Iranian Journal of Mathematical Sciences and Informatics
Vol. 20, No. 1 (2025), pp 193-204
DOI: 10.61186/ijmsi.20.1.193

A Fuzzy Multivariate Regression Model to Control Outliers and Multicollinearity Based on Exact Predictors and Fuzzy Responses

Gholamreza Hesamian[a], Mohamad Ghasem Akbari[b], Mehdi Shams[c*]

[a]Department of Statistics, Payame Noor University, Tehran 19395-3697, Iran
[b]Department of Statistics, Ferdowsi University of Mashhad, Mashhad, Iran
[c]Department of Statistics, University of Kashan, Kashan, Iran

E-mail: gh.hesamian@pnu.ac.ir
E-mail: g_z_akbari@birjand.ac.ir
E-mail: mehdishams@kashanu.ac.ir

Abstract. Multivariate regression is an approach for modeling the linear relationship between several variables. This paper proposed a ridge methodology with a kernel-based weighted absolute error target with exact predictors and fuzzy responses. Some standard goodness-of-fit criteria were also used to examine the performance of the proposed method. The effectiveness of the proposed method was then illustrated through two numerical examples including a simulation study. The effectiveness and advantages of the proposed fuzzy multiple linear regression model were also examined and compared with some well-established methods through some common goodness-of-fit criteria. The numerical results indicated that our prediction/estimation gives more accurate results in cases where multicollinearity and/or outliers occur in the data set.

Keywords: Goodness-of-fit measure, Robust, Multicollinearity, Kernel function, Outlier.

2000 Mathematics subject classification: 62A86, 60A86, 62Jxx.

*Corresponding Author

Received 16 July 2020; Accepted 08 October 2023

1. Introduction

Multivariate regression analysis is a powerful technique for predicting the response variable from two or more predictors. However, at the presence of outliers and/or multicollinearity in the data set, the conventional multiple linear regression methods may result in poor prediction and wrong signs in estimators [12]. Fuzzy regression analysis has been introduced by [24] and successfully employed for numerous real applications of regression analysis in case where the observed data are fuzzy values instead of exact ones. Many studies have been conducted on fuzzy multiple regression models. The observations of the predictors of such methods are either fuzzy (for instance, see [1, 3, 4, 5, 7, 9, 10, 15, 20]) or real numbers (for instance, see [11, 14, 17, 19]). In addition, at the presence of outliers, robust fuzzy linear regression analysis techniques have drawn attention and a significant progress has been made over the past decades [8, 16, 22]. Moreover, [2] proposed a semi-parametric partial linear regression model with fuzzy predictors and fuzzy response at the presence outliers and multicollinearity in the predictors.

Inspired by [2], this study is an attempt to improve the conventional fuzzy multiple regression models in case where both outliers and multicollinearity occur in data set in cases where the observed data set includes exact predictors and fuzzy responses. Then the performance of the proposed method was compared with some common fuzzy multiple regression models in terms of several goodness-of-fit criteria via a simulated study and an applied example.

The rest of this paper is organized as follows: section 2 reviews some concepts including fuzzy numbers and a distance measure. In Section 3, a fuzzy multiple regression model is estimated using a ridge-based method adopted with a weighted absolute error target function with exact predictors and fuzzy responses. Section 4 illustrates two numerical examples including a simulation study and an applied example to evaluate the effectiveness and performance of the proposed method relative to other fuzzy multiple regression methods in terms of some common performance measures. The main contributions of this paper are summarized in Section 5.

2. Fuzzy numbers

This section reviews some basic definitions of fuzzy numbers based on [9] which will be used in next sections. A fuzzy set $\widetilde{A}$ of $\mathbb{R}$ (the real line) is defined by its membership function $\mu_{\widetilde{A}} : \mathbb{R} \to [0,1]$. In addition, a fuzzy set $\widetilde{A}$ of $\mathbb{R}$ is called a fuzzy number if it is normal, i.e. there is a unique $x^*_{\widetilde{A}} \in \mathbb{R}$ so that $\mu_{\widetilde{A}}(x^*_{\widetilde{A}}) = 1$, and for every $\alpha \in [0,1]$, the set $\widetilde{A}[\alpha] = \{x \in \mathbb{R} : \mu_{\widetilde{A}}(x) \geq \alpha\}$ is a nonempty compact interval in $\mathbb{R}$. A fuzzy number $\widetilde{A}$ is called a LR-fuzzy number if there are real numbers a, $l_a \geq 0, r_a \geq 0$ and strictly decreasing and continuous functions $L, R : [0,1] \to [0,1]$ such that:

$$\mu_{\widetilde{A}}(x) = \begin{cases} L(\frac{a-x}{l_a}) & a - l_a \leq x \leq a, \\ R(\frac{x-a}{r_a}) & a < x \leq a + r_a, \\ 0 & x \in \mathbb{R} - [a - l_a, a + r_a]. \end{cases}$$

In this case $\widetilde{A}$ is simply denoted by $(a; l_a, r_a)_{LR}$. The most commonly used LR-fuzzy numbers are triangular fuzzy numbers (**TFN**s) which are employed for handling the impression included in data set. The membership function of a triangular fuzzy number, denoted by $\widetilde{A} = (a; l_a, r_a)_T$, is given by:

$$\mu_{\widetilde{A}}(x) = \begin{cases} \frac{x-(a-l_a)}{l_a} & a - l_a \leq x \leq a, \\ \frac{a+r_a-x}{r_a} & a \leq x \leq a + r_a, \\ 0 & x \in \mathbb{R} - [a - l_a, a + r_a]. \end{cases}$$

Some common operations between the two triangular fuzzy numbers $\widetilde{A} = (a; l_a, r_a)_T$ and $\widetilde{B} = (b; l_b, r_b)_T$ are defined as follows [18]:

1) (Addition) $\widetilde{A} \oplus \widetilde{B} = (a + b; l_a + l_b, r_a + r_b)_T$.
2) (scalar multiplication):

$$\lambda \otimes \widetilde{A} = \begin{cases} (\lambda a; \lambda l_a, \lambda r_a)_T, & if\ \lambda > 0, \\ (\lambda a; -\lambda r_a, -\lambda l_a)_T, & if\ \lambda < 0. \end{cases} \tag{2.1}$$

In addition, a square error distance between the two **FN**s of $\widetilde{A}$ and $\widetilde{B}$ was employed in this paper; defined as:

$$D_p(\widetilde{A}, \widetilde{B}) = (\int_0^1 2\alpha(\frac{|\widetilde{A}_\alpha^L - \widetilde{B}_\alpha^L|^{1/p} + |\widetilde{A}_\alpha^U - \widetilde{B}_\alpha^U|^{1/p}}{2})^p d\alpha)^{1/p}. \tag{2.2}$$

Any **FN**s of $\widetilde{A}$, $\widetilde{B}$ and $\widetilde{C}$ satisfy the following conditions:

- $D_p(\widetilde{A}, \widetilde{B}) = 0$ if and only if $\widetilde{A} = \widetilde{B}$,
- $D_p(\widetilde{A}, \widetilde{B}) = D_p(\widetilde{B}, \widetilde{A})$,
- $D_p(\widetilde{A}, \widetilde{C}) \leq \big(D_p(\widetilde{A}, \widetilde{B}) + D_p(\widetilde{B}, \widetilde{C})\big)$.

3. Fuzzy multiple regression model

Consider the following fuzzy multiple linear regression model based on n statistical units of $(\widetilde{y}_i, \boldsymbol{x}_i = (x_{1i}, x_{2i}, \ldots, x_{ki})^T)$:

$$\widetilde{y}_i = \oplus_{j=1}^k (\widetilde{\beta}_j \otimes x_{ij}) \oplus \widetilde{\epsilon}_i,\ i = 1, 2, \ldots, n, \tag{3.1}$$

where

(1) $\widetilde{y}_i = (y_i; l_{y_i}, r_{y_i})_T$ denote fuzzy responses,
(2) x_{ij} are predictors,
(3) $\widetilde{\beta}_j = (\beta_j; l_{\beta_j}, r_{\beta_j})_T$ present unknown fuzzy coefficients,
(4) $\widetilde{\epsilon}_i$ indicate fuzzy error terms.

To estimate the coefficients of model (3.1), an extended ridge estimator of $\widetilde{\beta}_1, ..., \widetilde{\beta}_k$ was considered based on a weighted-variant of least absolute estimation:

$$(\widehat{\boldsymbol{\beta}}, \boldsymbol{L}_{\widehat{\boldsymbol{\beta}}}, \boldsymbol{R}_{\widehat{\boldsymbol{\beta}}}) = \min_{\boldsymbol{\beta}\in\mathbb{R}^k, \boldsymbol{L}_{\boldsymbol{\beta}}\in(0,\infty)^k, \boldsymbol{R}_{\boldsymbol{\beta}}\in(0,\infty)^k} \quad \frac{1}{n}\sum_{i=1}^n w_h(\boldsymbol{x}_i) D_1(\widetilde{y}_i, \oplus_{j=1}^k(\widetilde{\beta}_j \otimes x_{ij}))$$
$$s.t. \qquad \max\{\boldsymbol{\beta}^\top \boldsymbol{\beta}, \boldsymbol{L}_{\boldsymbol{\beta}}^\top \boldsymbol{L}_{\boldsymbol{\beta}}, \boldsymbol{R}_{\boldsymbol{\beta}}^\top \boldsymbol{R}_{\boldsymbol{\beta}}\} \leq \lambda^2, \tag{3.2}$$

(1) D_1 is the absolute error distance between the two fuzzy numbers $\widetilde{A}$ and $\widetilde{B}$.
(2) $w_h(\boldsymbol{x}_i) = \frac{\sum_{j=1}^n K(\frac{||\boldsymbol{x}_j - \boldsymbol{x}_i||}{h})}{\sum_{i=1}^n \sum_{j=1}^n K(\frac{||\boldsymbol{x}_j - \boldsymbol{x}_i||}{h})}$.
(3) $K(.)$ is a kernel satisfying Mercer's conditions [26] and $h > 0$ controls the degree of smoothing called the bandwidth of K relevant to x_i.
(4) $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$, $\boldsymbol{L}_{\boldsymbol{\beta}} = (l_{\beta_1}, \ldots, l_{\beta_k})^T$ and $\boldsymbol{R}_{\boldsymbol{\beta}} = (r_{\beta_1}, \ldots, r_{\beta_k})^T$.

Remark 3.1. A common criterion to detect the multicollinearity in a conventional multiple regression is variance inflation factor (VIF) [21]. The VIF [13] of each variable can be computed using $VIF_j = \frac{1}{1-R_j^2}$ where R_j^2 is the coefficient of determination by regressing the jth predictor on the remaining predictors. As a rule of thumb, a VIF value that exceeds 5 indicates a problematic amount of collinearity which is equivalent to $R_j^2 > 0.80$. Moreover, an extended Cook's distance criterion was employed to detect the outliers [2] in this paper. For this purpose, the potential outliers were selected using a $\overline{D}$-chart with lower and upper control limits of $LCL = \overline{D} - 3\frac{S_D}{\sqrt{n}}$ and $UCL = \overline{D} + 3\frac{S_D}{\sqrt{n}}$.

Remark 3.2. To examine the performance of the proposed fuzzy robust varying coefficient regression model, following commonly used performance measures were employed to estimate the predicting accuracy and compare different models [9]:

(1) Root mean square error :

$$RMSE = \sqrt{\frac{\sum_{i=1}^n D_2^2(\widetilde{y}_i, \widetilde{\widehat{y}}_i)}{n}}. \tag{3.3}$$

(2) Mean absolute relative error:

$$MARE = \frac{1}{n}\sum_{i=1}^n \frac{\int_0^1 |\widetilde{y}_i(x) - \widetilde{\widehat{y}}_i(x)|dx}{\int_0^1 \widetilde{y}_i(x)dx}. \tag{3.4}$$

(3) Similarity measure:

$$MSM = \frac{1}{n}\sum_{j=1}^n S_{UI}(\widetilde{\widehat{y}}_j, \tilde{y}_j), \tag{3.5}$$

where

$$S_{UI}(\widetilde{\widehat{y}}_j, \tilde{y}_j) = \frac{Card(\widetilde{\widehat{y}}_j \cap \tilde{y}_j)}{Card(\widetilde{\widehat{y}}_j \cup \widehat{y}_j)},$$

in which $\cap$, $\cup$ denote the intersection and union operators on the space of fuzzy numbers, respectively; and $Card(\widetilde{A})$ denotes the cardinal number of $\widetilde{A}$.

(4) Coefficient of determination (COD):

$$COD = \frac{\sum_{i=1}^{n} D_2^2(\widetilde{y}_i, \widetilde{\widehat{y}}_i)}{\sum_{i=1}^{n} D_2^2(\widetilde{y}_i, \widetilde{\overline{\widehat{y}}})}. \tag{3.6}$$

It should be noted that $0 \leq COD \leq 1$. A COD value of 0 indicates that the regression model dose not fit the set of data points while the value of 1 indicates that the regression model perfectly fits the set of data points.

(5) Cross validation [25]:

$$CV = \frac{1}{n} \sum_{j=1}^{n} D_2^2\big(\tilde{y}_j, \widetilde{\widehat{y}}^{(j)}\big), \tag{3.7}$$

where $\widetilde{\widehat{y}}^{(j)}$, $j = 1, 2, \ldots, n$ denotes the estimated value of fuzzy response related to the smoothing parameter h based on the data $(\widetilde{y}_i, \boldsymbol{x}_i^T)$ in which $i \neq j$.

3.1. **Algorithm for estimating model's components.** Assume that both outliers and multicollinearity occur in data set. To estimate the unknown components of model (3.1), the vector of fuzzy parameters of $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_1, \widetilde{\beta}_2, \ldots, \widetilde{\beta}_k)^T$ and bandwidth $h > 0$ should be simultaneously estimated based on fuzzy observations $(\widetilde{y}_1, \mathbf{x}_1^\top)$, ..., $(\widetilde{y}_n, \mathbf{x}_n^\top)$ and a specified kernel function K. Note that gaussian and triweight kernels were utilized in this paper as two popular ones. Since all the above target functions are connected to each other, a hybrid optimization algorithm is required for their estimation. However, since the smoothing parameter h often strongly influences the degree of $\widetilde{\boldsymbol{\beta}}$ estimation, one of the main objectives of kernel curve fitting approach is to choose the optimal value for the smoothing parameter h. For this purpose, a common criterion called "generalized cross validation" (**GCV**) [25] was employed. The optimal vector of bandwidth $\boldsymbol{h}$ can be then evaluated as:

$$h_{opt} = arg \min_h \frac{\sum_{i=1}^{n} d(\widetilde{y}_i, \widetilde{\widehat{y}}_i)}{n(1 - tr(W_p))^2}, \tag{3.8}$$

where

$$W_p = \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \ddots & \vdots \\ w_{n1} & \cdots & w_{nn} \end{pmatrix}; \; w_{ij} = \frac{K\left(\dfrac{t_i - t_j}{h}\right)}{\sum_{j=1}^{n} K\left(\dfrac{t_i - t_j}{h}\right)},$$

and $tr(A)$ denotes the trace of an $n \times n$ square matrix A. Now , the following algorithm can be suggested in this paper to find the fuzzy coefficient $\widetilde{\boldsymbol{\beta}}$, optimal values for bandwidth h and tuning parameter λ:

Step 1: Let $\lambda^{(0)}$ be the optimal value of tuning parameter λ using the training Cross-validation method based on the ordinary multiple regression model $y_i = \oplus_{j=1}^{k}(\beta_j \otimes x_{ij}) + \epsilon_i$, $i = 1, 2, \ldots, n$.

Step 2 Let $h^{(0)} = n^{-0.2}$ be an initial value for the bandwidth.

Step 3: According to initial values of $h^{(0)}$ and $\lambda^{(0)}$, the optimal value $\widetilde{\widehat{\boldsymbol{\beta}}}^{(0)}$ can be evaluated by Eq. (3.2).

Step 4: Using the initial values of $\lambda^{(0)}$ and optimal value of $\widetilde{\widehat{\boldsymbol{\beta}}}^{(0)}$, the optimal value of $h^{(1)}$ can be selected by minimizing GVC criterion given in Eq. (3.8).

Step 5: Using the optimal values of $\widehat{h}^{(1)}$ and initial value of $\widetilde{\widehat{\boldsymbol{\beta}}}^{(0)}$, the optimal value $\widehat{\lambda}^{(1)}$ can be selected by maximizing MSM criterion.

Step 6: Using the optimal value of $\widehat{h}^{(1)}$ and $\widehat{\lambda}^{(1)}$, the optimal value of $\widetilde{\widehat{\boldsymbol{\beta}}}^{(1)}$ can be evaluated by Eq. (3.2).

Step 7: For every $\varepsilon > 0$, if

$$d^*(\widetilde{\widehat{\boldsymbol{\beta}}}^{(1)}, \widetilde{\widehat{\boldsymbol{\beta}}}^{(0)}) < \varepsilon,$$

then $\widehat{h}^{(1)}$, $\widehat{\lambda}^{(1)}$ and $\widetilde{\widehat{\boldsymbol{\beta}}}^{(1)}$ are the optimal solutions, otherwise, return to Step 4 and repeat the algorithm until

$$d^*(\widetilde{\widehat{\boldsymbol{\beta}}}^{(i+1)}, \widetilde{\widehat{\boldsymbol{\beta}}}^{(i)}) < \varepsilon,$$

where

$$d^*(\widetilde{\widehat{\boldsymbol{\beta}}}^{(i+1)}, \widetilde{\widehat{\boldsymbol{\beta}}}^{(i)}) = \max\{||\widehat{\boldsymbol{\beta}}^{(i+1)} - \widehat{\boldsymbol{\beta}}^{(i)}||, ||\boldsymbol{L}_{\widehat{\boldsymbol{\beta}}^{(i+1)}} - \boldsymbol{L}_{\widehat{\boldsymbol{\beta}}^{(i)}}||, ||\boldsymbol{R}_{\widehat{\boldsymbol{\beta}}^{(i+1)}} - \boldsymbol{R}_{\widehat{\boldsymbol{\beta}}^{(i)}}||\}.$$

Therefore, $\widetilde{\widehat{\boldsymbol{\beta}}}^{(i+1)}$, $\widehat{\lambda}^{(i+1)}$ and $\widehat{h}^{(i)}$ are the optimal estimators.

4. Numerical examples

is examined via some numerical examples in data set possessing outliers and multicollinearity. Moreover, the proposed regression model is also examined with three well-established fuzzy multiple regression models introduced by [6], [23] and [27]. They showed the superiority of their method over other fuzzy multiple linear regression models. We compared our fuzzy regression model with them to investigate the effectiveness of the proposed method under both outliers and multicollinearity. For this purpose, the performance measures given in Remark 3.2 were applied.

Example 4.1. (**A simulation study**) Here, a set of $m = 10$ simulated data set with size of $n = 200$ was generated according to the following fuzzy multiple linear regression model:

$$\widetilde{y}_i = \oplus_{j=0}^{5}(\widetilde{\beta}_j \otimes x_{ij}) \oplus \widetilde{\epsilon}_i, \tag{4.1}$$

where $\widetilde{\beta}_0 = (0; 0.5, 0.5)_T$, $\widetilde{\beta}_1 = (1; 0.1, 0.2)_T$, $\widetilde{\beta}_2 = (2; 0.2, 0.4)_T$, $\widetilde{\beta}_3 = (3; 0.3, 0.6)_T$, $\widetilde{\beta}_4 = (4; 0.4, 0.8)_T$ and $\widetilde{\beta}_5 = (5; 0.5, 1)_T$. For this purpose, in each simulation process, a random sample $(\widetilde{y}_i, (x_{i1}, x_{i2}, ..., x_{i5})^T)$ with size of $n = 500$ is generated by the following steps:

Step 1:

(1) Generate a random sample with size of $n = 200$ of exact predictors $(x_{i1}, \ldots, x_{i5}) \sim N_5(\mathbf{0}, \Sigma)$ in which $[\Sigma_{lk}] = Cov(x_{il}, x_{ik}) = 0.5^{|l-k|}$.

(2) Let $\widetilde{z}_i = \oplus_{j=0}^{5}(\widetilde{\beta}_j \otimes x_{ij})$.

Step 2:

(1) For a randomly selected $I_1 = \{i_1^1, i_2^1, \ldots, i_{15}^1\} \subseteq \{1, 2, 3, \ldots, 200\}$, let $\widetilde{y}_i = \widetilde{z}_i \oplus (90; u_1^i, u_2^i)_T$ where $i \in I_1$.

(2) For a randomly selected $I_2 = \{i_1^2, i_2^2, \ldots, i_{15}^2\} \subseteq \{1, 2, 3, \ldots, 200\} - I_1$, let $\widetilde{y}_i = \widetilde{z}_i \oplus (130; u_1^i, u_2^i)_T$ where $i \in I_2$ and u_1^i, u_2^i are observed random variables from $U(2, 4)$ and $U(3, 5)$, respectively.

Performance measures	$\overline{MSM}$	$\overline{MARE}$	$\overline{RMSE}$	$\overline{CV}$	$\overline{COD}$
Proposed (Triweight kernel)	0.71	16.50	18.31	14.89	0.72
Proposed (Normal kernel)	0.69	16.80	18.96	15.83	0.70
Choi and Yoon	0.55	21.66	23.52	21.01	0.61
Zeng, Feng and Li	0.58	22.17	21.38	19.10	0.62
Kelkinnama and Taheri	0.45	26.37	25.47	27.28	0.56

Table 1. Mean performance measure values of 10 simulated data set with size of $n = 100$ along with some common fuzzy regression models in Example 4.1.

Step 3: Let $\widetilde{y}_i = \widetilde{z}_i$ for $i \in \{1, 2, 3, \ldots, 200\} - (I_1 \cup I_2)$.

Using the above simulation process, one could imagine 1- a set of potential outliers in each simulation data set and 2- multicollinearity between predictors. In this regard, the mean values of performance measures along with some common fuzzy multivariate regression models are summarized in Table 1 and 2. Comparing the various methods, in terms of the applied goodness-of-fit criteria it is seen that 1- the proposed method gives more accurate results compared to the other methods according to both Gaussian and triweight kernels, and 2) the proposed method adopted with triweight kernel provides a better result than gaussian kernel for this simulated data set. Specifically, consider such comparison based on the 5^{th} simulated data and triweight kernel. For this purpose, the

Method	Coefficients	MSM	$MARE$	$RMSE$	CV	COD
Proposed (Triweight kernel) $\widehat{\lambda}^2 = 15$ and $\widehat{h} = 5$	$\widetilde{\widehat{\beta}}_0 = (3.554; 0, 8.896)_T$, $\widetilde{\widehat{\beta}}_1 = (0.829; 0, 2.623)_T$ $\widetilde{\widehat{\beta}}_2 = (0.720; 0.0003, 1.613)_T$, $\widetilde{\widehat{\beta}}_3 = (0.390; 0, 0)_T$ $\widetilde{\widehat{\beta}}_4 = (0.406; 0, 0.0003)_T$, $\widetilde{\widehat{\beta}}_5 = (4.838; 0, 0)_T$	0.76	15.24	16.50	13.52	0.79
Proposed (Normal kernel) $\widehat{\lambda}^2 = 20$ and $\widehat{h} = 4$	$\widetilde{\widehat{\beta}}_0 = (3.468; 0, 4.233)_T$, $\widetilde{\widehat{\beta}}_1 = (0.755; 0, 0.855)_T$ $\widetilde{\widehat{\beta}}_2 = (0.634; 0.024, 0.681)_T$, $\widetilde{\widehat{\beta}}_3 = (0.393; 0, 0.373)_T$ $\widetilde{\widehat{\beta}}_4 = (0.436; 0.006, 0.423)_T$, $\widetilde{\widehat{\beta}}_5 = (6.653; 0, 0.566)_T$	0.72	17.28	17.28	15.02	0.72
Choi and Yoon	$\widetilde{\widehat{\beta}}_0 = (5.299; 0.010, 3.102)_T$, $\widetilde{\widehat{\beta}}_1 = (1.104; 0.0079, 0.839)_T$ $\widetilde{\widehat{\beta}}_2 = (0.536; 0, 0.839)_T$, $\widetilde{\widehat{\beta}}_3 = (0.400; 0, 0.017)_T$ $\widetilde{\widehat{\beta}}_4 = (1.207; 0.200, 0.008)_T$, $\widetilde{\widehat{\beta}}_5 = (2.627; 0, 0)_T$	0.58	21.66	23.52	21.07	0.60
Zeng, Feng and Li	$\widetilde{\widehat{\beta}}_0 = (4.034; 0.1, 0.830)_T$, $\widetilde{\widehat{\beta}}_1 = (1.036; 0, 0.829)_T$ $\widetilde{\widehat{\beta}}_2 = (0.903; 0.058, 2.882)_T$, $\widetilde{\widehat{\beta}}_3 = (1.603; 0, 0.207)_T$ $\widetilde{\widehat{\beta}}_4 = (1.202; 0, 0)_T$, $\widetilde{\widehat{\beta}}_5 = (4.902; 0.034, 0)_T$	0.57	22.17	21.38	19.10	0.61
Kelkinnama and Taheri	$\widetilde{\widehat{\beta}}_0 = (15.023; 0.935, 10.66)_T$, $\widetilde{\widehat{\beta}}_1 = (3.539; 1.842, 1.628)_T$ $\widetilde{\widehat{\beta}}_2 = (1.723; 0, 0.623)_T$, $\widetilde{\widehat{\beta}}_3 = (-1.935; 0.005, 0)_T$ $\widetilde{\widehat{\beta}}_4 = (1.0452; 1.526, 0.016)_T$, $\widetilde{\widehat{\beta}}_5 = (3.772; 0, 0)_T$	0.49	33.55	24.01	25.60	0.59

TABLE 2. Performance measures corresponding to 5^{th} simulated sample in Example 4.1.

variance inflation factors (**VIF**) of the predictors $x_1 - x_5$ are $\text{VIF}_{x_1} = 3.623$, $\text{VIF}_{x_2} = 7.164$, $\text{VIF}_{x_3} = 7.264$, $\text{VIF}_{x_4} = 6.526$ and $\text{VIF}_{x_5} = 4.224$. This indicates a very strong multicollinearity between the predictors. In addition, according to $\overline{D}$-chart in Fig. 1, it can be seen that the data set contains some potential outliers. Comparing the various methods, in terms of the goodness-of-fit criteria MSM, $MARE$ and $RMSE$, it is also observed that the proposed method performs well compared to the other methods.

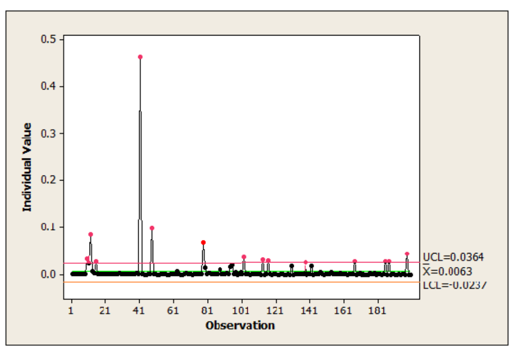

FIGURE 1. $\overline{D}$-chart for detecting the outliers in Example 4.1.

EXAMPLE 4.2. Suppose that the data set given in Table 3 denotes the house prices in Japan [6]. We wish to estimate the fuzzy regression model for house prices $\widetilde{y}$ based on x_1: materials, x_2: the lot of first, x_3: second floor, x_4:

Observation	$\widetilde{y}$	x_1	x_2	x_3	x_4	x_5
1	$(6060; 550)_T$	1	38.09	36.43	5	1
2	$(7100; 50)_T$	1	6210	26.50	6	1
3	$(8080; 400)_T$	1	63.70	44.71	7	1
4	$(8260; 150)_T$	1	74.52	38.09	8	1
5	$(8650; 750)_T$	1	75.38	41.40	7	2
6	$(8520; 450)_T$	2	52.99	26.49	4	2
7	$(9170; 700)_T$	2	62.93	26.49	5	2
8	$(10310; 200)_T$	2	72.04	33.12	6	3
9	$(10920; 600)_T$	2	76.12	43.06	6	2
10	$(12030; 100)_T$	2	90.26	42.64	7	2
11	$(13940; 350)_T$	3	85.70	31.33	7	3
12	$(14200; 250)_T$	3	95.27	27.64	6	3
13	$(16010; 300)_T$	3	105.98	27.64	6	3
14	$(16320; 500)_T$	3	79.25	66.81	6	3
15	$(16990; 650)_T$	3	120.5	32.25	6	3

TABLE 3. Data set in Example 4.1.

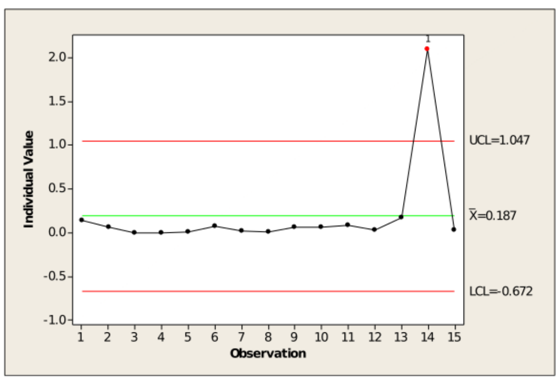

FIGURE 2. $\overline{D}$-chart for detecting the outliers in Example 4.2.

number of general rooms and x_5: Japanese style rooms. The house prices are reported in symmetric **TFN**s. For this data set, $\overline{D}$-chart was plotted in Fig. 2 to detect the outliers. As it is observed, the data set contains an outlier. In addition, **VIF** values for predictors of $x_1 - x_5$ were evaluated as $\text{VIF}_{x_1} = 8.111$, $\text{VIF}_{x_2} = 4.28$, $\text{VIF}_{x_3} = 1.278$, $\text{VIF}_{x_4} = 2.459$ and $\text{VIF}_{x_5} = 6.109$. It can be concluded that there exists a very strong multicollinearity between $x_1 - x_5$. Therefore, we face a situation in which data set involves both multicollinearity and outliers. In this regard, the results of the proposed method (with triweight and gaussian kernels) along with some common fuzzy multivariate regression models were summarized in Table 4. Comparing the various methods, it can be seen that the proposed method based on both kernels gives more accurate results compared to the other methods. In addition, triweight kernel leads to a better performances than gaussian kernel.

Method	Coefficients	MSM	$MARE$	$RMSE$	CV	COD
Proposed (Gaussian kernel) $\widehat{\lambda}^2 = 350$ and $\widehat{h} = 0.01$	$\widetilde{\widehat{\beta}}_0 = (-497.565; 31.386, 136.417)_T$, $\widetilde{\widehat{\beta}}_1 = (2132.017; 45.957, 136.417)_T$ $\widetilde{\widehat{\beta}}_2 = (94.628; 0.188, 0)_T$, $\widetilde{\widehat{\beta}}_3 = (76.656; 0.718, 0)_T$ $\widetilde{\widehat{\beta}}_4 = (-436.907; 4.7860, 0)_T$, $\widetilde{\widehat{\beta}}_5 = (-34.816; 76.631, 200.419)_T$	0.76	6.74	5.10	4.20	0.85
Proposed (Triweight kernel) $\widehat{\lambda}^2 = 450$ and $\widehat{h} = 0.005$	$\widetilde{\widehat{\beta}}_0 = (-497.565; 31.386, 136.417)_T$, $\widetilde{\widehat{\beta}}_1 = (2132.17; 45.975, 136.417)_T$ $\widetilde{\widehat{\beta}}_2 = (94.628; 0.1888, 0)_T$, $\widetilde{\widehat{\beta}}_3 = (76.656; 0.718, 0)_T$ $\widetilde{\widehat{\beta}}_4 = (-436.907; 4.786, 0)_T$, $\widetilde{\widehat{\beta}}_5 = (-34.816; 76.631; 200.419)_T$	0.85	5.03	4.23	3.47	0.90
Choi and Yoon	$\widetilde{\widehat{\beta}}_0 = (-8.571; 0, 0)_T$, $\widetilde{\widehat{\beta}}_1 = (94.372; 0, 179.470)_T$ $\widetilde{\widehat{\beta}}_2 = (130.104; 0, 179.470)_T$, $\widetilde{\widehat{\beta}}_3 = (44.877; 0, 0)_T$ $\widetilde{\widehat{\beta}}_4 = (-127.101; 143.616, 0)_T$, $\widetilde{\widehat{\beta}}_5 = (76.970; 0, 133.380)_T$	0.65	8.89	7.35	6.63	0.65
Zeng, Feng and Li	$\widetilde{\widehat{\beta}}_0 = (-1828.256; 40.637, 0)_T$, $\widetilde{\widehat{\beta}}_1 = (2635.497; 14.626, 2752.240)_T$ $\widetilde{\widehat{\beta}}_2 = (82.0876; 0.0229, 57.069)_T$, $\widetilde{\widehat{\beta}}_3 = (72.474; 1.750, 46.180)_T$ $\widetilde{\widehat{\beta}}_4 = (-155.228; 0.143, 0)_T$, $\widetilde{\widehat{\beta}}_5 = (-158.613; 28.484, 0)_T$	0.69	7.006	6.53	5.16	0.74
Kelkinnama and Taheri	$\widetilde{\widehat{\beta}}_0 = (-26.446; 0, 0)_T$, $\widetilde{\widehat{\beta}}_1 = (427.191; 0, 549.855)_T$ $\widetilde{\widehat{\beta}}_2 = (126.445; 0, 0)_T$, $\widetilde{\widehat{\beta}}_3 = (68.766; 0, 0)_T$ $\widetilde{\widehat{\beta}}_4 = (-456.136; 68.395, 0)_T$, $\widetilde{\widehat{\beta}}_5 = (296.697; 0, 327.512)_T$	0.67	7.39	6.47	5.32	0.72

Table 4. Coefficients of model and performance measures corresponding to some fuzzy regression techniques in Example 4.2.

5. Conclusion

This study improved the conventional fuzzy multiple linear regression model at the presence of outliers and/or multicollinearity in data set. For this purpose, having a set of fuzzy responses and exact predictors, unknown fuzzy coefficients were estimated based on an optimization algorithm via a weighted absolute error distance measure adopted with a ridge estimation in the fuzzy environment. The proposed regression model was also examined and compared with several existing fuzzy multiple regression models in terms of some common goodness-of-fit criteria used in the fuzzy environment. The numerical results clearly indicated higher efficiency of our fuzzy multiple regression model compared to the other fuzzy multiple regression models in cases where outliers and/or multicollinearity occur in data set. The proposed method was examined for two popular kernels (Gaussian and triweight). However, the best results were obtained when triweight kernel was used. Future researches should be focused on extending the proposed model for other types of imprecision such as intuitionistic fuzzy numbers.

Acknowledgments

The Authors would like to thank the reviewers for the thoughtful comments and the constructive suggestions which helped to improve the quality of this study.

References

1. M. G. Akbari, G. Hesamian, Linear Model with Exact Inputs and Interval-valued Fuzzy Outputs, *IEEE Transactions on Fuzzy Systems*, **26**, (2018), 518-530.
2. M. G. Akbari, G. Hesamian, A Partial-robust-ridge Based-regression Model with Fuzzy Predictors-Responses, *Journal of Computational and Applied Mathematics*, **351**, (2019), 290-301.
3. G. Alfonso, A. F. R. L. de Hierro, C. Roldan, A Fuzzy Regression Model based on Finite Fuzzy Numbers and Its Application to Real-World Financial Data, *Journal of Computational and Applied Mathematics*, **318**, (2017), 47-58.
4. J. Chachi, M. Roozbeh, A Fuzzy Robust Regression Approach Applied to Bedload Transport Data, *Communications in Statistics-Simulation and Computation*, **46**, (2017), 1703-1714.
5. S. H. Choi, J. H. Yoon, Fuzzy Regression Based on Non-Parametric Methods, *Wseas Transaction on Systems and Control*, **13**, (2018), 20-25.
6. S. H. Choi, J. H. Yoon, General Fuzzy Regression Using Least Squares Method, *International Journal of Systems Science*, **41**, (2010), 477-485.
7. A. F. R. L. de Hierro, J. Martinez-Moreno, C. Aguilar-Pena, C. R. L. de Hierro, Estimation of a Fuzzy Regression Model Using Fuzzy Distances, *IEEE Transactions on Fuzzy Systems*, **24**, (2016), 344-359.
8. P. D'Urso, R. Massari, A. Santoro, Robust Fuzzy Regression Analysis, *Information Sciences*, **181**, (2011), 4154-4174.
9. G. Hesamian, M. G. Akbari, M. Asadollahi, Fuzzy Semi-Parametric Partially Linear Model with Fuzzy Inputs and Fuzzy Outputs, *Expert Systems with Applications*, **71**, (2017), 230-239.
10. G. Hesamian, M. G. Akbari, Fuzzy Quantile Linear Regression Model Adopted with a Semi-Parametric Technique based on Fuzzy Predictors and Fuzzy Responses, *Expert Systems with Applications*, **118**, (2019), 585–597.
11. G. Hesamian, M. G. Akbari, A Robust Varying Coefficient Approach to Fuzzy Multiple Regression Model, *Journal of Computational and Applied Mathematics*, **375**, (2020), 1-13.
12. A. E. Hoerl, R. W. Kennard, Ridge Regression: Biased Estimation for Non-Orthogonal Problems, *Technometrics*, **12**, (1970), 55-67.
13. G. James, D. Witten, T. Hastie, R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R,* 8^{th} *ed.*, Springer-Verlag, New York, 2017.
14. H. Y. Jung, J. H. Yoon, S. H. Choi, Fuzzy Linear Regression Using Rank Transform Method, *Fuzzy sets and systems*, **274**, (2015), 97-108.
15. U. T. Khan, C. Valeo, A New Fuzzy Linear Rgression Approach for Dissolved Oxygen Prediction, *Hydrological Sciences Journal*, **60**, (2015), 1096-1119.
16. K. Kula, A. Apaydin, Fuzzy Robust Regression Analysis based on Ranking of Fuzzy Sets, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **16**, (2008), 663-681.

17. K. S. Kula, F. Tank, T. E. Dalkylyc, A Study on Fuzzy Robust Regression and Its Application to Insurance, *Mathematical and Computational Applications*, **17**, (2012), 223-234.
18. K. H. Lee, *First Course on Fuzzy Theory and Applications*, Springer-Verlag, Berlin, 2005.
19. W. J. Lee, H. Y. Jung, J. H. Yoon, S. H. Choi, The Statistical Inferences of Fuzzy Regression Based on Bootstrap Techniques, *Soft Computing*, **19**, (2015), 883-890.
20. J. Li, W. Zeng, J. Xie, Q. Yin, A New Fuzzy Regression Model based on Least Absolute Deviation, *Engineering Applications of Artificial Intelligence*, **52**, (2016), 54-64.
21. S. Sheather, A Modern Approach to Regression with R, Springer Science and Business Media, Wiely, New York, 2009.
22. B. Y. Sohn, Robust Fuzzy Linear Regression Based on M-estimators, *Journal of Applied Mathematics and Computing*, **18**, (2005), 596-597
23. S. M. Taheri, M. Kelkinnama, Fuzzy Linear Regression Based on Least Absolute Deviations. *Iranian Journal of Fuzzy Systems*, **9**, (2012), 121-140.
24. H. Tanaka, I. Hayashi, J. Watada, Possibilistic Linear Regression Analysis for Fuzzy Data, *European Journal of Operational Research*, **40**, (1989), 389-396.
25. G. Wahba, *Spline Models for Observational Data*, Society for industrial and applied mathematics (Siam), Taiwan, 1990.
26. L. Wasserman, *All of Nonparametric Statistics*, Springer, New York, 2007.
27. W. Zeng, Q. Feng, J. Li, Fuzzy Least Absolute Linear Regression, *Applied Soft Computing*, **52**, (2017), 1009-1019.