FUZZY LINEAR REGRESSION

Amrender Kumar
Indian Agricultural Statistics Research Institute, New Delhi-11012
akjha@iasri.res.in

1. Introduction

For many years statistical linear regression has been used in almost every field of science. The purpose of regression analysis is to explain the variation of a dependent variable Y in terms of the variation of explanatory variables X as \( Y = f(X) \) where \( f(X) \) is a linear function. The use of statistical linear regression is bounded by some strict assumptions about the given data, that is, the unobserved error term are mutually independent and identically distributed. As a result, the statistical regression model can be applied only if the given data are distributed according to a statistical model, and the relation between \( x \) and \( y \) is crisp, if linguistic data are obtained, symbolic numbers are used to represent qualitative terms, for example, number 4 for “excellent”, 3 for “very good”, 2 for “good”, and 1 for “fair”. In many real world problems, an oversimplification of data could leave out important information for regression models. Some observations can be described only in linguistic terms (such as fair, good, and excellent). For such data, fuzzy set theory provides a means for modelling such linguistic variables utilizing fuzzy membership functions. Fuzzy regression was deal with fuzzy data. Regression is based on probability theory whereas the fuzzy regression are based on possibility theory & fuzzy set theory.

Zadeh (1965) describes the fuzzy uncertainty with ambiguity and vagueness and introduces the theory of fuzzy to build such a system as needed to deal with ambiguous and vague sentences or information. Tanaka et al. (1982) explains fuzzy uncertainty of dependent variables with the fuzziness of response functions or regression coefficients in regression model and introduces initially the fuzzy regression model. The fuzzy regression model may be roughly classified by conditions of independent and dependent variables into three categories, as follows:

(i) Input and output data are both non-fuzzy number
(ii) Input data is non-fuzzy number but output data is fuzzy number
(iii) Input and output data are both fuzzy number.

2. Fuzzy linear regression methodology:

Fuzzy linear regression (FLR) is a fuzzy type of classical regression analysis in which some elements of the model are represented by fuzzy numbers. It is used in evaluating the functional relationship between the dependent and independent variables in a fuzzy environment. Fuzzy linear regression (FLR) was first introduced by Tanaka et al. (1982), some of the strict assumptions of the statistical model are relaxed. The basic model assumes a fuzzy linear function as

\[
\hat{Y} = \hat{A}_0 X_0 + \hat{A}_1 X_1 + \cdots + \hat{A}_N X_N = \hat{A} X
\]

where \( \hat{X} = [X_0, X_1, \ldots, X_N]^T \) is a vector of independent variables; \( \hat{A} = [\hat{A}_0, \hat{A}_1, \ldots, \hat{A}_N]^T \) is a vector of fuzzy coefficients presented in the form of symmetric triangular fuzzy numbers denoted by \( \hat{A}_j = (\alpha_j, \epsilon_j, \delta_j) \) with its membership function described as


\[ u_{x_j}(a_j) = \begin{cases} 
1 - \frac{|\alpha_j - a_j|}{c_j}, & \alpha_j - a_j \ll a_j \ll \alpha_j + a_j \forall j = 1, 2, 3, \ldots N \\
0, & \text{otherwise} 
\end{cases} \]

where \( \alpha_j \) is its central value and \( c_j \) is the spread value. Therefore, formula (1) can be rewritten as

\[ \hat{Y}_i = (\alpha_0, c_0)X_0 + (\alpha_1, c_1)X_1 + \ldots + (\alpha_N, c_N)X_N \quad (3) \]

The above fuzzy regression analysis assumes the crisp input and output data, while the relation between the input and output data is denoted by a fuzzy function of which the distribution of the parameter is a possibility function. The membership function of fuzzy number \( \hat{Y}_i \) is given as

\[ u_{x_j}(a_j) = \begin{cases} 
1 - \frac{|\alpha_j - a_j|}{c_j|X|}, & X \neq 0 \\
1, & X = 0, Y = 0 \forall j = 1, 2, 3, \ldots M \\
0, & X = 0, Y = 0 \quad (4) 
\end{cases} \]

where \( c^F = [c_0, c_1, \ldots, c_N] \) and each value of dependent variable can be estimated as a fuzzy number \( \hat{Y}_i = (Y_i^L, Y_i^U) \) where the lower bound \( \hat{Y}_i \) is \( Y_i^L = \sum_{j=0}^{N}(\alpha_j - a_j)X_{ij} \); the central value of \( \hat{Y}_i \) is \( Y_i^{U^C} = \sum_{j=0}^{N}a_jX_{ij} \) and the upper bound of \( \hat{Y}_i \) is \( Y_i^U = \sum_{j=0}^{N}(\alpha_j + a_j)X_{ij} \).

In order to get the fuzzy regression with minimized fuzziness, the objective function is to minimize the total spread of the fuzzy number \( \hat{Y}_i \) as

\[ \text{Min} c^F|X| = \text{Min} \left( \sum_{j=0}^{N} |X_{ij}| \right) \quad (5) \]

and the constraints require that each observation \( Y_i \) has the least \( h \) degree of belonging to \( \hat{Y}_i \).

This leads to the following linear programming problem

\[ \text{Min} \left( \sum_{j=0}^{N} |X_{ij}| \right) \]

such that

\[ \sum_{j=0}^{N} \alpha_j x_{ij} + (1 - h) \sum_{j=0}^{N} c_j |x_{ij}| \gg Y_i \]

\[ \sum_{j=0}^{N} \alpha_j x_{ij} - (1 - h) \sum_{j=0}^{N} c_j |x_{ij}| \ll Y_i \]

90
Since the fuzzy regression analysis can be applied to many real-life problems in which the strict assumptions of classical regression analysis cannot be satisfied, there are many researchers devoted to the field of fuzzy linear regression. There are basically, three different approaches of fuzzy regression, first, a fuzzy regression method that is based on minimizing fuzziness for model fitting. Second, fuzzy regression using least squares of errors as a criterion. Third, fuzzy regression is implemented by using interval analysis.

**Exercise**

The following data pairs of \((X_i : Y_i; i=1, 2, \ldots 8)\) are used to demonstrate the crisp X and crisp Y data case [Chang and Ayyub (2001)]

<table>
<thead>
<tr>
<th>X</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
<th>12</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>14</td>
<td>16</td>
<td>14</td>
<td>18</td>
<td>18</td>
<td>22</td>
<td>18</td>
<td>22</td>
</tr>
</tbody>
</table>

The interval coefficient model for the given data is

\[
Y(x_j) = A_0 + A_1 X_j
\]

\[
Y(x_j) = < a_{0c}, a_{0w} > + < a_{1c}, a_{1w} > X_j
\]

The fitting of this model under fuzzy framework amounts to solving the following LP problem by minimizing

\[
\text{Min } \sum_{j=1}^{m} (a_{0w} + a_{1w} |x_{ij}|) \quad \text{or } \min 8a_{0w} + (2 + 4 + \ldots + 16)a_{1w}
\]

Subject to constraints

\[
(a_{0c} + a_{1c} x_j) - (a_{0w} + a_{1w} |x_{ij}|) \leq y_j
\]

\[
(a_{0c} + a_{1c} x_j) + (a_{0w} + a_{1w} |x_{ij}|) \geq y_j
\]

\[
a_{1w} \geq 0
\]

\[
a_{0w} \geq 0
\]

\[
j = 1, 2, \ldots, 8
\]

In this case, there are \((8 + 8 + 2)\) i.e 18 constraints equations with four unknown parameters and 16 slack and surplus variables, This can be written as
a_{0c} + 2a_{1c} - a_{0w} - 2a_{1w} \leq 14

\ldots

a_{0c} + 16a_{1c} - a_{0w} - 16a_{1w} \leq 30
a_{0c} + 2a_{1c} + a_{0w} + 2a_{1w} \geq 14

\ldots

a_{0c} + 16a_{1c} + a_{0w} + 16a_{1w} \geq 30
a_{0w} \geq 0 \text{ and } a_{1w} \geq 0

By solving this LP problem (using SAS-Proc LP), the fitted model with interval coefficients

Y = (12.00, 1.00) + (0.63, 0.13)X
Y = <a_{0c}, a_{0w}> + <a_{1c}, a_{1w}>X

The form of simple regression model is

Y = 12.93 + 0.54X

The comparison of interval coefficient method with regression model are given below

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Interval coefficient method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>Upper</td>
</tr>
<tr>
<td>8.77</td>
<td>19.22</td>
</tr>
<tr>
<td>10.12</td>
<td>20.02</td>
</tr>
<tr>
<td>11.38</td>
<td>20.90</td>
</tr>
<tr>
<td>12.54</td>
<td>21.87</td>
</tr>
<tr>
<td>13.60</td>
<td>22.95</td>
</tr>
<tr>
<td>14.59</td>
<td>24.11</td>
</tr>
<tr>
<td>15.47</td>
<td>25.37</td>
</tr>
<tr>
<td>16.27</td>
<td>26.72</td>
</tr>
</tbody>
</table>

References

