A prediction error for a linear regression model with fuzzy random elements

Maria Brigida Ferraro

Abstract The concept of a fuzzy random variable (FRV) arises in order to jointly consider two kinds of uncertainty: the randomness and the imprecision. A linear regression model with fuzzy random elements and the related least squares estimation problem are briefly recalled. To check the adequacy of the model from a predictive point of view a prediction error is introduced and investigated. In details, it is defined by means of an appropriate metric and it is estimated by means of a cross-validation procedure.

Key words: LR fuzzy random variable, linear regression model, prediction error, cross-validation

1 Preliminaries

Given a universe U of elements, a fuzzy set $\widetilde{A}$ is defined through the so-called *membership function* $\mu_{\widetilde{A}} : U \to [0,1]$. For a generic $x \in U$, the membership function expresses the extent to which x belongs to $\widetilde{A}$. Such a degree ranges from 0 (complete non-membership) to 1 (complete membership) [6].

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 $\mathscr{F}_{LR}$. An *LR* fuzzy number $\widetilde{A}$ is determined by three real-valued parameters, namely, the center (A^m) and the (non-negative) left and right spreads (A^l and A^r, respectively). The membership function of $\widetilde{A} \in \mathscr{F}_{LR}$ can be written as

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

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

where the functions L and R are particular decreasing shape functions from R^+ to $[0,1]$ such that $L(0) = R(0) = 1$ and $L(x) = R(x) = 0, \forall x \in R \setminus [0,1]$, and 1_I is the indicator function of a set I. $\widetilde{A}$ is a *triangular* fuzzy number if $L(z) = R(z) = 1 - z$, for $0 \leq z \leq 1$. The operations considered in $\mathscr{F}_{LR}$ are the natural extensions of the Minkowski sum and the product by a positive scalar for interval. In particular, the sum of $\widetilde{A}$ and $\widetilde{B}$ in $\mathscr{F}_{LR}$ 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 $\widetilde{A} \in \mathscr{F}_{LR}$ by a scalar $\gamma > 0$ is the *LR* fuzzy number $\gamma\widetilde{A}$ so that

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

Yang & Ko [5] define a distance between two *LR* fuzzy numbers $\widetilde{X}$ and $\widetilde{Y}$ as follows

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

where the parameters $\lambda = \int_0^1 L^{-1}(\omega)\mathrm{d}\omega$ and $\rho = \int_0^1 R^{-1}(\omega)\mathrm{d}\omega$ represent the influence of the shape of the membership function. For what follows it is necessary to embed the space $\mathscr{F}_{LR}$ into $\mathbb{R}^3$ by preserving the metric. For this reason a generalization of the Yang and Ko metric can be derived [1]. Given $a = (a_1, a_2, a_3)$ and $b = (b_1, b_2, b_3) \in \mathbb{R}^3$, it is $D^2_{\lambda\rho}(a,b) = (a_1 - b_1)^2 + ((a_1 - \lambda a_2) - (b_1 - \lambda b_2))^2 + ((a_1 + \rho a_3) - (b_1 + \rho b_3))^2$, where $\lambda, \rho \in R^+$. $D^2_{\lambda\rho}$ will be used in the sequel as a tool for quantifying errors in the regression models we are going to introduce.

Let $(\Omega, \mathscr{A}, P)$ be a probability space. In this context, a mapping $\widetilde{X} : \Omega \to \mathscr{F}_{LR}$ is an *LR* FRV if $(X^m, X^l, X^r) : \Omega \to \mathbb{R} \times \mathbb{R}^+ \times \mathbb{R}^+$ is a random vector [4]. The expectation of an *LR* FRV $\widetilde{X}$ is the unique fuzzy set $E(\widetilde{X})$ ($\in \mathscr{F}_{LR}$) whose parameters are $E(X^m)$, $E(X^l)$ and $E(X^r)$.

2 A linear regression model with fuzzy random elements

Consider a random experiment in which an *LR* fuzzy response variable $\widetilde{Y}$ and p *LR* fuzzy explanatory variables $\widetilde{X}_1, \widetilde{X}_2, ..., \widetilde{X}_p$ are observed on a random sample of n statistical units, $\{\widetilde{Y}_i, \widetilde{X}_{1i}, \widetilde{X}_{2i}, ..., \widetilde{X}_{pi}\}_{i=1,...,n}$. We consider the shape of the membership functions as fixed, 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. We face the non-negativity constraints of the spreads of the response variable by introducing two invertible functions $g : (0,+\infty) \longrightarrow \mathbb{R}$ and $h : (0,+\infty) \longrightarrow \mathbb{R}$. We jointly consider three regression models whose responses are, respectively, the center of the fuzzy response variable and two transforms of the left and the right spreads, and the explanatory variables are the center, the left and the right spreads of each fuzzy explanatory variables. We have

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

where $\underline{X} = (X_1^m, X_1^l, X_1^r, ..., X_p^m, X_p^l, X_p^r)$ is the row-vector of length $3p$ of all the components of the explanatory variables, ε_m, ε_l and ε_r are real-valued random variables with $E(\varepsilon_m|\underline{X}) = E(\varepsilon_l|\underline{X}) = E(\varepsilon_r|\underline{X}) = 0$, $\underline{a}_m = (a_{mm}^1, a_{ml}^1, a_{mr}^1, ..., a_{mm}^p, a_{ml}^p, a_{mr}^p)$, $\underline{a}_l = (a_{lm}^1, a_{ll}^1, a_{lr}^1, ..., a_{lm}^p, a_{ll}^p, a_{lr}^p)$ and $\underline{a}_r = (a_{rm}^1, a_{rl}^1, a_{rr}^1, ..., a_{rm}^p, a_{rl}^p, a_{rr}^p)$ are row-vectors of length $3p$ of the parameters related to $\underline{X}$. Finally, b_m, b_l, b_r denote the intercepts.
Under the assumptions of model (2), the solution of the LS problem is $\widehat{\underline{a}}_m^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}\underline{Y}^{mc}$, $\widehat{\underline{a}}_l^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}g(\underline{Y}^l)^c$, $\widehat{\underline{a}}_r^{'} = (\mathbf{X}^{c'}\mathbf{X}^c)^{-1}\mathbf{X}^{c'}h(\underline{Y}^r)^c$, $\hat{b}_m = \overline{Y^m} - \overline{\underline{X}}\,\hat{\underline{a}}_m^{'}$, $\hat{b}_l = \overline{g(Y^l)} - \overline{\underline{X}}\,\hat{\underline{a}}_l^{'}$ and $\hat{b}_r = \overline{h(Y^r)} - \overline{\underline{X}}\,\hat{\underline{a}}_r^{'}$, where $\underline{Y}^{mc} = \underline{Y}^m - \underline{1}\overline{Y^m}$, $g(\underline{Y}^l)^c = g(\underline{Y}^l) - \underline{1}\overline{g(Y^l)}$, $h(\underline{Y}^r)^c = h(\underline{Y}^r) - \underline{1}\overline{h(Y^r)}$ 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 $\underline{X}$. 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 [2].

3 Prediction error

From a predictive point of view to check the adequacy of our model it is important to introduce a prediction error. We need a training set to estimate the regression parameters and a test set to evaluate the regression model by means of the prediction error. We indicate with $\left\{\widetilde{Y}_i^{TR}, \widetilde{X}_{1i}^{TR}, \widetilde{X}_{2i}^{TR}, ..., \widetilde{X}_{pi}^{TR}\right\}_{i=1,...,n_{TR}}$ the training set and with $\{\widetilde{Y}_i^{TS}, \widetilde{X}_{1i}^{TS}, \widetilde{X}_{2i}^{TS}, ..., \widetilde{X}_{pi}^{TS}\}_{i=1,...,n_{TS}}$ the test set. By means of the distance $D_{\lambda\rho}^2$, the prediction error is defined as the expected value of the distance between the observed values of the fuzzy response in the test set and the fitted values of the response computed by means of the estimators obtained in the training set and the explanatory variables observed in the test set. In details,

$$\begin{aligned} PE = E\Bigg(& \left\| \underline{Y}^{mTS} - \left(\mathbf{X}^{TS}\widehat{\underline{a}}_m^{TR'} + \underline{1}\,\widehat{b}_m^{TR} \right) \right\|^2 \\ & + \left\| \left(\underline{Y}^{mTS} - \lambda g(\underline{Y}^{lTS}) \right) - \left(\mathbf{X}^{TS}\widehat{\underline{a}}_m^{TR'} + \underline{1}\,\widehat{b}_m^{TR} - \lambda \left(\mathbf{X}^{TS}\widehat{\underline{a}}_l^{TR'} + \underline{1}\,\widehat{b}_l^{TR} \right) \right) \right\|^2 \\ & + \left\| \left(\underline{Y}^{mTS} + \rho h(\underline{Y}^{rTS}) \right) - \left(\mathbf{X}^{TS}\widehat{\underline{a}}_m^{TR'} + \underline{1}\,\widehat{b}_m^{TR} + \rho \left(\mathbf{X}^{TS}\widehat{\underline{a}}_r^{TR'} + \underline{1}\,\widehat{b}_r^{TR} \right) \right) \right\|^2 \Bigg), \end{aligned} \tag{3}$$

where $\underline{a}_m^{TR}$, $\underline{a}_l^{TR}$, $\underline{a}_r^{TR}$, b_m^{TR}, b_l^{TR} and b_r^{TR} are the estimators of the regression parameters obtained in the training set. In practice, there are different approaches to estimate the prediction error. Here a K-fold cross-validation procedure is performed [3]. It consists in splitting the data into K roughly equal-sized parts. For the k-th part we calculate the predicted/fitted values of the response considering the regression parameters estimated by using the remaining $K-1$ parts. That is, the k-th part is considered as test set and the remaining $K-1$ parts are the training set. This procedure is repeated K times. In details, the estimated prediction error by means of a cross validation procedure is

$$\widehat{PE}_{CV} = \frac{1}{K}\sum_{k=1}^{K} Err_k, \tag{4}$$

where

$$\begin{aligned} Err_k = & \left(\left\| \underline{Y}_k^{mTS} - \left(\mathbf{X}_k^{TS} \widehat{\underline{a}}_m^{TR-k'} + \underline{1}\widehat{b}_m^{TR-k} \right) \right\|^2 \right. \\ & + \left\| \left(\underline{Y}_k^{mTS} - \lambda g(\underline{Y}_k^{lTS}) \right) - \left(\mathbf{X}_k^{TS} \widehat{\underline{a}}_m^{TR-k'} + \underline{1}\widehat{b}_m^{TR-k} - \lambda \left(\mathbf{X}_k^{TS} \widehat{\underline{a}}_l^{TR-k'} + \underline{1}\widehat{b}_l^{TR-k} \right) \right) \right\|^2 \\ & \left. + \left\| \left(\underline{Y}_k^{mTS} + \rho h(\underline{Y}_k^{rTS}) \right) - \left(\mathbf{X}_k^{TS} \widehat{\underline{a}}_m^{TR-k'} + \underline{1}\widehat{b}_m^{TR-k} + \rho \left(\mathbf{X}_k^{TS} \widehat{\underline{a}}_r^{TR-k'} + \underline{1}\widehat{b}_r^{TR-k} \right) \right) \right\|^2 \right), \end{aligned} \tag{5}$$

with $\widehat{\underline{a}}_m^{TR-k}$, $\widehat{\underline{a}}_l^{TR-k}$, $\widehat{\underline{a}}_r^{TR-k}$, $\widehat{b}_m^{TR-k}$, $\widehat{b}_l^{TR-k}$, $\widehat{b}_r^{TR-k}$ that are the regression parameters estimated in the training set obtained by removing the k-th part and $\{\underline{Y}_k^{mTS}, \underline{Y}_k^{lTS}, \underline{Y}_k^{rTS}, \mathbf{X}_k^{TS}\}$ is the test set obtained by considering the k-th part.

References

1. Ferraro, M.B., Coppi, R., Gonzalez-Rodriguez, G., Colubi, A.: A linear regression model for imprecise response. Int. J. Approx. Reason. **51**, 759–770 (2010).
2. Ferraro, M.B., Giordani, P.: A multiple linear regression model for LR fuzzy random variables. Metrika (2012) doi: 10.1007/s00184-011-0367-3.
3. Hastie, T., Tibshirani, R.J., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York (2009).
4. Puri, M.L., Ralescu, D.A.: Fuzzy random variables. J. Math. Anal. Appl. **114**, 409–422 (1986).
5. 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).
6. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 338–353 (1965).