Multiple set arithmetic-based linear regression models for interval-valued variables

Angela Blanco-Fernández [1], Ana Colubi [1], Marta García-Bárzana [1] and Erricos J. Kontoghiorghes [2]

[1] Department of Statistics, Oviedo University, Spain

[2] Department of Commerce, Finance and Shipping, Cyprus University of Technology, and School of Economics and Finance, Queen Mary University of London,UK

Address for correspondence: Angela Blanco-Fernández, Department of Statistics, Oviedo University, C/ Calvo Sotelo s/n, Oviedo 33007, Spain.

E-mail: `blancoangela@uniovi.es`.

Phone: (+34) 985 103 126.

Fax: (+34) 985 103 354.

Abstract: New flexible regressions to model linear relationships between interval-valued random variables are presented. These new models account for cross relationships between midpoints and spreads of the intervals in a unique equation based on the interval arithmetic. The estimation problem, which can be written as a contrained minimization problem, is theoretically analyzed and empirically tested. Numerically stable general expressions of the estimators are provided. The simple linear regression entails less computational complexity and a more efficient estimation algorithm is developed. The main differences between the new and the existing methods are highlighted in a real-life application. It is shown that the new model provides the most accurate results by preserving the coherency with the interval nature of the data.

Key words: interval-valued data; least-squares estimators; linear modelling; multiple regression; set arithmetic

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 et al. (2005). Censoring and grouping processes also produce intervals; see Černý et al. (2011); Joly et al. (2009); Zhang (2009), among others. Symbolic Data Analysis (SDA) considers intervals for summarizing information stored in large data sets, as in Billard et al. (2000); Lima Neto et al. (2010). 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 et al. (2004); González-Rodríguez et al. (2007). This work focuses on this latter approach and its aim is to develop new regression models for the 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 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 (see D'Urso, 2003; Lima Neto et al., 2010). In a different context, possibilistic regression models are considered (Boukezzoula et al., 2011; Černý et al., 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 interval-valued random variables 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).

Here extensions of the set arithmetic-based linear models 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. On the other hand, multiple regression models allowing several explanatory variables to model the response are formalized.

The rest of the paper is organized as follows: In Section 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 Section 3. The least-squares estimation problem is analyzed and numerically stable expressions are derived. In Section 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 $\mathcal{K}_{\mathrm{c}}(\mathbb{R}) = \{[a_1, a_2] : a_1, a_2 \in \mathbb{R}, a_1 \leq a_2\}$. Each interval $A \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ can be parametrized in terms of its midpoint, $\mathrm{mid}\,A = (\sup A + \inf A)/2$, and its spread, $\mathrm{spr}\,A = (\sup A - \inf A)/2$. The notation $A = [\mathrm{mid}A \pm \mathrm{spr}A]$ will be used. An alternative representation for intervals is the so-called canonical decomposition, introduced in Blanco-Fernández et al. (2011), given by $A = \mathrm{mid}A[1 \pm 0] + \mathrm{spr}A[0 \pm 1]$. 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 $\mathcal{K}_{\mathrm{c}}(\mathbb{R})$. In terms of the (mid, spr)-representation these operations can be jointly expressed as

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

for any $A, B \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ and $\lambda \in \mathbb{R}$. The space $(\mathcal{K}_{\mathrm{c}}(\mathbb{R}), +, \cdot)$ 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 $\mathrm{spr}B \leq \mathrm{spr}A$ (see Blanco-Fernández et al., 2011 for details).

For every $A, B \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$, an L_2-type generic metric is introduced in Trutschnig et al. (2009) as $d_\theta(A, B) = ((\mathrm{mid}A - \mathrm{mid}B)^2 + \theta\,(\mathrm{spr}A - \mathrm{spr}B)^2)^{\frac{1}{2}}$, for an arbitrary

$\theta \in (0, \infty)$. 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 $\boldsymbol{x} : \Omega \to \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ is a random interval iff $\operatorname{mid} \boldsymbol{x}, \operatorname{spr} \boldsymbol{x} : \Omega \to \mathbb{R}$ are real random variables and $\operatorname{spr} \boldsymbol{x} \geq 0$. 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 $(\mathcal{K}_{\mathrm{c}}(\mathbb{R}), d_\theta)$, 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 $\mathcal{K}_{\mathrm{c}}(\mathbb{R})$ 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 $\operatorname{Cov}(x, y)$ 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 Section 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 $\boldsymbol{\varepsilon} : \Omega \to \mathcal{K}_{\mathrm{c}}(\mathbb{R})$ is 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 interval linear model, called model M, has been formalized in Blanco-Fernández et al. (2011) as $\boldsymbol{y} = [b_1\,\text{mid}\boldsymbol{x} \pm b_2\,\text{spr}\boldsymbol{x}] + \boldsymbol{\varepsilon}$, $b_1 \in \mathbb{R}$, $b_2 \geq 0$. The transferred linear relationships are in this case $\text{mid}\boldsymbol{y} = b_1\text{mid}\boldsymbol{x} + \text{mid}\boldsymbol{\varepsilon}$ and $\text{spr}\boldsymbol{y} = b_2\text{spr}\boldsymbol{x} + \text{spr}\boldsymbol{\varepsilon}$, with $b_1 \neq b_2$ in general.

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 et al. (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 offer 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 $\mathrm{M}_{\mathbf{G}}$

A novel multiple linear regression model for intervals is presented. It arises as a natural extension of the model M 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:

$$\boldsymbol{y} = [(b_1 \, \mathrm{mid} x^t + b_4 \, \mathrm{spr} x^t) \pm (b_2 \, \mathrm{spr} x^t + b_3 \, |\mathrm{mid} x^t|)] + \boldsymbol{\varepsilon} \tag{3.1}$$

where $b_1, b_4 \in \mathbb{R}^k$, $b_2, b_3 \in \mathbb{R}^{k^+}$, $\mathrm{mid}\, x = (\mathrm{mid}\, \boldsymbol{x}_1, \mathrm{mid}\, \boldsymbol{x}_2, \ldots, \mathrm{mid}\, \boldsymbol{x}_k)^t \in \mathbb{R}^k$ (analogously $\mathrm{spr}\, x$), and $\boldsymbol{\varepsilon}$ is a random interval-valued error such that $E(\boldsymbol{\varepsilon}|x) = \Delta \in \mathcal{K}_{\mathrm{c}}(\mathbb{R})$. 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 (3.1) are

$$\mathrm{mid}\, \boldsymbol{y} = \mathrm{mid}\, x^t \, b_1 + \mathrm{spr}\, x^t \, b_4 + \mathrm{mid}\, \boldsymbol{\varepsilon} \ , \tag{3.2a}$$

$$\mathrm{spr}\, \boldsymbol{y} = \mathrm{spr}\, x^t \, b_2 + |\mathrm{mid}\, x^t| \, b_3 + \mathrm{spr}\, \boldsymbol{\varepsilon} \ . \tag{3.2b}$$

Thus, both variables $\mathrm{mid}\boldsymbol{y}$ and $\mathrm{spr}\boldsymbol{y}$ are modelled from the complete information provided by the independent random intervals in x, characterized by the random vector $(\mathrm{mid} x, \mathrm{spr} x)$. 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:

$$\boldsymbol{y} = X^{Bl}\, B + \boldsymbol{\varepsilon}\ , \tag{3.3}$$

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 (3.3) 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 LS estimation searches for $\widehat{B}$ and $\widehat{\Delta}$ minimizing $d_\theta^2(y, X^{eBl}A+1^nC)$ 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 $\mathrm{spr}\, X$) so that the following conditions are to be included in the minimization problem:

$$\mathrm{spr}\, X\, a_2 + |\mathrm{mid}\, X|\, a_3 \leq \mathrm{spr}\, y. \tag{3.4}$$

Analogously to what happens in classical regression, the estimate of the (interval-valued) intercept term Δ can be obtained first. If $\widehat{B}$ verifies (3.4), 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

$$\widehat{\Delta} = \overline{\boldsymbol{y}} -_H \overline{X^{eBl}}\widehat{B} \ . \tag{3.5}$$

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

$$d_\theta^2(y -_H X^{eBl}A, \overline{\boldsymbol{y}} -_H \overline{X^{eBl}}A) \tag{3.6}$$

$$\text{subject to}$$

$$\operatorname{spr} X\, a_2 + |\operatorname{mid} X|\, a_3 \leq \operatorname{spr} y$$

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 3.1 *The least-squares estimators of the pairs of regression parameters (b_1, b_4) and (b_2, b_3) in (3.1) 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 $v_m = \operatorname{mid} y - \overline{\operatorname{mid}\mathbf{y}} 1^n \in \mathbb{R}^n$, $v_s = \operatorname{spr} y - \overline{\operatorname{spr}\mathbf{y}} 1^n \in \mathbb{R}^n$, $F_m = \operatorname{mid} X^{eBl} - 1^n(\overline{\operatorname{mid} X^{eBl}}) \in \mathbb{R}^{n\times 2k}$, $F_s = \operatorname{spr} X^{eBl} - 1^n(\overline{\operatorname{spr} X^{eBl}}) \in \mathbb{R}^{n\times 2k}$, $\operatorname{mid} X^{eBl} = (\operatorname{mid} X, \operatorname{spr} X) \in \mathbb{R}^{n\times 2k}$, $\operatorname{spr} X^{eBl} = (\operatorname{spr} X, |\operatorname{mid} X|) \in \mathbb{R}^{n\times 2k}$ *and* $D = \left(-I_{2k}\ ,\ spr X^{eBl}\right)^t \in \mathbb{R}^{(2k+n)\times 2k}$.

Proof 3.1 The problem (3.6) is solved by transforming it to an equivalent quadratic optimization problem, as follows:

$$\min_{A_m \in\ \mathbb{R}^{2k},\ A_s \in\ \Gamma} \|v_m - F_m A_m\|^2 + \theta\, \|v_s - F_s A_s\|^2 \tag{3.7}$$

$$\Gamma = \{A_s \in \mathbb{R}^{2k} : D\, A_s \leq d\}$$

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 $d = \Big(0_{2k}\ ,\ \mathrm{spr}y\Big)^t \in \mathbb{R}^{(2k+n)\times 1}$.

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

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

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

$$\omega = M\,\lambda + q$$
$$s.t. \quad \omega, \lambda \geq 0\ ,\ \omega_j \lambda_j = 0\ ,\ j = 1, \ldots, n+1\ ,$$

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 (3.7) is $(\widehat{b}_2, \widehat{b}_3) = H^{-1}\,(c - D^t\,\lambda)$. □

Observe that $(\widehat{b}_1, \widehat{b}_4)$ coincides with the OLS estimator of the classical multiple regression model (3.2a). 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

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

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 and special caution should be taken in reaching the correct solution. For instance, analytic expressions for its expectation and standard error are difficult to obtain. In Efron et al. (1993) it is proposed a bootstrap algorithm to estimate these moments. Applied to $(\widehat{b}_2, \widehat{b}_3)$, it is summarized as follows:

Algorithm 2: Bootstrap estimation of $E(\widehat{b}_l)$ and $\mathrm{se}(\widehat{b}_l)$, for $l = 2, 3$.

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, $\{(\boldsymbol{y}_j^*, \boldsymbol{x}_{1,j}^*, \ldots, \boldsymbol{x}_{k,j}^*)\}_{j=1}^n$, by re-sampling uniformly and with replacement from the original sample.

2. Compute the bootstrap replica of the regression estimator, $\widehat{b}_l^{*(t)}$, $t = 1, \ldots, T$.

3. Estimate the mean and the standard error of $\widehat{b}_l$ by the sample mean and the sample deviation of $\{\widehat{b}_l^{*(t)}\}_{t=1}^T$, i.e.

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

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

It is shown in Efron et al. (1993) that a number of bootstrap iterations T between 25 and 200 of the algorithm generally provides good approximations. In Section 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 et al., 1996) to (3.7) 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 3.2 *The least-squares estimators of the model* M_G *(3.1) can be equivalently computed as*

$$(\widehat{b}_1, \widehat{b}_4) = R_m^{-1} \widetilde{y}_{m_1}$$

$$(\widehat{b}_2, \widehat{b}_3) = R_s^{-1} \widetilde{y}_{s_1} - R_s^{-1}\, R_s^{-t} D^t \lambda,$$

where $Q_m^t(F_m|v_m) = \left(\begin{array}{c|c} R_m & \widetilde{y}_{m_1} \\ 0 & \widetilde{y}_{m_2} \end{array}\right)$ *and* $Q_s^t(F_s|v_s) = \left(\begin{array}{c|c} R_s & \widetilde{y}_{s_1} \\ 0 & \widetilde{y}_{s_2} \end{array}\right)$ *are the QR decomposition of* $(F_m|v_m)$ *and* $(F_s|v_s)$*, respectively.*

Proof 3.2 The QR decompositions of $(F_m|v_m)$ and $(F_s|v_s)$ work with the orthogonal matrices Q_m, $Q_s \in \mathbb{R}^{n\times n}$ and the upper triangular matrices R_m, $R_s \in \mathbb{R}^{2k\times 2k}$.

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 (3.7) is reformulated as

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

$$= \min_{A_s\in\Gamma} A_s^t \underbrace{R_s^t R_s}_{H} A_s - 2A_s^t \underbrace{R_s^t \widetilde{y}_{s_1}}_{c} + \|\widetilde{y}_{s_1}\|_2^2 + \|\widetilde{y}_{s_2}\|_2^2.$$

Consider applying Lemke's algorithm to $M = DH^{-1}D^t = D(R_s^t R_s)^{-1}D^t =$ $= (DR_s^{-1})(DR_s^{-1})^t$ and $q = d - DH^{-1}c = d - D(R_s^t R_s)^{-1}R_s^t \widetilde{y}_{s_1} = d - DR_s^{-1}\widetilde{y}_{s_1}$. It follows that the solution is given by $(\widehat{b}_2, \widehat{b}_3) = H^{-1}(c - D^t\lambda) = R_s^{-1}\widetilde{y}_{s_1} - (R_s^t R_s)^{-1}D^t\lambda =$ $R_s^{-1}\widetilde{y}_{s_1} - R_s^{-1}\, R_s^{-t}D^t\lambda$.

□

Remark 3.1 The separate minimization of the problem (3.7) 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 (Sinova et al., 2012).

3.3 Fast algorithm for the simple linear model

When the simple case is considered (i.e. the linear modelling of the interval response $\boldsymbol{y}$ by means of one explanatory interval $\boldsymbol{x}$), the estimation of the model M_G can be solved through a different process. It makes use of graphical ideas in $\mathbb{R}^2$ to obtain the estimates of the parameters with a lower computational cost.

The model M_G relating $\boldsymbol{y}$ in terms of $\boldsymbol{x}$ takes the expression

$$\boldsymbol{y} = b_1\,\boldsymbol{x}^M + b_2\,\boldsymbol{x}^S + b_3\,\boldsymbol{x}^C + b_4\,\boldsymbol{x}^R + \boldsymbol{\varepsilon}\ , \tag{3.8}$$

with $b_i \in \mathbb{R}$, $i = 1, 2, 3, 4$ and $E(\boldsymbol{\varepsilon}|\boldsymbol{x}) = \Delta \in \mathcal{K}_c(\mathbb{R})$.

Once $\widehat{\Delta}$ is obtained as in (3.5), the minimization problem which solves the LS estimation of $B = (b_1, b_2, b_3, b_4)^t \in \mathbb{R}^4$ is

$$\min_{\substack{(a,d) \in \mathbb{R}^2 \\ (b,c) \in \Gamma_G}} \frac{1}{n}\sum_{j=1}^{n} d_\theta^2\Big(\boldsymbol{y}_j -_H (a\boldsymbol{x}_j^M + b\boldsymbol{x}_j^S + c\boldsymbol{x}_j^C + d\boldsymbol{x}_j^R), \overline{\boldsymbol{y}} -_H (a\overline{\boldsymbol{x}^M} + b\overline{\boldsymbol{x}^S} + c\overline{\boldsymbol{x}^C} + d\overline{\boldsymbol{x}^R})\Big)\ , \tag{3.9}$$

where $\Gamma_G = \{(b,c) \in [0,\infty)\times[0,\infty) : b\,\mathrm{spr}\boldsymbol{x}_j + c|\mathrm{mid}\boldsymbol{x}_j| \leq \mathrm{spr}\boldsymbol{y}_j, \forall j = 1,\ldots,n\}$.

The minimization over (a,d) is solved without restrictions and it leads to the following estimators of the coefficients (b_1, b_4):

$$(\widehat{b}_1, \widehat{b}_4)^t = S_1^{-1} z_1. \tag{3.10}$$

Here $z_1 = (\widehat{\sigma}_{\boldsymbol{x}^M,\boldsymbol{y}}, \widehat{\sigma}_{\boldsymbol{x}^R,\boldsymbol{y}})^t$ and S_1 corresponds to the sample covariance matrix of the interval-valued random vector $(\boldsymbol{x}^M, \boldsymbol{x}^R)$.

The minimization over (b,c) in the feasible set Γ_G, which is nonempty, closed and convex, is solved by using graphical ideas. The addend of the function in (3.9) to be minimized over (b,c) can be expressed as the globally convex function

$$g(b,c) = b^2\widehat{\sigma}^2_{\boldsymbol{x}^S} + c^2\widehat{\sigma}^2_{\boldsymbol{x}^C} + 2bc\widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{x}^C} - 2b\widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{y}} - 2c\widehat{\sigma}_{\boldsymbol{x}^C,\boldsymbol{y}}\ .$$

If the global minimum of the function g is so that $(b^*, c^*)^t \notin \Gamma_G$, then the local minimum of g over Γ_G is unique, and it is located on the boundary of Γ_G. The boundary of Γ_G, denoted by $\mathrm{fr}(\Gamma_G)$, verifies that $\mathrm{fr}(\Gamma_G) = L_1 \cup L_2 \cup L_3$, where L_i, $i = 1,2,3$ are the following sets:

- $L_1 = \left\{(0,c)\,|\,0 \leq c \leq r_0 = \min_{j=1,\ldots,n} \frac{\mathrm{spr}\boldsymbol{y}_j}{|\mathrm{mid}\boldsymbol{x}_j|}\right\}$,
- $L_2 = \{(b, \min_{j=1\ldots n}\{-u_j b + v_j\})\,|\,0 \leq b \leq s_0\}$, and
- $L_3 = \left\{(b,0)\,|\,0 \leq b \leq s_0 = \min_{j=1,\ldots,n} \frac{\mathrm{spr}\boldsymbol{y}_j}{\mathrm{spr}\boldsymbol{x}_j}\right\}$, with

$$u_j = \frac{\mathrm{spr}\boldsymbol{x}_j}{|\mathrm{mid}\boldsymbol{x}_j|} \quad \text{and} \quad v_j = \frac{\mathrm{spr}\boldsymbol{y}_j}{|\mathrm{mid}\boldsymbol{x}_j|} \text{ for all } j = 1,\ldots,n.$$

The set L_2 is composed on several straight segments from some of the straight lines $\{l_j : c = -u_j b + v_j\}_{j=1}^n$. If $|\mathrm{mid}\boldsymbol{x}_j| = 0$ for any $j \in \{1,\ldots,n\}$, then the corresponding straight line is $b = \text{spr}\boldsymbol{y}_j/\text{spr}\boldsymbol{x}_j$ for $\text{spr}\boldsymbol{x}_j \neq 0$. Thus, it is a vertical line, which could take part in L_2 only if $\text{spr}\boldsymbol{y}_j/\text{spr}\boldsymbol{x}_j = s_0$. Moreover, if $\text{spr}\boldsymbol{x}_j = 0$ too, then the sample interval $\boldsymbol{x}_j$ is reduced to the real value $\boldsymbol{x}_j = 0$, so it does not take part in the construction of Γ_G. In Figure 1 the feasible set and its boundary in a practical example are illustrated graphically. The sample data corresponds to the applicative example shown in Section 4.2.

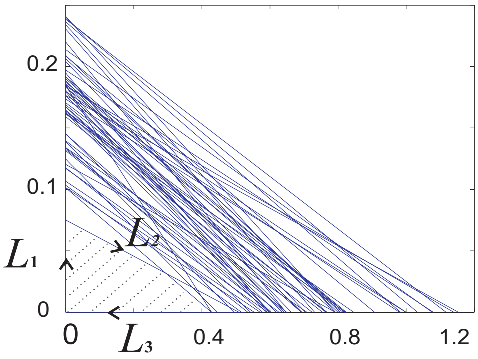

Figure 1: Γ_G for the sample data in Example 4.2.

In order to find the exact solution of $\min_{(b,c)\in\Gamma_G} g(b,c)$ the global minimum of g should be computed and, if needed, the local minimum over $L_t, t = 1,2,3$. Following this graphical approach, the iterative process to obtain the solution is detailed in Algorithm 1.

Algorithm 1

1. Compute the global minimum of g, $\widehat{\nu} = S_2^{-1} z_2$, with $z_2 = (\widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{y}}, \widehat{\sigma}_{\boldsymbol{x}^C,\boldsymbol{y}})^t$ and S_2 the sample covariance matrix of $(\boldsymbol{x}^S, \boldsymbol{x}^C)$.

   If $\widehat{\nu} \in \Gamma_G$, **then** $\widehat{\nu}$ is the solution, **else goto** Step 2.

2. Compute $s_0 = \min_{j=1,\dots,n} \text{spr}\boldsymbol{y}_j/\text{spr}\boldsymbol{x}_j$ and $r_0 = \min_{j=1,\dots,n} \text{spr}\boldsymbol{y}_j/|\text{mid}\boldsymbol{x}_j|$. Identify the straight line $l_{(v)}$ in the set $\{l_j : c = -u_j b + v_j\}_{j=1}^n$ such that $(0, r_0) \in l_{(v)}$.

If there exists more than one line in these conditions, then $l_{(v)}$ is the one for which the value $-\text{spr}\boldsymbol{y}_j/|\text{mid}\boldsymbol{x}_j|$ is lowest.

3. Let $R = \{l_{(v)}\}$, $C = \{0\}$, $D = \{(v)\}$, i=1 and $l_{(i)} = l_{(v)}$.

   If $(s_0, 0) \in l_{(v)}$, **then** let $C = \{0, s_0\}$, redefine $R = \{l^1\}, C = \{x^0, x^1\}$, let t=1 and **goto** Step 7 **else goto** Step 4.

4. Compute $(b_{(i,j)}, c_{(i,j)})$ the intersection points of $l_{(i)}$ and each line in $\{l_j : c = -u_j b + v_j\}_{j=1}^n$ such that $j \notin D$. Take the line l_{j^*} such that $b_{(i,j^*)} = \min\{b_{(i,j)} : b_{(i,j)} > C(i)\}$. If there exists more than one line in these conditions, choose as l_{j^*} the one for which the value $-\text{spr}\boldsymbol{y}_{j^*}/|\text{mid}\boldsymbol{x}_{j^*}|$ is lowest.

5. Let $R = R \cup \{l_{j*}\}$, $C = C \cup \{b_{(i,j*)}\}$, $D = D \cup \{j*\}, i = i+1, l_{(i)} = l_{j*}$.

   If $(s_0, 0) \in l_{(i)}$, **then** let $C = C \cup \{s_0\}$ and **goto** Step 6 **else goto** Step 4.

6. Redefine $R = \{l_{(v)}, l_{j_1^*}, l_{j_2^*}, \ldots, l_{j_p^*}\}$ and $C = \{0, b_{(1,j_1^*)}, b_{(j_1^*,j_2^*)}, \ldots, b_{(j_{p-1}^*,j_p^*)}, s_0\}$ as $\{l^1, l^2, l^3, \ldots, l^{t-1}, l^t\}$ and $\{x^0, x^1, x^2, \ldots, x^{t-1}, x^t\}$,respectively. **Goto** Step 7.

7. For $i = 1, \ldots, t$, compute the local minimum of g over the segment corresponding to the line l^i on $[x^{i-1}, x^i]$, given by the expressions $\begin{cases} b_*^i = \max\left\{x^{i-1}, \min\{b^i, x^i\}\right\} \\ c_*^i = -u_i b_*^i + v_i \end{cases}$

   where $b^i = \dfrac{u_i v_i \widehat{\sigma}^2_{\boldsymbol{x}^C} - v_i \widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{x}^C} - u_i \widehat{\sigma}_{\boldsymbol{x}^C,\boldsymbol{y}} + \widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{y}}}{\widehat{\sigma}^2_{\boldsymbol{x}^S} + u_i^2 \widehat{\sigma}^2_{\boldsymbol{x}^C} - 2u_i \widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{x}^C}}$.

   Compute $g(b_*^i, c_*^i)$.

   Take (b_{L_2}, c_{L_2}) the point in $\{(b_*^i, c_*^i)\}_{i=1}^t$ for which the value $g(b_*^i, c_*^i)$ is lowest. Note that (b_{L_2}, c_{L_2}) is the local minimum of g over L_2.

8. Compute (b_{L_1}, c_{L_1}) the local minimum of g over L_1, given by the expressions

$$\begin{cases} b_{L_1} = 0 \\ c_{L_1} = \max\left\{0, \min\left\{\dfrac{\widehat{\sigma}_{\boldsymbol{x}^C,\boldsymbol{y}}}{\widehat{\sigma}^2_{\boldsymbol{x}^C}}, r_0\right\}\right\} \end{cases}$$

Compute $g(b_{L_1}, c_{L_1})$.

9. Compute (b_{L_3}, c_{L_3}) the local minimum of g over L_3, given by the expressions
$$\begin{cases} b_{L_3} = \max\left\{0, \min\left\{\dfrac{\widehat{\sigma}_{\boldsymbol{x}^S,\boldsymbol{y}}}{\widehat{\sigma}^2_{\boldsymbol{x}^S}}, s_0\right\}\right\} \\ c_{L_3} = 0 \end{cases}$$
Compute $g(b_{L_3}, c_{L_3})$.

10. Take (b^*, c^*) the point in $\{(b_{L_i}, c_{L_i})\}_{i=1}^3$ whose value $g(b_{L_i}, c_{L_i})$ is lowest. Note that (b^*, c^*) is the local minimum of g on $\mathrm{fr}(\Gamma_G)$.

The worst-case computational complexity of Algorithm 1 is $O(n^2)$, meanwhile the worst-case complexity of Lemke's algorithm is $O(2^n)$. The straight lines in $\{l_j : c = -u_j b + v_j : j \neq (v), (h)\}_{j=1}^n$ such that $-u_j b_{(v,h)} + v_j > c_{(v,h)}$ do not take part on the construction of $\mathrm{fr}(\Gamma_G)$. Thus, they can be ignored from Step 4 to the end of the algorithm. However, for practical examples with moderate sample sizes n, this reduction will result in a negligible improvement on the computational efficiency of the algorithm.[2]

As happens in the multiple case, the estimates of the regression parameters do not depend on the constant θ chosen for the metric d_θ.

Expression in (3.10) jointly with the application of Algorithm 1 provide the exact solution for the LS estimation of the regression parameters of the simple model M_G (3.8). The computational complexity of these estimation method is lower than the optimization programming methods employed for the multiple model. This feature supports the application of Algorithm 1 for the estimation of the Model M_G when only one explanatory variable is involved.

[2] An implementation of Algorithm 1 for R sofware version 2.15.2 is available in http://bellman.ciencias.uniovi.es/SMIRE/Applications.html

3.4 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 $b_3 = b_4 = (0, \ldots, 0)$. Nevertheless, the extension of the basic simple model in González-Rodríguez et al. (2007) is not directly obtained from (3.1). The reason is that taking $b_1 = b_2$, and since $b_2 \geq 0$ without loss of generality in (3.1), then $b_1 \geq 0$ too. Thus, according to (3.2a), 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:

$$\boldsymbol{y} = x^t b + \boldsymbol{\varepsilon} \ , \tag{3.11}$$

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

$$\mathrm{mid}\boldsymbol{y} = \mathrm{mid}(x^t)\, b + \mathrm{mid}\, \boldsymbol{\varepsilon} \ , \tag{3.12a}$$

$$\mathrm{spr}\boldsymbol{y} = \mathrm{spr}(x^t)\, |b| + \mathrm{spr}\, \boldsymbol{\varepsilon} \ . \tag{3.12b}$$

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 (3.11). 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 $\widehat{\Delta} = \boldsymbol{y} -_H \overline{x^t\widehat{b}}$ and $\widehat{b}$ is found through the following quadratic optimization problem subject to linear constraints:

$$\min_{a\in\Gamma_1} = \|v_m - G_m\, a\| + \theta\|v_s - G_s\, a\| \ ,$$

where v_m and v_s are as in (3.7), $G_m = \mathrm{mid}X - 1^n(\overline{\mathrm{mid}X})$, $G_s = \mathrm{spr}X - 1^n(\overline{\mathrm{spr}X}) \in \mathbb{R}^{n\times k}$, $a \in \mathbb{R}^k$, and

$$\Gamma_1 = \{d \in (\mathbb{R}^k)^+ : \mathrm{spr}X\, d \leq \mathrm{spr}y\} \ .$$

Standard numerical optimization methods can be used to solve this problem.

3.5 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

$$R^2 = 1 - \frac{\sum_{j=1}^n d_\theta^2(\boldsymbol{y}_j, \widehat{\boldsymbol{y}}_j)}{\sum_{j=1}^n d_\theta^2(\boldsymbol{y}_j, \overline{\boldsymbol{y}})} \ . \tag{3.13}$$

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 (3.13) differs in general from SSR/SST.

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

$$\mathrm{MSE}_{\mathrm{model}} = \frac{\sum_{j=1}^n d_\theta^2(\boldsymbol{y}_j, \widehat{\boldsymbol{y}}_j)}{n} \ . \tag{3.14}$$

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: 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 et al., 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 a 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.[3]

The estimation of the new flexible model M_G does not depend on θ (see Remark 3.1). However, the estimated basic models recalled in Section 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 $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$

[3] The results are obtained by using the R implementation algorithms provided in http://bellman.ciencias.uniovi.es/SMIRE/Applications.html.

and an interval error $\boldsymbol{\varepsilon}$ will be considered. Let $\operatorname{mid}\boldsymbol{x}_1 \sim \mathcal{N}(1,2)$, $\operatorname{spr}\boldsymbol{x}_1 \sim \mathcal{U}(0,10)$, $\operatorname{mid}\boldsymbol{x}_2 \sim \mathcal{N}(2,1)$, $\operatorname{spr}\boldsymbol{x}_2 \sim \mathcal{X}_4^2$, $\operatorname{mid}\boldsymbol{x}_3 \sim \mathcal{N}(1,3)$, $\operatorname{spr}\boldsymbol{x}_3 \sim \mathcal{U}(0,5)$, $\operatorname{mid}\boldsymbol{\varepsilon} \sim \mathcal{N}(0,1)$ and $\operatorname{spr}\boldsymbol{\varepsilon} \sim \mathcal{X}_1^2$. Different linear expressions with the investigated structures will be considered.

- Model M_1: According to the multiple basic linear model presented in (3.11), $\boldsymbol{y}$ is defined by the expression:

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

- Model M_2: A simple linear relationship in terms of the simple model M_G is defined by considering only $\boldsymbol{x}_1$ as independent interval for modelling $\boldsymbol{y}$ through the expression:

$$\boldsymbol{y} = -2\boldsymbol{x}_1^M + 2\boldsymbol{x}_1^S + \boldsymbol{x}_1^C + 0.5\ \boldsymbol{x}_1^R + \varepsilon.$$

- Model M_3: A multiple flexible linear regression model following (3.3) 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 $s = 10000$ random samples have been generated for different sample sizes n. The estimates of the regression parameters have been computed for each iteration from their expressions in Prop. 3.1. Table 1 shows the estimated mean value and standard error of the LS estimators, computed through the corresponding analytic expressions, for $(\widehat{b}_1, \widehat{b}_4)$, and through the bootstrap algorithm shown for $(\widehat{b}_2, \widehat{b}_3)$. Besides, the estimated MSE of each estimator is computed as

$$\widehat{\mathrm{MSE}}(\widehat{b}_l) = \Big(\sum_{i=1}^{s}((\widehat{b}_l)_i - b_l)^2\Big)/s\ .$$

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	$\widehat{b}_1$	-1.9993	0.0989	0.0097	-1.9995	0.0506	0.0025	-1.9997	0.0225	0.00052
	$\widehat{b}_2$	1.9514	0.0604	0.0053	1.9723	0.0293	0.0015	1.9879	0.0125	0.00023
	$\widehat{b}_3$	0.9105	0.1399	0.0240	0.9493	0.0637	0.0054	0.9732	0.0277	0.00125
	$\widehat{b}_4$	0.5009	0.0663	0.0044	0.4997	0.0349	0.0012	0.5001	0.0155	0.00021
M_3	$\widehat{b}_1^1$	-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

The findings show 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 Figure 2 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 the models M_2 and M_3 in Figures 3 and 4, respectively.

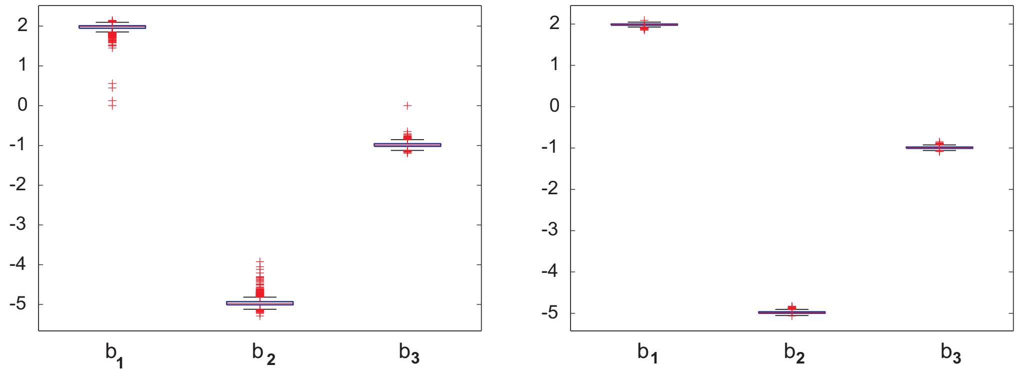

Figure 2: Box plot of the LS estimators for model M_1, n=30 (left); n=100 (right)

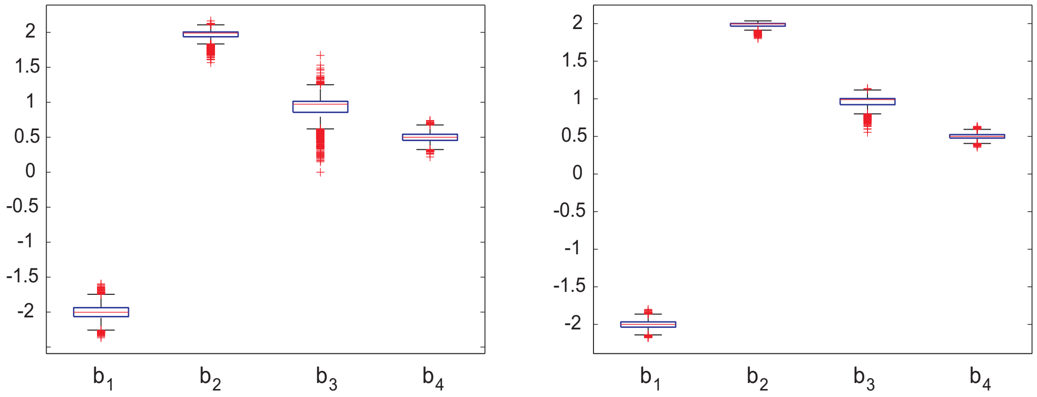

Figure 3: Box plot of the LS estimators for model M_2, n=30 (left); n=100 (right)

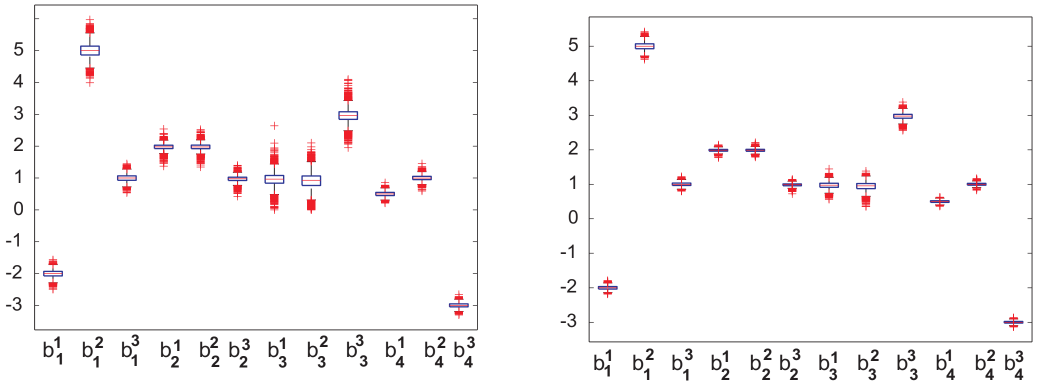

Figure 4: Box plot of the LS estimators for model M_3, n=30 (left); n=100 (right)

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 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); González-Rodríguez et al. (2007). 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 first 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. 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 simple linear model based on interval arithmetic can be formalized between the random intervals $\boldsymbol{y}$ and $\boldsymbol{x}_1$ and estimated from the available interval sample set. The estimation results for all the alternatives, both the existing methods recalled in Section 2 and the new simple model M_G introduced in Section 3.3 are gathered in Table 2, jointly with the corresponding values of R^2 and MSE.

The inclusion of the random interval $\boldsymbol{x}_2$ to model $\boldsymbol{y}$, in addition to $\boldsymbol{x}_1$, is possible by considering a multiple interval model for $(\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2)$. As before, different alternatives to estimate the linear relationship from the interval data sample could be followed. A comparison among existing methods and the new multiple interval model M_G is shown in Table 3.

Several comments can be extracted from these results. The classical procedure and the models by Lima Neto et al. (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. But 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 (3.13) and (3.14), respectively. Both in the simple and the multiple case the poorest goodness of fit corresponds to the basic interval model. This clearly shows that the condition of identical regression parameters for modelling $\mathrm{mid}\boldsymbol{y}$ and $\mathrm{spr}\boldsymbol{y}$ 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 case. 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 models. However, from these separate fitting models 30 of the 59 (in the simple case) and 29 of the 59 (in the multiple case) sample individuals do not fulfil the existence of the interval residuals; for instance, in the simple case, $\mathrm{spr}\boldsymbol{y}_1 = 19.5 < -0.0428\mathrm{mid}\boldsymbol{x}_{1,1} + 0.0366\mathrm{spr}\boldsymbol{x}_{1,1} + 29.8177 = 24.5968$. 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 et al. (2010) fail in the existence of the sample interval residuals too. The estimation procedures of the M_G models proposed here provide accurate fitting results in addition to interval coherency. It is important to recall that the formalization of the proposed models 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.

A final comment regards the comparison between the simple models and the multiple counterparts. In all the approaches it is shown that the difference in the R^2 value between the multiple and the simple case is not large, which indicates that the pulse rate has low fitting power in this application.

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. A particular algorithm is proposed, which is computationally preferable, when considering the simple model. This algorithm strongly relies on the geometry of the feasible set and it cannot be generalized into the multiple case in an easy way. The extension of the basic simple model in González-Rodríguez et al. (2007), which is not based on the canonical decomposition, requires a different approach. The solution to the estimation problem can also be obtained through KKT conditions.

The practical applicability of the proposed models 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.

Due to the large amount of regression parameters involved in the proposed models, it might be interesting to apply sparse techniques to the estimation processes in order to identify the components of the regressors which do not contribute significantly. The inclusion of robust techniques in the estimation problem to deal with the possible presence of extreme values or changes of data is also an important point to consider.

Acknowledgements

The research in this paper has been partially supported by the Spanish Government through MTM2013-44212-P Grant.

References

Billard, L. and Diday, E. (2000) Regression analysis for interval-valued data. Data Analysis, Classification and Related Methods. In Kiers, H.A.L. et al., eds., Proc. 7th Conference IFCS, 1, pages 369-374.

Blanco-Fernández, A., Corral, N., and González-Rodríguez, G. (2011). Estimation of a flexible simple linear model for interval data based on set arithmetic, *Computational Statistics & Data Analysis*, **55(9)**, 2568–2578.

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

Boukezzoula, R., Galichet, S., and Bisserier, A. (2011). A MidpointRadius approach to regression with interval data. *International Journal of Approximate Reasoning*, **52(9)**, 1257–1271.

Černý, M. and Rada, M. (2011). On the possibilistic approach to linear regression with rounded or interval-censored data. *Measurements Science Review*, **11(2)**, 34–40.

Diamond, P. (1990). Least squares fitting of compact set-valued data. *Journal of Mathematical Analysis and Applications*, **147**, 531–544.

D'Urso, P.P. (2003). Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. *Computational Statistics & Data Analysis*, **42**, 47-72.

D'Urso, P.P. and Giordani, P. (2004). A least squares approach to principal component analysis for interval valued data. *Chemometrics and Intelligent Laboratory Systems*, **70**, 179–192.

Efron, B. and Tibshirani, R. (1993) *An introduction to the Bootstrap.* , New York: Chapman & Hall.

Freedman, D.A. (1981). Bootstrapping Regression Models. *The Annals of Statistics*, **9(6)**, 1218–1228.

Gil, M.A., González-Rodríguez, G., Colubi, A., and Montenegro, M. (2007). Testing linear independence in linear models with interval-valued data. *Computational Statistics & Data Analysis*, **51**, 3002–3015.

Golub, H.G., Van Loan, C.F. (1996). *Matrix Computations.* Baltimore: Johns Hopkins University Press.

González-Rodríguez, G., Blanco, A., Corral, N., and Colubi, A. (2007). Least squares estimation of linear regression models for convex compact random sets. *Advances in Data Analysis and Classification*, **1**, 67–81.

Higham, N.J. (1996). *Accuracy and Stability of Numerical Algorithms.* Philadelphia. Society for Industrial and Applied Mathematics.

Jahanshahloo, G.R., Hosseinzadeh Lotfi, F., Rostamy Malkhalifeh, M., and Ahadzadeh Namin, M. (2008). A generalized model for data envelopment analysis with interval data. *Applied Mathematical Modelling*, **33**, 3237–3244.

Johnston, J. (1972) *Econometric Methods.* New York: McGraw-Hill Book Co.

Joly, P., Durand, C., Helmer, C, and Commenges, D. (2009). Estimating life expectancy of demented and institutionalized subjects from interval-censored observations of a multi-state model. *Statistical Modelling*, **9(4)**, pp. 345–360.

Körner, R. (1997). On the variance of fuzzy random variables. *Fuzzy Sets and Systems*, **92**, 83–93.

Lauro, C.N. and 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*, pages 173–184. Springer.

Lemke, C.E. (1962). A method of solution for quadratic programs. *Management Science*, **8(4)**, 442–453.

Liew, C.K. (1976). Inequality constrained least-squares estimation. *Journal of the American Statistical Association*, **71**, 746–751.

Lima Neto, E.A. and de Carvalho, F.A.T. (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, **54**, 333–347.

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

Sinova, B., Colubi, A., Gil, M.A., and González-Rodríguez, G. (2012). Interval arithmetic-based linear regression between interval data: Discussion and sensitivity analysis on the choice of the metric. *Information Sciences*, **199**, 109–124.

Srivastava, M.S. and Srivastava, V.K. (1986). Asymptotic distribution of least squares estimator and a test statistic in linear regression models. *Economic Letters*, **21**, 173–176.

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

Wets, R.J.B. (1991). Constrained estimation: consistency and asymptotics. *Applied stochastic models and data analysis*, **7**, 17–32.

Zhang, Z. (2009). Linear transformation models for interval-censored data: prediction of survival probability and model checking. *Statistical Modelling 2009*, **9(4)**, 321-343.

Table 2: Estimation results in Example 4.2 - Simple models

Model	Estimated interval model	Separate estimated models for *mid* and *spr* variables	R^2	$\mathrm{MSE}_{\mathrm{model}}$
Classical-midpoints	---	$\widehat{\mathrm{mid}\boldsymbol{y}} = 0.4539\mathrm{mid}\boldsymbol{x}_1 + 16.8582$	0.4317	87.82
Basic model	$\widehat{\boldsymbol{y}} = 0.4286\boldsymbol{x}_1 + [10.7167, 30.4358]$	$\mathrm{mid}\widehat{\boldsymbol{y}} = 0.4286\mathrm{mid}\boldsymbol{x}_1 + 20.5762$ $\mathrm{spr}\widehat{\boldsymbol{y}} = 0.4286\mathrm{spr}\boldsymbol{x}_1 + 9.8595$	0.4137	97.39
Model M	$\widehat{\boldsymbol{y}} = 0.4539\boldsymbol{x}_1^M + 0.2570\boldsymbol{x}_1^S$ $+[1.0164, 32.7000]$	$\mathrm{mid}\widehat{\boldsymbol{y}} = 0.4539\mathrm{mid}\boldsymbol{x}_1 + 16.8582$ $\mathrm{spr}\widehat{\boldsymbol{y}} = 0.2570\mathrm{spr}\boldsymbol{x}_1 + 15.8814$	0.4495	66.42
Lima Neto et al. model	---	$\widehat{\mathrm{mid}\boldsymbol{y}} = 0.4539\mathrm{mid}\boldsymbol{x}_1 + 16.8582$ $\widehat{\mathrm{spr}\boldsymbol{y}} = 0.6842\mathrm{spr}\boldsymbol{x}_1 + 0.9443$	0.4378	73.10
D'Urso model	---	$\mathrm{mid}\widehat{\boldsymbol{y}} = 0.5397\mathrm{mid}\boldsymbol{x}_1 - 0.4614\mathrm{spr}\boldsymbol{x}_1 + 20.3560$ $\widehat{\mathrm{spr}\boldsymbol{y}} = -0.0428\mathrm{mid}\boldsymbol{x}_1 + 0.0366\mathrm{spr}\boldsymbol{x}_1 + 29.8177$	0.4789	61.48
Simple M_G (3.8)	$\widehat{\boldsymbol{y}} = 0.5399\boldsymbol{x}_1^M + 0.2570\boldsymbol{x}_1^S$ $-0.4411\boldsymbol{x}_1^R + [3.7911, 35.4747]$	$\mathrm{mid}\widehat{\boldsymbol{y}} = 0.5399\mathrm{mid}\boldsymbol{x}_1 - 0.4411\mathrm{spr}\boldsymbol{x}_1 + 19.6329$ $\mathrm{spr}\widehat{\boldsymbol{y}} = 0.2570\mathrm{spr}\boldsymbol{x}_1 + 15.8418$	0.4994	60.08

Table 3: Estimation results in Example 4.2 - Multiple models

Model	Estimated interval model	Separate estimated models for *mid* and *spr* variables	R^2	$\mathrm{MSE}_{\mathrm{model}}$
Classical-midpoints	---	$\widehat{\mathrm{mid}\boldsymbol{y}} = 0.4497\mathrm{mid}\boldsymbol{x}_1 + 0.0517\mathrm{mid}\boldsymbol{x}_2 + 13.6263$	0.4337	86.87
Basic model (3.11)	$\widehat{\boldsymbol{y}} = 0.4094\boldsymbol{x}_1 + 0.0463\boldsymbol{x}_2$ $+[10.3630, 29.5168]$	$\mathrm{mid}\widehat{\boldsymbol{y}} = 0.4094\mathrm{mid}\boldsymbol{x}_1 + 0.0463\mathrm{mid}\boldsymbol{x}_2 + 19.9399$ $\mathrm{spr}\widehat{\boldsymbol{y}} = 0.4094\mathrm{spr}\boldsymbol{x}_1 + 0.0463\mathrm{spr}\boldsymbol{x}_2 + 9.5769$	0.4221	89.37
Lima Neto et al. model	---	$\widehat{\mathrm{mid}\boldsymbol{y}} = 0.4497\mathrm{mid}\boldsymbol{x}_1 + 0.0517\mathrm{mid}\boldsymbol{x}_2 + 13.6263$ $\widehat{\mathrm{spr}\boldsymbol{y}} = 0.4847\mathrm{spr}\boldsymbol{x}_1 + 0.3605\mathrm{spr}\boldsymbol{x}_2 + 0.4947$	0.4401	69.96
D'Urso model	---	$\widehat{\mathrm{mid}\boldsymbol{y}} = 0.5434\mathrm{mid}\boldsymbol{x}_1 + 0.0188\mathrm{mid}\boldsymbol{x}_2$ $-0.4615\mathrm{spr}\boldsymbol{x}_1 + 0.1003\mathrm{spr}\boldsymbol{x}_2 + 16.3601$ $\widehat{\mathrm{spr}\boldsymbol{y}} = -0.0357\mathrm{mid}\boldsymbol{x}_1 - 1.24 \times 10^{-3}\mathrm{mid}\boldsymbol{x}_2$ $+0.0304\mathrm{spr}\boldsymbol{x}_1 - 6.60 \times 10^{-3}\mathrm{spr}\boldsymbol{x}_2 + 29.2195$	0.4837	60.92
Multiple M_G (3.1)	$\widehat{\boldsymbol{y}} = 0.5435\boldsymbol{x}_1^M + 0.0190\boldsymbol{x}_2^M$ $+0.2588\boldsymbol{x}_1^S + 0.1685\boldsymbol{x}_2^S$ $+2.73 \times 10^{-19}\boldsymbol{x}_1^C - 0.4446\boldsymbol{x}_1^R$ $+0.1113\boldsymbol{x}_2^R + [3.2032, 27.8373]$	$\mathrm{mid}\widehat{\boldsymbol{y}} = 0.5435\mathrm{mid}\boldsymbol{x}_1 + 0.0190\mathrm{mid}\boldsymbol{x}_2$ $-0.4446\mathrm{spr}\boldsymbol{x}_1 + 0.1113\mathrm{spr}\boldsymbol{x}_2 + 15.5203$ $\mathrm{spr}\widehat{\boldsymbol{y}} = 0.2588\mathrm{spr}\boldsymbol{x}_1 + 0.1685\mathrm{spr}\boldsymbol{x}_2$ $+2.73 \times 10^{-19}\lvert\mathrm{mid}\boldsymbol{x}_1\rvert + 12.3170$	0.5083	59.02