LEAST ABSOLUTE DEVIATION ESTIMATOR IN FUZZY REGRESSION

KYUNG JOONG KIM*, DONG HO KIM AND SEUNG HOE CHOI

Abstract. In this paper we consider a fuzzy least absolute deviation method in order to construct fuzzy linear regression model with fuzzy input and fuzzy output. We also consider two numerical examples to evaluate an effectiveness of the fuzzy least absolute deviation method and the fuzzy least squares method.

AMS Mathematics Subject Classification : 62J05, 65K10

Key words and phrase : Fuzzy regression model, fuzzy least squares method, fuzzy least absolute deviation method.

1. Introduction

In this paper we consider the fuzzy linear regression model when the variables themselves are fuzzy;

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \quad i = 1, 2, \cdots, n. \]

where \( Y_i \) and \( X_{ij}, j = 1, \cdots, p \) are the fuzzy number and regression coefficient \( \beta_j \) is a crisp (non-fuzzy) number.

In the fuzzy regression analysis which was initially developed by Tanaka et al.[10], a deviation between an observed value and an estimated one is assumed to be due to some indefiniteness of the system structure or due to fuzziness of regression coefficients. The goal of fuzzy regression analysis is to find a regression model that minimizes the total difference between the estimated fuzzy outputs and the observed fuzzy outputs.

Savic and Pedrycz[7] have proposed two consecutive steps by combining the method of least squares and the minimum fuzziness criterion so as to construct a fuzzy regression model having minimal difference. The first step employs the
ordinary least squares method in order to find the fuzzy center values of fuzzy regression coefficients. The second step takes advantage of the minimum fuzziness criterion to find the fuzzy widths of fuzzy regression coefficients. Moreover, Diamond[3] has introduced the fuzzy least squares regression analysis, while Kim and Bishu[6] suggested another approach based on the criterion of minimizing the difference between the membership values of the observed and estimated fuzzy response.

Most studies on fuzzy regression analysis [6,7,8,9,10] have emphasized the fuzziness of the response variable alone. Only a few of these authors [3,5], have investigated the fuzzy regression model when both response and explanatory variables are fuzzy numbers. A common characteristic of the studies on the fuzzy regression model with the crisp independent variable is that the regression coefficients were treated as fuzzy numbers. So, in estimation of fuzzy coefficient the spread of the estimated response variable becomes wider as the magnitude of the explanatory variables increases, even though the spreads of the observed responses are roughly constant or decreasing.

To offset complementary this drawback, Kao and Chyu[5] considered the fuzzy regression model with fuzzy dependent variables, fuzzy independent variables and crisp regression coefficients. Since the regression coefficients are crisp, the problem that the spreads of the estimated fuzzy number are increasing can be avoided.

On the other hand, Kao and Chyu, Diamond, and many other authors have made use of the fuzzy least squares method so as to construct a fuzzy regression model having fuzzy input-output data. However, since the least squares method is sensitive to outlier, it is a very poor estimator for many non-Gaussian forms. So, in the regression analysis the search for robust procedure has considerable interest in statistical methods based on Least Absolute Deviation (LAD) using the $L_1$-norm rather than Least Squares (LS) using the $L_2$-norm. Choi[2] pointed that the LAD estimators based on a sample median are more efficient than the least squares estimators based on sample mean in the case of heavy-tailed or skewed error distribution. Therefore, we need to consider the LAD estimator in fuzzy regression model in order to minimize the total difference between the estimated fuzzy outputs and the observed fuzzy outputs.

We propose a two-stage method to construct the fuzzy linear regression model in section 2. The first stage is to defuzzify fuzzy observations to crisp values and to apply the LAD estimators so as to calculate regression coefficients. The second stage is then to determine the fuzzy error term which will minimize the total difference between the observed response and estimated ones. In the following section we compare the total error for Fuzzy Least Squares (FLS) method and Fuzzy Least Absolute Deviation (FLAD) method using two numerical examples.

2. Fuzzy Least Absolute Deviation Method
In the first place, we will deal with the basic concepts of fuzzy set and fuzzy number in order to construct the fuzzy regression model. For this purpose, let $X_{ij}$ and $Y_i$ be fuzzy numbers having the membership functions $\mu_{X_{ij}}$ and $\mu_{Y_i}$, respectively.

**Definition 2.1.** Let $X$ be a collection of objects $x$. A fuzzy subset $A$ in $X$ is defined as a set of ordered pairs:

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

where $\mu_A$ is a function from $X$ to the closed interval $[0, 1]$. Here, $\mu_A$ is called the **membership function or grade function** and $\mu_A(x)$ is called the **grade of $x$ in $A$**.

**Definition 2.2.** A fuzzy number $A$ is a fuzzy subset of the real line $\mathbb{R}$ with membership function satisfying the following criteria:

(i) $\alpha$-level set of $A$, $A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}$, is a closed interval.

(ii) There exists exactly one $x_0 \in \mathbb{R}$ with $\mu_A(x_0) = 1$.

(iii) $\mu_A(\lambda x_1 + (1 - \lambda)x_2) \geq \min \{\mu_A(x_1), \mu_A(x_2)\}$, for every $x_1, x_2 \in X$ and $\lambda \in [0, 1]$.

(iv) $\mu_A(x)$ is piecewise continuous.

For the special fuzzy number, let $L$ and $R$ be non-incremental in the interval $[0, \infty)$ and satisfy

$L(0) = R(0) = 1, \quad L(1) = R(1) = 0, \quad L(x) = L(-x) \quad \text{and} \quad R(x) = R(-x)$.

**Definition 2.3.** A fuzzy number $A$ is said to be the LR-fuzzy number with the following membership function:

$$\mu_A(x) = \begin{cases} 
L\left(\frac{m-x}{l_A}\right) & \text{if } m - l_A \leq x < m, \\
R\left(\frac{x-m}{r_A}\right) & \text{if } m \leq x < m + r_A, \\
0 & \text{otherwise,}
\end{cases}$$

for $x \in \mathbb{R}$, where $m \in \mathbb{R}$, $l_A \geq 0$ and $r_A \geq 0$.

The LR-fuzzy number $A$ can be also represented as $A = (l_A, m, r_A)_{LR}$ with $m$ as the center or mode and $l_A$ and $r_A$ as the left and right spread of the fuzzy number $A$, respectively. Specially, every triangular fuzzy number $A = (l_A, m, r_A)_T$ is represented by the shape functions $L(x) = R(x) = 1 - x$. Dubios
and Prade[2] defined the elementary operations for two LR-fuzzy numbers and proved the following theorem.

**Theorem 2.1.** Let $A = (l_A, m_A, r_A)_{LR}$ and $B = (l_B, m_B, r_B)_{LR}$ be fuzzy numbers in $X$.

(i) $(l_A, m_A, r_A)_{LR} + (l_B, m_B, r_B)_{LR} = (l_A + l_B, m_A + m_B, r_A + r_B)_{LR}$.

(ii) $-(l_A, m_A, r_A)_{LR} = (r_A, -m_A, l_A)_{LR}$.

(iii) $(l_A, m_A, r_A)_{LR} - (m_B, l_B, r_B)_{LR} = (l_A + r_B, m_A - m_B, r_A + l_B)_{LR}$.

Since it is to easy to extend triangular fuzzy numbers to LR-fuzzy numbers, in this paper we shall always assume that all observations are triangular fuzzy numbers defined as $X_{ij} = (l_{x_{ij}}, x_{ij}, r_{x_{ij}})$ and $Y_i = (l_{y_i}, y_i, r_{y_i})$.

Now, we consider a two-stage method in order to construct the fuzzy regression model. The first stage is to de-fuzzify the fuzzy observations to crisp (non-fuzzy) values and then to apply the LAD method to estimate the regression coefficients.

Let $x_{ijc}$ and $y_{ic}$ be the de-fuzzified values of $X_{ij}$ and $Y_i$, respectively. The centroid method calculates $x_{ijc}$ and $y_{ic}$ via the following formulae:

$$x_{ijc} = \frac{\int_{-\infty}^{\infty} x \mu_{X_{ij}}(x) \, dx}{\int_{-\infty}^{\infty} \mu_{X_{ij}}(x) \, dx} = \frac{1}{3} (l_{x_{ij}} + x_{ij} + r_{x_{ij}}),$$

$$y_{ic} = \frac{\int_{-\infty}^{\infty} y \mu_{Y_i}(y) \, dy}{\int_{-\infty}^{\infty} \mu_{Y_i}(y) \, dy} = \frac{1}{3} (l_{y_i} + y_i + r_{y_i}).$$

Since the LS method is sensitive to outlier, we consider the LAD method of obtaining a fuzzy regression model more effective than fuzzy regression model using LS method. The LAD estimator $\hat{\beta}_i$ for the regression parameters $\beta_i$, based on the set of crisp values $(x_{ijc}, y_{ic})$, is defined as the value that minimizes the absolute deviation

$$\sum_{i=1}^{n} \left| y_{ic} - \sum_{j=1}^{n} \beta_j x_{ijc} \right|. \quad (1)$$

Next, we have to consider the estimator for the difference between the observed fuzzy response $Y_i$ and the fuzzy number $\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$.

For this, let $\hat{E} = (-l, 0, r)$ be an estimator for

$$E_i = Y_i - \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}.$$
Then, we can think the estimated fuzzy response as
\[ \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip} + \hat{E}. \]

As an aim of fuzzy regression analysis minimizes the fuzziness in the predicted value of the dependent variable, therefore, in order to minimize the spread of the fuzzy error term \( \hat{E} = (-l, 0, r) \) we shall use the value to determine the minimal difference of membership values between the observed and estimated fuzzy numbers as a measure. The difference between their membership values, presented by Kim and Bishu[6], is the value area which can be calculated as
\[
D_i = \int_{S_{Y_i} \cup S_{\hat{Y}_i}} |\mu_{Y_i}(y) - \mu_{\hat{Y}_i}(y)| dy,
\]
where \( S_{Y_i} \) and \( S_{\hat{Y}_i} \) are the supports of \( \mu_{Y_i} \) and \( \mu_{\hat{Y}_i} \), respectively. Since smaller values of \( D_i \) indicate that the fuzzy regression model fits the data better, we have to find the estimates for \( l \) and \( r \) which will minimize the total difference between the observed and estimated responses. Let \( l_{min} \) and \( r_{min} \) be the smallest left and right spreads of the observed responses, respectively. The problem of determining the estimator of the fuzzy error term \( E \) can be formulated as the following mathematical program:
\[
\min \sum_{i=1}^{n} \int_{S_{Y_i} \cup S_{\hat{Y}_i}} |\mu_{Y_i}(y) - \mu_{\hat{Y}_i}(y)| dy
\]
such that \( \hat{Y}_i = \hat{a}_1 X_{i1} + \cdots + \hat{a}_p X_{ip} + (-l, 0, r), \)
\[
(m_{\hat{y}_i} - l_{\hat{y}_i}) \geq l_{min}, \quad (r_{\hat{y}_i} - m_{\hat{y}_i}) \geq r_{min}.
\]

3. Numerical examples

In this section, we examine two numerical examples to investigate the performance of the LAD estimators in fuzzy linear regression. The first example has fuzzy observations only for the response variable and the second example has fuzzy observations for both the explanatory and response variables.

Example 1. Tanaka et al.[10] designed an example to illustrate their regression model for dealing with the problem of crisp independent variable and fuzzy dependent variable. By applying the method of this study, the fuzzy regression models constructed are summarized as following:
(i) $Y_{FMLS} = (3.11, 4.95, 6.84) + (1.55, 1.71, 1.82)x.$
(ii) $Y_{KC} = 4.95 + 1.71x + (-3.01, 0, 1.80).$
(iii) $Y_{FLAD} = (3.46, 5.56, 7.71) + (1.43, 1.49, 1.54)x.$

The right half of Table 3.1 shows the errors of the five observations for the three models. The total error of the FLAD method is 9.147 which is obviously better than the total errors of 10.026 calculated from the FMLS method and 9.678 calculated from the Kao-Chyu method.

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Response variable</th>
<th>FMLS</th>
<th>Kao-Chyu</th>
<th>FLAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(6.2, 8.0, 9.8)</td>
<td>2.207</td>
<td>2.7826</td>
<td>1.6850</td>
</tr>
<tr>
<td>2</td>
<td>(4.2, 6.4, 8.6)</td>
<td>3.025</td>
<td>2.5951</td>
<td>3.2818</td>
</tr>
<tr>
<td>3</td>
<td>(6.9, 9.5, 12.1)</td>
<td>1.042</td>
<td>0.5559</td>
<td>0.9949</td>
</tr>
<tr>
<td>4</td>
<td>(10.9, 13.5, 16.1)</td>
<td>2.902</td>
<td>3.3569</td>
<td>3.1849</td>
</tr>
<tr>
<td>5</td>
<td>(10.6, 13.0, 15.4)</td>
<td>0.850</td>
<td>0.3874</td>
<td>0.0000</td>
</tr>
<tr>
<td>Total error</td>
<td></td>
<td>10.026</td>
<td>9.6779</td>
<td>9.1467</td>
</tr>
</tbody>
</table>

Table 3.1 : Numerical data and the estimation errors for Example 1

**Example 2.** Consider the example designed by Sakawa and Yano[8]. All observations for the response and explanatory variables are triangular fuzzy numbers as shown in the left half of Table 3.2. By applying the method of this study to the data in Table 3.2, we obtain the following:

(i) $Y_{KC} = 3.5724 + 0.5193X + (-0.24, 0, 0.24).$
(ii) $Y_{Diamond} = (3.2636, 3.5632, 3.8628) + 0.5206X.$
(iii) $Y_{FLAD} = 3.94 + 0.44X + (-0.28, 0, 0.28).$

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Response variable</th>
<th>Errors in estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Diamond</td>
<td>Kao-Chyu</td>
</tr>
<tr>
<td>(1.5, 2.0, 2.5)</td>
<td>(3.5, 4.0, 4.5)</td>
<td>0.633</td>
</tr>
<tr>
<td>(3.0, 3.5, 4.0)</td>
<td>(5.0, 5.5, 6.0)</td>
<td>0.453</td>
</tr>
<tr>
<td>(4.5, 5.5, 6.5)</td>
<td>(6.5, 7.5, 8.5)</td>
<td>1.613</td>
</tr>
<tr>
<td>(6.5, 7.0, 7.5)</td>
<td>(6.0, 6.5, 7.0)</td>
<td>1.165</td>
</tr>
<tr>
<td>(8.0, 8.5, 9.0)</td>
<td>(8.0, 8.5, 9.0)</td>
<td>0.770</td>
</tr>
<tr>
<td>(9.5, 10.5, 11.5)</td>
<td>(7.0, 8.0, 9.0)</td>
<td>1.977</td>
</tr>
<tr>
<td>(10.5, 11.0, 11.5)</td>
<td>(10.0, 10.5, 11.0)</td>
<td>1.368</td>
</tr>
<tr>
<td>(12.0, 12.5, 13.0)</td>
<td>(9.0, 9.5, 10.0)</td>
<td>1.452</td>
</tr>
<tr>
<td>Total error</td>
<td></td>
<td>9.431</td>
</tr>
</tbody>
</table>

Table 3.2 : Numerical data and the estimation errors for Example 2
The right half of Table 3.2 shows the results. For the Diamond method, the Kao-Chyu method, and the LAD method, the total errors are 9.431, 7.4881, and 6.2362, respectively, in favor of the LAD method.

4. Conclusion

The estimated fuzzy numbers from the new method on the basis of the LAD estimator resulted in smaller errors than those from the other methods based on the LS estimator in both examples. So, like as the ordinary regression analysis, it is expected that for the cases containing outliers the LAD method using the least absolute deviation method is more effective than the LS method using the least squares method in the fuzzy regression analysis.

REFERENCES


Kyung Joong Kim obtained a Ph.D on numerical analysis from Yonsei University, Korea. Since the beginning of 2004, he has a permanent position at Hankuk Aviation University. His principal current research interests are cubature rules for multivariate integrals, exponential-fitting quadrature rules and interpolation theories.

Department of General Studies, Hankuk Aviation University, Koyang 412-791, Korea.
e-mail: kj_kim@hau.ac.kr

Dong Ho Kim obtained a Ph.D on numerical analysis from Yonsei University, Korea in 2001. Since the beginning of 2003, he is a research professor of Natural Science Research
Institute at Yonsei University. His principal current research interest is a posteriori error estimator for mixed finite element methods for partial differential equations.

**e-mail:** dhkimm@yonsei.ac.kr

**Seung Hoe Choi** obtained a Ph.D from Yonsei University. Since 1996 he has been at the Hankuk Aviation University. His research interests focus on the estimation and the test of hypothesis in nonlinear regression model and the fuzzy statistics.

**e-mail:** shchoi@hau.ac.kr