Computational Statistics and Data Analysis 55 (2011) 2568–2578

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Estimation of a flexible simple linear model for interval data based on set arithmetic

Angela Blanco-Fernández [a,*], Norberto Corral [a], Gil González-Rodríguez [b]

[a] Dpto. de Estadística, I.O y D.M., Oviedo University, C/ Calvo Sotelo s/n, Oviedo 33007, Spain

[b] European Centre for Soft Computing, C/ Gonzalo Gutiérrez Quirós s/n, Mieres-Asturias 33600, Spain

ARTICLE INFO

Article history:
Received 28 April 2010
Received in revised form 3 March 2011
Accepted 11 March 2011
Available online 22 March 2011

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

ABSTRACT

The estimation of a simple linear regression model when both the independent and dependent variable are interval valued is addressed. The regression model is defined by using the interval arithmetic, it considers the possibility of interval-valued disturbances, and it is less restrictive than existing models. After the theoretical formalization, the least-squares (LS) estimation of the linear model with respect to a suitable distance in the space of intervals is developed. The LS approach leads to a constrained minimization problem that is solved analytically. The strong consistency of the obtained estimators is proven. The estimation procedure is reinforced by a real-life application and some simulation studies.

1. Introduction

Sometimes the exact value of a variable may be hidden for confidentiality reasons, or may be identified imprecisely due to uncertainty in its quantification. In such cases, one may know the interval that contains a representative data point instead of its precise value (see, for instance, Giordani and Kiers, 2006; Ham and Hsiao, 1984; Rivero and Valdes, 2008). On the other hand, it is also of interest to focus on characteristics which are *essentially* interval valued, like economical fluctuations, numerical ranges (in the sense of the range of variation (minimum, maximum) of a magnitude over the course of a certain period of time), subjective perceptions, grouped data, and so on (see Diamond, 1990; Gil et al., 2002, 2007).

When the data for the regressors and/or for the dependent variable in a regression model are intervals, the regression analysis becomes more complex (see Zhang and Sun, 2010; Hashimoto et al., 2011). Frequently, only certain representative values of the intervals, such as the midpoints (see, for instance, Billick et al., 1979) are considered, disregarding part of the information this way. These approaches produce in general inconsistent or inaccurate solutions (see Hasselblad et al., 1980). In particular, their application in the context of linear regression models is criticized in Hsiao (1983).

Interval data are often assumed only for the dependent variable, and in the context of grouped data (see, for instance, Chesher and Irish, 1987; Hong and Tamer, 2003; Steward, 1983). Manski and Tamer (2002) consider either an interval-valued regressor or outcome, but not both variables at the same time. When both the regressor and the dependent variable take interval values, the estimation of real linear models involving separately the centres (or midpoints) and semi-ranges (or spreads) has been suggested (see Lima Neto et al., 2008; Lima Neto and de Carvalho, 2010 and the references therein). However, these approaches are not always well-suited to the inferential scenario, because the models do not prevent the values of the dependent variable from being ill-defined as interval elements. Thus, although the problem is easily overcome in the fitting process, the estimates may refer to an ill-defined population model.

* Corresponding author. Tel.: +34 985 102958; fax: +34 985 103354.
E-mail addresses: blancoangela@uniovi.es (A. Blanco-Fernández), norbert@uniovi.es (N. Corral), gil.gonzalez@softcomputing.es (G. González-Rodríguez).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.03.005

Table 1
Data set on the range of variation of blood pressures of 59 patients.

X	Y	X	Y	X	Y
[11.8, 17.3]	[6.3, 10.2]	[11.9, 21.2]	[4.7, 9.3]	[9.8,16.0]	[4.7,10.8]
[10.4, 16.1]	[7.1, 10.8]	[12.2, 17.8]	[7.3, 10.5]	[9.7,15.4]	[6.0,10.7]
[13.1, 18.6]	[5.8, 11.3]	[12.7, 18.9]	[7.4, 12.5]	[8.7, 15.0]	[4.7, 8.6]
[10.5, 15.7]	[6.2, 11.8]	[11.3, 21.3]	[5.2, 11.2]	[14.1, 25.6]	[7.7, 15.8]
[12.0, 17.9]	[5.9, 9.4]	[14.1, 20.5]	[6.9, 13.3]	[10.8, 14.7]	[6.2, 10.7]
[10.1, 19.4]	[4.8, 11.6]	[9.9, 16.9]	[5.3, 10.9]	[11.5, 19.6]	[6.5, 11.7]
[10.9, 17.4]	[6.0, 11.9]	[12.6, 19.7]	[6.0, 9.8]	[9.9, 17.2]	[4.2, 8.6]
[12.8, 21.0]	[7.6, 12.5]	[9.9, 20.1]	[5.5, 12.1]	[11.3, 17.6]	[5.7, 9.5]
[9.4, 14.5]	[4.7, 10.4]	[8.8, 22.1]	[3.7, 9.4]	[11.4, 18.6]	[4.6, 10.3]
[14.8, 20.1]	[8.8, 13.0]	[11.3, 18.3]	[5.5, 8.5]	[14.5, 21.0]	[10.0, 13.6]
[11.1, 19.2]	[5.2, 9.6]	[9.4, 17.6]	[5.6, 12.1]	[12.0, 18.0]	[5.9, 9.0]
[11.6, 20.1]	[7.4, 13.3]	[10.2, 15.6]	[5.0, 9.4]	[10.0, 16.1]	[5.4, 10.4]
[10.2, 16.7]	[3.9, 8.4]	[10.3, 15.9]	[5.2, 9.5]	[15.9, 21.4]	[9.9, 12.7]
[10.4, 16.1]	[5.5, 9.8]	[10.2, 18.5]	[6.3, 11.8]	[13.8, 22.1]	[7.0, 11.8]
[10.6, 16.7]	[4.5, 9.5]	[11.1, 19.9]	[5.7, 11.3]	[8.7, 15.2]	[5.0, 9.5]
[11.2, 16.2]	[6.2, 11.6]	[13.0, 18.0]	[6.4, 12.1]	[12.0, 18.8]	[5.3, 10.5]
[13.6, 20.1]	[6.7, 12.2]	[10.3, 16.1]	[5.5, 9.7]	[9.5, 16.6]	[5.4, 10.0]
[9.0, 17.7]	[5.2, 10.4]	[12.5, 19.2]	[5.9,10.1]	[9.2,17.3]	[4.5, 10.7]
[11.6, 16.8]	[5.8, 10.9]	[9.7, 18.2]	[5.4,10.4]	[8.3,14.0]	[4.5, 9.1]
[9.8, 15.7]	[5.0, 11.1]	[12.7, 22.6]	[5.7, 10.1]		

Gil et al. (2001, 2002, 2007) consider an affine transformation, which relates random intervals, based on the natural interval arithmetic, and avoids ill-defined models. Nevertheless, González-Rodríguez et al. (2007) shows that, contrary to what happens in the case of real-valued random variables, the optimal solutions for the affine transformation model can produce estimates non-coherent for the regression parameters. This shortcoming has been overcome in the latter work by solving the least-squares (LS) problem over a suitable feasible set.

The model in González-Rodríguez et al. (2007) implies that the midpoints and the spreads of the variables are related by means of two linear functions in which the absolute value of the slope parameter is the same. There are some practical situations in which it is reasonable to assume that the midpoints and the spreads of the variables may increase/decrease with the same absolute ratio. However, in general, this consideration imposes some restrictions. The aim of this paper is to present and estimate a more flexible simple linear model between random interval variables.

1.1. Motivating example

To motivate the methodology, the analysis of the (linear) relationship between the fluctuation of the systolic and diastolic blood pressure for the patients in a hospital is considered. Data have been supplied by the Department of Nephrology of the Hospital *Valle del Nalón* in Asturias, Spain (see, for instance, Gil et al., 2002). Values of the blood pressures for each patient are measured at different points in a day. For some purposes, physicians focus on the range of variation (minimum-maximum) of these magnitudes which are registered for the patient on a concrete day. In these cases, devices used for measuring the pressure are programmed to record not the whole registers for a day, but only the lowest and highest measurements of pressure during the day. Therefore, the range of variation of each pressure over one day can be modelled by means of a real compact interval defined from the corresponding extreme values registered during that day.

Data in Table 1 correspond to X = "range of variation of the systolic blood pressure over a day" and Y = "range of variation of the diastolic blood pressure over the same day". The sample data consist of 59 patients of the Hospital *Valle del Nalón*, taken from a population of 3000 who are hospitalized per year in this hospital. Fig. 1 represents the scatter plot of both variables, by means of crosses. For each cross, the intersection represents the pair of midpoints of the corresponding intervals, and the length of the arms of the cross is the diameter of each interval.

The rest of the paper is organized as follows: Section 2 describes the simple linear model and some of its main features. In Section 3, the LS estimators of the linear model are obtained by finding the solutions to a constrained minimization problem. Both the theoretical and the empirical performance of the solutions are shown in Section 4. Specifically, the strong consistency of the estimators is proven and some illustrative simulation studies are performed. Moreover, the method is applied to the sample data presented in Section 1.1 in order to check its practical applicability in a real-life case study. Finally, Section 5 concludes and give some future directions.

2. A new interval regression model: the model M

The formalization of the linear model for an interval-valued regressor and an interval-valued-dependent variable is based on the (mid, spr)-representation of the intervals, i.e. $A = [\mathrm{mid}\, A \pm \mathrm{spr}\, A]$ for each nonempty real compact interval, with $\mathrm{spr}\, A \geq 0$. This formalization will be used throughout the paper. In general, for statistical developments, it is more suitable to work with this representation instead of the (inf, sup)-representation, $A = [\inf A, \sup A]$ with $\inf A \leq \sup A$, since the non-negativity condition for the spread is usually easier to handle than the order restriction of infima and suprema.