

A class of linear regression models for imprecise random elements

Renato Coppi, Maria Brigida Ferraro and Paolo Giordani

Abstract The linear regression model of a fuzzy response variable on a set of real and/or fuzzy explanatory variables is investigated. The notion of *LR* fuzzy random variable is introduced in this connection, leading to the probabilization of the center and the left and right spreads of the response variable. A specific metric is suggested for coping with this type of variables. A class of linear regression models is then proposed for the center and for suitable transforms of the spreads in order to satisfy the non-negativity conditions for the latter ones. A Least Squares solution for estimating the parameters of the models is derived, along with a goodness of fit measure and hypothesis tests. The results of a real-life application are finally illustrated.

Key words: Fuzzy random variables, Imprecise data and fuzzy methods, Linear regression analysis

1 Introduction

In the literature a great deal of attention has been paid to the management of uncertain information. We can roughly distinguish two sources of uncertainty, namely, randomness and imprecision. In the case of randomness the information is uncertain because we do not know the (precise) outcome of a (random) mechanism. In general, randomness is limited to the data generation process and it can be dealt with by means of probability theory (probabilistic uncertainty). In contrast with random-

Renato Coppi
SAPIENZA Università di Roma, P.le A. Moro 5, 00185, Roma e-mail: renato.coppi@uniroma1.it

Maria Brigida Ferraro
SAPIENZA Università di Roma, P.le A. Moro 5, 00185, Roma e-mail: mariabrigida.ferraro@uniroma1.it

Paolo Giordani
SAPIENZA Università di Roma, P.le A. Moro 5, 00185, Roma e-mail: paolo.giordani@uniroma1.it

ness, imprecision is connected to the uncertainty concerning the placement of an outcome in a given class and, thus, it can be seen as non-probabilistic uncertainty [1].

The different sources of uncertainty are not exclusive but can occur together. A possible way to cope with imprecision is represented by fuzzy set theory [8]. This allows us to express the imprecise information in terms of fuzzy sets. When such information is also affected by randomness, the concept of fuzzy random variable (FRV) can be adopted [5, 6]. In this work we aim at investigating the linear regression problem for a very common class of FRVs, namely the *LR* family.

2 Preliminaries

Given a universe U of elements, a fuzzy set $\tilde{A}$ is defined through the so-called *membership function* $\mu_{\tilde{A}} : U \rightarrow [0, 1]$. For a generic $x \in U$, the membership function expresses the extent to which x belongs to $\tilde{A}$. Such a degree ranges from 0 (complete non-membership) to 1 (complete membership).

A particular class of fuzzy sets is the LR family, whose members are the so-called *LR fuzzy numbers*. The space of the LR fuzzy numbers is denoted by $\mathcal{F}_{LR}$. A nice property of the LR family is that its elements can be determined uniquely in terms of the mapping $s : \mathcal{F}_{LR} \rightarrow \mathbb{R}^3$, i.e., $s(\tilde{A}) = s_{\tilde{A}} = (A^m, A^l, A^r)$. This implies that $\tilde{A}$ can be expressed by means of three real-valued parameters, namely, the center (A^m) and the (non-negative) left and right spreads (A^l and A^r , respectively). In what follows it is indistinctly used $\tilde{A} \in \mathcal{F}_{LR}$ or (A^m, A^l, A^r) . The membership function of $\tilde{A} \in \mathcal{F}_{LR}$ can be written as

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

where the functions L and R are particular decreasing shape functions from $\mathbb{R}^+$ to $[0, 1]$ such that $L(0) = R(0) = 1$ and $L(x) = R(x) = 0, \forall x \in \mathbb{R} \setminus [0, 1]$, and 1_I is the indicator function of a set I . $\tilde{A}$ is a *triangular* fuzzy number if $L(z) = R(z) = 1 - z$, for $0 \leq z \leq 1$.

The operations considered in $\mathcal{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for intervals. Going into detail, the sum of $\tilde{A}$ and $\tilde{B}$ in $\mathcal{F}_{LR}$ is the LR fuzzy number $\tilde{A} + \tilde{B}$ so that

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

and the product of $\tilde{A} \in \mathcal{F}_{LR}$ by a positive scalar γ is

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

Yang & Ko [7] define a distance between two LR fuzzy numbers $\tilde{X}$ and $\tilde{Y}$ as follows

$$D_{LR}^2(\tilde{X}, \tilde{Y}) = (X^m - Y^m)^2 + [(X^m - \lambda X^l) - (Y^m - \lambda Y^l)]^2 + [(X^m + \rho X^r) - (Y^m + \rho Y^r)]^2,$$

where the parameters $\lambda = \int_0^1 L^{-1}(\omega) d\omega$ and $\rho = \int_0^1 R^{-1}(\omega) d\omega$ play the role of taking into account the shape of the membership function. For instance, in the triangular case, it is $\lambda = \rho = \frac{1}{2}$. For what follows it is necessary to embed the space $\mathcal{F}_{LR}$ into $\mathbb{R}^3$ by preserving the metric. For this reason a generalization of the Yang and Ko metric can be derived [2]. Given $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3) \in \mathbb{R}^3$, it is

$$D_{\lambda\rho}^2(a, b) = (a_1 - b_1)^2 + ((a_1 - \lambda a_2) - (b_1 - \lambda b_2))^2 + ((a_1 + \rho a_3) - (b_1 + \rho b_3))^2,$$

where $\lambda, \rho \in \mathbb{R}^+$. $D_{\lambda\rho}^2$ will be used in the sequel as a tool for quantifying errors in the regression models we are going to introduce.

Let $(\Omega, \mathcal{A}, P)$ be a probability space. In this context, a mapping $\tilde{X} : \Omega \rightarrow \mathcal{F}_{LR}$ is an LR FRV if the s -representation of $\tilde{X}$, $(X^m, X^l, X^r) : \Omega \rightarrow \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ is a random vector [6]. As for non-fuzzy random variables, it is possible to determine the moments for an LR FRV. The expectation of an LR FRV $\tilde{X}$ is the unique fuzzy set $E(\tilde{X}) (\in \mathcal{F}_{LR})$ such that $(E(\tilde{X}))_\alpha = E(X_\alpha)$ provided that $E\|\tilde{X}\|_{D_{LR}}^2 = E(X^m)^2 + E(X^m - \lambda X^l)^2 + E(X^m + \rho X^r)^2 < \infty$, where X_α is the α -level of fuzzy set $\tilde{X}$, that is, $X_\alpha = \{x \in \mathbb{R} | \mu_{\tilde{X}}(x) \geq \alpha\}$, for $\alpha \in (0, 1]$, and $X_0 = cl(\{x \in \mathbb{R} | \mu_{\tilde{X}} \geq 0\})$. Also, on the basis of the mapping s , we can observe that $s_{E(\tilde{X})} = (E(X^m), E(X^l), E(X^r))$. The variance of $\tilde{X}$ can be defined as $\sigma_{\tilde{X}}^2 = var(\tilde{X}) = E[(D_{LR}^2(\tilde{X}, E(\tilde{X})))]$. In terms of s , it is $\sigma_{\tilde{X}}^2 = var(\tilde{X}) = E \langle s_{\tilde{X}} - s_{E(\tilde{X})}, s_{\tilde{X}} - s_{E(\tilde{X})} \rangle_{LR} = E(X^m - EX^m)^2 + E(X^m - EX^m - \lambda(X^l - EX^l))^2 + E(X^m - EX^m + \rho(X^r - EX^r))^2$. In a similar way, the covariance between two FRVs $\tilde{X}$ and $\tilde{Y}$ is $\sigma_{\tilde{X}, \tilde{Y}} = cov(\tilde{X}, \tilde{Y}) = E \langle s_{\tilde{X}} - s_{E(\tilde{X})}, s_{\tilde{Y}} - s_{E(\tilde{Y})} \rangle_{LR} = E((X^m - EX^m)(Y^m - EY^m)) + E((X^m - EX^m - \lambda(X^l - EX^l))(Y^m - EY^m - \lambda(Y^l - EY^l))) + E((X^m - EX^m + \rho(X^r - EX^r))(Y^m - EY^m + \rho(Y^r - EY^r)))$.

3 A regression model with LR fuzzy random response and real explanatory variables

Consider a random experiment in which an LR fuzzy response variable $\tilde{Y}$ and p real explanatory variables $X_1, X_2, \dots, X_p$ are observed on n statistical units, $\{\tilde{Y}_i, \underline{X}_i\}_{i=1, \dots, n}$, where $\underline{X}_i = (X_{1i}, X_{2i}, \dots, X_{pi})'$, or in a compact form $(\tilde{Y}, \mathbf{X})$. Since $\tilde{Y}$ 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 $X_1, X_2, \dots, X_p$ through a classical regression model. However, the restriction of non-negativity satisfied by Y^l and Y^r may entail some difficulties. We thus propose

modelling a transform of the left spread and a transform of the right spread of the response through linear regressions (on the explanatory variables $X_1, X_2, \dots, X_p$). This can be represented in the following way, letting $g : (0, +\infty) \rightarrow \mathbb{R}$ and $h : (0, +\infty) \rightarrow \mathbb{R}$ be invertible:

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

where $\varepsilon_m, \varepsilon_l$ and ε_r are real-valued random variables with $E(\varepsilon_m|\underline{X}) = E(\varepsilon_l|\underline{X}) = E(\varepsilon_r|\underline{X}) = 0$, $\underline{a}_m = (a_{m1}, \dots, a_{mp})'$, $\underline{a}_l = (a_{l1}, \dots, a_{lp})'$, $\underline{a}_r = (a_{r1}, \dots, a_{rp})'$ are the $(p \times 1)$ -vectors of the parameters related to the vector $\underline{X}$ and b_m, b_l, b_r are the intercepts. 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 $g=h=\log$). We use this approach in this example to transform the spreads into real variables without the restriction of non-negativity.

It can be shown that the population parameters can be expressed, as usual, in terms of some moments involving the considered random variables. In fact, let $\tilde{Y}$ 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

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

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

The estimators of the population parameters are based on the Least Squares (LS) criterion. In this case, using the generalized Yang-Ko metric $D_{\lambda\rho}^2$ 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 $\hat{b}_r$ in order to

$$\min \Delta_{\lambda\rho}^2 = \min D_{\lambda\rho}^2((\underline{Y}^m, g(\underline{Y}^l), h(\underline{Y}^r)), ((\underline{Y}^m)^*, g^*(\underline{Y}^l), h^*(\underline{Y}^r))) \quad (3)$$

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

Under the assumptions of model (2), the solution of the LS problem is

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

where $\underline{Y}^{mc} = \underline{Y}^m - \underline{1}\overline{Y^m}$, $g(\underline{Y}^l)^c = g(\underline{Y}^l) - \underline{1}g(\overline{Y^l})$, $h(\underline{Y}^r)^c = h(\underline{Y}^r) - \underline{1}h(\overline{Y^r})$ are the centered values of the response variables, $\underline{X}^c = \underline{X} - \underline{1}\overline{X}$ is the centered matrix of the explanatory variables and, $\overline{Y^m}$, $\overline{g(Y^l)}$, $\overline{h(Y^r)}$ and $\overline{X}$ denote, respectively, the sample means of Y^m , $g(Y^l)$, $h(Y^r)$ and $\underline{X}$. The estimators $\hat{a}_m$, $\hat{a}_l$, $\hat{a}_r$, $\hat{b}_m$, $\hat{b}_l$ and $\hat{b}_r$ are unbiased and strongly consistent. It is worth mentioning that the significance of the regression parameters can be tested. Since there are not generalized models for FRVs that can be used in practice and asymptotic tests work suitably for large size samples, a bootstrap approach is adopted. For more details, see [2].

4 A regression model with LR fuzzy random response and explanatory variables

Model (2) can be generalized to the case in which the available information refers to an LR fuzzy response variable $\tilde{Y}$ and p LR fuzzy explanatory variables $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$ observed on a sample of n statistical units, $\{\tilde{Y}_i, \tilde{X}_{1i}, \tilde{X}_{2i}, \dots, \tilde{X}_{pi}\}_{i=1, \dots, n}$. We are interested in analyzing the relationship between $\tilde{Y}$ and $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$. This can be done by modelling the center and the spreads of $\tilde{Y}$ by means of the centers and the spreads of $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$. However, in doing so, attention should be paid to the non-negativity of the spreads of $\tilde{Y}$. Analogously to the previous model, we introduce the functions g and h already defined. The linear regression model can be formalized as

$$\begin{cases} Y^m = \underline{X}' \underline{a}_m + b_m + \varepsilon_m, \\ g(Y^l) = \underline{X}' \underline{a}_l + b_l + \varepsilon_l, \\ h(Y^r) = \underline{X}' \underline{a}_r + b_r + \varepsilon_r, \end{cases} \quad (4)$$

where $\underline{X} = (X_1^m, X_1^l, X_1^r, \dots, X_p^m, X_p^l, X_p^r)'$ is the $(3p \times 1)$ -vector of all the components of the explanatory variables, ε_m , ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|\underline{X}) = E(\varepsilon_l|\underline{X}) = E(\varepsilon_r|\underline{X}) = 0$, $\underline{a}_m = (a_{mm}^1, a_{ml}^1, a_{mr}^1, \dots, a_{mm}^p, a_{ml}^p, a_{mr}^p)'$, $\underline{a}_l = (a_{lm}^1, a_{ll}^1, a_{lr}^1, \dots, a_{lm}^p, a_{ll}^p, a_{lr}^p)'$, $\underline{a}_r = (a_{rm}^1, a_{rl}^1, a_{rr}^1, \dots, a_{rm}^p, a_{rl}^p, a_{rr}^p)'$ are the $(3p \times 1)$ -vectors of the parameters related to $\underline{X}$ and b_m , b_l , b_r are the intercepts. The generic a_{ij}^t is the regression coefficient between the component $i \in \{m, l, r\}$ of $\tilde{Y}$ (where m , l and r refer to the center Y^m and the transforms of the spreads $g(Y^l)$ and $h(Y^r)$, respectively) and the component $j \in \{m, l, r\}$ of the explanatory variables $\tilde{X}^t$, $t = 1, \dots, p$, (where m , l and r refer to the corresponding center, left spread and right spread). For example, a_{mr}^3 represents the relationship between the right spread of the explanatory variable $\tilde{X}^3$ (X_3^r) and the center of the response, Y^m .

Also in this case, the estimation problem of the regression parameters is faced by means of the LS criterion. The LS problem consists in looking for $\hat{a}_m$, $\hat{a}_l$, $\hat{a}_r$, $\hat{b}_m$, $\hat{b}_l$ and $\hat{b}_r$ such that

$$\Delta_{\lambda p}^2 = D_{\lambda p}^2((Y^m, g(Y^l), h(Y^r)), ((Y^m)^*, g^*(Y^l), h^*(Y^r)))$$

is minimized, where $\underline{Y}^m$, $g(\underline{Y}^l)$ and $h(\underline{Y}^r)$ are the $(n \times 1)$ -vectors of the observed values and $(\underline{Y}^m)^* = \underline{X}\underline{a}_m + \underline{1}b_m$, $g^*(\underline{Y}^l) = \underline{X}\underline{a}_l + \underline{1}b_l$ and $h^*(\underline{Y}^r) = \underline{X}\underline{a}_r + \underline{1}b_r$ are the theoretical ones being $\underline{X} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n)'$ the matrix of the explanatory variables of order $(n \times 3p)$. The solution of the LS problem is

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

where $\underline{Y}^{mc}$, $g(\underline{Y}^l)^c$, $h(\underline{Y}^r)^c$ and $\underline{X}^c$ are defined analogously to Section 3. Once again, the parameters significance can be analyzed by means of a bootstrap test [4].

5 Goodness of fit

Model (2) is a particular case of model (4) so, for brevity of exposition, we consider only the latter model. Since the total variation of the response can be written in terms of variances and covariances of real random variables, by taking advantage of their properties it can be decomposed in the variation not depending on the model and that explained by the model. In particular, let $\tilde{Y}$ and $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$ be LR FRVs satisfying the linear model (4) so that the errors are uncorrelated with $\underline{X}$, by indicating $Y^T = (Y^m, g(Y^l), h(Y^r))$, we obtain

$$E \left[D_{\lambda\rho}^2(Y^T, E(Y^T)) \right] = E \left[D_{\lambda\rho}^2(Y^T, E(Y^T|\underline{X})) \right] + E \left[D_{\lambda\rho}^2(E(Y^T|\underline{X}), E(Y^T)) \right]. \quad (5)$$

Let $\tilde{Y}$ be the LR FRV of the linear model (4). Based on (5), it is possible to define the following determination coefficient:

$$R^2 = \frac{E \left[D_{\lambda\rho}^2(E(Y^T|\underline{X}), E(Y^T)) \right]}{E \left[D_{\lambda\rho}^2(Y^T, E(Y^T)) \right]} = 1 - \frac{E \left[D_{\lambda\rho}^2(Y^T, E(Y^T|\underline{X})) \right]}{E \left[D_{\lambda\rho}^2(Y^T, E(Y^T)) \right]}. \quad (6)$$

This coefficient measures the degree of linear relationship. As in the classical case, it takes values in $[0, 1]$.

In order to estimate the determination coefficient, it is worth introducing the decomposition of the total sum of squares. Let $\tilde{Y}$ and $\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p$ be LR FRVs satisfying the linear model (4) observed on n statistical units, $\{\tilde{Y}_i, \tilde{X}_{1i}, \tilde{X}_{2i}, \dots, \tilde{X}_{pi}\}_{i=1, \dots, n}$. The total sum of squares, SST , is equal to the sum of the residual sum of squares, SSE , and the regression sum of squares, SSR , that is,

$$SST = SSE + SSR, \quad (7)$$

where $SST = D_{\lambda\rho}^2((\underline{Y}^m, g(\underline{Y}^l), h(\underline{Y}^r)), (\overline{\underline{Y}^m}, \overline{g(\underline{Y}^l)}, \overline{h(\underline{Y}^r)}))$, $SSE = D_{\lambda\rho}^2((\underline{Y}^m, g(\underline{Y}^l), h(\underline{Y}^r)), (\widehat{\underline{Y}^m}, \widehat{g(\underline{Y}^l)}, \widehat{h(\underline{Y}^r)}))$ and $SSR = D_{\lambda\rho}^2((\widehat{\underline{Y}^m}, \widehat{g(\underline{Y}^l)}, \widehat{h(\underline{Y}^r)}), (\overline{\underline{Y}^m}, \overline{g(\underline{Y}^l)}, \overline{h(\underline{Y}^r)}))$,

with $\widehat{Y^m} = \mathbf{X}\widehat{a_m} + \widehat{1b_m}$, $\widehat{g(Y^l)} = \mathbf{X}\widehat{a_l} + \widehat{1b_l}$, $\widehat{h(Y^r)} = \mathbf{X}\widehat{a_r} + \widehat{1b_r}$ denoting the vectors of the estimated values.

Under the assumptions of model (4), the estimator of the determination coefficient R^2 is

$$\widehat{R}^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST}.$$

It can be shown that $\widehat{R}^2$ is a strongly consistent estimator. Note that the hypothesis testing problem concerning the linear independence can be approached by bootstrapping. For more details, we refer to [3, 4].

6 An application

In a recent study about the student satisfaction of a course the subjective judgements/perceptions were observed on a sample of $n = 64$ students. To formalize the problem we defined $\Omega = \{\text{sets of students attending the course}\}$ endowed with the Borel σ -field. Since the observations were arbitrarily chosen, P is the uniform distribution over Ω . For any $i \in \Omega$, four characteristics were observed. These were the overall assessment of the course, the assessment of the teaching staff, the assessment of the course content and the average mark. The first three subjective judgements/perceptions were managed in terms of triangular fuzzy variables (hence $\lambda = \rho = 1/2$). In fact, to represent them, the students were invited to draw a triangular fuzzy number for every characteristic. The considered support went from 0 (dissatisfaction) to 100 (full satisfaction). The students were informed to place the center where they wished to represent their average judgement/perception and the lower and upper bounds of the triangular fuzzy number where they wished to represent their worst and best judgement/perception, respectively. Note that the students were informed to compute the average, minimum and maximum values w.r.t. the variability of their subjective judgements/perceptions depending on the different course contents and/or members of the teaching staff. The average mark is a crisp variable. For analyzing the linear relationship of the overall assessment of the course ($\tilde{Y}$) on the assessment of the teaching staff ($\tilde{X}_1$), the assessment of the course contents ($\tilde{X}_2$) and the average mark (X_3), the linear regression model in (4) was employed. To overcome the problem about the non-negativity of spreads estimates, we used the logarithmic transformation ($g = h = \log$). Through the *LS* procedure we obtained the following estimated model

$$\begin{cases} \widehat{Y}^m = 1.08X_1^m + 0.13X_1^l - 0.07X_1^r \\ \quad - 0.17X_2^m - 0.89X_2^l + 0.66X_2^r - 1.12X_3 + 34.06 \\ \widehat{Y}^l = \exp(0.01X_1^m + 0.02X_1^l + 0.02X_1^r \\ \quad + 0.00X_2^m + 0.03X_2^l + 0.01X_2^r - 0.00X_3 + 0.67) \\ \widehat{Y}^r = \exp(0.00X_1^m + 0.03X_1^l - 0.02X_1^r \\ \quad - 0.01X_2^m + 0.03X_2^l + 0.01X_2^r + 0.04X_3 + 1.01) \end{cases}$$

We also found $\widehat{R}^2 = 0.7727$, hence approximately 77% of the total variation of the overall assessment of the course was explained by the model. Furthermore, by applying the bootstrap procedure to test the linear independence (with 1000 bootstrap samples) a p -value equal to 0 was obtained, so the null hypothesis of linear independence should be rejected.

7 Concluding remarks

In this work a class of new linear regression models for LR fuzzy variables has been introduced, by taking into account different kinds of uncertainty, through a formalization in terms of FRVs. Furthermore, the problem of the non-negativity of the spreads of the response has been dealt with by using reasonable transformation functions. The suitability of the obtained results has been analyzed by means of a real-life application.

References

1. Coppi, R., Gil, M.A., Kiers, H.A.L.: The fuzzy approach to statistical analysis. *Comput. Statist. Data Anal.* **51**, 1–14 (2006).
2. Ferraro, M.B., Coppi, R., Gonzalez-Rodriguez, G., Colubi, A.: A linear regression model for imprecise response. *Int. J. Approx. Reason.* (2010) doi: 10.1016/j.ijar.2010.04.003.
3. Ferraro, M.B., Colubi, A., Gonzalez-Rodriguez, G., Coppi, R.: A determination coefficient for a linear regression model with imprecise response. *Environmetrics* (accepted with minor revisions) (2010).
4. Ferraro, M.B., Giordani, P.: A multiple linear regression model for LR fuzzy random variables. *Tech. Rep. n. 3*, Dept. Statistics, Probability and Applied Statistics, Sapienza University of Rome. Submitted (2010).
5. González-Rodríguez, G., Blanco, A., Colubi, A., Lubiano, M.A.: Estimation of a simple linear regression model for fuzzy random variables. *Fuzzy Sets Syst.* **160**, 357–370 (2009).
6. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. *J. Math. Anal. Appl.* **114**, 409–422 (1986).
7. Yang, M.S., Ko, C.H.: On a class of fuzzy c -numbers clustering procedures for fuzzy data. *Fuzzy Sets Syst.* **84**, 49–60 (1996).
8. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965).