On the Estimation of the Regression Model M for Interval Data

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

Abstract A linear regression model for interval data based on the natural interval-arithmetic has recently been proposed. Interval data can be identified with 2-dimensional points in $\mathbb{R} \times \mathbb{R}^+$, since they can be parametrized by its mid-point and its semi-amplitude or spread, which is non-negative. The model accounts separately for the contribution of the mid-points and the spreads through a single equation. The least squares estimation becomes a quadratic optimization problem subject to linear constraints, which guarantee the existence of the residuals. Several estimators are discussed. Namely, a closed-form estimator, the restricted least-squares estimator, an empirical estimator and an estimator based on separate models for mids and spreads have been investigated. Real-life examples are considered. Simulations are performed in order to assess the consistency and the bias of the estimators. Results indicate that the numerical and the closed-form estimator are appropriate in most of cases, while the empirical estimator and the one based on separate models are not always suitable.

1 Introduction

Often experimental researches involves non-perfect data, as missing data, or censored data. In particular, closed and bounded real-valued sets in $\mathbb{R}^p$ are useful to model information which also representing linguistic descriptions, fluctuations, grouped data images, to name but a few. Interval data are a specific case of this kind of elements. The study of linear regression models working with interval-valued variables has been addressed *mainly* by two ways:

[1] Department of Statistics and Operational Research, University of Oviedo, Spain, garciabarmarta.uo@uniovi.es · colubi@uniovi.es
[2] Department of Commerce, Finance and Shipping, Cyprus University of Technology, Cyprus, erricos@cut.ac.cy

C. Borgelt et al. (Eds.): Towards Advanced Data Analysis, STUDFUZZ 285, pp. 43–52.
springerlink.com

(a) in terms of the separate models involving some interval components (as the midpoint and the range or the minimum and the maximun) (see Billard and Diday, 2003; Lima Neto *et al.*, 2005 and references therein) which most of the times work with symbolic interval variables; and (b) in terms of arithmetic set-based unified models (as in Diamond 1990, Gil *et al.* 2001, 2002, 2007, González-Rodríguez *et al.* 2007, Blanco-Fernández *et al.* 2011, among others). The main difference between both views is that the first approach usually fits the separate models by numerical or classical tools, but without the usual probabilistic assumptions for the regression model. This provides good fittings but non-obvious easy ways of making inferences. On the other hand, the second approach provides a natural framework to develop inferences, although the least squares approach becomes a minimization problem with strong constraints.

In Blanco-Fernández *et al.* (2011) a flexible simple linear regression model was introduced, the so-called Model M. This model is *flexible* in the sense that it accounts for relationship between mid points and the radius of the involved random intervals. A comparison of several regression estimators of Model M will be addressed.

The rest of the paper is organized as follows: in Section 2 some preliminary about the Model M will be introduced. In Section 3 four estimation approaches of Model M will be described. In Section 4 a real-life example is analyzed to compare the behaviour of the estimators. Finally, Section 5 cointains some conclusions.

2 The Model M for Random Intervals

Hereafter, the intervals that will be considered are elements in the space $\mathcal{K}_c(\mathbb{R}) = \{[a_1, a_2] : a_1, a_2 \in \mathbb{R}, a_1 \leq a_2\}$. An interval $A \in \mathcal{K}_c(\mathbb{R})$ can be expressed in terms of its minimun and maximun or in terms or its middle point (mid) and the radius (spr). The second characterization is more usual in regression studies, as it involves non-negativity constraints which are easier to handle than the order contraints involved in the first characterization.
There is another representation for the intervals which will be used, namely, the *canonical decomposition*, defined as $A = \operatorname{mid} A\,[1 \pm 0] + \operatorname{spr} A\,[0 \pm 1]$ (see Blanco-Fernández *et al.*, 2011).

The arithmetics which will be used are the *Minkowski addition* $A + B = \{a + b : a \in A, b \in B\}$ and the product by scalars $\lambda A = \{\lambda a : a \in A\}$, with $A, B \in \mathcal{K}_c(\mathbb{R})$ and $\lambda \in \mathbb{R}$. The space $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not linear as the existence of the symmetric element with respect to the addition is not guaranteed in general, in the sense that $A + (-A) \neq \{0\}$ unless A is a singleton. A new concept of difference agreeing with the natural dif-

ference, the so-called *Hukuhara difference*, is introduced. It is defined as $A -_H B = [\inf A - \inf B, \sup A - \sup B]$ if and only if $\operatorname{spr} B \leq \operatorname{spr} A$.

Remark: If $\operatorname{spr} B > \operatorname{spr} A$, then the Hukuhara difference does not exist.

The distance used is the so-called d_τ (see Trutschnig *et al.*, 2009) defined as

$$d_\tau(A, B) = \sqrt{(1-\tau)(\operatorname{mid} A - \operatorname{mid} B)^2 + \tau\,(\operatorname{spr} A - \operatorname{spr} B)^2}$$

for all $A, B \in \mathcal{K}_c(\mathbb{R})$.
Random intervals emerged as a generalization of the real-valued random variables. Then, $\boldsymbol{y}$ is a random interval if it is $\mathcal{B}_{d_\tau}|\mathcal{A}$ measurable, being $\mathcal{B}_{d_\tau}$ the Borel σ-algebra and $\mathcal{A}$ the σ-algebra of the probabilistic space $(\Omega, \mathcal{A}, P)$.

Notation: Random intervals will be denoted with boldlowercase letters $(\boldsymbol{x})$, vectors with lowercase letters (x) and matrices with uppercase letters (X). The (Aumann) expect value is defined as $E(\boldsymbol{x}) = [E(\operatorname{mid}\boldsymbol{x}) \pm E(\operatorname{spr}\boldsymbol{x})]$, whenever $\operatorname{mid}\boldsymbol{x}$ and $\operatorname{spr}\boldsymbol{x} \in L^1(\Omega, \mathcal{A}, P)$. The Aumann expectation fulfils Fréchet principle and the Fréchet variance associated with this expectation is defined as

$$Var_\tau(\boldsymbol{x}) = \sigma^2_{\boldsymbol{x},\tau} = E(d_\tau(\boldsymbol{x}, E(\boldsymbol{x}))) = (1-\tau)\,\sigma^2_{\operatorname{mid}\boldsymbol{x}} + \tau\,\sigma^2_{\operatorname{spr}\boldsymbol{x}}$$

whenever $\operatorname{mid}\boldsymbol{x}$ and $\operatorname{spr}\boldsymbol{x}$ are integrably bounded.
As $(\mathcal{K}_c(\mathbb{R}), +, \cdot)$ is not a linear space, the covariance cannot be defined by mimiching the usual expression involving the arithmetic in $\mathcal{K}_c(\mathbb{R})$. However, it can be defined in $\mathbb{R}^2$ and we get the following expression

$$Cov_\tau(\boldsymbol{x}, \boldsymbol{y}) = \sigma_{\boldsymbol{x},\boldsymbol{y}} = (1-\tau)\sigma_{\operatorname{mid}\boldsymbol{x},\operatorname{mid}\boldsymbol{y}} + \tau\,\sigma_{\operatorname{spr}\boldsymbol{x},\operatorname{spr}\boldsymbol{y}}$$

whenever $\|\operatorname{mid}\boldsymbol{x}\|^2_\tau, \|\operatorname{mid}\boldsymbol{y}\|^2_\tau, \|\operatorname{spr}\boldsymbol{x}\|^2_\tau, \|\operatorname{spr}\boldsymbol{y}\|^2_\tau \in L^1(\Omega, \mathcal{A}, P)$.

Model M will relate a response random interval $\boldsymbol{y} : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ with an explanatory random interval $\boldsymbol{x} : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ as follows

$$\boldsymbol{y} = \boldsymbol{x}^M\,\alpha_1 + \boldsymbol{x}^S\,\alpha_2 + \varepsilon \tag{1}$$

where $\boldsymbol{x}^M = \operatorname{mid}\boldsymbol{x}[1\pm 0] = [\operatorname{mid}\boldsymbol{x}, \operatorname{mid}\boldsymbol{x}]$, $\boldsymbol{x}^S = \operatorname{spr}\boldsymbol{x}[0\pm 1] = [-\operatorname{spr}\boldsymbol{x}, \operatorname{spr}\boldsymbol{x}]$, α_1, α_2 and $\varepsilon \in \mathcal{K}_c(\mathbb{R})$ (see Blanco-Fernández et al, 2011).
The Model can be written in the matricial way as

$$\boldsymbol{y} = x^{Bl}\,b_\alpha + \varepsilon \tag{2}$$

with $x^{Bl} = (\boldsymbol{x}^M \,|\, \boldsymbol{x}^S) \in \mathcal{K}_c(\mathbb{R})^{1\times 2}$, $b_\alpha = (\alpha_1 \,|\, \alpha_2)^t \in \mathbb{R}^{2\times 1}$ and
$\varepsilon : \Omega \longrightarrow \mathcal{K}_c(\mathbb{R})$ being a random interval such that $E(\varepsilon|\boldsymbol{x}) = \Delta \in \mathcal{K}_c(\mathbb{R})$.

Remark: A property of this model is that it is not identifiable due to the fact that $\boldsymbol{x}^S = -\boldsymbol{x}^S$. However, the coefficient α_2 can be considered, without loss of generality, a non-negative vector in $\mathbb{R}$ and the space in which the

solutions to the estimation problem are, can be restricted to $\mathbb{R}^+$. In this way, the model is identifiable.

Model M entails the following separate models

$$\begin{aligned} \operatorname{mid} \boldsymbol{y} &= \alpha_1 \left(\operatorname{mid} \boldsymbol{x}\right) + \operatorname{mid} \varepsilon \\ \operatorname{spr} \boldsymbol{y} &= |\alpha_2| \left(\operatorname{spr} \boldsymbol{x}\right) + \operatorname{spr} \varepsilon. \end{aligned} \tag{3}$$

Remark: By the assumption that α_2 can be considered non-negative, the second expression can be written as

$$\operatorname{spr} \boldsymbol{y} = \alpha_2 (\operatorname{spr} \boldsymbol{x}) + \operatorname{spr} \varepsilon.$$

Thus, it is feasible to consider the estimation of α_1 and α_2 through the estimation of the separate models.

3 Estimation of the Model M

Four estimators of the regression coefficients will be considered. The first one based on the fitting of the separate models introduced in (3). Separate models have already considered to relate interval-valued variables (see Lima Neto & Carvalho 2010 among others). In this case the proposed separate models are:

$$\operatorname{mid} \boldsymbol{y} = \boldsymbol{x^c}\, b^m + \varepsilon^m \tag{4}$$

$$\operatorname{spr} \boldsymbol{y} = \boldsymbol{x^s}\, b^s + \varepsilon^s, \tag{5}$$

where $\boldsymbol{x^c} = (1, \operatorname{mid} \boldsymbol{x})$ and $\boldsymbol{x^s} = (1, \operatorname{spr} \boldsymbol{x}) \in \mathcal{K}_c(\mathbb{R})^{1\times 2}$, b^m and $b^s \in \mathbb{R}^{2\times 1}$, $\boldsymbol{y} \in \mathcal{K}_c(\mathbb{R})$ and ε^m, $\varepsilon^s \in \mathbb{R}$. Lima Neto & Carvalho impose the condition that $b^s \geq 0$ to avoid spreads ill-defined. However, b^m has no constraint to be fulfilled.

Then, let $\{\boldsymbol{y_i}, \boldsymbol{x_i}\}_{i=1,\ldots,n}$ be a random simple sample of intervals, the estimator of b^m will be:

$$\widehat{b^m} = \left[(x^c)^t \, (x^c)\right]^{-1} (x^c)^t \operatorname{mid} y \tag{6}$$

where $\operatorname{mid} y \in \mathbb{R}^{n\times 1}$ and

$$x^c = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \operatorname{mid} \boldsymbol{x_1} & \operatorname{mid} \boldsymbol{x_2} & \ldots & \operatorname{mid} \boldsymbol{x_n} \end{pmatrix} \in \mathbb{R}^{n\times 2}.$$

Parameter b^s is estimated according to Lawson and Hanson algorithm (see Lawson and Hanson, 1974) for constrained LS problems. Then the estimator of both parameters will be denoted by $\widehat{b_{sep}} = (\widehat{b^m}, \widehat{b^s})$.

Remark: The main drawback of using the separate models to estimate the coefficients is that (5) is not a linear model, due to the non-negativity constraint of the variables. Additionally, the linear independence between

the residuals and the independent variables implies further restrictions on the residuals. Thus, inferences are not straight-forward deduced.

It is possible to obtain another estimator of b_α by using sample moments. Hence, it is introduced the following proposition:

Proposition 1. *Given the random interval* $\boldsymbol{y}$ *and the vector of random intervals* x^{bl} *in the conditions of the Model* M*, the coefficients' vector* b_α *can be expressed by:*

$$b_\alpha = Cov_\tau(\boldsymbol{y}, x^{bl}) Cov_\tau(x^{bl}, x^{bl})^{-1}.$$

According to Proposition 1, an empirical estimator could be proposed based on the sample moments, namely:

$$\widehat{b_{emp}} = Cov_\tau(y, X^{bl})\, Cov_\tau(X^{bl}, X^{bl})^{-1} \tag{7}$$

with $X^{bl} \in \mathcal{K}_c(\mathbb{R})^{2\times n}$ and $y \in \mathcal{K}_c(\mathbb{R})^{1\times n}$.

The least squares estimation of b_α and the parameter Δ will be carried out from the information provided by the simple random sample of random intervals $\{\boldsymbol{y_i}, \boldsymbol{x_i}\}_{i=1,\ldots,n}$ obtained from the model:

$$y = X^{bl}\, \widehat{b_\alpha} + \hat{\varepsilon} \tag{8}$$

being

$$X^{bl} = (x^M \mid x^S) \in \mathcal{K}_c(\mathbb{R})^{n\times 2}$$

and

$$\widehat{b_\alpha} = (\hat{\alpha_1} \mid \hat{\alpha_2})^t \in \mathbb{R}^{2\times 1}.$$

It is neccesary to assure the existence of the residuals, or in other words, that the Hukuhara's difference $y -_H (X^{bl}\, \widehat{b_\alpha})$ exists. Then the expression of the constraints is:

$$\operatorname{spr}(\hat{\alpha_1}\, x^M + \hat{\alpha_2}\, x^S) \leq \operatorname{spr} y$$

which is equivalent to

$$sign(\hat{\alpha_2}) \circ |\alpha_2| \operatorname{spr} x \leq \operatorname{spr} y \equiv \hat{\alpha_2} \operatorname{spr} x \leq \operatorname{spr} y.$$

In order to assure the existance of the residuals, the least squares problem will be written as a minimization problem with linear constraints. Specifically, the aim will be to find feasible estimates of b_α and Δ minimizing the not explained variability, that is,

$$\min_{c_2 \in \Gamma} d_\tau^2(y, X^{bl}\, c + 1\Delta) \tag{9}$$

where $c = (c_1, c_2)^t \in \mathbb{R}^{2\times 1}$ and $\Gamma = \{c_2 \in [0, \infty)/c_2 \operatorname{spr} x \leq \operatorname{spr} y\}$.

Introducing the following notation, the minimization problem (9) will be transcribed into another one with some useful properties.

$$v_m = \operatorname{mid} y - \overline{\operatorname{mid} \boldsymbol{y}}1 \ ; \ F_m = \operatorname{mid} X^{bl} - (\overline{\operatorname{mid} X^{bl}})1 \qquad (10)$$
$$v_s = \operatorname{spr} y - \overline{\operatorname{spr} \boldsymbol{y}}1 \ ; \ F_s = \operatorname{spr} X^{bl} - (\overline{\operatorname{spr} X^{bl}})1,$$

where $v_m, v_s \in \mathbb{R}^{n\times 1}$ and $F_m, F_s \in \mathbb{R}^{n\times 2}$. Then, the minimization problem can be written as:

$$\min_{c_2 \in \Gamma} (1-\tau)(v_m - F_m\, c)^t (v_m - F_m\, c) + \tau\, (v_s - F_s\, c)^t (v_s - F_s\, c). \qquad (11)$$

Two possible ways of solving the problem have been proposed. The first one results in a numerical estimator and the second one in an exact expression. Concerning the first approach, as the objective function is a quadratic function and Γ is a set of linear constraints, Karush-Kuhn-Tucker (KKT) Theorem assures the existence of solution and by means of the numerical estimator, which will be denoted in the sequel by $\widehat{b_{kkt}}$, an estimation of the solution will be obtained.
On the other hand, a closed expression to estimate the regression coefficients has been obtained in Blanco-Fernández *et al.* (2011). It is given in the following proposition and will be the last one to be compared later on.

Proposition 2. *Under the conditions of Model M, the LS regression coefficients estimator is* $\widehat{b_{exact}} = (\hat{\alpha_1}, \hat{\alpha_2})$*, where:*

$$\hat{\alpha_1} = \frac{Cov(\boldsymbol{x}^M, \boldsymbol{y})}{Var(\boldsymbol{x}^M)}$$
$$\hat{\alpha_2} = \min\left\{\hat{a^0}, \max\left\{0, \frac{Cov(\boldsymbol{x}^S, \boldsymbol{y})}{Var(\boldsymbol{x}^S)}\right\}\right\}$$

being $\hat{a^0} = \min\left\{\frac{\operatorname{spr} y_i}{\operatorname{spr} x_i}\right\}\ \forall i \in \{1, \ldots, n\}$.

According to Blanco-Fernández *et al.* (2011), given $\widehat{b_\alpha}$ any estimator of b_α, it can be proved that the estimator for the residual, Δ, is:

$$\widehat{\Delta} = \overline{y} -_H \overline{X^{Bl}}\, \widehat{b_\alpha},$$

or alternatively as

$$\widehat{\Delta} = \overline{y} -_H (\overline{x^M}\, \widehat{\alpha_1} + \overline{x^S}\, \widehat{\alpha_2}).$$

Indeed, as the existence of Hukuhara's difference $y -_H X^{Bl}\, \widehat{b_\alpha}$ is guaranteed, $d^2_\tau(y, X^{Bl}\, \widehat{b_\alpha} + 1\Delta) = d^2_\tau(y -_H X^{Bl}\, \widehat{b_\alpha}, 1\Delta)$ and applying Fréchet principle, it is obtained

$$\widehat{\Delta} = \overline{y -_H (x^M\, \widehat{\alpha_1} + x^S\, \widehat{\alpha_2})} = \overline{y} -_H (\overline{x^M}\, \widehat{\alpha_1} + \overline{x^S}\, \widehat{\alpha_2}).$$

4 Applications: A Comparative Study

The first example is concerned with the relationship between the systolic and diastolic pressures in some patients in the hospital Valle del Nalón, in Asturias. The pulse rate as well as both pressure ranges along a day will be modelled by random intervals, where the endpoints of the interval are the minimum and maximum respectively. The mathematical structure will be given by $\Omega = \{3000 \text{ patients of the hospital}\}$, the Borel σ-algebra and a probability P which is uniformly distributed.

Table 1 represents the data of the sample of 56 patients. For this example the constraint $\operatorname{spr} \boldsymbol{x}\, b_\alpha \leq \operatorname{spr} \boldsymbol{y}$ is fulfilled for the 56 patients. Table 2 summarizes the estimates for α_1 and α_2. For the separate models approach, b_0^m and b_0^s refer to the real-valued intercepts while for the rest of the procedures Δ denotes the interval-valued intercept.

Table 1 $\boldsymbol{y}$: diastolic blood preassure (mmHg) and $\boldsymbol{x}$: systolic blood pressure (mmHg)

$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x}$	$\boldsymbol{y}$	$\boldsymbol{x}$	$\boldsymbol{y}$
118-173	63-102	119-212	47-93	98-160	47-108
104-161	71-118	122-178	73-105	138-221	70-118
131-186	58-113	127-189	74-125	97-154	60-107
105-157	62-118	113-213	52-112	87-152	50-95
120-179	59-94	141-205	69-133	87-150	47-86
101-194	48-116	99-169	53-109	120-188	53-105
109-174	60-119	126-191	60-98	141-256	77-158
128-210	76-125	99-201	55-121	95-166	54-100
94-145	47-104	88-221	37-94	108-147	62-107
148-201	88-130	94-176	56-121	92-172	45-107
111-192	52-96	102-156	50-94	115-196	65-117
116-201	74-133	103-159	52-95	83-140	45-91
102-167	39-84	102-185	63-118	99-172	42-86
104-161	55-98	111-199	57-113	113-176	57-95
106-167	45-95	130-180	64-121	114-186	46-103
112-162	62-116	103-161	55-97	145-210	100-136
136-201	67-122	125-192	59-101		
90-177	52-104	97-182	54-104		
116-168	58-109	100-161	54-104		
98-157	50-111	159-214	90-127		

All the estimates for α_1 are equal. However, the situation for α_2 is different. Focussing on $\widehat{b_{emp}}$ and $\widehat{b_{exact}}$, it can be seen that they are equal because the sample values fulfil the constraints to assure the existence of the residuals. However, in general, they do not need to be the same value (as shown in next example, due to the fact that $\widehat{b_{exact}}$ was defined to fulfil the constraint, whereas $\widehat{b_{emp}}$ was not). The estimate obtained from the KKT approach is the same as well, but this one was obtained by a numerical approximation. Then we can conclude that the numerical approximation is really close to the exact

Table 2 Estimations of the parameters $\alpha_1, \alpha_2, \Delta$ and b_0^m, b_0^s

Estimator	α_1	α_2	$\Delta/\mathbf{b_0^m - b_0^s}$
$\widehat{\mathbf{b_{exact}}}$	0.4539	0.2570	[1.0164,32.7000]
$\widehat{\mathbf{b_{kkt}}}$	0.4539	0.2570	[1.0164,32.7000]
$\widehat{\mathbf{b_{emp}}}$	0.4539	0.2570	[1.0164,32.7000]
$\widehat{\mathbf{b_{sep}}}$	0.4539	0.6842	16.8582-0.9443

one. $\widehat{b_{sep}}$ reaches a really high value in the estimation of α_2, which seem to denote that this estimator is not a good one, when it is applied to Model M.

The second example is concerned with the relationship between the familiar average income ($\boldsymbol{y}$) and the percentage of people with higher education ($\boldsymbol{x}$) in EEUU in 2006 (http://fact?nder.census.gov). The difference between this example and the previous one is that not all the values of the sample fulfil the constraint $\operatorname{spr} \boldsymbol{x}\, b_\alpha \leq \operatorname{spr} \boldsymbol{y}$. Table 3 displays the data of the sample of 50 people. Then, Table 4 summarizes the values of the different estimates for α_1 and α_2. Again, estimates of α_1 are equal for all the approaches. However, the estimate of α_2 is different for all the approaches excepting the exact and the KKT-based methods.

5 Conclusions

Some approaches to estimate the regression coefficients have been proposed and the comparison between them have been made by means of some examples. According to the empirical results, estimator $\widehat{b_{sep}}$ does not provide good results, which is natural, as they do not account for the specific features of the unified model that has been considered. Thus, $\widehat{b_{sep}}$ will often divert from the $\widehat{b_{exact}}$.

The performance of the empirical estimator depends on the data which has been used. If the data satisfies the constraint to assure the existence of the residuals, then the estimator is similar to the exact one. Otherwise, it is an erroneous estimator, as it provides wrong estimates for α_2, the coefficient accompanying the spreads. In any case, the estimator could be used for large samples, as it approaches to the populational parameter consistenly.
Finally, the numerical estimator $\widehat{b_{kkt}}$ is an adequate, as it reaches values which are really close to the exact ones.

Table 3 $\boldsymbol{y}$: familiar average income, $\boldsymbol{x}$: percentage of people with higher education

State	$\boldsymbol{y}$	$\boldsymbol{x}$	State	$\boldsymbol{y}$	$\boldsymbol{x}$
Alabama	48.460-49.954	7.5-7.9	Alaska	67.501-72.243	8.8-10.2
Arizona	55.063-56.355	8.9-9.5	Arkansas	44.28-45.906	5.9-6.5
California	64.150-64.976	10.3-10.5	Colorado	63.639-65.589	12.1-12.7
Connect.	77.203-79.105	14.0-14.8	Delaware	60.406-64.84	9.9-11.1
Columbia	57.076-65.134	24.4-26.4	Florida	54.043-54.847	8.8-9.0
Georgia	55.503-56.721	9.0-9.4	Hawaii	68.823-71.731	9.3-10.3
Idaho	50.612-52.668	6.8-7.4	Illinois	62.592-63.650	10.6-11.0
Indiana	55.322-56.240	7.8-8.2	Iowa	55.158-56.312	7.1-7.7
Kansas	56.159-57.555	9.5-10.1	Kentucky	48.044-49.408	8.0-8.4
Louisiana	47.467-49.055	6.6-7.0	Maine	51.820-53.766	8.5-9.3
Maryland	76.988-78.690	15.4-16.0	Massach.	73.710-75.216	15.4-15.8
Michigan	57.461-58.531	9.0-9.4	Minnesota	66.324-67.294	9.4-9.8
Mississippi	41.797-43.813	5.8-6.4	Missouri	52.465-53.587	8.5-8.9
Montana	50.177-51.835	7.8-9.0	Nebraska	56.291-57.589	8.0-8.8
Nevada	60.629-62.303	6.9-7.5	N.Hampshire	70.065-72.287	10.6-11.8
N.Jersey	77.226-78.524	12.2-12.6	N.Mexico	46.84749.551	10.5-11.3
N.York	61.774-62.502	13.2-13.4	N.Carolina	51.855-52.817	8.1-8.5
N.Dakota	53.918-56.852	5.9-7.1	Ohio	55.760-56.536	8.1-8.5
Oklahoma	47.179-48.731	7.0-7.4	Oregon	55.166-56.680	9.7-10.3
Pennsylv.	57.787-58.509	9.4-9.8	R.Island	62.762-66.704	10.7-11.9
S.Carolina	49.677-50.991	7.7-8.1	S.Dakota	52.870-54.742	6.7-7.7
Tennes.	49.240-50.368	7.3-7.7	Texas	52.080-52.630	7.9-8.1
Utah	57.306-58.976	9.0-9.8	Vermont	56.752-59.574	12.1-13.5
Virginia	66.263-67.509	12.9-13.5	Washington	63.055-64.355	10.5-10.9
W.Virginia	43.189-44.835	6.3-6.9	Wisconsin	60.172-61.096	8.2-8.6
Wyoming	55.797-59.213	6.8-8.0			

Table 4 Estimates of the parameters

Estimator	$\hat{\alpha_1}$	$\hat{\alpha_2}$	$\hat{\mathbf{\Delta}}/\mathbf{b}_0^{\hat{\mathbf{m}}} - \mathbf{b}_0^{\hat{\mathbf{s}}}$
$\widehat{\mathbf{b_{exact}}}$	2.9767	1.3817	[29.7003,30.5269]
$\widehat{\mathbf{b_{kkt}}}$	2.9767	1.3817	[29.7003,30.5269]
$\widehat{\mathbf{b_{emp}}}$	2.9767	2.3947	[30.0204,30.2068]
$\widehat{\mathbf{b_{sep}}}$	2.9767	2.6276	30.1136-0.0196

Acknowledgements This research has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-02 and the Short-Term Scientific Missions associated with the COST Action IC0702 Ref. 010611-008222. Their financial support is gratefully acknowledged. Moreover, part of the work was performed while the first two authors have been visiting the Cyprus University of Technology which partly has supported this research.

References

1. Billard L, Diday E (2003) From the Statistics of data to the Statistics of knowledge: Symbolic Data Analysis. *J Amer Stat Assoc* 98:470–487
2. Blanco-Fernández A, Corral N, González-Rodríguez G (2011) Estimation of a flexible simple linear model for interval data based on set arithmetic. *Comp Stat Data Anal* 55(9):2568–2578
3. Diamond P (1990) Least squares fitting of compact set-valued data. *J Math Anal Appl* 147:531–544
4. Gil MA, Lubiano A, Montenegro M, López-García MT (2002). Least squares fitting of an affine function and strength of association for interval-valued data. *Metrika* 56:97–111
5. González-Rodríguez G, Blanco A, Corral N, Colubi A (2007) Least squares estimation of linear regression models for convex compact random sets. *Adv D Anal Class* 1:67–81
6. Lima Neto EA, DeCarvalho FAT, Freire ES (2005) Applying constrained linear regression models to predict interval-valued data. *LNAI* 3698:92–106
7. Lima Neto EA, DeCarvalho FAT (2010) Constrained linear regression models for symbolic interval-valued variables. *Comp Stat Data Anal* 54:333–347
8. 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. *Inform Sci* 179(23):3964–3972
9. Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs Reprinted with a detailed "new developments" appendix in 1996 by SIAM Publications, Philadelphia
10. Gil MA, López MT, Lubiano MA, Montenegro M (2001) Regression and correlation analyses of a linear relation between random intervals. *Test* 10(1):183–201 (doi:10.1007/BF02595831)
11. Gil MA, González-Rodríguez G, Colubi A, Montenegro M (2007) Testing linear independence in linear models with interval-valued data. *Computational Statistics and Data Analysis* 51(6):3002–3015. (doi:10.1016/j.csda.2006.01.015)