A Mathematical Programming Method for Formulating a Fuzzy Regression Model Based on Distance Criterion

Liang-Hsuan Chen and Chan-Ching Hsueh

Abstract—Fuzzy regression models are useful to investigate the relationship between explanatory and response variables with fuzzy observations. Different from previous studies, this correspondence proposes a mathematical programming method to construct a fuzzy regression model based on a distance criterion. The objective of the mathematical programming is to minimize the sum of distances between the estimated and observed responses on the X axis, such that the fuzzy regression model constructed has the minimal total estimation error in distance. Only several c-cuts of fuzzy observations are needed as inputs to the mathematical programming model; therefore, the applications are not restricted to triangular fuzzy numbers. Three examples, adopted in the previous studies, and a larger example, modified from the crisp case, are used to illustrate the performance of the proposed approach. The results indicate that the proposed model has better performance than those in the previous studies based on either distance criterion or Kim and Bishu’s criterion. In addition, the efficiency and effectiveness for solving the larger example by the proposed model are also satisfactory.

Index Terms—Fuzzy regression model, fuzzy sets, mathematical programming.

I. INTRODUCTION

Regression analysis has a widespread application in various fields, such as business, engineering, and economics, to explore the statistical relationship between input (independent or explanatory) and output (dependent or response) variables. In general, the data to apply to the analysis should be measurable quantitatively and follow some probability distribution. However, in the real world, human estimation is, often, influential in the decision-making processes, and the data provided will be fuzzy in nature. For example, the observations are represented in linguistic terms, such as “about five inches,” and “approximately equal to 100 pounds.” The emergence of fuzzy regression is suitable to construct the relationship between input and output variables in the fuzzy environment. Since Tanaka et al. [1] first proposed the fuzzy linear regression model, it has motivated a number of researchers’ interest in this issue in both applications and methodologies.

After the first study, Tanaka [2], Tanaka et al. [3], and Tanaka and Watada [4] made some improvements. However, they all suffered some problems, such as sensitiveness to outliers [5], possibly producing infinite solutions, and producing a wider spread of the estimated output when more data are involved in the model [6], [7]. In addition to the Tanaka approach, there are several studies on fuzzy regression analysis [8]–[17]. Some of them applied the concept of least squares in the classical regression analysis [6], [7], [9], [10]. Several studies only incorporated the fuzziness of response variables into the model [11], [13]. Moreover, the regression coefficients were fuzzy numbers in the fuzzy regression models established in several studies [4], [5], [13], [18]. Applying these models to estimations will make the spread of the estimated responses wider when the magnitude of the independent variables increases, even though the spreads of the observed responses are approximately constant or actually decreasing.

In order to solve the problems outlined above, Kao and Chyu [6], [7] proposed updated fuzzy regression models with crisp regression coefficients. In order to develop the models in [6], a two-stage approach was proposed. They employed the least squares method to determine the regression coefficients using the defuzzified observations at the first stage. A mathematical programming model was then constructed at the next stage to determine the fuzzy error term using the evaluation criterion proposed by Kim and Bishu [18] as the objective of the programming problem. This approach is favored when compared with the methods in [9], [14], and [18]. As to the models in [7], the sums of fuzzy squared errors are minimized with respect to the crisp regression coefficients by applying Zadeh’s extension principle [19] and a ranking method for fuzzy numbers [20]. Kao and Chyu [7] then established a nonlinear programming problem to determine the crisp regression coefficients. After that, the average of the individual errors between the observed and estimated responses was treated as the fuzzy error term. Actually, this method is also a two-stage approach. For comparisons with Tanaka’s models [3], Kim and Bishu’s criterion was still employed. The results show that the proposed approach is better than Tanaka’s method.

Although Kao and Chyu’s approaches [6], [7] have shown the advantages of fuzzy regression analysis, they still have some problems. Their methods are actually two-stage approaches, leading to inefficiency in the resolution processes. The other problem is the use of Kim and Bishu’s criterion, which minimizes the total difference of the membership values between the observed and estimated fuzzy responses. In order to determine the fuzzy error term to minimize the difference in the second stage, the mathematical programming model presented by Kao and Chyu [6] is not suitable for general cases, since the model is designed based on a particular case of the relative position of the observed and estimated responses. In addition, Kim and Bishu’s measure sometimes cannot reflect the actual difference. For the ith observation, the measure can be formulated as follows:

\[ D_i = \int_{S_{\hat{\mu}_i} \cup S_{\hat{\varphi}_i}} |\mu_{\hat{\varphi}_i}(y) - \mu_{\hat{\mu}_i}(y)| \, dy \]  

where \( \mu_{\hat{\varphi}_i}(y) \) and \( \mu_{\hat{\mu}_i}(y) \) are the membership functions of the observed and estimated responses, respectively, \( S_{\hat{\mu}_i} \) and \( S_{\hat{\varphi}_i} \) are the supports of \( \mu_{\hat{\mu}_i}(y) \) and \( \mu_{\hat{\varphi}_i}(y) \), respectively, and \( D_i \) is the error in estimation. The fuzzy regression model fits the data better if the measure of \( D_i \) is smaller. Based on (1), if \( \mu_{\hat{\varphi}_i}(y) = 0 \) and \( \mu_{\hat{\mu}_i}(y) \) do not intersect each other, i.e., \( \mu_{\hat{\varphi}_i}(y) = 0 \) and \( \mu_{\hat{\mu}_i}(y) = 0 \) for some values of \( y \), the measure of \( D_i \) will be constant no matter what the distance between the two functions is. Under such circumstances, the measure cannot reflect the actual difference.

Different from the existing methods, this correspondence develops a mathematical programming model to determine the crisp regression coefficients and fuzzy adjustment variable simultaneously based on the concept of minimizing the sum of distances between the observed and estimated fuzzy responses. Since the distance factor is incorporated as the objective of the programming problem, the fuzzy regression model can have the minimum sum of distances between the observed and estimated fuzzy responses, i.e., a minimal total error, and can thus alleviate the aforementioned problems. In Section II, the mathematical
programming is described and established by using the $\alpha$-cuts of each fuzzy observation. In Section III, we then use three examples to compare with some other methods to illustrate the advantages of the proposed method. In addition, a larger example is also used to demonstrate the efficiency and effectiveness of the proposed approach. Finally, conclusions are made in Section IV.

II. MATHEMATICAL PROGRAMMING

In a classical regression analysis, the model formulates the relationship between several explanatory variables and the response variable as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \epsilon_i, \quad i = 1, \ldots, n \quad (2)$$

where $X_{ij}$ and $Y_i$ are the $j$th explanatory variable and the response variable, respectively, in the $i$th case, $\beta_j$ is the corresponding parameter of the $j$th explanatory variable, and $\epsilon_i$ is the error term associated with the $i$th observation. To determine the estimates of parameters, a group of observation data is collected, and normally, the least squares method is used.

If some of the observations of $X_{ij}$ and $Y_i$ are fuzzy, the classical linear regression model becomes a fuzzy one. A fuzzy linear regression model can be expressed as

$$\tilde{Y}_i = \beta_0 + \beta_1 \tilde{X}_{i1} + \beta_2 \tilde{X}_{i2} + \cdots + \beta_p \tilde{X}_{ip} + \tilde{\epsilon}_i, \quad i = 1, \ldots, n \quad (3)$$

where $\tilde{X}_{ij}$, $\tilde{Y}_i$, and $\tilde{\epsilon}_i$ are fuzzy numbers with the membership functions $\mu_{\tilde{X}_{ij}}$, $\mu_{\tilde{Y}_i}$, and $\mu_{\tilde{\epsilon}_i}$, respectively. Fuzzy numbers are generally defined as normal and convex sets [21]. In general, triangular fuzzy numbers are commonly used in fuzzy decision-making problems. For a triangular fuzzy number $\tilde{X}_{ij}$, the membership function is

$$\mu_{\tilde{X}_{ij}}(x) = \begin{cases} \frac{x - X_{ijl}}{X_{ijm} - X_{ijl}}, & X_{ijl} \leq x \leq X_{ijm} \\ \frac{X_{ijm} - x}{X_{iju} - X_{ijm}}, & X_{ijm} \leq x \leq X_{iju} \end{cases} \quad (4)$$

where the range $[X_{ijl}, X_{iju}]$ is the possible interval of the value; $X_{ijl}$, $X_{ijm}$, and $X_{iju}$ are the smallest, most possible, and largest value, respectively, and the most possible is also called the mean value with the membership degree equal to one [21], i.e., $\mu_{\tilde{X}_{ij}}(X_{ijm}) = 1$.

In (3), all observations are represented as fuzzy numbers, since crisp numbers can be treated as the particular case of a fuzzy number. In addition, the parameters in (3) are to avoid the problem of increasing the spreads of the fuzzy responses in estimation, from which the previous studies suffered [4], [5], [13], [18]. To determine the estimates of the parameters $\beta_j$, the least squares method is usually applied in the literature [6]–[9], [18], [22] to construct the relationship model between the explanatory and response variables.

Instead of using the least squares method, in this correspondence, we propose a mathematical programming method to find out the estimates of the parameters based on the idea of minimizing the sum of distances between the observed and estimated fuzzy responses. Let $b_0, b_1, \ldots, b_p$ be the estimates for the regression parameters $\beta_0, \beta_1, \ldots, \beta_p$, respectively. Based on the concepts of classical statistics, the estimated fuzzy response is expressed as

$$\tilde{\hat{Y}}_i = b_0 + b_1 \tilde{X}_{i1} + b_2 \tilde{X}_{i2} + \cdots + b_p \tilde{X}_{ip}, \quad i = 1, \ldots, n \quad (5)$$

However, the predicted or estimated responses will be crisp in (5) if the explanatory variables are crisp, and a large fuzzy error will arise. In order to reduce this weakness, in this correspondence, an adjustment term $\hat{\delta}$ is added to the aforementioned model as follows:

$$\hat{\tilde{Y}}_i = b_0 + b_1 \tilde{X}_{i1} + b_2 \tilde{X}_{i2} + \cdots + b_p \tilde{X}_{ip} + \hat{\delta}, \quad i = 1, \ldots, n \quad (6)$$

where $\hat{\delta}$ is a fuzzy number with the membership function $\mu_{\hat{\delta}}$. When $\tilde{X}_{ij}$ and $\tilde{Y}_i$ are triangular fuzzy numbers, the adjustment term will also be a triangular one, which is defined as $\hat{\delta} = (\delta_l, \delta_m, \delta_u)$ based on the fuzzy arithmetic operations [23]. The three parameters in the membership function $\mu_{\hat{\delta}}$, i.e., $\delta_l$, $\delta_m$, and $\delta_u$, are treated as variables to be determined. Therefore, the problem here is to find the estimates for the regression parameters and those of the adjustment term to provide the best explanation for the relationship between the explanatory and response variables. In other words, the estimates to be determined should minimize the fuzzy error. Referring to the absolute difference in Kim and Bishu’s criterion (1), this correspondence defines the fuzzy error as the minimum of the sum of absolute values of differences between the estimated and observed fuzzy responses, which can be formulated as

$$\min \sum_i \left| \tilde{\hat{Y}}_i - \tilde{\tilde{Y}}_i \right|. \quad (7)$$

The aforementioned minimum error is, obviously, the function of the estimates of the parameters in a fuzzy regression model. Since (7) involves several fuzzy numbers, it is difficult to solve directly to determine the estimates. The approach of this correspondence is to apply $\alpha$-cuts of each fuzzy number to represent it. Actually, one fuzzy number can be described by a group of its $\alpha$-cuts [23]. The advantage of using $\alpha$-cuts to express a fuzzy number is that the fuzzy problem can be converted into a crisp one, so that it can be formulated and solved easily. Referring to Kao and Chyu’s notations [7], the $\alpha$-cuts of $\tilde{X}_{ij}$, $\tilde{Y}_i$, and $\hat{\delta}$ are symbolized as $(\tilde{X}_{ij})_\alpha$, $(\tilde{Y}_i)_\alpha$, and $(\hat{\delta})_\alpha$, respectively, and can be represented as

$$(\tilde{X}_{ij})_\alpha = \left[ (X_{ijl})_\alpha, (X_{ijm})_\alpha, (X_{iju})_\alpha \right] = \left[ \min_{x_{ij}} \{ x_{ij} \in S_{\tilde{X}_{ij}} \mid \mu_{\tilde{X}_{ij}}(x) \geq \alpha \} , \max_{x_{ij}} \{ x_{ij} \in S_{\tilde{X}_{ij}} \mid \mu_{\tilde{X}_{ij}}(x) \geq \alpha \} \right]'$$

$$(\tilde{Y}_i)_\alpha = \left[ (Y_i)_\alpha \right]' = \left[ \min_{y_i} \{ y_i \in S_{\tilde{Y}_i} \mid \mu_{\tilde{Y}_i}(y) \geq \alpha \} , \max_{y_i} \{ y_i \in S_{\tilde{Y}_i} \mid \mu_{\tilde{Y}_i}(y) \geq \alpha \} \right]'$$

$$(\hat{\delta})_\alpha = \left[ (\hat{\delta})_\alpha \right]' = \left[ \min_\hat{\delta} \{ \hat{\delta} \in S_{\hat{\delta}} \mid \mu_{\hat{\delta}}(\hat{\delta}) \geq \alpha \} , \max_\hat{\delta} \{ \hat{\delta} \in S_{\hat{\delta}} \mid \mu_{\hat{\delta}}(\hat{\delta}) \geq \alpha \} \right]'$$

where $\alpha \in [0, 1]$; $[(X_{ijl})_\alpha, (X_{ijm})_\alpha, (X_{iju})_\alpha]$ and $[(Y_i)_\alpha]$ are the lower and upper bounds of $(\tilde{X}_{ij})_\alpha$, $Y_i$, and $\hat{\delta}$, respectively, at
Cuts, which is formulated as

\[ D \] 

of the two fuzzy numbers at the 

- level can be 

the estimated fuzzy response in (6) are 

respectively. For a particular 

level, the lower and upper bounds of 

the estimated fuzzy response at the 

level as 

\[ S \] 

relatively. For a particular 

level, the lower and upper bounds of 

the estimated fuzzy response in (6) are 

and 

respectively. For a particular 

level, the lower and upper bounds of 

the estimated fuzzy response at the 

level can be expressed as

\[
\begin{align*}
\tilde{Y}_i^L(k) & = b_0 + b_1(X_{i1})_{\alpha k}^L + b_2(X_{i2})_{\alpha k}^L \\
& \quad + \cdots + b_p(X_{ip})_{\alpha k}^L + (\delta)_{\alpha k}^L \\
\tilde{Y}_i^U(k) & = b_0 + b_1(X_{i1})_{\alpha k}^U + b_2(X_{i2})_{\alpha k}^U \\
& \quad + \cdots + b_p(X_{ip})_{\alpha k}^U + (\delta)_{\alpha k}^U,
\end{align*}
\]

respectively. Obviously, (8a) and (8b) will merge to be a traditional regression model when all fuzzy observations degenerate to be crisp, such that the adjustment term can be removed. Referring to (7) and based on the operations of fuzzy equations [23], we define the fuzzy error between \( \tilde{Y}_i^L \) and \( \tilde{Y}_i^U \) at the \( k \)th level as 

\[
(\tilde{Y}_i^L - \tilde{Y}_i^U)_L \]

which is actually the average of the distance between the lower bounds and that between the upper bounds of the two fuzzy numbers at the \( k \)th level on the \( X \) axis. The aforementioned distance idea can also be seen in Fig. 1. For the 4th case, the fuzzy error is the average of 

\[
(\tilde{Y}_i^L - \tilde{Y}_i^U)_L \]

for several \( \alpha \) levels, since a group of \( \alpha \)-cuts can describe a fuzzy number. Let \( D_k \) be the average error for the \( k \)th case by \( m \) \( \alpha \)-cuts, which is formulated as

\[
D_k = \frac{1}{2m} \sum_{i=1}^{m} \left( \left( \tilde{Y}_i^L(k) - Y_i(k) \right)_L + \left( \tilde{Y}_i^U(k) - Y_i(k) \right)_U \right),
\]

Equation (9) is actually called the Hamming distance in distance measure [23], and obviously, the measure \( D_k \) is equivalent to the absolute difference of two real numbers, if both the estimated and observed fuzzy numbers degenerate to be crisp ones. The total error for all observations will be \( \sum D_k \), which is obviously the function of the estimates \( b_0, b_1, \ldots, b_p, \delta_1, \delta_m \), and \( \delta_k \). The minimum fuzzy error can be achieved by minimizing the total error \( \sum D_k \).

In order to determine the estimates of fuzzy regression models in obtaining the best explanation for the fuzzy regression model, a mathematical programming problem is formulated as follows:

\[
\begin{align*}
\min & \sum_{i=1}^{n} D_i \\
\text{s.t.} & \begin{align*}
\left( \tilde{Y}_i^L(k) \right)_{\alpha k}^L & = b_0 + b_1(X_{i1})_{\alpha k}^L + b_2(X_{i2})_{\alpha k}^L \\
& \quad + \cdots + b_p(X_{ip})_{\alpha k}^L + (\delta)_{\alpha k}^L \\
\left( \tilde{Y}_i^U(k) \right)_{\alpha k}^U & = b_0 + b_1(X_{i1})_{\alpha k}^U + b_2(X_{i2})_{\alpha k}^U \\
& \quad + \cdots + b_p(X_{ip})_{\alpha k}^U + (\delta)_{\alpha k}^U,
\end{align*}
\end{align*}
\]

Obviously, this minimization problem is a constrained nonlinear problem, and general commercial software packages can be used to solve it. The solution of (10) can achieve the minimum sum of errors in terms of distance on the \( X \) axis, unlike the Kim and Bishu’s criterion (1), and therefore, the resulting average estimation error for all observations will be \( (1/n) \sum D_i \). From the concept of the expected error in the regression analysis, the measure \( (1/n) \sum D_i \) can be defined as the estimate of the expected error for the fuzzy estimation in terms of distance criterion. Theoretically, better estimates can be obtained when more \( \alpha \)-cuts are applied. However, based on the similar experience of Chen and Lu [24] using the distance criterion in the ranking problem of fuzzy numbers, three or four \( \alpha \)-cuts are sufficient when the membership functions of fuzzy numbers are triangular.

III. ILLUSTRATIVE EXAMPLES

In order to illustrate the performance of the proposed mathematical programming model in constructing fuzzy regression models, three examples discussed in the literature are solved in this section. The performance of the proposed models is compared with that achieved by the other models based on Kim and Bishu’s criterion and the distance criterion proposed in this section. The performance comparison among the models is carried out by using the training data that are used for constructing all the models. The observations in the first example are crisp for the explanatory variable and fuzzy for the response variable, whereas those for both explanatory and response variables are fuzzy in the other two examples. The third example has two explanatory variables, different from the other two examples with only one explanatory variable. In addition, to demonstrate the efficiency and effectiveness of the proposed models, a larger data set containing three explanatory variables and 50 observations is also adopted. The four examples are solved on a laptop, using the well-known LINGO software [25].

Example 1: An example, designed by Tanaka et al. [3], has been used to illustrate the performance of fuzzy regression models by several researchers [6], [9], [18], [26]. This example includes one explanatory variable with five crisp observations and one response variable with fuzzy ones, which is listed in the left part of Table I. The fuzzy observations are represented as triangular fuzzy numbers. Tanaka et al. [3] discussed three types of fuzzy regression models in their study, i.e., Min problem, Max problem, and Conjunction problem. Kao and Chyu [6] and Nasrabadi and Nasrabadi [26] adopted the Min problem at \( \alpha = 0 \) for the comparisons, by which the fuzzy regression model (THW) is

\[
\tilde{Y}_{\text{THW}} = (0, 3.850, 7.700) + 2.100X.
\]

Authorized licensed use limited to: National Cheng Kung University. Downloaded on October 20, 2009 at 22:24 from IEEE Xplore. Restrictions apply.
The aforementioned model has a crisp slope and a fuzzy intercept. Using the data in this example, Kim and Bishu [18] also constructed a fuzzy regression model (KB) as

\[ \hat{Y}_{KB} = (3.110, 4.950, 6.840) + (1.550, 1.710, 1.820)X. \]  \hspace{1cm} (12)

Presenting a least squares approach for fuzzy observations, Diamond [9] constructed the fuzzy regression model (DM) as

\[ \hat{Y}_{DM} = (3.110, 4.950, 6.790) + (1.550, 1.710, 1.870)X. \]  \hspace{1cm} (13)

Nasrabadi and Nasrabadi [26] proposed an algorithm according to a mathematical programming-based approach to establish the fuzzy regression model (NN) as follows:

\[ \hat{Y}_{NN} = (2.360, 4.680, 7.000) + 1.730X. \]  \hspace{1cm} (14)

Instead of the fuzzy estimates constructed in the literature, Kao and Chyu [6] proposed a two-stage approach based on Kim and Bishu’s criterion to build up the fuzzy regression model (KC) with crisp coefficients and a fuzzy error term, which is formulated as

\[ \hat{Y}_{KC} = 4.950 + 1.710X + (-3.010, 0, 1.800). \]  \hspace{1cm} (15)

To avoid the problem wherein the estimation of the spread of the estimated fuzzy responses becomes wider as the magnitude of the explanatory variables increases, this correspondence proposed a mathematical programming model based on the distance criterion to construct the fuzzy regression model (DT). Using two \( \alpha \)-cuts, i.e., \( \alpha = 0 \) and 1, the model is

\[ \hat{Y}_{DT} = 4.750 + 1.650X + (-2.400, 0, 2.400). \]  \hspace{1cm} (16)

Obviously, (16) is similar to (15), although the intercept and slope are somewhat smaller than those of (15), whereas the fuzzy adjustment term in (16) is larger than the fuzzy error term in (15) based on the fuzzy numbers ranking methods [24]. It is also noted that the spread of the estimated fuzzy response is almost equal, 4.8, for (15) and (16). However, the fuzzy adjustment term of \( \hat{Y}_{DT} \) is a symmetrical triangular fuzzy number. Comparing with (15), this makes our model more reasonable based on the fuzzy arithmetic operations [23], since the fuzzy responses observed are symmetrical triangular fuzzy numbers.

To compare the performance of the aforementioned models in estimation, distance criterion and Kim and Bishu’s criterion are applied to calculate the total error in estimating the fuzzy responses. In applying the distance criterion, two \( \alpha \)-cuts, i.e., \( \alpha = 0 \) and 1, are used to calculate the estimation errors for all models. Table I shows the errors in estimating fuzzy responses for all models based on different criteria. Obviously, our approach (DT) is the best among all the models, producing 5.6 for the distance criterion and 9.066 for Kim and Bishu’s criterion.

**Example 2:** Sakawa and Yano [14] designed an example to illustrate their fuzzy regression model. The example contains eight pairs of fuzzy observations represented as symmetrical triangular fuzzy numbers, which is shown in the left part of Table II. To apply Sakawa and Yano’s approach, the degree of conformity between the observed and estimated responses should be specified in advance. Referring to Kao and Chyu’s experience [6], the value of the degree of conformity is set to be 0.5, and the fuzzy regression model (SY) constructed is

\[ \hat{Y}_{SY} = (3.031, 3.201, 3.371) + (0.498, 0.579, 0.659)X. \]  \hspace{1cm} (17)

The fuzzy regression coefficient \( b_1 \) is required to be crisp in employing the Diamond method [9] when the observations of the explanatory variable are fuzzy numbers. Using the data set in this example, the fuzzy regression model (DM) is

\[ \hat{Y}_{DM} = (3.263, 3.563, 3.863) + 0.521X. \]  \hspace{1cm} (18)

For this example, Kao and Chyu (KC) [6] as well as Nasrabadi and Nasrabadi (NN) [26] built up their fuzzy regression models as

\[ \hat{Y}_{KC} = 3.572 + 0.519X + (-0.240, 0, 0.240) \]  \hspace{1cm} (19)

\[ \hat{Y}_{NN} = 3.577 + (0.447, 0.547, 0.647)X \]  \hspace{1cm} (20)

respectively. Applying the proposed mathematical programming model (10) to this example based on two \( \alpha \)-cuts as Example 1, the estimates of parameters are solved and the fuzzy regression model (DT) is

\[ \hat{Y}_{DT} = 1.981 + 0.444X + (1.686, 1.964, 2.242). \]  \hspace{1cm} (21)

It is noted that all of the aforementioned estimated fuzzy responses are symmetrical triangular fuzzy numbers, since the response and explanatory variables are also the same type of fuzzy numbers in this example. Applying the aforementioned models based on the two criteria, the errors in estimation are listed in Table II. The same as in the previous example, two \( \alpha \)-cuts are used to calculate the
estimation errors based on the distance criterion. The table shows that the proposed approach is the most favored among the models in terms of the distance criterion (total error of DT = 5.556) and Kim and Bishu’s criterion (total error of DT = 6.039).

Example 3: Wu [27] recently proposed a method to construct the fuzzy regression model. To determine the estimators of parameters in the model, the fuzzy least squares are adopted by applying the extension principle. An example containing 15 fuzzy observations with two explanatory variables and one response variable, which is shown in the left part of Table III, is designed to illustrate this approach. All fuzzy observations are triangular fuzzy numbers. The applications of Wu’s method generated an interval of values for each estimator at each $\alpha$ level. For comparison, the most possible value, i.e., $\alpha = 1$, is adopted to establish the model as

$$\hat{Y}_{wU} = 3.453 + 0.496 \tilde{X}_1 + 0.009 \tilde{X}_2.$$  (22)

Using the 15 fuzzy observations, the proposed mathematical programming model based on two $\alpha$-cuts, as in the previous examples, is solved, and the model (DT) is

$$\hat{Y}_{DT} = 3.110 + 0.489 \tilde{X}_1 + 0.009 \tilde{X}_2 - (-12.243, -3.839, 10.974).$$  (23)

Comparing (22) and (23), the estimate $b_1$ is almost the same, and $b_2$ is identical for the two models. It is noticed that the fuzzy adjustment term in (23) is a minus item in the model and contains the negative smallest and mean values. Examining the estimation errors in Table III, the performance of the proposed approach in estimation is, obviously, better than that of Wu’s method using both criteria.

To compare the performance of the proposed approach with the other methods, two $\alpha$-cuts, i.e., $\alpha = 0$ and 1, are used in the three examples. As mentioned before, more $\alpha$-cuts may achieve more accurate models, but the computational efforts will increase. To examine the effects of various $\alpha$-cuts on the estimation errors based on distance criterion, several different numbers of $\alpha$-cuts, $m = 5, 7, 9, 11$, where $\alpha_k = k/(m-1)$, $k = 0, \ldots, m - 1$, are applied to (10) using the fuzzy observations of Examples 2 and 3. Figs. 2 and 3 demonstrate

<table>
<thead>
<tr>
<th>$\tilde{X}_1$</th>
<th>$\tilde{X}_2$</th>
<th>$\hat{Y}$</th>
<th>Distance criterion</th>
<th>Kim &amp; Bishu criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>(151, 274, 322)</td>
<td>(1432, 2450, 3461)</td>
<td>(111, 162, 194)</td>
<td>SY</td>
<td>5.173</td>
</tr>
<tr>
<td>(101, 180, 291)</td>
<td>(2448, 3254, 4463)</td>
<td>(88, 120, 161)</td>
<td>DM</td>
<td>11.278</td>
</tr>
<tr>
<td>(221, 375, 539)</td>
<td>(2592, 3802, 5116)</td>
<td>(161, 223, 288)</td>
<td>KC</td>
<td>14.202</td>
</tr>
<tr>
<td>(128, 205, 313)</td>
<td>(1414, 2838, 3252)</td>
<td>(83, 131, 194)</td>
<td>NN</td>
<td>2.229</td>
</tr>
<tr>
<td>(62, 86, 112)</td>
<td>(1024, 2347, 3766)</td>
<td>(51, 67, 83)</td>
<td>DT</td>
<td>4.856</td>
</tr>
<tr>
<td>(132, 265, 362)</td>
<td>(2136, 3782, 5991)</td>
<td>(124, 169, 213)</td>
<td>SY</td>
<td>13.346</td>
</tr>
<tr>
<td>(66, 98, 152)</td>
<td>(1687, 3908, 4325)</td>
<td>(62, 81, 102)</td>
<td>DM</td>
<td>7.365</td>
</tr>
<tr>
<td>(151, 330, 463)</td>
<td>(1524, 2450, 3864)</td>
<td>(138, 192, 241)</td>
<td>KC</td>
<td>19.484</td>
</tr>
<tr>
<td>(115, 195, 291)</td>
<td>(1216, 2137, 3161)</td>
<td>(82, 116, 159)</td>
<td>NN</td>
<td>8.963</td>
</tr>
<tr>
<td>(35, 53, 71)</td>
<td>(1432, 2560, 3782)</td>
<td>(41, 55, 71)</td>
<td>DT</td>
<td>3.223</td>
</tr>
<tr>
<td>(307, 430, 584)</td>
<td>(2592, 4020, 5562)</td>
<td>(168, 252, 367)</td>
<td>SY</td>
<td>8.047</td>
</tr>
<tr>
<td>(121, 236, 370)</td>
<td>(1734, 2660, 4094)</td>
<td>(111, 144, 198)</td>
<td>KC</td>
<td>15.043</td>
</tr>
<tr>
<td>(103, 157, 211)</td>
<td>(1426, 2088, 3312)</td>
<td>(78, 103, 148)</td>
<td>NN</td>
<td>6.174</td>
</tr>
<tr>
<td>(216, 370, 516)</td>
<td>(1785, 2605, 4042)</td>
<td>(167, 212, 267)</td>
<td>DT</td>
<td>17.922</td>
</tr>
</tbody>
</table>

| Total estimation error | 150.670 | 112.674 | 284.658 | 217.739 |
the resulting total estimation errors ($\sum D_i$) for Examples 2 and 3, respectively, using various $\alpha$-cuts for various models. The figures illustrate that the changes in the total estimation errors are insignificant for different models using various $\alpha$-cuts based on distance criterion, notably remaining equal for DM, KC, and DT in Example 2 due to the symmetrical triangular fuzzy numbers. Obviously, the total estimation errors will converge to a stable level when the number of $\alpha$-cuts increases. The proposed approach still performs the best in these experiments.

**Example 4:** In order to demonstrate the applicability of the proposed models, a larger data set is adopted to examine the efficiency and effectiveness. This data set was provided by Chatterjee et al. [28] to investigate the relationship between per capita expenditure on public education ($Y$) for each of the 50 states in the U.S. in 1975 and three variables, namely: 1) per capita income ($X_1$) in 1973; 2) the percentage of population under 18 years of age ($X_2$) in 1974; and 3) the percentage of the population living in urban areas ($X_3$) in 1970. For the purpose of demonstrating the proposed models in this correspondence, suppose that the data set has measurement errors in the data collection processes, so that the observations in the data set are represented as fuzzy numbers. Let the original data be the mean values, and the (left spread, right spread) of $Y$, $X_1$, $X_2$, and $X_3$ be (12% or 14%, 8% or 10%), (8% or 10%, 16% or 18%), (6% or 8%, 14% or 16%), and (10% or 12%, 6% or 8%) of the corresponding mean value, respectively. Therefore, there are four combinations for each explanatory and response variable in the fuzzy version, and one of them for each variable is determined by means of simulations. Obviously, the resulting fuzzy observations are nonsymmetrical triangular fuzzy numbers. The fuzzified observations can be found in Table IV. Following the previous examples, two $\alpha$-cuts are used and put into (10). The model is solved by LINGO software in a Microsoft Windows XP Professional environment running on a laptop with a central processing unit of 1.4 GHz and a random access memory of 512 MB. The model is built up as

$$\hat{Y}_{DT} = -355.584 + 0.063\tilde{X}_1 + 9.886\tilde{X}_2$$
$$+ 0.328\tilde{X}_3 - (-11.229, 3.521, 81.731)$$

by taking the software runtime of 23 s. The efficiency seems satisfactory. The estimated fuzzy responses, the estimation error in terms of distance for each fuzzy observation, and the total estimation error (1513.840) are listed in Table IV.

Using the same data set, Birkes and Dodge [29] applied nonparametric regression analysis to construct the model as

$$\hat{Y}_{BD} = -422.9 + 0.05868X_1 + 12.54X_2 + 0.3372X_3.$$  

As mentioned before, $D_i$ can measure the absolute difference of two real numbers, if both the estimated and observed fuzzy numbers degenerate to be crisp ones. For comparison, the total estimation error from (25) in terms of distance is calculated as 1545.525, which is larger than that from (24). Furthermore, to demonstrate the effectiveness of the proposed model, a 10-fold cross validation is made by using the fuzzified data set in this example. Two $\alpha$-cuts of each fuzzified observation are used as the testing/training data. The proposed model has the better performance with average estimation error of 35.697 in terms of distance than the multiple linear regression model with the average error of 36.176.

**IV. CONCLUSION**

A number of approaches to construct fuzzy regression models have been presented by previous studies to investigate the relationship between explanatory and response variables with fuzzy observations. Some studies [6], [7] have emphasized the importance of avoiding the problem of increasing the spread of the estimated responses as the magnitude of the explanatory variables increases, even though the spreads of the observed responses were kept roughly constant. Based on such a concept, this correspondence proposes a mathematical programming model to obtain the estimates of fuzzy regression coefficients based on the distance criterion.

Unlike Kim and Bishu’s criterion adopted by previous studies [3], [6], [9], [14], [18], [26], [27], the use of the distance criterion in this correspondence can measure the total error in estimation, even without an overlapped area between the estimated and observed responses. The fuzzy regression model constructed by the proposed mathematical programming model can have the minimum total error in estimation in terms of distance. The advantages of the proposed mathematical programming approach are described as follows. First, the proposed approach can also avoid the problem of increasing the spread of the estimated response due to the use of fuzzy coefficients in the previous studies, since the coefficients are crisp. Second, only several $\alpha$-cuts of fuzzy observations are needed to build up the proposed mathematical programming model. This means that triangular fuzzy observations are not necessary, and even the exact membership functions are not required. Under the circumstance that the exact membership functions are unknown, the only problem in constructing the mathematical
programming is that the membership function of the fuzzy adjustment term will need to be approximately determined. Although the proposed mathematical model is a nonlinear problem, it is easily solved with a general software package, in this correspondence, LINGO. The performance of the proposed approach is illustrated by three examples from the existing relevant literature. The results show that our approach is better than the previous studies based on both the distance criterion and Kim and Bishu’s criterion. In addition, the computational efficiency and effectiveness of the proposed approach seems satisfactory as was demonstrated by using a larger data set as the fourth example.

REFERENCES


