Metrika (2012) 75:1049–1068
DOI 10.1007/s00184-011-0367-3

A multiple linear regression model for imprecise information

Maria Brigida Ferraro · Paolo Giordani

Received: 18 September 2010 / Published online: 23 July 2011

Abstract In standard regression analysis the relationship between the (response) variable and a set of (explanatory) variables is investigated. In the classical framework the response is affected by probabilistic uncertainty (randomness) and, thus, treated as a random variable. However, the data can also be subjected to other kinds of uncertainty such as imprecision. A possible way to manage all of these uncertainties is represented by the concept of fuzzy random variable (FRV). The most common class of FRVs is the LR family (LR FRV), which allows us to express every FRV in terms of three random variables, namely, the center, the left spread and the right spread. In this work, limiting our attention to the LR FRV class, we consider the linear regression problem in the presence of one or more imprecise random elements. The procedure for estimating the model parameters and the determination coefficient are discussed and the hypothesis testing problem is addressed following a bootstrap approach. Furthermore, in order to illustrate how the proposed model works in practice, the results of a real-life example are given together with a comparison with those obtained by applying classical regression analysis.

Keywords LR fuzzy data · Regression models · Least squares approach · Bootstrap procedure

1 Introduction

In several situations the description of real world phenomena can be done by using numerical values. However, in some cases, it may happen that the available information is affected by imprecision. A possible way to cope with imprecision is represented

M. B. Ferraro (✉) · P. Giordani
Dip. Scienze Statistiche, Sapienza Università di Roma, P.le A. Moro 5, 00185 Roma, Italy
e-mail: mariabrigida.ferraro@uniroma1.it

by fuzzy set theory (Zadeh 1965). This allows us to express imprecise information in terms of fuzzy sets (see, for instance, Näther and Wünsche 2007; Arnold and Stahlecker 2010; Ramos-Guajardo et al. 2010; Arefi et al. 2011). In contrast with the 'black and white' nature of classical sets (given a universe of values U, an element $x \in U$ either belongs or does not belong to the classical set, say $A \subseteq U$), fuzzy sets are characterized by a 'gray scale' nature allowing us to express a degree according to which an element x belongs to the fuzzy set $\tilde{A} \subseteq U$, where the symbol tilde is used to denote that a fuzzy set is involved. Such a degree, usually called membership function of $x \in U$ in $\tilde{A}$ and denoted by $\mu_{\tilde{A}}(x)$, ranges from 0 (complete non-membership) to 1 (complete membership). For instance, if we are interested in managing the attribute "bad" in terms of fuzzy numbers (fuzzy set $\tilde{A}$), we can say that, in the scale [0,100] (universe U), the term "bad" corresponds to a fuzzy set in the interval [20, 45]. This means that, within these values, the membership function takes values strictly higher than zero. The higher (and closer to one) $\mu_{\tilde{A}}(x)$ is, the better $x \in U$ describes the term "bad". For instance, $\mu_{\tilde{A}}(26) = 0.9$ is the degree of truth (0.9) of "bad" concerning number 26 (26 characterizes "bad" with a degree equal to 0.9).

It should be underlined that there are, at least, two different interpretations of a fuzzy datum and, therefore, of a fuzzy variable. One point of view consists in looking at a fuzzy datum as a convenient representation of an underlying crisp datum, whenever one does not have sufficient information for singling it out, therefore being compelled to describe it with some imprecision. In this case (see, for instance, Kruse and Meyer 1987), the two components (the ill-known crisp datum and its imprecision) need a different treatment, due to their distinct nature. An opposite view of fuzzy datum, followed in this paper, looks at the imprecision embodied in it as an intrinsic property. In this case the membership function should be considered as a whole and dealt with as an entity in itself, with reference to both its mathematical representation and its utilization within a statistical model. This is, for instance, the view hold by Puri and Ralescu (1986) when they introduced the notion of Fuzzy Random Variable (FRV). The need for FRVs arises when the data are not only affected by imprecision but also by randomness. Imprecision and randomness are different sources of uncertainty, which may affect the data. They are not exclusive but can occur together. See, for more details about the different sources of uncertainty, Klir (2006).

In this work we aim at investigating the linear regression problem when the data are random and imprecise managing them as FRVs. In the literature, there exist several works devoted to the linear regression problem for imprecise data. At least three approaches can be distinguished. The first one is the possibilistic approach. Originally introduced by Tanaka et al. (1982), its basic idea is that the regression model is intrinsically fuzzy because there does not exist a "true" relationship between the response variable and the explanatory ones. This is done by detecting fuzzy regression coefficients such that the fuzziness of the estimated response variable is minimized. Other works about possibilistic regression can be found in, e.g., Tanaka and Watada (1988); Tanaka et al. (1995) and Guo and Tanaka (2006). Another approach is the least squares one, in which a suitable dissimilarity measure between the observed and the estimated response variable must be introduced and the model parameters are estimated by minimizing such a dissimilarity measure. See, for instance, Celminš (1987); Diamond (1988); Chang and Lee (1996); D'Urso (2003); Coppi et al. (2006);

Bargiela et al. (2007); Lu and Wang (2009). Generally speaking, the possibilistic and least squares approaches could also be used when the fuzzy data are affected by randomness. Unfortunately, this is simply done by overlooking it. The third line of research, which we may call fuzzy-probabilistic approach, consists in explicitly taking into account randomness for estimating the regression parameters and assessing their statistical properties. Works belonging to this approach can be found in, e.g., Körner and Näther (1998); Krätschmer (2006a,b); Näther (2006); González-Rodríguez et al. (2009); Ferraro et al. (2010). Note, however, that a few assignments of the previously mentioned papers to a given approach could be debatable.

In this paper we approach the linear regression problem from the fuzzy-probabilistic viewpoint. Limiting our attention to the so-called LR fuzzy family, this is achieved by proposing a new linear regression model exploiting the potentialities of FRVs. As we will see, the parameters can be expressed in terms of the moments of real random variables. In order to estimate the parameters a closed form solution will be provided and their statistical properties will be investigated. The paper is organized as follows. In the next section, the concepts of (LR) fuzzy sets and FRVs are recalled. Section 3 focuses on the proposed linear regression model. In Sect. 4 we discuss the estimation of the model parameters. In Sect. 5, by using a bootstrap approach, the hypothesis testing problem is addressed and a simulation experiment is carried out in order to evaluate the adequacy of the bootstrap tests. Section 6 contains the results of a real-life application concerning collected fuzzy data about the evaluation of a course by a sample of students along with a comparison of the performance of the proposed model with the one of classical regression. Finally, some concluding remarks are made.

2 Preliminaries

We already saw that a fuzzy set $\widetilde{A}$ is a subset of the universe U defined through the so-called *membership function* $\mu_{\widetilde{A}}(x), \forall x \in U$, expressing the extent to which x belongs to $\widetilde{A}$. 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 $s : \mathcal{F}_{LR} \to \mathbb{R}^3$, i.e., $s(\widetilde{A}) = s_{\widetilde{A}} = (A^m, A^l, A^r)$. This implies that $\widetilde{A}$ 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 $\widetilde{A} \in \mathcal{F}_{LR}$ or (A^m, A^l, A^r).

The arithmetics considered in $\mathcal{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for interval. Going into detail, the sum of $\widetilde{A}$ and $\widetilde{B}$ in $\mathcal{F}_{LR}$ is the LR fuzzy number $\widetilde{A} + \widetilde{B}$ 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 $\widetilde{A} \in \mathcal{F}_{LR}$ by a scalar $\gamma > 0$ is

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

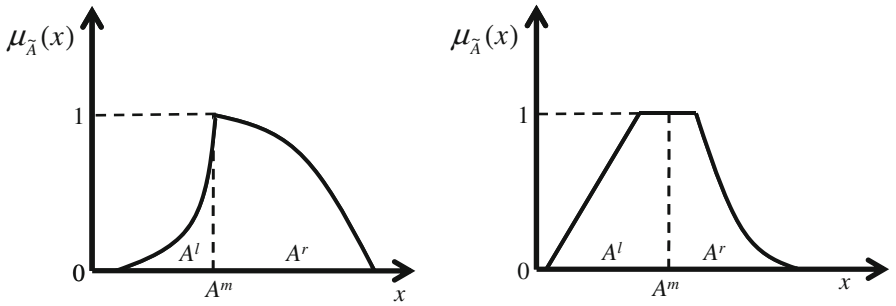

Fig. 1 Examples of LR fuzzy numbers

The membership function of $\widetilde{A} \in \mathcal{F}_{LR}$ can be written as

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

where the functions $L, R : \mathbb{R} \to [0, 1]$ 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 (see Fig. 1).
$\widetilde{A}$ is a *triangular* fuzzy number if (1) takes the form

$$\mu_{\widetilde{A}}(x) = \begin{cases} 0 & x \leq A^m - A^l, \\ 1 - \frac{A^m - x}{A^l} & A^m - A^l \leq x \leq A^m, \\ 1 - \frac{x - A^m}{A^r} & A^m \leq x \leq A^m + A^r, \\ 0 & x \geq A^m + A^r. \end{cases} \tag{2}$$

It is convenient to mention that the functions L and R must be chosen in advance by the researcher and, in general, the same shape functions are fixed for all the available data. The role of such functions is to take into account suitably the level of imprecision embodied in the data. The α-level set ($0 < \alpha \leq 1$) of $\widetilde{A}$ can be defined as the non-empty compact convex subset of $\mathbb{R}$, A_α, such that $A_\alpha = \left\{x \in U : \mu_{\widetilde{A}}(x) \geq \alpha\right\}$. If $\alpha = 0, A_0 = cl(\{x \in \mathbb{R} : \mu_{\widetilde{A}}(x) > 0\})$. For more details one can refer to Zimmermann (2001). With particular reference to fuzzy arithmetics with LR fuzzy numbers see Hanss (2005).

A distance for LR fuzzy numbers has been introduced by Yang and Ko (1996). It is

$$\begin{aligned} D^2_{LR}(\widetilde{A}, \widetilde{B}) = (A^m - B^m)^2 + [(A^m - \lambda A^l) - (B^m - \lambda B^l)]^2 \\ + [(A^m + \rho A^r) - (B^m + \rho B^r)]^2. \end{aligned} \tag{3}$$

In (3), the parameters $\lambda = \int_0^1 L^{-1}(\omega)\mathrm{d}\omega$ and $\rho = \int_0^1 R^{-1}(\omega)\mathrm{d}\omega$ play the role of taking into account the shape of the membership function. For instance, if the membership function takes the form reported in (2), it is $\lambda = \rho = \frac{1}{2}$. As it will be clear, 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 has been derived (see Ferraro et al. 2010). Given $\underline{a} = (a_1, a_2, a_3)$ and $\underline{b} = (b_1, b_2, b_3) \in \mathbb{R}^3$, it is

$$D^2_{\lambda\rho}(\underline{a}, \underline{b}) = (a_1 - b_1)^2 + ((a_1 - \lambda a_2) - (b_1 - \lambda b_2))^2 + ((a_1 + \rho a_3) - (b_1 + \rho b_3))^2, \tag{4}$$

where $\lambda, \rho \in \mathbb{R}^+$. The distance in (4) will be used in the following as a tool for quantifying errors in the regression model we are going to introduce.

The definition of α-level set is connected to that of FRV in Puri and Ralescu's sense. Note that in the following we limit our attention to FRVs of LR type (in brief LR FRV). Let (Ω, A, P) be a probability space, an LR FRV is a mapping $\widetilde{X} : \Omega \to \mathcal{F}_{LR}$ such that the α-level set X_α is a random compact convex set for any $\alpha \in [0, 1]$ (see, for further details, Puri and Ralescu 1985, 1986). As for non-fuzzy random variables, it is possible to determine the moments of a FRV. The expectation of an LR FRV $\widetilde{X}$, $E(\widetilde{X})$, is the fuzzy set in $\mathcal{F}_{LR}$ (EX^m, EX^l, EX^r). With respect to (3) the variance of $\widetilde{X}$ is $\sigma^2_{\widetilde{X}} = var(\widetilde{X}) = E[(D^2_{LR}(\widetilde{X}, E(\widetilde{X}))]$ (see Ferraro et al. 2010).

3 The linear regression model for LR FRVs

The available information refers to an LR fuzzy response variable $\widetilde{Y}$ and p LR fuzzy explanatory variables $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$ observed on a random sample of n statistical units, $\{\widetilde{Y}_i, \widetilde{X}_{1i}, \widetilde{X}_{2i}, \ldots, \widetilde{X}_{pi}\}_{i=1,\ldots,n}$. We are interested in analyzing the relationship between $\widetilde{Y}$ and $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$. We consider L and R as fixed functions, so the fuzzy response and the fuzzy explanatory variables are determined only by means of three parameters, namely the center and the left and right spreads. The idea is to model the center and the spreads of $\widetilde{Y}$ by means of the centers and the spreads of $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$. However, in doing so, attention should be paid to the non-negativity of the spreads of $\widetilde{Y}$. To overcome this problem one can either solve a non-negative regression problem (see, e.g., Lawson and Hanson 1995) or model a transform of the spreads of $\widetilde{Y}$ (the new "response variable") by means of the centers and the spreads of $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$. The former choice is a numerical procedure yielding a dependence between the errors and the explanatory variables (Liew 1976) and not allowing to formalize a realistic theoretical model and to obtain a complete analytical solution. For this reason we propose to consider the latter choice introducing two invertible functions $g : (0, +\infty) \longrightarrow \mathbb{R}$ and $h : (0, +\infty) \longrightarrow \mathbb{R}$. The linear regression model can be formalized as

$$\begin{cases} Y^m = \underline{X}\,\underline{a}'_m + b_m + \varepsilon_m, \\ g(Y^l) = \underline{X}\,\underline{a}'_l + b_l + \varepsilon_l, \\ h(Y^r) = \underline{X}\,\underline{a}'_r + b_r + \varepsilon_r, \end{cases} \tag{5}$$

where $\underline{X} = (X_1^m, X_1^l, X_1^r, \ldots, X_p^m, X_p^l, X_p^r)$ is the row-vector of lenght $3p$ of all the components of the explanatory variables, ε_m, ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|\underline{X}) = E(\varepsilon_l|\underline{X}) = E(\varepsilon_r|\underline{X}) = 0$, $\underline{a}_m = (a_{mm}^1, a_{ml}^1, a_{mr}^1, \ldots, a_{mm}^p, a_{ml}^p, a_{mr}^p)$, $\underline{a}_l = (a_{lm}^1, a_{ll}^1, a_{lr}^1, \ldots, a_{lm}^p, a_{ll}^p, a_{lr}^p)$ and $\underline{a}_r = (a_{rm}^1, a_{rl}^1, a_{rr}^1, \ldots, a_{rm}^p, a_{rl}^p, a_{rr}^p)$ are row-vectors of length $3p$ of the parameters related to $\underline{X}$. The generic a_{ij}^t is the regression coefficient between the component $i \in \{m, l, r\}$ of $\widetilde{Y}$ (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 $\widetilde{X}^t$, $t = 1, \ldots, p$, (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 $\widetilde{X}^3(X_3^r)$ and the center of the response, Y^m. Finally, b_m, b_l, b_r denote the intercepts. Therefore, by means of (5), we aim at studying the relationship between the response and the explanatory variables taking into account also the information provided by the spreads of the explanatory variables, which are usually arbitrarily ignored.

The covariance matrix of $\underline{X}$ is denoted by $\Sigma_{\underline{X}} = E[(\underline{X} - E\underline{X})'(\underline{X} - E\underline{X})]$ and Σ stands for the covariance matrix of $(\varepsilon_m, \varepsilon_l, \varepsilon_r)$, with variances, $\sigma_{\varepsilon_m}^2, \sigma_{\varepsilon_l}^2$ and $\sigma_{\varepsilon_r}^2$, strictly positive and finite. The population parameters can then be expressed, as usual, in terms of some moments related to real random variables. We get

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

The above expressions are useful to prove some statistical properties of the estimators introduced in the next section.

Remark 1 When the explanatory variables are real-valued, the model in (5) reduces to the regression model proposed by Ferraro et al. (2010).

3.1 The determination coefficient

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 $\widetilde{Y}$ and $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$ be LR FRVs satisfying the linear model in (5) so that the errors are uncorrelated with $\underline{X}$, by indicating $Y^T = (Y^m, g(Y^l), h(Y^r))$,

we obtain

$$E\left[D^2_{\lambda\rho}(Y^T, E(Y^T))\right] = E\left[D^2_{\lambda\rho}(Y^T, E(Y^T|\underline{X}))\right] + E\left[D^2_{\lambda\rho}(E(Y^T|\underline{X}), E(Y^T))\right]. \quad (6)$$

Based on the decomposition of the total variation (6), it is possible to define the following determination coefficient.

Definition 1 Let $\widetilde{Y}$ be the LR FRV of the linear model in (5), the determination coefficient can be defined as

$$R^2 = \frac{E\left[D^2_{\lambda\rho}(E(Y^T|\underline{X}), E(Y^T))\right]}{E\left[D^2_{\lambda\rho}(Y^T, E(Y^T))\right]} = 1 - \frac{E\left[D^2_{\lambda\rho}(Y^T, E(Y^T|\underline{X}))\right]}{E\left[D^2_{\lambda\rho}(Y^T, E(Y^T))\right]}. \quad (7)$$

This coefficient measures the degree of linear relationship. As in the classical case, it takes values in [0,1]. Concerning the spreads, model (5) is linear in the transformed scales represented by functions g and h, so R^2 refers specifically to the chosen scales. When $R^2 = 0$ the regression model is not able to explain the variability of the response variable (linear independence). Conversely, $R^2 = 1$ implies that the regression model accounts for the whole variability of the response variable (perfect fit). The closer R^2 is to 1, the better the model explains the variability of the response variable.

4 The estimation problem

4.1 Estimation of the regression parameters

The estimation problem of the regression parameters is faced by means of the Least Squares (LS) criterion. Accordingly, the parameters of model (5) are estimated by minimizing the sum of the squared distances between the observed and theoretical values of the response variable. However, as already noted, suitable transforms of the spreads are considered in (5). This allows us to use (4) in the objective function of the problem. Therefore, the LS problem consists in looking for $\widehat{\underline{a}}_m, \widehat{\underline{a}}_l, \widehat{\underline{a}}_r, \widehat{b}_m, \widehat{b}_l$ and $\widehat{b}_r$ such that

$$\begin{aligned}\Delta^2_{\lambda\rho} &= D^2_{\lambda\rho}((\underline{Y}^m, g(\underline{Y}^l), h(\underline{Y}^r)), ((\underline{Y}^m)^*, g(\underline{Y}^l)^*, h(\underline{Y}^r)^*)) \\ &= \sum_{i=1}^{n} D^2_{\lambda\rho}((Y_i^m, g(Y_i^l), h(Y_i^r)), ((Y_i^m)^*, g(Y_i^l)^*, h(Y_i^r)^*)) \end{aligned} \quad (8)$$

is minimized, where $\underline{Y}^m$, $g(\underline{Y}^l)$ and $h(\underline{Y}^r)$ are the vectors of length n of the observed values and $(\underline{Y}^m)^* = \mathbf{X}\underline{a}_m' + \underline{1}b_m$, $g(\underline{Y}^l)^* = \mathbf{X}\underline{a}_l' + \underline{1}b_l$ and $h(\underline{Y}^r)^* = \mathbf{X}\underline{a}_r' + \underline{1}b_r$ are the theoretical ones being $\mathbf{X} = (\underline{X}_1, \underline{X}_2, \ldots, \underline{X}_n)'$ the $n \times 3p$ matrix of the explanatory variables. Finally $\underline{1}$ is the unit vector of length n.

Proposition 1 *The solution of the LS problem is*

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

where

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

are the centered values of the response variables,

$$\mathbf{X}^c = \mathbf{X} - \underline{1}\,\overline{\underline{X}}$$

is the centered matrix of the explanatory variables and, $\overline{Y^m}$, $\overline{g(Y^l)}$, $\overline{h(Y^r)}$ *and* $\overline{\underline{X}}$ *denote, respectively, the sample means of* Y^m, $g(Y^l)$, $h(Y^r)$ *and* $\underline{X}$.

Proof In order to solve the minimization problem in (8) and to find the parameters estimators, we follow the usual procedure of equating to zero the partial derivatives of the objective function with respect to (w.r.t.) the parameters to be estimated, although we have to take into account that the regression parameters are related to some others.

It is easy to show that the objective function in (8) can be rewritten as

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

where $\|\cdot\|^2$ denotes the squared Euclidean norm. After a little algebra, it can be exploited as

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

$$+2\rho\left(\underline{Y}^m - \mathbf{X}\underline{a}_m^{'} - \underline{1}b_m\right)^{'}\left(h(\underline{Y}^r) - \mathbf{X}\underline{a}_r^{'} - \underline{1}b_r\right). \tag{9}$$

Starting from the estimation of b_l and b_r, we equate to zero the partial derivatives w.r.t b_l and b_r, respectively. It is easy to find that the minimum is attained at

$$b_l = \overline{g(Y^l)} - \overline{\underline{X}}\,\underline{a}_l^{'} - \frac{1}{\lambda}\overline{Y^m} + \frac{1}{\lambda}\overline{\underline{X}}\,\underline{a}_m^{'} + \frac{1}{\lambda}b_m, \tag{10}$$

$$b_r = \overline{h(Y^r)} - \overline{\underline{X}}\,\underline{a}_r^{'} + \frac{1}{\rho}\overline{Y^m} - \frac{1}{\rho}\overline{\underline{X}}\,\underline{a}_m^{'} - \frac{1}{\rho}b_m. \tag{11}$$

Since b_l and b_r depend on b_m, we have to substitute (10) and (11) in (9) before equating to zero the partial derivative of the objective function w.r.t. b_m. As a result, we obtain

$$b_m = \overline{Y^m} - \overline{\underline{X}}\,\underline{a}_m^{'}.$$

Since the parameters b_m, b_l and b_r are expressed in terms of $\underline{a}_m$, $\underline{a}_l$ and $\underline{a}_r$, to go on with the estimation procedure it is important to take this into account by substituting b_m, b_l and b_r in the objective function.
We consider the centered vectors $\underline{Y}^{mc}$, $g(\underline{Y}^l)^c$, $h(\underline{Y}^r)^c$ and the centered matrix $\mathbf{X}^c$ to make it simpler to analyze the objective function that can be expressed as

$$\begin{aligned}\Delta^2_{\lambda\rho} = {} & 3\left(\underline{Y}^{mc} - \mathbf{X}^c\underline{a}_m^{'}\right)^{'}\left(\underline{Y}^{mc} - \mathbf{X}^c\underline{a}_m^{'}\right)\\ & +\lambda^2\left(g(\underline{Y}^l)^c - \mathbf{X}^c\underline{a}_l^{'}\right)^{'}\left(g(\underline{Y}^l)^c - \mathbf{X}^c\underline{a}_l^{'}\right)\\ & +\rho^2\left(h(\underline{Y}^r)^c - \mathbf{X}^c\underline{a}_r^{'}\right)^{'}\left(h(\underline{Y}^r)^c - \mathbf{X}^c\underline{a}_r^{'}\right)\\ & -2\lambda(\underline{Y}^{mc} - \mathbf{X}^c\underline{a}_m^{'})^{'}\left(g(\underline{Y}^l)^c - \mathbf{X}^c\underline{a}_l^{'}\right)\\ & +2\rho(\underline{Y}^{mc} - \mathbf{X}^c\underline{a}_m^{'})^{'}\left(h(\underline{Y}^r)^c - \mathbf{X}^c\underline{a}_r^{'}\right).\end{aligned} \tag{12}$$

Following the usual reasoning it is easy to check that

$$\underline{a}_l^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}g(\underline{Y}^l)^c - \frac{1}{\lambda}(\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}\underline{Y}^{mc} + \frac{1}{\lambda}\underline{a}_m^{'}, \tag{13}$$

$$\underline{a}_r^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}h(\underline{Y}^r)^c + \frac{1}{\rho}(\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}\underline{Y}^{mc} - \frac{1}{\rho}\underline{a}_m^{'}. \tag{14}$$

The last step is the estimation of $\underline{a}_m$. Since this vector appears in (13) and (14) we need to substitute (13) and (14) in (12). By equating to 0 the partial derivative of (12) w.r.t. $\underline{a}_m$ we get

$$\widehat{\underline{a}}_m^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}\underline{Y}^{mc}.$$

By making all the appropriate substitutions we also find

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

□

Remark 2 Since the LS estimators are written in terms of sample moments and taking into account the expression of the theoretical values, it can be shown that they are unbiased and strongly consistent.

Remark 3 Once the parameters of the model are determined, the estimated values of the response variable can be computed as follows. First of all, the estimated centers are $\widehat{Y}_i^m = \underline{X}_i\widehat{\underline{a}}_m^{'} + \widehat{b}_m$, $i = 1, \ldots, n$. The estimated transforms of the left and right spreads are $\widehat{g(Y_i^l)} = \underline{X}_i\widehat{\underline{a}}_l^{'} + \widehat{b}_l$ and $\widehat{h(Y_i^r)} = \underline{X}_i\widehat{\underline{a}}_r^{'} + \widehat{b}_r$, respectively, from which we get the estimated spreads $\widehat{Y}_i^l = g^{-1}(\underline{X}_i\widehat{\underline{a}}_l^{'} + \widehat{b}_l)$ and $\widehat{Y}_i^r = h^{-1}(\underline{X}_i\widehat{\underline{a}}_r^{'} + \widehat{b}_r)$, respectively, $i = 1, \ldots, n$. Therefore, the estimated response variable is an LR fuzzy number and the shape of its membership function is inherited from that of the observed response variable.

4.2 Estimation of the determination coefficient

In order to estimate the determination coefficient, it is worth introducing the next proposition about the decomposition of the total sum of squares.

Proposition 2 *Let $\widetilde{Y}$ and $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$ be LR FRVs satisfying the linear model in (5) observed on n statistical units, $\{\widetilde{Y}_i, \underline{X}_{1i}, \underline{X}_{2i}, \ldots, \underline{X}_{pi}\}_{i=1,\ldots,n}$. 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,*

$$SST = SSE + SSR, \tag{15}$$

where

$$\begin{aligned}
SST = &\left\|\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m}\right\|^2 + \left\|\left(\underline{Y}^m - \lambda g(\underline{Y}^l)\right) - \left(\underline{1}\,\overline{\underline{Y}^m} - \lambda\underline{1}\,\overline{g(\underline{Y}^l)}\right)\right\|^2\\
&+ \left\|\left(\underline{Y}^m + \rho h(\underline{Y}^r)\right) - \left(\underline{1}\,\overline{\underline{Y}^m} + \rho\underline{1}\,\overline{h(\underline{Y}^r)}\right)\right\|^2,\\
SSE = &\left\|\underline{Y}^m - \widehat{\underline{Y}^m}\right\|^2 + \left\|\left(\underline{Y}^m - \lambda g(\underline{Y}^l)\right) - \left(\widehat{\underline{Y}^m} - \lambda\widehat{g(\underline{Y}^l)}\right)\right\|^2\\
&+ \left\|\left(\underline{Y}^m + \rho h(\underline{Y}^r)\right) - \left(\widehat{\underline{Y}^m} + \rho\widehat{h(\underline{Y}^r)}\right)\right\|^2,
\end{aligned}$$

$$SSR = \left\| \widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right\|^2 + \left\| \left(\widehat{\underline{Y}^m} - \lambda \widehat{g(\underline{Y}^l)} \right) - \left(\underline{1}\,\overline{\underline{Y}^m} - \lambda \underline{1}\,\overline{g(\underline{Y}^l)} \right) \right\|^2 + \left\| \left(\widehat{\underline{Y}^m} + \rho \widehat{h(\underline{Y}^r)} \right) - \left(\underline{1}\,\overline{\underline{Y}^m} + \rho \underline{1}\,\overline{h(\underline{Y}^r)} \right) \right\|^2,$$

with $\widehat{\underline{Y}^m}$, $\widehat{g(\underline{Y}^l)}$, $\widehat{h(\underline{Y}^r)}$ *being the vectors of the estimated values, that is,*

$$\widehat{\underline{Y}^m} = \mathbf{X}\underline{\widehat{a}}_m^{'} + \underline{1}\,\widehat{b}_m, \quad \widehat{g(\underline{Y}^l)} = \mathbf{X}\underline{\widehat{a}}_l^{'} + \underline{1}\,\widehat{b}_l, \quad \widehat{h(\underline{Y}^r)} = \mathbf{X}\underline{\widehat{a}}_r^{'} + \underline{1}\,\widehat{b}_r.$$

Proof The total sum of squares can be written as

$$\begin{aligned} SST = {} & 3 \left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right)^{'} \left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right) \\ & + \lambda^2 \left(g(\underline{Y}^l) - \underline{1}\,\overline{g(\underline{Y}^l)} \right)^{'} \left(g(\underline{Y}^l) - \underline{1}\,\overline{g(\underline{Y}^l)} \right) \\ & + \rho^2 \left(h(\underline{Y}^r) - \underline{1}\,\overline{h(\underline{Y}^r)} \right)^{'} \left(h(\underline{Y}^r) - \underline{1}\,\overline{h(\underline{Y}^r)} \right) \\ & - 2\lambda \left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right)^{'} \left(g(\underline{Y}^l) - \underline{1}\,\overline{g(\underline{Y}^l)} \right) \\ & + 2\rho \left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right)^{'} \left(h(\underline{Y}^r) - \underline{1}\,\overline{h(\underline{Y}^r)} \right). \end{aligned} \tag{16}$$

By subtracting and adding $\widehat{\underline{Y}^m}$ in $\left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right)$, we get that $\left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right)^{'} \left(\underline{Y}^m - \underline{1}\,\overline{\underline{Y}^m} \right)$ is equal to

$$\begin{aligned} & \left(\underline{Y}^m - \widehat{\underline{Y}^m} + \widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right)^{'} \left(\underline{Y}^m - \widehat{\underline{Y}^m} + \widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right) \\ & = \left(\underline{Y}^m - \widehat{\underline{Y}^m} \right)^{'} \left(\underline{Y}^m - \widehat{\underline{Y}^m} \right) + \left(\widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right)^{'} \left(\widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right) \qquad (17) \\ & + 2 \left(\underline{Y}^m - \widehat{\underline{Y}^m} \right)^{'} \left(\widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right). \qquad (18) \end{aligned}$$

The two terms of (17) are the first terms of SSE and SSR, respectively. Now we prove that the term in (18) is equal to 0. Since $\widehat{\underline{Y}^m} = \mathbf{X}\underline{\widehat{a}}_m^{'} + \underline{1}\,\widehat{b}_m$ where $\underline{\widehat{a}}_m^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}\underline{Y}^{mc}$ and $\widehat{b}_m = \overline{Y^m} - \overline{\underline{X}}\,\underline{\widehat{a}}_m^{'}$, it results

$$\begin{aligned} & \left(\underline{Y}^m - \widehat{\underline{Y}^m} \right)^{'} \left(\widehat{\underline{Y}^m} - \underline{1}\,\overline{\underline{Y}^m} \right) \\ & = \left(\underline{Y}^m - \mathbf{X}\underline{\widehat{a}}_m^{'} - \underline{1}\,\overline{Y^m} + \underline{1}\,\overline{\underline{X}}\,\underline{\widehat{a}}_m^{'} \right)^{'} \left(\mathbf{X}\underline{\widehat{a}}_m^{'} + \underline{1}\,\overline{Y^m} - \underline{1}\,\overline{\underline{X}}\,\underline{\widehat{a}}_m^{'} - \underline{1}\,\overline{\underline{Y}^m} \right) \\ & = \left(\underline{Y}^{mc} - \mathbf{X}^c\underline{\widehat{a}}_m^{'} \right)^{'} \left(\mathbf{X}^c\underline{\widehat{a}}_m^{'} \right) = \underline{\widehat{a}}_m\mathbf{X}^c\underline{Y}^{mc} - \underline{\widehat{a}}_m\mathbf{X}^c\underline{Y}^{mc} = 0. \end{aligned}$$

By using the same procedure for the other terms in (16), namely by subtracting and adding the corresponding estimate in each term, the thesis follows. □

Definition 2 Let $\widetilde{Y}$ and $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$ be LR FRVs satisfying the linear model in (5) observed on n statistical units, $\{\widetilde{Y}_i, \widetilde{X}_{1i}, \widetilde{X}_{2i}, \ldots, \widetilde{X}_{pi}\}_{i=1,\ldots,n}$. The estimator of the determination coefficient R^2 is

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

It represents the part of the total sum of squares explained by the regression model, so it can be considered as a goodness-of-fit measure and it takes values in [0, 1]. Furthermore, it can be shown that $\widehat{R}^2$ is a strong consistent estimator.

5 Hypothesis testing

5.1 Hypothesis testing on the regression parameters

The parameters $\underline{a}_m, \underline{a}_l$ and $\underline{a}_r$ express the strength of the relationship between the response variable and the explanatory ones. Testing the explicative power of $\underline{X}$ consists in testing that the vectors of coefficients $\underline{a}_m, \underline{a}_l$ and $\underline{a}_r$ are equal to $\underline{0}$. In general it is possible to test the null hypothesis

$$H_0 : \begin{pmatrix} \underline{a}_m' \\ \underline{a}_l' \\ \underline{a}_r' \end{pmatrix} = \begin{pmatrix} \underline{k}_m' \\ \underline{k}_l' \\ \underline{k}_r' \end{pmatrix}$$

against the alternative

$$H_1 : \begin{pmatrix} \underline{a}_m' \\ \underline{a}_l' \\ \underline{a}_r' \end{pmatrix} \neq \begin{pmatrix} \underline{k}_m' \\ \underline{k}_l' \\ \underline{k}_r' \end{pmatrix},$$

where $\underline{k}_m, \underline{k}_l$, and $\underline{k}_r$ are real-valued vectors. Starting from Ferraro et al. (2010), the test statistic to be used is $T_n = V_n' V_n$, where

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

It is important to stress that, since there are not generalized models for FRVs that can be used in practice and an asymptotic test works suitably for large size samples, the hypothesis testing problem is approached by bootstrapping (see, for more details, Efron and Tibshirani 1993). The non-parametric bootstrap test is based on the following algorithm:

Bootstrap algorithm

Step 1: Compute the estimates $\widehat{\underline{a}}_m, \widehat{\underline{a}}_l, \widehat{\underline{a}}_r$ and the value of the statistic

$$T_n = V_n' V_n.$$

Step 2: Compute the bootstrap population fulfilling the null hypothesis,

$$\left\{(\underline{X}_i, Z_i^m, Z_i^l, Z_i^r)\right\}_{i=1,\ldots,n}, \tag{19}$$

where

$$\begin{aligned} Z_i^m &= Y_i^m - \underline{X}_i \widehat{\underline{a}}_m^{'} + \underline{X}_i \underline{k}_m^{'}, \\ Z_i^l &= g(Y_i^l) - \underline{X}_i \widehat{\underline{a}}_l^{'} + \underline{X}_i \underline{k}_l^{'}, \\ Z_i^r &= h(Y_i^r) - \underline{X}_i \widehat{\underline{a}}_r^{'} + \underline{X}_i \underline{k}_r^{'}. \end{aligned}$$

Step 3: Draw a sample of size n with replacement

$$\left\{(X_i^*, Z_i^{m*}, Z_i^{l*}, Z_i^{r*})\right\}_{i=1,\ldots,n},$$

from the bootstrap population (19).

Step 4: Compute the bootstrap estimates $\widehat{\underline{a}}_m^*, \widehat{\underline{a}}_l^*, \widehat{\underline{a}}_r^*$ and the value of the bootstrap statistic

$$T_n^* = V_n^{*'} V_n^*.$$

Step 5: Repeat Steps 3 and 4 a large number B of times to get a set of B estimators, denoted by $\left\{T_{n1}^*, \ldots, T_{nB}^*\right\}$.

Step 6: Compute the bootstrap p-value as the proportion of values in $\left\{T_{n1}^*, \ldots, T_{nB}^*\right\}$ being greater than T_n.

5.2 Hypothesis testing on a single parameter

A particular case of the above hypothesis test on the regression parameters is referred to testing the significance of a single regression parameter. In this way it is possible to check if a given component of the explanatory variables is significantly related to the LR fuzzy response variable. For example, let $\widetilde{Y}$ and $\widetilde{X}_1, \widetilde{X}_2, \ldots, \widetilde{X}_p$ be LR FRVs satisfying the linear model in (5), to test the significance of the left spread of the explanatory variable $\widetilde{X}_1$ w.r.t. the center of the response variable $\widetilde{Y}$, it is tested the following hypothesis

$$H_0 : a_{ml}^1 = 0$$

against the alternative

$$H_1 : a^1_{ml} \neq 0.$$

As for the previous hypothesis test, according to the bootstrap approach, the above described algorithm can be adopted. The most relevant difference consists in considering a bootstrap population $\left\{(\underline{X}_i, Z_i^m, Z_i^l, Z_i^r)\right\}_{i=1,\ldots,n}$ fulfilling the null hypothesis which is now

$$\begin{aligned} Z_i^m &= Y_i^m - \widehat{a}^1_{ml} X_1^l, \\ Z_i^l &= g(Y_i^l), \\ Z_i^r &= h(Y_i^r). \end{aligned}$$

The significance of the intercepts can also be tested by suitable modifications of the bootstrap population fulfilling the null hypothesis and of the test statistic.

5.3 Linear independence test

In this section a bootstrap linear independence test is introduced on the basis of Ferraro et al. (2011). To test the null hypothesis $H_0 : R^2 = 0$ against the alternative $H_1 : R^2 > 0$, the test statistic $T_n = n\widehat{R}^2$ is used. Once again, a bootstrap algorithm can be adopted. To obtain a bootstrap population fulfilling the null hypothesis, the residual variables $Z^m = Y^m - \underline{X}\,\underline{\widehat{a}}_m^{'}$, $Z^l = g(Y^l) - \underline{X}\,\underline{\widehat{a}}_l^{'}$ and $Z^r = h(Y^r) - \underline{X}\,\underline{\widehat{a}}_r^{'}$ must be considered. A sample of size n with replacement $\left\{(\underline{X}_i^*, Z_i^{m*}, Z_i^{l*}, Z_i^{r*})\right\}_{i=1,\ldots,n}$ from the bootstrap population is drawn and the bootstrap statistic to be used is

$$T_n^* = n\frac{\sum_{i=1}^n D^2_{\lambda\rho}(\widehat{Z_i^{*T}}, \overline{Z^{*T}})}{\sigma^2_{Y^T}},$$

where $Z^{*T}_i = (Z_i^{m*}, Z_i^{l*}, Z_i^{r*})$.

5.4 Simulation study

Several bootstrap algorithms have been proposed to obtain bootstrap p-values for testing hypotheses about the determination coefficient and the regression parameters of (5). By means of a simulation experiment we aim at investigating whether the obtained p-values work as such, that is, if we find a bootstrap p-value equal to 0.05 we would like to conclude from this that the true p-value (i.e., that obtained if we knew the distribution function) is 0.05. The simulation study concerns the test on a single regression parameter and the linear independence test. During the experiment we employ $B = 1{,}000$ replications of the bootstrap estimator and we carry out 10.000 iterations of the test at three different nominal significance levels ($\alpha = 0.01, 0.05, 0.1$)

Table 1 Empirical percentages of rejection under the hypothesis $H_0 : a^1_{mm} = 0$

$n \setminus \alpha \times 100$	1	5	10
30	0.75	4.13	8.81
50	1.06	5.67	10.14
100	1.28	5.55	10.85
200	1.28	5.42	10.57
300	1.16	5.62	10.12

Table 2 Empirical percentages of rejection under the hypothesis of linear independence

$n \setminus \alpha \times 100$	1	5	10
30	0.31	2.6	6.87
50	0.79	4.74	9.59
100	1.13	5.65	10.63
200	1.31	5.35	10.77
300	1.09	4.92	10.01

for different sample sizes ($n = 30, 50, 100, 200, 300$). We consider the case of two LR fuzzy explanatory variables $\widetilde{X}_1$ and $\widetilde{X}_2$. We deal with the following real random variables: X_1^m and X_2^m, behaving as $N(0, 1)$ random variables, X_1^l and X_2^l as χ^2_1, X_1^r and X_2^r as χ^2_2, Y_m as $N(0, 1)$, $Y_2 = g(Y_l)$ and $Y_3 = h(Y_r)$ as $N(0, 0.5)$.

With respect to the hypothesis testing on a single parameter, we want to test $H_0 : a^1_{mm} = 0$ against the alternative $H_1 : a^1_{mm} \neq 0$. The empirical percentages of rejection under H_0 are given in Table 1. With respect to the linear independence test ($H_0 : R^2 = 0$ against $H_1 : R^2 > 0$), the empirical percentages of rejection under H_0 are reported in Table 2. All in all, from Tables 1 and 2 we can conclude that the bootstrap p-values are fairly good approximations of the true p-values in most cases. As one may expect, this especially holds for increasing values of n ($n > 30$).

6 A real-case study

In order to evaluate the students' satisfaction of a course their subjective judgements/perceptions are observed on a sample of $n = 64$ students. To formalize the problem we define $\Omega = \{sets\ of\ students\ that\ attend\ the\ course\}$ endowed with the Borel σ-field. Since the observations are arbitrarily chosen, P is the uniform distribution over Ω. For any $i \in \Omega$, four characteristics are observed. These are the overall assessment of the course, the assessment of the teaching staff, the assessment of the course content and the average mark (single-valued variable). We assume that the students' judgements/perceptions (first three variables) are intrinsically imprecise. Therefore, we manage them in terms of fuzzy variables, in particular of triangular type (hence $\lambda = \rho = 1/2$). We obtain membership functions by collecting them in a direct way. In fact, to represent the subjective judgements/perceptions, the students are invited to draw a triangular fuzzy number for every characteristic on a segment from 0 (dissatisfaction) to 100 (full satisfaction). More specifically, the students are informed to

Table 3 Overall assessment of the course (Y^m, Y^l, Y^r), assessment of the teaching staff (X_1^m, X_1^l, X_1^r), assessment of the course content (X_2^m, X_2^l, X_2^r) of the course, average mark (X_3)

Y^m	Y^l	Y^r	X_1^m	X_1^l	X_1^r	X_2^m	X_2^l	X_2^r	X_3
93	7	7	87	9	7	75	10	8	27
90	10	10	80	10	10	60	10	30	26
80	20	10	80	10	20	40	20	13	27
76	18	14	77	17	15	50	15	15	28
52	11	12	75	10	5	88	18	2	28.5
90	10	10	86	12	11	80	13	17	28.5
90	10	10	94	7	6	67	10	14	27.5
80	10	20	90	10	10	81	16	19	28
80	10	10	80	10	10	80	10	10	28
70	10	15	80	10	20	50	10	10	28.7
80	3	3	93	4	7	72	6	8	29
..	..	..	..	..	..	..	..	..	..

choose the support of the fuzzy datum (0-level set) as the set of all values that the student considers to be compatible with her/his subjective judgement/perception and the center as the most compatible one (1- level set). One may wonder whether it is fruitful drawing a fuzzy set rather than simply choosing a single value expressing the compatibility with the individual judgement/perception. In our way of thinking there does not exist a single value able to well characterize the individual judgement/perception. In fact, the students may not feel that a single value is capable to *fully* capture her (his) judgement/perception. Also note that this way of fuzzifying has been already suggested by González-Rodríguez et al. (2011).

For analyzing the linear relationship of the overall assessment of the course ($\widetilde{Y}$) on the assessment of the teaching staff ($\widetilde{X}_1$), the assessment of the course contents ($\widetilde{X}_2$) and the average mark (X_3) (see Table 3), the proposed linear regression model is employed.

To overcome the problem about the non-negativity of spreads estimates, we use the logarithmic transformation (that is, $g = h = \ln$). Through the LS procedure we obtain the following estimated model

$$\begin{cases} \widehat{Y^m} = 1.08X_1^m + 0.13X_1^l - 0.07X_1^r \\ \quad -0.17X_2^m - 0.89X_2^l + 0.66X_2^r - 1.12X_3 + 34.06 \\ \widehat{Y^l} = \exp(0.01X_1^m + 0.02X_1^l + 0.02X_1^r \\ \quad +0.00X_2^m + 0.03X_2^l + 0.01X_2^r - 0.00X_3 + 0.67) \\ \widehat{Y^r} = \exp(0.00X_1^m + 0.03X_1^l - 0.02X_1^r \\ \quad -0.01X_2^m + 0.03X_2^l + 0.01X_2^r + 0.04X_3 + 1.01) \end{cases} \tag{20}$$

For the estimated model it results $\widehat{R}^2 = 0.77$, hence approximately 77% of the total variation of the overall assessment of the course is explained by the model. Furthermore, by applying the bootstrap procedure to test the linear independence (with

Table 4 Hypothesis testing on each regression parameter

	$p-$value		$p-$value		$p-$value
a^1_{mm}	0	a^1_{lm}	0.094	a^1_{rm}	0.630
a^1_{ml}	0.721	a^1_{ll}	0.386	a^1_{rl}	0.203
a^1_{mr}	0.753	a^1_{lr}	0.088	a^1_{rr}	0.234
a^2_{mm}	0.036	a^2_{lm}	0.852	a^2_{rm}	0.026
a^2_{ml}	0.0001	a^2_{ll}	0.017	a^2_{rl}	0.091
a^2_{mr}	0	a^2_{lr}	0.149	a^2_{rr}	0.099
a^3_m	0.227	a^3_l	0.915	a^3_r	0.233
b_m	0.213	b_l	0.419	b_r	0.334

The underlined values are significant at $\alpha = 0.05$

$B = 1{,}000$) a p-value equal to 0 is obtained, so the null hypothesis should be rejected. We then test the significance of every single regression parameter by computing the bootstrap $p-$values ($B = 1{,}000$) given in Table 4.

With respect to the model for the center of $\widetilde{Y}$, we can see that, considering a significance level $\alpha = 0.05$, both the centers of $\widetilde{X}_1$ and $\widetilde{X}_2$ are significant. As one may expect, the center of X_1 is positively related to the center of the response ($a^1_{mm} = 1.08$). Surprisingly, this is not the case for the center of X_2($a^2_{mm} = -0.17$), namely, as the assessment of the course contents increases, the overall assessment of the course decreases. Furthermore, also the spreads of $\widetilde{X}_2$ significantly affect the response Y^m. In details, higher values of the left spreads (that is, in case of more imprecision on the lower values of the assessment of the course content) lead to lower values of the center of the response ($a^2_{ml} = -0.89$). The opposite comment holds for the right spread ($a^2_{mr} = 0.66$). The models for the left and right spreads of the response provide additional information about the imprecision of the predicted values. From Table 4 we can observe that only some of the components of $\widetilde{X}_2$ are significantly related to the transformed spreads of $\widetilde{Y}$. In particular, there exists a significant positive relationship between the left spread of $\widetilde{Y}$ and the one of $\widetilde{X}_2$ ($a^2_{ll} = 0.03$). As the imprecision on the lower values of $\widetilde{X}_2$ increases, the one of $\widetilde{Y}$ also increases (in the logarithmic scale). Moreover, a significant negative relationship between Y^r and X^m_2 is found. All in all, we can conclude that the spreads information of the explanatory variables, which is usually arbitrarily ignored, plays a relevant role in explaining the response variable $\widetilde{Y}$. This holds for the model for the center of $\widetilde{Y}$ as well as for the models for the spreads. Therefore, in case of imprecise data, the use of the spreads information seems to be advisable.

To further investigate the potentialities of our model, we compare the above-described results with those obtained by applying classical regression. In other words, we aim at studying whether considering the imprecision of the explanatory variables makes the proposed model more powerful than the classical regression model, which ignores the embodied imprecision. We think that the latter is inappropriate since the analysis of the intrinsic data complexity would be missed, hence leading to conclusions that may be incomplete at best. We now corroborate this claim by inspecting

the predictive power of the here-proposed regression model in comparison with the one of classical linear regression. Specifically, in order to check it, we compare the predicted power of the model related to the center Y^m (the first model in (5)) with the one of the classical regression model with "explanatory variables" X_1^m, X_2^m and X_3. We obtain the following estimated model

$$\widehat{Y^m} = 1.21X_1^m - 0.26X_2^m - 1.59X^3 + 41.02$$

and, by means of classical t-tests, we find that all the explanatory variables are significant (considering $\alpha = 0.05$). It is interesting to stress that, once again, the estimated coefficient of X_2^m (-0.26) indicates that the assessment of the course contents is negatively related to the overall assessment of the course. In order to evaluate the predictive power of the models the K-fold cross-validation procedure is performed (see, for more details, Hastie et al. 2009), where only the significant variables for each model are considered. It consists in splitting the data into K roughly equal-sized parts. For the k-th part we calculate the predicted values of the response considering the regression parameters estimated by using the remaining $K - 1$ parts. By indicating with $\widehat{Y}_i^{m(-k(i))}$ the fitted response, computed with the kth part of the data removed, the cross-validation estimate of the prediction error is

$$CV = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i^m - \widehat{Y}_i^{m(-k(i))}\right)^2.$$

We set $K = 8$ (obtaining 8 subsamples of size 8) and it results that the CV of the proposed model (39.05) is lower than the CV of the classical one (62.76). We can thus conclude that, in this application, the proposed model works better than the classical one for predictive purposes.

7 Concluding remarks

In this work a new linear regression model for data affected by different sources of uncertainty has been introduced. In particular, through a formalization in terms of FRVs, we have coped with the imprecision and the randomness embodied in the data. In order to handle the non-negativity of the spreads of the response we have considered suitable transformation functions allowing us to get analytic estimators of the regression parameters according to a least squares approach. Some inferential procedures for such estimators have been developed. In particular, tests on their significance have been developed by bootstrapping. The goodness of fit of the model has been evaluated by proposing a suitable determination coefficient and the corresponding estimator together with a bootstrap linear independence test. An application to real data affected by imprecision and randomness has shown the capability of the proposed regression model. In fact, we found that it is more powerful than the classical regression technique for predictive purposes. Therefore, if imprecise and random data occur, our regression model seems to be a valuable choice. On the basis of these good results, in the near future, it will be interesting to develop a suitable selection procedure to obtain

the appropriate number of explanatory variables and to address the multicollinearity problem. Future research may also focus on proposing a linearity test in order to check the linearity of the relationship between the response and the explanatory variables.

Acknowledgments The first author has been partially supported by the Spanish Ministry of Education and Science Grant MTM2009-09440-C02-02. The second author has been partially supported by the Sapienza Research Grant "Modelli di gestione dell'incertezza nell'analisi di dati osservazionali e sperimentali (Progetti di Ricerca - Anno 2010)". Their financial support is gratefully acknowledged.

References

Arefi M, Viertl R, Taheri SM (2011) Fuzzy density estimation. Metrika. doi:10.1007/s00184-010-0311-y
Arnold BF, Stahlecker P (2010) A surprising property of uniformly best linear affine estimation in linear regression when prior information is fuzzy. J Stat Plan Inference 140:954–960
Bargiela A, Pedrycz W, Nakashima T (2007) Multiple regression with fuzzy data. Fuzzy Sets Syst 158:2169–2188
Celminš A (1987) Multidimensional least-squares fitting of fuzzy models. Math Model 9:669–690
Chang PT, Lee ES (1996) A generalized fuzzy weighted least-squares regression. Fuzzy Sets Syst 82:289–298
Coppi R, D' Urso P, Giordani P, Santoro A (2006) Least squares estimation of a linear regression model with LR fuzzy response. Comput Stat Data Anal 51:267–286
Diamond P (1988) Fuzzy least squares. Inf Sci 46:141–157
D' Urso P (2003) Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Comput Stat Data Anal 42:47–72
Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, New York
Ferraro MB, Colubi A, González-Rodríguez G, Coppi R (2011) A determination coefficient for a linear regression model with imprecise response. Environmetrics 22:516–529
Ferraro MB, Coppi R, González-Rodríguez G, Colubi A (2010) A linear regression model for imprecise response. Int J Approx Reason 51:759–770
González-Rodríguez G, Blanco A, Colubi A, Lubiano MA (2009) Estimation of a simple linear regression model for fuzzy random variables. Fuzzy Sets Syst 160:357–370
González-Rodríguez G, Colubi A, Gil MA (2011) Fuzzy data treated as functional data: a one-way ANOVA test approach. Comput Stat Data Anal. doi:10.1016/j.csda.2010.06.013
Guo P, Tanaka H (2006) Dual models for possibilistic regression analysis. Comput Stat Data Anal 51:253–266
Hanss M (2005) Applied fuzzy arithmetic—an introduction with engineering applications. Springer, Berlin
Hastie T, Tibshirani RJ, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
Klir GJ (2006) Uncertainty and information: foundations of generalized information theory. Wiley, New York
Körner R, Näther W (1998) Linear regression with random fuzzy variables: extended classical estimates, best linear estimates, least squares estimates. Inf Sci 109:95–118
Krätschmer V (2006a) Strong consistency of least-squares estimation in linear regression models with vague concepts. J Multivar Anal 97:633–654
Krätschmer V (2006b) Limit distributions of least squares estimators in linear regression models with vague concepts. J Multivar Anal 97:1044–1069
Kruse R, Meyer KD (1987) Statistics with vague data. Kluwer, Dortrecht
Lawson CL, Hanson RJ (1995) Solving least squares problems. Classics in applied mathematics 15. SIAM, Philadelphia
Liew CK (1976) Inequality constrained least-squares estimation. J Am Stat Assoc 71:746–751
Lu J, Wang R (2009) An enhanced fuzzy linear regression model with more flexible spreads. Fuzzy Sets Syst 160:2505–2523
Näther W (2006) Regression with fuzzy random data. Comput Stat Data Anal 51:235–252
Näther W, Wünsche A (2007) On the conditional variance of fuzzy random variables. Metrika 65:109–122

Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. Ann Probab 13:1373–1379

Puri ML, Ralescu DA (1986) Fuzzy random variables. J Math Anal Appl 114:409–422

Ramos-Guajardo AB, Colubi A, González-Rodríguez G, Gil MA (2010) One-sample tests for a generalized Fréchet variance of a fuzzy random variable. Metrika 71:185–202

Tanaka H, Ishibuchi H, Yoshikawa S (1995) Exponential possibility regression analysis. Fuzzy Sets Syst 69:305–318

Tanaka H, Uejima S, Asai K (1982) Linear regression analysis with fuzzy model. IEEE Trans Syst Man Cybern 12:903–907

Tanaka H, Watada J (1988) Possibilistic linear systems and their application to the linear regression model. Fuzzy Sets Syst 27:275–289

Yang MS, Ko CH (1996) On a class of fuzzy *c*-numbers clustering procedures for fuzzy data. Fuzzy Sets and Syst 84:49–60

Zadeh LA (1965) Fuzzy sets. Inf Control 8:338–353

Zimmermann HJ (2001) Fuzzy set theory and its applications. Kluwer, Dordrecht