# Interval-Valued Linear Model

**Xun Wang**
Beijing University of Technology
Beijing, China
wangxun@emails.bjut.edu.cn

**Shoumei Li**
Beijing University of Technology
Beijing, China
lisma@bjut.edu.cn

**Thierry Denoeux**
Universitié de Technologie de
Compiègne, Heudiasyc, CNRS
Compiègne, France
thierry.denoeux@hds.utc.fr

## Abstract

This paper introduces a new type of statistical model: the interval-valued linear model, which describes the linear relationship between an interval-valued output random variable and real-valued input variables. Firstly, we discuss the notions of variance and covariance of set-valued and interval-valued random variables. Then, we give the definition of the interval-valued linear model and its least square estimation, as well as some properties of the least square estimation. Thirdly, we show that, whereas the best linear unbiased estimation does not exist, the best binary linear unbiased estimator exists and it is just the least square estimator. Finally, we present simulation experiments and an application example regarding temperature of cities affected by their latitude, which illustrates the application of our model.

**Keywords.** Interval-valued linear model, least square estimation, best binary linear unbiased estimation, $D_p$ metric.

## 1 Introduction

Traditional statistical models have played a significant role in a wide range of areas. However, in real life situations, many problems cannot be handled by traditional statistical models due to imperfectness of data. Therefore, specialized statistical techniques are needed. In many practical cases, we often have to face a particular kind of imperfect data: interval-valued data (e.g., [8], [9] and [13]).

Interval-valued data may represent uncertainty or variability. In the former case, the interval data represent incomplete observations, i.e., we just know the true data belong to a range (an interval), rather than the precise values. For example, assume that researchers test the service life of a group of products, such as light bulbs. Since testing time is very long, they cannot stay in the laboratory at any time. They could come to the laboratory to see how many bulbs are burnt out every two or three hours. Then, the data regarding service life of bulbs they get are interval-valued. In contrast, in the variability case, an interval is not interpreted as a set containing a single true value, but the observation themselves are interval-valued. For instance, a weather forecast typically provides the highest and lowest temperature of the next day, which is an interval including almost all the useful information about tomorrow's temperature. This interval reflects variability of temperature of one day.

The linear model is probably the simplest and most common statistical model. It describes a random output variable determined by a few input variables and an error term in a linear way. In this paper, we consider the situation in which observations are interval-valued, i.e., the random variable is an interval-valued random variable, which is determined by real-valued variables in a linear way. This interval-valued linear model could play a significant role in dealing with imperfect data, e.g., to investigate how (interval-valued) temperature is impacted by (point-valued) intensity of solar radiation, air pressure, latitude of location , or the statistical relationship between interval-valued service life of light bulbs and point-valued properties of materials used in making bulbs.

Interval-valued random variables are a special kind of set-valued random variables, whose values are compact convex subsets of the real line $\mathbb{R}^1$. Since we have at our disposal many results on the theory of set-valued random variables (e.g., [16], [17] and [26]), this is a suitable framework to tackle the problem addressed in this paper. For a long time, however, there has been only a few works to discuss the variance and covariance of set-valued random variables, since the difference between two sets is difficult to define and the hyperspace (e.g., the space of all intervals) is not linear with respect to addition and multiplication. Vital [21] studied the metric for compact convex sets via the support functions. In 2005, Yang and Li [24], Yang [25] investigated the $d_p$ metric for sets and the $D_p$ metric in the space of set-valued random variables; they proposed to use the $D_p$ metric to define the variance and covariance of set-valued and interval-valued random variables, which proved to be a good approach to deal with this problem. In Chapter 5 of [25], Yang also built a linear regression model with interval-valued regression coefficients. The underlying space in [24] and [25] is $\mathbb{R}^d$. In 2008, Blanco et al. [4] defined the $d_K$-variance for interval-valued random variables with underlying space $\mathbb{R}^1$, which is a special case of [24] and [25].

Some other works about interval-valued and set-valued statistical models are as follows. Tanaka and Lee [19] introduced the interval linear regression model, which is not based on the interval-valued random variable framework, and estimated the coefficients using a quadratic optimization method. Blanco-Fernandez et al. [5] and Sinova et al. [18] investigated the linear relationship between two interval-valued random variables, considering the input variable as two real-valued random variables (center and radius of the interval). They gave the least square estimation of the coefficients under the $d_2$ metric of intervals. Blanco-Fernandez et al. [6] studied the strong consistency and asymptotic distributions of the least square estimator. Beresteanu and Molinari [3] investigated inference for partially observed models via the asymptotic approach; they supposed the observations to be uncertain and proposed an estimation method for the real-valued parameters. Hsu and Wu [14] investigated interval-valued time series and gave three evaluation criteria of estimation and forecast efficiency for interval-valued time series. Wang and Li [22] introduced a new type of interval-valued time series (the interval autoregressive time series model) and proposed methods for parameter estimation and forecasting based on the evaluation criteria in [14]. Wang and Li [23] investigated set-valued and interval-valued stationary time series, based on the definition of variance and covariance of set-valued and interval-valued random variables introduced in [24] and [25].

In this paper, we start with the set-valued framework and consider the interval-valued random variable as a special case of set-valued random variable. We then introduce the interval-valued linear model and its least square estimation, prove its unbiasedness and discuss the best binary unbiased estimation. Treating an interval-valued random variable as two separate point-valued random variables (the left- and right-endpoints of the interval, or the center and radius of the interval) is deemed to be unreasonable. One reason is that it is quite easy to obtain estimation or forecast results such that the left-endpoint is larger than the right-endpoint or the center is negative, because these two linear models are unrelated. In this paper, we also show the limitation of using two separate linear models in terms of forecast efficiency via a simulation experiment.

The organization of this paper is as follows. In Section 2, we define the variance and covariance of set-valued random variables based on the $d_p$ metric for sets and the $D_p$ metric for interval-valued random variables. In Section 3, we introduce the interval-valued linear model and its least square estimator (LSE), prove the unbiasedness of this LSE and give the covariance matrix of this estimator. In Section 4, we show that the best linear unbiased estimation does not exist in general, but the best binary linear unbiased estimation (BBLUE) exists and is unique, and the BBLUE is just the LSE. In Section 5, we present a simulation study to show the methodology, and illustrate the efficiency of estimations introduced in Sections 3 and 4. We then present another simulation experiment to compare our model with using two separate linear models. Finally, in Section 6, we use the interval-valued linear model to investigate the relationship between city temperature and latitude. This example also shows how this model can be used to deal with some practical problems.

Due to page limitation, we have to omit all the proofs of theorems in Sections 3 and 4 in this paper.

## 2 Variance and Covariance of Set-Valued Random Variables

### 2.1 $d_p$ Metric between Sets

In this section, we assume that $(\Omega, \mathcal{A}, P)$ is a probability space, $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ is a Banach space, $\mathbf{K}(\mathcal{X})$ is the family of all nonempty closed subsets of $\mathcal{X}$ and $\mathbf{K}_{kc}(\mathcal{X})$ is the family of all nonempty compact convex subsets of $\mathcal{X}$.

For any $A, B \in \mathbf{K}(\mathcal{X}), \lambda \in \mathbb{R}$, define

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

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

If $A, B \in \mathbf{K}_{kc}(\mathcal{X})$, then $A + B \in \mathbf{K}_{kc}(\mathcal{X})$.

For each $A \in \mathbf{K}_{kc}(\mathcal{X})$, the support function is defined by

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

where $\mathcal{X}^*$ is the dual space of $\mathcal{X}$, i.e., the set of all bounded linear functionals on $\mathcal{X}$. For example, if $\mathcal{X} = \mathbb{R}^1$, $\mathcal{X}^* = \mathbb{R}^1$. Take an interval $[a, b]$ with

$0 \leq a < b$, $x \in \mathbb{R}^1$, then $s(x, [a,b]) = \begin{cases} bx, & x \geq 0 \\ ax, & x < 0 \end{cases}$. Regarding the support function, we have the following properties:

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

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

For $1 \leq p < \infty$, take $A, B \in \mathbf{K}_{kc}(\mathcal{X})$. We define the metric $d_p$ on $\mathbf{K}_{kc}(\mathcal{X})$ ([1], [16], [24]) by

$$d_p(A,B) = \left[ \int\limits_{S^*} |s(x^*, A) - s(x^*, B)|^p d\mu \right]^{1/p},$$

where $S^*$ is the unit sphere of $\mathcal{X}^*$, i.e., $S^* = \{x^* \in \mathcal{X}^* : \|x^*\|_{\mathcal{X}^*} = 1\}$, $\mu$ is a measure on $(\mathcal{X}^*, \mathcal{B}(\mathcal{X}^*))$.

**Remark 2.1.** If $\mathcal{X} = \mathbb{R}^1$, then $\mathbf{K}_{kc}(\mathbb{R}^1) = \{[a,b] : -\infty < a \leq b < \infty\}$ is the family of all intervals on $\mathbb{R}^1$. If $A_1 = [a_1, b_1] = (c_1; r_1)$, $A_2 = [a_2, b_2] = (c_2; r_2)$, where $c_i = (a_i + b_i)/2$ and $r_i = (b_i - a_i)/2$ for $i = 1, 2$, then

$$A_1 + A_2 = [a_1 + a_2, b_1 + b_2] = (c_1 + c_2; r_1 + r_2)$$

$$kA_1 = (kc_1; |k| r_1)$$

and

$$\begin{aligned} d_p(A_1, A_2) &= [|a_2 - a_1|^p + |b_2 - b_1|^p]^{1/p} \\ &= [|(c_2 - c_1) - (r_2 - r_1)|^p \\ &\quad + |(c_2 - c_1) + (r_2 - r_1)|^p]^{1/p}. \end{aligned}$$

### 2.2 $D_p$ Metric Space of Set-Valued Random Variables

A set-valued mapping $F : \Omega \to \mathbf{K}(\mathcal{X})$ is called a set-valued random variable (e.g., [11], [16]) if, for each open subset $O$ of $\mathcal{X}$, $F^{-1}(O) \in \mathcal{A}$, where $F^{-1}(O) = \{\omega \in \Omega : F(\omega) \cap O \neq \emptyset\}$ and $\emptyset$ is the empty set. Any two set-valued random variables are considered *identical* if $F_1(\omega) = F_2(\omega)$ for almost every $\omega \in \Omega$ (for short, denoted by "$a.s.(P)$").

Let $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]$ denote the family of set-valued random variables taking values in $\mathbf{K}_{kc}(\mathcal{X})$.

The $D_p$ metric with respect to set-valued random variables is defined by

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

where $F_1, F_2 \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]$ ([24]).

**Remark 2.2.** If $\mathcal{X} = \mathbb{R}^1$, $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$ is the family of all interval-valued random variables. For $F_i \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$, $F_i(\omega) = [f_i(\omega), g_i(\omega)] = (c_i(\omega); r_i(\omega))$, where $f_i(\omega), g_i(\omega)$ are random variables and $f_i(\omega) \leq g_i(\omega)$, and $c_i(\omega) = (f_i(\omega) + g_i(\omega))/2, r_i(\omega) = (g_i(\omega) - f_i(\omega))/2$, $i = 1, 2$. By the definition of $D_p$, we have

$$\begin{aligned} & D_p(F_1(\omega), F_2(\omega)) \\ = \; & [E|f_2(\omega) - f_1(\omega)|^p + E|g_2(\omega) - g_1(\omega)|^p]^{1/p} \\ = \; & [E|(c_2(\omega) - c_1(\omega)) - (r_2(\omega) - r_1(\omega))|^p \\ & + E|(c_2(\omega) - c_1(\omega)) + (r_2(\omega) - r_1(\omega))|^p]^{1/p}. \end{aligned}$$

Let $\mathcal{L}^p[\Omega, \mathbf{K}_{kc}(\mathcal{X})] = \{F : E[\|F\|_{d_p}^p] < +\infty, F \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]\}$. Then we have the following theorem:

**Theorem 2.1.** *$(\mathcal{L}^p[\Omega, \boldsymbol{K}_{kc}(\mathbb{R}^d)], D_p)$ is a complete metric space for each $1 \leq p < \infty$.* [24]

### 2.3 Variance and Covariance of Set-Valued Random Variables

The expectation of set-valued random variable $F$ was introduced by Aumann [2].

**Definition 2.1.** *For each integrable bounded set-valued random variable $F$, which means $\sup\{\|f\| : f \in F\}$ has finite expectation, the Aumann integral of $F$, denoted by $E[F]$, is defined by*

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

*where $S_F = \{f : f(\omega) \in F(\omega)\ a.s.(P), and\ f$ is integrable$\}$ is called the selection of set-valued random variable $F$, $\int_\Omega f dP$ is the usual Bochner integral.*

The properties of the expectation of set-valued random variables have been discussed in [11] and [16].

However, since the space of subsets of $\mathcal{X}$ is not a linear space with respect to the addition and multiplication, the minus between two sets is difficult to define. Thus, extending the important notions of variance and the covariance to set-valued random variables is not a trivial task. Yang and Li [24] proposed to define variance and covariance using the $D_p$ metric on $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^d)]$, based on the fact that the support function of sets is subtractive. Later, Wang and Li [23] extended these definitions to the more general space $\mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathcal{X})]$.

**Definition 2.2.** *For each set-valued random variable $F \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$, the variance of $F$, denoted by* $\mathrm{Var}(F)$, *is defined as*

$$\begin{aligned} \mathrm{Var}(F) &= [D_2(F, E(F))]^2 \\ &= E\left\{ \int\limits_{S^*} [s(x^*, F(\omega)) - s(x^*, E(F(\omega)))]^2 d\mu \right\}. \end{aligned}$$

*For two set-valued random variables $F_1, F_2 \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$, the covariance of $F_1$ and $F_2$, denoted* *by* $\mathrm{Cov}(F_1, F_2)$*, is defined as*

$$\begin{aligned} & \mathrm{Cov}(F_1, F_2) \\ = \ & E\Bigg\{ \int\limits_{S^*} [s(x^*, F_1(\omega)) - s(x^*, E(F_1))] \\ & [s(x^*, F_2(\omega)) - s(x^*, E(F_2))] d\mu \Bigg\}. \end{aligned}$$

*The correlation coefficient of* $F_1$ *and* $F_2$*, denoted by* $\rho(\mathrm{F}_1, \mathrm{F}_2)$*, is defined as*

$$\rho(F_1, F_2) = \frac{\mathrm{Cov}(F_1, F_2)}{\sqrt{\mathrm{Var}(F_1) \cdot \mathrm{Var}(F_2)}}.$$

The variance, covariance and correlation coefficient of set-valued random variables have the following properties. The proofs of Theorem 2.3-2.6 can be found in [23].

**Theorem 2.2.** *The variance* $\mathrm{Var}(F)$ *of* $F \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$ *has the following properties:*

*(1)* $\mathrm{Var}(C) = 0$ *for any constant* $C \in \boldsymbol{K}_k(\mathcal{X})$*.*

*(2)* $\mathrm{Var}(aF) = a^2 \mathrm{Var}(F)$ *for any* $a \geq 0$*.*

*(3)* $\mathrm{Var}(F_1 + F_2) = \mathrm{Var}(F_1) + 2\mathrm{Cov}(F_1, F_2) + \mathrm{Var}(F_2)$*.*

*(4) (Chebyshev Inequality)* $P(d_2(F, E(F)) \geq \varepsilon)) \leq \mathrm{Var}(F)/\varepsilon^2$*, for any* $\varepsilon > 0$*.*

**Theorem 2.3.** *The covariance* $\mathrm{Cov}(F_1, F_2)$ *of* $F_1, F_2 \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$ *has the following properties:*

*(1)* $\mathrm{Cov}(aF_1, F_2) = \mathrm{Cov}(F_1, aF_2) = a\mathrm{Cov}(F_1, F_2)$ *for any* $a \geq 0$*.*

*(2)* $\mathrm{Cov}(F_1 + F_2, F_3) = \mathrm{Cov}(F_1, F_3) + \mathrm{Cov}(F_2, F_3)$*,* $\mathrm{Cov}(F_1, F_2 + F_3) = \mathrm{Cov}(F_1, F_2) + \mathrm{Cov}(F_1, F_3)$*.*

**Theorem 2.4.** *For any two interval-valued random variables* $X_1(\omega) = [a_1(\omega), b_1(\omega)] = (c_1(\omega); r_1(\omega))$ *and* $X_2(\omega) = [a_2(\omega), b_2(\omega)] = (c_2(\omega); r_2(\omega))$*, where* $c_i(\omega) = (a_i(\omega) + b_i(\omega))/2$ *is the center and* $r_i(\omega) = (b_i(\omega) - a_i(\omega))/2$ *is the radius of* $X_i(\omega)$*,* $i = 1, 2$*, the following equalities hold:*

$$\begin{aligned} & \mathrm{Cov}(X_1(\omega), X_2(\omega)) \\ = \ & \mathrm{Cov}(a_1(\omega), a_2(\omega)) + \mathrm{Cov}(b_1(\omega), b_2(\omega)) \\ = \ & 2\mathrm{Cov}(c_1(\omega), c_2(\omega)) + 2\mathrm{Cov}(r_1(\omega), r_2(\omega)). \end{aligned}$$

**Theorem 2.5.** *The correlation coefficient* $\rho$ *of* $F_1, F_2 \in \mathcal{U}[\Omega, \boldsymbol{K}_{kc}(\mathcal{X})]$ *has the following properties:*

*(1)* $|\rho| \leq 1$*.*

*(2) If* $F_1$ *and* $F_2$ *are independent, then* $\rho = 0$*.*

*(3)* $\rho(F_1, F_2) = 1$ *if and only if* $F_2 + \lambda E(F_1) = E(F_2) + \lambda F_1$*, a.s.(P),* $\rho(F_1, F_2) = -1$ *if and only if* $F_2 + \lambda F_1 = E(F_2) + E(\lambda F_1)$*, a.s.(P), where* $\lambda = \sqrt{\mathrm{Var}(F_2)/\mathrm{Var}(F_1)}$*.*

**Remark 2.3.** For an interval-valued random variable $F \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$, denoted as $F(\omega) = [f(\omega), g(\omega)] = (c(\omega); r(\omega))$, where $f(\omega), g(\omega)$ are real-valued random variables and $f(\omega) \leq g(\omega)$, $c(\omega) = (f(\omega) + g(\omega))/2, r(\omega) = (g(\omega) - f(\omega))/2$, by the definition of Aumann integral and variance of set-valued random variables, we have

$$E(F(\omega)) = [E(f(\omega)), E(g(\omega))] = (E(c(\omega)); E(r(\omega)))$$

and

$$\begin{aligned} & \mathrm{Var}(\mathrm{F}(\omega)) \\ = \ & E(|f(\omega) - E(f)|^2) + E(|g(\omega) - E(g)|^2) \\ = \ & E(|c(\omega) - E(c) - (r(\omega) - E(r))|^2) \\ & + E(|c(\omega) - E(c) + (r(\omega) - E(r))|^2). \end{aligned}$$

For interval-valued random variables $F_1, F_2 \in \mathcal{U}[\Omega, \mathbf{K}_{kc}(\mathbb{R}^1)]$,

$$\begin{aligned} & \mathrm{Cov}(\mathrm{F}_1(\omega), \mathrm{F}_2(\omega)) \\ = \ & E(|f_1(\omega) - E(f_1)||f_2(\omega) - E(f_2)|) \\ & + E(|g_1(\omega) - E(g_1)||g_2(\omega) - E(g_2)|) \\ = \ & E(|c_1(\omega) - E(c_1) - (r_1(\omega) - E(r_1))| \\ & |c_2(\omega) - E(c_2) - (r_2(\omega) - E(r_2))|) \\ & + E(|c_1(\omega) - E(c_1) + (r_1(\omega) - E(r_1))| \\ & |c_2(\omega) - E(c_2) + (r_2(\omega) - E(r_2))|). \end{aligned}$$

## 3 Interval-Valued Linear Model and Least Square Estimation

In this section, we consider an interval-valued linear model with the following general form

$$E(y) = X\beta, \tag{1}$$

where $y = (y_1, y_2, \cdots, y_n)^T$ is an $n \times 1$ vector of interval-valued observations, $X = (x_{ij})_{i=1,j=1}^{n,p}$ is an $n \times p$ design matrix, $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a $p \times 1$ interval-valued parameter vector.

**Definition 3.1.** *If* $(y_i; x_{i1}, x_{i2}, \cdots, x_{ip})$*,* $i = 1, 2, \cdots, n$ *is a sample of interval-valued linear model (1), the least square estimator of unknown parameters* $\beta$ *is the estimator which minimizes* $d_2(y, X\beta)$*.*

By the definition of the $d_p$ metric, we have

$$\begin{aligned} & d_2^2(y, X\beta) \\ = \ & \sum_{i=1}^{n} d_2^2(y_i, x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots, + x_{ip}\beta_p) \end{aligned}$$

$$\begin{aligned} &= \sum_{i=1}^{n} \Big[(c_{y_i} - x_{i1}c_{\beta_1} - \cdots - x_{ip}c_{\beta_p}) \\ &\quad -(r_{y_i} - |x_{i1}|r_{\beta_1} - \cdots - |x_{ip}|r_{\beta_p})\Big]^2 \\ &\quad +\sum_{i=1}^{n} \Big[(c_{y_i} - x_{i1}c_{\beta_1} - \cdots - x_{ip}c_{\beta_p}) \\ &\quad +(r_{y_i} - |x_{i1}|r_{\beta_1} - \cdots - |x_{ip}|r_{\beta_p})\Big]^2 \\ &= 2\sum_{i=1}^{n} \Big[(c_{y_i} - x_{i1}c_{\beta_1} - \cdots - x_{ip}c_{\beta_p})^2 \\ &\quad +(r_{y_i} - |x_{i1}|r_{\beta_1} - \cdots - |x_{ip}|r_{\beta_p})^2\Big], \end{aligned}$$

where $c_A, r_A$ represent the center and radius of interval $A$, respectively. This is a quadratic function of $c_{\beta_1}, \cdots, c_{\beta_p}, r_{\beta_1}, \cdots, r_{\beta_p}$ and $d_2^2(y, X\beta) \geq 0$, so there exists a minimum value, which satisfies

$$\frac{\partial d_2^2(y, X\beta)}{\partial c_{\beta_j}} = 0, \ \frac{\partial d_2^2(y, X\beta)}{\partial r_{\beta_j}} = 0, \ j = 1, 2, \cdots, p,$$

that is

$$\begin{cases} \sum\limits_{i=1}^{n}(c_{y_i} - x_{i1}c_{\beta_1} - \cdots - x_{ip}c_{\beta_p})(-x_{ij}) = 0 \\ \sum\limits_{i=1}^{n}(r_{y_i} - |x_{i1}|r_{\beta_1} - \cdots - |x_{ip}|r_{\beta_p})(-x_{ij}) = 0, \end{cases}$$

$j = 1, 2, \cdots, p$. Rewriting these equations in matrix form, we get:

$$\begin{cases} X^T c_y = X^T X c_\beta \\ |X|^T r_y = |X|^T |X| r_\beta, \end{cases} \tag{2}$$

where $|X| = (|x_{ij}|)_{i=1,j=1}^{n,p}$.

From the above discussions, we have the following theorem.

**Theorem 3.1.** *If $rank(X) = rank(|X|) = p$, the least square estimator for the interval-valued linear model (1), denoted as $\hat{\beta}_{LS}$, is unique, and*

$$\hat{\beta}_{LS} = ((X^T X)^{-1} X^T c_y; (|X|^T |X|)^{-1} |X|^T r_y). \tag{3}$$

Furthermore, we can obtain the following theorems.

**Theorem 3.2.** *The LSE $\hat{\beta}_{LS}$ is an unbiased estimator of $\beta$.*

**Theorem 3.3.** *If $E(y) = X\beta$, $rank(X) = rank(|X|) = p$ and $\mathrm{Cov(c_y)} = \sigma_1^2 \mathrm{I_n}$, $\mathrm{Cov(r_y)} = \sigma_2^2 \mathrm{I_n}$, then the covariance matrix of $\hat{\beta}_{LS}$ is*

$$\mathrm{Cov}(\hat{\beta}_{\mathrm{LS}}) = 2\sigma_1^2 (\mathrm{X^T X})^{-1} + 2\sigma_2^2 (\mathrm{|X|^T |X|})^{-1}.$$

# 4 Best Linear Unbiased and Binary Linear Unbiased Estimation

## 4.1 Best Linear Unbiased Estimation

Given $n$ interval-valued data from the interval-valued linear model (1), $y_i = [a_{y_i}, b_{y_i}] = (c_{y_i}; r_{y_i}), i = 1, 2, \cdots, n$, the best linear unbiased estimator is a linear combination of $y_1, y_2, \cdots, y_n$

$$\hat{\beta}_j = \lambda_{j1} y_1 + \lambda_{j2} y_2 + \cdots + \lambda_{jn} y_n \doteq \lambda_j^T y, \tag{4}$$

$j = 1, 2, \cdots, p$, and the estimation is unbiased, that is,

$$E(\hat{\beta}_j) = \beta_j.$$

Assume $\beta_j = [a_{\beta_j}, b_{\beta_j}] = (c_{\beta_j}; r_{\beta_j})$. By (1) and (4), we have

$$\begin{aligned} E(\hat{\beta}_j) &= \lambda_j^T E(y) \\ &= \lambda_j^T (X c_\beta; |X| r_\beta) = (\lambda_j^T X c_\beta; |\lambda_j|^T |X| r_\beta), \end{aligned}$$

where $|\lambda_j| = (|\lambda_{j1}|, |\lambda_{j2}|, \cdots, |\lambda_{jn}|)^T$. Therefore we obtain

$$E(\hat{\beta}) = (\Lambda X c_\beta; |\Lambda||X| r_\beta), \tag{5}$$

where $\Lambda = \begin{pmatrix} \lambda_1^T \\ \lambda_2^T \\ \vdots \\ \lambda_p^T \end{pmatrix} = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pn} \end{pmatrix}$

and $|\Lambda| = \begin{pmatrix} |\lambda_{11}| & |\lambda_{12}| & \cdots & |\lambda_{1n}| \\ |\lambda_{21}| & |\lambda_{22}| & \cdots & |\lambda_{2n}| \\ \cdots & \cdots & \cdots & \cdots \\ |\lambda_{p1}| & |\lambda_{p2}| & \cdots & |\lambda_{pn}| \end{pmatrix}$.

On the other hand, since $\hat{\beta}$ is unbiased, we get

$$E(\hat{\beta}) = (c_{\beta_j}; r_{\beta_j}). \tag{6}$$

Therefore, by (5) and (6), we have

$$\Lambda X = I_p, \ \ |\Lambda||X| = I_p. \tag{7}$$

Unfortunately, the solution of (7) does not exist in general. For the case $p > 1$, consider the interval-valued linear regression model as an example:

$$E(y) = \beta_1 + \beta_2 X_2,$$

where $X_2 = (x_{12}, x_{22}, \cdots, x_{n2})$.

Let $\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1n} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2n} \end{pmatrix}$ and $X = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix}^T$, then the second equation of (7) is

$$\sum_{i=1}^{n} |\lambda_{1i}| = 1, \ \ \sum_{i=1}^{n} |\lambda_{1i}||x_{2i}| = 0,$$

$$\sum_{i=1}^{n} |\lambda_{2i}| = 0, \quad \sum_{i=1}^{n} |\lambda_{2i}||x_{2i}| = 1.$$

It is obvious that these equations are contradictory.

For the case $p = 1$, $E(y) = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} \beta_1$, then (7) becomes

$$\sum_{i=1}^{n} \lambda_{1i} x_{i1} = 1, \quad \sum_{i=1}^{n} |\lambda_{1i}||x_{i1}| = 0.$$

Therefore, a linear unbiased estimator exists if and only if $x_{i1} \geq 0, i = 1, 2, \cdots, n$.

### 4.2 Best Binary Linear Unbiased Estimation

From the above discussions, we know that, for the interval-valued linear model (1), the best linear unbiased estimation does not exist in general, which is a major difference with the traditional linear model. However, for the interval-valued linear model, we could introduce another notion: the binary best linear unbiased estimation, which has some interesting statistical properties.

**Definition 4.1.** *The binary linear combination of interval-valued data* $y_i = [a_{y_i}, b_{y_i}] = (c_{y_i}; r_{y_i}), i = 1, 2, \cdots, n$ *with coefficients* $k_i, l_i$ $(l_i \geq 0)$ *is defined as*

$$\sum_{i=1}^{n} (k_i c_{y_i}; l_i r_{y_i}) = \left( \sum_{i=1}^{n} k_i c_{y_i}; \sum_{i=1}^{n} l_i r_{y_i} \right).$$

**Definition 4.2.** *An estimator of an interval-valued parameter is called binary linear estimator, if it is a binary linear combination of interval-valued observations. Assume* $\hat{\theta}$ *is a binary linear estimator of interval-valued parameter* $\theta$*, if* $\hat{\theta}$ *is unbiased and for any binary linear unbiased estimator* $\theta^*$ *of* $\theta$*,*

$$\mathrm{Var}(\theta^*) \geq \mathrm{Var}(\hat{\theta}),$$

$\hat{\theta}$ *is called best binary linear unbiased estimator of* $\theta$*, denoted as BBLUE.*

If $\theta$ is a $p \times 1$ vector of interval-valued parameter, $\mathrm{Var}(\theta^*) \geq \mathrm{Var}(\hat{\theta})$ in this definition means that $\mathrm{Cov}(\theta^*) - \mathrm{Cov}(\hat{\theta})$ is a nonnegative definite matrix.

**Theorem 4.1.** *If* $E(y) = X\beta$*,* $rank(X) = rank(|X|) = p$ *and* $\mathrm{Cov(c_y)} = \sigma_1^2 \mathrm{I_n}$*,* $\mathrm{Cov(r_y)} = \sigma_2^2 \mathrm{I_n}$*, then the least square estimator* $\hat{\beta}_{LS}$ *is the unique BBLUE.*

**Theorem 4.2.** *If* $E(y) = X\beta$*,* $rank(X) = rank(|X|) = p$ *and* $\mathrm{Cov(c_y)} = \sigma_1^2 \mathrm{I_n}$*,* $\mathrm{Cov(r_y)} = \sigma_2^2 \mathrm{I_n}$*, then for for all* $\alpha \in \mathbb{R}^p$*,* $\alpha^T \hat{\beta}_{LS}$ *is the unique BBLUE of* $\alpha^T \beta$*.*

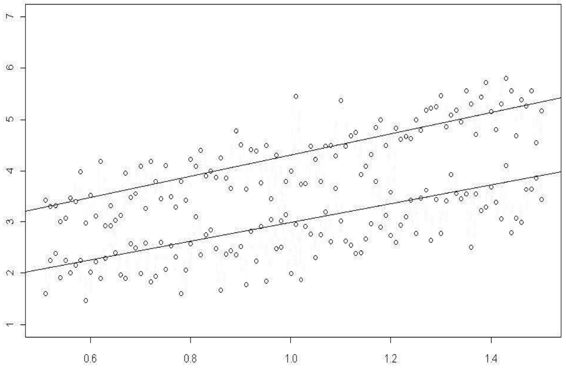

Figure 1: Points indicate 100 observations and the two lines represent the interval-valued linear regression function: $y = [1.06, 2.02] + [1.66, 2.32]x$.

## 5 Simulation Results

### 5.1 Test of Estimation Efficiency

In this section, we illustrate the interval-valued linear regression model by simulation. Let $\beta_1 = [1, 2] = (1.5; 0.5)$, $\beta_2 = [1.7, 2.3] = (2; 0.3)$ and

$$\begin{aligned} y_i &= \beta_1 + x_i \beta_2 + \varepsilon_i \\ &= (1.5 + 2x_i + c_{\varepsilon_i}; 0.5 + 0.3x_i + r_{\varepsilon_i}), \end{aligned}$$

$i = 1, 2, \cdots, n$, where $c_{\varepsilon_i}, r_{\varepsilon_i}$ are $N(0, 0.3^2)$ normal independent random variables, so that $E(y_i) = \beta_1 + E(x_i)\beta_2$. Therefore, we have

$$Ey = E \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = X \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}.$$

Firstly, we let the quantity of observations $n$ be 100, $x_i = 0.5 + 0.01i$, $i = 1, 2, \cdots, 100$. In one experiment, we get a least square estimator $\hat{\beta}_{LS}$ of $\beta_1, \beta_2$. Figure 1 shows the simulation experiment, in which $\hat{\beta}_{LS} = ([1.06, 2.02], [1.66, 2.32])^T$. In Figure 1, the points show the simulated data $y_i(x_i) = [1, 2] + [1.7, 2.3]x_i + \varepsilon_i$ $, x_i = 0.5 + 0.01i$, $i = 1, 2, \cdots, 100$ and the two lines represent the interval-valued linear regression function computed by LSE (3): $y = [1.06, 2.02] + [1.66, 2.32]x$.

We repeated this experiment 1000 times, average value of $\hat{\beta}_{LS}^{(1)}$ was $[0.9959131, 1.996367] = (1.49614; 0.5002269)$, with a sample mean square error (sample MSE) equal to 0.0442. The average value of 1000 $\hat{\beta}_{LS}^{(2)}$ was $[1.706118, 2.300196] =$

Table 1: Average value and sample MSE of $\hat{\beta}_{LS}^{(1)}$.

| | mean value of $\hat{\beta}_{LS}^{(1)}$ | sample MSE of $\hat{\beta}_{LS}^{(1)}$ |
|---|---|---|
| n=100 | [0.9959131,1.996367] | 0.0442 |
| n=200 | [1.002874,1.995194] | 0.0236 |
| n=300 | [1.002542,2.006844] | 0.0154 |

Table 2: Average value and sample MSE of $\hat{\beta}_{LS}^{(2)}$.

| | mean value of $\hat{\beta}_{LS}^{(2)}$ | sample MSE of $\hat{\beta}_{LS}^{(2)}$ |
|---|---|---|
| n=100 | [1.706118,2.300196] | 0.0446 |
| n=200 | [1.705211,2.299007] | 0.0220 |
| n=300 | [1.699598,2.295972] | 0.0142 |

$(2.003157; 0.297039)$ with a sample MSE is 0.0446. Here the sample mean square error of $\beta$ is defined by $\frac{1}{1000}\sum_{i=1}^{1000} d_2^2(\beta, \hat{\beta}_{LS})$.

Then we let the quantity of observations $n$ be 200 and 300. Regarding $X$, we let

$$x_i = 0.5 + 0.01i,\ i = 1, 2, \cdots, 100,$$

$$x_i = x_{i-100},\ i = 101, 102, \cdots, 200,$$

$$x_i = x_{i-200},\ i = 201, 202, \cdots, 300.$$

Similarly, we obtained estimators of $\hat{\beta}_{LS}^{(1)}, \hat{\beta}_{LS}^{(2)}$ by the same method. The results are presented in Tables 1 and 2, which give the average value and the sample MSE of 1000 estimators of $\hat{\beta}_{LS}^{(1)}$ (real value is $[1, 2]$) and $\hat{\beta}_{LS}^{(2)}$ (real value is $[1.7, 2.3]$) respectively. We can see that the sample MSE decreases as the number of observations increases.

### 5.2 Comparison with Other Models

When handling the point-valued input and interval-valued output data, an easy and intuitive solution is to fit the left- and right-endpoints (or the center and the radius) of the interval-valued data to two point-valued linear model, respectively (e.g., [5],[14] and [18]). As a matter of fact, it is easy to see these two methods are equivalent. As already mentioned in the introduction, a drawback of using two separate point-valued linear model is that it is possible to obtain an inter-valued estimation or forecast result such that the left-endpoint is larger than the right-endpoint (or the radius is negative). In this section, we present the advantage of our model from another view via a simulation experiment: comparing the efficiency of the forecast.

We generated the data in the same way as in Section 5.1 with $\beta_1 = [1, 2] = (1.5; 0.5)$, $\beta_2 = [1.7, 2.1] = (1.9; 0.2)$ and

$$y_i = \beta_1 + x_i\beta_2 + \varepsilon_i, \tag{8}$$

in which $x_i = (-3 : 0.05 : 6)$ and $c_{\varepsilon_i}$, $r_{\varepsilon_i}$ are $N(0, 0.1^2)$ independent random variables.

We then obtained the parameter estimation using the least square estimation for interval-valued linear model (3): $\hat{\beta}_{LS} = ([0.9979, 2.0062], [1.7017, 2.1000])^T$, and the regression function

$$y = [0.9979, 2.0062] + [1.7017, 2.1000]x. \tag{9}$$

In a second step, we fit $(a_{y_i}, x_i)$ and $(b_{y_i}, x_i)$, where $a_{y_i}$ and $b_{y_i}$ are the left- and right-endpoints of $y_i$, using two traditional point-valued linear models. Using the least square estimation for the traditional linear model, we obtain two fitted lines with equations:

$$\begin{cases} a_y = 0.6398 + 1.8061x \\ b_y = 2.3642 + 1.9956x. \end{cases} \tag{10}$$

Finally, we generated some new data from (8) and use (9) and (10) to forecast the output respectively. Letting $x_i = (-3 : 0.2 : 6)$, we put $x_i$ back to (8), we obtain the (real) interval-valued output $y_i, i = 1, 2, \cdots, 46$. Then, we substitute $x_i = (-3 : 0.2 : 6)$ back to (9) and (10) and obtain the forecasts of $y_i, i = 1, 2, \cdots, 46$ using the interval-valued LS estimation (denoted by $\tilde{y}_i$) and two endpoints point-valued LS estimation (denoted by $\hat{y}_i$), respectively. The MSE of $\tilde{y}_i$ was $\frac{1}{46}\sum_{n=1}^{46} d_2^w(\tilde{y}_i, y_i) = 0.0352$ and the MSE of $\hat{y}_i$ was $\frac{1}{46}\sum_{n=1}^{46} d_2^w(\hat{y}_i, y_i) = 0.1290$. The box plots in Figure 2 show the median, the 25th and 75th percentiles and the extreme data points of the 46 forecasts using interval-valued linear model and using two separate linear models. Since the data are randomly generated, the above procedure (from data generation to forecast) is repeated 30 times, so that mean values of the MSEs of the forecasts may be computed, which are 0.0388 (using the interval-valued LS estimation) and 0.1321 (using two endpoints point-valued LS estimation). Obviously, we can see that the interval-valued linear model is better in the sense that it has smaller forecasting error.

## 6 Application to Real Data

In this section, we use the interval-valued linear model to investigate the relationship between temperature and latitude. The data we gather are the highest and

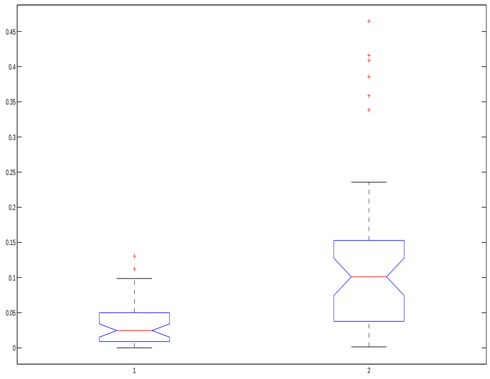

Figure 2: Box plots of forecasts results using interval-valued linear model (left) and left- and right-endpoints point-valued linear models (right).

Table 3: Temperatures and latitudes of 15 European cities on 14-th of August, 2012.

| City | Latitude (°) | Highest Temp. ($^\circ C$) | Lowest Temp. ($^\circ C$) |
|---|---|---|---|
| Athens | 38 | 24 | 34 |
| Madrid | 40.4 | 19 | 31 |
| Istanbul | 41 | 23 | 30 |
| Roma | 41.9 | 23 | 33 |
| Marsaille | 43.3 | 19 | 31 |
| Geneve | 46.25 | 13 | 28 |
| Paris | 48.8 | 19 | 26 |
| Brussel | 50.8 | 14 | 25 |
| London | 51.5 | 14 | 21 |
| Berlin | 52.5 | 13 | 23 |
| Moscow | 55.75 | 14 | 24 |
| Stockholm | 59.3 | 12 | 20 |
| St. Petersburg | 59.9 | 13 | 22 |
| Bergen | 60.4 | 14 | 20 |
| Reykjavik | 64 | 11 | 17 |

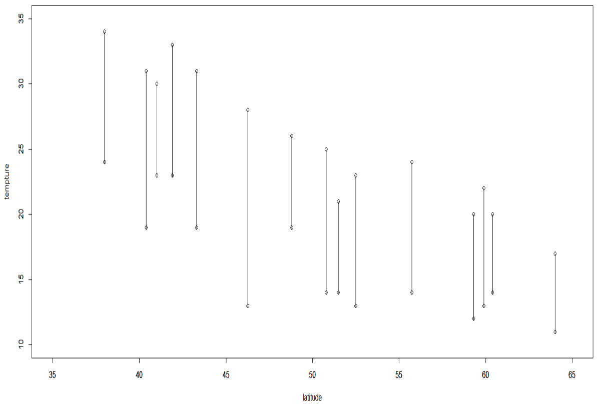

Figure 3: Temperatures (in the form of interval) of 15 European cities. Each line segment represents the temperature interval of a city.

the lowest temperatures of 15 cities in Europe on 14-th of August, 2012, as shown in Table 3 and Figure 3.

Suppose that temperature (interval-valued, $y$) and latitude (real-valued, $x$) follow the interval-valued linear model (1), that is

$$E(y_i) = \beta_1 + x_i\beta_2, i = 1, 2, \cdots, 15.$$

By least square estimation (3), which is also the best linear unbiased estimation by Theorem 4.1, we can get estimators of $\beta_1, \beta_2$. The linear relationship between temperature $y$ and latitude $x$ is

$$y = [39.03 - 0.45x, 56.01 - 0.60x],$$

which is also shown in Figure 4. From Figure 4, we can see that, as latitude increases the temperature decreases, and the daily difference in temperature also tends to decrease.

## 7 Conclusions

The linear model, which describes a random variable determined by a few variables and error in a linear way, plays an important role in statistics. However, in the real world, there are also a great deal of phenomena that are better described by an interval-valued random variable determined by a few real-valued random variables, e.g., temperature, stock price, service life of a kind of products. The relation between the interval-valued data and a few real-valued data can sometimes be expressed by a linear model. Therefore, we need a new type of statistical model to describe this kind of relation. In this paper, we introduced such a statistical model: the interval-valued linear

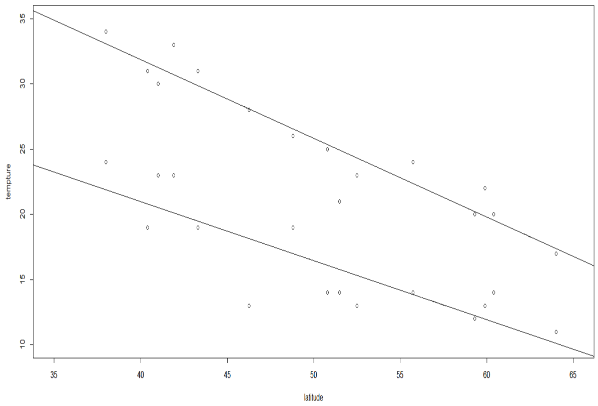

Figure 4: Data and linear relationship of temperature and latitude of 15 cities in Europe on 14-th of August, 2012. The two lines mean interval-valued linear regression function $y = [39.03196 - 0.451684x, 56.00954 - 0.6037982x]$.

model, which considers interval-valued observations determined by real-valued variables in a linear way.

Interval-valued random variables are a special kind of set-valued random variables, whose values are compact convex subsets of $\mathbb{R}^1$. In this paper, we investigated the theory in the general set-valued framework first, before focusing on the interval-valued random variables, in order to obtain some theoretical results in a wider range. In particular, we recalled the definition of variance and covariance of set-valued random variables based on the $d_p$ metric of sets and the $D_p$ metric of interval-valued random variables. We then introduced the interval-valued linear model and its least square estimation (LSE), proved the unbiasedness of the LSE and gave the covariance matrix of this estimator. We also showed that the best linear unbiased estimation does not exist in general, but the best binary linear unbiased estimation (BBLUE) exists and is unique, and the BBLUE is just the LSE. The performances of this estimator were illustrated using simulation experiments, and compared to those of the simple approach that consists in fitting two separate linear models using the endpoints of output intervals. The obtained results suggest that our approach yields better forecasting performance. Finally, we gave an example of the interval-valued linear model explaining how temperature is related by latitude. This short example shows how our model can be used and what type of practical problem can be solved using the interval-valued linear model.

## References

[1] Aubin, J. P. and H. Franbowska, Set-Valued Analysis, Birkhauser, 1990.

[2] Aumann, R., Integrals of set valued functions, J. Math. Anal. Appl., vol: 12, pp. 1-12, 1965.

[3] Beresteanu, A. and F. Molinari, Asymptotic properties for a class of partially identified models, Econometrica, vol: 76, pp. 763-814, 2008.

[4] Blanco, A., N. Corral, G. Gonzalez-Redriguez and M. A. Lubiano, Some properties of the $d_K$-variance for interval-valued sets, D. Dubois et al. (Eds.): Soft Methods for Hand. Var. and Imprecision, ASC 48, pp. 331-337, 2008.

[5] Blanco-Fernandez, A., N. Corral and G. Gonzalez-Redriguez, Estimation of a flexible simple linear model for interval data based on set arithmetic, Computational Statistics and Data Analysis, vol: 55, pp. 2568-2578, 2011.

[6] Blanco-Fernandez, A., A. Colubi and G. Gonzalez-Redriguez, Confidence sets in a linear regression model for interval data, Journal of Statistical Planning and Inference, vol: 142, pp. 1320-1329, 2012.

[7] Clarke, B. R., Linear Model: the Theory and Application of Analysis of Variance, Wiley, 2008.

[8] Denoeux, T. and M.-H. Masson, Multidimensional scaling of interval-valued dissimilarity data, Pattern Recognition Letters, 21: 83-92, 2000.

[9] Denoeux, T. and M.-H. Masson, Principal component analysis of fuzzy data using autoassociative neural networks, IEEE Transactions on Fuzzy Systems, 12 (3): 336-349, 2004

[10] Diamond, P. and P. Kloeden, Metric Space of Fuzzy Sets, World Scientific, 1994.

[11] Hiai, F. and H. Umegaki, Integrals, conditional expectations and martingales of multivalued functions, J. Multivar. Anal., vol: 7, pp. 149-182, 1977.

[12] Maia, A., F. Carvalho and T. B. Ludermir, Forecasting models for interval-valued time series, Neurocomputing vol: 71 pp. 3344-3352, 2008.

[13] Masson, M.-H. and T. Denoeux, Multidimensional scaling of fuzzy dissimilarity data, Fuzzy Sets and Systems, 128 (3): 339-352, 2002.

[14] Hsu, H.L. and B. Wu, Evaluating forecasting performance for interval data, Computers and Mathematics with Applications, vol: 56, pp. 2155-2163, 2008.

[15] Lai, T. L. and H. Xing, Statistical Model and Methods for Financial Markets, Springer, 2007.

[16] Li, S., Y. Ogura and V. Kreinovich, Limit Theorems and Applications of Set-Valuded and Fuzzy

Set-Valued Random Variables, Kluwer Academic Publishers (Now Springer), Dordrecht, 2002.

[17] Molchanov, I., Theory of Random Sets, Springer, 2005.

[18] Sinova, B., A. Colubi, M. A. Gil and G. Gonzalez-Rodriguez, Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric, Information Sciences, vol: 199, pp. 109-124, 2012.

[19] Tanaka, H. and H. Lee, Interval regression analysis by quadratic programming approach, IEEE Transactions on Fuzzy Systems, vol: 6, no. 4, 1998.

[20] Tseng, F., G. Tzeng, H. Wu and B. Yuan, Fuzzy ARIMA model for forecasting the foreign exchange market, Fuzzy Sets and Systems, vol: 118, pp. 9-19, 2001.

[21] Vital, R.A., $L_p$ metrics for compact, convex sets, Journal of Approximation Theory, vol: 45, issue 3, pp. 280-287, 1985.

[22] Wang, X. and S. Li, The interval autoregressive time series model, in the proceeding of IEEE-FUZZ International Conference, pp. 2528-2533, 2011.

[23] Wang, X. and S. Li, Stationary set-valued and interval-valued time series, preprint, 2011.

[24] Yang, X. and S. Li, The $D_p$-metric space of set-valued random variables and its application to covariances, International Journal of Innovative Computing, Information and Control, vol: 1, pp. 73-82, 2005.

[25] Yang, X, The $D_p$-metric space of set-valued random variables and its applications, Dissertation for Sciences Master's Degree, in May, 2005.

[26] Zhang, W., S. Li, Z. Wang and Y. Gao, Set-Valued Stochastic Processes, Science Publisher (in Chinese), 2007.