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 2010

Barbara GŁADYSZ*

A METHOD FOR DETECTING OUTLIERS IN FUZZY REGRESSION

In this article we propose a method for identifying outliers in fuzzy regression. Outliers in a sample may have an important influence on the form of the regression equation. For this reason there is great scientific interest in this issue. The method presented is analogous to the method of finding outliers based on the studentized distribution of residuals. In order to identify outliers, regression models are constructed with an additional explanatory variable for each observation. Next, the significance of a fuzzy regression coefficient is analysed considering this additional explanatory variable. Illustrative examples are presented.

Keywords: *fuzzy regression, outliers, possibility theory*

1. Introduction

In 1965 Zadeh proposed his possibility theory, [12]. We will present the basic notions of this theory. First, we will present the concept of a fuzzy variable. Let X be a single valued variable whose value is not precisely known. The possibility distribution for X is a normal, quasi concave and upper semi continuous function $\mu_X : \Re \to [0,1]$, see X [1], [13]. The value $\mu_X(x)$ for $x \in \Re$ denotes the possibility of the event that the fuzzy variable X takes the value of x. We denote it as follows:

$$\mu(x) = Pos(X = x). \tag{1}$$

An L–L fuzzy variable is one whose possibility distribution can be expressed in the following form:

* Institute of Organization and Management, Department of Management, Wrocław University of Technology, ul. Smoluchowskiego 25, 50-372 Wrocław, e-mail: barbara.gladysz@pwr.wroc.pl

$$\mu(x) = L\left(\left|\frac{x - m_x}{l_x}\right|\right) \tag{2}$$

where: m_x is a constant and L is a continuous, symmetric function which attains the value 1 for argument 0 and is increasing in the domain of non-negative numbers (m_x is called a centre of fuzzy variable X, l_x – a spread of fuzzy variable X).

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

Examples of such $L(x)$ functions are $L(x) = \max\{0, 1 - |x|^p\}$, $L(x) = (1 + |x|^p)^{-1}$, $L(x) = \exp(- |x|^p)$.

An L–L fuzzy variable will be denoted as $X = (m_X, l_X)$.

Consider two fuzzy variables X, Y with possibility distributions $\mu_X(x)$, $\mu_Y(y)$, respectively. The possibility distributions of the fuzzy variables $Z = X + Y$ and $V = XY$ are defined by means of Zadeh's extension principle [12] as follows:

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

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

We are interested in comparing X to Y, i.e. we want to characterize the possibility of the event that the value taken by X will be greater (not smaller) than the value taken by Y. To describe the possibility of these events Dubois and Prade proposed the following indices [1]:

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

$$Pos(X > Y) = \sup_x \sup_{y \geq x}(\min \mu_X(x), 1 - \mu_Y(y))). \tag{6}$$

We want now to characterise the possibility of the event that the realisation of the variable X will be equal to the variable Y. This possibility is defined as follows [1]:

$$Pos(X = Y) = \min(Pos(X \geq Y), Pos(Y \geq X)). \tag{7}$$

The measure of necessity of the event that the variable X is different to the variable Y is

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

Both indices take a value from the interval [0, 1].

2. Detection of outliers in fuzzy regression

A fuzzy regression is a linear dependence

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

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 the explanatory variables can be real numbers. There are many methods of estimating a fuzzy regression equation. A review of this field can be found in [4].

Let us consider the fuzzy regression model which was introduced by TANAKA et al. [11]. Let us assume that we have observations (y_i, x_i), $y_i, x_i \in \Re$, where y_i is an observation of a triangular fuzzy variable Y_i, $i = 1, ..., n$. The model for estimating the coefficients of fuzzy regression (9) is formulated as:

$$\sum_{j=0}^{k} l_{A_j} \mid x_{ji} \mid \rightarrow \min \tag{10}$$

with the constraints:

$$y_i \le \sum_{j=0}^{k} a_j x_{ji} + \sum_{j=0}^{k} l_{A_j} \mid x_{ji} \mid L^{-1}(\lambda), \quad \text{for} \quad j = 0, ..., k , \tag{11}$$

$$y_i \ge \sum_{j=0}^{k} a_j x_{ji} - \sum_{j=0}^{k} l_{A_j} \mid x_{ji} \mid L^{-1}(\lambda), \quad \text{for} \quad j = 0, ..., k , \tag{12}$$

$$l_{A_j} \ge 0 \quad \text{for} \quad j = 0, ..., k , \tag{13}$$

$$L(x) = \max\{0, 1 - \mid x \mid\} . \tag{14}$$

The form of the optimal solution to the fuzzy regression problem defined by (10)–(14) depends on the assumed λ-level. If we know the optimal solution for a given λ-level, we can find the optimal solution for other λ-levels. Namely, if the coefficients $\widetilde{A}_j = (a_j, l_{a_j})$ give the optimal solution for the λ-level, then the optimal solution for the λ'-level is $\widetilde{A}_j = \left(a_j, \dfrac{L^{-1}(\lambda)}{L^{-1}(\lambda')} l_{a_j} \right)$, $j = 0, ..., k$.

Let us assume that we have fuzzy observations $[Y_i]_\lambda = [L_{Y_i}^{-1}(\lambda), R_{Y_i}^{-1}(\lambda)]$ of the dependent variable for a given λ-level, $i = 1, ..., n$. Let us define $[\hat{Y}_i]_\lambda = [L_{\hat{Y}_i}^{-1}(\lambda),$

$R_{\hat{Y}_i}^{-1}(\lambda)]$ to be estimators of $[Y_i]_\lambda$, $i = 1, ..., n$. The upper fuzzy regression model is given by the following:

$$\sum_{i=1}^{n} [R_{\hat{Y}_i}^{-1}(\lambda) - L_{\hat{Y}_i}^{-1}(\lambda)] \to \min \tag{15}$$

with the constraints

$$L_{\hat{Y}_i}^{-1}(\lambda) \le L_{Y_i}^{-1}(\lambda) \quad \text{for } i = 1, ..., n, \tag{16}$$

$$R_{Y_i}^{-1}(\lambda) \le R_{\hat{Y}_i}^{-1}(\lambda) \quad \text{for } i = 1, ..., n. \tag{17}$$

Many papers consider the issue of identifying outliers in the models of fuzzy regression given by (10)–(14) and (15)–(17). Here, we present some chosen methods in which outliers are identified by using soft boundaries in the regression model: see [2], [7], [8].

To identify outliers, PETERS [8] uses soft limitations on the objective function (10) and constraints (11)–(12), thus constructing the following model

$$\overline{\Lambda} = \frac{1}{n} \sum_{i=1}^{n} \Lambda_i \to \max \tag{18}$$

with the constraints:

$$(1 - \overline{\Lambda}) p_0 - \sum_{i=1}^{n} \sum_{j=0}^{k} l_{A_j} | x_{ij} | \ge -d_0 , \tag{19}$$

$$(1 - \Lambda_i) p_i + \sum_{j=0}^{k} a_j x_{ij} + \sum_{j=0}^{k} l_{A_j} x_{ij} \ge y_i \quad \text{for } i = 1, ..., n, \tag{20}$$

$$(1 - \Lambda_i) p_i - \sum_{j=0}^{k} a_j x_{ij} + \sum_{j=0}^{k} l_{A_j} x_{ij} \ge -y_i \quad \text{for } i = 1, ..., n, \tag{21}$$

$$l_{A_j} \ge 0 \quad \text{for } j = 0, ..., k , \tag{22}$$

$$0 \le \Lambda_i \le 1 \quad \text{for } i = 1, ..., n, \tag{23}$$

$$| L^{-1}(\lambda) | = 1 . \tag{24}$$

where p_0 is the width of the "tolerance interval" for the objective function and p_i is the "tolerance interval" for the observation y_i.

The parameters p_0 and p_i, $i = 1, ..., n$, must be given all together. The parameter d_0 describes the tolerance for the value of the objective function. The suggested value is $d_0 = 0$, which means that the preferences are for a model of crisp type $\left(\sum_{i=1}^{n} \sum_{j=0}^{k} l_{A_j} \mid x_{ij} \mid = 0 \right)$. In the model defined by (14)–(24) the variable $\overline{\Lambda}$ describes a compromise between the objective function (minimization of the width of regression coefficients) and limiting the number of "wrong" observations. Such a formulation of the problem implies that a trade-off between the number of "right" and "wrong" observations will be achieved. Each observation has an influence on the objective function of weight $1/n$. We may also consider the weighted average $\overline{\Lambda} = \sum_{i=1}^{n} w_i \Lambda_i$, where $\sum_{i=1}^{n} w_i = 1$.

Definition 1 [8]. An outlier is an observation which is characterised by a low value of Λ_i.

ÖZELKAN and DUCKSTEIN [7] propose a method for identifying outliers using a multiobjective fuzzy regression model in which, beside the criterion of minimising the sum of the widths of the estimators of the dependent variable, the criterion of minimising the sum of the deviations of the observations from the regression equation. Let us assume that we have fuzzy observations of the dependent variable $[Y_i]_\lambda = [L_{Y_i}^{-1}(\lambda), R_{Y_i}^{-1}(\lambda)]$ for a given λ-level, $i = 1, ..., n$. Let $[\hat{Y}_i]_\lambda = [L_{\hat{Y}_i}^{-1}(\lambda), \ R_{\hat{Y}_i}^{-1}(\lambda)]$. The two-criterion model for detecting outliers with soft boundaries proposed in [7] is

$$\sum_{i=1}^{n} \sum_{j=0}^{k} [R_{\hat{Y}_i}^{-1}(\lambda) - L_{\hat{Y}_i}^{-1}(\lambda)] \to \min , \tag{25}$$

$$\sum_{i=1}^{n} [\varepsilon_{L_i}^{p} + \varepsilon_{R_i}^{p}] \to \min \tag{26}$$

with the constraints:

$$L_{\hat{Y}_i}^{-1}(\lambda) - L_{Y_i}^{-1}(\lambda) \le \varepsilon_{L_i} \quad \text{for} \ \ i = 1, ..., n , \tag{27}$$

$$R_{Y_i}^{-1}(\lambda) - R_{\hat{Y}_i}^{-1}(\lambda)] \le \varepsilon_{R_i} \quad \text{for} \ \ i = 1, ..., n , \tag{28}$$

$$\varepsilon_{L_i}, \ \varepsilon_{R_i} \ge 0 \quad \text{for} \ \ i = 1, ..., n \tag{29}$$

where p is an integer.

In the model given by (25)–(29) both of the constraints (27)–(28) are active, although not all observations are used in the determination of the values of the regression coefficients. Non-dominated Pareto-optimal solutions are considered as a solution of this model.

Definition 2 [7]. An outlier is an observation which is characterised by a significantly large value of ε_{L_i} (ε_{R_i}).

Özelkan and Duckstein proved the following property.

Lemma 1 [7]

The models of Tanaka et al., Peters and the classical regression model are particular cases of the model given by (25)–(29).

Another multicriteria model for the identification of outliers was proposed by GŁADYSZ and KUCHTA [2], in which the criterion $\sum_{i=1}^{n}[\varepsilon_{L_i}^p + \varepsilon_{R_i}^p] \to \min$ is replaced by

$$Me[\varepsilon_{L_i}^p + \varepsilon_{R_i}^p] \to \min . \tag{30}$$

In this article we will propose a method for detecting outliers using n regression models each with an additional explanatory variable corresponding to a given observation of the dependent variable. An observation is considered to be an outlier if the regression component corresponding to the appropriate additional variable has a significant influence on the predicted value of the dependent variable.

3. Method of identifying outliers

When we estimate the regression parameters

$$\hat{y} = a_0 + a_1 x_1 + ... + a_k x_k \tag{31}$$

using the classical least squares method, it may happen that one or more observations have an significant influence on the values of the regression coefficients. Such observations are called outliers. The basic question of data analysis where outliers occur is the following question: which observations are outliers? There are many methods of identifying outliers proposed in the literature, see e.g. [3], [9]. We will describe one of them. It is a method that analyses the changes in the predictions of the dependent variable caused by the exclusion of a particular observation from a data set. The analysis is carried out based on residuals from prediction and studentized residuals. Both kinds of residuals can be defined using a binary variable and the fact that a studentized residual is the residual from prediction divided by its standard error [5]. To determine a studentized residual for the i-th observation, we construct a regression model based on the set of all observations, adding to the set of explanatory variables a variable d_i of the form:

$$d_i = \begin{cases} 1 & \text{for } i\text{-th observation} \\ 0 & \text{for other observations} \end{cases}. \tag{32}$$

The regression coefficient corresponding to the variable d_i is called a predictive residual and the *t*-Student statistic is called a studentized residual. This statistic has a *t*-Student distribution with $(n - k - 1)$ degrees of freedom. If the studentized residual (*t*-Student statistic) belongs to the critical set, the *i*-th observation is inferred to be an outlier.

In this article we propose a method of detecting outliers in fuzzy regression analogous to the method described above for classical regression. Let us consider the fuzzy regression (9)

$$\hat{Y} = A_0 + A_1 X_1 + ... + A_k X_k .$$

To verify whether the *i*-th observation is an outlier, we build a fuzzy regression model based on the set of all observations, adding to the set of explanatory variables a variable d_i described by formula (32). Hence, we determine the coefficients of fuzzy regression for the model

$$\hat{Y} = A_0 + A_1 X_1 + ... + A_k X_k + A_D d_i . \tag{33}$$

Let us characterise the regression model (33) as a sum of two components:

$$\hat{Y} = \hat{Y}_0 + A_D d_i , \tag{34}$$

where $\hat{Y}_0 = A_0 + A_1 X_1 + ... + A_k X_k$.

Next, we analyse whether the second component $A_D d_i$ of the model (34) has a significant influence on the prediction of the value of the dependent variable for the *i*-th observation. If its influence is significant, the *i*-th observation is classified as an outlier. Let us define an outlier in the following way.

Definition 3. The *i*-th observation is classified as an outlier when the following occurs for the model (34)

$$Pos(\hat{Y} = \hat{Y}_0) \le \lambda_0 . \tag{35}$$

The parameter λ_0 is specified subjectively by a decision-maker.

Inequality (35) is equivalent to the inequality

$$Nec(\hat{Y} \ne \hat{Y}_0) \ge 1 - \lambda_0 , \tag{36}$$

which means that the necessity that the prediction $\hat{Y}$ is not equal to $\hat{Y}_0$ is not lower than λ_0.

Let us formulate a linear model of coefficients estimation in fuzzy regression (33) for the case of real data. Its form will be the following

$$\sum_{j=0}^{k} l_{A_j} \mid x_{ij} \mid + l_{A_D} \to \min \tag{37}$$

with the constraints:

$$y_i \le \sum_{j=0}^{k} a_j x_{ij} + \sum_{j=0}^{k} l_{A_j} \mid x_{ij} \mid L^{-1}(\lambda) \quad \text{for } i = 1, ..., n,\ i \ne i_0, \tag{38}$$

$$y_i \ge \sum_{j=0}^{k} a_j x_{ij} - \sum_{j=0}^{k} l_{A_j} \mid x_{ij} \mid L^{-1}(\lambda) \quad \text{for } i = 1, ..., n,\ i \ne i_0, \tag{39}$$

$$y_i \le \sum_{j=0}^{k} a_j x_{ij} + \sum_{j=0}^{k} l_{A_j} \mid x_{ij} \mid L^{-1}(\lambda) + a_D + l_{A_D} L^{-1}(\lambda) \quad \text{for } i = i_0, \tag{40}$$

$$y_i \ge \sum_{j=0}^{k} a_j x_{ij} - \sum_{j=0}^{k} l_{A_j} \mid x_{ij} \mid L^{-1}(\lambda) + a_D - l_{A_D} L^{-1}(\lambda) \quad \text{for } i = i_0, \tag{41}$$

$$l_{A_j},\ l_{A_D} \ge 0 \quad \text{for } j = 0, ..., k\ , \tag{42}$$

$$L(x) = \max\{0, 1 - \mid x \mid\}\ . \tag{43}$$

In the model given by (37)–(43), the constraints (40)–(41) for observation i_0 are soft boundary constraints. So the method of identifying outliers proposed in this article involves relaxation of the constraints. A relaxing variable is a fuzzy variable. An observation is treated as an outlier if the corresponding relaxing variable has a significant influence on prediction.

A linear model for estimating regression coefficients can be constructed in a similar way to the regression model given by (25)–(29) for fuzzy data.

4. Examples

To illustrate the method of detecting outliers proposed in this article we will present its implementation for a set of real and fuzzy data.

4.1. Example 1

The data are presented in Table 1. All the observations are real numbers. The fifth observation is an outlier.

Table 1. Data for Example 1

i	1	2	3	4	5	6	7	8	9	10
x	1	2	3	4	5	6	7	8	9	10
y	1.5	2.3	2.7	4.4	9.4	6.3	6.5	7.8	8.5	10.5

Source: [8].

Let us consider the following fuzzy regression

$$\hat{Y} = A_0 + A_1 x \,. \tag{44}$$

For the data from Table 1, Tanaka's regression model (44) takes the form

$$\hat{Y} = 2.25, 2.4) + (0.95, 0)x \,.$$

We will determine ten regression models (44) each with one additional variable d_i for observation $i = 1, ..., 10$.

$$\hat{Y} = A_0 + A_1 x + A_D d_i \,. \tag{45}$$

The results of such estimation are presented in Table 2. For $i = 1, ..., 8$, $i \neq 5$ the same regression models were obtained in the form of $\hat{Y} = (2.25, 2.4) + (0.95, 0)x + (0, 0)d_i$. In these models the coefficient of the variable $A_D = (0, 0)$. So the element $A_D d_i$ in the model (45) for observations $i = 1, ..., 8$, $i \neq 5$ is insignificant.

Table 2. Regression coefficients for the model (45) and the significance of the component $A_D d_i$

i	A_0	A_1	A_D	Pos $(\hat{Y} = \hat{Y}_0)$	Nec $(\hat{Y} \neq \hat{Y}_0)$
1	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
2	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
3	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
4	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
5	(0.175, 0.325)	(0,975, 0.325)	(4.35, 0)	0	1
6	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
7	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
8	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
9	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0
10	(2.25, 2.4)	(0.95, 0)	(0, 0)	1	0

Source: Author's own work.

For the fifth observation we obtain the following regression equation

$$\hat{Y} = (0.175,\ 0.325) + (0.975,\ 0.325)x + (4.35,\ 0)d_5 \tag{46}$$

where the measure of significance of the element $A_D d_5$ is $\mathrm{Nec}(\hat{Y} \neq \hat{Y}_0) = 1$. So the element (4.35, 0)d_5 in the model (46) significantly influences the predicted value of the dependent variable. The fifth observation has been detected as an outlier. The results of estimation are presented in Figures 1 and 2. Tanaka's regression model calculated for the complete data set is presented in Figure 1. It should be noted that the outlier (fifth observation) implies a big width using Tanaka's model. Figure 2 shows the model estimated after excluding the fifth observation from the data set.

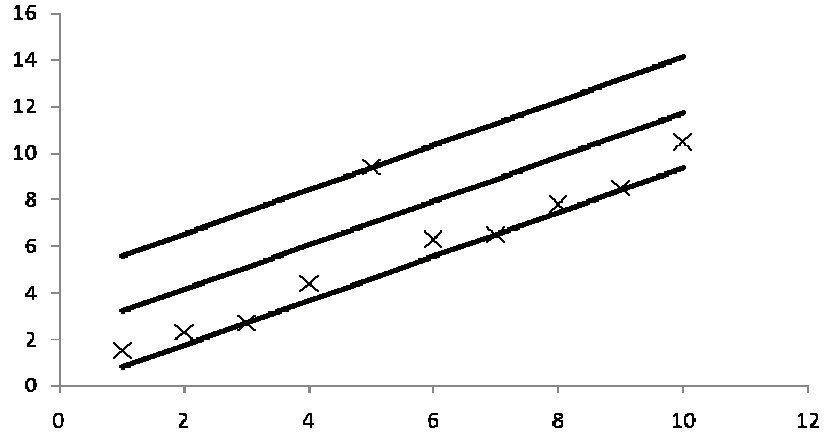

Fig. 1. Observations and fuzzy regression (44) for the complete data set ($\lambda = 0$, $\lambda = 1$)

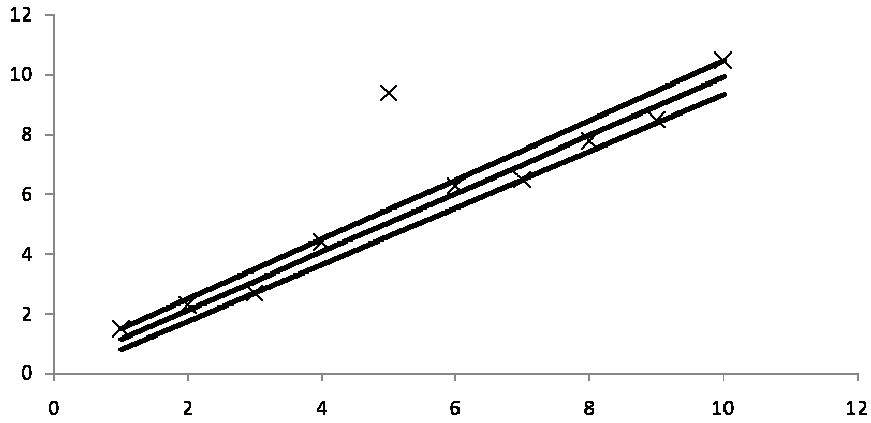

Fig. 2. Observations and fuzzy regression (44) for the data without the fifth observation ($\lambda = 0$, $\lambda = 1$)

If we construct regression models using: the Peters model, the Özelkan–Duckstein model and the Gładysz–Kuchta model, for the data from Table 1 we get the following results:

- the Peters model

$$\hat{\tilde{Y}} = (0.229,\ 0.271) + (0.974,\ 0.025)x,$$

- Özelkan and Duckstein's multicriterion model

$$\hat{\tilde{Y}} = (0.258,\ 0.408) + (0.967,\ 0.017)x,$$

- Gładysz and Kuchta's multicriterion model

$$\hat{\tilde{Y}} = (0.258,\ 0.408) + (0.967,\ 0.017)x.$$

The data and their weights Λ_i (the Peters model) and the values of the relaxing variables ε_{Li}, ε_{Ri} are presented in Table 3. The results from the Gładysz–Kuchta model are identical to those obtained from the Özelkan–Duckstein model. For the fifth observation (the outlier) the weight Λ_5 is relatively small compared to the other weights and the value of the relaxation variable ε_{R5} is much bigger than the value of the relaxation variable for the other observations.

Table 3. Indicators of outliers for Example 1

x	y	Λ	ε_L	ε_R
1	1.5	1	0	0
2	2.3	1	0	0
3	2.7	0.989	0	0
4	4.4	1	0	0
5	9.4	0.610	0	3.817
6	6.3	1	0	0
7	6.5	0.990	0	0
8	7.8	1	0	0
9	8.5	1	0	0
10	10.5	1	0	0

Source: [7].

All these three methods identify the fifth observation as an outlier, as does the method proposed in this article.

4.2. Example 2

Let us consider the data from [1, 0], presented in Table 4. Both the observations of the dependent variable and the independent variable are symmetric triangular fuzzy numbers ($L(x) = \max\{0, 1 - |x|\}$). The fifth observation is an outlier.

Let us consider the following fuzzy regression model

$$\hat{Y} = A_0 + a_1 X \,, \tag{47}$$

where the coefficient A_0 is a fuzzy coefficient and the coefficient a_1 of the fuzzy explanatory variable is a real coefficient.

Table 4. Data for Example 2

i	x	l_x	y	l_y
1	2.0	0.5	4.0	0.5
2	3.5	0.5	5.5	0.5
3	5.5	1.0	7.5	1.0
4	7.5	0.5	6.5	0.5
5	8.5	0.5	18.5	0.5
6	10.5	1.0	8.0	1.0
7	11.0	0.5	10.5	0.5
8	12.5	0.5	9.5	0.5

Source: [10].

To detect outliers, let us introduce additional variables d_i according to formula (32) and construct regression models:

$$\hat{Y} = A_0 + a_1 X + A_D d_i \,. \tag{48}$$

Table 5. Regression coefficients for model (48) and significance of the component $A_D d_i$

i	A_0	A_1	A_D	$\mathrm{Pos}(\hat{Y} = \hat{Y}_0)$	$\mathrm{Nec}(\hat{Y} \neq \hat{Y}_0)$
1	(10.05, 6.85)	0.40	(–0.30, 0)	0.97	0.03
2	(9.70, 6.86)	0.43	(–0.30, 0)	0.98	0.02
3	(9.70, 6.86)	0.44	(–0.30, 0)	0.98	0.02
4	(9.70, 6.86)	0.44	(–0.30, 0)	0.98	0.02
5	(3.75, 1.50)	0.50	(10.50, 0.25)	0.00	1.00
6	(8.45, 6.65)	0.60	(–0.50, 0)	0.97	0.03
7	(9.70, 6.86)	0.44	(2.56, 0)	0.82	0.18
8	(9.70, 6.86)	0.44	(0.91, 0)	0.94	0.06

Source: Author's own work.

The coefficients of regression in (48) for $i = 1, ..., n$ are presented in Table 5. Table 5 also shows the measures for the event that the component $A_D d_i$ in model (48) is not significant. Let us assume that $\lambda_0 = 0.6$. It can be observed that the component $A_D d_i$

has an important influence on predicting the dependent variable only for the fifth observation. For the other observations $\mathrm{Nec}(\hat{Y} \neq \hat{Y}_0) \leq 1 - \lambda_0$.

So the fifth observation has been detected as an outlier. The results of estimation are shown in Figures 3 and 4. Figure 3 shows the fuzzy regression model for the complete set of data and Figure 4 illustrates the estimated model after excluding the fifth observation from the data set.

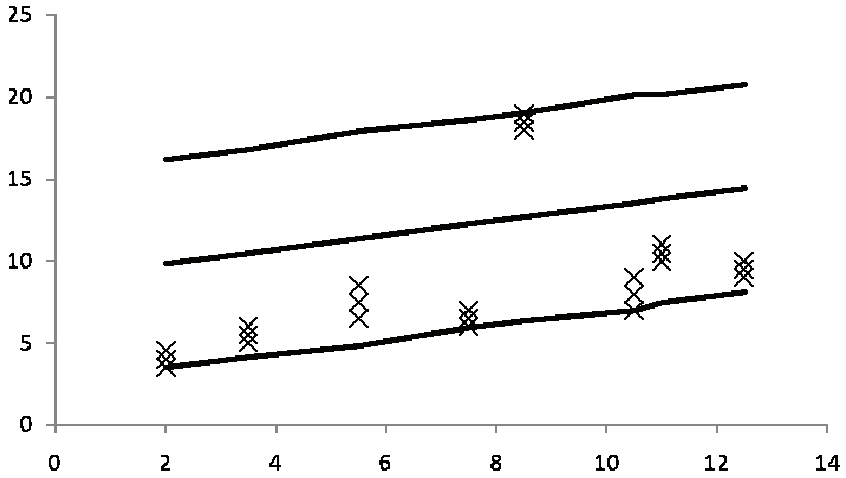

Fig. 3. Observations and fuzzy regression (47) for the complete data set ($\lambda = 0$ and $\lambda = 1$)

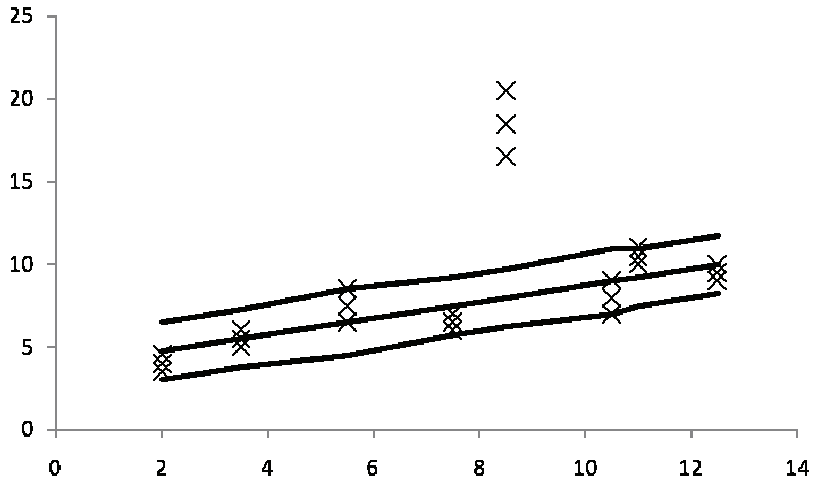

Fig. 4. Observations and fuzzy regression for model (47) for the complete data set ($\lambda = 0$ and $\lambda = 1$)

When we construct the Özelkan–Duckstein model and the Gładysz–Kuchta model for the data from Table 4, we obtain the following models:

- Özelkan–Duckstein multicriterion model

$$\hat{\tilde{Y}} = (3.944,\ 0.278) + 0.444\tilde{X},$$

- Gładysz–Kuchta multicriterion model

$$\hat{\tilde{Y}} = (3.19,\ 1.4) + 0.6\tilde{X}.$$

The relaxation variables ε_{Li}, ε_{Ri} are presented in Table 6.

Table 6. Indicators of outliers for Example 2

Özelkan–Duckstein		Gładysz–Kuchta	
ε_L	ε_R	ε_L	ε_R
0.83	0.00	0.53	0.00
0.00	0.00	0.09	0.00
0.00	1.39	0.01	0.00
0.78	0.00	0.00	0.01
0.00	10.78	0.00	14.21
0.89	0.00	0.50	0.00
0.00	1.67	0.00	0.53
0.00	0.00	0.00	0.09

Source: Author's own work.

For the fifth observation the value of the relaxation variable is much bigger than the values of the relaxing variables for other observations, both for the Özelkan–Duckstein model and the Gładysz–Kuchta model. Both methods identified the fifth observation as an outlier, as does the method presented in this article.

5. Conclusions

In this article we have proposed a method of detecting outliers in fuzzy regression. In order to identify outliers, regression models are constructed each with an additional explanatory variable corresponding to a given observation. Next, the significance of the fuzzy regression coefficient corresponding to the additional explanatory variable is analysed. If its influence is significant, this observation is treated as an outlier. Two examples were presented. In the first example the observations are real numbers (realisations of fuzzy numbers). In the second example the data are fuzzy numbers.

References

[1] DUBOIS D., PRADE H., *Possibility Theory: An Approach to Computerized Processing of Uncertainty*, Plenum Press New York, 1988.
[2] GŁADYSZ B., KUCHTA D., *Multi criteria programming in robust estimation for interval data*, Foundation of Computing and Decision Sciences, 2008, 33 (2), 193–207.
[3] HUBER P.J., *Robust statistics*, John Wiley & Sons, Hoboken, New Jersey, 2004.
[4] KACPRZYK J., FEDRIZZI M., *Fuzzy Regression Analysis*, Omnitech Press Warsaw and Physica-Verlag Heilderberg, Warsaw, 1992.
[5] MADDALA G.S., *Ekonometria*, PWN, Warszawa, 2006.
[6] NASRABADI M.M., NASRABADI E., NASRABADI A.R., *Fuzzy linear regression analysis: A multi-objective programming approach*, Applied Mathematics and Computation, 2005, 163, 245–251.
[7] ÖZELKAN E.C., DUCKSTEIN L., *Multi-objective fuzzy regression: a general framework*, Computers and Operation Research, 2002, 27, 635–652.
[8] PETERS G., *Fuzzy linear regression with fuzzy intervals*, Fuzzy Sets and Systems, 1994, 63, 45–55.
[9] ROUSSEEUW P., LEROY A.M., *Robust regression and outlier detection*, John Wiley & Sons, 1987.
[10] SAKAWA M., YANO H., *Multiobjective fuzzy linear regression analysis for fuzzy input-output data*, Fuzzy Sets and Systems, 1992, 47, 173–181.
[11] TANAKA H., UEJIMA S., ASAI K., *Linear regression analysis with fuzzy model*, IEEE Transaction on Systems Man and Cybernetics, 1982, 12, 903–907.
[12] ZADEH L.A., *Fuzzy Sets*, Information and Control, 1965, (8), 338–353.
[13] ZADEH L.A., *Fuzzy sets as a basis of theory of possibility*, Fuzzy Sets and Systems, 1978, (1), 3–28.