O P E R A T I O N S R E S E A R C H A N D D E C I S I O N S

No. 2 2011

Barbara GŁADYSZ*
Dorota KUCHTA*

A METHOD OF VARIABLE SELECTION FOR FUZZY REGRESSION – THE POSSIBILITY APPROACH

A method of variable selection for fuzzy regression has been proposed. Using the method, the significance of fuzzy regression coefficients has been examined. The method presented is equivalent to the method of variable selection for classical regression based on an analysis of the confidence intervals for their coefficients. Illustrative examples are presented.

Key words: *fuzzy regression, possibility distribution of a fuzzy variable, variable significance*

1. Introducion

Possibility theory was proposed by Zadeh (1965). We will present here the basic elements of this theory. First, we will present the notion of a fuzzy variable. Let X be a variable whose value is unknown. Let $\mu_X : \Re \to [0,1]$ be a normal, quasi-concave, upper semi-continuous function, called the possibility distribution of the variable X [Dubois, Prade 1988], [Zadeh 1978]. The value $\mu_X(x)$ for $x \in \Re$ denotes the possibility of the event that the variable X takes the value x. This is formulated in the following way:

$$\mu_X(x) = \mathrm{Pos}(X = x) \tag{1}$$

If $\mu_X(x)$ fulfills the following conditions:

- normality: there exists an x such that $\mu_X(x) = 1$,
- convexity: the λ-level sets $[X]_\lambda = \{x : \mu(x) \geq \lambda\}$ of X are convex ($0 < \lambda \leq 1$),

*Institute of Organisation and Management, Wrocław University of Technology, ul. Smoluchowskiego 25, 50-372 Wrocław, Department of Management, e-mail addresses: Barbara.Gladysz@pwr.wroc.pl, Dorota.Kuchta@pwr.wroc.pl

• continuity: $\mu_X(x)$ is piecewise continuous, then the variable X is called a fuzzy variable.

For a given fuzzy variable and a given λ, the λ-cut of the fuzzy variable is defined as $[X]_\lambda = \{x: \mu_X(x) \geq \lambda\}$.

An *L-L* fuzzy variable is a special case of a fuzzy variable. It has the possibility distribution

$$\mu_X(x) = L\left(\left|\frac{x - m_X}{l_X}\right|\right) \tag{2}$$

where: $L(x) = L(-x)$, m_X – the centre of the fuzzy variable X, l_X – the width of the fuzzy variable X.

The following are examples of the $L(x)$ function: $L(x) = \max\{0,1 - |x|^p\}$, $L(x) = (1 + |x|^p)^{-1}$, $L(x) = \exp(-|x|^p)$. An *L-L* fuzzy variable will be denoted by $X = (m_X, l_X)$.

Let X, Y be two fuzzy variables with possibility distributions $\mu_X(x)$, $\mu_Y(y)$, respectively. Then, according to the Zadeh extension principle [Zadeh 1965], the possibility distributions of the sum $Z = X + Y$ and of the product $V = XY$ can be written in the following way:

$$\mu_Z(z) = \sup_{z=x+y}\left(\min\left(\mu_X(x), \mu_Y(y)\right)\right) \tag{3}$$

$$\mu_V(v) = \sup_{v=xy}\left(\min\left(\mu_X(x), \mu_Y(y)\right)\right) \tag{4}$$

Let us suppose that we want to compare two fuzzy variables, i.e. that we want to determine the possibility of the occurrence of the event that the realization of the variable X (i.e. the actual value taken by X) will be higher (not smaller) than the realization of the variable Y. To determine this possibility Dubois and Prade proposed the following measures [Dubois, Prade 1988]:

$$\mathrm{Pos}(X \geq Y) = \sup_{x \geq y}\left(\min\left(\mu_X(x), \mu_Y(y)\right)\right) \tag{5}$$

$$\mathrm{Pos}(X > Y) = \sup_x \inf_{y \geq z}\left(\min\left(\mu_X(x), 1 - \mu_Y(y)\right)\right) \tag{6}$$

Let us suppose now that we want to determine the possibility of the event that the realization of variable X will be equal to that of variable Y. This is defined in the following way [Dubois, Prade 1988]:

$$\mathrm{Pos}(X = Y) = \min\left(\mathrm{Pos}(X \geq Y), \mathrm{Pos}(Y \geq X)\right) \tag{7}$$

Dubois and Prade also proposed the necessity measure of the respective event which is stronger than the possibility measure and is defined to be the complement of the possibility measure of the opposite event [Dubois, Prade 1988]. Thus we have:

$$\mathrm{Nec}(X \neq Y) = 1 - \mathrm{Pos}(X = Y) \tag{8}$$

Both the possibility and the necessity measures take on values from the interval [0, 1].

2. The choice of independent variables for fuzzy regression

Fuzzy regression is based on a linear dependence:

$$\hat{Y} = A_0 + A_1 X_1 + \cdots + A_k X_k \tag{9}$$

in which the dependent variable Y, the independent variables X_1, …, X_k and the regression coefficients A_0, A_1, …, A_k are fuzzy variables. In special cases, the regression coefficients or independent variables may be crisp numbers. The first work in the domain of fuzzy regression is the paper of Tanaka et al. [Tanaka et al., 1982]. The literature devoted to this problem is today very numerous. Individual fuzzy regression models differ from each other in the assumptions regarding the data and optimisation criteria. We can distinguish two basic streams here. In the first one, developing the proposal of Tanaka et al., the fuzzy regression coefficients are estimated using methods from linear programming [Sakawa, Yano 1992]. In the latter one, these coefficients are estimated by means of the method of least squares (e.g. [Celmins 1987], [Diamond 19988], [D'Urso, Gastaldi 2002], [Körner, Näther 1998]). A review of fuzzy regression methods can be found in [Kacprzyk, Fedrizzi 1992] and [Gładysz 2011].

In fuzzy regression, as in the case of classical regression, it is important to select an appropriate regression model: the form of the regression model and the set of independent variables. The following papers, among others, consider this problem: [D'Urso, Santaro 2006], [Wang, Tsuar 2000]. Wang and Tsuar propose a variable selection method in which the partial correlation coefficient is analyzed. D'Urso and Santaro propose variable selection procedures based on the multiple determination coefficient, adjusted multiple determination coefficient and the so called Mallows index.

In this paper, a method of selecting independent variables for fuzzy regression based on the significance of a coefficient will be presented. The method is equivalent to the method of selecting independent variables for classical regression based on the significance of a Student t-test [Maddala 2001].

A significance measure for fuzzy regression coefficients was proposed in [Gładysz, Kuchta 2009]. The coefficient A_j in fuzzy regression (9) is significantly different from zero, if the independent variable X_j linked to it has a significant influence on the value of the dependent variable Y, i.e.

$$\mathrm{Pos}\left(A_j = 0\right) \le \lambda_0 \tag{10}$$

where λ_0 is a parameter given by the decision maker.

The proposed method for selecting independent variables for fuzzy regression is as follows:

Algorithm 1

STEP 1. Let S be the set of independent variables. Construct a fuzzy regression model in which the variables from S are the independent variables.

STEP 2. Check the significance of the fuzzy regression coefficients. If there are insignificant variables in the fuzzy regression model, i.e. such that $\mathrm{Pos}(A_j = 0) > \lambda_0$, remove them from the set of independent variables and go back to STEP 1. Otherwise STOP.

The above algorithm will be applied to the construction of two fuzzy regression models: a fuzzy regression of the energy load and a fuzzy regression of the costs of sales representatives. In both fuzzy regressions, the observations of the dependent and independent variables are crisp and the regression coefficients will be determined by means of the method proposed by Savic and Pedrycz (1991). According to this method, it is assumed that the regression coefficients A_j, $j = 0, \ldots, k$, are symmetric triangular fuzzy numbers (i.e. such L-L fuzzy numbers that $L(x) = \max\{0, 1 - |x|\}$). The form of the regression model is determined in two phases:

In the former phase, the centrers of the coefficients A_j are estimated by means of the method of least squares.

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=0}^{k} a_j x_{ji}\right)^2 \longrightarrow \min \tag{11}$$

In the latter phase, using the centres $a_j^{\bullet}$ found by means of the method of least squares, the widths of the coefficients are determined according to the linear model:

$$\sum_{j=0}^{k} l_j \longrightarrow \min \tag{12}$$

together with the constraints:

$$(1-\lambda)\sum_{j=0}^{k} l_{A_j}\left|x_{ji}\right| + \sum_{j=0}^{k} a_j^{\bullet} x_{ji} \geq y_i \qquad \text{for} \qquad i = 1, ..., n \tag{13}$$

$$(\lambda-1)\sum_{j=0}^{k} l_{A_j}\left|x_{ji}\right| + \sum_{j=0}^{k} a_j^{\bullet} x_{ji} \geq y_i \qquad \text{for} \qquad i = 1, ..., n \tag{14}$$

$$l_{A_j} \geq 0 \qquad \text{for} \qquad j = 0, ..., k \tag{15}$$

The Savic and Pedrycz method is a combination of the ordinary method of least squares and an optimization method based on the possibility measure.

2.1. Example 1

Let us construct a fuzzy regression model for the energy load at noon in September. Taking into account the nature of the time series of energy load, the following potential independent variables were assumed: the energy load in the past, calendar data describing the nature of energy demand throughout the week and a temperature variable. The data are September observations over a period of four years. The model has been constructed based on the data from the first three years. The data from the fourth year served to evaluate the correctness of the forecast.

First we determine a fuzzy regression model based on the set of all the potential independent variables. Taking $\lambda = 0$, we obtain the following regression model:

$$\begin{aligned}\widehat{L12_t} = {} & (104\,791.2, 0) + (-18\,169.1, 0)\, Sunday \\ & + (-4144.4, 0)\, Day + (0.1726, 0)\, L12_{t-1} + (0.7936, 0)\, L7_t \\ & + (-0.2204, 0.005)\, L7_{t-1} + (-1237.15, 2726.6)\, Temp\end{aligned} \tag{16}$$

where: Lh_t – the system load at hour h on day t, Sunday (1 – for Sundays, 0 – for other days), *Day* – day type (1 – for Saturdays, 2 – for Sundays, 3 – for Mondays, 4 – for other days), *Temp* – temperature difference (maximum temperature – minimum temperature on the current day).

The value of the objective function (12) for this fuzzy regression model equals 2 571 505.69. Let us assume that $\lambda_0 = 0.1$. The significance measures for the individual fuzzy regression coefficients (16) are presented in Table 1.

The variable *Temp* is insignificant, because Pos((–1237.15, 2726.6) = 0) = 0.54 > λ_0. We remove the variable *Temp* from the set of independent variables. Then we construct a fuzzy regression model based on the other variables:

$$\begin{aligned}\widehat{L12_t} &= (83\,079.9, 11\,786.4) + (-16\,531.9, 0)\, Sunday \\ &+ (-4028.8, 0)\, Day + (0.2572, 0)\, L12_{t-1} \\ &+ (0.8517, 0)\, L7_t + (-0.33286, 0.107)\, L7_{t-1}\end{aligned} \tag{17}$$

In the fuzzy regression model (17) all the variables are significant, as we have $\mathrm{Pos}(A_j = 0) = 0$ for each coefficient. The value of the objective function (12) for this fuzzy regression model equals 2888911.46.

Table 1. Significance of the fuzzy regression coefficients (Eq. (16))

Variable	Pos ($A_j = 0$)
Const	0
Sunday	0
Day	0
$L12_{t-1}$	0
$L7$	0
$L7_{t-1}$	0
Temp	0.54

Now we are going to build a forecast for the following September on the basis of (17). This is presented in Table 2 and Figure 1.

Table 2. Meteorological data and the forecasts

Day of the month	Day of the week	Forecast $L12$ Centre	Forecast $L12$ width	$\mathrm{Pos}(\widehat{L12} = L12)$
1	2	3	4	5
1	Saturday	216876.6	30832.4	0.97
2	Sunday	189718.6	29384.6	0.93
3	Monday	220433.9	27776.1	0.95
4	Tuesday	217694.2	30600.6	0.94
5	Wednesday	219955.7	30964.7	0.85
6	Thursday	220659.2	31372.3	0.68
7	Friday	222485.4	31532.5	0.92
8	Saturday	221214.3	31710.4	0.95
9	Sunday	188521.3	29879.0	0.99
10	Monday	221372.3	28084.9	0.92
11	Tuesday	220647.6	30851.9	0.54
12	Wednesday	225424.1	31348.4	0.89
13	Thursday	224794.3	32233.2	0.88
14	Friday	226328.0	31809.8	0.96
15	Saturday	222819.5	32206.4	0.86
16	Sunday	183774.5	30598.7	1.00

Table 2 continued

1	2	3	4	5
17	Monday	221262.3	27597.3	0.94
18	Tuesday	229310.5	30991.6	0.88
19	Wednesday	227406.5	32482.1	0.97
20	Thursday	222944.3	32505.8	0.92
21	Friday	225824.5	31959.1	0.97
22	Saturday	225426.3	32343.1	0.65
23	Sunday	194121.7	30950.6	0.45
24	Monday	223977.7	28959.8	0.53
25	Tuesday	233017.5	31738.8	0.68
26	Wednesday	237561.9	32638.3	0.63
27	Thursday	238269.0	33298.4	0.71
28	Friday	237717.1	33644.5	0.69
29	Saturday	238123.7	33626.9	0.87
30	Sunday	204499.4	32264.1	0.97

Source: authors' work.

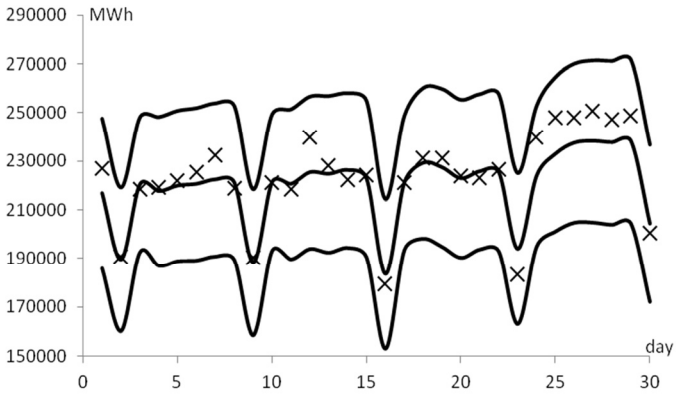

Fig. 1. The fuzzy forecast and the actual values of the energy load in September (authors' work)

The forecast can be considered to fit the actual data well – the smallest value of $\mathrm{Pos}(\widehat{L12} = L12)$ equals 0.45 for the 23rd of September. For the other days we have $\mathrm{Pos}(\widehat{L12} = L12) \geq 0.53$.

2.2. Example 2

We will now implement the proposed method of variable selection to the fuzzy regression of the costs of sales representatives in a certain company. This cost is com-

posed of travel expenses, telephone expenses, the operating expenses of company cars, salaries and related costs, entertainment expenses and other costs. Salaries and related costs constitute about 65% of the total cost, and their value depends on the sales figure. The second largest cost component is the operating expenses of company cars – they constitute about 30% of the total cost. The other cost items constitute about 5% of the total cost. The sales representatives act in five regions. They may also use company cars for private travel. The given cost structure and the fact that the sales representatives work in various regions made us take the following independent variables to construct a fuzzy regression model: the sales figure, the number of customers served by the given sales representative, the number of kilometers travelled and the region in which each representative works. The sales representatives of the company in question serve five regions (variables OI, OII, $OIII$, OIV, for the fifth region $OI = OII = OIII = OIV = 0$). Due to the fact that the bonuses for sales results are paid to the sales representatives with a delay of two months, the values of the independent variables come from period $t - 2$. The data are presented in Table 3.

Table 3. The data for the construction of the fuzzy regression model for cost

Sales representative	Cost	No. of customers (L)	Sales figure (S)	Duty kilometers (km S)	Private kilometers (km P)	OI	OII	$OIII$	OIV	OV
1	3591.16	57	116728.8	5411	459	1	0	0	0	0
2	3808.41	60	106020.2	4395	81	1	0	0	0	0
3	3530.35	49	23240.8	4246	486	0	1	0	0	0
4	3591.16	58	120113.7	3479	756	0	1	0	0	0
5	3808.41	57	205990.1	4890	540	0	0	1	0	0
6	3530.35	58	210290.1	4195	1107	0	0	1	0	0
7	3591.16	53	163997.4	2438	216	0	0	0	1	0
8	3808.41	62	104513.7	5259	621	0	0	0	1	0
9	3530.35	47	200692.7	4530	1107	0	0	0	0	1
10	3591.16	53	156497.4	3290	675	0	0	0	0	1
11	3808.41	47	171007.7	4265	567	0	0	0	0	1
12	3530.35	42	114741.6	5085	324	0	0	0	0	1

Source: M. Troska, *Econometric modelling of cost drivers in Activity Based Costing*, PhD thesis, Wroclaw University of Technology, Wroclaw, 2009 (in Polish).

First we will determine a fuzzy regression model based on the set of all the potential independent variables. Taking $\lambda = 0.5$, we obtain the following regression model:

$$\begin{aligned}\widehat{Cost_t} = {} & (2172.27, 0) + (22.23, 0)\, L_{t-2} + (0.0017, 0.0013)\, S_{t-2} \\ & + (0.93, 0.13)\, kmS_{t-2} + (-0.43, 0)\, kmP_{t-2} + (-306.18, 0)\, OI \\ & + (-19.27, 35.23)\, OII + (-212.60, 0.005)\, OIII + (-162.34, 0)\, OIV\end{aligned} \tag{18}$$

The value of the objective function (12) for this fuzzy regression model equals 1487.11. Let us assume that $\lambda_0 = 0.1$. Table 4 contains the significance measures for the coefficients of the fuzzy regression model (18).

Table 4. Significance of the fuzzy regression coefficients (Eq. (18))

Variable	Pos($A_j = 0$)
Const	0
Nb_ customers	0
Sales figure	0
km_s	0
km_p	0
OI	0
OII	0.45
OIII	0
OIV	0

The coefficient of the variable *OII*, representing the second region, is insignificant, since we have Pos((–19.27, 3523) = 0) > 0.1. We remove this variable from the set of potential independent variables. The fuzzy regression model determined based on the other variables takes the following form:

$$\begin{aligned}\widehat{Cost}_t = {} & (2156.91, 94.45) + (21.88, 0)\, L_{t-2} + (0.0019, 0)\, S_{t-2} \\ & + (0.096, 0)\, kmS_{t-2} + (-0.43, 0)\, kmP_{t-2} + (-298.53, 0)\, OI \\ & + (-213.53, 0)\, OIII + (-153.81, 0)\, OIV\end{aligned} \tag{19}$$

All the variables in the fuzzy regression model (19) are significant, since for each coefficient in Eq. (19) we have Pos($A_j = 0$) = 0. The value of the objective function for this fuzzy regression model equals 1488.19.

The fuzzy regression model (19) shows that in fact the representatives acting in two regions generate similar costs – region OII and region OV. Neither are present in (19), so they have to be, to some extent, similar, and at the same time more expensive than the other regions – which are represented in (19) by crisp negative coefficients. Also, the other regions can be ranked with respect to the cost they generate per sales representative – this ranking is unequivocal, as in the case discussed the source of fuzziness lies only in the constant of the system (the widths of the other fuzzy coefficients in (19) are zero or close to zero). Thus we obtain the following fuzzy representation and ranking of the cost of the sales representative acting in the specific regions:

Table 5. Fuzzy cost of sales representatives acting in the regions OI, OII, OIII, OIV and OV

Region	Cost
OI	(1858.374, 94.95)
OII	(2156.907, 94.95)
OIII	(1943.38, 94.95)
OIV	(2003.01, 94.95)
OV	(2156.907, 94.95)

The most expensive regions are thus OII and OV, the least expensive is OI, the width of all the fuzzy costs in Table 5 are equal to 94.25, which is about 4–5% of the value of centres. This means that we have fairly crisp knowledge about the cost of the sales representatives in each region.

This width, thus the fuzziness, comes from the constant in (19). This constant represents the cost independent of the factors captured by the independent variables and in fact here is responsible for almost all the cost – the absolute values of the differences between the centre of the constant in (19) and the centres of the fuzzy numbers in Table 5 are not greater than 14% of the centre of the constant. This means that in the case analyzed a major part of the cost of a sales representative is constant and does not depend in a substantial way on the level of professional or private travel, or on the number of customers or the sales figures. Also, the influence of the region is relatively weak.

It is understandable that the coefficient associated with private travel in (19) has a negative coefficient: the more a sales representative uses the car for private purposes, the less time and inclination he has for travelling for business purposes. Also, if he travels for private purposes, he has to pay for the operating expenses proportionally to the number of private kilometers, and thus his total cost decreases.

3. Conclusions

In this paper we have proposed an algorithm for variable selection in fuzzy regression. The algorithm is based on the possibility measure, estimating the possibility of a fuzzy regression coefficient to be equal to zero. The algorithm was applied to modeling energy load and the cost of sales representatives. Unfortunately, the companies did not make more data available for a more reliable verification of the fuzzy regression model, but these two examples clearly show that the proposed method offers a fairly easy tool to estimate a fuzzy regression model for uncertain and changeable magni-

tudes. This in turn shows us which factors influence a given magnitude and also which factors are responsible for its fuzziness, i.e. its uncertainty and changeability.

References

[1] DUBOIS D., PRADE H., *Possibility Theory, An Approach to Computerized Processing of Uncertainty*, Plenum Press, New York, 1988.

[2] CELMINS A., *Multidimensional least-squares fitting of fuzzy models*. Mathematical Modelling, 1987, 9, 669–690.

[3] DIAMOND P., *Fuzzy least squares*, Information Sciences, 1988, 46, 141–157.

[4] D'URSO P., GASTALDI T., *An "ordewise" polynomial regression procedure for fuzzy data*, Fuzzy Sets and Systems, 2002, 130, 1–19.

[5] D'URSO P., SANTARO A., *Goodness of fit and variable selection in fuzzy regression multiple linear regression*, Fuzzy Sets and Systems, 2006, 157 (19), 2627–2647.

[6] GŁADYSZ B., *Interval and Fuzzy Regression*, PWN, Warsaw, 2011 (in Polish).

[7] GŁADYSZ B., KUCHTA D., *Least squares method for L-R fuzzy variables*, W.V. Di Gesu, S.K. Pal, A. Petrosino (Eds.), *Lecture Notes in Computer Science*, Lecture Notes in Artificial Inteligence, LNAI, 2009, 5571, 36–43.

[8] KACPRZYK J., FEDRIZZI M., *Fuzzy Regression Analysis*, Omnitech Press Warsaw, and Physica Heilderberg, 1992.

[9] KÖRNER R., NÄTHER W., *Linear regression with random fuzzy variables*. *Extended classical estimates, bestlinear estimates, least squares estimates*, Information Sciences, 1998, 109, 95–118.

[10] MADDALA G.S., *Introduction to Econometrics*, Wiley, 2001.

[11] SAKAWA M., YANO H., *Multiobjective fuzzy linear regression analysis for fuzzy input-output data*, Fuzzy Sets and Systems, 1992, 47, 173–181.

[12] SAVIC D.A., PEDRYCZ W., *Evaluation of fuzzy linear regression modes*, Fuzzy Sets and Systems, 1991, 23, 51–63.

[13] TROSKA M., *Econometric modelling of cost drivers in the activity based costing*, PhD Thesis, Wroclaw University of Technology, Wroclaw, 2009 (in Polish).

[14] WANG H.-F., TSUAR R.-CH., *Bicriteria variable selection in a fuzzy regression equation*, Computers & Mathematics with Applications, 2000, 40 (6–7), 877–883.

[15] ZADEH L.A., *Fuzzy Sets*, Information and Control, l8, 1965, 338–353.

[16] ZADEH L.A., *Fuzzy sets as a basis of theory of possibility*, Fuzzy Sets and Systems, 1978, 1, 3–28.