Fuzzy regression model using fuzzy partition

To cite this article: Hye-Young Jung et al 2019 J. Phys.: Conf. Ser. 1334 012019

View the article online for updates and enhancements.
Fuzzy regression model using fuzzy partition

Hye-Young Jung¹, Woo-Joo Lee² and Seung Hoe Choi†

¹Faculty of Liberal Education, Seoul National University, Seoul, 08826, South Korea
²Department of Mathematics, Yonsei University 50 Yonsei-Ro, Seoul, South Korea
†School of Liberal Arts and Science, Korea Aerospace University, South Korea

E-mail: ¹shchoi@kau.ac.kr

Abstract. In this paper, we introduce a novel method to estimate fuzzy regression model when the center and spread have different pattern. We combine the least absolute deviations estimation for the center with discriminant analysis for spread. Also, we use the fuzzy partition to categorize spreads of the dependent variable into several classes. A numerical study is given to show that our method outperforms the estimation methods based on a fuzzy regression model with the assumption of equivalent functions for the center and spread.

1. Introduction

The regression model is one of the most famous tools with many successful applications in various fields, such as economics, medicine, and engineering. In regression analysis, the involved experimental data are assumed to be precise, and numerous statistical treatments have been proposed for modeling precisely defined crisp data [1]. However, we encounter sometimes non-precise data. To solve this problem, the fuzzy set theory has been developed by Zadeh in [2]. The fuzzy set is a useful tool for mathematically expressing the concept of non-precise data. Fuzzy regression model is a regression model with data and/or parameters expressed as fuzzy sets. After Tanaka in [12] proposed the fuzzy regression model, many extensions and applications have been developed [1-14]. If the center of the fuzzy number used in the fuzzy regression analysis varies and the spread is repeated by the same number, it may be more effective to estimate the fuzzy regression model by separately estimating the center and spread without using the same function.

In this paper, we propose a novel estimation method to estimate the center and spread by combining the least absolute deviations (LAD) estimation method for the center with discriminant analysis for spreads. Also, we use the fuzzy partition to categorize given spreads into several classes.

This paper is organized as follows: In Section 2, some preliminary concepts required in order to develop the main results are presented. In Section 3, we discussed the problem of constructing an estimation method about the parameters of fuzzy regression model when the center and spread show different patterns. Then, we proposed a new algorithm for the estimation of the fuzzy regression model by combining the LAD estimation method and the discriminant analysis. In Section 4, A numerical example is also given to explain our results and compare with the existing methods.

2. Preliminaries

The fuzzy set theory provides a suitable framework for dealing with non-precise data. Following [3] and [4], we introduce some definitions regarding the fuzzy sets and the fuzzy numbers.
A fuzzy set $A$ is a set of ordered pairs

$$A = \{(x, \mu_A(x)) : x \in A\},$$

where $\mu_A : X \to [0, 1]$ is a membership function of $A$.

The support of $A$ defined on $\mathbb{R}$ is a crisp set defined as

$$\text{supp} A = \{x | x \in \mathbb{R}, \mu_A(x) > 0\}.$$ 

A fuzzy number, $A$, is a normal and convex subset of the real line, $\mathbb{R}$, with bounded support. As a special case, a fuzzy number $A$, denoted by $A = (a, l_a, r_a)_{LR}$, is said to be a $LR$-fuzzy number if its membership function is denoted by

$$\mu_A(x) = \begin{cases} 
L_A \left( \frac{a-x}{l_a} \right) & \text{for } 0 \leq a-x \leq l_a, \\
R_A \left( \frac{x-a}{r_a} \right) & \text{for } 0 \leq x-a \leq r_a, \\
0 & \text{for otherwise,}
\end{cases}$$

where $a$ is the center, $l_a$ and $r_a$ are the left spread and the right spread. $L_A$ and $R_A$ are functions verifying the properties of the class of fuzzy sets such that $L(0) = R(0) = 1$ and $L(x) = R(x) = 0$, $x \in \mathbb{R} \setminus [0, 1)$. In particular, if $L(x) = R(x) = 1 - x$ in $A = (a, l_a, r_a)_{LR}$, then $A$ is called a triangular fuzzy number and is denoted by $A = (a, l_a, r_a)_T$.

3. Fuzzy regression based on fuzzy partition

Fuzzy linear regression is formulated as follows:

$$Y_i = A_0 + A_1 X_{i1} + \cdots + A_p X_{ip} + E_i, \quad i = 1, \ldots, n$$

where $Y_i = (m_Y, l_Y, r_Y)_T$, $A_i = (m_A_i, l_A_i, r_A_i)_T$ and $X_{ik} = (m_{X_{ik}}, l_{X_{ik}}, r_{X_{ik}})_T$ are fuzzy numbers.

Generally, the spread of the fuzzy numbers in the fuzzy sample varies greatly. However, a set of the spread of the fuzzy numbers in the fuzzy sample obtained from the survey or the preference or quality of goods consists of repeating numbers. If the center of the fuzzy number used in the fuzzy regression analysis varies and the spread is repeated by the same number, it may be more effective to estimate the fuzzy regression model by separately estimating the center and spread. In particular, if the spread of the dependent variable of the fuzzy regression model is expressed only in a few numbers due to repetition, it may be effective estimation method for the spread to use the discriminative analysis proposed by Fisher [15, 16].

Discriminant analysis is a technique that is used to predict the probability of belonging to a given category based on one or multiple independent variables when the dependent variable is categorical and the independent variable is interval in nature. If the spreads of the dependent variable given in the fuzzy regression model are represented by only a few numbers due to repetition, it may be effective to divide the interval that contains all spreads into subintervals of equal width without overlapping. This is because if the width is not constant, effective discriminative analysis results cannot be obtained. For this purpose, this paper uses fuzzy partition, which generalizes the mathematical partition. The fuzzy partition allows the use of the distance between the width of a given dependent variable and the center of the fuzzy set constituting the fuzzy partition.

If the set $\{A_i : A_i \subseteq X, i = 1, \ldots, n + 1\}$ is the fuzzy partition of the universe $X$, then the membership degree $A_i(x)$ satisfies the following condition

$$\sum_{i=1}^{n+1} A_i(x) = 1 \quad \text{for } \forall x \in X,$$

Figure 1. Fuzzy partition

where $\hat{A}_i$ is the support of a fuzzy set $A_i$.

In this paper, the following four steps are used to predict the dependent variable of the fuzzy regression model.

The proposed algorithm to predict value of the dependent variable

Step 1. Define the fuzzy partition for the left spread of dependent variable:
Let $l_{y_i^{(m)}}$ be the maximum value of $S_l$, $l_{y_i^{(1)}}$ be the minimum value of $S_l$, and $m_l$ be the number of fuzzy sets constituting the fuzzy partition. The fuzzy partition of the set $S_l = \{l_{yi} : i = 1, \cdots, m_l\}$ which is the set of left spread of a given dependent variable can be defined as follows.

1. \[
\frac{(l_{y_i^{(n)}}-l_{y_i^{(1)}})}{l} = (m_l-1)
\]

2. \[
\{L_i : \hat{L}_i \subseteq [l_{y_i^{(1)}}, l_{y_i^{(n)}}], i = 1, \cdots, m_l\}, \quad L_i = \left(m_l, \frac{1}{2}, \frac{1}{2}\right) \quad (2 \leq i \leq m_l - 1), \quad m_{l_i} = l_{y_i^{(1)}} + \frac{(i-1)}{2}, \quad L_1 = (l_{y_i^{(1)}}, 0, \frac{1}{2}), \quad L_{m_l} = (l_{y_i^{(n)}}, 0, 0)
\]

Step 2. Define the fuzzy partition for the right spread of dependent variable:
Let $r_{y_i^{(m)}}$ be the maximum value of $S_r$, $r_{y_i^{(1)}}$ be the minimum value of $S_r$, and $m_r$ be the number of fuzzy sets constituting the fuzzy partition. The fuzzy partition of the set $S_r = \{r_{yi} : i = 1, \cdots, n\}$ which is the set of right spread of a given dependent variable can be defined as follows.

1. \[
\frac{(r_{y_i^{(n)}}-r_{y_i^{(1)}})}{l} = (m_r-1)
\]

2. \[
\{R_i : \hat{R}_i \subseteq [r_{y_i^{(1)}}, r_{y_i^{(n)}}], i = 1, \cdots, m_r\}, \quad R_i = \left(m_r, \frac{1}{2}, \frac{1}{2}\right) \quad (2 \leq i \leq m_r - 1), \quad m_{r_i} = r_{y_i^{(1)}} + \frac{(i-1)}{2}, \quad R_1 = (r_{y_i^{(1)}}, 0, \frac{1}{2}), \quad R_{m_r} = (r_{y_i^{(n)}}, \frac{1}{2}, 0)
\]

Step 3. Predict the center of the dependent variable:
Using the set of the center of the dependent variable $\{(m_{X_i1}, \cdots, m_{X_{ip}}, m_{y_i}) : i = 1, \cdots, n\}$, obtain the predicted value $m_{y_i}$ of the center of the dependent variable by minimizing the following objective function

\[
\sum_{i=1}^{n} |m_{Y_i} - m_{A_0} - m_{A_1}m_{X_{i1}} - \cdots - m_{A_p}m_{X_{ip}}|
\]

Step 4. Predict the spreads of the dependent variable:
Using the set $\{(m_{y_i}, l_{X_{i1}}, \cdots, l_{X_{ip}}) : i = 1, \cdots, n\}$ (or $\{(m_{y_i}, r_{X_{i1}}, \cdots, r_{X_{ip}}) : i = 1, \cdots, n\}$) and the discriminant analysis, obtain the predicted value $l_{y_i}$ (or $r_{y_i}$) for the spread of the dependent variable.
4. Numerical Example and comparison study

In this section, example is illustrated to compare the proposed method with other methods using the least squares estimation. Diamond estimated the fuzzy regression model by minimizing the sum of the square of the end point of support and the center of residuals [7]. Chachi and Taheri estimated the fuzzy regression model by minimizing the sum of the square of the left and right endpoints of the \(\alpha\)-level of residuals [4]. Two performance measures are used to compare the effectiveness of the estimated fuzzy regression model. One is the \(D(Y, \hat{Y})\) comparing the difference between the observed value and the estimated value, and the other is \(I(Y, \hat{Y})\) comparing the overlapping area.

The performance measure \(D(Y, \hat{Y})\) based on the difference between the observed value and the predicted value is defined as follows.

\[
D(Y, \hat{Y}) = \sum_{i=1}^{n} m_d(Y_i, \hat{Y}_i),
\]

where

\[
m_d(Y_i, \hat{Y}_i) = \frac{\int_{-\infty}^{\infty} |\mu_{Y_i}(x) - \mu_{\hat{Y}_i}(x)| \, dx}{\int_{-\infty}^{\infty} |\mu_{Y_i}(x)| \, dx} + h_d(Y_i(0), \hat{Y}_i(0))
\]

and \(h_d(Y_i(0), \hat{Y}_i(0)) = \inf\{\inf\{|a - b| : a \in Y_i(0)\} : b \in \hat{Y}_i(0)\}\). The more efficient model has \(M_d\) value closer to zero.

The performance measure \(I(Y, \hat{Y})\) based on the overlapping area between the observed value and the predicted value is defined as follows.

\[
I(Y, \hat{Y}) = \sum_{i=1}^{n} c(Y_i, \hat{Y}_i),
\]

where

\[
c(Y_i, \hat{Y}_i) = \frac{\int_{-\infty}^{\infty} \min\{\mu_{Y_i}(x), \mu_{\hat{Y}_i}(x)\} \, dx}{\int_{-\infty}^{\infty} \mu_{Y_i}(x) \, dx + \int_{-\infty}^{\infty} \mu_{\hat{Y}_i}(x) \, dx - \int_{-\infty}^{\infty} \min\{\mu_{Y_i}(x), \mu_{\hat{Y}_i}(x)\} \, dx}.
\]

The more efficient method has the smaller value of \(D(Y, \hat{Y})\) and the larger value of \(I(Y, \hat{Y})\).

Example D’Urso and Chachi and Taheri consider a fuzzy regression model, which consist of two input fuzzy variables and a fuzzy output variables. The data in Table 1 are the performance of the 30 good-quality Roman restaurants.

From table 1, we obtain \(S_l = \{0, 0.25, 0.5, 0.75\}\) which is the set of left spread of dependent variable.

This shows that the spreads of 30 samples are expressed in only four numbers. The fuzzy partition of \(S_l\) can be defined as follows:

\[
L_k = \begin{cases} 
(0, 0, 0.25)_T & k = 1, \\
((k - 1)0.25, 0.25, 0.25)_T & 2 \leq k \leq 4, \\
(1, 0.25, 0)_T & k = 5.
\end{cases}
\]
Table 1. Fuzzy observed value and predicted value

<table>
<thead>
<tr>
<th>no</th>
<th>X_{i1}</th>
<th>X_{i2}</th>
<th>Y_i</th>
<th>\bar{Y}_i</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(5.75, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>2</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(5.25, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>3</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.08, 0.25, 0.50)_Т</td>
</tr>
<tr>
<td>4</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(9, 0.00, 1.00)_Т</td>
<td>(9, 0.00, 1.00)_Т</td>
<td>(7.92, 0.75, 1.00)_Т</td>
</tr>
<tr>
<td>5</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7.42, 0.50, 1.00)_Т</td>
</tr>
<tr>
<td>6</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(5, 0.00, 1.00)_Т</td>
<td>(6.08, 0.25, 0.50)_Т</td>
</tr>
<tr>
<td>7</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>8</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(5, 0.00, 1.00)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>9</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>10</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.08, 0.25, 0.50)_Т</td>
</tr>
<tr>
<td>11</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>12</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.00, 0.50, 1.00)_Т</td>
</tr>
<tr>
<td>13</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(9, 0.00, 1.00)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>14</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>15</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>16</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>17</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.08, 0.25, 0.50)_Т</td>
</tr>
<tr>
<td>18</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>19</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(6.50, 0.75, 1.25)_Т</td>
</tr>
<tr>
<td>20</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(9, 0.00, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>21</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>22</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(7.00, 0.75, 1.25)_Т</td>
</tr>
<tr>
<td>23</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(9, 0.00, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.50, 0.75, 1.25)_Т</td>
</tr>
<tr>
<td>24</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>25</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>26</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>27</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(5, 0.00, 1.00)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>28</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(8, 0.75, 1.00)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7.00, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>29</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6.50, 0.50, 1.25)_Т</td>
</tr>
<tr>
<td>30</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(7, 0.50, 1.25)_Т</td>
<td>(6, 0.25, 0.50)_Т</td>
<td>(6.08, 0.25, 0.50)_Т</td>
</tr>
</tbody>
</table>

The Fisher linear discriminant functions which classify the centers \( m_{l_k} \) of fuzzy numbers \( L_k \) using the spreads of independent variables \( l_{X_1} \) and \( l_{X_2} \) are as follows:

\[
\begin{align*}
  m_{l_1} &= -16 + 46.33 l_{X_1} + 10.98 l_{X_2} \\
  m_{l_2} &= -11.55 + 37.38 l_{X_1} + 11.6 l_{X_2} \\
  m_{l_3} &= -16.56 + 47.34 l_{X_1} + 13.11 l_{X_2} \\
  m_{l_4} &= -17.838 + 50.07 l_{X_1} + 11.64 l_{X_2}
\end{align*}
\]

From table 1, we obtain \( S_r = \{0.5, 1, 1.25\} \) which is the set of right spread of dependent variable. This shows that the spreads of 30 samples are expressed in only three numbers. The fuzzy partition of \( S_r \) can be defined as follows:

\[
  R_k = \begin{cases} 
    (0.25, 0, 0.25)_T & k = 1, \\
    (0.25k, 0.25, 0.25)_T & 2 \leq k \leq 4, \\
    (1.25, 0.25, 0)_T & k = 5.
  \end{cases}
\]

The results according to the Fisher linear discriminant function, which categorizes the center \( m_{r_k} \) of the fuzzy number \( R_k \), are given in Table 1. The result of least absolute deviation estimation using the centers \( \{(m_{X_{i1}}, m_{X_{i2}}, m_{Y_i}) : i = 1, \ldots, 30\} \) of the variables given in Table 1 is as follows:
Table 2. The results of the performance measures for Data in Table 1

<table>
<thead>
<tr>
<th></th>
<th>Diamond’s Method</th>
<th>Chachi and Taheri Method</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>$D(Y, \hat{Y})$</td>
<td>1.7442</td>
<td>1.9495</td>
<td>1.7372</td>
</tr>
<tr>
<td>$I(Y, \hat{Y})$</td>
<td>0.2668</td>
<td>0.3785</td>
<td>0.3798</td>
</tr>
</tbody>
</table>

$m_{\hat{Y}_i} = 0.0705 + 0.4181m_{X_{i1}} + 0.5005m_{X_{i2}}$

Table 1 shows the estimated value $\hat{Y}_i$ using the LAD method and the discriminant analysis. The results of the performance measures $D(Y, \hat{Y})$ and $I(Y, \hat{Y})$ for data given in Table 1 are in Table 2.

Table 2 shows that the proposed method has the smallest sum of areas for residual and the largest overlapping areas between observations and forecasts.

5. Conclusion

In this paper, we have proposed a novel algorithm to estimate the fuzzy regression model by separating the center and spread without using the same function. We used the fuzzy partition to classify spreads of the given data into several classes. Then, we combined the LAD estimation method for the center of the dependent variable with discriminant analysis for the spread of the dependent variable to predict the value of the dependent variable. The results of numerical example showed that the proposed estimation algorithm have better performance than the existing estimation methods using the same function for center and spread. This means that the proposed estimation algorithm can provide more efficient estimation results when the number of spreads of dependent variable is less than the sample size or the number of centers of dependent variable.

Acknowledgment. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No.2017R1C1B1005069, No. 2017R1D1A1B03029559)

References