day.

The linear model is probably the simplest and most frequently-used statistical model. It describes a random output variable influenced by a few input variables and an error term in a linear way. In this paper, we consider the situation of interval-valued observations, i.e., the output 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 real line $\mathbb{R}^1$. Since we have at our disposal many results on the theory of set-valued random variables[18,19,29], this is a suitable framework to tackle the problem addressed in this paper. Until recently, however, there has been only a few works discussing 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[23] studied the metric for compact convex sets via the support functions. In 2005, Yang and Li[27], Yang[28] 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 Yang[28], the author also built a linear regression model with interval-valued regression coefficients. The underlying space in the above two papers is $\mathbb{R}^d$. In 2008, Blanco et al.[4] defined d_K-variance for interval-valued random variables with underlying space being $\mathbb{R}^1$, which is a special case of Yang and Li[27] and Yang[28].

Other authors studied interval-valued and set-valued statistical models. Tanaka and Lee[21] 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.[20] 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 LSE of the coefficients under the d_2 metric of intervals. Blanco-Fernandez et al.[6] studied the strong consistency and asymptotic distributions of the LSE. 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[24] introduced a new type of interval-valued time series (the interval autoregressive time series model) and gave the estimation method of parameters and forecast method based on the evaluation criteria. Wang and Li[25] investigated set-valued and interval-valued stationary time series, which is based on the definition of variance and covariance of set-valued and interval-valued random variables introduced in Yang and Li[27] and Yang[28].

In this paper, we start with the set-valued framework and consider interval-valued random variables as a special case. We then introduce the interval-valued linear model and its LSE, 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) has some drawbacks. One reason is that it is possible to obtain estimation or forecast results such that the left-endpoint is larger than the right-endpoint, 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.

This paper is a complete version of the results presented by the authors[26]. 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

LSE, prove the unbiasedness of this LSE and give the covariance matrix of this estimator. Section 4 shows that the best linear unbiased estimation does not exist in general, but the best binary linear unbiased estimation (BBLUE) exists, is unique and equal to the LSE. In Section 5, we present a simulation study to show the methodology, and illustrate the efficiency of estimation method 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.

2. Variance and Covariance of Set-Valued Random Variables

2.1. d_p Metric of Sets

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

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

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

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

and denote

$$A \oplus B = \mathrm{cl}\{a + b : a \in A, b \in B\}.$$

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

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

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

where $\mathscr{X}^*$ is the dual space of $\mathscr{X}$, i.e., the set of all bounded linear functionals on $\mathscr{X}$. For example, if $\mathscr{X} = \mathbb{R}^1$, $\mathscr{X}^* = \mathbb{R}^1$. Take an interval $[a, b]$ with $0 \leqslant a < b$, $x \in \mathbb{R}^1$, then the support function is $s(x, [a, b]) = \begin{cases} bx, & x \geqslant 0 \\ ax, & x < 0 \end{cases}$. The support function has the following properties:

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

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

For $1 \leqslant p < \infty$, take $A, B \in \mathbf{K}_{kc}(\mathscr{X})$. We define the metric d_p on $\mathbf{K}_{kc}(\mathscr{X})$[2,18,27] by

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

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

Remark 2.1. If $\mathscr{X} = \mathbb{R}^1$, then $\mathbf{K}_{kc}(\mathbb{R}^1) = \{[a, b] : -\infty < a \leqslant b < \infty\}$ is the family of all intervals on $\mathbb{R}^1$. If $A_1, A_2 \in \mathbf{K}_{kc}(\mathbb{R}^1)$ with $A_1 = [a_1, b_1] = (c_1; r_1)$, $A_2 = [a_2, b_2] = (c_2; r_2)$, 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

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

Theorem 2.1. [27] *$(\boldsymbol{K}_{kc}(\mathbb{R}^d), d_p)$ is a complete, separable metric space for each $p \in [1, \infty)$.*

2.2. D_p Metric Space of Set-Valued Random Variables

A set-valued mapping $F : \Omega \to \mathbf{K}(\mathscr{X})$ is called a set-valued random variable[11,18] if, for each open subset O of $\mathscr{X}$, $F^{-1}(O) \in \mathscr{A}$, 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 $\mathscr{U}[\Omega, \mathbf{K}_{kc}(\mathscr{X})]$ denote the family of set-valued random variables taking values in $\mathbf{K}_{kc}(\mathscr{X})$. The D_p metric[27] 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 $F_1, F_2 \in \mathscr{U}[\Omega, \mathbf{K}_k(\mathscr{X})]$.