A STEPWISE PROCEDURE TO SELECT VARIABLES IN A FUZZY LEAST SQUARE REGRESSION MODEL

Francesco Campobasso and Annarita Fanizzi
Department of Statistical Sciences "Carlo Cecchi", University of Bari, Bari, Italy

Keywords: Fuzzy least square regression, Multivariate generalization, Asymmetric fuzzy intercept, Total sum of squares, Goodness of fit, Stepwise procedure.

Abstract: Fuzzy regression techniques can be used to fit fuzzy data into a regression model. Diamond treated the case of a simple model introducing a metrics into the space of triangular fuzzy numbers. In previous works we provided some theoretical results about the estimates of a multiple regression model with a non-fuzzy intercept; in this paper we show how the sum of squares of the dependent variable can be decomposed in exactly the same way as the classical OLS estimation procedure only when the intercept is fuzzy asymmetric. Such a decomposition allows us to introduce a stepwise procedure which simplifies, in terms of computational, the identification of the most significant independent variables in the model.

1 INTRODUCTION

Modalities of quantitative variables are commonly given as exact single values, although sometimes they cannot be precise. The imprecision of measuring instruments and the continuous nature of some observations, for example, prevent researcher from obtaining the corresponding true values.

On the other hand qualitative variables are commonly expressed using common linguistic terms, which also represent verbal labels of sets with uncertain borders.

The appropriate way to manage such an uncertainty of observations is provided by using fuzzy numbers.

In 1988 P. M. Diamond introduced a metric onto the space of triangular fuzzy numbers and derived the expression of the estimated coefficients in a simple fuzzy regression of an uncertain dependent variable on a single uncertain independent variable.

Starting from a multivariate generalization of this regression, we provided in previous works some results on the decomposition of the deviance of the dependent variable according to Diamond's metric.

2 THE FUZZY LEAST SQUARE REGRESSION

A *triangular fuzzy number* $\tilde{X} = (x, x_L, x_R)_T$ for the variable X is characterized by a function $\mu_{\tilde{X}} : X \to [0,1]$, like the one represented in Fig. 1, that expresses the *membership degree* of any possible value of X to $\tilde{X}$.

The accumulation value x is considered the *core* of the fuzzy number, while $\underline{\xi} = x_R - x$ and $\overline{\xi} = x - x_L$ are considered the *left spread* and the *right spread* respectively.

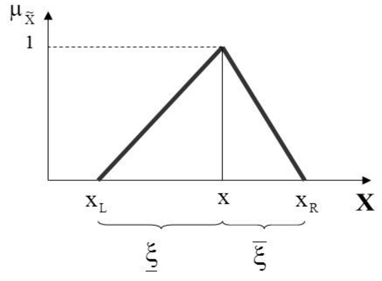

Figure 1: Representation of a triangular fuzzy number.

Note that x belongs to $\tilde{X}$ with the highest degree (equal to 1), while the other values included between

Campobasso F. and Fanizzi A..
A STEPWISE PROCEDURE TO SELECT VARIABLES IN A FUZZY LEAST SQUARE REGRESSION MODEL.
DOI: 10.5220/0003720504170426
In *Proceedings of the International Conference on Evolutionary Computation Theory and Applications* (FCTA-2011), pages 417-426
ISBN: 978-989-8425-83-6

the *left extreme* x_L and the *right extreme* x_R belong to $\tilde{X}$ with a gradually lower degree.

The set of triangular fuzzy numbers is closed under addition: given two triangular fuzzy numbers $\tilde{X} = (x, x_L, x_R)_T$ and $\tilde{Y} = (y, y_L, y_R)_T$, their sum $\tilde{Z}$ is still a triangular fuzzy number $\tilde{Z} = \tilde{X} + \tilde{Y} = (x + y, x_L + y_L, x_R + y_R)_T$. Moreover the opposite of a triangular fuzzy number $\tilde{X} = (x, x_L, x_R)_T$ is $-\tilde{X} = (-x, -x_R, -x_L)_T$.

It follows that, given n fuzzy numbers $\tilde{X}_i = (x_i, x_{Li}, x_{Ri})_T$, i =1, 2, .., n, their average is $\overline{X} = \frac{\sum \tilde{X}_i}{n} = \left(\frac{\sum x_i}{n}, \frac{\sum x_{Li}}{n}, \frac{\sum x_{Ri}}{n}\right)_T$.

Diamond (1988) introduced a metrics onto the space of triangular fuzzy numbers; according to this metrics, the squared distance between $\tilde{X}$ and $\tilde{Y}$ is

$$d(\tilde{X}, \tilde{Y})^2 = d\big((x, x_L, x_R)_T, (y, y_L, y_R)_T\big)^2 =$$
$$(x-y)^2 + (x_L - y_L)^2 + (x_R - y_R)^2.$$

The same Author treated the fuzzy regression model of a dependent variable $\tilde{Y}$ on a single independent variable $\tilde{X}$, which can be written as

$\tilde{Y} = a + b\tilde{X}$, $\quad a, b \in \mathbb{R}$,

when the intercept a is non-fuzzy, as well as

$\tilde{Y} = \tilde{A} + b\tilde{X} \quad a, b \in \mathbb{R}$,

when the intercept $\tilde{A} = (a, a_L, a_R)_T$ is fuzzy, where it is $a_L = a - \underline{\gamma}$, $a_R = a - \overline{\gamma}$ and $\underline{\gamma}$, $\overline{\gamma} > 0$.

The expression of the corresponding parameters is derived from minimizing the sum $\sum d(\tilde{Y}_i, \tilde{Y}_i^*)^2$ of the squared distances between theoretical and empirical values in n observed units of the fuzzy dependent variable $\tilde{Y}$ with respect to a and b.

Such a sum takes different forms according to the signs of the coefficient b, as the product of a fuzzy number $\tilde{X} = (x, x_L, x_R)_T$ and a real number k depends on whether the latter is positive or negative. by subtracting the right spread from the core.

Diamond demonstrated that the optimization problem has a unique solution under certain conditions.

In previous works we provided some theoretical results about the estimates of the regression coefficients and about the decomposition of the sum of squares of the dependent variable (Campobasso, Fanizzi and Tarantini, 2009) in a multiple regression model. In particular we treated the case of a non-fuzzy intercept, as well as the case of a fuzzy intercept, which seems more appropriate (Campobasso and Fanizzi, 2011) for some reasons which will be clearer later.

3 A MULTIVARIATE GENERALIZATION OF THE REGRESSION MODEL

3.1 A Generalization of the Model Including a Non-fuzzy Intercept

Let's assume to observe a fuzzy dependent variable $\tilde{Y}_i = (y_i, y_{Li}, y_{Ri})_T$ and two fuzzy independent variables, $\tilde{X}_i = (x_i, x_{Li}, x_{Ri})_T$ and $\tilde{Z}_i = (z_i, z_{Li}, z_{Ri})_T$, on a set of n units. The linear regression model is given by

$\tilde{Y}_i$ *= a +b $\tilde{X}_i$ +c $\tilde{Z}_i$, $\quad$ i=1,2, ...,n; a,b,c ∈ IR.

The corresponding parameters are determined by minimizing the sum of Diamond's distances between theoretical and empirical values of the dependent variable

$$\sum d(\tilde{Y}_i, a + b\tilde{X}_i + c\tilde{Z}_i)^2 \tag{1}$$

respect to a, b and c. As we stated above, such a sum assumes different expressions according to the signs of the regression coefficients b and c. This generates the following four cases

<u>*Case 1*</u>: b>0, c>0

$$\sum d(\tilde{Y}_i, a + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a - bx_{Li} - cz_{Li})^2 +$$
$$+ (y_{Ri} - a - bx_{Ri} - cz_{Ri})^2]$$

<u>*Case 2*</u>: b<0, c>0

$$\sum d(\tilde{Y}_i, a + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a - bx_{Ri} - cz_{Li})^2 +$$
$$+ (y_{Ri} - a - bx_{Li} - cz_{Ri})^2]$$

<u>*Case 3*</u>: b>0, c<0

$$\sum d(\tilde{Y}_i, a + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a - bx_{Li} - cz_{Ri})^2 +$$
$$+ (y_{Ri} - a - bx_{Ri} - cz_{Li})^2]$$

<u>*Case 4*</u>: b<0, c<0

$$\sum d(\tilde{Y}_i, a + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a - bx_{Ri} - cz_{Ri})^2 +$$
$$+ (y_{Ri} - a - bx_{Li} - cz_{Li})^2]$$

Let's consider, as an example, case 3 and let's express it in matricial terms. The expression to be minimized is given by

$$G(\beta)=\|y-X\beta\|^2+\|y_L-X_L\beta\|^2+\|y_R-X_R\beta\|^2=$$
$$=(y-X\beta)'(y-X\beta)+(y_L-X_L\beta)'(y_L-X_L\beta)+ \qquad (2)$$
$$+(y_R-X_R\beta)'(y_R-X_R\beta)$$

where

$\mathbf{y}$ = [y_i], is the n-dimensional vector of cores of the dependent variable;

$\mathbf{y_L}$ = [y_{Li}] and $\mathbf{y_R}$ = [y_{Ri}] are the n-dimensional vectors of lower extremes and upper extremes of the dependent variable respectively;

X is the n×3 matrix of cores of the independent variables, formed by vectors **1**, **x** = [x_i], **z** = [z_i];

$\mathbf{X_L}$ is the n×3 matrix of lower bounds of the independent variables, formed by vectors **1**, $\mathbf{x_L}$ = [x_{Li}], $\mathbf{z_R}$ = [z_{Ri}];

$\mathbf{X_R}$ is the n×3 matrix of upper bounds of the independent variables (analogous to $\mathbf{X_L}$), formed by vectors **1**, $\mathbf{x_R}$, $\mathbf{z_L}$;

$\boldsymbol{\beta}$ is the vector (a, b, c) '.

The estimates of the regression coefficients are derived from minimizing G($\boldsymbol{\beta}$) with respect to $\boldsymbol{\beta}$ i.e. from seeking the solutions of the system

$$[\mathbf{X'X}+\mathbf{X_L}'\mathbf{X_L}+\mathbf{X_R}'\mathbf{X_R}]\boldsymbol{\beta}-[\mathbf{y}'\mathbf{X}+\mathbf{y_L}'\mathbf{X_L}+\mathbf{y_R}'\mathbf{X_R}]=0$$

and in particular we obtain

$$\boldsymbol{\beta}=[\mathbf{X'X}+\mathbf{X_L}'\mathbf{X_L}+\mathbf{X_R}'\mathbf{X_R}]^{-1}[\mathbf{X'y}+\mathbf{X_L}'\mathbf{y_L}+\mathbf{X_R}'\mathbf{y_R}].$$

Similarly to OLS estimation procedure, the optimization problem admits a single and finite solution if $[\mathbf{X'X}+\mathbf{X_L}'\mathbf{X_L}+\mathbf{X_R}'\mathbf{X_R}]$ is invertible and the hessian matrix is definite positive.

The found solution $\boldsymbol{\beta}^*$=(a^*, b^*, c^*)', is admissible if the signs of the regression coefficients are coherent with basic assumptions (b >0, c <0).

In the remaining three cases the expression (2) to be minimized is obtained after replacing $\mathbf{z_R}$ by $\mathbf{z_L}$ in $\mathbf{X_L}$ and $\mathbf{z_L}$ by $\mathbf{z_R}$ in $\mathbf{X_R}$ (case 1), $\mathbf{x_L}$ by $\mathbf{x_R}$ and $\mathbf{z_R}$ by $\mathbf{z_L}$ in $\mathbf{X_L}$ and also $\mathbf{x_R}$ by $\mathbf{x_L}$ and $\mathbf{z_L}$ by $\mathbf{z_R}$ in $\mathbf{X_R}$ (case 2), x_L by $\mathbf{x_R}$ in $\mathbf{X_L}$ and $\mathbf{x_R}$ by x_L in $\mathbf{X_R}$ (case 4) respectively.

The optimum solution corresponds to that (admissible) one which makes minimum (1) among all.

The generalization of such a procedure to the case of several independent variables is immediate and that the number of solutions to analyse, in order to identify the optimum one, growths exponentially with the considered number of variables. For example, if the model includes k independent variables, 2^k possible cases must be taken into account, which derive from combining the signs of the regression coefficients.

3.2 A Generalization of the Model Including a Fuzzy Intercept

Now we analyze an extension of the model with a *fuzzy* intercept, which seems more appropriate than the non-fuzzy one as it expresses the average value of the dependent variable (which is also *fuzzy*) when the independent variables equal zero.

For this purpose we start from the results obtained by Diamond in the case of the univariate regression model with a fuzzy intercept.

3.2.1 The Univariate Model

Let's regress, for example, the dependent variable $\tilde{Y}_i=(y_i,y_{Li},y_{Ri})_T$ on a single independent variable $\tilde{X}_i=(x_i,x_{Li},x_{Ri})_T$ in a set of n units. If we consider a symmetric fuzzy intercept $\tilde{A}=(a,a_L,a_R)_T$, where $a_L=a-\gamma$, $a_R=a+\gamma$ and $\gamma>0$ (if $\gamma=0$, $\tilde{A}$ would be no more fuzzy), the model assumes the following expression:

$$\tilde{Y}_i{}^*=\tilde{A}+b\tilde{X}_i \qquad i = 1, 2, ..., n;\ a, b \in \mathbb{R}.$$

The fuzzy regression parameters are determined by minimizing the sum of the squared Diamond's distances between theoretical and empirical fuzzy values of the dependent variable

$$\sum d(\tilde{A}+b\tilde{X}_i,\tilde{Y}_i)^2$$

respect to a, b and γ.

The function to minimize assumes different expressions according to the sign of the regression coefficients b. Supposing that b > 0, the estimates of a,b and γ are obtained as solutions a^*, b^* and γ^* of the system of equations

$$\begin{cases} a\sum(x_i+x_{Li}+x_{Ri})+\gamma\sum(x_{Ri}-x_{Li})+b\sum(x_i^2+x_{Li}^2+x_{Ri}^2)= \\ \quad =\sum(y_ix_i+y_{Li}x_{Li}+y_{Ri}x_{Ri}) \\ 2n\gamma=\sum[(y_{Ri}-y_{Li})-b(x_{Ri}-x_{Li})] \\ na=\frac{1}{3}\sum[y_i+y_{Li}+y_{Ri}-b(x_i+x_{Li}+x_{Ri})]. \end{cases}$$

Otherwise, supposing b<0, the estimates of a, b and γ are obtained as solutions a_*, b_* and γ_* of the system of equations

$$\begin{cases} a\sum(x_i+x_{Li}+x_{Ri})-\gamma\sum(x_{Ri}-x_{Li})+ \\ b\sum(x_i^2+x_{Li}^2+x_{Ri}^2)=\sum(x_iy_i+y_{Li}x_{Ri}+y_{Ri}x_{Li}) \\ 2n\gamma=\sum[(y_{Ri}-y_{Li})+b(x_{Ri}-x_{Li})] \\ na=\frac{1}{3}\sum[y_i+y_{Li}+y_{Ri}-b(x_i+x_{Li}+x_{Ri})]. \end{cases}$$

As Diamond shows (1988), the solution to such a problem of minimization exists and is unique if the following conditions occur simultaneously:

a) either $b^* < 0$ or $b_* > 0$;

b) $\sum\left[(x_{Ri}-x_{Li})-\frac{1}{n}(x_{Ri}-x_{Li})\right]\left[(y_{Ri}-y_{Li})-\frac{1}{n}(y_{Ri}-y_{Li})\right]\geq 0$;

c) $b^* > b_*$.

3.2.2 The Multivariate Model

Now we generalize the regression model with a fuzzy intercept to the case of more than a single independent variable.

Assuming to regress a dependent variable $\tilde{Y}_i = (y_i, y_{Li}, y_{Ri})_T$ on two independent variables $\tilde{X}_i = (x_i, x_{Li}, x_{Ri})_T$ and $\tilde{Z}_i = (z_i, z_{Li}, z_{Ri})_T$ in a set of n units, the linear regression model including a fuzzy asymmetric intercept $\tilde{A} = (a, a_L, a_R)_T$, where $a_L = a - \underline{\gamma}$, $a_R = a - \bar{\gamma}$ and $\underline{\gamma}$, $\bar{\gamma} > 0$ (if $\underline{\gamma} = \bar{\gamma} =$ 0, $\tilde{A}$ would be no more fuzzy), assumes the following expression:

$$\tilde{Y}_i^* = \tilde{A} + b\,\tilde{X}_i + c\,\tilde{Z}_i, \quad i = 1, 2, ..., n;\ a, b, c \in \mathbb{R}.$$

Note that the asymmetric intercept is more appropriate the symmetric one, which evidently fits the data in a less efficient way.

The corresponding estimates of the parameters are again determined by minimizing the sum of the squared Diamond's distances between empirical and theoretical values of the dependent variable

$$\sum d(\tilde{Y}_i, \tilde{A} + b\tilde{X}_i + c\tilde{Z}_i)^2 \qquad (3)$$

respect to a, b, c, $\underline{\gamma}$ and $\bar{\gamma}$. The function to minimize assumes different expressions according to the signs of the regression coefficients b and c.

Case 1: b>0, c>0

$$\sum d(\tilde{Y}_i, \tilde{A} + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a_L - bx_{Li} - cz_{Li})^2 +$$
$$+ (y_{Ri} - a_R - bx_{Ri} - cz_{Ri})^2]$$

Case 2: b<0, c>0

$$\sum d(\tilde{Y}_i, \tilde{A} + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a_L - bx_{Ri} - cz_{Li})^2 +$$
$$+ (y_{Ri} - a_R - bx_{Li} - cz_{Ri})^2]$$

Case 3: b>0, c<0

$$\sum d(\tilde{Y}_i, \tilde{A} + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a_L - bx_{Li} - cz_{Ri})^2 +$$
$$+ (y_{Ri} - a_R - bx_{Ri} - cz_{Li})^2]$$

Case 4: b<0, c<0

$$\sum d(\tilde{Y}_i, \tilde{A} + b\tilde{X}_i + c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a_L - bx_{Ri} - cz_{Ri})^2 +$$
$$+ (y_{Ri} - a_R - bx_{Li} - cz_{Li})^2]$$

Let's consider, as an example, case 3 and let's express it in matricial terms. The expression to be minimized is given by

$$G(\beta) = \|y - X\beta\|^2 + \|y_L - X_L\beta\|^2 + \|y_R - X_R\beta\|^2 =$$
$$= (y - X\beta)'(y - X\beta) + (y_L - X_L\beta)'(y_L - X_L\beta) + \qquad (4)$$
$$+ (y_R - X_R\beta)'(y_R - X_R\beta)$$

where

y = [y_i], is the n-dimensional vector of cores of the dependent variable;

$\mathbf{y_L}$ = [y_{Li}] and $\mathbf{y_R}$ = [y_{Ri}] are the n-dimensional vectors of lower extremes and upper extremes of the dependent variable respectively;

X is the n×5 matrix of cores of the independent variables, formed by vectors **1**, **x** = [x_i], **z** = [z_i] and two vectors **0**;

$\mathbf{X_L}$ is the n×5 matrix of lower bounds of the independent variables, formed by vectors **1**, $\mathbf{x_L}$ = [x_{Li}], $\mathbf{z_R}$ = [z_{Ri}] and **-1, 0**;

$\mathbf{X_R}$ is the n×5 matrix of upper bounds of the independent variables (analogous to $\mathbf{X_L}$), formed by vectors **1**, $\mathbf{x_R}$, $\mathbf{z_L}$ and **0**, **1**;

β is the vector (a, b, c, $\underline{\gamma}$, $\bar{\gamma}$) '.

The estimates of the regression coefficients are derived from minimizing G(β) with respect to β i.e. from seeking the solutions of the system

$$[\mathbf{X'X} + \mathbf{X_L}'\mathbf{X_L} + \mathbf{X_R}'\mathbf{X_R}]\boldsymbol{\beta} - [\mathbf{y}'\mathbf{X} + \mathbf{y_L}'\mathbf{X_L} + \mathbf{y_R}'\mathbf{X_R}] = 0$$

and in particular we obtain

$$\boldsymbol{\beta} = [\mathbf{X'X} + \mathbf{X_L}'\mathbf{X_L} + \mathbf{X_R}'\mathbf{X_R}]^{-1}[\mathbf{X'y} + \mathbf{X_L}'\mathbf{y_L} + \mathbf{X_R}'\mathbf{y_R}].$$

Similarly to OLS estimation procedure, the optimization problem admits a single and finite solution if $[\mathbf{X'X} + \mathbf{X_L}'\mathbf{X_L} + \mathbf{X_R}'\mathbf{X_R}]$ is invertible and the hessian matrix is definite positive.

The found solution $\boldsymbol{\beta}^*$=(a^*, b^*, c^*, $\underline{\gamma}^*$, $\bar{\gamma}^*$)', is admissible if the signs of the regression coefficients are coherent with basic assumptions, that is b >0, c <0 and $\underline{\gamma}$, $\bar{\gamma}$ > 0.

In the remaining three cases the expression (4) to be minimized is obtained after appropriately replacing the vectors of the left and right extremes in the matrices as described above, according to the case considered. The optimum solution corresponds to that (admissible) one which makes minimum (3) among all.

When the intercept is symmetric, we estimate a parameter less than the previous model, because the spreads left and right coincide (Campobasso and Fanizzi, 2011). Note that the matrices **X**, $\mathbf{X_L}$ and $\mathbf{X_R}$, relative to independent variables, and the vector of parameters **β** change their expression. In particular we have that

X is the n×4 matrix of cores of the independent variables, formed by vectors **1**, **x** = [x_i], **z** = [z_i] and **0**;

$\mathbf{X_L}$ is the n×4 matrix of lower bounds of the independent variables, formed by vectors **1**, $\mathbf{x_L}$ = [x_{Li}], $\mathbf{z_R}$ = [z_{Ri}] and **-1**;

$\mathbf{X_R}$ is the n×4 matrix of upper bounds of the independent variables (analogous to $\mathbf{X_L}$), formed by vectors **1**, $\mathbf{x_R}$, $\mathbf{z_L}$ and **1**;

β is the vector (a, b, c, γ) '.

4 DECOMPOSITION OF THE TOTAL SUM OF SQUARES OF THE DEPENDENT VARIABLE

In this section two important theoretical results will be demonstrated: the first one regards the inequality between theoretical and empirical averages of the fuzzy dependent variable (unlike in the classical OLS estimation procedure); the second one regards the decomposition of the total sum of squares of the dependent variable, which involves other two additive components besides the regression and the residual sum of squares.

4.1 The Model Including a Non-fuzzy Intercept

Let's consider, only for example, the sum of Diamond's distances between theoretical and empirical values of the dependent variable in the case 3:

$$\sum d(\tilde{Y}_i, a+b\tilde{X}_i+c\tilde{Z}_i)^2 =$$
$$= \sum[(y_i - a - bx_i - cz_i)^2 + (y_{Li} - a - bx_{Li} - cz_{Ri})^2 +$$
$$+ (y_{Ri} - a - bx_{Ri} - cz_{Li})^2]$$

Setting equal to 0 the derivate of $\sum d(\tilde{Y}_i, a+b\tilde{X}_i+c\tilde{Z}_i)^2$ respect to a, b and c, we can obtain the following system of equations:

$$\begin{cases} -2\sum[(y_i - a - bx_i - cz_i) + (y_{Li} - a - bx_{Li} - cz_{Ri}) + \\ \quad + (y_{Ri} - a - bx_{Ri} - cz_{Li})] = 0 \\ -2\sum[(y_i - a - bx_i - cz_i)x_i + (y_{Li} - a - bx_{Li} - cz_{Ri})x_{Li} + \\ \quad + (y_{Ri} - a - bx_{Ri} - cz_{Li})x_{Ri}] = 0 \\ -2\sum[(y_i - a - bx_i - cz_i)z_i + (y_{Li} - a - bx_{Li} - cz_{Ri})z_{Ri} + \\ \quad + (y_{Ri} - a - bx_{Ri} - cz_{Li})z_{Li}] = 0 \end{cases}$$

Such a system can be written as

$$\begin{cases} \sum(a + bx_i + cz_i) + \sum(a + bx_{Li} + cz_{Ri}) + \\ + \sum(a + bx_{Ri} + cz_{Li}) = \sum(y_i + y_{Li} + y_{Ri}) \\ \sum(a + bx_i + cz_i)x_i + \sum(a + bx_{Li} + cz_{Ri})x_{Li} + \\ + \sum(a + bx_{Ri} + cz_{Li})x_{Ri} = \sum y_i x_i + \sum y_{Li}x_{Li} + \sum y_{Ri}x_{Ri} \\ \sum(a + bx_i + cz_i)z_i + \sum(a + bx_{Li} + cz_{Ri})z_{Ri} + \\ + \sum(a + bx_{Ri} + cz_{Li})z_{Li} = \sum y_i z_i + \sum y_{Li}z_{Ri} + \sum y_{Ri}z_{Li} \end{cases}$$

Recalling that the theoretical values of the fuzzy dependent variable are $y_i^* = a + bx_i + cz_i$, $y_{Li}^* = a + bx_{Li} + cz_{Ri}$ and $y_{Ri}^* = a + bx_{Ri} + cz_{Li}$, we obtain

$$\begin{cases} \sum(y_i^* + y_{Li}^* + y_{Ri}^*) = \sum(y_i + y_{Li} + y_{Ri}) \\ \sum y_i^* x_i + y_{Li}^* x_{Li} + y_{Ri}^* x_{Ri} = \sum y_i x_i + y_{Li}x_{Li} + y_{Ri}x_{Ri} \\ \sum y_i^* z_i + y_{Li}^* z_{Ri} + y_{Ri}^* z_{Li} = \sum y_i z_i + y_{Li}z_{Ri} + y_{Ri}z_{Li}^L \end{cases} \quad (5)$$

The first equation of the system (5) shows that the total sum of lower extremes, cores and upper extremes of the theoretical values of the dependent variable coincides with the same amount referred to the empirical values. This equation does not allow us to say that theoretical and empirical averages of the fuzzy dependent variable coincide.

Let's examine how the total sum of squares of dependent variable

$$\text{Tot SS} = \sum[(y_i - \bar{y})^2 + (y_{Li} - \bar{y}_L)^2 + (y_{Ri} - \bar{y}_R)^2]$$

can be decomposed according to Diamond's metric.

Adding and subtracting the corresponding theoretical value within each square and developing all the squares, the total deviance can be expressed as:

$$\text{Tot SS} = \sum[(y_i - y_i^* + y_i^* - \bar{y})^2 + (y_{Li} - y_{Li}^* + y_{Li}^* - \bar{y}_L)^2 +$$
$$+ (y_{Ri} - y_{Ri}^* + y_{Ri}^* - \bar{y}_R)^2] =$$
$$= \sum[(y_i - y_i^*)^2 + (y_i^* - \bar{y})^2 + 2(y_i - y_i^*)(y_i^* - \bar{y}) +$$
$$+ (y_{Li} - y_{Li}^*)^2 + (y_{Li}^* - \bar{y}_L)^2 + 2(y_{Li} - y_{Li}^*)(y_{Li}^* - \bar{y}_L) +$$
$$+ (y_{Ri} - y_{Ri}^*)^2 + (y_{Ri}^* - \bar{y}_R)^2 + 2(y_{Ri} - y_{Ri}^*)(y_{Ri}^* - \bar{y}_R)].$$

Adding and subtracting the theoretical average values of the lower extremes, of the cores and of the upper extremes of the dependent variable within each square and developing all the squares, the previous expression becomes

$$\begin{aligned}&\mathrm{TotSS}=\sum[(y_i-y_i^*)^2+(y_i^*-\bar y^*+\bar y^*-\bar y)^2+\\&+2(y_i-y_i^*)(y_i^*-\bar y)+(y_{Li}-y_{Li}^*)^2+(y_{Li}^*-\bar y_L^*+\bar y_L^*-\bar y_L)^2+\\&+2(y_{Li}-y_{Li}^*)(y_{Li}^*-\bar y_L)+(y_{Ri}-y_{Ri}^*)^2+\\&+(y_{Ri}^*-\bar y_R^*+\bar y_R^*-\bar y_R)^2+2(y_{Ri}-y_{Ri}^*)(y_{Ri}^*-\bar y_R)]=\\&=\sum[(y_i-y_i^*)^2+(y_i^*-\bar y^*)^2+(\bar y^*-\bar y)^2+2(y_i^*-\bar y^*)(\bar y^*-\bar y)+\\&+2(y_i-y_i^*)(y_i^*-\bar y)+(y_{Li}-y_{Li}^*)^2+(y_{Li}^*-\bar y_L^*)^2+(\bar y_L^*-\bar y_L)^2+\\&+2(y_{Li}^*-\bar y_L^*)(\bar y_L^*-\bar y_L)+2(y_{Li}-y_{Li}^*)(y_{Li}^*-\bar y_L)+(y_{Ri}-y_{Ri}^*)^2+\\&+(y_{Ri}^*-\bar y_R^*)^2+(\bar y_R^*-\bar y_R)^2+2(y_{Ri}^*-\bar y_R^*)(\bar y_R^*-\bar y^R)+\\&+2(y_{Ri}-y_{Ri}^*)(y_{Ri}^*-\bar y_R)]\end{aligned}$$

where:

$$\mathrm{RegSS}=\sum d(\tilde Y_i^*,\bar Y)^2=\sum[(y_i^*-\bar y)^2+(y_{Li}^*-\bar y_L)^2+(y_{Ri}^*-\bar y_R)^2]$$

represents the regression sum of squares, while

$$\mathrm{ResSS}=\sum d(\tilde Y_i,\tilde Y_i^*)^2=\sum[(y_i-y_i^*)^2+(y_{Li}-y_{Li}^*)^2+(y_{Ri}-y_{Ri}^*)^2]$$

represents the residual sum of squares, and

$$nd(\bar Y^*,\bar Y)^2=n[(\bar y^*-\bar y)^2+(\bar y_L^*-\bar y_L)^2+(\bar y_R^*-\bar y_R)^2]$$

represents the distance between theoretical and empirical average values of dependent variable.

Synthetically the expression of Tot SS can be written as:

$$\text{Tot SS} = \text{Reg SS} + \text{Res SS} + nd(\bar Y,\bar Y^*)^2+\eta$$

where:

$$\begin{aligned}&\eta=2\sum[(y_i^*-\bar y^*)(\bar y^*-\bar y)+(y_{Li}^*-\bar y_L^*)(\bar y_L^*-\bar y_L)+\\&(y_{Ri}^*-\bar y_R^*)(\bar y_R^*-\bar y_R)]+2\sum[(y_i-y_i^*)(y_i^*-\bar y)+\\&+(y_{Li}-y_{Li}^*)(y_{Li}^*-\bar y_L)+(y_{Ri}-y_{Ri}^*)(y_{Ri}^*-\bar y_R)].\end{aligned}$$

As the sums of deviations of each component from its average equal zero, then it is

$$\sum[(y_i^*-\bar y^*)(\bar y^*-\bar y)+(y_{Li}^*-\bar y_L^*)(\bar y_L^*-\bar y_L)+(y_{Ri}^*-\bar y_R^*)(\bar y_R^*-\bar y_R)]=0$$

and the amount η is reduced to

$$\begin{aligned}&\eta=2\sum[(y_i-y_i^*)(y_i^*-\bar y)+(y_{Li}-y_{Li}^*)(y_{Li}^*-\bar y_L)+\\&+(y_{Ri}-y_{Ri}^*)(y_{Ri}^*-\bar y_R)]=\\&=2\sum[(y_i-y_i^*)y_i^*-(y_i-y_i^*)\bar y+(y_{Li}-y_{Li}^*)y_{Li}^*-(y_{Li}-y_{Li}^*)\bar y_L+\\&+(y_{Ri}-y_{Ri}^*)y_{Ri}^*-(y_{Ri}-y_{Ri}^*)\bar y_R].\end{aligned}$$

Moreover, as it is $y_i^*=a+bx_i+cz_i$, $y_{Li}^*=a+bx_{Li}+cz_{Ri}$ and $y_{Ri}^*=a+bx_{Ri}+cz_{Li}$, it is also

$$2\sum(y_i-y_i^*)y_i^*+2\sum(y_{Li}-y_{Li}^*)y_{Li}^*+2\sum(y_{Ri}-y_{-Ri}^*)y_{Ri}^*=0.$$

By replacing expressions of the theoretical values in the latter equation, we obtain

$$\begin{aligned}&\eta=2\sum[a(y_i+y_{Li}+y_{Ri})-a(y_i^*+y_{Li}^*+y_{Ri}^*)+\\&+b(y_ix_i+y_{Li}x_{Li}+y_{Ri}x_{Ri})-b(y_i^*x_i+y_{Li}^*x_{Li}+y_{Ri}^*x_{Ri})+\\&+c(y_iz_i+y_{Li}z_{Ri}+y_{Ri}z_{Li})-c(y_i^*x_i+y_{Li}^*x_{Li}+y_{Ri}^*x_{Ri})]+\\&-2\sum[(y_i-y_i^*)\bar y+(y_{Li}-y_{Li}^*)\bar y_L+(y_{Ri}-y_{Ri}^*)\bar y_R].\end{aligned}$$

According to the condition (5) the last expression can be reduced to

$$\eta=-2\sum[(y_i-y_i^*)\bar y+(y_{Li}-y_{Li}^*)\bar y_L+(y_{Ri}-y_{Ri}^*)\bar y_R].$$

Note that, if the residual sum of squares equals zero, also η and $d(\bar Y,\bar Y^*)^2$ equal zero, because theoretical and empirical average values of the dependent variable coincide for each observation. Therefore:

- if the regression sum of squares equals zero, then the model has no forecasting capability, because the sum of the components of the i-th theoretical value equals the sum of the components of the empirical average value (i = 1 ,..., n). Actually it is for each i

$$\sum y_i^*+y_{Li}^*+y_{Ri}^*=\sum y_i+y_{Li}+y_{Ri}\ \Rightarrow$$

$$\Rightarrow\ ny_i^*+ny_{Li}^*+ny_{Ri}^*=n\bar y+n\bar y_L+n\bar y_R\ \Rightarrow$$

$$\Rightarrow\ y_i^*+y_{Li}^*+y_{Ri}^*=\bar y+\bar y_L+\bar y_R\,;$$

- if the residual sum of squares equals zero, the relationship between the dependent variable and the independent ones is well represented by the estimated model. In this case, the total sum of squares is entirely explained by the regression sum of squares.

4.2 The Model Including a Fuzzy Intercept

Let's consider, only for example, the sum of Diamond's distances between theoretical and empirical values of the dependent variable in the case 3 for a model with fuzzy intercept:

$$\begin{aligned}&\sum d(\tilde Y_i,\tilde A+b\tilde X_i+c\tilde Z_i)^2=\\&=\sum[(y_i-a-bx_i-cz_i)^2+(y_{Li}-a_L-bx_{Li}-cz_{Ri})^2+\\&+(y_{Ri}-a_R-bx_{Ri}-cz_{Li})^2]\end{aligned}$$

By minimizing such a quantity with respect to a, b, c, $\underline{\gamma}$ and $\overline{\gamma}$ (remember that $a_L=a-\underline{\gamma}$ and $a_R=a+\overline{\gamma}$) we can obtain the following system of equations

$$\begin{cases} -2\sum[(y_i - a - bx_i - cz_i) + (y_{Li} - a + \underline{\gamma} - bx_{Li} - cz_{Ri}) + \\ \quad + (y_{Ri} - a - \bar{\gamma} - bx_{Ri} - cz_{Li})] = 0 \\ -2\sum[(y_i - a - bx_i - cz_i)x_i + (y_{Li} - a + \underline{\gamma} - bx_{Li} - cz_{Ri})x_{Li} + \\ \quad + (y_{Ri} - a - \bar{\gamma} - bx_{Ri} - cz_{Li})x_{Ri}] = 0 \\ -2\sum[(y_i - a - bx_i - cz_i)z_i + (y_{Li} - a + \underline{\gamma} - bx_{Li} - cz_{Ri})z_{Ri} + \\ \quad + (y_{Ri} - a - \bar{\gamma} - bx_{Ri} - cz_{Li})z_{Li}] = 0 \\ 2\sum(y_{Li} - a + \underline{\gamma} - bx_{Li} - cz_{Ri}) = 0 \\ -2(y_{Ri} - a - \bar{\gamma} - bx_{Ri} - cz_{Li}) = 0 \end{cases}$$

Such a system can be written as

$$\begin{cases} \sum(a + bx_i + cz_i) + \sum(a - \underline{\gamma} + bx_{Li} + cz_{Ri}) + \\ + \sum(a + \bar{\gamma} + bx_{Ri} + cz_{Li}) = \sum(y_i + y_{Li} + y_{Ri}) \\ \sum(a + bx_i + cz_i)x_i + \sum(a - \underline{\gamma} + bx_{Li} + cz_{Ri})x_{Li} + \\ + \sum(a + \bar{\gamma} + bx_{Ri} + cz_{Li})x_{Ri} = \sum y_i x_i + \sum y_{Li} x_{Li} + \sum y_{Ri} x_{Ri} \\ \sum(a + bx_i + cz_i)z_i + \sum(a - \underline{\gamma} + bx_{Li} + cz_{Ri})z_{Ri} + \\ + \sum(a + \bar{\gamma} + bx_{Ri} + cz_{Li})z_{Li} = \sum y_i z_i + \sum y_{Li} z_{Ri} + \sum y_{Ri} z_{Li} \\ \sum(a - \underline{\gamma} + bx_{Li} + cz_{Ri}) = \sum y_{Li} \\ \sum(a + \bar{\gamma} + bx_{Ri} + cz_{Li}) = \sum y_{Ri} \end{cases}$$

Recalling that the theoretical values of the fuzzy dependent variable are $y^*_i = a + bx_i + cz_i$, $y^*_{Li} = a - \underline{\gamma} + bx_{Li} + cz_{Ri}$ and $y^*_{Ri} = a + \bar{\gamma} + bx_{Ri} + cz_{Li}$ respectively, we obtain

$$\begin{cases} \sum(y^*_i + y^*_{Li} + y^*_{Ri}) = \sum(y_i + y_{Li} + y_{Ri}) \\ \sum y^*_i x_i + \sum y^*_{Li} x_{Li} + \sum y^*_{Ri} x_{Ri} = \\ = \sum y_i x_i + \sum y_{Li} x_{Li} + \sum y_{Ri} x_{Ri} \\ \sum y^*_i z_i + \sum y^*_{Li} z_{Ri} + \sum y^*_{Ri} z_{Li} = \\ = \sum y_i z_i + \sum y_{Li} z_{Ri} + \sum y_{Ri} z_{Li} \\ \sum y^*_{Li} = \sum y_{Li} \\ \sum y^*_{Ri} = \sum y_{Ri} \end{cases} \tag{6}$$

The first equation shows that the total sum of cores and extremes of the theoretical values of the dependent variable coincides with the same amount referred to the empirical values. The combination of the first equation with the last two allows us to state that theoretical and empirical values of the average fuzzy dependent variable coincide, like it happens in the classic OLS estimation procedure.

Let's examine how the total sum of squares of dependent variable can be decomposed according to Diamond's metric:

$$\mathrm{TotSS} = \sum[(y_i - \bar{y}\ )^2 + (y_{Li} - \bar{y}_L)^2 + (y_{Ri} - \bar{y}_R)^2].$$

Adding and subtracting the corresponding theoretical value within each square and developing all the squares, the total deviance can be expressed as:

$$\begin{aligned} \mathrm{Tot\,SS} &= \sum[(y_i - y^*_i + y^*_i - \bar{y}\ )^2 + (y_{Li} - y^*_{Li} + y^*_{Li} - \bar{y}_L)^2 + \\ &+ (y_{Ri} - y^*_{Ri} + y^*_{Ri} - \bar{y}_R)^2] = \\ &= \sum[(y_i - y^*_i)^2 + (y^*_i - \bar{y})^2 + 2(y_i - y^*_i)(y^*_i - \bar{y}\ ) + \\ &+ (y_{Li} - y^*_{Li})^2 + (y^*_{Li} - \bar{y}_L)^2 + 2(y_{Li} - y^*_{Li})(y^*_{Li} - \bar{y}_L) + \\ &+ (y_{Ri} - y^*_{Ri})^2 + (y^*_{Ri} - \bar{y}_R)^2 + 2(y_{Ri} - y^*_{Ri})(y^*_{Ri} - \bar{y}_R)]. \end{aligned}$$

Adding and subtracting the theoretical average values of the lower extremes, of the cores and of the upper extremes of the dependent variable within each square and developing all the squares, the previous expression becomes

$$\begin{aligned} \mathrm{TotSS} &= \sum[(y_i - y^*_i)^2 + (y^*_i - \bar{y}^* + \bar{y}^* - \bar{y})^2 + \\ &+ 2(y_i - y^*_i)(y^*_i - \bar{y}) + (y_{Li} - y^*_{Li})^2 + (y^*_{Li} - \bar{y}^*_L + \bar{y}^*_L - \bar{y}_L)^2 + \\ &+ 2(y_{Li} - y^*_{Li})(y^*_{Li} - \bar{y}_L) + (y_{Ri} - y^*_{Ri})^2 + \\ &+ (y^*_{Ri} - \bar{y}^*_R + \bar{y}^*_R - \bar{y}_R)^2 + 2(y_{Ri} - y^*_{Ri})(y^*_{Ri} - \bar{y}_R)] = \\ &= \sum[(y_i - y^*_i)^2 + (y^*_i - \bar{y}^*)^2 + (\bar{y}^* - \bar{y})^2 + 2(y^*_i - \bar{y}^*)(\bar{y}^* - \bar{y}) + \\ &+ 2(y_i - y^*_i)(y^*_i - \bar{y}) + (y_{Li} - y^*_{Li})^2 + (y^*_{Li} - \bar{y}^*_L)^2 + (\bar{y}^*_L - \bar{y}_L)^2 + \\ &+ 2(y^*_{Li} - \bar{y}^*_L)(\bar{y}^*_L - \bar{y}_L) + 2(y_{Li} - y^*_{Li})(y^*_{Li} - \bar{y}_L) + (y_{Ri} - y^*_{Ri})^2 + \\ &+ (y^*_{Ri} - \bar{y}^*_R)^2 + (\bar{y}^*_R - \bar{y}_R)^2 + 2(y^*_{Ri} - \bar{y}^*_R)(\bar{y}^*_R - \bar{y}^R) + \\ &+ 2(y_{Ri} - y^*_{Ri})(y^*_{Ri} - \bar{y}_R)] \end{aligned}$$

where:

$$\mathrm{RegSS} = \sum d(\tilde{Y}^*_i, \bar{Y})^2 = \sum[(y^*_i - \bar{y})^2 + (y^*_{Li} - \bar{y}_L)^2 + (y^*_{Ri} - \bar{y}_R)^2]$$

represents the regression sum of squares, while

$$\mathrm{Res\,SS} = \sum d(\tilde{Y}_i, \tilde{Y}^*_i)^2 = \sum[(y_i - y^*_i)^2 + (y_{Li} - y^*_{Li})^2 + (y_{Ri} - y^*_{Ri})^2]$$

represents the residual sum of squares. Moreover, according to the conditions (6), it is

$$\begin{aligned} &\sum[(\bar{y}^* - \bar{y})^2 + 2(y_i - y^*_i)(y^*_i - \bar{y}) + 2(y^*_i - \bar{y}^*)(\bar{y}^* - \bar{y}) + \\ &+ (\bar{y}^*_L - \bar{y}_L)^2 + 2(y_{Li} - y^*_{Li})(y^*_{Li} - \bar{y}_L) + +2(y^*_{Li} - \bar{y}^*_L)(\bar{y}^*_L - \bar{y}_L) + \\ &+ (\bar{y}^*_R - \bar{y}_R)^2 + 2(y_{Ri} - y^*_{Ri})(y^*_{Ri} - \bar{y}_R) + 2(y^*_{Ri} - \bar{y}^*_R)(\bar{y}^*_R - \bar{y}_R)] = 0 \end{aligned}$$

Therefore the expression of the total sum of squares of the dependent variable can be reduced to

$$\mathrm{Tot\,SS} = \mathrm{Reg\,SS} + \mathrm{Res\,SS}.$$

Ultimately the total sum of squares consists only of two addends, the regression sum of square and the residual one, like in the classic OLS estimation procedure, when the intercept has the same form of the dependent variable.

Note that, when the intercept has not the same form of the dependent variable, theoretical and empirical average values of the latter do not coincide for each observation; rather the total sum of lower extremes, cores and upper extremes of the theoretical values coincides with the same amount referred to the empirical values:

$$\begin{cases}\Sigma(y_i^* + y_{Li}^* + y_{Ri}^*) = \Sigma(y_i + y_{Li} + y_{Ri}) \\ \Sigma y_i^* x_i + \Sigma y_{Li}^* x_{Li} + \Sigma y_{Ri} x_{Ri} = \Sigma y_i x_i + \Sigma y_{Li} x_{Li} + \Sigma y_{Ri} x_{Ri} \\ \Sigma y_i^* z_i + \Sigma y_{Li}^* z_{Ri} + \Sigma y_{Ri}^* z_{Li} = \Sigma y_i z_i + \Sigma y_{Li} z_{Ri} + \Sigma y_{Ri} z_{Li} \\ \Sigma(y_{Ri}^* - y_{Ri}) = \Sigma(y_{Li}^* - y_{Li})\end{cases}$$

In this case the total sum of squares of the dependent variable consists of two other components in addition to the regression sum of square and the residual one: the first is residual in nature and is characterized by an uncertain sign, the second is equal to n times the distance between theoretical and empirical average values of the dependent variable.

5 A FUZZY MODEL FIT INDEX

We have just demonstrated that the total sum of squares of the dependent variable consists only of two addends, the regression sum of square and the residual one, when the intercept is fuzzy asymmetric. This is because theoretical and empirical average values of the dependent variable coincide and, therefore, both the total sum of squares and the regression one can be expressed in terms of distance between empirical values and their averages.

Under these circumstances, the greater the regression sum of squares the better the model fits the data.

When there are more addends of the total sum of squares than those just mentioned, an increase in the regression sum of square does not necessarily imply a better fit to observed data: this is because the theoretical average value, from which the regression sum of squares is calculated, may be very different from the empirical one. On the contrary a decrease in the residual sum of squares necessarily implies a better fit to observed data.

In order to assess the goodness of fit of the regression model, we propose the following index, for simplicity called Fuzzy Fit Index (FFI), which is common to all three models:

$$\mathrm{FFI} = 1 - \frac{\mathrm{Res\,SS}}{\mathrm{Tot\,SS}} = 1 - \frac{\sum d(\tilde{Y}_i, \tilde{Y}_i^*)^2}{\sum d(\tilde{Y}_i, \overline{Y})^2}$$

where $\overline{Y}^* = (\bar{y}^*, \bar{y}_L^*, \bar{y}_R^*)_T$ and $\overline{Y} = (\bar{y}, \bar{y}_L, \bar{y}_R)_T$ denote the fuzzy theoretical average and the fuzzy empirical average of the dependent variable respectively.

The more this index is next to 1, the smaller the residual sum of squares is and the better the model fits the observed data.

With specific reference to the model with a symmetric (both fuzzy and not) intercept, if the residual sum of squares decreases, also the distance between theoretical and empirical fuzzy averages of the dependent variable decreases, as well as the component η of the total sum of squares. It follows ultimately that the forecasting capability of the model increases.

6 A STEPWISE FORWARD PROCEDURE TO SELECT INDEPENDENT VARIABLES

The selection of the most significant independent variables presents greater difficulties from a computational point of view in the case of a fuzzy regression model than in the classic one.

In classical regression analysis, if the number p of independent variables is limited, the optimal subset of them can be selected by examining in succession at most $\sum \frac{p!}{k!(p-k)!}$ models, from the simple ones (k = 1) to the saturated one (k = p).

The fuzzy approach makes the search for optimal combinations of explanatory variables more complex from a computational point of view.

The total number of the potential hyperplanes to be tested increases exponentially with the number p of the starting variables considered: in fact, for each subset of q≤p variables, 2^q different hyperplanes result from all combinations of the signs assumed by the corresponding regression coefficients.

In order to avoid complications related to the above checks, we introduce a stepwise procedure which enables us to find the optimal combination of the starting variables by including only one of them at a time. At each iteration the procedure selects the variable which helps to explain the total sum of squares of the dependent variable more than the other variables not yet included in the model and which is also less correlated with the ones already included. This allows us to estimate $2\sum_{k=0}^{p-1}(p-k)2^k$ model at most.

More specifically, in the first step $\tilde{X}_{(1)}$ is included in the equation if it presents the highest correlation with the dependent variable $\tilde{Y}$; in the q.th step $\tilde{X}_{(q)}$ is selected to enter the model if its explanatory contribution to the sum of squares of $\tilde{Y}$ is higher than the other variables not yet included and also than an arbitrary threshold value. Such a contribution can be measured as the increase in the FFI due to the introduction of $\tilde{X}_{(q)}$ into the equation, equal to $FFI_{y;1,2,...,q}$ - $FFI_{y;1,2,...,q-1}$ (where the two terms of the subtraction represent the proportion of the sum of squares of $\tilde{Y}$ explained by the model including $\tilde{X}_{(q)}$ and not). The higher the threshold value, the easier the procedure inhibits the entry of new independent variables, because of the increases in the fraction of the total variability which should be explained.

Once $\tilde{X}_{(q)}$ is selected, its originality is evaluated through the so called tolerance $T_q=1-FFI_{q;1,2,...,q-1}$, where $FFI_{q;1,2,...,q-1}$ represents the share of variability of $\tilde{X}_{(q)}$ explained by the q-1 independent variables already in the model. The tolerance ranges between 0 and 1, depending on the degree of linear correlation of $\tilde{X}_{(q)}$ with the other variables; therefore, only if T_q exceeds a threshold between 0 and 1, $\tilde{X}_{(q)}$ will become part of the model. A high value of the threshold allows to select very original variables, but it can also stop the process right from the initial steps; on the contrary, a low value allows most of the variables enter into the equation only if they explain a significant fraction of variability of $\tilde{Y}$. The described procedure stops when none of the variables not yet included in the equation may introduce a significant contribution to the model, or if none of the candidate variables to enter is significantly original.

For an application of this procedure see Montrone, Campobasso, Perchinunno and Fanizzi, 2011, which elaborates on data revealed by the EU-SILC survey of 2006 regarding the perception of poverty by Italian families. For this purpose, by using the editor of *Matlab*, we generated a function which requires, as input parameters, the matrices of cores, left extremes and right extremes both of the dependent and of the independent fuzzy variables.

A more accurate procedure provides the possibility of eliminating at each iteration variables already included in the model, whose explanatory contribution is subrogated by the combination of the independent variables introduced later.

In particular, unlike the procedure just described, we can verify at each iteration that the explanatory contribution of the variable $\tilde{X}_{(i)}$ (i = 1, 2, ..q-1) is still significant, once the candidate variable $\tilde{X}_{(q)}$ is inserted. In the q.th step such a contribution can be measured by the reduction of FFI in the elimination of the variable $\tilde{X}_{(i)}$ from the model, equal to $FFI_{y;1,2,...,q}$ - $FFI_{y;1,2,...,q\ (-i)}$ (where the two terms of the subtraction represent the proportion of the sum of squares of $\tilde{Y}$ explained by the model including all the variable and without the variable $\tilde{X}_{(i),}$ respectively). So, the variable $\tilde{X}_{(i)}$ remains in the model if the percentage of the sum of squares explained by the model including all variables is higher than the model without the variable $\tilde{X}_{(i)}$ and also arbitrary threshold value.

7 CONCLUSIONS

In this work we first explicit the expressions of the estimated parameters of a multivariate fuzzy regression model with a fuzzy asymmetric intercept. Such an intercept is more appropriate than a non-fuzzy on, as it is to be estimated by the average value of the dependent variable (which is also fuzzy) when the independent variables equal zero.

Moreover we verify that the sum of squares of the dependent variable consists simply in the regression sum of squares and the residual one, like it happens in the classic OLS estimation procedure, only when the intercept is fuzzy asymmetric triangular. Conversely, when the intercept is symmetric (both fuzzy and not), the analysis of the forecasting capability of the model is more difficult. This happens because of the presence of two additional components of the sum of squares: the first one which is related to the difference between the theoretical and the empirical average values of the dependent variable, the second one which is residual in nature and is characterized by an uncertain sign.

The selection of the most significant independent variables in a fuzzy regression model presents computational difficulties due to the large number of potential hyperplanes to be tested. We propose to overcome such difficulties through a stepwise procedure, based on a fuzzy version of the R^2 index.

In each step a single variable is included between the starting ones, according to two basic criteria: its explanatory contribution to the model and its originality with respect to the other variables already included in the model.

A more accurate procedure provides the possibility of eliminating at each iteration variables already included in the model, whose explanatory contribution is subrogated by the combination of the independent variables introduced later.

The forecasting capability of the proposed fuzzy regression model has been successfully verified in a recent application to data revealed by the EU-SILC survey of 2006, regarding the perception of poverty by Italian families. In that circumstance we have used the editor of *Matlab* and, in particular, we have generated a function which requires, as input parameters, the matrices of cores, left extremes and right extremes both of the dependent and of the independent fuzzy variables.

Some improvements to the model mainly concern the shape of the membership function different from the triangular one.

REFERENCES

Bilancia, M., Campobasso, F., Fanizzi, A., 2010. The pricing of risky securities in a Fuzzy Least Square Regression model. In *Advances in Data Analysis and Classification 2010.* Springer Berlin-Heidelberg-New York,.

Campobasso, F., Fanizzi, A., Tarantini, M., 2009. Some results on a multivariate generalization of the Fuzzy Least Square Regression. In *Proceedings of the International Conference on Fuzzy Computation,* Madeira.

Campobasso, F., Fanizzi, A., 2011. A Fuzzy Approach To The Least Squares Regression Model With A Symmetric Fuzzy Intercept. In *Proceedings of the 14th Applied Stochastic Model and Data Analysis Coinference*, Roma.

Campobasso, F., Perchinunno, P., Fanizzi, A., 2008. Homogenous Urban Poverty Clusters within the city of Bari. In *Lecture Notes in Computer Science ICCSA 2008*. Springer.

Diamond, P. M., 1988. Fuzzy Least Square. In *Information Sciences*.

Kao, C., Chyu, C. L., 2003. Least-squares estimates in fuzzy regression analysis. In *European Journal of Operational Research*.

Montrone, S., Campobasso, F., Perchinunno, P., Fanizzi, A., 2011. A Fuzzy Approach to the Small Area Estimation of Poverty in Italy. In *Advances in Intelligent Decision Technologies – Proceedings of the Second KES International Symposium IDT 2010,* Springer.

Montrone, S., Campobasso, F., Perchinunno, P., Fanizzi, A., 2011. An Analysis of Poverty in Italy through a fuzzy regression model. In *Lecture Notes in Computer Science ICCSA 2011*, Springer.

Montrone, S., Perchinunno, P., Di giuro, A., Torre, C. M., Rotondo, F., 2011. Identification of hot spot of social and housing difficulty in urban areas. In *Lecture Notes in Computer Science ICCSA 2011*, Springer.

Takemura, K., 2005. Fuzzy least squares regression analysis for social judgment study. In *Journal of Advanced Intelligent Computing and Intelligent Informatics*.