International Journal of Approximate Reasoning 51 (2010) 759–770

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

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

A linear regression model for imprecise response

M.B. Ferraro [a,*], R. Coppi [a], G. González Rodríguez [b], A. Colubi [c]

[a] Dipartimento di Statistica, Probabilità e Statistiche Applicate, Sapienza Università di Roma, Italy
[b] European Centre for Soft Computing, Mieres, Spain
[c] Departamento de Estadística e I.O. y D.M., Universidad de Oviedo, Spain

ARTICLE INFO

Article history:
Received 25 March 2009
Received in revised form 23 March 2010
Accepted 4 April 2010
Available online 20 April 2010

Keywords:
Least-squares approach
Asymptotic distribution
LR fuzzy data
Interval data
Regression models

ABSTRACT

A linear regression model with imprecise response and *p* real explanatory variables is analyzed. The imprecision of the response variable is functionally described by means of certain kinds of fuzzy sets, the *LR* fuzzy sets. The *LR* fuzzy random variables are introduced to model usual random experiments when the characteristic observed on each result can be described with fuzzy numbers of a particular class, determined by 3 random values: the center, the left spread and the right spread. In fact, these constitute a natural generalization of the interval data. To deal with the estimation problem the space of the *LR* fuzzy numbers is proved to be isometric to a closed and convex cone of $\mathbb{R}^3$ with respect to a generalization of the most used metric for *LR* fuzzy numbers. The expression of the estimators in terms of moments is established, their limit distribution and asymptotic properties are analyzed and applied to the determination of confidence regions and hypothesis testing procedures. The results are illustrated by means of some case-studies.

1. Introduction

Different elements of a statistical problem may be imprecisely observed or defined. This has led to the development of various theories able to cope with an uncertainty which is not necessarily due to randomness: e.g. the methods based on imprecise probabilities (see, for instance, [37]), subjective probabilities (see, for instance, [34]) belief functions (see, for instance, [29]) or diverse approaches for fuzzy statistical analysis (see, for instance, [3,5] or [7]). In this paper we will consider a regression problem for a random experiment in which a fuzzy response and real-valued explanatory variables are observed. Actually, in many practical applications in public health, medical science, ecology, social or economic problems, many useful variables are vague, and the researchers find it easier to reflect the vagueness through fuzzy data than to discard it and obtain precise data. In addition, it is often less expensive to obtain an imprecise observation than to look for precise measurements of the variable of interest (see, for instance, [16]). Formally, any $[0,1]$-valued function determines a fuzzy set. However, in practice, the usual membership functions belong to some specific classes easier to fix and handle. In particular, the class of fuzzy numbers consisting of upper semi-continuous $[0,1]$-valued functions with compact support is rich enough to cover most of the applications (see, for instance [24] or [9]). However, this class is still very general, and many practitioners prefer to use simple shapes, as triangular or, slightly more general ones, as *LR*-fuzzy numbers, which are considered flexible enough to represent accurately their real-life data. For example, in agriculture quantitative soil data are unavailable over vast areas and imprecise measures, that can be modelled through *LR* fuzzy sets, are used (see [22]). Also in medical science symptoms,

[*] Corresponding author.
E-mail addresses: mariabrigida.ferraro@uniroma1.it (M.B. Ferraro), renato.coppi@uniroma1.it (R. Coppi), gil.gonzalez@softcomputing.es (G. González Rodríguez), colubi@uniovi.es (A. Colubi).

0888-613X/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2010.04.003

diagnosis and phenomena of disease may often lead to *LR* data (see, for instance, [6]). *LR*-type fuzzy data may also arise in other contexts, like image processing or artificial intelligence (see, for instance, [32,33]).

LR fuzzy sets are a generalization of intervals. Epidemiological research often entails the analysis of failure times subject to grouping, and the analysis with interval-grouped data is numerically simple and statistically meaningful (see [30,13,2]).

There are two different main lines concerning fuzzy regression problems in the literature; namely, the so-called fuzzy or possibilistic regression introduced by Tanaka [35] and widely analyzed since then (see, for instance, [36,25] and references therein) and the so-called least squares problems involving fuzzy data, also widely studied by Diamond [8], Näther [27], Krätschmer [20], González-Rodríguez et al. [15] and references therein. When the first one involves fuzzy data, an imprecise model focused on inclusion relations between actual and estimated outputs to explain the relationship is searched (see, for instance, [36]). In contrast, in the second approach, classical least squares fitting or statistical estimation problems for standard models involving fuzzy random variables are taken into consideration. Modelling this situation might be viewed as an extension of the classical error in variables models admitting measurement errors which are of nonrandom nature. From another point of view, this problem can be managed in a standard framework by means of an appropriate metric and through the concepts coherent with the space structure. Many of the above-mentioned fuzzy regression analyses have mainly focused on non probabilistic models. The only source of uncertainty accounted for in this case was the vagueness/imprecision of the data and/or of the regression parameters, and appropriate techniques of fuzzy/possibilistic analysis were utilized in this respect. Only a few papers have been devoted to regression methods able to cope with both imprecision and randomness (due to the data generation process). Among these we mention González-Rodríguez et al. [15], Näther [27] and Körner [18,19]. The present work is framed in the latter context. Specifically, a generalization of the work of Coppi et al. [4] is considered.

Coppi et al. [4] have proposed a linear regression model with crisp inputs and *LR* fuzzy response. The basic idea consists in modelling the centers of the response variable by means of a classical regression model, and simultaneously modelling the left and the right spread of the response through simple linear regressions on its estimated center. The study in Coppi et al. [4] is mainly descriptive, and the authors impose a non-negativity condition to the numerical minimization problem to avoid negative estimated spreads. In this work we propose an alternative model to overcome the non-negativity condition, because the inferences for models with non-negativity restrictions are more complex and less efficient (see, for instance, [23,12]).

The model may be looked at in the context of a multivariate regression problem. From a semantic viewpoint, it differs from the classical econometric models. In fact the equations related to the centers and the spreads jointly refer to a unique fuzzy variable, and therefore to a unique phenomenon which may be jointly affected by several (crisp) variables. Consequently, the approach we propose allows us to express and analyze the model within the context of classical multiple and multivariate regression models (see, for instance, [28]). Furthermore, unlike the model proposed by Coppi et al., in the proposed model the left and the right spread of the response are not modelled through simple linear regressions on its estimated center but by means of simple linear regressions on the explanatory variables. In this way, the parameter identification problem is avoided.

The rest of the paper is organized as follows. In Section 2 the way of modelling the imprecise response through *LR* fuzzy random variables is formalized. In Section 3 the variance of an *LR* fuzzy random variable is defined and some properties are proved. In Section 4 the new linear regression model is introduced, and the least squares estimators of the parameters are found and analyzed. Section 5 deals with asymptotic confidence regions and asymptotic hypothesis tests for the regression parameters. In Section 6 a real-life example with *LR* fuzzy data and another with interval data are illustrated. Finally, Section 7 contains some remarks and future directions.

2. Modelling the imprecise data

2.1. Fuzzy sets

A fuzzy set A of $\mathbb{R}$ may be simply defined as a mapping $A : \mathbb{R} \to [0, 1]$ verifying some conditions (see, for instance, Zadeh [39]). In practice there are some experiments whose results can be described by means of fuzzy sets of a particular class, determined by 3 values: the center, the left spread and the right spread. This type of fuzzy datum is called *LR fuzzy number* and is defined in the following way (see Fig. 1)

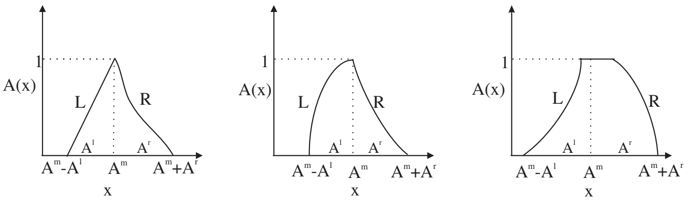

Fig. 1. Examples of *LR* membership functions.

$$A(x)=\begin{cases} L\left(\frac{A^m-x}{A^l}\right) & x\leqslant A^m, \quad A^l>0,\\ 1_{\{A^m\}}(x) & x\leqslant A^m, \quad A^l=0,\\ R\left(\frac{x-A^m}{A^r}\right) & x>A^m, \quad A^r>0,\\ 0 & x>A^m, \quad A^r=0,\end{cases}$$

where $A^m\in\mathbb{R}$ is the center, $A^l, A^r\geqslant 0$ are, respectively, the left and the right spread. $L, R:\mathbb{R}\to[0,1]$ are convex upper semi-continuous functions so that $L(0) = R(0) = 1$ and $L(z) = R(z) = 0$, for all $z\in\mathbb{R}\setminus[0,1]$, and 1_I is the indicator function of a set I.

Remark 1. An interval I is a particular kind of LR fuzzy set where the membership function is the characteristic function 1_I, that is equal to 1, for all $x\in I$, and 0 otherwise ($L = R = I_{[0,1]}$, $A^m = (\inf I + \sup I)/2$ and $A^l = A^r = (\sup I - \inf I)/2$).

Let $\mathscr{F}_{LR}$ be the class of LR fuzzy numbers. Since any $A\in\mathscr{F}_{LR}$ can be represented by means of a 3-tuple (A^m, A^l, A^r), we define the mapping $s:\mathscr{F}_{LR}\to\mathbb{R}^3$ such that $s(A) = s_A = (A^m, A^l, A^r)$.

In what follows we use without distinction $A\in\mathscr{F}_{LR}$ or its s-representation (A^m, A^l, A^r).

The natural sum and the product by a scalar in $\mathscr{F}_{LR}$ extend the Minkowski sum and the product by a positive scalar for intervals, that is, $A + \gamma B$, for $\gamma > 0$, is the fuzzy set in $\mathscr{F}_{LR}$ such that

$$(A^m,A^l,A^r)+\gamma(B^m,B^l,B^r)=(A^m+\gamma B^m,A^l+\gamma B^l,A^r+\gamma B^r).$$

The function s is obviously *semi-linear*, because $s(A) + s(B) = s(A + B)$ and $\gamma s(A) = s(\gamma A)$, if $\gamma > 0$.

Yang and Ko [38] have defined a distance D^2_{LR} between two LR fuzzy numbers $A, B\in\mathscr{F}_{LR}$ as follows:

$$\begin{aligned} D^2_{LR}(A,B)&=(A^m-B^m)^2+((A^m-\lambda A^l)-(B^m-\lambda B^l))^2+((A^m+\rho A^r)-(B^m+\rho B^r))^2\\ &=3(A^m-B^m)^2+\lambda^2(A^l-B^l)+\rho^2(A^r-B^r)^2-2\lambda(A^m-B^m)(A^l-B^l)+2\rho(A^m-B^m)(A^r-B^r), \end{aligned} \qquad (1)$$

where $\lambda=\int_0^1 L^{-1}(\omega)d\omega$ and $\rho=\int_0^1 R^{-1}(\omega)d\omega$ represent the influence of the shape of the membership function on the distance. As a result $(\mathscr{F}_{LR}, D^2_{LR})$ is a metric space.

Remark 2. In the space of LR fuzzy numbers $\mathscr{F}_{LR}$ the δ^2_2-distance (see, for details, [27]) and D^2_{LR} only differ from each others for multiplicative constants. To be more precise, the δ^2_2-distance is given by

$$\delta_2(X,Y)=\left(p\int_0^1\int_{\mathbb{S}^{p-1}}(supp_X(u,\alpha)-supp_Y(u,\alpha))^2d\mu(u)d\alpha\right)^{1/2},$$

where *supp* is a mapping that generalizes level-wise the support function. In the case of two LR fuzzy numbers $A, B\in\mathscr{F}_{LR}$ it is defined as follows:

$$\delta^2_2(A,B)=(A^m-B^m)^2+\frac{1}{2}L_2(A^l-B^l)^2+\frac{1}{2}R_2(A^l-B^l)^2-L_1(A^m-B^m)(A^l-B^l)+R_1(A^m-B^m)(A^r-B^r),$$

where $L_2=\int_0^1(L^{-1})^2(\omega)d\omega, R_2=\int_0^1(R^{-1})^2(\omega)d\omega$, $L_1=\int_0^1 L^{-1}(\omega)d\omega$ and $R_1=\int_0^1 R^{-1}(\omega)d\omega$.

The space $\mathscr{F}_{LR}$ can be embedded into $\mathbb{R}^3$ endowed with a generalization of the Yang and Ko distance, $D_{\lambda\rho}$, by preserving the metric. Furthermore, $\mathscr{F}_{LR}$ is isometric to a closed convex cone of the Hilbert space $(\mathbb{R}^3,\langle\cdot,\cdot\rangle_{\lambda\rho})$, where $\langle\cdot,\cdot\rangle_{\lambda\rho}$ is the inner product related to $D_{\lambda\rho}$.

From now on, we will consider the operation $\langle A,B\rangle_{LR} = \langle s_A,s_B\rangle_{LR}$, which is not exactly an inner product due to the lack of linearity, but it has interesting properties.

2.2. Fuzzy random variables

Kwakernaak [21], Puri & Ralescu [31] and Klement et al. [17] have introduced the concept of Fuzzy Random Variable (FRV) as an extension of both, random variables and random sets.

Let $(\Omega,\mathscr{A},P)$ be a probability space. In this context, a mapping $X:\Omega\to\mathscr{F}_{LR}$ is an FRV if the s-representation of X, $(X^m,X^l,X^r):\Omega\to\mathbb{R}\times\mathbb{R}^+\times\mathbb{R}^+$ is a random vector (see [31]). It should be noted that in our approach X is not an ill-measured real random variable but a random element assuming "purely" fuzzy values (see, also [14]).

Example 1. An example of FRVs is introduced in Colubi [3]. In a recent study about the reforestation in a given area of Asturias (Spain), carried out in the INDUROT institute (University of Oviedo), the quality of the trees has been analyzed. This characteristic has not been assigned on the basis of an underlying real-valued magnitude, but rather on the basis of subjective judgements/perceptions, through the observation of the leaf structure, the root system, the relationship height/diameter, and so on. The experts used a fuzzy-valued scale to represent their perceptions, besides linguistic labels, because the usual categorical scale (very low, low, medium, high, very high) was not able to capture the perceptions. The considered support goes from 0 (absence of quality) to 100 (perfect quality). It is possible to have different values for the same linguistic label. Some possible fuzzy values are represented in Fig. 2. This variable has been observed on 238 trees. Thus Ω = {*sets of*

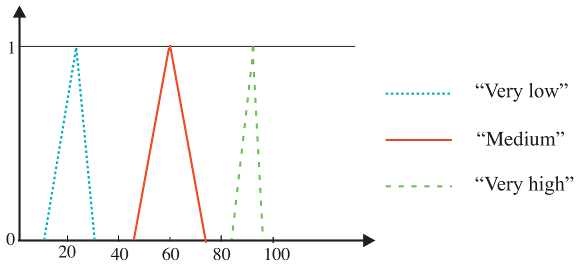

Fig. 2. Values of the "quality" of three different trees.

trees in a given area of Asturias} endowed with the Borel σ-field. Since the observations were randomly chosen, P is the uniform distribution over Ω. For any $i \in \Omega$, several characteristics are to be observed. Here we have considered the quality as an LR triangular fuzzy random variable ($\lambda = \rho = 1/2$) (see Table 1).

The expected value of an FRV is defined by means of the generalized Aumann integral [1], that is, the expected value of the FRV X is the unique fuzzy set $E(X)$ in $\mathcal{F}_{LR}$, if $E\|X\|^2_{LR} < \infty$ (see [31]). Equivalently, $E(X)$ is the fuzzy set in $\mathcal{F}_{LR}$ whose s-representation is equal to (EX^m, EX^l, EX^r).

3. The variance

The notion of variance for FRVs has been previously established in terms of several metrics (see [18,26]). By following the same ideas, we can also consider it in the sense of the D_{LR} metric.

Definition 1. The variance of an LR fuzzy random variable $X = (X^m, X^l, X^r)$ with $E\|X\|^2_{LR} < \infty$ is defined by $Var(X) = ED^2_{LR}(X, EX) = E\langle s_X - s_{EX}, s_X - s_{EX}\rangle_{LR}$.

It can be easily checked that

$$\begin{aligned} Var(X) &= E\left[3(X^m - EX^m)^2 + \lambda^2(X^l - EX^l)^2 + \rho^2(X^r - EX^r)^2\right] + E\left[-2\lambda(X^m - EX^m)(X^l - EX^l) + 2\rho(X^m - EX^m)(X^r - EX^r)\right] \\ &= 3Var(X^m) + \lambda^2 Var(X^l) + \rho^2 Var(X^r) - 2\lambda Cov(X^m, X^l) + 2\rho Cov(X^m, X^r). \end{aligned}$$

This notion of variance satisfies the same suitable properties of the usual variance in $\mathbb{R}$. In particular, it verifies the *Fréchet principle* (see [11]) because $E[D^2_{LR}(X, A)]$ is minimized for $A = EX$, which makes coherent the application of least squares techniques in regression problems.

4. Least squares estimators

Consider a random experiment in which an LR fuzzy response variable Y and p real explanatory variables $X_1, X_2, \ldots, X_p$ are observed on n statistical units, $\{Y_i, \underline{X}_i\}_{i=1,\ldots,n}$, where $\underline{X}_i = (X_{1i}, X_{2i}, \ldots, X_{pi})'$, or in a compact form $(\underline{Y}, \mathbf{X})$. Since Y is determined by (Y^m, Y^l, Y^r), the proposed regression model concerns the real-valued random variables in this tuple. The center Y^m can be related to the explanatory variables $X_1, X_2, \ldots, X_p$ through a classical regression model. However, the restriction of non-negativity satisfied by Y^l and Y^r entails some difficulties (see [4]). One solution is to consider a model with the restriction of non-negativity but, when a variable has this kind of restrictions, the errors of the model may be dependent on the explanatory variable, and the classical methods are not efficient (see, for instance, [23,12]). In addition, in presence of non-negativity restrictions most of the works in the literature are numerical procedures while in this paper the idea is to formalize a realistic theoretical model and to obtain a complete analytical solution.

We propose modelling a transformation of the left spread and a transformation of the right spread of the response through simple linear regressions (on the explanatory variables $X_1, X_2, \ldots, X_p$). The same explanatory variables have been considered for the three simultaneous equations because of the nature of the problem we have considered. Namely, to analyze how the fuzzy response variable depends on the crisp explanatory variables $X_1, X_2, \ldots, X_p$. This can be represented in the following way, letting $g : (0, +\infty) \to \mathbb{R}$ and $h : (0, +\infty) \to \mathbb{R}$ be invertible:

$$\begin{cases} Y^m = \underline{X}'\underline{a}_m + b_m + \varepsilon_m, \\ g(Y^l) = \underline{X}'\underline{a}_l + b_l + \varepsilon_l, \\ h(Y^r) = \underline{X}'\underline{a}_r + b_r + \varepsilon_r, \end{cases} \tag{2}$$

where $\varepsilon_m, \varepsilon_l$ and ε_r are real-valued random variables with $E(\varepsilon_m|\underline{X}) = E(\varepsilon_l|\underline{X}) = E(\varepsilon_r|\underline{X}) = 0$, $\underline{a}_m = (a_{m1}, \ldots, a_{mp})'$, $\underline{a}_l = (a_{l1}, \ldots, a_{lp})'$ and $\underline{a}_r = (a_{r1}, \ldots, a_{rp})'$ are the $(p \times 1)$-vectors of the parameters related to the vector $\underline{X}$. The covariance matrix of the vector of explanatory variables $\underline{X}$ will be denoted by $\Sigma_{\underline{X}}$ and Σ will stand for the covariance matrix of $(\varepsilon_m, \varepsilon_l, \varepsilon_r)$, whose variances, $\sigma^2_{\varepsilon_m}, \sigma^2_{\varepsilon_l}$ and $\sigma^2_{\varepsilon_r}$, are strictly positive and finite. Since the expected values of $\varepsilon_m, \varepsilon_l$ and ε_r given $\underline{X}$ are equal to 0 it results that $\varepsilon_m, \varepsilon_l$ and ε_r are uncorrelated with the explanatory variables.

Remark 3. The original model in Coppi et al. [4] may be stated in a general case in terms of a simultaneous equation model in the form

$$\mathbf{B}\underline{Z} = \Gamma\underline{X} + \mathbf{U},$$

where $\mathbf{B}$ and Γ are suitable matrices of coefficients and $\mathbf{U}$ is a matrix of "residual" variables, and the parameter identification problem would arise (see, for instance, [28]). However, model (2) in this formulation implies that $\underline{Z} = (Y^m, g(Y^l), h(Y^r))$ and $\mathbf{B} = \mathbf{I}$, and then this problem is overcome.

Example 2. We consider a simplification of the data introduced in Colubi [3] (see Table 1). We use the new linear regression model to analyze the part of the *quality*, Y, of *238* trees explained by the *height*, X.

In presence of constrained variables, a common approach consists in transforming the constrained variable into an unconstrained one by means of the logarithmic transformation (that is g=h=ln). We will use this approach in this example to transform the spreads into real variables without the restriction of non-negativity.

In Proposition 1 we show that the population parameters can be expressed, as usual, in terms of some moments involving the considered random variables.

Proposition 1. *Let Y be an LR fuzzy random variable and $\underline{X}$ the vector of p real random variables satisfying the linear model (2), then we have that*

$$\begin{aligned}
\underline{a}_m &= \{\Sigma_{\underline{X}}\}^{-1} E[(\underline{X} - E\underline{X})(Y^m - EY^m)],\\
\underline{a}_l &= \{\Sigma_{\underline{X}}\}^{-1} E\left[(\underline{X} - E\underline{X})(g(Y^l) - Eg(Y^l))\right],\\
\underline{a}_r &= \{\Sigma_{\underline{X}}\}^{-1} E[(\underline{X} - E\underline{X})(h(Y^r) - Eh(Y^r))],\\
b_m &= E(Y^m|\underline{X}) - E\underline{X}'\{\Sigma_{\underline{X}}\}^{-1} E[(\underline{X} - E\underline{X})(Y^m - EY^m)],\\
b_l &= E(g(Y^l)|\underline{X}) - E\underline{X}'\{\Sigma_{\underline{X}}\}^{-1} E\left[(\underline{X} - E\underline{X})(g(Y^l) - Eg(Y^l))\right],\\
b_r &= E(h(Y^r)|\underline{X}) - E\underline{X}'\{\Sigma_{\underline{X}}\}^{-1} E[(\underline{X} - E\underline{X})(h(Y^r) - Eh(Y^r))],
\end{aligned}$$

where $\Sigma_{\underline{X}} = E[(\underline{X} - E\underline{X})(\underline{X} - E\underline{X})']$.

The estimators of the population parameters will be based on the Least Squares (LS) criterion. As mentioned above, the use of this criterion is justified by the properties of the variance, among which we find the *Fréchet principle*. In addition, it should be remarked that the lack of realistic parametric models for the distribution of FRVs prevents us from using other approaches, as maximum likelihood. In this case, using the generalized Yang-Ko metric $D^2_{\lambda\rho}$ written in vector terms, the LS problem consists in looking for $\hat{\underline{a}}_m,\ \hat{\underline{a}}_l,\ \hat{\underline{a}}_r, \hat{b}_m, \hat{b}_l$ and $\hat{b}_r$ solutions of the following problem:

$$\min \Delta^2_{\lambda\rho} = \min D^2_{\lambda\rho}((\underline{Y}^m, g(\underline{Y}^l), h(\underline{Y}^r)), ((\underline{Y}^m)^*, g^*(\underline{Y}^l), h^*(\underline{Y}^r))), \tag{3}$$

where $(\underline{Y}^m)^* = \mathbf{X}\underline{a}_m + \underline{1}b_m$, $g^*(\underline{Y}^l) = \mathbf{X}\underline{a}_l + \underline{1}b_l$ and $h^*(\underline{Y}^r) = \mathbf{X}\underline{a}_r + \underline{1}b_r$ are the $(n \times 1)$-vectors of the predicted values.

Table 1
Quality (Y^m, Y^l, Y^r) and Height (X) of 238 trees in Asturias.

Y^m (center)	Y^l (left spread)	Y^r (right spread)	X (cm)
45	12.5	15	170
25	15	12.5	245
17.5	7.5	12.5	190
20	11.25	15	130
55	15	12.5	230
23.75	11.25	18.75	90
56.25	18.75	13.75	195
13.75	8.75	8.75	75
26.25	13.75	8.75	184
62.5	10	7.5	215
75	12.5	10	245
67.5	12.5	12.5	220
32.5	22.5	10	195
40	15	10	160
52.5	12.5	17.5	213
55	15	17.5	215
77.5	12.5	12.5	370
85	5	5	230
50	20	20	234
...	...	...	...

The function to minimize

$$\Delta^2_{\lambda\rho} = \|\underline{Y}^m - (\underline{Y}^m)^*\|^2 + \left\|\left(\underline{Y}^m - \lambda g(\underline{Y}^l)\right) - \left((\underline{Y}^m)^* - \lambda g^*(\underline{Y}^l)\right)\right\|^2 + \|(\underline{Y}^m + \rho h(\underline{Y}^r)) - ((\underline{Y}^m)^* + \rho h^*(\underline{Y}^r))\|^2$$

becomes

$$\begin{aligned}\Delta^2_{\lambda\rho} = {} & 3(\underline{Y}^m - \mathbf{X}\underline{a}_m - \underline{1}b_m)'(\underline{Y}^m - \mathbf{X}\underline{a}_m - \underline{1}b_m) + \lambda^2\left(g(\underline{Y}^l) - \mathbf{X}\underline{a}_l - \underline{1}b_l\right)'\left(g(\underline{Y}^l) - \mathbf{X}\underline{a}_l - \underline{1}b_l\right) \\ & + \rho^2(h(\underline{Y}^r) - \mathbf{X}\underline{a}_r - \underline{1}b_r)'(h(\underline{Y}^r) - \mathbf{X}\underline{a}_r - \underline{1}b_r) - 2\lambda(\underline{Y}^m - \mathbf{X}\underline{a}_m - \underline{1}b_m)'\left(g(\underline{Y}^l) - \mathbf{X}\underline{a}_l - \underline{1}b_l\right) \\ & + 2\rho(\underline{Y}^m - \mathbf{X}\underline{a}_m - \underline{1}b_m)'(h(\underline{Y}^r) - \mathbf{X}\underline{a}_r - \underline{1}b_r).\end{aligned} \tag{4}$$

Proposition 2. *Under the assumptions of model (2), the solutions of the LS problem are*

$$\begin{aligned}&\widehat{\underline{a}}_m = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{Y}^m},\\ &\widehat{\underline{a}}_l = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'g\widetilde{(\underline{Y}^l)},\\ &\widehat{\underline{a}}_r = (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'h\widetilde{(\underline{Y}^r)},\\ &\hat{b}_m = \overline{Y^m} - \overline{\underline{X}'}\,\hat{\underline{a}}_m,\\ &\hat{b}_l = \overline{g(Y^l)} - \overline{\underline{X}'}\,\hat{\underline{a}}_l,\\ &\hat{b}_r = \overline{h(Y^r)} - \overline{\underline{X}'}\,\hat{\underline{a}}_r,\end{aligned}$$

where, as usual, $\overline{Y^m}, \overline{g(Y^l)}, \overline{h(Y^r)}$ *and* $\overline{\underline{X}}$ *are, respectively, the sample means of* Y^m, $g(Y^l)$, $h(Y^r)$ *and* $\underline{X}$,

$$\begin{aligned}&\widetilde{\underline{Y}^m} = \underline{Y}^m - \underline{1}\overline{Y^m},\\ &g\widetilde{(\underline{Y}^l)} = g(\underline{Y}^l) - \underline{1}\overline{g(Y^l)},\\ &h\widetilde{(\underline{Y}^r)} = h(\underline{Y}^r) - \underline{1}\overline{h(Y^r)}\end{aligned}$$

are the centered values of the response and

$$\widetilde{\mathbf{X}} = \mathbf{X} - \underline{1}\overline{\underline{X}'},$$

the centered matrix of the explanatory variables.

Proposition 3. *Under the assumptions of model (2), the estimators* $\widehat{\underline{a}}_m, \widehat{\underline{a}}_l, \widehat{\underline{a}}_r, \widehat{b}_m, \widehat{b}_l$ *and* $\widehat{b}_r$ *are unbiased and strongly consistent.*

For inferential purposes it is useful to provide an approximation to the distribution of the estimators. The above-mentioned lack of realistic parametric models for the distribution of the FRVs makes it worth to look for the asymptotic distribution of the estimators.

Proposition 4. *Under the assumptions of model (2), as* $n \to \infty$,

$$\sqrt{n}\begin{pmatrix}\widehat{\underline{a}}_m - \underline{a}_m\\ \widehat{\underline{a}}_l - \underline{a}_l\\ \widehat{\underline{a}}_r - \underline{a}_r\end{pmatrix} \xrightarrow{D} N\begin{pmatrix} & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_m}\\ \underline{0}, & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_l}\\ & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_r}\end{pmatrix}. \tag{5}$$

Since the probability distribution function that has generated the data set is unknown, in practice we propose to use a bootstrap procedure to evaluate the accuracy of the estimators, by means of the estimates of the standard errors (see [10]).

5. Confidence regions and hypothesis testing on the regression parameters

In addition to the estimation of the regression parameters, the confidence regions and the hypothesis testing procedures are introduced. Starting from the asymptotic distribution (5) it is easily obtained the following $100(1-\alpha)$ confidence region for the parameters $(\underline{a}_m, \underline{a}_l, \underline{a}_r)'$

$$\left[\begin{pmatrix}\widehat{\underline{a}}_m\\ \widehat{\underline{a}}_l\\ \widehat{\underline{a}}_r\end{pmatrix} - \frac{c_{\alpha/2}}{\sqrt{n}}, \begin{pmatrix}\widehat{\underline{a}}_m\\ \widehat{\underline{a}}_l\\ \widehat{\underline{a}}_r\end{pmatrix} + \frac{c_{\alpha/2}}{\sqrt{n}}\right],$$

where $c_{\alpha/2}$ is a $\alpha/2$-quantile of a $N\begin{pmatrix} & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_m}\\ \underline{0}, & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_l}\\ & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_r}\end{pmatrix}$.

In order to test the null hypothesis H_0: $(\underline{a}_m, \underline{a}_l, \underline{a}_r)' = (\underline{k}_m, \underline{k}_l, \underline{k}_r)'$ against the alternative H_1: $(\underline{a}_m, \underline{a}_l, \underline{a}_r)' \neq (\underline{k}_m, \underline{k}_l, \underline{k}_r)'$, where $\underline{k}_m$, $\underline{k}_l$, and $\underline{k}_r$ are vectors of constant values in $\mathbb{R}$, the test statistic $T_n = V_n' V_n$, where

$$V_n = \sqrt{n}\begin{pmatrix} \widehat{\underline{a}}_m - \underline{k}_m \\ \widehat{\underline{a}}_l - \underline{k}_l \\ \widehat{\underline{a}}_r - \underline{k}_r \end{pmatrix},$$

can be used. It is possible to define a rejection region for the null hypothesis, that is:

Proposition 5. *In testing the above-defined null hypothesis at the nominal significance level* α, H_0 *should be rejected if* $T_n > c_\alpha$, *where* c_α *is a* α*-quantile of the asymptotic distribution of* T_n, *that is* $f_1(V)(V \sim N\begin{pmatrix} & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_m} \\ \underline{0}, & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_l} \\ & (\Sigma_{\underline{X}})^{-1}\sigma^2_{\varepsilon_r} \end{pmatrix}$ *and* $f_1(A) = A'A$).

The unknown population variance can be approximated by means of the sample one and Slutzky's theorem guarantees the asymptotic convergence of the standardized statistic to a normal distribution.

6. Empirical results

To illustrate the application of the regression model introduced in this work we consider the following examples.

Example 3. We consider the data of Example 1. For analyzing the part of the quality explained by the height of the trees we use the new regression model and we obtain the following estimated models

$$\begin{cases} \widehat{Y^m} = 0.1558X + 18.7497, \\ \widehat{Y^l} = \exp(-0.00017X + 2.5780), \\ \widehat{Y^r} = \exp(-0.00067X + 2.6489). \end{cases} \tag{6}$$

The value of the estimated parameter $\hat{a}_m$ equal to 0.1558 represents a positive linear relationship between the response and the explanatory variable. In particular, the quality is expected to increase of about 0.16 for any additional cm of the height.

The estimated spreads of the response variable, $\widehat{Y^l}$ and $\widehat{Y^r}$, represent the imprecision of the quality estimated by the new model. In Fig. 3 the extreme values of the 0-level and the single-value of the 1-level of the quality by the height are indicated, respectively, by means of the vertical segments and the dots, while the estimated centers and the estimated spreads are represented by the solid line and the dashed line.

To evaluate the accuracy of these estimates we drew 800 bootstrap samples of size $n = 238$ with replacement from our data set. For each bootstrap replication we calculated the estimate of the parameters of the linear regression model. By means of the 800 replications of the estimation procedure we compute the estimate of the standard errors $\widehat{se}$ of the parameters and we obtained

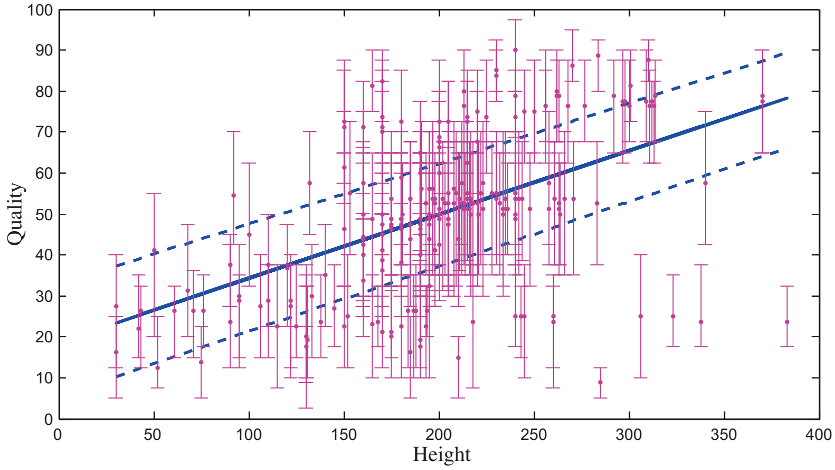

Fig. 3. The observed extreme values of the 0-level and the single-value of the quality by the height of the trees, and the estimated linear regression models.

$$\widehat{se}(\hat{a}_m) = 0.0210, \quad \widehat{se}(\hat{a}_l) = 0.0004, \quad \widehat{se}(\hat{a}_r) = 0.0004,$$
$$\widehat{se}(\hat{b}_m) = 3.9745, \quad \widehat{se}(\hat{b}_l) = 0.0821, \quad \widehat{se}(\hat{b}_r) = 0.0839.$$

Hence two kinds of uncertainty have been taken into account: the imprecision of the estimated quality and the stochastic uncertainty of the regression model represented by the above values.

To construct a confidence band for the vector of parameters (a_m, a_l, a_r), the covariance matrix of the vector (ε_m, ε_l, ε_r) was replaced by the covariance matrix of the residuals $\widehat{\varepsilon_{mi}} = \widehat{Y_i^m} - Y_i^m, \widehat{\varepsilon_{li}} = g(\widehat{Y_i^l}) - g(Y_i^l), \widehat{\varepsilon_{ri}} = h(\widehat{Y_i^r}) - h(Y_i^r)$, and the variance of the explanatory variable, σ_X^2, has been estimated by means of the sample variance $\widehat{\sigma_X}^2 = 3715.9$. A confidence band of approximate level $\alpha = 0.05$ has been found, that is,

$$\left[\begin{pmatrix} -28.9355 \\ -0.0133 \\ -0.0122 \end{pmatrix}, \begin{pmatrix} 29.2470 \\ 0.0130 \\ 0.0109 \end{pmatrix}\right].$$

When testing if the vector of regression parameters $(a_m, a_l, a_r)'$ is equal to $(0, 0, 0)'$, a p-value equal to *0* is obtained. Hence this hypothesis (related to the linear independence) should be rejected.

Example 4. In this example we are interested in analyzing the dependence relationship of the Retail Trade Sales (in millions of dollars) of the U.S. in 2002 by kind of business on the number of employees (see http://www.census.gov/econ/www/). The Retail Trade Sales are intervals in the period: January 2002 through December 2002 (see Table 2). For each interval we consider the center and the spreads and we apply the new regression model in order to evaluate the dependence relationship. As in Example 3 we transformed the spreads by means of the logarithmic transformation.

By means of the least squares estimation we obtained the following predicted values

$$\begin{cases} \widehat{Y^m} = 0.0181X - 672.731, \\ \widehat{Y^l} = \exp(0.000002482X + 5.9244), \\ \widehat{Y^r} = \exp(0.000002482X + 5.9244). \end{cases}$$

The value 0.0181 indicates the strength of the relationship between the response and the explanatory variable, in particular, the retail trade sales are expected to increase of about 18100 dollars for any additional employee.

Also in this case we evaluate the accuracy of the estimators by means of a bootstrap procedure with 800 replications. It is easy to check that

$$\widehat{se}(\hat{a}_m) = 0.0015, \quad \widehat{se}(\hat{a}_l) = 0.0000, \quad \widehat{se}(\hat{a}_r) = 0.0000,$$
$$\widehat{se}(\hat{b}_m) = 412.0407, \quad \widehat{se}(\hat{b}_l) = 0.2151, \quad \widehat{se}(\hat{b}_r) = 0.2151.$$

Table 2
The retail trade sales and the number of employees of 22 kinds of business in the U.S. in 2002.

Kind of business	Retail trade sales	Number of employees
Automotive parts, acc., and tire stores	4638–5795	453,468
Furniture stores	4054–4685	249,807
Home furnishings stores	2983–5032	285,222
Household appliance stores	1035–1387	69,168
Computer and software stores	1301–1860	73,935
Building mat. and supplies dealers	14508–20727	988,707
Hardware stores	1097–1691	142,881
Beer, wine, and liquor stores	2121–3507	133,035
Pharmacies and drug stores	11964–14741	783,392
Gasoline stations	16763–23122	926,792
Men’s clothing stores	532–1120	62,223
Family clothing stores	3596–9391	522,164
Shoe stores	1464–2485	205,067
Jewelry stores	1304–5810	148,752
Sporting goods stores	1748–3404	188,091
Book stores	968–1973	133,484
Discount dept. stores	9226–17001	762,309
Department stores	5310–14057	668,459
Warehouse clubs and superstores	13162–22089	830,845
All other gen. merchandize stores	2376–4435	263,116
Miscellaneous store retailers	7862–10975	792,361
Fuel dealers	1306–3145	98,574

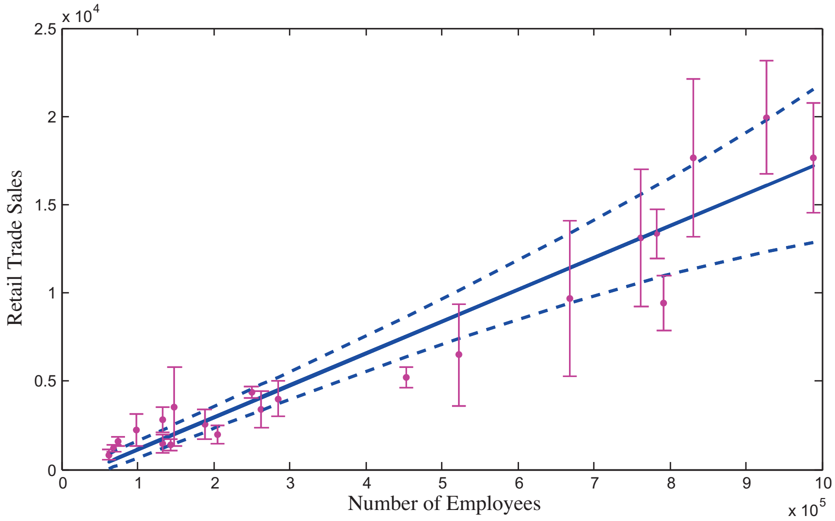

Fig. 4. The observed interval retail trade sales by number of employees and the estimated linear regression models.

The intercept term $\hat{b}_m$ is affected by a high degree of uncertainty, while the uncertainty of $\hat{a}_l$ and $\hat{a}_r$, which represent the relationship between the explanatory variable and the logarithmic transformation of the spreads of the response, is practically equal to 0.

As Fig. 4 shows, the predicted values of the spreads grow as the number of employees increases. Also in this case the null hypothesis that all the regression parameters are equal to *0* should be rejected.

7. Concluding remarks

The main objective of the present work was to provide the researcher with a viable means for analyzing regression relationships when vagueness/imprecision and randomness act jointly on the observed data.

When modelling statistical relationships between imprecise and real elements by means of classical techniques, one of the main difficulties is related to the condition of non-negativity of the spreads. In this paper by means of the introduction of the functions *g* and *h* which transform the spreads into real numbers and through an appropriate metric, we have obtained a simple solution, expressed as a function of the sample moments, which furthermore is unbiased, consistent and useful in practice.

Based on an asymptotic distribution of the parameters, confidence regions have been constructed and hypothesis testing procedures have been analyzed. We propose bootstrapping for estimating the standard errors of the estimators. Further bootstraps procedures could be also considered for interval estimation and hypothesis testing.

This new linear regression model can be used for all kinds of *LR* functional data and in particular for interval-grouped data.

The linear regression model proposed in this paper can be generalized to other useful types of random sets, e.g. trapezoidal fuzzy sets, or considering nonlinear regression.

A further field of research consists in the study of appropriate functions *g* and *h* that can be used for a wide class of practical problems, by considering the model, for instance, in a semiparametric setting.

Acknowledgements

Part of this research was carried out while Maria Brigida Ferraro worked at Departamento de Estadística e I. O. y D. M., University of Oviedo, Spain, partially supported by the COST Action IC0702. We would like to acknowledge our colleagues in the SMIRE group (http://bellman.ciencias.uniovi.es/smire/) for making possible this stay and for their suggestions to improve this paper. This research has been partially supported by the Spanish Ministry of Science and Innovation Grants MTM2009-09440-C02-01 and MTM2009-09440-C02-02, the Principality of Asturias Grants IB09-042C1 and IB09-042C2, the COST Action IC0702 and the Italian Ministry of Education, University and Research Grant PRIN 2005. Their financial support is gratefully acknowledged.

Appendix A

Proof of Proposition 1. Under the assumptions in this theorem, it can be simply checked that

$$\begin{aligned} E[(\underline{X}-E\underline{X})(Y^m-EY^m)] &= E[(\underline{X}-E\underline{X})(\underline{X}'\,\underline{a}_m+b_m+\varepsilon_m-E\underline{X}'\,\underline{a}_m+b_m+E\varepsilon_m)] \\ &= E\left[(\underline{X}-E\underline{X})(\underline{X}-E\underline{X})'\underline{a}_m+(\underline{X}-E\underline{X})(\varepsilon_m-E\varepsilon_m)\right] \\ &= E\left[(\underline{X}-E\underline{X})(\underline{X}-E\underline{X})'\underline{a}_m\right]+E[(\underline{X}-E\underline{X})(\varepsilon_m-E\varepsilon_m)]. \end{aligned}$$

Since ε_m is uncorrelated with the vector of explanatory variables $\underline{X}$, it results that

$$\underline{a}_m=\left\{\Sigma_{\underline{X}}\right\}^{-1}E[(\underline{X}-E\underline{X})(Y^m-EY^m)]$$

and

$$b_m=E(Y^m|\underline{X})-E\underline{X}'\left\{\Sigma_{\underline{X}}\right\}^{-1}E[(\underline{X}-E\underline{X})(Y^m-EY^m)].$$

Analogously, following the same reasoning we obtain the remaining expressions.

Proof of Proposition 2. In case of symmetric *LR* fuzzy variables, the least squares problem can be obviously divided into two independent parts. In general, even if in the minimization problem there are terms that consider the interaction between the center and the spreads of the response variable, by means of simple calculations, the least squares problem can be divided into three independent parts.

Proof of Proposition 3. To prove the unbiasedness of the estimators we have to analyze their expected values. Starting from $\hat{a}_m$ we have

$$E\left(\widehat{\underline{a}}_m|\underline{X}\right)=E\left[(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{Y}^m}|\underline{X}\right].$$

Since $\widetilde{\underline{Y}^m}=\widetilde{\mathbf{X}}\underline{a}_m+\widetilde{\underline{\varepsilon}}_m$, where $\widetilde{\underline{\varepsilon}}_m$ is the $(n\times 1)$-vector of the centered errors, we obtain

$$E\left(\widehat{\underline{a}}_m|\underline{X}\right)=E\left[(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\left(\widetilde{\mathbf{X}}\underline{a}_m+\widetilde{\underline{\varepsilon}}_m\right)|\underline{X}\right]=E\left[(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\underline{a}_m|\underline{X}\right]+E\left[(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_m|\underline{X}\right]$$

and, taking into account that the errors are uncorrelated with the explanatory variables, the thesis is proved, that is,

$$E\left(\widehat{\underline{a}}_m\right)=E\left[E\left(\widehat{\underline{a}}_m|\underline{X}\right)\right]=\underline{a}_m.$$

Analogously, it is possible to check that $E(\widehat{\underline{a}}_l)=\underline{a}_l$ and $E(\widehat{\underline{a}}_r)=\underline{a}_r$.

Furthermore

$$E\left(\widehat{b}_m|\underline{X}\right)=E\left(\overline{Y^m}|\underline{X}\right)-E\left(\overline{\underline{X}'}\,\hat{\underline{a}}_m|\underline{X}\right)$$

and since the sample means are unbiased estimators of the expectations, it is checked that $E(\widehat{b}_m|\underline{X})=b_m$, hence $E(\widehat{b}_m)=E\ [E(\widehat{b}_m|\underline{X})]=b_m$ and, by means of similar reasoning, it is checked the unbiasedness of $\widehat{b}_l$ and $\widehat{b}_r$.

The consistency is easily deduced from the expressions of the estimators and from the properties of population moments.

Proof of Proposition 4. Starting from the expression of $\widehat{\underline{a}}_m$, $\widehat{\underline{a}}_l$ and $\widehat{\underline{a}}_r$ in terms of sample moments

$$\begin{pmatrix}\widehat{\underline{a}}_m\\ \widehat{\underline{a}}_l\\ \widehat{\underline{a}}_r\end{pmatrix}=\begin{pmatrix}(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{Y}^m}\\ (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{g(\underline{Y}^l)}\\ (\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{h(\underline{Y}^r)}\end{pmatrix}$$

and taking into account that $\widetilde{\underline{Y}^m}=\widetilde{\mathbf{X}}\underline{a}_m+\widetilde{\underline{\varepsilon}}_m,\widetilde{g(\underline{Y}^l)}=\widetilde{\mathbf{X}}\underline{a}_l+\widetilde{\underline{\varepsilon}}_l$ and $\widetilde{h(\underline{Y}^r)}=\widetilde{\mathbf{X}}\underline{a}_r+\widetilde{\underline{\varepsilon}}_r$, it is easy to check that

$$\begin{pmatrix}\widehat{\underline{a}}_m\\ \widehat{\underline{a}}_l\\ \widehat{\underline{a}}_r\end{pmatrix}=\begin{pmatrix}\underline{a}_m+(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_m\\ \underline{a}_l+(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_l\\ \underline{a}_r+(\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}})^{-1}\widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_r\end{pmatrix}.$$

In this way, we have that

$$\sqrt{n}\begin{pmatrix}\widehat{\underline{a}}_m-\underline{a}_m\\ \widehat{\underline{a}}_l-\underline{a}_l\\ \widehat{\underline{a}}_r-\underline{a}_r\end{pmatrix}=\left((\underline{1}'\underline{1})^{-1}\widetilde{\mathbf{X}}'\widetilde{\mathbf{X}}\right)^{-1}(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}\widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_m\\ \widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_l\\ \widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_r\end{pmatrix}$$

and then,

$$\begin{aligned}(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}\widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_m\\ \widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_l\\ \widetilde{\mathbf{X}}'\widetilde{\underline{\varepsilon}}_r\end{pmatrix}&=(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}(\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_m\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_l\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_m\end{pmatrix}\\ &+(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}((\underline{1}E\underline{X}')'-(\underline{1}\overline{\underline{X}'})')\underline{\varepsilon}_m\\ ((\underline{1}E\underline{X}')'-(\underline{1}\overline{\underline{X}'})')\underline{\varepsilon}_l\\ ((\underline{1}E\underline{X}')'-(\underline{1}\overline{\underline{X}'})')\underline{\varepsilon}_m\end{pmatrix}\\ &-(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}(\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{1}\overline{\varepsilon_m}\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{1}\overline{\varepsilon_l}\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{1}\overline{\varepsilon_r}\end{pmatrix}\\ &-(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}((\underline{1}E\underline{X}')'-(\underline{1}\overline{\underline{X}'})')\underline{1}\overline{\varepsilon_m}\\ ((\underline{1}E\underline{X}')'-(\underline{1}\overline{\underline{X}'})')\underline{1}\overline{\varepsilon_l}\\ ((\underline{1}E\underline{X}')'-(\underline{1}\overline{\underline{X}'})')\underline{1}\overline{\varepsilon_r}\end{pmatrix}.\end{aligned}$$

Furthermore, as $n\to\infty$, the last three terms of the sum tend almost surely to $\underline{0}$ ($(3p\times 1)$-null vector) and

$$\left\{\begin{pmatrix}(\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_m\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_l\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_r\end{pmatrix}\right\}$$

is a sequence of random vectors i.i.d., centered at $\underline{0}$, whose covariance matrix is $\Sigma_{\underline{X}}\Sigma$, so applying the Central Limit Theorem it results that

$$(\underline{1}'\underline{1})^{-1/2}\begin{pmatrix}(\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_m\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_l\\ (\mathbf{X}'-(\underline{1}E\underline{X}')')\underline{\varepsilon}_r\end{pmatrix}\xrightarrow{D}N\begin{pmatrix} & \Sigma_{\underline{X}}\sigma^2_{\varepsilon_m}\\ \underline{0}, & \Sigma_{\underline{X}}\sigma^2_{\varepsilon_l}\\ & \Sigma_{\underline{X}}\sigma^2_{\varepsilon_r}\end{pmatrix}.$$

Hence

$$\sqrt{n}\begin{pmatrix}\widehat{\underline{a}}_m-\underline{a}_m\\ \widehat{\underline{a}}_l-\underline{a}_l\\ \widehat{\underline{a}}_r-\underline{a}_r\end{pmatrix}\xrightarrow{D}N\begin{pmatrix} & \left(\Sigma_{\underline{X}}\right)^{-1}\sigma^2_{\varepsilon_m}\\ \underline{0}, & \left(\Sigma_{\underline{X}}\right)^{-1}\sigma^2_{\varepsilon_l}\\ & \left(\Sigma_{\underline{X}}\right)^{-1}\sigma^2_{\varepsilon_r}\end{pmatrix}.$$

References

[1] R.J. Aumann, Integrals of set-valued functions, J. Math. Anal. Appl. 12 (1965) 1–12.

[2] L. Billard, E. Diday, From statistics of data to the statistics of knowledge: symbolic data analysis, J. Amer. Stat. Assoc. 98 (2003) 470–487.

[3] A. Colubi, Statistical inference about the means of fuzzy random variables: applications to the analysis of fuzzy- and real-valued data, Fuzzy Sets Syst. 160 (2009) 344–356.

[4] R. Coppi, P. D'Urso, P. Giordani, A. Santoro, Least squares estimation of a linear regression model with *LR* fuzzy response, Comp. Stat. Data Anal. 51 (2006) 267–286.

[5] R. Coppi, Management of uncertainty in statistical reasoning: the case of regression analysis, Int. J. Approx. Reason. 47 (2008) 284–305.

[6] L. Di Lascio, L. Ginolfi, A. Albunia, G. Galardi, F. Meschi, A fuzzy-based methodology for the analysis of diabetic neuropathy, Fuzzy Sets Syst. 129 (2002) 203–228.

[7] T. Denoeux, M.H. Masson, P.A. Hébert, Nonparametric rank-based statistics and significance tests for fuzzy data, Fuzzy Sets Syst. 153 (2005) 1–28.

[8] P. Diamond, Fuzzy least squares, Inform. Sci. 46 (1988) 141–157.

[9] D. Dubois, M.A. Lubiano, H. Prade, M.A. Gil, P. Grzegorzewski, O. Hryniewicz, Soft methods for handling variability and imprecision, Advances in Soft Computing, Springer, Berlin, 2008.

[10] B. Efron, R.J. Tibshirani, An Introduction to The Bootstrap, Chapman & Hall, New York, 1993.

[11] M. Fréchet, Les éléments aléatoires de natures quelconque dans un áspace distancié, Ann. Inst. H. Poincaré 10 (1948) 215–310.

[12] A.R. Gallant, T.M. Gerig, Computations for constrained linear models, J. Econom. 12 (1980) 59–89.

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

[14] G. González-Rodríguez, A. Colubi, P. D'Urso, M. Montenegro, Multi-sample test-based clustering for fuzzy random variables, Int. J. Approx. Reason. 50 (2009) 721–731.

[15] G. González-Rodríguez, A. Blanco, M.A. Lubiano, Estimation of a simple linear regression model for fuzzy random variables, Fuzzy Sets Syst. 160 (2009) 357–370.
[16] P.J. Heagerty, S.R. Lele, A composite likelihood approach to binary spatial data, J. Amer. Stat. Assoc. 93 (1998) 1099–1111.
[17] E. Klement, M.L. Puri, D.A. Ralescu, Limit theorems for fuzzy random variables, Proc. Roy. Soc. London Ser. A 1832 (1986) 171–182.
[18] R. Körner, Linear Models with Random Fuzzy Variables, Ph.D Thesis, Faculty of Mathematics and Computer Science, Freiberg University of Mining and Technology, 1997.
[19] R. Körner, On the variance of fuzzy random variables, Fuzzy Sets Syst. 92 (1997) 83–93.
[20] V. Krätschmer, Limit distribution of least squares estimators in linear regression models with vague concepts, J. Multivariate Anal. 97 (2006) 1044–1069.
[21] H. Kwakernaak, Fuzzy random variables-I. Definitions and theorems, Inform. Sci. 15 (1978) 1–29.
[22] P. Lagacherie, D.R. Cazemier, R. Martin-Clouaire, T. Wassenaar, A spatial approach using imprecise soil data for modelling crop yields over vast areas, Agric. Ecosyst. Environ. 81 (2000) 5–16.
[23] C.K. Liew, Inequality constrained least-squares estimation, J. Amer. Stat. Assoc. 71 (1976) 746–751.
[24] M. López-Díaz, M.A. Gil, P. Grzegorzewski, O. Hryniewicz, J. Lawry, Soft methodology and random information systems, Advances in Soft Computing, Springer-Verlag, Berlin, 2004.
[25] J. Lu, R. Wang, An enhanced fuzzy linear regression model with more flexible spreads, Fuzzy Sets Syst. 160 (2009) 2505–2523.
[26] M.A. Lubiano, M.A. Gil, M. López-Díaz, M.T. López, The $\vec{\lambda}$-mean squared dispersion associated with a fuzzy random variable, Fuzzy Sets Syst. 111 (2000) 307–317.
[27] W. Näther, Regression with fuzzy random data, Comp. Stat. Data Anal. 51 (2006) 235–252.
[28] F. Peracchi, Econometrics, Wiley, 2001.
[29] S. Petit-Renaud, T. Denoeux, Nonparametric regression analysis of uncertain and imprecise data using belief functions, Int. J. Approx. Reason. 35 (2004) 1–28.
[30] C.B. Pipper, C. Ritz, Checking the grouped data version of the Cox model for interval-grouped survival data, Scand. J. Stat. 34 (2007) 405–418.
[31] M.L. Puri, D.A. Ralescu, Fuzzy random variables, J. Math. Anal. Appl. 114 (1986) 409–422.
[32] J. Ranilla, L.J. Rodríguez-Muñiz, A heuristic approach to learning rules from fuzzy database, IEEE Intell. Syst. 22 (2007) 62–68.
[33] M. Sezgin, B. Sankur, Survey over image thresholding techniques and quantitative performance evaluation, J. Electron. Imaging 13 (2004) 146–168.
[34] N.D. Singpurwalla, J.M. Booker, Membership functions and probability measures of fuzzy sets, J. Amer. Stat. Assoc. 99 (2004) 867–877.
[35] H. Tanaka, Fuzzy data analysis by possibilistic linear models, Fuzzy Sets Syst. 24 (1987) 363–375.
[36] H. Tanaka, P. Guo, Possibilistic regression analysis, in: D.A. Ralescu, B. Bertoluzza, M.A. Gil (Eds.), Statistical Modeling Analysis and Management of Fuzzy Data, Physica-Verlag, Wurzburg, 2002, pp. 239–254.
[37] P. Walley, A bounded derivative model for prior ignorance about a real-valued parameter, Scand. J. Stat. 24 (1996) 463–483.
[38] M.S. Yang, C.H. Ko, On a class of fuzzy *c*-numbers clustering procedures for fuzzy data, Fuzzy Sets Syst. 84 (1996) 49–60.
[39] L.A. Zadeh, Fuzzy sets, Inform. Control 8 (1965) 338–353.