Computational Statistics (2020) 35:755–773
https://doi.org/10.1007/s00180-019-00910-1

ORIGINAL PAPER

Multiple linear regression models for random intervals: a set arithmetic approach

Marta García-Bárzana[1] · Ana Belén Ramos-Guajardo[2] · Ana Colubi[3] · Erricos J. Kontoghiorghes[4]

Received: 13 February 2019 / Accepted: 22 June 2019 / Published online: 27 June 2019

Abstract

Some regression models for analyzing relationships between random intervals (i.e., random variables taking intervals as outcomes) are presented. The proposed approaches are extensions of previous existing models and they account for cross relationships between midpoints and spreads (or radii) of the intervals in a unique equation based on the interval arithmetic. The estimation problem, which can be written as a constrained minimization problem, is theoretically analyzed and empirically tested. In addition, numerically stable general expressions of the estimators are provided. The main differences between the new and the existing methods are highlighted in a real-life application, where it is shown that the new model provides the most accurate results by preserving the coherency with the interval nature of the data.

Keywords Interval-valued data · Least-squares estimators · Linear modelling · Multiple regression · Set arithmetic

✉ Ana Belén Ramos-Guajardo
ramosana@uniovi.es

Marta García-Bárzana
martagb5@gmail.com

Ana Colubi
Ana.Colubi@wirtschaft.uni-giessen.de

Erricos J. Kontoghiorghes
e.kontoghiorghes@dcs.bbk.ac.uk

[1] Reasearch and Development Department, ArcelorMittal, Asturias, Spain

[2] Department of Statistics, Oviedo University, Oviedo, Spain

[3] Justus Leibig Univerity Giessen, Giessen, Germany

[4] School of Economics and Finance, Queen Mary University of London, London, UK

1 Introduction

The statistical treatment of interval-valued data has been extensively considered in the last years, as it appears in multiple experimental scenarios. Sometimes a real random variable is imprecisely observed, so that the experimental data are recorded as the real intervals which may contain the precise values of the variable in each individual (see, for instance, Jahanshahloo et al. 2008; Lauro and Palumbo 2005; Park et al. 2016). Censoring and grouping processes also produce intervals (see Boruvka and Cook 2015; Černý and Rada 2011; Yu et al. 2014; Zhang 2009, among others). Symbolic data analysis (SDA) considers intervals for summarizing information stored in large data sets, as in Billard and Diday (2000), Lima Neto and De Carvalho (2010), and Lima Neto and Dos Anjos (2015). Additionally, *essentially* interval experimental data can be obtained. This is the case of fluctuations, ranges of values (in the sense of the range of variation between a minimum and a maximum of a magnitude on a certain period of time) or subjective perceptions; some examples can be found in Diamond (1990), D'Urso and Giordani (2004), González-Rodríguez et al. (2007), Ramos-Guajardo et al. (2014) and Ramos-Guajardo and Grzegorzewski (2016). This work focuses on this latter approach and its aim is to develop new regression models for the random intervals, also called interval-valued variables, which are the random elements modelling the experiment on target.

Several alternatives have been previously proposed to face linear regression problems for interval-valued data. Separate models for the center points of the intervals (formally called midpoints) and their radii (formally called spreads) can be used, but in this case the non-negativity constraints satisfied by the spread variables preclude of treating the problem within the context of classical linear regression (D'Urso 2003; Lima Neto and De Carvalho 2010). Thus, although the usual fitting techniques are used, the associated inferences are no longer valid. In a different context, possibilistic regression models are considered Boukezzoula et al. (2011) and Černý and Rada (2011), when the intervals represent the imprecision in the measurement of real values, and this imprecision is transferred to the regression model and its estimators. Finally, the set arithmetic-based approach consists in the formalization of a linear relationship between random intervals associated with a given probability space in terms of the interval arithmetic. Thus, the estimators of such coefficients can be interpreted in the classical sense (Blanco-Fernández et al. 2011, 2013; Diamond 1990; González-Rodríguez et al. 2007).

The aim of this work is to extend the models above based on the interval arithmetic to a more general framework by proposing several multiple regression models. Thus, extensions for the simple linear regression models within the framework of the works in Blanco-Fernández et al. (2011) and González-Rodríguez et al. (2007) are twofold investigated. On one hand, whereas the previous regression models relate the response midpoints (respectively spreads) by means of the explanatory midpoints (respectively spreads), the new model is able to accommodate cross-relationships between midpoints and spreads in a unique equation based on the set arithmetic. As the model in Blanco-Fernández et al. (2011), the new one is based on the so-called *canonical decomposition* of the intervals. In addition, the model in Blanco-Fernández et al. (2011) is considered to be more flexible than previous works since it takes into account possible interval-valued disturbances and it is less restrictive than previous existing models whereas the new work is even more flexible since it allows to analyze new relationships between the involved variables.

On the other hand, due to the essential differences of the model in González-Rodríguez et al. (2007) and those based on the canonical decomposition, multiple regression models allowing several explanatory variables to model the response are formalized. The least-squares (LS) estimation problems associated with the proposed regression models are solved even when the multiple linear regression problem is more difficult to handle than the simple one as it happens in the real framework.

Some empirical results and a real-life application are presented in order to show the applicability and the differences among the proposed models. Specifically, the relationship between the daily fluctuations of the systolic and diastolic blood pressures and the pulse rate over a sample of patients in a hospital of Asturias, in the north of Spain, has been analyzed. This real-life example has been previously considered in Blanco-Fernández et al. (2011), Gil et al. (2007) and González-Rodríguez et al. (2007), and the dataset including the minima and maxima of the corresponding intervals can be directly downloaded from http://bellman.ciencias.uniovi.es/SMIRE/Hospital.html.

The rest of the paper is organized as follows: In Sect. 2 preliminaries concerning the interval framework are presented and some previous linear models for intervals are revised. Extensions of those linear models are introduced in Sect. 3. The least-squares estimation problem is analyzed and numerically stable expressions are derived. In Sect. 4 the empirical performance and the practical applicability of the models are shown and compared with existing techniques through some simulation studies and real-life examples. Section 5 includes some conclusions and future directions.

2 Preliminaries

The considered interval experimental data are elements belonging to the space $\mathcal{K}_c(\mathbb{R}) = \{[a_1, a_2] : a_1, a_2 \in \mathbb{R}, a_1 \leq a_2\}$. Each interval $A \in \mathcal{K}_c(\mathbb{R})$ can be parametrized in terms of its midpoint, $\text{mid}A = (\sup A + \inf A)/2$, and its spread, $\text{spr}A = (\sup A - \inf A)/2$. The notation $A = [\text{mid}A \pm \text{spr}A]$ will be used. An alternative representation for intervals is the so-called canonical decomposition, introduced in Blanco-Fernández et al. (2011), given by $A = \text{mid}A[1 \pm 0] + \text{spr}A[0 \pm 1]$. It allows the consideration of the *mid* and *spr* components of A separately within the interval arithmetic. The Minkowski addition and the product by scalars constitute the natural arithmetic on $\mathcal{K}_c(\mathbb{R})$. In terms of the (mid, spr)-representation these operations can be jointly expressed as

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

for any $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. The space $(\mathcal{K}_c(\mathbb{R}), +, \cdot p)$ is not linear but semilinear (or conical), due to the lack of symmetric element with respect to the addition. If C verifying that $A = B + C$ exists, then C is called the Hukahara difference $(A -_H B)$ between the pair of intervals A and B. The interval C exists iff $\text{spr}B \leq \text{spr}A$ (see Blanco-Fernández et al. 2011, for details).

For every $A, B \in \mathcal{K}_c(\mathbb{R})$, an L_2-type generic metric has been introduced in Trutschnig et al. (2009) as $d_\theta(A, B) = ((\text{mid}A - \text{mid}B)^2 + \theta\,(\text{spr}A - \text{spr}B)^2)^{\frac{1}{2}}$, for an arbitrary $\theta \in (0, \infty)$ which measures the importance given to the spreads in relation to the one given to the midpoints. The value $\theta = 1/3$ is often considered as the natural election, because it corresponds to compute and weigh uniformly all the differences between the points of the intervals.

Given a probability space $(\Omega, \mathcal{A}, P)$, the mapping $\boldsymbol{x}\colon \Omega \to \mathcal{K}_c(\mathbb{R})$ is a random interval (or an interval-valued variable) iff $\text{mid}\,\boldsymbol{x}, \text{spr}\,\boldsymbol{x}\colon \Omega \to \mathbb{R}$ are real random variables and $\text{spr}\,\boldsymbol{x} \geq 0$. Random intervals will be denoted with bold lowercase letters, $\boldsymbol{x}$, random interval-valued vectors by non-bold lowercase letters, x, and interval-valued matrices with uppercase letters, X.

The expected value of $\boldsymbol{x}$ is defined in terms of the well-known Aumann expectation for intervals. It can be expressed as $E(\boldsymbol{x}) = [E(\text{mid}\boldsymbol{x}) \pm E(\text{spr}\boldsymbol{x})]$. It exists and $E(\boldsymbol{x}) \in \mathcal{K}_c(\mathbb{R})$ iff $\text{mid}\boldsymbol{x}$ and $\text{spr}\boldsymbol{x} \in L^1(\Omega, \mathcal{A}, P)$. The variance of $\boldsymbol{x}$ can be defined as the usual Fréchet variance (Näther 1997) associated with the Aumann expectation in the metric space $(\mathcal{K}_c(\mathbb{R}), d_\theta)$, i.e. $\sigma^2_{\boldsymbol{x}} = E(d^2_\theta(\boldsymbol{x}, E(\boldsymbol{x})))$, whenever $\text{mid}\boldsymbol{x}$ and $\text{spr}\boldsymbol{x} \in L^2(\Omega, \mathcal{A}, P)$. However, the conical structure of the space $\mathcal{K}_c(\mathbb{R})$ entails some differences while trying to define the usual covariance (Körner 1997). In terms of the d_θ-metric it has the expression $\sigma_{\boldsymbol{x},\boldsymbol{y}} = \sigma_{\text{mid}\,\boldsymbol{x},\text{mid}\boldsymbol{y}} + \theta\sigma_{\text{spr}\,\boldsymbol{x},\text{spr}\boldsymbol{y}}$, whenever those classical covariances exist. The expression $\text{Cov}(x, y)$ denotes the covariance matrix between two random interval-valued vectors $x = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ and $y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_k)$.

Several linear regression models for intervals based on the set arithmetic have been previously considered. They are briefly recalled and a comparison study with the new approach is addressed in Sect. 4. The basic simple linear model proposed in González-Rodríguez et al. (2007) is formalized as $\boldsymbol{y} = b\boldsymbol{x} + \boldsymbol{\varepsilon}$ with $b \in \mathbb{R}$ and $\boldsymbol{\varepsilon} : \Omega \to \mathcal{K}_c(\mathbb{R})$ being an interval-valued random error such that $E[\boldsymbol{\varepsilon}|\boldsymbol{x}] = \Delta \in \mathcal{K}_c(\mathbb{R})$. It only involves one regression parameter to model the dependency between the variables and thus, it induces quite restrictive separate models for the *mid* and *spr* components of the intervals. Namely, $\text{mid}\,\boldsymbol{y} = b\text{mid}\,\boldsymbol{x} + \text{mid}\,\boldsymbol{\varepsilon}$ and $\text{spr}\boldsymbol{y} = |b|\text{spr}\,\boldsymbol{x} + \text{spr}\,\boldsymbol{\varepsilon}$.

A more flexible linear model, called model M, has been introduced in Blanco-Fernández et al. (2011). It is defined in terms of the *canonical decomposition* as follows:

$$\boldsymbol{y} = b_1 \text{mid}\,\boldsymbol{x}\,[1 \pm 0] + b_2 \text{spr}\,\boldsymbol{x}\,[0 \pm 1] + \gamma\,[1 \pm 0] + \boldsymbol{\varepsilon}, \tag{1}$$

where $b_1, b_2 \in \mathbb{R}$ are the regression coefficients, $\gamma \in \mathbb{R}$ is an intercept term influencing the *mid* component of $\boldsymbol{y}$ and $\boldsymbol{\varepsilon}$ is a random interval error satisfying that $E[\boldsymbol{\varepsilon}|\boldsymbol{x}] = [-\delta, \delta]$, with $\delta \geq 0$. From (1) the linear relationships $\text{mid}\,\boldsymbol{y} = b_1\text{mid}\,\boldsymbol{x} + \gamma + \text{mid}\,\boldsymbol{\varepsilon}$ and $\text{spr}\boldsymbol{y} = |b_2|\text{spr}\,\boldsymbol{x} + \text{spr}\,\boldsymbol{\varepsilon}$ are transferred, where b_1 and b_2 may be different. The least-squares estimation leads to analytic expressions of the regression parameters of model M (see Blanco-Fernández et al. 2011). Specifically, the expressions for the regression estimators of model M are the following ones:

$$\widehat{b_1} = \frac{\widehat{\sigma}_{\text{mid}\,\boldsymbol{x},\text{mid}\,\boldsymbol{y}}}{\widehat{\sigma}^2_{\text{mid}\,\boldsymbol{x}}} \quad \text{and} \quad \widehat{b_2} = \min\left\{\widehat{s_0}, \max\left\{0, \frac{\widehat{\sigma}_{\text{spr}\,\boldsymbol{x},\text{spr}\,\boldsymbol{y}}}{\widehat{\sigma}^2_{\text{spr}\,\boldsymbol{x}}}\right\}\right\},$$

where $\widehat{s_0} = \min\{\operatorname{spr} \boldsymbol{y}_i / \operatorname{spr} \boldsymbol{x}_i \colon \operatorname{spr} \boldsymbol{x}_i \neq 0\}$. In addition, the strong consistency of the previous estimators is also proven in that work.

Given a sample data set of intervals it is also possible to fit the separate models for the *mid* and the *spr* components, as previously proposed in Lima Neto and De Carvalho (2010) and references therein. Alternatively, D'Urso (2003) presents several linear regression models for the so-called LR fuzzy numbers and therefore also for the particular case of intervals. In this case, possible cross-relationships between midpoints and spreads of the intervals are considered. It is important to observe that these approaches are different from the set arithmetic-based one from the statistical basis. They are considered from a descriptive point of view, since no probabilistic assumptions on the random intervals are established. Thus, it may be infeasible to study statistical properties of the estimators and inferential studies in this setting. For instance, since the independence or the uncorrelation of the regressor and the error term are not guaranteed, a problem of model identification may appear. As a conclusion, although the proposed estimation for these separate models offers an alternative to find a linear fitting on the available data set of intervals, the solutions to these problems cannot be identified with those of the theoretical linear models based on interval arithmetic.

3 A multiple flexible linear model: model $\mathrm{M_G}$

A novel multiple linear regression model for intervals is presented. It arises as a natural extension of the model M developed in Blanco-Fernández et al. (2011) both into the multiple case and into a more flexible scenario.

3.1 Population model

Let $\boldsymbol{y}$ be a response random interval and let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k$ be k explanatory random intervals. It is assumed that the real-valued random variables mid and spread associated with all the random intervals are not degenerated, the considered random intervals have finite and strictly positive variance and the var–cov matrix of the explanatory variables is invertible. The set arithmetic-based multiple flexible linear regression model, denoted by M_G, is formalized as follows:

$$\boldsymbol{y} = [(b_1 \operatorname{mid} x^t + b_4 \operatorname{spr} x^t) \pm (b_2 \operatorname{spr} x^t + b_3 \, |\operatorname{mid} x^t|)] + \boldsymbol{\varepsilon} \tag{2}$$

where $b_1, b_4 \in \mathbb{R}^k$, $b_2, b_3 \in \mathbb{R}^{k^+}$, $\operatorname{mid} x = (\operatorname{mid} \boldsymbol{x}_1, \operatorname{mid} \boldsymbol{x}_2, \ldots, \operatorname{mid} \boldsymbol{x}_k)^t \in \mathbb{R}^k$ (analogously $\operatorname{spr} x$), and $\boldsymbol{\varepsilon}$ is a random interval-valued error such that $E(\boldsymbol{\varepsilon}|x) = \Delta \in \mathcal{K}_c(\mathbb{R})$. From this condition, it is straightforward to see that x an ε are uncorrelated, i.e. $\sigma_{\varepsilon, x_i} = 0$, for all $i = 1, \ldots, k$. The separate linear relationships for the *mid* and *spr* components of the intervals transferred from (2) are

$$\operatorname{mid} \boldsymbol{y} = \operatorname{mid} x^t \, b_1 + \operatorname{spr} x^t \, b_4 + \operatorname{mid} \boldsymbol{\varepsilon}, \tag{3a}$$

$$\operatorname{spr} \boldsymbol{y} = \operatorname{spr} x^t \, b_2 + |\operatorname{mid} x^t| \, b_3 + \operatorname{spr} \boldsymbol{\varepsilon}. \tag{3b}$$

Thus, both variables mid $\boldsymbol{y}$ and spr $\boldsymbol{y}$ are modelled from the complete information provided by the independent random intervals in x, characterized by the random vector (mid$x$, spr$x$). An immediate conclusion from this property is that model M_G allows more flexibility on the possible linear relationship between the random intervals than the preceding set arithmetic-based models. However, the inclusion of more coefficients increases the difficulty of the estimation process, as happens in classical regression problems.

For a simpler notation, let us define the intervals $x^M = [\mathrm{mid}\, x^t, \mathrm{mid}\, x^t]$, $x^S = [-\,\mathrm{spr}\, x^t, \mathrm{spr}\, x^t]$, $x^C = [-\,|\mathrm{mid}\, x^t|, |\mathrm{mid}\, x^t|]$ and $x^R = [\mathrm{spr}\, x^t, \mathrm{spr}\, x^t]$. Thus, the model M_G is equivalently expressed in matrix notation as

$$\boldsymbol{y} = X^{Bl}\, B + \boldsymbol{\varepsilon}, \tag{4}$$

where $X^{Bl} = (x^M|x^S|x^C|x^R) \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})^{1\times 4k}$ and $B = (b_1|b_2|b_3|b_4)^t$. The associated regression function is $E(\boldsymbol{y}|\boldsymbol{x}_1 = x_1, \ldots, \boldsymbol{x}_k = x_k) = X^{Bl}\, B + \Delta$.

Let $\{(\boldsymbol{y}_j, \boldsymbol{x}_{1,j}, \ldots, \boldsymbol{x}_{k,j})\}_{j=1}^n$ be a simple random sample obtained from the random intervals $(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$. Then,

$$y = X^{eBl} B + \varepsilon,$$

where $y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^t$, $X^{eBl} = (X^M|X^S|X^C|X^R) \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})^{n\times 4k}$, B as in (4) and $\varepsilon = (\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_n)^t$ is such that $E(\varepsilon|x) = 1^n \Delta$. X^M is the $(n \times k)$-interval-valued matrix such that $(X^M)_{j,i} = [\mathrm{mid}\, \boldsymbol{x}_{i,j}, \mathrm{mid}\, \boldsymbol{x}_{i,j}]$ (analogously X^S, X^C and X^R).

3.2 Least squares estimation of the model

The least squares estimation (LS estimation, for short) searches for $\widehat{B}$ and $\widehat{\Delta}$ minimizing $d_\theta^2(y, X^{eBl} A + 1^n C)$ for $A \in \mathbb{R}^{4k\times 1}$, $C \in \mathcal{K}_c(\mathbb{R})$ and guaranteeing the existence of the residuals $\varepsilon = y -_H X^{eBl} A$. It is easy to see that $\mathrm{spr}(X^{eBl} A) = \mathrm{spr}\, X\, a_2 + |\mathrm{mid}\, X|\, a_3$, (with $(\mathrm{mid}\, X)_{j,i} = \mathrm{mid}\, \boldsymbol{x}_{i,j}$, and analogously spr X) so that the following conditions are to be included in the minimization problem:

$$\mathrm{spr}\, X\, a_2 + |\mathrm{mid}\, X|\, a_3 \le \mathrm{spr}\, y. \tag{5}$$

Analogously to what happens in classical regression, the estimate of the (interval-valued) intercept term Δ can be firstly obtained. If $\widehat{B}$ verifies (5), then the minimum value of $d_\theta^2(y, X^{eBl}\widehat{B} + 1^n C)$ over $C \in \mathcal{K}_c(\mathbb{R})$ is attained at

$$\widehat{\Delta} = \overline{\boldsymbol{y}} -_H \overline{X^{eBl}} \widehat{B}\ . \tag{6}$$

As a result, the LS estimate of the regression parameter B is obtained by minimizing

$$\begin{gathered} d_\theta^2(y -_H X^{eBl} A, \overline{\boldsymbol{y}} -_H \overline{X^{eBl}} A) \\ \text{subject to} \\ \mathrm{spr}\, X\, a_2 + |\mathrm{mid}\, X|\, a_3 \le \mathrm{spr}\, y \end{gathered} \tag{7}$$

with $A = (a_1|a_2|a_3|a_4)$ such that $a_1, a_4 \in \mathbb{R}^k$ and $a_2, a_3 \in \mathbb{R}^{k^+}$.

Proposition 1 *The least-squares estimators of the pairs of regression parameters* (b_1, b_4) *and* (b_2, b_3) *in* (2) *are*

$$(\widehat{b}_1, \widehat{b}_4) = (F_m^t\, F_m)^{-1} F_m^t\, v_m$$

and

$$(\widehat{b}_2, \widehat{b}_3) = (F_s^t F_s)^{-1}\, (F_s^t\, v_s - D^t\, \lambda),$$

respectively, where $v_m = \text{mid}y - \overline{\text{mid}\,\boldsymbol{y}}1^n \in \mathbb{R}^n$, $v_s = \text{spr}y - \overline{\text{spr}\,\boldsymbol{y}}1^n \in \mathbb{R}^n$, $F_m = \text{mid}X^{eBl} - 1^n\overline{(\text{mid}X^{eBl})} \in \mathbb{R}^{n\times 2k}$, $F_s = \text{spr}X^{eBl} - 1^n\overline{(\text{spr}X^{eBl})} \in \mathbb{R}^{n\times 2k}$, $\text{mid}X^{eBl} = (\text{mid}X, \text{spr}X) \in \mathbb{R}^{n\times 2k}$, $\text{spr}X^{eBl} = (\text{spr}X, |\text{mid}X|) \in \mathbb{R}^{n\times 2k}$, $D = \Big(-I_{2k}\ ,\ sprX^{eBl}\Big)^t \in \mathbb{R}^{(2k+n)\times 2k}$ *and* $\lambda \in \mathbb{R}^{2k+n}$ *is the vector minimizing the Linear Complementary problem in* (9).

Proof The problem (7) is solved by transforming it to an equivalent quadratic optimization problem, as follows:

$$\min_{A_m\in\,\mathbb{R}^{2k},\ A_s\in\,\Gamma} \|v_m - F_m A_m\|^2 + \theta\,\|v_s - F_s A_s\|^2$$
$$\Gamma = \{A_s \in \mathbb{R}^{2k} : D\,A_s \le d\} \tag{8}$$

being $A_m = (a_1|a_4)^t \in \mathbb{R}^{2k\times 1}$, $A_s = (a_2|a_3)^t \in \mathbb{R}^{2k\times 1}$ and $d = \Big(0_{2k}\ ,\ \text{spr}y\Big)^t \in \mathbb{R}^{(2k+n)\times 1}$.

This problem can be solved separately for A_m and A_s. On one hand, $(\widehat{b}_1, \widehat{b}_4)$ derives directly from the minimization of the unconstrained quadratic form $\|v_m - F_m A_m\|^2$ for $A_m \in \mathbb{R}^{2k}$. On the other hand, the minimization problem $\|v_s - F_s A_s\|^2$ over $A_s \in \Gamma$ admits the following equivalent formulation

$$\min \frac{1}{2} A_s^t\, H\, A_s - c^t\, A_s$$
$$s.t. \quad DA_s \le d$$

being $H = F_s^t F_s \in \mathbb{R}^{2k\times 2k}$ and $c = F_s^t\, v_s \in \mathbb{R}^{2k\times 1}$. This problem has the structure of a *linear complementary problem (LCP)*, since it can be expressed as

$$\omega = M\,\lambda + q$$
$$s.t.\omega, \lambda \ge 0\ ,\ \omega_j\lambda_j = 0\ ,\ j = 1, \ldots, n+1\ , \tag{9}$$

with $M = D\,H^{-1}\,D^t$ and $q = d - D\,H^{-1}\,c$. Lemke's or Dantzig–Cottle's algorithms can be used to obtain by an iterative process the value λ minimizing the LCP (see Lemke 1962; Liew 1976 for further details). Once λ is computed, the close form of the solution in (8) is $(\widehat{b}_2, \widehat{b}_3) = H^{-1}\,(c - D^t\,\lambda)$. □

The previous estimator based on Lemke's method has a computational complexity of an $O((2k+n)^3) \sim O(n^3)$, as the number of observations n is usually much greater than the number of variables k.

Moreover, observe that $(\widehat{b}_1, \widehat{b}_4)$ coincides with the *ordinary least squares* (OLS) estimator of the classical multiple regression model (3a). Therefore, it is guaranteed that it is an unbiased, consistent and efficient estimator of the vector of regression coefficients (b_1, b_4), i.e. $E(\widehat{b}_1, \widehat{b}_4) = (b_1, b_4)$, $(\widehat{b}_1, \widehat{b}_4) \overset{n\to\infty}{\longrightarrow} (b_1, b_4)$, and $Var(\widehat{b}_1, \widehat{b}_4) \overset{n\to\infty}{\longrightarrow} 0$. Besides, the analytic expression of its standard error is

$$se(\widehat{b}_1, \widehat{b}_4) = \left(\sqrt{\sigma^2 (F_m^t F_m)^{-1}_{11}}, \sqrt{\sigma^2 (F_m^t F_m)^{-1}_{22}}\right).$$

The result is immediate from the Gauss–Markov Theorem (Johnston 1972). The availability of a closed form of the estimator $(\widehat{b}_2, \widehat{b}_3)$ greatly benefits the development of further statistical studies on the linear model, as inferences, linear independence, etc. Nonetheless, as the computation of λ is done in an iterative way this entails some computational costs difficulties to develop inferences for $(\widehat{b}_2, \widehat{b}_3)$. Therefore, an alternative estimator for the constrained parameters (b_2, b_3), called minimum-distance estimator, $\widehat{b}^{min}$, is introduced. It is computed by following Algorithm 1:

Algorithm 1: The minimum distance estimator $\widehat{b}^{min}$

1. Compute the Ordinary Least Squares estimate as in Proposition 3.1, i.e.,

$$\widehat{b}^{OLS} = (F_s^t\, F_s)^{-1} F_s^t\, v_s$$

2. In order to fulfill the non-negativity constraints, compute the non-negative OLS

$$\widehat{b}^* = \max\left\{0, \widehat{b}^{OLS}\right\}$$

3. Find the vector $\widehat{b}^{min} = (\widehat{b}_1^{min}, \widehat{b}_2^{min})$ that minimizes

$$\underset{b\in\Gamma}{\operatorname{argmin}} \|\widehat{b}^* - b\|^2$$

The goal is to find the closest point within the feasible region to the non-negative OLS, $\widehat{b}^* = (\widehat{b}_2^*, \widehat{b}_3^*)$.

The procedure to solve the minimization problem in Step 3 of Algorithm 1 is open. Nonetheless, since the goal is to minimize a norm subject to some inequality constraints, it is possible to use quadratic programming procedures, applying KKT conditions and using a numerical method. The numerical method applied here is the Matlab 2015b code lsqlin.m, based on an active-set method, and ensuring the finding of a global minimum in a finite number of steps. This two-step algorithm consists of an initial phase devoted to compute a feasible point (if it exists) and a second phase generating an iterative sequence of feasible points that converge to the solution (Gillis 2012). Besides, the computational complexity is known to be polynomial to the number of variables and the number of observations as an $O(n\,(2k)^3)$, which compared to the computational complexity of the previous estimator in Proposition 1, $O(n^3)$, is much lower whenever the number of observations is greater than the number of variables.

Theorem 1 *Under the conditions of Model 3.3,* $(\widehat{b}_2^{min}, \widehat{b}_3^{min})$ *is a consistent estimator, i.e.,* $(\widehat{b}_2^{min}, \widehat{b}_3^{min}) \overset{n\to\infty}{\longrightarrow} (b_2, b_3)$ *a.s.-[P].*

Proof Since $(b_2, b_3) \in \Gamma$, by definition, the distance between $(\widehat{b}_2^{*}, \widehat{b}_3^{*})$ with (b_2, b_3) is always greater than or equal to the distance between $(\widehat{b}_2^{*}, \widehat{b}_3^{*})$ with its closest point within such region, i.e., $(\widehat{b}_2^{min}, \widehat{b}_3^{min})$. That is,

$$\|(\widehat{b}_2^{*}, \widehat{b}_3^{*}) - (\widehat{b}_2^{min}, \widehat{b}_3^{min})\|^2 \leq \|(\widehat{b}_2^{*}, \widehat{b}_3^{*}) - (b_2, b_3)\|^2. \tag{10}$$

The consistency of $(\widehat{b}_2^{*}, \widehat{b}_3^{*})$ to (b_2, b_3) is proven as follows:

$$(\widehat{b}_2^{*}, \widehat{b}_3^{*}) = \max\{0, (F_s^t\, F_s)^{-1} F_s^t\, v_s\}$$

can be equivalently expressed as

$$(\widehat{b}_2^{*}, \widehat{b}_3^{*}) = \max\{0, \mathrm{Cov}(F_s, F_s)^{-1}\mathrm{Cov}(v_s, F_s)\}.$$

It is well known that

$$(b_2, b_3) = \mathrm{Cov}(f_s, f_s)^{-1}\mathrm{Cov}(v_s, f_s),$$

where $f_s = \mathrm{spr}X^{Bl} - 1^n E(\mathrm{spr}X^{Bl})$.

Since $(b_2, b_3) \geq \mathbf{0}$, then $\mathrm{Cov}(v_s, f_s) \geq 0$. It is well-known the strong consistency of $\mathrm{spr}X^{eBl}$ to $\mathrm{spr}X^{Bl}$ and thus it is also the consistency of $\mathrm{Cov}(F_s, F_s)$ and $\mathrm{Cov}(v_s, F_s)$ towards its populational values, $\mathrm{Cov}(f_s, f_s)$ and $\mathrm{Cov}(v_s, f_s)$, respectively. Besides, since the maximum is a continuous function, by applying the Continuous Mapping Theorem it holds that

$$(\widehat{b}_2^{*}, \widehat{b}_3^{*}) = \max\{0, \mathrm{Cov}(F_s, F_s)^{-1}\mathrm{Cov}(v_s, F_s)\}$$

and

$$\max\{0, \mathrm{Cov}(f_s, f_s)^{-1}\mathrm{Cov}(v_s, f_s)\} = \mathrm{Cov}(f_s, f_s)^{-1}\mathrm{Cov}(v_s, f_s) = (b_2, b_3),$$

so that

$$(\widehat{b}_2^{*}, \widehat{b}_3^{*}) \xrightarrow{n\to\infty} (b_2, b_3) \quad \text{a.s.} - [\text{P}].$$

Thus, combining (10) with the consistency of $(\widehat{b}_2^{*}, \widehat{b}_3^{*})$ to (b_2, b_3) it is obtained that $\|(\widehat{b}_2^{*}, \widehat{b}_3^{*}) - (\widehat{b}_2^{min}, \widehat{b}_3^{min})\|^2 \xrightarrow{n\to\infty} 0$ a.s.-[P], which proves the consistency of $(\widehat{b}_2^{min}, \widehat{b}_3^{min})$ to (b_2, b_3). □

Analytic expressions for the expectation and the standard error of the minimum-distance estimator are difficult to obtain. In Efron and Tibshirani (1993) it is proposed a bootstrap algorithm to estimate these moments. Applied to $(\widehat{b}_2, \widehat{b}_3)$, which is how $(\widehat{b}_2^{min}, \widehat{b}_3^{min})$ is denoted from now on, it is summarized as follows:

Algorithm 2: Bootstrap estimation of $\boldsymbol{E(\widehat{b}_l)}$ and se$\boldsymbol{(\widehat{b}_l)}$, for $\boldsymbol{l = 2, 3}$.
Let $\{(\boldsymbol{y}_j, \boldsymbol{x}_{1,j}, \ldots, \boldsymbol{x}_{k,j})\}_{j=1}^{n}$ be a simple random sample from the random intervals $(\boldsymbol{y}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k)$ and let $T \in \mathbb{N}$ be large enough.

1. Obtain T bootstrap samples of size n, $\{(\boldsymbol{y}_j^{\mathcal{B}}, \boldsymbol{x}_{1,j}^{\mathcal{B}}, \ldots, \boldsymbol{x}_{k,j}^{\mathcal{B}})\}_{j=1}^{n}$, by re-sampling uniformly and with replacement from the original sample.
2. Compute the bootstrap replica of the regression estimator, $\widehat{b}_l^{\mathcal{B}(t)}$, $t = 1, \ldots, T$.
3. Estimate the mean and the standard error of $\widehat{b}_l$ by the sample mean and the sample deviation of $\{\widehat{b}_l^{\mathcal{B}(t)}\}_{t=1}^{T}$, i.e.

$$\widehat{E}(\widehat{b}_l) = \overline{\widehat{b}_l^{\mathcal{B}}} = \frac{\sum_{t=1}^{T} \widehat{b}_l^{\mathcal{B}(t)}}{T}, \quad \text{and}$$

$$\widehat{\text{se}}(\widehat{b}_l) = \sqrt{\frac{\sum_{t=1}^{T} \left(\widehat{b}_l^{\mathcal{B}(t)} - \overline{\widehat{b}_l^{\mathcal{B}}}\right)^2}{T-1}}\,.$$

It is shown in Efron and Tibshirani (1993) that a number of bootstrap iterations T between 25 and 200 of the algorithm generally provides good approximations. In Sect. 4 some practical and simulated results are shown.

It is possible to obtain more numerically stable expressions for the estimators by applying the QR decomposition (see Golub and Van Loan 1996) to (8) and taking benefit from the triangular structure of the leading matrices. In fact, the set of triangular matrices is an stable subspace for products and inverses. Therefore, the computation of the inverses can be solved as a triangular system by back or forward-substitution—for upper or lower triangular matrices, respectively—(see Higham 1996).

Proposition 2 *The least-squares estimators of the model* M_G (2) *can be equivalently computed as* $(\widehat{b}_1, \widehat{b}_4) = R_m^{-1}\widetilde{y}_{m_1}$ *and* $(\widehat{b}_2, \widehat{b}_3)$ *by applying Algorithm 1 starting from* $R_s^{-1}\widetilde{y}_{s_1}$, *where given the QR decompositions of* $(F_m|v_m)$ *and* $(F_s|v_s)$, $Q_m,\ Q_s \in \mathbb{R}^{n\times n}$ *are orthogonal matrices and* $R_m,\ R_s \in \mathbb{R}^{2k\times 2k}$ *are upper triangular ones.*

Proof The first quadratic problem can be written as

$$\min_{A_m\in\mathbb{R}^{2k}} \|Q_m^T(F_m A_m - v_m)\|_2^2 = \min_{A_m\in\mathbb{R}^{2k}} \|R_m A_m - \widetilde{y}_{m_1}\|_2^2 + \|\widetilde{y}_{m_2}\|_2^2\,,$$

whose solution is $(\widehat{b}_1, \widehat{b}_4) = R_m^{-1}\widetilde{y}_{m_1}$. The second quadratic problem in (8) is written as

$$\min_{A_s\in\Gamma} \|Q_s^t(F_s A_s - v_s)\|_2^2 = \min_{A_s\in\Gamma} \|R_s A_s - \widetilde{y}_{s_1}\|_2^2 + \|\widetilde{y}_{s_2}\|_2^2$$

so that, departing from the non-negative OLS, $(\widehat{b}_2^*, \widehat{b}_3^*) = \max\{0, R_m^{-1}\widetilde{y}_{m_1}\}$ the estimator is obtained solving $\|(\widehat{b}_2^*, \widehat{b}_3^*) - (b_2, b_3)\|^2$ with Algorithm 1. □

Remark 1 The separate minimization of the problem (8) entails that the regression estimates do not depend on the value of the constant θ chosen for the metric. Thus, sensitivity analysis for the estimation process of the model M_G is not required, as happens with other models (see, for instance, Sinova et al. 2012).

3.3 Other models

The model M_G provides directly the extension to the multiple case for the simple linear model M addressed in Blanco-Fernández et al. (2011), by taking $b_3 = b_4 = (0, \overset{k}{\ldots}, 0)$. However, the extension of the basic simple model in González-Rodríguez et al. (2007) is not directly obtained from (2). The reason is that taking $b_1 = b_2$, and since $b_2 \geq 0$ without loss of generality in (2), then $b_1 \geq 0$ too. Thus, according to (3a), the linear relationship between the midpoints of the response and the explanatory intervals is always increasing. Clearly this is more restrictive than the relationship for *mid* variables transferred from the basic model. The extension of the basic simple regression model to the multiple case is formalized as follows:

$$\boldsymbol{y} = x^t b + \boldsymbol{\varepsilon}, \tag{11}$$

with $b = (b_1, b_2, \ldots, b_k)^t \in \mathbb{R}^k$ and $\boldsymbol{\varepsilon}$ such that $E(\boldsymbol{\varepsilon}|x) = \varDelta \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$. The following separate models are transferred:

$$\mathrm{mid}\boldsymbol{y} = \mathrm{mid}(x^t)\, b + \mathrm{mid}\, \boldsymbol{\varepsilon} \tag{12a}$$
$$\mathrm{spr}\boldsymbol{y} = \mathrm{spr}(x^t)\, |b| + \mathrm{spr}\, \boldsymbol{\varepsilon}. \tag{12b}$$

Extending directly the estimation method of the simple model proposed in González-Rodríguez et al. (2007) would lead to a computationally infeasible combinatorial problem. Alternatively, quadratic optimization techniques can be used to the estimation of (11). It is easy to show that the absolute value of $\widehat{b}$ and its sign can be estimated separately, by taking into account that $\widehat{b} = |\widehat{b}| \circ \mathrm{sign}(\widehat{b})$ and $\mathrm{sign}(\widehat{b})_i = \mathrm{sign}(\widehat{\mathrm{Cov}}(\mathrm{mid}\boldsymbol{y}, \mathrm{mid}x_i))$ for each $i = 1\ldots, k$. By following an analogous reasoning than for the model M_G, the LS estimation of the regression parameters guaranteeing the existence of the residuals gives $\widehat{\varDelta} = \boldsymbol{y} -_H \overline{x^t \widehat{b}}$ and $\widehat{b}$ is found through the following quadratic optimization problem subject to linear constraints:

$$\min_{a \in \varGamma_1} = \|v_m - G_m\, a\| + \theta \|v_s - G_s\, a\|,$$

where v_m and v_s are as in (8), $G_m = \mathrm{mid}X - 1^n(\overline{\mathrm{mid}X})$, $G_s = \mathrm{spr}X - 1^n(\overline{\mathrm{spr}X}) \in \mathbb{R}^{n\times k}$, $a \in \mathbb{R}^k$, and

$$\varGamma_1 = \{d \in (\mathbb{R}^k)^+ : \mathrm{spr}X\, d \leq \mathrm{spr}y\}.$$

In addition, standard numerical optimization methods can be used to solve this problem.

3.4 Goodness of the estimated linear model

Some classical concepts to measure the goodness of an estimated model can be defined in the interval framework, by taking into account the semilinear structure of the space of intervals. For instance, the determination coefficient of an estimated interval linear model, related to the proportion of variability of the interval response unexplained by the estimated model, can be defined in terms of the d_θ distance by means of expression

$$R^2 = 1 - \frac{\sum_{j=1}^{n} d_\theta^2(\boldsymbol{y}_j, \widehat{\boldsymbol{y}}_j)}{\sum_{j=1}^{n} d_\theta^2(\boldsymbol{y}_j, \overline{\boldsymbol{y}})} \,. \tag{13}$$

It is important to remark that the classical decomposition of the total sum of squares $\mathrm{SST} = \sum_{j=1}^{n} d_\theta^2(\boldsymbol{y}_j, \overline{\boldsymbol{y}})$ as $\mathrm{SSR} + \mathrm{SSE} = \sum_{j=1}^{n} d_\theta^2(\widehat{\boldsymbol{y}}_j, \overline{\boldsymbol{y}}) + \sum_{j=1}^{n} d_\theta^2(\boldsymbol{y}_j, \widehat{\boldsymbol{y}}_j)$ does not hold in this framework. Thus, R^2 in (13) differs in general from SSR/SST.

The mean square error (MSE) of the estimated linear models can also be computed in terms of the metric d_θ for intervals as

$$\mathrm{MSE}_{\mathrm{model}} = \frac{\sum_{j=1}^{n} d_\theta^2(\boldsymbol{y}_j, \widehat{\boldsymbol{y}}_j)}{n} \,. \tag{14}$$

Once the estimation problem is solved, the statistical analysis of the proposed interval linear models continues with the development of inferential studies on the models as, for example, confidence sets and hypothesis testing for the regression parameters, linearity testing, among others. Due to the lack of realistic general parametric models for random intervals, asymptotic and/or bootstrap techniques are generally applied in inferences (see, for instance, Gil et al. 2007). On one hand, classical procedures can be applied to the regression parameters whose LS estimators are not affected by the conditions assuring the interval coherence (Freedman 1981; Srivastava and Srivastava 1986). On the other hand, a thorough investigation is required for the case of constrained statistical inferences to the constrained regression estimators.

4 Empirical results

The practical applicability and the empirical behaviour of the proposed estimation procedures are illustrated is this section. For the sake of comparison with existing techniques, an interval dataset employed in previous interval regression problems is considered. Additionally, some simulations are performed in order to show the general performance of the methodology. The results are obtained by using the R implementation algorithms provided in http://bellman.ciencias.uniovi.es/SMIRE/Applications.html.

The estimation of the new flexible model M_G does not depend on θ (see Remark 1). However, the estimated basic models recalled in Sect. 2 depend on θ, as well as the computation of R^2 and $\mathrm{MSE}_{\mathrm{model}}$ for all the cases do. The usual value $\theta = 1/3$ for the metric d_θ is fixed.

4.1 Simulation results

The empirical performance of the regression estimators for the proposed linear models is investigated by means of simulations. Three independent random intervals $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ and an interval error $\boldsymbol{\varepsilon}$ will be considered. Let $\text{mid}\,\boldsymbol{x}_1 \sim \mathcal{N}(1,2)$, $\text{spr}\,\boldsymbol{x}_1 \sim \mathcal{U}(0,10)$, $\text{mid}\,\boldsymbol{x}_2 \sim \mathcal{N}(2,1)$, $\text{spr}\,\boldsymbol{x}_2 \sim \mathcal{X}_4^2$, $\text{mid}\,\boldsymbol{x}_3 \sim \mathcal{N}(1,3)$, $\text{spr}\,\boldsymbol{x}_3 \sim \mathcal{U}(0,5)$, $\text{mid}\,\boldsymbol{\varepsilon} \sim \mathcal{N}(0,1)$ and $\text{spr}\,\boldsymbol{\varepsilon} \sim \mathcal{X}_1^2$. Different linear expressions with the investigated structures will be considered.

- Model M_1: According to the multiple basic linear model presented in (11), $\boldsymbol{y}$ is defined by the expression:

$$\boldsymbol{y} = 2\boldsymbol{x}_1 - 5\boldsymbol{x}_2 - \boldsymbol{x}_3 + \varepsilon.$$

- Model M_2: A multiple flexible linear regression model following (4) is defined as:

$$\begin{aligned}\boldsymbol{y} = &-2\boldsymbol{x}_1^M + 5\boldsymbol{x}_2^M - \boldsymbol{x}_3^M + 2\boldsymbol{x}_1^S + 2\boldsymbol{x}_2^S + \boldsymbol{x}_3^S + \boldsymbol{x}_1^C + \boldsymbol{x}_2^C + 3\boldsymbol{x}_3^C \\ &+ 0.5\boldsymbol{x}_1^R + \boldsymbol{x}_2^R - 3\boldsymbol{x}_3^R + \varepsilon.\end{aligned}$$

From each linear model $s = 10{,}000$ random samples have been generated for different sample sizes n. Table 1 shows the estimated mean value and standard error of the LS estimators. Besides, the estimated MSE of each estimator is computed as

$$\widehat{\text{MSE}}(\widehat{b_l}) = \left(\sum_{i=1}^{s} ((\widehat{b_l})_i - b_l)^2\right) / s.$$

The findings display that the LS estimators of the models behave empirically good, since the mean values of the estimates are always closer to the corresponding regression parameters and the standard error approximates zero, as the sample size n increases. Moreover, the values for the estimated MSE tend to zero as n increases too, which agrees with the empirical consistency of the estimators.

The empirical performance of the regression estimators can also be checked graphically. In Fig. 1 the box-plots of the s estimates of the model M_1 are presented for $n = 30$ (left-side plot) and $n = 100$ (right-side plot) sample observations. In all the cases the boxes reduce their width around the true value of the corresponding parameter on the population linear model as the sample size n increases, which illustrates the empirical consistency of the estimators. Analogous conclusions are obtained for model M_2 in Fig. 2.

4.2 Comparative example

A methodological example concerning the relationship between the daily fluctuations of the systolic and diastolic blood pressures and the pulse rate over a sample of

Table 1 Empirical behaviour of the regression estimators

	$\widehat{b}_l$	$n = 30$			$n = 100$			$n = 500$		
		$\widehat{E}(\widehat{b}_l)$	$\widehat{\text{se}}(\widehat{b}_l)$	$\widehat{\text{MSE}}(\widehat{b}_l)$	$\widehat{E}(\widehat{b}_l)$	$\widehat{\text{se}}(\widehat{b}_l)$	$\widehat{\text{MSE}}(\widehat{b}_l)$	$\widehat{E}(\widehat{b}_l)$	$\widehat{\text{se}}(\widehat{b}_l)$	$\widehat{\text{MSE}}(\widehat{b}_l)$
M_1	$\widehat{b}_1$	1.9737	0.0490	0.0011	1.9382	0.0234	0.0007	1.9853	0.0099	0.00013
	$\widehat{b}_2$	− 4.9164	0.0724	0.0069	− 5.0211	0.0299	0.0014	− 4.9938	0.0127	0.00025
	$\widehat{b}_3$	− 1.0879	0.0509	0.0023	− 0.9561	0.0262	0.0007	− 0.9998	0.0106	0.00012
M_2	$\widehat{b}_1^1$	− 2.0064	0.1145	0.0111	− 2.0010	0.0520	0.0028	− 2.0000	0.0224	0.00055
	$\widehat{b}_1^2$	5.0128	0.2278	0.0474	4.9990	0.1041	0.0127	4.9992	0.0449	0.00201
	$\widehat{b}_1^3$	− 0.9970	0.0760	0.0053	− 1.0000	0.0347	0.0012	− 1.0004	0.0149	0.00024
	$\widehat{b}_2^1$	1.9672	0.0925	0.0095	1.9775	0.0408	0.0021	1.9884	0.0169	0.00031
	$\widehat{b}_2^2$	1.9703	0.1043	0.0098	1.9797	0.0434	0.0021	1.9890	0.0171	0.00042
	$\widehat{b}_2^3$	0.9275	0.1831	0.0357	0.9582	0.0822	0.0073	0.9771	0.0336	0.00140
	$\widehat{b}_3^1$	0.9352	0.2049	0.0414	0.9597	0.0908	0.0092	0.9789	0.0365	0.00162
	$\widehat{b}_3^2$	0.8841	0.2593	0.0773	0.9205	0.1198	0.0190	0.9585	0.0486	0.00327
	$\widehat{b}_3^3$	2.9664	0.1486	0.0198	2.9719	0.0638	0.0042	2.9856	0.0257	0.00081
	$\widehat{b}_4^1$	0.4958	0.0775	0.0052	0.4989	0.0358	0.0013	0.5001	0.0156	0.00027
	$\widehat{b}_4^2$	0.9969	0.0872	0.0063	0.9979	0.0377	0.0014	0.9997	0.0159	0.00032
	$\widehat{b}_4^3$	− 3.0004	0.1552	0.0200	− 3.0007	0.0716	0.0052	− 2.9975	0.0311	0.00104

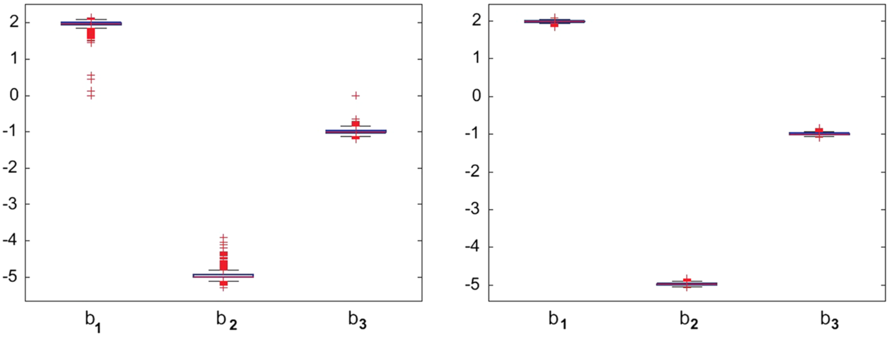

Fig. 1 Box plot of the LS estimators for model M_1, $n = 30$ (left); $n = 100$ (right)

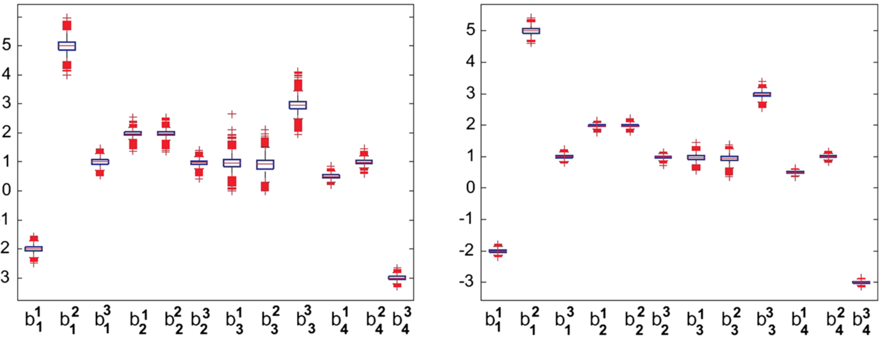

Fig. 2 Box plot of the LS estimators for model M_2, $n = 30$ (left); $n = 100$ (right)

patients in the Hospital *Valle del Nalón*, in Spain, is considered. This real-life example has been previously explored in Blanco-Fernández et al. (2011), Gil et al. (2007) and González-Rodríguez et al. (2007), and the data can be found in http://bellman.ciencias.uniovi.es/SMIRE/Hospital.html. From a population of 3000 inpatients, random intervals $\boldsymbol{y} =$"fluctuation of the diastolic blood pressure of a patient over a day", $\boldsymbol{x}_1 =$"fluctuation of the systolic blood pressure over the same day" and $\boldsymbol{x}_2 =$"pulse range variation over the same day" are defined. The Nephrology Unit of the hospital has supplied a random sample for $(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2)$ which is available in the references cited above.

Consider the problem of modelling in a linear fashion the daily range of the diastolic blood pressure of a patient. This is performed in terms of the patient's corresponding systolic pressure fluctuation and the pulse range variation. Classical regression techniques could be applied by summarizing the sample intervals into point data, the midpoints in general. Alternatively, midpoints and spreads of the response can be estimated by means of separate models. Moreover, a multiple linear model based on interval arithmetic can be formalized between the random intervals $\boldsymbol{y}$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ and estimated from the available interval sample set. The estimation results for all the alternatives, both the existing methods recalled in Sect. 2 and the new multiple interval model M_G is shown in Table 2.

Table 2 Estimation results in Example 4.2—multiple models

Model	Estimated interval model	Separate estimated models for *mid* and *spr* variables	R^2	$\text{MSE}_{\text{model}}$
Classical-midpoints	–	$\widehat{\text{mid}\,\boldsymbol{y}} = 0.4497\text{mid}\,\boldsymbol{x}_1 + 0.0517\text{mid}\,\boldsymbol{x}_2 + 13.6263$	0.4337	86.87
Basic model (11)	$\widehat{\boldsymbol{y}} = 0.4094\boldsymbol{x}_1 + 0.0463\boldsymbol{x}_2$	$\text{mid}\,\widehat{\boldsymbol{y}} = 0.4094\text{mid}\,\boldsymbol{x}_1 + 0.0463\text{mid}\,\boldsymbol{x}_2 + 19.9399$	0.4221	89.37
	$+ [10.3630, 29.5168]$	$\text{spr}\widehat{\boldsymbol{y}} = 0.4094\text{spr}\,\boldsymbol{x}_1 + 0.0463\text{spr}\,\boldsymbol{x}_2 + 9.5769$		
Lima Neto et al. model	–	$\widehat{\text{mid}\,\boldsymbol{y}} = 0.4497\text{mid}\,\boldsymbol{x}_1 + 0.0517\text{mid}\,\boldsymbol{x}_2 + 13.6263$	0.4401	69.96
		$\widehat{\text{spr}\boldsymbol{y}} = 0.4847\text{spr}\boldsymbol{x}_1 + 0.3605\text{spr}\boldsymbol{x}_2 + 0.4947$		
D'Urso model	–	$\widehat{\text{mid}\,\boldsymbol{y}} = 0.5434\text{mid}\,\boldsymbol{x}_1 + 0.0188\text{mid}\,\boldsymbol{x}_2$	0.4837	60.92
		$- 0.4615\text{spr}\,\boldsymbol{x}_1 + 0.1003\text{spr}\,\boldsymbol{x}_2 + 16.3601$		
		$\widehat{\text{spr}\,\boldsymbol{y}} = - 0.0357\text{mid}\,\boldsymbol{x}_1 - 1.24 \times 10^{-3}\text{mid}\,\boldsymbol{x}_2$		
		$+ 0.0304\text{spr}\,\boldsymbol{x}_1 - 6.60 \times 10^{-3}\text{spr}\,\boldsymbol{x}_2 + 29.2195$		
Multiple M_G (2)	$\widehat{\boldsymbol{y}} = 0.5435\boldsymbol{x}_1^M + 0.0190\boldsymbol{x}_2^M$	$\text{mid}\widehat{\boldsymbol{y}} = 0.5435\text{mid}\,\boldsymbol{x}_1 + 0.0190\text{mid}\,\boldsymbol{x}_2$	0.5083	59.02
	$+ 0.2588\boldsymbol{x}_1^S + 0.1685\boldsymbol{x}_2^S$	$- 0.4446\text{spr}\,\boldsymbol{x}_1 + 0.1113\text{spr}\,\boldsymbol{x}_2 + 15.5203$		
	$+ 2.73 \times 10^{-19}\boldsymbol{x}_1^C - 0.4446\boldsymbol{x}_1^R$	$\text{spr}\widehat{\boldsymbol{y}} = 0.2588\text{spr}\,\boldsymbol{x}_1 + 0.1685\text{spr}\,\boldsymbol{x}_2$		
	$+ 0.1113\boldsymbol{x}_2^R + [3.2032, 27.8373]$	$+ 2.73 \times 10^{-19}\lvert\text{mid}\,\boldsymbol{x}_1\rvert + 12.3170$		

Several comments can be extracted from these results. The classical procedure and the models by Lima Neto and De Carvalho (2010) and D'Urso (2003) do not provide an interval estimated equation to relate the intervals, but separate fitting real-valued equations for *mid* and *spr* variables (only for *mids* in the classical approach). The estimated model for the *mid* variables coincides with the classical OLS estimation for the model M and the Lima-Neto models. Nevertheless, it is not the case for the estimated relationship for the *spr* variables, due to the consideration of different conditions in the estimation process. The determination coefficient and the MSE of all the models are computed by formulas (13) and (14), respectively. Moreover, the poorest goodness of fit corresponds to the basic interval model. This clearly shows that the condition of identical regression parameters for modelling mid$\boldsymbol{y}$ and spr$\boldsymbol{y}$ is too restrictive in this application. All the remainder models for intervals behave better than the classical estimation. This might be due to the loss of the information from the spreads in this latter approach. The highest value of R^2 is obtained for the M_G models, both in the simple and the multiple cases. This is coherent to the great flexibility on the obtained relationships to estimate both *mid* and *spr* components of $\boldsymbol{y}$. It is shown that the separate models by D'Urso (2003) reach a value for the determination coefficient slightly lower than the M_G model. However, from these separate fitting models 29 of the 59 sample individuals do not fulfil the existence of the interval residuals. Thus, these solutions are not valid as regression estimates of an interval model formalized theoretically for relating linearly the random intervals $\boldsymbol{y}$, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. The separate estimated models by Lima Neto and De Carvalho (2010) fail in the existence of the sample interval residuals too. The estimation procedures of the M_G model proposed here provides accurate fitting results in addition to interval coherency. It is important to recall that the formalization of the proposed model in a probabilistic framework allows us to develop further statistical analysis on the regression problem for these variables based on the available interval dataset. This is the case of constructing confidence intervals for the regression parameters, testing the explicative power of the regressors, to name but a few.

5 Conclusions

Previous simple linear regression models for interval-valued data based on the set arithmetic are extended. As a result, new models arise representing not only an extension but a generalization of the previous ones, allowing to study new relationships between the variables. In all cases the search of the LS estimators involves minimization problems with constraints. The constraints are necessary to assure the existence of the residuals and thus, the coherency of the estimated model with the population one.

A flexible multiple model based on the canonical decomposition and allowing cross-relationships between midpoints and spreads is presented. The LS estimates can be found by transforming the quadratic problem into a linear complementary problem and solving it by means of Lemke's algorithm. For consistency purposes, the minimum-distance estimator is presented. Moreover, the latter has a lower computational complexity. For all these reasons, it is the preferable alternative to be considered.

The practical applicability of the proposed model is illustrated by means of some examples. The estimation results have been compared with classical regression techniques, as well as with existing regression analysis methods for interval-valued data, reaching the new estimators better results. Simulation studies show the empirical validity of the estimation process for all the models.

The development of inferential studies for the models, as the development of confidence sets for the regression parameters and hypothesis testing about the theoretical models, are to be addressed as future research.

Acknowledgements The research in this paper has been partially supported by the Spanish Government through MINECO-18-MTM2017-89632-P Grant and by the COST Action 1408. Their financial support is greatfully acknowledged.

References

Billard L, Diday E (2000) Regression analysis for interval-valued data. Data analysis, classification and related methods. In: Kiers HAL et al (eds) Proceedings of 7th conference IFCS, vol 1, pp 369–374

Blanco-Fernández Á, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. Comput Stat Data Anal 55(9):2568–2578

Blanco-Fernández Á, Colubi A, García-Bárzana M (2013) A set arithmetic-based linear regression model for modelling interval-valued responses through real-valued variables. Inf Sci 247(20):109–122

Boruvka A, Cook RJ (2015) A Cox–Aalen model for interval-censored data. Scand J Stat 42(2):414–426

Boukezzoula R, Galichet S, Bisserier A (2011) A midpoint radius approach to regression with interval data. Int J Approx Reason 52(9):1257–1271

Černý M, Rada M (2011) On the possibilistic approach to linear regression with rounded or interval-censored data. Meas Sci Rev 11(2):34–40

Diamond P (1990) Least squares fitting of compact set-valued data. J Math Anal Appl 147:531–544

D'Urso PP (2003) Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput Stat Data Anal 42:47–72

D'Urso PP, Giordani P (2004) A least squares approach to principal component analysis for interval valued data. Chemom Intell Lab 70:179–192

Efron B, Tibshirani R (1993) An introduction to the bootstrap. Chapman & Hall, New York

Freedman DA (1981) Bootstrapping regression models. Ann Stat 9(6):1218–1228

Gil MA, González-Rodríguez G, Colubi A, Montenegro M (2007) Testing linear independence in linear models with interval-valued data. Comput Stat Data Anal 51:3002–3015

Gillis N (2012) Sparse and unique nonnegative matrix factorization through data preprocessing. J Mach Learn Res 13:3349–3386

Golub HG, Van Loan CF (1996) Matrix computations. Johns Hopkins University Press, Baltimore

González-Rodríguez G, Blanco Á, Corral N, Colubi A (2007) Least squares estimation of linear regression models for convex compact random sets. Adv Data Anal Classif 1:67–81

Higham NJ (1996) Accuracy and stability of numerical algorithms. Society for Industrial and Applied Mathematics, Philadelphia

Jahanshahloo GR, Hosseinzadeh Lotfi F, Rostamy Malkhalifeh M, Ahadzadeh Namin M (2008) A generalized model for data envelopment analysis with interval data. Appl Math Model 33:3237–3244

Johnston J (1972) Econometric methods. McGraw-Hill Book Co., New York

Körner R (1997) On the variance of fuzzy random variables. Fuzzy Set Syst 92:83–93

Lauro CN, Palumbo F (2005) Principal component analysis for non-precise data. New developments in classification and data analysis. In: Studies in classification, data analysis and knowledge organization. Springer, pp 173–184

Lemke CE (1962) A method of solution for quadratic programs. Manag Sci 8(4):442–453

Liew CK (1976) Inequality constrained least-squares estimation. J Am Stat Assoc 71:746–751

Lima Neto EA, De Carvalho FAT (2010) Constrained linear regression models for symbolic interval-valued variables. Comput Stat Data Anal 54:333–347

Lima Neto EA, Dos Anjos UU (2015) Regression model for interval-valued variables based on copulas. J Appl Stat 42(9):2010–2029

Näther W (1997) Linear statistical inference for random fuzzy data. Statistics 29(3):221–240

Park C, Yongho J, Kee-Hoon K (2016) An exploratory data analysis in scale-space for interval-valued data. J Appl Stat 43(14):2643–2660

Ramos-Guajardo AB, Grzegorzewski P (2016) Distance-based linear discriminant analysis for interval-valued data. Inf Sci 272:591–607

Ramos-Guajardo AB, Colubi A, González-Rodríguez G (2014) Inclusion degree tests for the Aumann expectation of a random interval. Inf Sci 288(20):412–422

Sinova B, Colubi A, Gil MA, González-Rodríguez G (2012) Interval arithmetic-based linear regression between interval data: discussion and sensitivity analysis on the choice of the metric. Inf Sci 199:109–124

Srivastava MS, Srivastava VK (1986) Asymptotic distribution of least squares estimator and a test statistic in linear regression models. Econ Lett 21:173–176

Trutschnig W, González-Rodríguez G, Colubi A, Gil MA (2009) A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread. Inf Sci 179(23):3964–3972

Wets RJB (1991) Constrained estimation: consistency and asymptotics. Appl Stoch Model Data Anal 7:17–32

Yu Q, Hsu Y, Yu K (2014) A necessary and sufficient condition for justifying non-parametric likelihood with censored data. Metrika 77(8):995–1011

Zhang Z (2009) Linear transformation models for interval-censored data: prediction of survival probability and model checking. Stat Model 9(4):321–343

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.