A comparative study of fuzzy multiple regression model and least square method

Ubale AB, Sananse SL

Abstract
The objectives of this study is to formulate a multiple fuzzy linear regression model using crisp input/output to investigate the relationship between explanatory and response variables to estimate the model parameters. For present study, fuzzy linear regression model proposed by Zadeh’s is used which is based on fuzzy linear function. Comparative study of fuzzy multiple regression model and conventional multiple regression model is done on the basis of coefficient of determination which is used as goodness of fit for both the models. Finally, a numerical example is provided for demonstration of the results. It is observed that the fuzzy multiple regression model is more suitable than the conventional multiple regression model resulting in higher coefficient of determination.

Keywords: Fuzzy linear regression, SST, SSR, Coefficient of determination, least square method.

1. Introduction
Regression analysis has a widespread application in various fields, such as business, engineering, agriculture, and economics, to explain the statistical relationship between explanatory and response variables. The use of statistical linear regression is bounded by some assumption about the given data that is error terms are mutually independent and identical distributed. Statistical regression model can be applied only if the given data are distributed according to statistical model, and the relation between explanatory and response variables is crisp. However, in real life situations decision-making processes the data will be fuzzy in nature. For example, the observations are represented in linguistic terms, vagueness, such kinds of data the fuzzy regression model is suitable to construct the relationship between input and output variables in the fuzzy environment.

The fuzzy linear regression model was first introduced by Tanaka et al. [1], by using linear programming problem to determine the regression coefficient as a fuzzy numbers. Later Tanaka [2], Tanaka and Watada [3], and Tanaka et al. [3] made some improvements. As pointed out by Redden and Woodall [5], their method is very sensitive to outliers. Moreover, the spread of the estimated response becomes wider as more observations are included in the model. The second approach developed by Diamond [6], which minimizes the total error some of square of the output is called the fuzzy least square method. On the basis of a possibilistic regression model.

2. Literature Review
Diamond [6] introduced the fuzzy regression model to minimize the sum of squares of differences for the centre of fuzzy of fuzzy numbers and the sum of squares of differences for spreads of fuzzy numbers. PierpaoloD’Urso, et al. [11]. Introduced new approach of fuzzy linear regression analysis. They had developed doubly linear adaptive fuzzy regression model, based on two linear models such as centre regression model and a spread regression model. The first one was explained the centre of the fuzzy observations, while the second one was for their spreads. They had observed that doubly linear adaptive fuzzy regression analysis had alternative methods for fuzzy linear regression analysis. Volker Krätschmer [12] developed new fuzzy linear regression models. He had generalized the type of single
Ordinary equation in linear regression models by incorporated the physical vagueness of the involved items in the form of fuzzy data for the variables. He had suggested that ordinary least-squares method was greater flexibility for modelling and estimation Yun-His, et al. [13]. Developed hybrid fuzzy least-squares regression. They were used weighted fuzzy-arithmetic mean and least-squares fitting criterion. They had compared hybrid regression with the ordinary regression and other fuzzy regression methods. They suggested that hybrid fuzzy regression model satisfied a limiting behaviour that the fuzziness decreases, the equations were similar to the results of the ordinary regression. Chiang Kao et al. [14], proposed two-stage fuzzy linear regression model. In first stage, the fuzzy observations was de-fuzzified so that the traditional least-squares method was applied to find a crisp regression line showed the general trend of the data. In the second stage, the error term of the fuzzy regression model was represents the fuzziness of the data and also determined to given the regression model was best explanatory power for the data. Finally they suggested that two-stage method had better performance than the Kim–Bishu’s and Diamond fuzzy linear regression model. Hsien-Chung Wu [15] proposed fuzzy estimates of regression parameters with the help of “Resolution Identity”. He said that the fuzzy estimates was constructed from the alpha level least-squares estimates used the alpha level real-valued data. Finally he had developed two computational procedures to solve the optimization problems.

Kyung. Kim, et al. [17]. Proposed fuzzy least absolute deviation method to construct fuzzy linear regression model with fuzzy input and fuzzy output. They had suggested that the fuzzy least absolute deviation method was more effective than the least square method used in the fuzzy regression analysis. Rajan Alex [16] introduced fuzzy regression and fuzzy inference. He was applied the two kinds of information resources, quantitative and qualitative information and also used simultaneously in practical prediction. Finally he suggested that fuzzy regression and fuzzy inference had better performance than pure regression or pure inference model.

Hye-Youngung, et al. [13], proposed rank transformation method. They investigated a method to obtain a predicted output with respect to a specific target value. They suggested that the rank transformation method in fuzzy regression model was better performance than the Chen and Hsueh and Diamond fuzzy linear regression models. Furkan Baseret et al. [20] applied hybrid fuzzy least squares regression analysis to predict future claim costs by used the concept of London Chain Ladder (LCL) method. They had suggested that the hybrid fuzzy least-squares regression model was taken both randomness and fuzziness type of uncertainty into a regression model. A.B Ubale, S.L Sananse [21] introduced brief research trends in application of fuzzy regression analysis in different field.

3. Materials and Methods

A regression Model that involves more than one explanatory or independent variables, called as multiple linear regression model. This model generalizes the simple linear regression in two ways. It allows the mean function \( E(y) \) to depend on more than one explanatory variables. Let \( y \) denotes the dependent (or study) variable that is linearly related to \( k \) independent (or explanatory) variables \( X_1, X_2, \ldots, X_n \) through the parameters \( \beta_1, \beta_2, \ldots, \beta_k \) and we write

\[
y = X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k + \varepsilon
\]

This is called as the multiple linear regression model. The parameters \( \beta_1, \beta_2, \ldots, \beta_k \) are the regression coefficients associated with \( X_1, X_2, \ldots, X_n \) respectively and \( \varepsilon \) is the random error component reflecting the difference between the observed and fitted linear relationship. These \( n \) equations can be written as

\[
\begin{pmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{pmatrix} =
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\begin{pmatrix}
\beta_0 \\
\beta_1 \\
\vdots \\
\beta_k
\end{pmatrix} +
\begin{pmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{pmatrix}
\]

Or

\[
y = X\beta + \varepsilon
\]

where \( y = (y_1, y_2, \ldots, y_n)' \) is a \( n \times 1 \) vector of \( n \) observation on study variable,

\[
X =
\begin{pmatrix}
x_{11} & x_{12} & \cdots & x_{1k} \\
x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\]

is a \( n \times k \) matrix of \( n \) observations on each of the \( k \) explanatory variables, \( \beta = (\beta_1, \beta_2, \ldots, \beta_k)' \) is a \( k \times 1 \) vector of regression coefficients and \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' \) is a \( n \times 1 \) vector of random error components or disturbance term.

If intercept term is present, take first column of \( X \) to be \((1,1,\ldots,1)'\). So that

\[
X =
\begin{pmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1k} \\
1 & x_{21} & x_{22} & \cdots & x_{2k} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{nk}
\end{pmatrix}
\]

In this case, there are \((k-1)\) intercept term.

### 3.1 Least Squares Methods (LS)

A general procedure for the estimation of regression coefficient vector is to minimize

\[
\sum_{i=1}^{n} M(e_i) = \sum_{i=1}^{n} M(y_i - x_i^t \beta_1 - x_i^t \beta_2 - \cdots - x_i^t \beta_k)
\]

for a suitably chosen function \( M \).

Let \( B \) be the set of all possible vectors \( \beta \). If there is no further information, then \( B \) is \( k \)-dimensional real Euclidean space.

The object is to find a vector \( b' = (b_1, b_2, \ldots, b_k) \) from \( B \) that minimizes the sum of squared deviations of \( e_i \)'s, i.e.,

\[
S(B) = \sum_{i=1}^{n} e_i^2 = e'e = (y - X\beta)'(y - X\beta)
\]

for given \( y \) and \( X \). A minimum will always exist as \( S(\beta) \) is a real valued, convex and differentiable function. Write \( S(\beta) = y'y + \beta'X'X\beta - 2\beta'X'y \)

Differentiate (4) with respect to \( \beta \)

\[
\frac{\partial S(\beta)}{\partial \beta} = 2X'X\beta - 2X'y
\]

\[
\frac{\partial^2 S(\beta)}{\partial \beta \partial \beta'} = 2X'X
\]

The normal equation is

\[
\frac{\partial S(\beta)}{\partial \beta} = 0
\]

\[
= X'X\beta = X'y
\]

Equation (5) is the least-squares normal equations. To solve the normal equations, multiply both sides of equation (5) by the inverse of \( X'X \). Thus, the least squares estimator of \( \beta \) is

\[
\hat{\beta} = (X'X)^{-1}X'y
\]
\[ R^2 = 1 - \frac{SS_{res}}{SS_T} \]  
\[ \text{Where} \]  
\[ SS_{res} : \text{sum of squares due to residuals}, \]  
\[ SS_T : \text{total sum of squares}, \]  
\[ R^2 \text{ measure the explanatory power of the model which in turn reflects the goodness of fit of the model.} \]

### 3.3 Estimate the parameters of Multiple Fuzzy regression

Fuzzy linear regression analysis is first proposed by H Tanaka [1], and applied for many researches by Hsiao-Fan Wang et al. [4]. The model is given below.

\[ \bar{Y} = \bar{A}X + \bar{A}_X + \cdots + \bar{A}_N X_N = \bar{A}X \]  
\[ \text{where} \]  
\[ X = [X_0, X_1, \ldots, X_N]^T \]  
\[ \text{is a vector of independent variables,} \]  
\[ \bar{A} = [\bar{A}_0, \bar{A}_1, \ldots, \bar{A}_N]^T \]  
\[ \text{is a vector of fuzzy coefficient presented in the form of triangular fuzzy numbers denoted by} \]  
\[ \bar{a}_j = (a_j, c_j, c_j) \]  
\[ \text{with its membership function described as} \]

\[ \mu_{\bar{a}_j}(a_j) = \begin{cases} 1 - \frac{|a_j - a|}{c_j} & , a_j - c_j \leq a_j \leq a_j + c_j, \forall j = 1, 2, \ldots, N. \\ 0, & \text{otherwise} \end{cases} \]

Where \( a_j \) the central value and \( c_j \) is the spread value.

Therefore the equation (8) can be written as

\[ \bar{Y}_i = (\bar{a}_0, c_0) + (\bar{a}_1, c_1)X_1 + (\bar{a}_2, c_2)X_2 + \cdots + (\bar{a}_N, c_N)X_N \]  
\[ \text{Equation (10) is fuzzy regression model for crisp input and crisp output data. By applying the Extension Principle [5], it constructs the membership function of fuzzy number \( \bar{Y}_i \) as shown in (11) and each value of dependent variable can be estimated as a fuzzy number is introduced in [9] as follows:} \]

\[ \mu(\bar{Y}_i) = \begin{cases} 1 - \frac{|Y_{iL} - X|}{c_i(X)} & X \neq 0, \\ 0, & X = 0, Y \neq 0 \\ 0, & X = 0, Y = 0 \end{cases} \]

To determine the fuzzy parameters while minimizing the total sum of the spreads of the estimated values for a certain \( h \) level, using a linear programming problem in [12] called Min problem, as follows:

\[ \text{Min} \sum_{i=1}^{m} \sum_{j=0}^{N} (c_j|X_{ij}|) \]
\[ \text{s. t.} \]
\[ N \sum_{j=0}^{N} \alpha_j X_{ij} + (1 - h) \sum_{j=0}^{N} c_j|X_{ij}| \geq Y_{id} = 1, 2, \ldots, M \]
\[ \sum_{j=0}^{N} \alpha_j X_{ij} - (1 - h) \sum_{j=0}^{N} c_j|X_{ij}| \leq Y_{id} = 1, 2, \ldots, M \]
\[ c_j \geq 0, \alpha \in R, X_{i0} = 1, (0 \leq h \leq 1) \]

In fuzzy linear regression, values of the vector independent variables \( X \) have its corresponding fuzzy numbers \( \bar{Y} \) in which, without any other information, the probabilities of occurrence of all points in \( \bar{Y} \) are assumed to be equal to the membership functions of the fuzzy parameters in a fuzzy linear regression model are symmetric, then the values of \( Y_{ih} \) is equal to \( \bar{Y} \) is estimated by introduced in [9] as follows:

Then the \( \bar{Y}_i \) we have \( Y_{iL} = Y_{iL}^{h=1} - c|X| \) and \( Y_{iU}^{h=1} = Y_{iU}^{h=1} + c|X| \)

Therefore

\[ \bar{Y}_i = \frac{Y_{iL}^{h=1} + Y_{iU}^{h=1}}{2} = \frac{[Y_{ih}^{h=1} - c|X|] + (Y_{iU}^{h=1} + c|X|)}{2} = Y_{iL}^{h=1} \]

We consider the M data points the \( Y_i \) are assumed to be equal to the membership functions of the fuzzy parameters in a fuzzy linear regression model are symmetric, then the values of \( Y_{ih}^{h=1} \) is equal to \( \bar{Y} \) is estimated by introduced in [9] as follows:

\[ \text{Then the} \bar{Y}_i \text{ have } Y_{iL} = Y_{iL}^{h=1} - c|X| \text{ and } Y_{iU}^{h=1} = Y_{iU}^{h=1} + c|X| \]

\[ \text{Therefore} \]

\[ \bar{Y}_i = \frac{Y_{iL}^{h=1} + Y_{iU}^{h=1}}{2} = \frac{[Y_{ih}^{h=1} - c|X|] + (Y_{iU}^{h=1} + c|X|)}{2} = Y_{iL}^{h=1} \]

Where the center values

\[ Y_{ih}^{h=1} = Y_{iL}^{h=1} + \frac{Y_{iU}^{h=1} - Y_{iL}^{h=1}}{2} \]

\[ \forall i = 1, 2, \ldots, M. \] Since the difference between \( Y_{iL} \) and \( Y_{iU} \)

Where \( n \) is the number of observation?

\[ \sum_{i=1}^{M} [(Y_{ih}^{h=1} - Y_{iL}^{h=1}) + (Y_{iU}^{h=1} - Y_{iL}^{h=1})] = \sum_{i=1}^{M} [(Y_{ih}^{h=1} - Y_{iL}^{h=1}) + (Y_{iU}^{h=1} - Y_{iL}^{h=1})]^2. \]
\[\sum_{i=1}^{M}(Y_i - Y_{L})^2 + \sum_{i=1}^{M}(Y_{U} - Y_i)^2 + 2\sum_{i=1}^{M}((Y_i - Y_{L} \cdot (Y_{U} - Y_i)) \tag{16}\]

Putting the values equation (14), (15) in equation (17) can be written as

\[2\sum_{i=1}^{M}((n_i \cdot \Delta/2), (n_i \cdot \Delta/2)) - 2\sum_{i=1}^{M}((a_i \cdot \Delta), (n_i \cdot \Delta - a_i \cdot \Delta)) \]

\[= 2\sum_{i=1}^{M}((n_i \cdot \Delta/2)^2 - a_i \cdot n_i \cdot \Delta^2 + a_i^2 \Delta^2) = 2\sum_{i=1}^{M}(n_i \cdot \Delta^2 - a_i \cdot \Delta)^2 \]

\[= 2\sum_{i=1}^{M}((Y_i \cdot n_{L} - \frac{n_{L}}{2} - (Y_{L} + a_i \cdot \Delta))^2 = 2\sum_{i=1}^{M}(Y_i \cdot n_{L}^1 - Y_{L}^1)^2. \tag{17}\]

By using equation (16) and (17),

\[\sum_{i=1}^{M}(Y_i - Y_{L})^2 + \sum_{i=1}^{M}(Y_{U} - Y_i)^2 = \sum_{i=1}^{M}(Y_i \cdot n_{L}^1 - Y_{L}^1)^2 + \sum_{i=1}^{M}(Y_i \cdot n_{L}^1 - Y_{L}^1)^2 + 2\sum_{i=1}^{M}(Y_i \cdot n_{L}^1 - Y_{L}^1)^2. \tag{18}\]

Let \(\sum_{i=1}^{M}(Y_i - Y_{L})^2\) be the total sum of squares (SST) of fuzzy regression interval, \(\sum_{i=1}^{M}(Y_{U} - Y_i)^2\) be the regression sum of squares (SSR) and \(2\sum_{i=1}^{M}(Y_i \cdot n_{L}^1 - Y_{L}^1)^2\) be the error sum of squares (SSE), then by using Equation (18) total sum of squares is as follows:

\[SST = SSR + SSE. \tag{19}\]

Here the total sum of squares (SST) measures the total variation of \(Y_i\) between lower and upper bounds. The error sum of squares (SSE) estimates the difference when we use \(Y_i \cdot n_{L}^1\) to estimate \(Y_i\) whereas regression sum of squares (SSR) represents the variation of \(Y_i \cdot n_{L}^1\) with respect to lower and upper bounds.

The measure of the degree of interpretation let us define an index of confidence by

\[IC = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{20}\]

which is similar to the coefficient of determinant, \(R^2\), in statistics. Since SST is a measure of the interval variation between \(Y_i \cdot n_{L}^1\) and, \(Y_{U}^1\) and SSR represents the variation in \(Y\) with respect to the central regression line \(Y_i \cdot n_{L}^1\) so IC measures the degree of the variation of \(Y\) between \(Y \cdot n_{L}^1\) and \(Y_{U}^1\) that can be explained by the centre regression line \(Y \cdot n_{L}^1\). Because \(0 \leq IC \leq 1\), it follows that \(0 \leq IC \leq 1\). So, it means that the higher the IC, the better is the \(Y_i \cdot n_{L}^1\) used to represent \(Y_i\).

### 3.4 Numerical example

Clerical employees of a large financial organization included questions related to employee satisfaction with their supervisor. There was a question designed to measure the overall performance of a supervisor, as well as questions that were related to specific activities involving interaction between supervisor and employee. An exploratory study has undertaken to try to explain the relationship between specific supervisor characteristics and overall satisfaction with supervisors as perceived by the employees. The response variable is overall rating of job being done by supervisor \((Y)\), and explanatory variables are handles employee complaints \((X_1)\), Does not allow special privileges \((X_2)\), opportunity to learn new things \((X_3)\), Too critical of poor performances \((X_4)\), Rate of advancing to better jobs \((X_5)\). The data set is taken from [19], given below.

**Table 1:** Supervisor Performance Data

<table>
<thead>
<tr>
<th>(Y)</th>
<th>(X_1)</th>
<th>(X_2)</th>
<th>(X_3)</th>
<th>(X_4)</th>
<th>(X_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>90</td>
<td>50</td>
<td>72</td>
<td>54</td>
<td>36</td>
</tr>
<tr>
<td>74</td>
<td>85</td>
<td>64</td>
<td>69</td>
<td>79</td>
<td>63</td>
</tr>
<tr>
<td>65</td>
<td>60</td>
<td>65</td>
<td>75</td>
<td>80</td>
<td>60</td>
</tr>
<tr>
<td>65</td>
<td>70</td>
<td>46</td>
<td>57</td>
<td>85</td>
<td>46</td>
</tr>
<tr>
<td>50</td>
<td>58</td>
<td>68</td>
<td>54</td>
<td>78</td>
<td>52</td>
</tr>
<tr>
<td>50</td>
<td>40</td>
<td>33</td>
<td>34</td>
<td>64</td>
<td>33</td>
</tr>
<tr>
<td>64</td>
<td>61</td>
<td>52</td>
<td>60</td>
<td>80</td>
<td>41</td>
</tr>
<tr>
<td>53</td>
<td>66</td>
<td>52</td>
<td>50</td>
<td>80</td>
<td>37</td>
</tr>
<tr>
<td>40</td>
<td>37</td>
<td>42</td>
<td>58</td>
<td>57</td>
<td>49</td>
</tr>
<tr>
<td>63</td>
<td>54</td>
<td>42</td>
<td>48</td>
<td>75</td>
<td>33</td>
</tr>
<tr>
<td>66</td>
<td>77</td>
<td>66</td>
<td>63</td>
<td>76</td>
<td>72</td>
</tr>
<tr>
<td>78</td>
<td>75</td>
<td>58</td>
<td>74</td>
<td>78</td>
<td>49</td>
</tr>
<tr>
<td>48</td>
<td>57</td>
<td>44</td>
<td>45</td>
<td>83</td>
<td>38</td>
</tr>
<tr>
<td>85</td>
<td>85</td>
<td>71</td>
<td>71</td>
<td>74</td>
<td>55</td>
</tr>
<tr>
<td>82</td>
<td>82</td>
<td>39</td>
<td>59</td>
<td>78</td>
<td>39</td>
</tr>
</tbody>
</table>
By using the equation (6) the parameters are estimated using MATLAB software as and multiple linear regression equation is obtained as follows:

\[ Y = 8.71 + 0.640X_1 - 0.211X_2 + 0.459X_3 + 0.056X_4 - 0.156X_5 \]  

(21)

By using equation (7), to determine coefficient of determination of multiple regression model is \( R^2 = 83.8\% \) which indicate that these five variables causes 83.8\% variability in the Overall rating of job being done by supervisor.

By using equation (12) the parameters are estimated using MATLAB software as and multiple linear regression equation is obtained as follows:

\[ Y = \left( 6.3969, 89.3259 \right) + \left( 0.6269, 0.0987 \right)X_1 + \left( -0.1632, 0.6580 \right)X_2 + \left( 0.6790, -0.6170 \right)X_3 + \left( -0.0066, -0.7288 \right)X_4 + \left( -0.3133, -0.5573 \right)X_5 \]  

(24)

And also by using equation (20), to determine coefficient of determination of multiple fuzzy regression model is \( R^2 = 88.08\% \) which indicate that these five variables causes 88.08\% variability in the Overall rating of job being done by supervisor.

By using equation (13), the parameters are estimated as follows:

\[ \bar{Y} = 8.71 + 0.640X_1 - 0.211X_2 + 0.459X_3 + 0.056X_4 - 0.156X_5 \]

\[ Y = \left[ 8.71 + 0.640X_1 - 0.211X_2 + 0.459X_3 + 0.056X_4 - 0.156X_5 \right] \]

which indicate that these five variables causes 83.8\% and 88.08\% variability in the Overall rating of job being done by supervisor.

4. Conclusion

In this paper least square regression analysis and multiple fuzzy regression model are compared using coefficient of determination for goodness of fit. The problem under study shows that the fuzzy multiple regression model is performing better than the least square method. Therefore, multiple fuzzy regression models can produce better prediction as compared to least square method.

5. References