CENSORED FUZZY REGRESSION MODEL

SEUNG HOE CHOI AND KYUNG JOONG KIM

Abstract. Various methods have been studied to construct a fuzzy regression model in order to present a fuzzy relation between a dependent variable and an independent variable. However, in the fuzzy regression analysis the value of the center point of estimated fuzzy output may be either greater than the value of the right endpoint or smaller than the value of the left endpoint. In the case, we cannot predict the fuzzy output properly.

This paper presents sufficient conditions to construct the fuzzy regression model using several methods investigated by some authors and then introduces the censored fuzzy regression model using the censored samples to manipulate the problem of crossing of the center and the end points of the estimated fuzzy number. Examples show that the censored fuzzy regression model is an extension of the fuzzy regression model and also it improves the problem of crossing.

1. Introduction

The equation in which the relation between the dependent and independent variables is expressed into a mathematical model, is called a regression model, being expressed, as follows:

\[ y_i = f(x_i, \theta_0) + \epsilon_i, \quad i = 1, 2, \ldots, n, \]

where \( x_i \) is an independent variable, \( y_i \) is a dependent variable, \( \epsilon_i \) is an error term, and \( \theta_0 \) is an unknown parameter.

It seems that in regression analysis a difference happens between the actual value of a response value and its estimator owing to errors arising from measuring the real value of the respond variable. Therefore, the value minimizing the difference between the real value and its estimator is used as the estimator of regression parameters. However, in the actual
state of things, all reactive variable values cannot be precisely measured. There exist ambiguity and vagueness in cases such as human subjective appraisal or judgment. In order to inquire into those regression models that have as their reactive variables such materials as include ambiguity and vagueness like linguistic variables, Tanaka et al. [15] proposed the following fuzzy regression model for the first time:

\[
Y_i = A_0 + A_1 x_{i1} + \cdots + A_p x_{ip}, \quad i = 1, 2, \ldots, n,
\]

where \( A_i \) and \( Y_i \) are fuzzy numbers and \( x_{ij} \) is a positive non-fuzzy number. And \( \mu_{Y_i} \) and \( \mu_{A_i} \) are membership function of \( Y_i \) and \( A_i \), respectively.

Several methods were presented to estimate fuzzy regression models. Developed in Tanaka [13, 14, 15, 16], Savic and Pedrycz [12], Kao and Chyu [8] and Chang and Ayyub [2] were numerical methods that minimize the fuzziness of the response variables, while studied in Diamond [5, 6, 7], Ming et al. [10], Chang [1], and Chen [4] were statistical methods that minimize the difference between the estimated and observed outputs. The fuzzy least squares method applies the least squares method used in ordinary regression models to fuzzy regression models and it estimates fuzzy models by first defining the distance between two fuzzy numbers and then generalizing to the distance in fuzzy regression models. Unlike ordinary regression models, the fuzzy output estimated using the least squares method must be fuzzy numbers. However, as in Examples 2.3, the value of the center point of estimated fuzzy output may be either smaller than the value of the endpoint on the left or greater than that on the right. That is, the example given in Section 2 is not fit to be explained with the fuzzy regression model. When analysis is inappropriate with a fuzzy regression model as in Example 2.3, such sufficient condition should be taken into account as can be described with the fuzzy regression model. Furthermore, in case that there arises a phenomenon in which the center point of the estimated fuzzy output is unable to satisfy the given sufficient condition because of cross endpoints on the right or the left side, new methods are required to solve such a problem.

This study presents the sufficient condition for the fuzzy regression model and introduces a model using some techniques of censoring to work out the problem that the center and end points in estimated fuzzy regression numbers cross each other. Moreover, it identifies the fact through examples that the fuzzy regression analysis using censored sample is an extension of fuzzy regression analysis and at the same time serves as a method to handle with the problem that the center point and the end points of the estimated fuzzy output cross.
2. Fuzzy regression model

We define \( A_j = (a_j, l_j, r_j) \) and \( Y_i = (y_i, l_{yi}, r_{yi}) \) as LR-fuzzy numbers to suggest the sufficient condition for the fuzzy regression model. Then, an \( \alpha \)-cut of fuzzy number \( A_j \), written \( A_j(\alpha) \) is defined as

\[
A_j(\alpha) = \{ x | \mu_{A_j} \geq \alpha \} = [a_j - l_j L(\alpha)^{-1}, a_j + r_j R(\alpha)^{-1}],
\]

where \( L \) and \( R \) are reference functions on the right and the left side, respectively. A triangular fuzzy number is represented by the shape function \( L(x) = R(x) = 1 - x \). In particular, \( \{ x | \mu_{A_j} > 0 \} \) is called the support of fuzzy number \( A_j \) and \( a_j - l_j \) and \( r_j - a_j \) are called a left spread and a right spread of fuzzy number \( A_j \), respectively. See Zimmerman [18]. The difference between two fuzzy numbers \( A_i \) and \( A_j \) equals

\[
[A_i(\alpha) - \bar{A}_j(\alpha), \bar{A}_i(\alpha) - A_j(\alpha)],
\]

where \( A_j(\alpha) = a_j - l_j L(\alpha)^{-1}, \bar{A}_j(\alpha) = a_j + r_j R(\alpha)^{-1} \) and \( 0 \leq \alpha \leq 1 \).

In order to construct the fuzzy regression model many scholars defined the distance between two fuzzy numbers defined the distance between two fuzzy numbers \( A_i \) and \( A_j \).

Diamond was defined the distance between two fuzzy numbers as

\[
D_d(A_i, A_j) = (A_i(1) - \bar{A}_j(1))^2 + (A_i(0) - A_j(0))^2 + (\bar{A}_i(0) - \bar{A}_j(0))^2,
\]

that of Ming et al. as

\[
D_m(A_i, A_j) = \left[ \int_0^1 \{(A_i(\alpha) - A_j(\alpha))^2 + (\bar{A}_i(\alpha) - \bar{A}_j(\alpha))^2\}d\alpha \right]^\frac{1}{2},
\]

that of Chang as

\[
D_{ch}(A_i, A_j) = \frac{\int_0^1 \{(A_i(\alpha) - A_j(\alpha))^2 + (\bar{A}_i(\alpha) - \bar{A}_j(\alpha))^2\}d\alpha}{\int_0^1 d\alpha},
\]

and that of Chen as

\[
D_c(A_i, A_j) = \frac{1}{2} \int_0^1 \{(A_i(\alpha) - A_j(\alpha)) + (\bar{A}_i(\alpha) - \bar{A}_j(\alpha))\}d\alpha.
\]

On the other hand, in estimating the fuzzy regression coefficients we can use the distance between two fuzzy numbers which is derive from the generalized Hausdorff distance as

\[
(\bar{A}_i(\alpha) - \bar{A}_j(\alpha))^2 + (\bar{A}_i(\alpha) - \bar{A}_j(\alpha))^2.
\]
However, in this case solution may not exist or the uniqueness of the solution cannot be guaranteed in a normal equation. Therefore, another distance between the fuzzy numbers needs to be defined to use the least squares method in the fuzzy regression analysis.

In this paper we construct the fuzzy regression model using the difference between fuzzy numbers as

\[ D_f(A_i, A_j) = ((\bar{A}_i(1) - \bar{A}_j(1))^2 + (\bar{A}_{\alpha}(\alpha) - \bar{A}_j(\alpha))^2 + (\bar{\bar{A}}_i(\alpha) - \bar{\bar{A}}_j(\alpha))^2, \]

which is a transformed equation of equation (2.1), to find the unique solution in a normal equation. Especially, the distance \( D_d(A_i, A_j) \) is easily seen to be a special case of \( D_f(A_i, A_j) \) when \( \alpha = 0 \).

As the results of the fuzzy simple regression model with two parameters can be naturally extend to the fuzzy model with several fuzzy regression coefficients, in this paper we consider the fuzzy simple regression model.

The least squares fuzzy estimators of the fuzzy regression coefficient, denoted by \( \hat{A} \), are the value which minimize the objective function

\[ Q_n(a, l, r) = \sum_{i=1}^{n} D_f(Y_i, A_0 + A_ix_i). \]

And we should consider the following conditions to provide for such a case not to occur in which the center point and the end points of the estimated fuzzy output cross.

Throughout this section, we want to make the following Conditions in the fuzzy regression model (1.1):

**CONDITION A.**

\[ A_1 : \sum_{i=1}^{n} (x_i - \bar{x}_n)(l_{yi} - \bar{y}_n) > 0. \]

\[ A_2 : \sum_{i=1}^{n} (x_i - \bar{x}_n)(r_{yi} - \bar{y}_n) > 0. \]

\[ A_3 : \sum_{i=1}^{n} (x_i - \bar{x}_n)(\bar{x}_n l_{yi} - x_i \bar{y}_n) > 0. \]

\[ A_4 : \sum_{i=1}^{n} (x_i - \bar{x}_n)(\bar{x}_n r_{yi} - x_i \bar{y}_n) > 0. \]

In the above conditions, \( \bar{x}_n, \bar{r}_n, \) and \( \bar{\bar{y}}_n \) signify the standard mean of input and of the right endpoint and the left endpoint in fuzzy output, respectively.
We present the sufficient conditions for the fuzzy regression model in the following theorem. For the proof, see Diamond [5].

**Theorem 2.1.** For the fuzzy regression model (1.1), suppose that Condition $A$ is fulfilled. Then the left spread and the right spread of the estimated fuzzy output are always greater than zero. That is, estimated fuzzy output satisfies the condition for fuzzy numbers.

As before mentioned, many scholars defined the distance between two fuzzy numbers, and estimated regression coefficients by applying the defined distance to the fuzzy regression model. We denote the value that minimizes objective function $\sum_{i=1}^{n} D_d(Y_i, A_0 + A_i x_i)$ using the fuzzy distance $D_d(Y_i, A_0 + A_i x_i)$ suggested by Diamond as $\hat{A}^d$. Similarly to $\hat{A}^d$, we denote the estimator suggested by Ming as $\hat{A}^m$, that by Chang as $\hat{A}^{ch}$, and that by Chen as $\hat{A}^c$.

The following theorem shows that fuzzy regression coefficients estimated by mutually different methods are the same.

**Theorem 2.2.** Suppose that the fuzzy regression model (1.1) satisfies Assumptions given in Theorem 2.1. Then, estimators $\hat{A}^d$, estimators $\hat{A}^m$, estimators $\hat{A}^{ch}$, and estimators $\hat{A}^c$ are all the same as the least squares fuzzy estimators $\hat{A}$ suggested in Theorem 2.1.

**Proof.** It suffices to show that the mutually different fuzzy estimators have same normal equation. For this, the derivatives of the function $Q_n(a, l, r)$ with respect to the six unknowns need to be derived, set to zero, and then solved for the six unknowns. From the derivatives of the function $Q_n(a, l, r)$ with respect to the unknowns $a_0$, $a_1$, $l_0$, $l_1$, $r_0$, $r_1$ and rearranging the terms of equation, we obtain the following six sets of normal equations which are well-known equations:

\[
na_0 + \sum_{i=1}^{n} x_i a_1 = \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} x_i a_0 + \sum_{i=1}^{n} x_i^2 a_1 = \sum_{i=1}^{n} x_i y_i, \\
nl_0 + \sum_{i=1}^{n} x_i l_1 = \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} x_i l_0 + \sum_{i=1}^{n} x_i^2 l_1 = \sum_{i=1}^{n} x_i y_i, \\
nr_0 + \sum_{i=1}^{n} x_i r_1 = \sum_{i=1}^{n} y_i, \quad \sum_{i=1}^{n} x_i r_0 + \sum_{i=1}^{n} x_i^2 r_1 = \sum_{i=1}^{n} x_i y_i.
\]

By simple calculation, we know that the sets of the normal equations for the objective functions $\sum_{i=1}^{n} D_d(Y_i, A_0 + A_i x_i)$, $\sum_{i=1}^{n} D_m(Y_i, A_0 + A_i x_i)$,
\[
\sum_{i=1}^{n} D_{ch}(Y_i, A_0 + A_i x_i), \text{ and } \sum_{i=1}^{n} D_c(Y_i, A_0 + A_i x_i) \text{ are all equal to the form of the normal equations given from above. So, the four fuzzy estimators are all identical.}
\]

Example 2.3 Table 2.1 shows the data suggested by Chang [3] and analyzed with Pierpaolo [11].

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Response variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>14.8</td>
<td>(59.7, 13.7, 18.3)</td>
</tr>
<tr>
<td>18.0</td>
<td>(63.5, 17.5, 20.5)</td>
</tr>
<tr>
<td>22.9</td>
<td>(66.8, 18.8, 19.2)</td>
</tr>
<tr>
<td>31.5</td>
<td>(70, 26, 26)</td>
</tr>
<tr>
<td>50.3</td>
<td>(67, 17, 21)</td>
</tr>
<tr>
<td>126.0</td>
<td>(62.2, 8.2, 11.8)</td>
</tr>
</tbody>
</table>

The fuzzy regression model using non-fuzzy input \(x_i\) and fuzzy output \(Y_i\) in Table 2.1 results in

\[
\hat{Y}_i = (44.716 + 0.075x_i, 65.478 - 0.014x_i, 88.341 - 0.091x_i).
\]

The left endpoint, 62.716, of the fuzzy output value for input value \(x = 240\) is larger than its center, 62.118 and the right endpoint, 59.949, of the fuzzy output value for input value \(x = 312\) is smaller than its center, 61.11. That is, the estimated fuzzy output \(\hat{Y}_i\) cannot be a fuzzy number because the center point is larger than the right endpoint or smaller than the left endpoint. The reason may be attributed to the fact that the values of \(\sum_{i=1}^{n} (x_i - \bar{x}_n)(l_{y_i} - \bar{l}_{y_n})\) and \(\sum_{i=1}^{n} (x_i - \bar{x}_n)(r_{y_i} - \bar{r}_{y_n})\) suggested by Condition A are -788 and -687, respectively, being smaller than 0.

As in Example 2.3, there exists a problem that the estimated fuzzy outputs do not satisfy the assumption for the fuzzy regression model. So, in this case we have to introduce new method to construct the fuzzy regression model.
3. Censored fuzzy regression model

As identified in Section 2, in order to estimate fuzzy regression coefficients using the least squares method that play an important role in the regression model, certain conditions must be satisfied. Therefore, new methods are required to be studied to estimate the data that are not able to satisfy the sufficient condition suggested in the previous section with the fuzzy regression model. In the fuzzy regression model, the center point of the estimated fuzzy output must be larger than the left endpoint and smaller than the right endpoint. That is to say, the right endpoint of the estimated fuzzy output must be at least larger than the center of the estimated fuzzy output but the left endpoint of the estimated fuzzy output may not be larger than the center of the estimated fuzzy output. In order to employ this nature, we apply censored data used in statistics to fuzzy regression analysis.

Tobin [17] applied the concept of censoring to regression analysis for the first time, see Kmenta [9]. He inquired into the relations between expenditure and income, recognized that expenditure in the home may be more than zero, utilizing censored data to estimate the regression model.

We call the fuzzy regression model using censored data a censored fuzzy regression model and we consider a two-stage procedure to estimate the fuzzy regression model using censored data, as follows:

Step I. We use the objective functions to estimate the center of fuzzy regression coefficient $A_j$ using the center of fuzzy numbers $Y_i$ and $A_0 + A_1 x_i$, as follows:

$$\sum_{i=1}^{n} \rho(y_i - a_0 - a_1 x_i) = \min!,$$

where $\rho$ is a continuous function on $R$.

In particular, the estimators $\hat{a}_i$ becomes the least square estimators in case of $\rho(x) = x^2$ and when $\rho(x) = |x|$, it becomes the least absolute deviation estimators.

Step II. We put $l_{\min}$ and $r_{\min}$ as the least left and right spreads among observed fuzzy outputs, respectively. We can obtain the right and the left spreads of the predicted fuzzy output using the following equation with the concept of censoring:

$$\sum_{i=1}^{n} \left( r_{y_i} - \max \{ \hat{a}_0 + \hat{a}_1 x_i + r, \hat{a}_0 + r_0 + (\hat{a}_1 + r_1) x_i \} \right)^2 = \min!$$
and
\[ \sum_{i=1}^{n} (l_{y_i} - \min \{ \hat{a}_0 + \hat{a}_1 x_i - l, \hat{a}_0 + l_0 + (\hat{a}_1 + l_1) x_i \})^2 = \min! \]
subject to
\[ r > \min \{ r_{y_1} - y_1, \ldots, r_{y_n} - y_n \} \text{ and } l > \min \{ y_1 - l_{y_1}, \ldots, y_n - l_{y_n} \}. \]
If we suppose that \( l > 0 \) and \( r > 0 \) in the above inequality, a case may occur in which the center and endpoint of the estimated fuzzy output may be nearly identical. So, we used the inequality to prevent the case in which the right and the left spreads approximate zero and to use the information of the given data.

Furthermore if the given data satisfy Condition A given in Section 2, when we put \( l \) and \( r \) suggested in Step II as 0 and 0, respectively, the fuzzy regression model can be estimated by the same method as that of the estimators made mention in Section 2.

In the following example we find the censored fuzzy regression model for the case satisfying the sufficient condition for the fuzzy regression model.

**Example 3.1.** The following data is an example given by Tanaka et al. [15], and the fuzzy regression model estimated using censored data is as follows:

\[ \hat{f}_{r}(x) = \begin{cases} 9.05 + 0.71x & \text{if } x \leq 2 \\ 7.05 + 1.71x & \text{if } x > 2, \end{cases} \]

\[ \hat{f}_{c}(x) = 4.45 + 1.71x, \]

and

\[ \hat{f}_{l}(x) = \begin{cases} 2.65 + 1.71x & \text{if } x \leq 4 \\ 5.21 + 1.07xx & \text{if } x > 4, \end{cases} \]

where \( \hat{f}_r, \hat{f}_c, \) and \( \hat{f}_l \) denote the function for right endpoint, the function for center point, and the function for left endpoint, respectively.

However, in the data presented in Table 3.1, Condition \( C_1 \) and \( C_2 \) are positive number and Condition \( C_3 \) and \( C_4 \) are negative number. As the data satisfy the conditions given in Theorem 2.1, the fuzzy regression model can be estimated by putting \( l = r = 0 \). Table 3.1 shows the result estimated using Step I and Step II in case of \( l = r = 0 \).
Example 3.1 implies that the censored fuzzy regression model extends the fuzzy regression analysis given by many authors.

4. Numerical example

In order to use, in fuzzy regression analysis, the data in which inputs are non-fuzzy numbers and outputs fuzzy numbers, certain conditions must be satisfied. However, even in the case that the sufficient condition is not satisfied, the fuzzy regression model can be estimated using censored data.

In this section, we first identify through the following example that in the censored fuzzy regression model the center and end points of the estimated fuzzy outputs do not cross at all and then compare two methods to estimate the center and end points of fuzzy regression coefficients with the least squares method and the least absolute deviation method.

Example 4.1. The data on the left of Table 4.1 was introduced by Chang et al. [3].

As in Example 2.3, it does not satisfy Condition $A_1$ and $A_2$, there occurs the phenomenon that the center and end points of the estimated fuzzy outputs cross each other in cases when the input is either 128 or 148. In order to solve the problem of crossing, the result estimated using the censored fuzzy regression model is as follows:

\[
\hat{f}_r(x) = \begin{cases} 
78.27145 + 0.1535x & \text{if } x < 126 \\
55.3673 + 0.3411x & \text{if } x \geq 126,
\end{cases}
\]

and

\[
\hat{f}_c(x) = 54.6672 + 0.3411x,
\]

\[
\hat{f}_l(x) = \begin{cases} 
15.6144 + 1.5675x & \text{if } x < 25.2 \\
35.9349 + 0.7611x & \text{if } 25.2 \leq x < 31.5 \\
49.1672 + 0.3411x & \text{if } x \geq 31.5.
\end{cases}
\]
Table 4.1: Numerical data and the estimation errors for Example 4.1

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Response variable</th>
<th>Errors in estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LS</td>
</tr>
<tr>
<td>7.4</td>
<td>(39.3, 13.8, 28.7)</td>
<td>30.3125</td>
</tr>
<tr>
<td>9.0</td>
<td>(40.5, 15.0, 31.5)</td>
<td>28.4118</td>
</tr>
<tr>
<td>11.4</td>
<td>(50.8, 18.6, 31.2)</td>
<td>10.3179</td>
</tr>
<tr>
<td>14.8</td>
<td>(59.7, 13.7, 18.3)</td>
<td>10.4149</td>
</tr>
<tr>
<td>15.7</td>
<td>(56.0, 18.0, 26.0)</td>
<td>6.8647</td>
</tr>
<tr>
<td>18.0</td>
<td>(70.5, 15.5, 13.5)</td>
<td>24.5294</td>
</tr>
<tr>
<td>22.9</td>
<td>(66.8, 18.8, 19.2)</td>
<td>12.7459</td>
</tr>
<tr>
<td>25.2</td>
<td>(66.3, 16.3, 23.7)</td>
<td>16.7136</td>
</tr>
<tr>
<td>31.5</td>
<td>(78.0, 20.0, 8.0)</td>
<td>18.0930</td>
</tr>
<tr>
<td>50.3</td>
<td>(80.0, 15.0, 8.0)</td>
<td>12.2070</td>
</tr>
<tr>
<td>62.9</td>
<td>(87.3, 12.0, 0.7)</td>
<td>16.6322</td>
</tr>
<tr>
<td>82.0</td>
<td>(86.0, 5.8, 1.0)</td>
<td>10.9901</td>
</tr>
<tr>
<td>90.0</td>
<td>(90.0, 8.0, 7.0)</td>
<td>12.3863</td>
</tr>
<tr>
<td>110.0</td>
<td>(92.2, 2.5, 0.7)</td>
<td>5.9887</td>
</tr>
<tr>
<td>126.0</td>
<td>(78.2, 0.6, 7.8)</td>
<td>44.9245</td>
</tr>
<tr>
<td>131.0</td>
<td>(97.0, 2.0, 2.0)</td>
<td>3.1468</td>
</tr>
<tr>
<td>152.0</td>
<td>(106.3, 1.8, 0.7)</td>
<td>4.4724</td>
</tr>
<tr>
<td>Total error</td>
<td></td>
<td>269.151</td>
</tr>
</tbody>
</table>

Unlike the fuzzy regression model, we known that the left spread and the right of the predicted censored fuzzy regression model are greater than zero from Example 4.1.

On the other hand, the last two columns of Table 4.1 show the errors obtained from using the Least Squares (LS) method and the Least Absolute Deviation (LAD) estimation. We can see that due to the abnormal position of the center point of the given data, errors of the least squares method are larger than those of the least absolute deviation method. Thus, it is well justified to use the censored data method we have developed newly.

5. Conclusion

In this paper several methods that estimate the fuzzy regression model initiated by Tanaka et. al show the same results. And it carefully considered the problem that the estimated fuzzy outputs in the fuzzy regression model cross. In order to solve the problem of the crossing of the estimated fuzzy outputs, the sufficient condition was suggested, and
in case that the suggested sufficient condition cannot be satisfied, the
censored fuzzy regression model using censored data was suggested.

We verified that the fuzzy regression model with censored data was
an extension of the fuzzy regression model and identified the solution of
the problem of the crossing of the fuzzy outputs from which the censored
regression model was estimated.

If the data for fuzzy regression model is unable to satisfy linearity,
it is not appropriate to use the fuzzy regression model. Therefore, we
need to study methods of how to interpret fuzzy data given using a
segmented regression model or a semi-parametric regression that are
used in regression analysis.

References

[1] Y. O. Chang, *Hybrid fuzzy least-squares regression analysis and its reliability mea-


Department of General Studies
Hankuk Aviation University
Koyang 411, Korea
*E-mail*: shchoi@hau.ac.kr