944. Artstein then proved a strong law of large numbers for random sets [2]; some details can be found in [3]. In 1965, Aumann introduced the integral of set-valued random variable [4], and Hiai later proposed a definition of conditional expectations [5]. On the basis of

the above theory, the convergence of set-valued martingales under different conditions was studied [6]–[11]. In recent years, a systematic theory for set-valued random variables was established [12], [13]. In 2007, Zhang *et al.* comprehensively summarized the basic theory of set-valued random variables and set-valued stochastic processes in [14]. All these contributions have laid the foundations of set-valued stochastic theory.

Since the hyperspace of sets is not linear with respect to addition and scalar multiplication, it is difficult to define the variance and covariance of set-valued random variables. In 2005, Yang and Li [15] proposed the D_p metric in the space of set-valued random variables, and defined the variance and covariance of set-valued random variable using this metric. Later, Blanco [16] defined the d_k -variance of interval-valued random variables with underlying space \mathbb{R}^1 , which is a special case of [15]. Based on the D_p metric, Wang et al. [17] defined a new type of interval-valued linear model, where the inputs are real-valued random variables and the output is an interval-valued random variable. They proved that the best linear unbiased estimation for this model does not exist, but the least square estimator is the best binary linear unbiased estimator. In this paper, we introduce another type of interval-valued linear model, in which the output and inputs are both interval-valued random variables.

Other approaches to the analysis of interval-valued data have been proposed. Billard [18] introduced a center model, in which only the midpoints of the intervals are used for fitting a linear regression model. The fitted model is then used to predict the lower and upper bounds of the intervals. Beside the center model, Lima Neto [19] proposed another linear model for the ranges of the intervals. They proved that this method reconstructs the boundaries of the intervals in a more efficient way than does the Billard method. However, the same problem happens with these methods, that is, the lower bound of the predicted interval may be greater than its upper bound. To avoid this problem, Lima Neto [20] introduced a constrained version of the center and range method, where the coefficients in the range model are forced to be nonnegative. However, they also recognized that nonnegativity constraints on the coefficients are inappropriate when the variables do have a negative relationship.

Based on the natural requirement that the lower bound of an interval should be less than or equal to its upper bound, some improvements are proposed in this paper. First, in our model the lower and upper bounds of the dependent interval-valued variable are both related to the upper and lower bounds of the independent interval-valued variables. Secondly, because the expectation of the errors is not zero any more, the least-squares estimates may be biased. To deal with this problem, we propose a two-step procedure. In the first step, we introduce an auxiliary variable, which can be treated as a new regressor in the regression model. The least square estimator is then computed in the second step. Results from numerical studies show that our method outperforms other interval-valued linear models.

This paper is organized as follows. In Section 2, some background knowledge about set-valued random variables and existing regression models is presented. Our new method is explained in Section 3. Section 4 describes the results of numerical and empirical studies. Concluding remarks are given in Section 5.

II. SET-VALUED RANDOM VARIABLES AND REGRESSION MODEL FOR INTERVAL-VALUED DATA

As interval-valued random variables are a special case of the set-valued random variables, we firstly recall the set-valued theory in this section. Some existing regression models for interval-valued data are then presented.

A. Set-valued random variables

1) *Operations in the hyperspace:* Suppose that (Ω, \mathcal{A}) is a measurable space, and \mathcal{X} is a metric space. We denote by $K(\mathcal{X})$ the family of all nonempty closed subsets of \mathcal{X} , $K_k(\mathcal{X})$ the family of all nonempty compact subsets of \mathcal{X} , $K_c(\mathcal{X})$ the family of all nonempty closed convex subsets of \mathcal{X} , $K_{kc}(\mathcal{X})$ the family of all nonempty compact and closed convex subsets of \mathcal{X} , and $\mathcal{O}(\mathcal{X})$ the family of all nonempty open subsets of \mathcal{X} .

For any $A, B \in K(\mathcal{X})$, $\lambda \in \mathbb{R}$, the addition and scalar multiplication can be defined, respectively, as

$$A + B = \{a + b : a \in A, b \in B\},$$

$$\lambda A = \{\lambda a : a \in A\}.$$

Remark that $(K(\mathcal{X}), +, \cdot)$ is not a linear space in general, since $A + (-A) \neq \{0\}$. When \mathcal{X} is an infinite-dimensional space, for $A, B \in K(\mathcal{X})$, $A + B \in K(\mathcal{X})$ may not be true, so it is represented by $A \oplus B = cl\{a + b : a \in A, b \in B\}$. But if $A, B \in K_{kc}(\mathcal{X})$ (or $K_k(\mathcal{X})$), $A + B \in K_{kc}(\mathcal{X})$ (or $K_k(\mathcal{X})$).

When \mathcal{X} is a separable Banach space, the dual space of \mathcal{X} is \mathcal{X}^* , i.e., the set of all bounded linear functionals on \mathcal{X} . Specially when $\mathcal{X} = \mathbb{R}^d$, that is \mathcal{X} is d dimensional Euclidean space, $\mathcal{X}^* = \mathbb{R}^d$.

For each $A \in K_c(\mathcal{X})$, the support function is defined by

$$s(x^*, A) = \sup\{x^*(a) : a \in A\}, \quad x^* \in \mathcal{X}^*.$$

It has the following properties:

$$s(x^*, A \oplus B) = s(x^*, A + B) = s(x^*, A) + s(x^*, B),$$

$$s(x^*, \lambda A) = \lambda s(x^*, A), \lambda \geq 0,$$

$$s(x^*, -A) = s(-x^*, A).$$

For $1 \leq p < \infty$, the metric d_p of sets $A, B \in K_{kc}(\mathcal{X})$ is defined by

$$d_p(A, B) = \left(\int_{S^*} |s(x^*, A) - s(x^*, B)|^p d\mu \right)^{\frac{1}{p}},$$

where S^* is the unit sphere of \mathcal{X}^* , i.e., $S^* = \{x^* \in \mathcal{X}^* : \|x^*\|_{\mathcal{X}^*} = 1\}$, and μ is a measure on $(\mathcal{X}^*, \mathcal{B}(\mathcal{X}^*))$. We denote the distance between A and $\{0\}$ by $\|A\|_{d_p} = d_p(\{0\}, A)$, where $\|\cdot\|_{d_p}$ is just a notation, not a norm, since $K_{kc}(\mathcal{X})$ is not a linear space in general.

Example 1. When $\mathcal{X} = \mathbb{R}^1$, $K_{kc}(\mathbb{R}^1)$ represents all closed intervals of \mathbb{R}^1 , namely

$$K_{kc}(\mathbb{R}^1) = \{[a, b] : 0 < a \leq b < \infty, a, b \in \mathbb{R}^1\}.$$

A scalar $a \in \mathbb{R}^1$ is actually a degenerate interval, that is $a = [a, a]$. For any two closed intervals $[a_i, b_i] \in K_{kc}(\mathbb{R}^1)$, $i = 1, 2$, $\lambda \in \mathbb{R}^1$, the addition and scalar multiplication of intervals are

$$[a_1, b_1] + [a_2, b_2] = [a_1 + a_2, b_1 + b_2],$$

$$\lambda[a_1, b_1] = \begin{cases} [\lambda a_1, \lambda b_1] & \lambda \geq 0, \\ [\lambda b_1, \lambda a_1] & \lambda < 0. \end{cases}$$

When $\mathcal{X} = \mathbb{R}^1$, we have $\mathbb{R}^{1*} = \mathbb{R}^1$. For each $A = [a, b] \in K_{kc}(\mathbb{R}^1)$, its support function is

$$s(x, A) = \begin{cases} bx & x \geq 0, \\ ax & x < 0, \end{cases}$$

where $x \in \mathbb{R}^1$, and the unit sphere of \mathbb{R}^1 is $S^* = \{-1, 1\}$. Thus for any $p \in [1, \infty)$, the d_p distance between any two intervals $A = [a_1, b_1]$, $B = [a_2, b_2]$ is

$$d_p(A, B) = (|b_1 - b_2|^p + |a_1 - a_2|^p)^{\frac{1}{p}}. \quad (1)$$

2) D_p metric of set-valued random variables:

Definition 1. A set-valued mapping $F : \Omega \rightarrow K(\mathcal{X})$ is called strongly measurable if $\forall C \in K(\mathcal{X})$, $F^{-1}(C) \in \mathcal{A}$, where $F^{-1}(C) = \{\omega \in \Omega : F(\omega) \cap C \neq \emptyset\}$. It is called weakly measurable if $\forall O \in \mathcal{O}(\mathcal{X})$, $F^{-1}(O) \in \mathcal{A}$. A weakly measurable set-valued mapping is also called a set-valued random variable for short (or a random set).

Theorem 1. A strongly measurable set-valued mapping is a set-valued random variable.

Two set-valued random variables are said to be identical if $F_1(\omega) = F_2(\omega)$ for almost every $\omega \in \Omega$. Let $\mathcal{U}[\Omega, K_k(\mathcal{X})]$ (or $\mathcal{U}[\Omega, K_{kc}(\mathcal{X})]$) be the family of set-valued random variables taking values in $K_k(\mathcal{X})$ (or $K_{kc}(\mathcal{X})$). The D_p metric of two

set-valued random variables $F_1, F_2 \in \mathcal{U}[\Omega, K_k(\mathcal{X})]$ is defined by

$$D_p(F_1, F_2) = [E(d_p^p(F_1(\omega), F_2(\omega)))]^{\frac{1}{p}}.$$

Example 2. Let $\mathcal{X} = \mathbb{R}^1$, and $F_i(\omega) = [a_i(\omega), b_i(\omega)]$, where $a_i(\omega)$ and $b_i(\omega)$ are two random variables such that $a_i(\omega) \leq b_i(\omega)$, $\forall \omega \in \Omega$, $i = 1, 2$. Denote by $c_i = \frac{a_i(\omega) + b_i(\omega)}{2}$ and $r_i = \frac{b_i(\omega) - a_i(\omega)}{2}$, respectively, the midpoint and range of F_i , $i = 1, 2$, then

$$D_p(F_1, F_2) = [E|a_2(\omega) - a_1(\omega)|^p + E|b_2(\omega) - a_1(\omega)|^p]^{\frac{1}{p}} \\ = [E|(c_2 - c_1) - (r_2 - r_1)|^p + E|(c_2 - c_1) + (r_2 - r_1)|^p]^{\frac{1}{p}}.$$

We define $\mathcal{L}^p[\Omega, K_{kc}(\mathcal{X})]$ as the set $\{F \in \mathcal{U}[\Omega, K_{kc}(\mathcal{X})] : E[|F|_{d_p}^p] < +\infty\}$. The space $(\mathcal{L}^p[\Omega, K_{kc}(\mathcal{R}^d)], D_p)$ is a complete metric space.

3) *Variance and covariance of set-valued random variables:* The expectation of set-valued random variable F was defined by Aumann [4] as

$$E[F] = \left\{ \int_{\Omega} f dP : f \in S_F \right\},$$

where $S_F = \{f : f(\omega) \in F(\omega) \text{ a.s.}(P), \text{ and } f \text{ is integrable}\}$ is called the selection set of the set-valued random variable F . Based on this definition, Yang and Li [15] proposed to define notions of variance and covariance via the D_p metric on $\mathcal{U}[\Omega, K_{kc}(\mathcal{X})]$, which is also used in Wang et al. [21]. Here, we recall these definitions.

For a set-valued random variable $F \in \mathcal{U}[\Omega, K_{kc}(\mathcal{X})]$, the variance of F is defined as

$$\text{Var}(F) = [D_2(F, E(F))]^2 \\ = E \left\{ \int_{S^*} [s(x^*, F(\omega)) - s(x^*, E(F(\omega)))]^2 du \right\}.$$

The covariance of two set-valued random variable $F_1, F_2 \in \mathcal{U}[\Omega, K_{kc}(\mathcal{X})]$ is defined as

$$\text{Cov}(F_1, F_2) = E \left\{ \int_{S^*} [s(x^*, F_1(\omega)) - s(x^*, E(F_1))] \right. \\ \left. \times [s(x^*, F_2(\omega)) - s(x^*, E(F_2))] du \right\}.$$

The correlation coefficient of F_1 and F_2 is defined as

$$\rho(F_1, F_2) = \frac{\text{Cov}(F_1, F_2)}{\sqrt{\text{Var}(F_1)\text{Var}(F_2)}},$$

where $\text{Var}(F_1)\text{Var}(F_2) \neq 0$.

The variance, covariance and correlation of set-valued random variables have the following properties:

- 1) $\text{Var}(C)=0$ for any constant $C \in K_{kc}(\mathcal{X})$.
- 2) $\text{Var}(aF) = a^2\text{Var}(F)$, for any $a > 0$.
- 3) $\text{Cov}(aF_1, F_2) = \text{Cov}(F_1, aF_2) = a\text{Cov}(F_1, F_2)$.
- 4) $\text{Var}(F_1 + F_2) = \text{Var}(F_1) + 2\text{Cov}(F_1, F_2) + \text{Var}(F_2)$.
- 5) $P(d_2(F, E(F)) \geq \varepsilon) \leq \frac{\text{Var}(F)}{\varepsilon^2}$, for any $\varepsilon > 0$.
- 6) $\text{Cov}(F_1 + F_2, F_3) = \text{Cov}(F_1, F_3) + \text{Cov}(F_2, F_3)$.
- 7) $|\rho| \leq 1$.
- 8) $\rho(F_1, F_2) = 0$ if F_1 and F_2 are independent.

Example 3. When $\mathcal{X} = \mathbb{R}^1$, and for any interval-valued random variable $F(\omega) = [a(\omega), b(\omega)] = (c(\omega); r(\omega))$, where $c(\omega)$ and $r(\omega)$ are the midpoint and range of F , respectively, then

$$E(F(\omega)) = [E(a(\omega)), E(b(\omega))], \\ \text{Var}(F(\omega)) = \text{Var}(a(\omega)) + \text{Var}(b(\omega)) \\ = 2\text{Var}(c(\omega)) + 2\text{Var}(r(\omega)).$$

And for any two interval-valued random variables $F_i(\omega) = [a_i(\omega), b_i(\omega)] = (c_i(\omega); r_i(\omega))$, $i = 1, 2$, where $c_i(\omega)$ and $r_i(\omega)$ are defined as before,

$$\text{Cov}(F_1(\omega), F_2(\omega)) = \\ \text{Cov}(a_1(\omega), a_2(\omega)) + \text{Cov}(b_1(\omega), b_2(\omega)) = \\ 2\text{Cov}(c_1(\omega), c_2(\omega)) + 2\text{Cov}(r_1(\omega), r_2(\omega)).$$

B. Linear regression model for interval-valued data

In this subsection, we briefly present three existing linear regression models for interval-valued data, and discuss advantages and disadvantages of these methods.

1) *Center model:* This model (also called CM method) was first proposed by Billard and Diday [18]. In this model only the midpoints of the intervals are considered to fitting a linear model. More precisely, suppose that the interval-valued variables Y can be explained by p interval-valued variables X_1, \dots, X_p . Typically we have a set of observation $\{(y_i, x_{i1}, \dots, x_{ip})\}_{i=1, \dots, n}$, where $x_{ij} = [a_{ij}, b_{ij}]$, $y_i = [y_{li}, y_{ui}]$, $i = 1, \dots, n, j = 1, \dots, p$. Let x_{ij}^c and y_i^c be, respectively, the midpoints of the intervals x_{ij} and y_i , defined as $x_{ij}^c = (a_{ij} + b_{ij})/2$, and $y_i^c = (y_{li} + y_{ui})/2$, $i = 1, \dots, n, j = 1, \dots, p$. The linear regression model is:

$$\mathbf{y}^c = \mathbf{X}^c \boldsymbol{\beta}^c + \boldsymbol{\varepsilon}^c, \quad (2)$$

where $\mathbf{y}^c = (y_1^c, \dots, y_n^c)^T$, and $\boldsymbol{\beta}^c = (\beta_0^c, \beta_1^c, \dots, \beta_p^c)^T$, $\mathbf{X}^c = (\mathbf{x}_1^c, \dots, \mathbf{x}_n^c)^T$, $\mathbf{x}_i^c = (1, x_{i1}^c, \dots, x_{ip}^c)^T$ ($i = 1, \dots, n$), $\boldsymbol{\varepsilon}^c = (\varepsilon_1^c, \dots, \varepsilon_n^c)^T$, $x_{ij}^c = (a_{ij} + b_{ij})/2$ and $y_i^c = (y_{li} + y_{ui})/2$. From (2), the sum of the squares of deviations is

$$S_{cm} = \sum_{i=1}^n (\varepsilon_i^c)^2, \quad (3)$$

which is the sum of the midpoint squared errors. It can also be interpreted as the sum of the square of the sum of the lower and upper bound errors.

The value of $\boldsymbol{\beta}$ minimizing (3) is the least square estimate. If \mathbf{X}^c has full rank $p + 1 \leq n$, then

$$\hat{\boldsymbol{\beta}}^c = ((\mathbf{X}^c)^T \mathbf{X}^c)^{-1} (\mathbf{X}^c)^T \mathbf{y}^c. \quad (4)$$

Based on $\hat{\boldsymbol{\beta}}^c$, given a new example $\mathbf{x} = (x_1, \dots, x_p)$ with $x_j = [a_j, b_j]$ ($j = 1, \dots, p$), the predicted value \mathbf{y} is

$$\hat{\mathbf{y}} = [\mathbf{x}_l^T \hat{\boldsymbol{\beta}}^c, \mathbf{x}_u^T \hat{\boldsymbol{\beta}}^c], \quad (5)$$

where $\mathbf{x}_l = (1, a_1, \dots, a_p)^T$, $\mathbf{x}_u = (1, b_1, \dots, b_p)^T$.

This is a simple model for fitting a linear regression model to interval-valued data. However, only the midpoints of the intervals are considered in the model; other information, like the range of the interval, is lost.

2) *Center and range method*: Based on the CM method, Lima Neto and De Carvalho [19] proposed the center and range model (CRM), in which both the centers and the ranges of the intervals are used for fitting a linear regression model. Let x_{ij}^c and y_i^c be defined as before. Meanwhile, let x_{ij}^r and y_i^r be, respectively, the range of the intervals x_{ij} and y_i , defined as $x_{ij}^r = (b_{ij} - a_{ij})/2$, $y_i^r = (y_{li} - y_{ui})/2$, $i = 1, \dots, n$, $j = 1, \dots, p$. The linear regression models are

$$\begin{aligned} \mathbf{y}^c &= \mathbf{X}^c \boldsymbol{\beta}^c + \boldsymbol{\varepsilon}^c, \\ \mathbf{y}^r &= \mathbf{X}^r \boldsymbol{\beta}^r + \boldsymbol{\varepsilon}^r, \end{aligned} \quad (6)$$

where $\mathbf{y}^r = (y_1^r, \dots, y_n^r)^T$, $\boldsymbol{\beta}^r = (\beta_0^r, \beta_1^r, \dots, \beta_p^r)^T$, $\mathbf{X}^r = (\mathbf{x}_1^r, \dots, \mathbf{x}_n^r)^T$, $\mathbf{x}_i^r = (1, x_{i1}^r, \dots, x_{ip}^r)^T$ ($i = 1, \dots, n$), $\boldsymbol{\varepsilon}^r = (\varepsilon_1^r, \dots, \varepsilon_n^r)^T$, and \mathbf{y}^c , \mathbf{X}^c , $\boldsymbol{\varepsilon}^c$ are defined in (2).

Thus, the sum of the squares of deviations is defined as

$$S_{crm} = \sum_{i=1}^n (\varepsilon_i^c)^2 + \sum_{i=1}^n (\varepsilon_i^r)^2, \quad (7)$$

which represents the sum of the midpoint square errors plus the sum of the range square errors. Assuming that the midpoint and range of the intervals are independent, minimizing the criterion above is equivalent to fitting two independent regression models over the midpoint and range of interval values.

If \mathbf{X}^c and \mathbf{X}^r have full rank $p+1 \leq n$, the least-square estimates of $\boldsymbol{\beta}^c$ is given by (4), and the least square estimate of $\boldsymbol{\beta}^r$ is

$$\hat{\boldsymbol{\beta}}^r = ((\mathbf{X}^r)^T \mathbf{X}^r)^{-1} (\mathbf{X}^r)^T \mathbf{y}^r. \quad (8)$$

The fitted value is

$$\hat{\mathbf{y}}^c = \mathbf{X}^c \hat{\boldsymbol{\beta}}^c, \quad \hat{\mathbf{y}}^r = \mathbf{X}^r \hat{\boldsymbol{\beta}}^r, \quad (9)$$

and $\hat{\mathbf{y}} = [\hat{\mathbf{y}}^c - \hat{\mathbf{y}}^r, \hat{\mathbf{y}}^c + \hat{\mathbf{y}}^r]$.

In comparison to the CM method, CRM fits two linear regression models using the information contained in the midpoints and ranges of the intervals in order to improve the model prediction performance. However, the CRM method cannot ensure the natural assumption that the predicted range $\hat{\mathbf{y}}^r$ should be not less than 0, namely, the predicted lower bound \hat{y}_l should be lower than or equal to the upper bound \hat{y}_u .

3) *Constrained center and range model*: To solve the deficiency of the CRM method, Lima Neto and De Carvalho later proposed a constrained version (called CCRM), to make sure that the lower bound of any prediction interval is less than or equal to its upper bound. The basic idea is to consider inequality constraints over the parameters $\boldsymbol{\beta}^r$ in (6), ensuring that the estimated values of $\hat{\mathbf{y}}^r$ will always be greater than or equal to zero. Remark that there are no restrictions on the parameters $\boldsymbol{\beta}^c$. The regression model of CCRM method is

$$\begin{aligned} \mathbf{y}^c &= \mathbf{X}^c \boldsymbol{\beta}^c + \boldsymbol{\varepsilon}^c, \\ \mathbf{y}^r &= \mathbf{X}^r \boldsymbol{\beta}^r + \boldsymbol{\varepsilon}^r, \\ \text{subject to } \beta_j^r &\geq 0, j = 1, \dots, p. \end{aligned} \quad (10)$$

The sum of squares of deviations is given by

$$S_{crm} = \sum_{i=1}^n (\varepsilon_i^c)^2 + \sum_{i=1}^n (\varepsilon_i^r)^2 \quad (11)$$

subject to $\beta_j^r \geq 0, j = 1, \dots, p$.

As no constraints are added to the center regression model, the least square estimates of $\boldsymbol{\beta}^c$ is given by (4) if \mathbf{X}^c has full rank. The estimation of $\boldsymbol{\beta}^r$ with constraints is obtained by adapting Lawson and Hanson's algorithm [22].

By ensuring all the parameters in the range model to be nonnegative, the CCRM method makes sure that the predicted upper bound of the interval is greater than or equal to the lower bound. But some problems still exists. Firstly, our ultimate aim is to ensure $\hat{\mathbf{y}}^r \geq 0$, rather than all the parameters $\hat{\beta}_j \geq 0, j = 1, \dots, p$. Secondly, the approach may be misleading if the relationship between variables is truly negative.

III. CONSTRAINED REGRESSION MODEL FOR INTERVAL-VALUED DATA

A. General framework and basic assumption

In order to overcome the limitations of the previous methods, we introduce a constrained regression model. More precisely, let X_1, \dots, X_p be p independent interval-valued variable, where $X_j = [X_{lj}, X_{uj}]$, $j = 1, \dots, p$, and let Y be the interval-valued dependent variable, where $Y = [Y_l, Y_u]$. We suppose that there is a linear relationship between Y and $X_j, j = 1, 2, \dots, p$,

$$\begin{pmatrix} Y_l \\ Y_u \end{pmatrix} = \begin{pmatrix} \beta_{l0} \\ \beta_{u0} \end{pmatrix} + \sum_{j=1}^p \begin{pmatrix} \beta_{l1}^j & \beta_{l2}^j \\ \beta_{u1}^j & \beta_{u2}^j \end{pmatrix} \begin{pmatrix} X_{lj} \\ X_{uj} \end{pmatrix} + \begin{pmatrix} \varepsilon_l \\ \varepsilon_u \end{pmatrix} \quad (12)$$

subject to $Y_l \leq Y_u$,

where $\varepsilon = (\varepsilon_l, \varepsilon_u)$ is an error term. Some explanations are in order:

- 1) In our model, the lower bound of Y is related to both the upper and lower bounds of X_j , and similarly for the upper bound of Y . Specially, when $\beta_{12} = \beta_{21} = 0$, the upper (resp., lower) bound of the interval is only related to the upper (resp., lower) bound of X_j .
- 2) As the upper bound of the interval should not be less than the lower bound, we cannot ignore the constraint condition $Y_l \leq Y_u$.

We define $\boldsymbol{\beta}_l = (\beta_{l0}, \beta_{l1}^1, \beta_{l1}^2, \beta_{l1}^3, \beta_{l1}^4, \dots, \beta_{l1}^p, \beta_{l2}^p)^T$, $\boldsymbol{\beta}_u = (\beta_{u0}, \beta_{u1}^1, \beta_{u1}^2, \beta_{u1}^3, \beta_{u1}^4, \dots, \beta_{u1}^p, \beta_{u2}^p)^T$, and $\mathbf{X} = (1, X_{l1}, X_{u1}, X_{l2}, X_{u2}, \dots, X_{lp}, X_{up})$. Then (12) can be written in matrix form

$$\begin{aligned} Y_l &= \mathbf{X} \boldsymbol{\beta}_l + \varepsilon_l \\ Y_u &= \mathbf{X} \boldsymbol{\beta}_u + \varepsilon_u \\ \text{subject to } Y_l &\leq Y_u. \end{aligned} \quad (13)$$

Suppose that we have observed n samples. Then we have the following equations:

$$\begin{aligned} Y_{li} &= \mathbf{X}_i \boldsymbol{\beta}_l + \varepsilon_{li} \\ Y_{ui} &= \mathbf{X}_i \boldsymbol{\beta}_u + \varepsilon_{ui} \\ \text{subject to } Y_{li} &\leq Y_{ui}, i = 1, 2, \dots, n, \end{aligned} \quad (14)$$

where $\mathbf{X}_i = (1, x_{li1}, x_{ui1}, x_{li2}, x_{ui2}, \dots, x_{lip}, x_{uip})$.

Before going on, we make some further assumptions about the sampling process:

- 1) We suppose that the probability density of $(\varepsilon_{li}, \varepsilon_{ui})$, denoted by $h(\varepsilon_{li}, \varepsilon_{ui})$, verifies $E(\varepsilon_{li}) = E(\varepsilon_{ui}) = 0$, $E(\varepsilon_{li}^2) = \sigma_l^2 < \infty$, $E(\varepsilon_{ui}^2) = \sigma_u^2 < \infty$, $E(\varepsilon_{li}\varepsilon_{ui}) = \sigma_{lu}^2$, and $E(\varepsilon_{li}\varepsilon_{uj}) = 0$, ($i \neq j$), $i, j = 1, 2, \dots, n$
- 2) We assume that the matrix $(\mathbf{X}_1^T, \mathbf{X}_2^T, \dots, \mathbf{X}_n^T)^T$ has full rank, its size being $n \times (2p + 1)$.

From now on, we discuss the expectation of the errors in (14) based on the condition $Y_{li} \leq Y_{ui}$. Let $G_{li} = \mathbf{X}_i\beta_l$, $G_{ui} = \mathbf{X}_i\beta_u$, $\Delta G_i = G_{ui} - G_{li}$, $\Delta Y_i = Y_{ui} - Y_{li}$. Due to the constraint $Y_{li} \leq Y_{ui}$, that is $\Delta Y_i \geq 0$, we have $G_{li} + \varepsilon_{li} \leq G_{ui} + \varepsilon_{ui}$, which means $\Delta\varepsilon_i = \varepsilon_{ui} - \varepsilon_{li} \geq G_{li} - G_{ui} = -\Delta G_i$.

For given X_i , Y_i and the constraints $\Delta Y_i \geq 0$, the conditional expectations of ε_{li} and ε_{ui} are

$$\begin{aligned} E[\varepsilon_{li} | \mathbf{X}_i, \Delta Y_i \geq 0] &= E[\varepsilon_{li} | \mathbf{X}_i, \Delta\varepsilon_i \geq -\Delta G_i], \\ E[\varepsilon_{ui} | \mathbf{X}_i, \Delta Y_i \geq 0] &= E[\varepsilon_{ui} | \mathbf{X}_i, \Delta\varepsilon_i \geq -\Delta G_i]. \end{aligned}$$

Since ε_{li} and ε_{ui} are dependent on $\Delta\varepsilon_i = \varepsilon_{ui} - \varepsilon_{li}$, when $P\{\Delta\varepsilon_i \geq -\Delta G_i\} \neq 0$, the conditional expectations are not zero any more, which can lead to systematic bias of the least square estimates. Thus, we consider a two-step procedure: we first need to modify the model in such a way that the expected errors are equal to zero, and then we can use the least square estimation, as done in [23].

Before deriving the parameter estimates, an additional assumption is first presented, that is, $h(\varepsilon_l, \varepsilon_u)$ is a bivariate normal distribution.

Assumption 1. *The error term $\varepsilon = (\varepsilon_l, \varepsilon_u)$ has a bivariate normal distribution, with joint probability density*

$$f(\varepsilon) = (2\pi)^{-1} |\Sigma|^{-1/2} \exp\{-\varepsilon^T \Sigma^{-1} \varepsilon / 2\},$$

$$\text{where } \Sigma = \begin{pmatrix} \sigma_l^2 & \rho\sigma_l\sigma_u \\ \rho\sigma_l\sigma_u & \sigma_u^2 \end{pmatrix}.$$

The motivation for this assumption is technical: the combination of normal distributions is still a normal distribution, and it is easy to compute conditional probability density under truncation conditions.

B. Preliminaries

For convenience, we omit the subscript i . The difference of the two equations in (13) is $Y_u - Y_l = \mathbf{X}\beta_u + \varepsilon_u - \mathbf{X}\beta_l - \varepsilon_l$, which can also be written in equivalent form

$$\Delta Y = \mathbf{X}\Delta\beta + \Delta\varepsilon, \quad (15)$$

As $\Delta\varepsilon$ is the difference between ε_u and ε_l , from the property of the normal distribution we know that $\Delta\varepsilon$ has a normal distribution, with $E(\Delta\varepsilon) = E(\varepsilon_u) - E(\varepsilon_l) = 0$, $Var(\Delta\varepsilon) = Var(\varepsilon_u) + Var(\varepsilon_l) - 2Cov(\varepsilon_u, \varepsilon_l) = \sigma_u^2 + \sigma_l^2 - 2\rho\sigma_u\sigma_l$. Let $\sigma_m^2 = Var(\Delta\varepsilon)$, then $\Delta\varepsilon \sim N(0, \sigma_m^2)$, or equivalently $\Delta\varepsilon/\sigma_m \sim N(0, 1)$.

Remark that if $\sigma_m = 0$, $\Delta\varepsilon$ boils down to zero almost everywhere, and (15) becomes a non random equation, in which case the condition $\Delta Y \geq 0$ is useless: we can directly

use least-square estimation. Thus, to make sure $\sigma_m > 0$, $\sigma_u = \sigma_l$ and $\rho = 1$ cannot hold at the same time.

Proposition 1. (1) $\forall c \in \mathbb{R}$, the probability density of a random variable $\Delta\varepsilon$ under the condition $\{\Delta\varepsilon \geq c\}$ is

$$f(z | \Delta\varepsilon \geq c) = \begin{cases} \frac{\phi(z/\sigma_m)}{\sigma_m(1 - \Phi(c/\sigma_m))} & z \geq c, \\ 0 & z < c. \end{cases}$$

(2) Let $\Delta Y = \Delta\varepsilon + \Delta G$, the probability density of ΔY under the condition $\{\Delta Y \geq 0\}$ is

$$g(\Delta y | \Delta Y \geq 0) = \frac{\phi(\Delta y/\sigma_m - \Delta G/\sigma_m)}{\sigma_m(1 - \Phi(-\Delta G/\sigma_m))}.$$

where ϕ and Φ are, respectively, the normal probability density and cumulative distribution functions.

Based on Assumption 1 and results in [24], we have the following lemma:

Lemma 1. *Under Assumption 1, the conditional expectation of the error term is*

$$\begin{aligned} E(\varepsilon_l | \mathbf{X}, \Delta\varepsilon \geq -\Delta G) &= \frac{\rho\sigma_l\sigma_u - \sigma_l^2}{\sigma_m} \tau, \\ E(\varepsilon_u | \mathbf{X}, \Delta\varepsilon \geq -\Delta G) &= \frac{\sigma_u^2 - \rho\sigma_l\sigma_u}{\sigma_m} \tau, \end{aligned}$$

where $\tau = \phi(z)/(1 - \Phi(z)) = \phi(z)/\Phi(-z)$, $z = -\Delta G/\sigma_m$. Parameter τ is called the inverse of Mill's ratio.

We define $C_l = \frac{\rho\sigma_l\sigma_u - \sigma_l^2}{\sigma_m}$ and $C_u = \frac{\sigma_u^2 - \rho\sigma_l\sigma_u}{\sigma_m}$. Then, the conditional expectations of Y_{li} and Y_{ui} are

$$\begin{aligned} E(Y_{li} | \mathbf{X}_i, \Delta Y_i \geq 0) &= \mathbf{X}_i\beta_l + C_l\tau_i, \\ E(Y_{ui} | \mathbf{X}_i, \Delta Y_i \geq 0) &= \mathbf{X}_i\beta_u + C_u\tau_i. \end{aligned}$$

Moreover, the regression model with constraints can be written as

$$\begin{aligned} Y_{li} &= E(Y_{li} | \mathbf{X}_i, \tau_i, \Delta Y_i \geq 0) + V_{li}, \\ Y_{ui} &= E(Y_{ui} | \mathbf{X}_i, \tau_i, \Delta Y_i \geq 0) + V_{ui}, \end{aligned} \quad (16)$$

where

$$\begin{aligned} V_{li} &= \varepsilon_{li} - C_l\tau_i, \quad V_{ui} = \varepsilon_{ui} - C_u\tau_i, \\ E(V_{li} | \mathbf{X}_i, \tau_i, \Delta\varepsilon_i \geq -\Delta G_i) &= 0, \\ E(V_{ui} | \mathbf{X}_i, \tau_i, \Delta\varepsilon_i \geq -\Delta G_i) &= 0, \\ E(V_{li}^2 | \mathbf{X}_i, \tau_i, \Delta\varepsilon_i \geq -\Delta G_i) &= \sigma_l^2 - \frac{\sigma_l^2(\rho\sigma_u - \sigma_l)^2}{\sigma_m^2} \frac{\Delta G_i}{\sigma_m} \tau_i, \\ E(V_{ui}^2 | \mathbf{X}_i, \tau_i, \Delta\varepsilon_i \geq -\Delta G_i) &= \sigma_u^2 - \frac{\sigma_l^2(\sigma_u - \rho\sigma_l)^2}{\sigma_m^2} \frac{\Delta G_i}{\sigma_m} \tau_i, \\ E(V_{li}V_{ui} | \mathbf{X}_i, \tau_i, \Delta\varepsilon_i \geq -\Delta G_i) &= \rho\sigma_l\sigma_u - \frac{\sigma_l\sigma_u(\rho\sigma_u - \sigma_l)(\sigma_u - \rho\sigma_l)}{\sigma_m^2} \frac{\Delta G_i}{\sigma_m} \tau_i. \end{aligned}$$

This model can be viewed as a multiple linear regression model with two dependent variables Y_l and Y_u , and $2p + 1$ independent variables $X_{l1}, X_{u1}, \dots, X_{lp}, X_{up}, \tau$. Once we have computed τ_i , we can obtain the least-square estimates $\hat{\beta}_l$ and $\hat{\beta}_u$. It should be noted that the least-square estimators are unbiased but inefficient, which is the consequence of heteroscedasticity. More details can be found in [23].

C. Two-step estimation

1) *First step*: In this step, we want to estimate τ_i . To this end, we model the range of the interval (14),

$$\Delta Y_i = \mathbf{X}_i \Delta \beta + \Delta \varepsilon_i. \quad (17)$$

From Proposition 1, the probability density of Δy_i is

$$f(\Delta Y_i | \Delta Y_i \geq 0) = \frac{1}{\sigma_m} \frac{\phi(\Delta Y_i / \sigma_m + (G_{li} - G_{ui}) / \sigma_m)}{1 - \Phi((G_{li} - G_{ui}) / \sigma_m)}. \quad (18)$$

Based on (18), we can construct the log-likelihood function of n observations

$$n^{-1} L(\Delta \mathbf{Y}; \Delta \beta, \sigma_m) = \frac{1}{n} \sum_{i=1}^n (\log f(\Delta Y_i | \Delta Y_i \geq 0)). \quad (19)$$

By maximizing (19), $\max_{\Delta \beta, \sigma_m} [n^{-1} L(\Delta \mathbf{Y}; \Delta \beta, \sigma_m)]$, we can obtain the ML estimators $\hat{\Delta \beta}$ and $\hat{\sigma}_m$. The ML estimators are substituted to obtain $\hat{\tau}_i$.

2) *Second step*: After plugging $\hat{\tau}_i$ into (16), we can obtain the following multiple linear regression model with two dependent variables:

$$\mathbf{Y} = \mathbf{X} \beta + \mathbf{V}, \quad (20)$$

where

$$\mathbf{Y} = \begin{pmatrix} y_{l1} & y_{l2} & \cdots & y_{ln} \\ y_{u1} & y_{u2} & \cdots & y_{un} \end{pmatrix}^T, \quad \mathbf{V} = \begin{pmatrix} V_{l1} & V_{l2} & \cdots & V_{ln} \\ V_{u1} & V_{u2} & \cdots & V_{un} \end{pmatrix}^T, \quad \mathbf{X} = \begin{pmatrix} 1 & x_{l11} & x_{u11} & \cdots & x_{l1p} & x_{u1p} & \hat{\tau}_1 \\ 1 & x_{l21} & x_{u21} & \cdots & x_{l2p} & x_{u2p} & \hat{\tau}_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 1 & x_{ln1} & x_{un1} & \cdots & x_{lnp} & x_{unp} & \hat{\tau}_n \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_{l0} & \beta_{l1}^p & \beta_{l2}^p & \cdots & \beta_{l1}^p & \beta_{l2}^p & C_l \\ \beta_{u0} & \beta_{u1}^p & \beta_{u2}^p & \cdots & \beta_{u1}^p & \beta_{u2}^p & C_u \end{pmatrix}^T.$$

The corresponding residual sum of squares is given by

$$\begin{aligned} S &= \sum_{i=1}^n d_2^2([y_{li}, y_{ui}], [\hat{y}_{li}, \hat{y}_{ui}]) \\ &= \sum_{i=1}^n (y_{li} - \hat{y}_{li})^2 + \sum_{i=1}^n (y_{ui} - \hat{y}_{ui})^2 \\ &= \sum_{i=1}^n (V_{li}^2 + V_{ui}^2). \end{aligned} \quad (21)$$

Minimizing (21) gives the least square estimation of β

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

TABLE I
CONFIGURATIONS OF DATA GENERATING PROCESSES FOR SIMULATION

| Parameters | DGP1 | DGP2 | DGP3 | DGP4 |
|-------------------------------|---------|---------|---------|---------|
| β_{l0} | 0 | 0 | 0 | 0 |
| β_{u0} | 0.2 | 0.2 | 0.2 | 0.2 |
| β_{l1} | 0.8 | 0.8 | 0.85 | 0.85 |
| β_{l2} | 0.1 | 0.1 | 0.15 | 0.15 |
| β_{21} | 0.1 | 0.1 | 0.1 | 0.1 |
| β_{22} | 0.8 | 0.8 | 0.8 | 0.8 |
| C_l | -0.2236 | -0.6524 | -0.2236 | -0.6524 |
| C_u | 0.2236 | -0.1603 | 0.2236 | -0.1603 |
| σ_l^2 | 0.1 | 0.5 | 0.1 | 0.5 |
| σ_u^2 | 0.1 | 0.1 | 0.1 | 0.1 |
| ρ | 0 | 0.8 | 0 | 0.8 |
| Sample sizes T=200, 500, 1000 | | | | |
| Number of replication=1000 | | | | |

IV. NUMERICAL STUDIES

To show the usefulness of proposed method in our paper, experiments on simulated and real interval-valued data are reported in this section.

A. Simulation Experiment

The data generating process (DGP) is governed by the regression model

$$\begin{pmatrix} Y_l \\ Y_u \end{pmatrix} = \begin{pmatrix} \beta_{l0} \\ \beta_{u0} \end{pmatrix} + \begin{pmatrix} \beta_{l1} & \beta_{l2} \\ \beta_{21} & \beta_{22} \end{pmatrix} \begin{pmatrix} X_l \\ X_u \end{pmatrix} + \begin{pmatrix} \varepsilon_l \\ \varepsilon_u \end{pmatrix} \quad (22)$$

subject to $Y_l \leq Y_u$,

where $\varepsilon = (\varepsilon_l, \varepsilon_u) \sim N(0, \Sigma)$.

To guarantee the inequality constraint $y_l \leq y_u$, which is also equivalent to $\Delta \varepsilon \geq -\Delta G$, we proceeded as follows. After obtaining $[x_l, x_u]$, we first calculated ΔG , $\Delta G = G_u - G_l$, where $G_u = \beta_{u0} + \beta_{21}x_l + \beta_{22}x_u$, $G_l = \beta_{l0} + \beta_{l1}x_l + \beta_{l2}x_u$. Then the error term $(\varepsilon_l, \varepsilon_u)$ was drawn from the truncated bivariate normal distribution with the truncation condition $\Delta \varepsilon \geq -\Delta G$. Finally, Y_l and Y_u were obtained, respectively, as $Y_l = G_l + \varepsilon_l$, $Y_u = G_u + \varepsilon_u$.

Through the above DGP, we can ensure that the error term has the truncated bivariate normal distribution. Furthermore, all the pairs $[Y_l, Y_u]$ satisfy inequality constraint.

We considered the four different configurations shown in Table I. In two configurations, the coefficients of the independent variables are symmetric ($\beta_{l1} = \beta_{22}$ and $\beta_{l2} = \beta_{21}$). Two different forms of covariance matrices Σ for error term were considered in our experiments: one with zero correlation coefficient ($\sigma_l^2 = \sigma_u^2 = 0.1$, $\rho = 0$) and the other with larger variance and highly correlated errors ($\sigma_l^2 = 0.5$, $\sigma_u^2 = 0.1$, $\rho = 0.8$). For each DGP, the results obtained by the two-step procedure (Two-step) and the classical least square (LS) estimation are displayed in Tables II and III. To illustrate the effect of sample size, three different sample size were considered: 200, 500 and 1000, with 1000 replications per DGP. Due to space limitation, only parts of the results are presented in Table II for DGP1 and Table III for DGP4.

Some conclusions can be drawn from Table III:

- 1) The means of LS estimates are close to the true values for DGP1; however for both DGP2 (with asymmetric coefficients) and DGP3 (with highly correlated errors),

TABLE II
SIMULATION RESULTS FOR DGP1

| Parameters | Two-step | | | | | |
|----------------------------|----------|--------|---------|--------|---------|--------|
| | T=200 | | T=500 | | T=1000 | |
| | Mean | RMSE | Mean | RMSE | Mean | RMSE |
| $\beta_{10} = 0$ | 0.0018 | 0.1362 | 0.0015 | 0.0814 | 0.0023 | 0.0569 |
| $\beta_{u0} = 0.2$ | 0.1969 | 0.1330 | 0.1984 | 0.0853 | 0.1995 | 0.5736 |
| $\beta_{11} = 0.8$ | 0.7989 | 0.0882 | 0.7993 | 0.0526 | 0.8017 | 0.3733 |
| $\beta_{12} = 0.1$ | 0.1008 | 0.0880 | 0.1002 | 0.0524 | 0.0982 | 0.3709 |
| $\beta_{21} = 0.1$ | 0.0989 | 0.0864 | 0.1001 | 0.0550 | 0.0998 | 0.0376 |
| $\beta_{22} = 0.8$ | 0.8012 | 0.0861 | 0.8000 | 0.0549 | 0.8002 | 0.0372 |
| $C_l = -0.2236$ | -0.2245 | 0.3227 | -0.2225 | 0.2050 | -0.2266 | 0.1362 |
| $C_u = 0.2236$ | 0.2264 | 0.3288 | 0.2250 | 0.2059 | 0.2241 | 0.1442 |
| | LS | | | | | |
| $\beta_{10} = 0$ | -0.0829 | 0.1000 | -0.0797 | 0.0874 | -0.0806 | 0.0845 |
| $\beta_{u0} = 0.2$ | 0.2811 | 0.0983 | 0.2811 | 0.0879 | 0.2810 | 0.0845 |
| $\beta_{11} = 0.8$ | 0.7475 | 0.0649 | 0.7474 | 0.0580 | 0.7469 | 0.0557 |
| $\beta_{12} = 0.1$ | 0.1528 | 0.0648 | 0.1524 | 0.0576 | 0.1529 | 0.0554 |
| $\beta_{21} = 0.1$ | 0.1527 | 0.0652 | 0.1532 | 0.0579 | 0.1524 | 0.0549 |
| $\beta_{22} = 0.8$ | 0.7470 | 0.0649 | 0.7470 | 0.0576 | 0.7475 | 0.0549 |
| Number of replication=1000 | | | | | | |

TABLE III
SIMULATION RESULTS FOR DGP4

| Parameters | Two-step | | | | | |
|----------------------------|----------|--------|---------|--------|---------|--------|
| | T=200 | | T=500 | | T=1000 | |
| | Mean | RMSE | Mean | RMSE | Mean | RMSE |
| $\beta_{10} = 0$ | -0.0001 | 0.2795 | -0.0044 | 0.1737 | -0.0069 | 0.1262 |
| $\beta_{u0} = 0.2$ | 0.2026 | 0.1340 | 0.1983 | 0.0791 | 0.1977 | 0.0583 |
| $\beta_{11} = 0.85$ | 0.8549 | 0.2319 | 0.8459 | 0.1500 | 0.8438 | 0.1066 |
| $\beta_{12} = 0.15$ | 0.1474 | 0.2079 | 0.1538 | 0.1346 | 0.1559 | 0.0961 |
| $\beta_{21} = 0.1$ | 0.1044 | 0.1205 | 0.0996 | 0.0728 | 0.0979 | 0.0530 |
| $\beta_{22} = 0.8$ | 0.7967 | 0.1066 | 0.8006 | 0.0643 | 0.8020 | 0.0470 |
| $C_l = -0.6524$ | -0.6644 | 0.2945 | -0.6550 | 0.1809 | -0.6506 | 0.1283 |
| $C_u = -0.1603$ | -0.1689 | 0.1706 | -0.1627 | 0.1011 | -0.1602 | 0.0716 |
| | LS | | | | | |
| $\beta_{10} = 0$ | -0.4328 | 0.4452 | -0.4327 | 0.4381 | -0.4322 | 0.4345 |
| $\beta_{u0} = 0.2$ | 0.0927 | 0.1214 | 0.0940 | 0.1117 | 0.0942 | 0.1086 |
| $\beta_{11} = 0.85$ | 0.4103 | 0.4454 | 0.4089 | 0.4435 | 0.4099 | 0.4413 |
| $\beta_{12} = 0.15$ | 0.5314 | 0.3876 | 0.5325 | 0.3852 | 0.5315 | 0.3827 |
| $\beta_{21} = 0.1$ | -0.0072 | 0.1135 | -0.0072 | 0.1098 | -0.0076 | 0.1089 |
| $\beta_{22} = 0.8$ | 0.8930 | 0.0999 | 0.8929 | 0.0958 | 0.8932 | 0.0946 |
| Number of replication=1000 | | | | | | |

the mean values of LS estimates are quite far from the true values. This is also true for the results for DGP4 (with both asymmetric coefficients and highly correlated errors).

- 2) The means of the two-step estimates are very close to the true values for all cases. The results get better when the sample size is increased, and the RMSE is much smaller.

The results obtained from two-step estimation are much better than those from LS estimation, which confirms our previous discussion. To further illustrate the effectiveness of the proposed method, we compared two-step (Two-step) estimation with the center and range model (CRM). The constrained center and range model (CCRM) was also taken into account.

To evaluate the performances of different methods, three criteria were used: (1) Root Mean Squared Error for upper and lower bounds of the intervals, (2) Coverage (CR) and Efficiency Rates (ER) of the predicted intervals, (3) Mean Distance Error (MDE) between the predicted and true intervals. For a sample of size n , assuming that the fitted values $\hat{y}_i = [\hat{y}_{li}, \hat{y}_{ui}] (i = 1, \dots, i = n)$ of the true values $y_i = [y_{li}, y_{ui}] (i = 1, \dots, n)$ have been obtained by each method, the criteria were defined as follows:

TABLE IV
CONFIGURATIONS OF DATA GENERATING PROCESSES FOR COMPARING

| | β_{10} | β_{u0} | β_{11} | β_{12} | β_{21} | β_{22} | σ_l | σ_u | ρ |
|------|--------------|--------------|--------------|--------------|--------------|--------------|------------|------------|--------|
| DGP1 | 0 | 0.2 | -0.8 | -0.1 | 0.1 | 0.8 | 0.1 | 0.1 | 0 |
| DGP2 | 0 | 0.2 | -0.8 | -0.1 | 0.1 | 0.8 | 0.5 | 0.1 | 0.8 |
| DGP3 | 0 | 0.2 | -0.85 | -0.15 | 0.1 | 0.8 | 0.1 | 0.1 | 0 |
| DGP4 | 0 | 0.2 | -0.85 | -0.15 | 0.1 | 0.8 | 0.5 | 0.1 | 0.8 |

TABLE V
EXPERIMENTAL EVALUATION FOR DGP1

| | RMSE | | CR & ER | | | MDE | |
|----------|---------------|---------------|---------|--------|---------------|---------------|---------------|
| | Lower | Upper | CR | ER | (CR+ER)/2 | $q = 1$ | $q = 2$ |
| CRM | 2.7092 | 2.6614 | 0.8555 | 0.7731 | 0.8143 | 2.3507 | 2.6854 |
| CCRM | 2.6787 | 2.6430 | 0.8543 | 0.7722 | 0.8132 | 2.3341 | 2.6609 |
| Two-step | 0.3281 | 0.3211 | 0.9620 | 0.9583 | 0.9601 | 0.2885 | 0.3246 |

- 1) RMSE:

$$RMSE_l = \sqrt{\sum_{i=1}^N (\hat{y}_{li} - y_{li})^2 / N},$$

$$RMSE_u = \sqrt{\sum_{i=1}^N (\hat{y}_{ui} - y_{ui})^2 / N};$$

- 2) CR and ER: $CR = \frac{1}{N} \sum_{i=1}^N w(y_i \cap \hat{y}_i) / w(y_i)$, $ER = \frac{1}{N} \sum_{i=1}^N w(y_i \cap \hat{y}_i) / w(\hat{y}_i)$, where $y_i \cap \hat{y}_i$ is the intersection of two intervals, and $w(\cdot)$ is the width of an interval. The average of the two rates $(CR + ER) / 2$ was used as an evaluation criterion.
- 3) MDE: Let $d_2(\hat{y}_i, y_i)$ be defined by (1), the mean distance error of the fitted and true intervals was defined as

$$MDE^q(\{\hat{y}_i\}, \{y_i\}) = \left[\sum_{i=1}^N d_2^q(\hat{y}_i, y_i) / \sqrt{2N} \right]^{1/q}.$$

In this paper, we considered two cases for q , namely, $q = 1$ and $q = 2$.

Again, four DGPs were considered, shown in Table IV. As in Table I, we compared these methods in different cases, correspond to symmetric/asymmetric coefficients, and correlated/uncorrelated errors. Moreover, to illustrate the fact that the constraints requiring the coefficients to be nonnegative in CCRM method are not suitable, negative values of the coefficients were considered in each DGP. Only parts of the results are reported in Table V for DGP1 and VI for DGP4.

From these results, we can see that, as the constrained version of CRM method, CCRM method does not make a big improvement, and sometimes is worse than CRM method, which confirms our point. Our method is obviously superior to the two reference methods for all cases.

TABLE VI
EXPERIMENTAL EVALUATION FOR DGP4

| | RMSE | | CR & ER | | | MDE | |
|----------|---------------|---------------|---------|--------|---------------|---------------|---------------|
| | Lower | Upper | CR | ER | (CR+ER)/2 | $q = 1$ | $q = 2$ |
| CRM | 2.8838 | 2.7666 | 0.8515 | 0.7677 | 0.8096 | 2.5054 | 2.8258 |
| CCRM | 2.8824 | 2.7962 | 0.8562 | 0.7767 | 0.8165 | 2.4818 | 2.8396 |
| Two-step | 0.7003 | 0.3136 | 0.9365 | 0.9370 | 0.9367 | 0.4490 | 0.5426 |

TABLE VII
EVALUATION FOR CARDIOLOGICAL INTERVAL DATA

| | RMSE | | CR & ER | | | MDE | |
|----------|---------------|---------------|---------|--------|-----------|---------------|---------------|
| | Lower | Upper | CR | ER | (CR+ER)/2 | $q=1$ | $q=2$ |
| CRM | 8.6026 | 8.6035 | 0.7199 | 0.6772 | 0.6985 | 7.9011 | 8.6030 |
| CCRM | 8.7186 | 8.5574 | 0.7208 | 0.6735 | 0.6972 | 7.9262 | 8.6384 |
| Two-step | 7.8304 | 8.3506 | 0.7043 | 0.6881 | 0.6962 | 7.2938 | 8.0947 |

B. Empirical Study

In this section, we apply the two-step procedure to real data. This is a cardiological interval dataset, including three attributes, namely: pulse rate $Y = [Y_l, Y_u]$, systolic blood pressure $X_1 = [X_{l1}, X_{u1}]$ and diastolic blood pressure $X_2 = [X_{l2}, X_{u2}]$ for eleven patients [18]. We aim to predict Y by the relationship

$$\begin{pmatrix} Y_l \\ Y_u \end{pmatrix} = \begin{pmatrix} \beta_{l0} \\ \beta_{u0} \end{pmatrix} + \sum_{i=1}^2 \begin{pmatrix} \beta_{l1}^i & \beta_{l2}^i \\ \beta_{u1}^i & \beta_{u2}^i \end{pmatrix} \begin{pmatrix} X_{li} \\ X_{ui} \end{pmatrix} + \begin{pmatrix} \varepsilon_l \\ \varepsilon_u \end{pmatrix}$$

subject to $Y_l \leq Y_u$,

where $\varepsilon = (\varepsilon_l, \varepsilon_u)$ is the error term.

The results obtained from our method, as well as the CRM and CCRM methods, are shown in Table VII. We can see that the CCRM method yields similar results to those of CRM. The two-step method is slightly better than the other two in terms of RMSE and MDE, and the three methods perform similarly in terms of average CR and ER.

V. CONCLUSION

Existing linear regression methods for interval-valued data, such as the CM and CRM methods, cannot ensure that the predicted lower bound of the interval is lower than or equal to its upper bound. To meet this condition, the CCRM method imposes nonnegativity constraint on the coefficients in the range model. However, we argue that this restriction is too strict.

A new constrained linear regression model for interval-valued data has been proposed in this paper. In our model, the lower and upper bounds of dependent variable are associated with both the lower and upper bounds of the independent variables. As a result, our model is more flexible than the other models. We provided a two-step procedure for this model, in which an auxiliary variable is first introduced before applying the least-square procedure. Experiments on both simulated data and real data have shown that our method outperforms the CRM and CCRM method.

There is still room for improvement in this method. For example, as the model is heteroscedastic, generalized least square estimation can be used, and some properties of the estimates should be deduced. Furthermore, other forms of error distribution could also be studied.

ACKNOWLEDGMENT

This research has been supported by NSFC (No.11571024) and by a grant from the Beijing Government as part of the Overseas Talents program.

REFERENCES

- [1] H. E. Robbins, "On the measure of a random set," *Annals of Mathematical Statistics*, vol. 15, no. 1, pp. 70–74, 1944.
- [2] Z. Artstein and R. A. Vitale, "A strong law of large numbers for random compact sets," *The Annals of Probability*, vol. 3, no. 5, pp. 879–882, 1975.
- [3] J. P. Aubin, *Set-Valued Analysis*. Boston: Birkhäuser, 1999.
- [4] R. J. Aumann, "Integrals of set-valued functions," *Journal of Mathematical Analysis and Applications*, vol. 12, no. 1, pp. 1–12, 1965.
- [5] F. Hiai and H. Umegaki, "Integrals, conditional expectations, and martingales of multivalued functions," *Journal of Multivariate Analysis*, vol. 7, no. 7, pp. 149–182, 1977.
- [6] C. Hess, "On multivalued martingales whose values may be unbounded: martingale selectors and Mosco convergence," *Journal of Multivariate Analysis*, vol. 39, no. 1, pp. 175–201, 1991.
- [7] N. S. Papageorgiou, "On the theory of Banach space valued multifunctions: 2. set valued martingales and set valued measures," *Journal of Multivariate Analysis*, vol. 17, no. 2, pp. 207–227, 1985.
- [8] —, "On the conditional expectation and convergence properties of random sets," *Transactions of the American Mathematical Society*, vol. 347, no. 7, pp. 2495–2515, 1995.
- [9] S. Li and Y. Ogura, "Convergence of set valued sub- and super-martingales in the Kuratowski-Mosco sense," *The Annals of Probability*, vol. 26, no. 3, pp. 1384–1402, 1998.
- [10] —, "Convergence of set-valued and fuzzy-valued martingales," *Fuzzy Sets and Systems*, vol. 101, no. 3, pp. 453–461, 1999.
- [11] S. Li, J. Li, and X. Li, "Stochastic integral with respect to set-valued square integrable martingales," *Journal of Mathematical Analysis and Applications*, vol. 370, no. 2, pp. 659–671, 2010.
- [12] S. Li, Y. Ogura, and V. Kreinovich, Eds., *Limit Theorems and Applications of Set-Valued and Fuzzy Set-Valued Random Variables*. London: Springer, 2002.
- [13] I. Molchanov, *Theory of Random Sets*. London: Springer, 2005.
- [14] W. Zhang, S. Li, and Y. Gao, *Set-Valued Stochastic Processes*. Science Publisher (in Chinese), 2007.
- [15] X. Yang and S. Li, "The d_p -metric space of set-valued random variables and its application to covariances," *Journal of Multivariate Analysis*, vol. 1, no. 1, pp. 73–82, 2005.
- [16] Á. Blanco, N. Corral, G. González-Rodríguez, and M. A. Lubiano, *Some Properties of the d_k -Variance for Interval-Valued Random Sets*. Berlin Heidelberg: Springer, 2008.
- [17] X. Wang, S. Li, and T. Denœux, "Interval-valued linear model," *International Journal of Computational Intelligence Systems*, vol. 8, no. 1, pp. 114–127, 2015.
- [18] L. Billard and E. Diday, *Regression Analysis for Interval-Valued Data*. Springer Berlin Heidelberg, 2000.
- [19] E. D. A. L. Neto and F. D. A. T. D. Carvalho, "Centre and range method for fitting a linear regression model to symbolic interval data," *Computational Statistics and Data Analysis*, vol. 52, no. 3, pp. 1500–1515, 2008.
- [20] —, "Constrained linear regression models for symbolic interval-valued variables," *Computational Statistics and Data Analysis*, vol. 54, no. 2, pp. 333–347, 2010.
- [21] X. Wang, Z. Zhang, and S. Li, "Set-valued and interval-valued stationary time series," *Journal of Multivariate Analysis*, vol. 145, pp. 208–223, 2016.
- [22] C. L. Lawson and R. J. Hanson, *Solving least squares problems*. Prentice-Hall, 1974.
- [23] J. J. Heckman, "Sample selection bias as a specification error," *Econometrica*, vol. 47, no. 1, pp. 153–161, 1979.
- [24] G. B. Nath, "Moments of a linearly truncated bivariate normal distribution," *Australian and New Zealand Journal of Statistics*, vol. 14, no. 2, pp. 97–102, 2008.