Information Sciences 199 (2012) 109–124

Contents lists available at SciVerse ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Interval arithmetic-based simple linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric

Beatriz Sinova *, Ana Colubi, María Ángeles Gil, Gil González-Rodríguez

Departamento de Estadística, I.O. y D.M., Universidad de Oviedo, 33071 Oviedo, Spain

ARTICLE INFO

Article history:
Received 21 May 2010
Received in revised form 19 November 2011
Accepted 23 February 2012
Available online 3 March 2012

Keywords:
Generalized mid/spread metric
Interval arithmetic
Linear regression analysis
Rrandom interval-valued set
t-Vector function

ABSTRACT

The prediction of a response random interval-valued set from an explanatory one has been examined in previous developments. These developments have considered an interval arithmetic-based linear model between the random interval-valued sets and a least squares regression analysis. The least squares approach involves a generalized L_2-metric between interval data; this metric weights squared distances between data location (mid-points/centers) and squared distances between data imprecision (spread/radius). As a consequence, estimators of the parameters in the linear model depend on the choice of the weights in the metric. To investigate about a suitable choice of weighting in the generalized mid/spread metric, a theoretical conclusion is first obtained. Finally, the impact of varying the weights is discussed by considering a Monte Carlo simulation study.

1. Introduction

In investigating the relationship between random elements, regression analysis enables to seek for the causal effect of one (or several) random element(s) upon another. Regression techniques have long been relevant to many fields. Most of the regression methods assume that the involved random elements can be formalized as real-valued random variables.

However, there exists an important number of practical situations for which involved attributes do not take real but interval values. In Lubiano [32], Ferson et al. [14–16], Billard and Diday [3], D'Urso and Giordani [11], Kreinovich et al. [28], and Chuang [8] one can find many instances of the usual sources of interval-valued data. Among them, intermittent measurements, censoring, data binning, cyclical fluctuations, ranges, and so on. So, the statistical analysis of these data becomes especially interesting in real-life.

The problem of linear regression analysis with interval data has been studied from different perspectives and in different frameworks (see, for instance, Diamond [9], Lubiano [32], Billard and Diday [2], Gil et al. [18–20], Manski and Tamer [33], Montenegro [34], De Carvalho et al. [10], and Lima Neto and De Carvalho [30,31]).

A least squares approach for an interval arithmetic-based linear model has been recently carried out (see González-Rodríguez et al. [24,23], Gil et al. [21], Blanco et al. [4], and Blanco-Fernández et al. [6]). This approach involves essential and distinctive features, like the following ones:

- The approach is based on the usual interval arithmetic to formalize the linear relationship between the response and explanatory random elements. Consequently, this approach looks jointly at the location and the imprecision characterizing interval data, instead of treating them separately.

* Corresponding author. Fax: +34 985103354.
E-mail address: sinovabeatriz@uniovi.es (B. Sinova).

0020-0255/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.02.040

- The so-called *t*-vector function or mid/spread characterization of the nonempty compact intervals enables to identify interval data with certain $\mathbb{R}^2$-valued data. This identification allows us to induce a generalized metric between intervals, as well as the model in the probabilistic setting for interval-valued random elements and the associated relevant summary measures of its distribution.
- The least squares methodology based on the above-mentioned arithmetic and generalized metric.

Estimators of the involved parameters in the linear model have been obtained and analyzed under general conditions [24,23,21,4–6]. The estimators depend on the metric between interval data which is considered to formalize the least squares approach. This metric generalizes the well-known in [1] (see also Trutschnig et al. [35] for a related detailed discussion).

As it has been outlined by Gil et al. [20] and Montenegro [34] the mid-point/spread (equivalent) expression of this metric has been crucial in interpreting and determining estimators of the parameters of the linear regression problem (see Gil et al. [20]), and in performing tests under linear model assumption (cf. González-Rodríguez et al. [24,23], Blanco et al. [4]). Kulpa [29] indicated the interest of the mid-point/spread tandem in some other implications from interval arithmetic. In a different approach to the regression between interval-valued data, its importance has been also pointed out later by other authors (see, for instance, De Carvalho et al. [10] and Lima Neto and De Carvalho [30,31]).

In Section 2 of this paper preliminary concepts and results will be presented. Section 3 recalls the regression problem between two interval-valued random elements, and the interval arithmetic-based linear model along with the associated parameters' estimators. In Section 4, a theoretical search of a suitable choice of the metric on the basis of the mean square error of the estimators is first carried out. The conclusions from the theoretical development will be corroborated by an empirical sensitivity analysis of estimators in Monte Carlo simulations from relevant representative situations. Some concluding remarks and future directions will be finally commented.

2. Preliminaries

The analysis of interval-valued data requires an adequate framework so that statistical developments, and especially inferential ones, can be well formalized. The space of *interval values* for data to be considered in the study is the class $\mathcal{K}_c(\mathbb{R})$ of nonempty compact intervals.

Remark 2.1. It should be emphasized that interval data can overlap and, hence, grouped data just mean a particular case of interval ones. In general, no constraints (but the nonempty compactness) will be assumed on interval data. However, in case interval values of the explanatory random element are nested and share the center, then the mid of this random term will be a degenerate random variable. In this case, the interval arithmetic-based regression approach does not make real interest, and even the separate regression study for the mids would not make sense too.

After presenting a motivating example, this section aims to recall the required tools for regression problems which will be formulated and discussed in Section 3.

2.1. Motivating case-studies

The next real-life examples motivate the interest of the regression problem between interval-valued random elements. The random elements the interest of the problem is focused on, correspond to intrinsically interval-valued ones.

Example 2.1. Data in Fig. 1 have been supplied in 1997 by the Nephrology Unit of the Hospital Valle del Nalón in Langreo (Asturias, Spain). The 'scatter diagrams' correspond to the "range of systolic blood pressure over the same day", *X*, and the

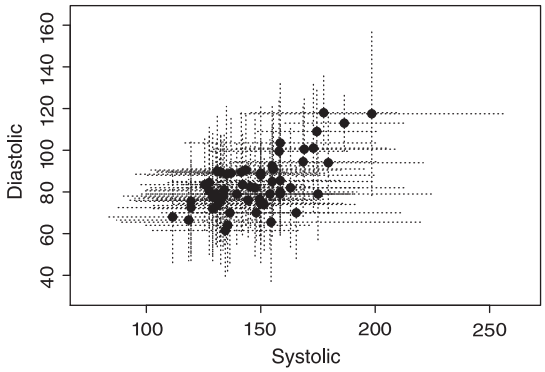

Fig. 1. Scatter diagrams of 'systolic *vs.* diastolic blood pressure'.