Information Sciences 247 (2013) 109–122

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables

Angela Blanco-Fernández *, Ana Colubi, Marta García-Bárzana

Department of Statistics, University of Oviedo, C/Calvo Sotelo s/n, Oviedo 33007, Spain

ARTICLE INFO

Article history:
Received 6 June 2012
Received in revised form 14 June 2013
Accepted 16 June 2013
Available online 25 June 2013

Keywords:
Linear regression
Set arithmetic
Interval data
Least-squares estimation

ABSTRACT

A new linear regression model for an interval-valued response and a real-valued explanatory variable is presented. The approach is based on the interval arithmetic. Comparisons with previous methods are discussed. The new linear model is theoretically analyzed and the regression parameters are estimated. Some properties of the regression estimators are investigated. Finally, the performance of the procedure is illustrated using both a real-life application and simulation studies.

1. Introduction

The scale of real intervals is appropriate for modelling several experimental data. Some examples are ranges of variation over a period of time (e.g. daily fluctuations of stock prices, temperatures and physical measurements), subjective perceptions or valuations, interval time sequences, and censored or grouped data. In D'Urso and Giordani [17], Ferson et al. [20], Gil et al. [21], Huber et al. [31], and Rho et al. [39] one can find many examples of the usual sources of experimental interval data. Moreover, interval data do not always represent experimental outcomes. In many cases intervals represent data in experimental scenarios which are imprecise or uncertain. In those situations, the experimental data are precise values. Yet, they are not measured precisely or they are hidden for confidentiality purpose. Therefore, only the intervals that contain the exact values are available [11,28,34,45]. Besides, certain kind of granular data can be represented by means of intervals [10]. Additionally, symbolic interval data usually summarize the information stored in large data sets. This results in a smaller and more manageable data set which preserves the essential information (see, for instance, Guo et al. [27] and references therein).

Linear regression problems with interval-valued data have been developed under different points of view. Generally, Symbolic Data Analysis considers the estimation of classical linear models which relate separately the midpoints (centers) and the spreads (radii) of the intervals [3,23,36]. No probabilistic assumptions are established for these models. Thus, the estimation process consists in a fitting problem and therefore, inferential studies on the models do not make sense. Moreover, the fitted parameters of the model for the spreads are not obtained analytically but numerically, so the study of statistical properties for them becomes difficult. An alternative approach to interval regression consists in relating the values of two (or more) intervals by means of an interval-valued function. If the relationship between the intervals is linear, the model can be easily defined in terms of the natural interval arithmetic [15,21,24,25,4,5]. Analogously to the classical linear

* Corresponding author. Tel.: +34 985 103126; fax: +34 985 103354.
E-mail address: blancoangela@uniovi.es (A. Blanco-Fernández).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2013.06.044

models for real-valued variables, the interval arithmetic-based models are formalized theoretically on a probabilistic scenario. Thus, it is possible to make inferences on the models, as proposed in Blanco-Fernández et al. [6] and Gil et al. [22].

The linear model proposed in this paper follows the latter approach of interval regression since it is formalized in terms of a set arithmetic. It has some distinctive features compared with the previous set arithmetic-based models: (i) it is designed to model an interval-valued response variable in terms of a real-valued regressor and (ii) the regression parameters are not real-valued but interval-valued.

The above features provide more flexibility in terms of modelling the whole interval values of the response through the real values of the explanatory variable. This is not the case for previous simple interval models. Ferraro et al. [19] developed a linear model for real-input and fuzzy-output. It could be applied to solve (i), since the interval-valued response can be considered as a particular case of a fuzzy random element. The advantage of the method presented here is that it does not involve any function to transform the spreads in variables taking values in $\mathbb{R}$. Further details are shown in Section 3.

When intervals represent imprecisely-measured (but crisp) values, there exists an alternative way to develop regression problems with interval (in general fuzzy-valued) data, by considering *possibilistic or fuzzy regression*. This approach was proposed first by Tanaka et al. [41]. It extends the imprecision and the vagueness to the system structure. Some additional papers in this alternative line are Tanaka et al. [42], Guo and Tanaka [26], Boukezzoula et al. [8], Huang [30] and Ĉerný et al. [9]. Yet, one should take into consideration that the aim of the possibilistic regression approaches compared to that of the regression models described above is completely different (for more details, see Coppi [13], Diamond [14], Hong et al. [29] and Näther [38]).

In econometric theory, there exists some work on partially identified models which includes the linear regression with interval data on the dependent variable (see, for instance, Beresteanu and Molinari [1], Bontemps et al. [7] and Manski and Tamer [37]). In this framework, intervals are considered to be censoring limits of a latent variable, which is unobservable in practice. Therefore, the regression parameter of the linear latent model is not a singleton anymore, but a set of parameters being consistent with the interval bounds. The preceding references address the estimation of this set of parameters, as well as asymptotic and inferential studies for this kind of models. These are established alternatives on the set identification literature. However, they tackle a different problem compared to the one considered in this paper. Specifically, the linear latent model is formalized as a classical model with real-valued variables and coefficients, which are partially observed and hence estimated through set-identification techniques. In this work the parameters and the response variable are interval-valued per se and our aim is to solve a classical estimation problem in a metric space. Our approach is not comparable with that of the previous literature, as the regression exhibits both a different statistical meaning and interpretation.

The rest of the paper is organized as follows: In Section 2 some preliminary concepts and results on the interval scenario are recalled. Several simple linear regression models for intervals that were previously investigated are revised. The limitations of those models when the regressor is real-valued are highlighted. The new proposed linear model is introduced in Section 3. Initially, the model is formalized on a given population, and then the main theoretical properties are shown. The least-squares (LS) estimation of the regression parameters is solved through a constrained minimization problem. In addition, some statistical properties of the regression estimators are investigated. The practical applicability and the empirical behavior of the estimation process are shown in Section 4, by means of a real-life example and some simulation studies. Finally, Section 5 includes the main concluding remarks and future directions.

2. Preliminary concepts and results

Interval data will be formalized as elements of

$$\mathcal{K}_c(\mathbb{R}) = \{[a,b] : a,b \in \mathbb{R}, a \leqslant b\}.$$

In terms of its endpoints, any interval $A \in \mathcal{K}_c(\mathbb{R})$ is parametrized by the two-dimensional vector $(\inf A, \sup A) \in \mathbb{R}^2$, being $\inf A \leqslant \sup A$. Alternatively, $(\mathrm{mid}\ A, \mathrm{spr}\ A) \in \mathbb{R} \times \mathbb{R}^+$, where $\mathrm{mid}\ A = (\sup A + \inf A)/2$ is the *midpoint* (or *center*), and $\mathrm{spr}\ A = (\sup A - \inf A)/2$ is the *spread* (or *radius*), also characterizes the interval A. The (*mid*, *spr*)–representation for intervals, $A = [\mathrm{mid}\ A \pm \mathrm{spr}\ A]$, is commonly employed due to the following reasons: (i) the non-negativity condition for the second component ($\mathrm{spr}\ A$) is more operative for computational developments than the order condition involved for the endpoints; (ii) it enables the embedding of the space $\mathcal{K}_c(\mathbb{R})$ into the subspace $\mathbb{R} \times \mathbb{R}^+$ of $\mathbb{R}^2$. This embedding would allow us to apply many classical properties on $\mathbb{R}^2$ in the statistical treatment of intervals. However, it will be always necessary to guarantee that the resulting elements remain in the subspace $\mathbb{R} \times \mathbb{R}^+$ (so they are associated with well-defined intervals); and (iii) it describes very intuitively the interval, since the *mid* – component express the location or position of the interval in the real line, and the *spr* – component provides information about the imprecision of the interval (in the sense of the difference with a precise quantity of $\mathbb{R}$).

The natural arithmetic on $\mathcal{K}_c(\mathbb{R})$ is defined by means of the Minkowski addition and the product by scalars, given by

$$A + B = \{a + b : a \in A, b \in B\} \quad \text{and} \quad \lambda A = \{\lambda a : a \in A\},$$

for all $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$, respectively. The operations can be jointly expressed in terms of the (mid, spr)-representation of the intervals as

$$A + \lambda B = [(\mathrm{mid}\ A + \lambda \mathrm{mid}\ B) \pm (\mathrm{spr}\ A + |\lambda| \mathrm{spr}\ B)].$$