A least-absolutes approach to multiple fuzzy regression

Chachi, Jalal
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 8415683111, Iran
E-mail: J.Chachi@math.iut.ac.ir

Taheri, S. Mahmoud[1,2]
[1]*Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 8415683111, Iran*
[2]*Department of Statistics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad, Mashhad, Iran*
E-mail: Taheri@cc.iut.ac.ir

A least-absolutes approach to multiple fuzzy regression modeling is introduced and investigated for the case of crisp input-fuzzy output data, by using the generalized Hausdorff-metric. A comparative study, based on three data set, including a real agricultural data set, and three well-known goodness of fit indices indicate that the proposed approach has certain advantages to some common methods in fuzzy regression modeling.

Keywords: Fuzzy regression; Hausdorff-metric; Imprecise data; Least-absolutes method.

1 Introduction

Fuzzy regression models were proposed to model the relationship between variables, when the data available are imprecise (fuzzy) quantities and/or the relationship between variables are fuzzy. The regression analysis with fuzzy data has been previously treated from different points of view as well as considering different kinds of input/output data. It has wide applications in many areas including agriculture, biology, business, economics, health sciences, and engineering. See, for example, [12, 14, 17] for some references and recent developments.

In this article, we develop a least-absolutes fuzzy regression model to handle the functional dependence of crisp input-fuzzy output variables. To do this, using the generalized Hausdorff-metric between fuzzy numbers as well as a linear programming method, we estimate the coefficients of the fuzzy regression model. The rest of this paper is organized as follows: The next section reviews some preliminary definitions and results, that used in the sequel. Section 3 provides a new multiple fuzzy least-absolutes regression model for crisp input-fuzzy output variables and in Section 4 some capability indices are provided to evaluate the goodness of fit of the proposed model. In Section 5, by using three numerical examples, we provide some comparative studies to show the performance of the proposed method. Finally, in Section 6 some concluding remarks are made.

2 Preliminaries

A fuzzy set $\tilde{A}$ on the universal set $\mathbb{X}$ is described by its membership function $\tilde{A}(x) : \mathbb{X} \to [0, 1]$. Hereafter we assume that $\mathbb{X} = \mathbb{R}$, the real line. The crisp set $A_\alpha = \{x \in \mathbb{R} : \tilde{A}(x) \geq \alpha\}$, $\alpha \in (0, 1]$, is called the α-cut of $\tilde{A}$, and for $\alpha = 0$, A_0 is the closure of set $\{x \in \mathbb{R} : \tilde{A}(x) > 0\}$.

A specific type of fuzzy set on $\mathbb{R}$ is triangular fuzzy number $\tilde{N} = (n, l, r)_T$ with central value $n \in \mathbb{R}$, left and right spreads $l, r \in \mathbb{R}^+$, for which the membership function is as follows [19]

$$\tilde{N}(x) = \frac{x-(n-l)}{l} I_{[n-l,n]}(x) + \frac{(n+r)-x}{r} I_{(n,n+r]}(x), \qquad x \in \mathbb{R}.$$

The α-cuts of triangular fuzzy number $\tilde{N}$ are $N_\alpha = [n-(1-\alpha)l, n+(1-\alpha)r]$. Triangular fuzzy number $\tilde{N}$ with $l = r = \lambda$ is called symmetric triangular fuzzy number and is abbreviated by $\tilde{N} = (n, \lambda)_T$.

A well-known result of fuzzy arithmetic is that if $\tilde{M} = (m, l_m, r_m)_T$, $\tilde{N} = (n, l_n, r_n)_T$, and λ is a real number, then

$$\begin{aligned}\lambda \otimes \tilde{M} &= \begin{cases} (\lambda n, \lambda l_m, \lambda r_m)_T & \text{if} \quad \lambda > 0, \\ \mathcal{I}_{\{0\}} & \text{if} \quad \lambda = 0, \\ (\lambda m, |\lambda| r_m, |\lambda| l_m)_T & \text{if} \quad \lambda < 0, \end{cases} \\ \tilde{M} \oplus \tilde{N} &= (m + n, l_m + l_n, r_m + r_n)_T.\end{aligned}$$

where $\mathcal{I}_{\{0\}}$ stands for the indicator function of the crisp zero [19].

A distance between fuzzy numbers: On the family of all fuzzy numbers several metrics can be defined (see [16]). But the generalized Hausdorff-metric, which is used in this paper, not only fulfill many good properties, but is also easy to calculate and handle when used for statistical purposes [2].

Definition 2.1 ([2]). *The generalized Hausdorff-metric between fuzzy numbers $\tilde{A}$ and $\tilde{B}$ is defined by*

$$\mathcal{D}_p(\tilde{A}, \tilde{B}) = \begin{cases} \left(\int_0^1 [d_H(A_\alpha, B_\alpha)]^p d\alpha\right)^{\frac{1}{p}} & \textit{if} \quad p \in [1, \infty), \\ \sup_{\alpha \in [0,1]} d_H(A_\alpha, B_\alpha) & \textit{if} \quad p = \infty, \end{cases}$$

where $d_H(A_\alpha, B_\alpha)$ is the Hausdorff-metric between crisp sets A_α and B_α, given by

$$d_H(A_\alpha, B_\alpha) = \max\{\sup_{b \in B_\alpha} \inf_{a \in A_\alpha} |a - b|, \sup_{a \in A_\alpha} \inf_{b \in B_\alpha} |a - b|\}.$$

In special case, if $I_1 = [a_1, a_2]$ and $I_2 = [b_1, b_2]$ are two intervals, then

$$d_H(I_1, I_2) = \max\{|a_1 - b_1|, |a_2 - b_2|\} = |midI_1 - midI_2| + |sprI_1 - sprI_2|,$$

where $midI_1 = \frac{a_1 + a_2}{2}$ and $sprI_1 = \frac{a_2 - a_1}{2}$ [16]. The generalized Hausdorff-metric between $\tilde{M} = (m, \lambda_m)_T$ and $\tilde{N} = (n, \lambda_n)_T$ is, then

$$\mathcal{D}_1(\tilde{M}, \tilde{N}) = |m - n| + 0.5|\lambda_m - \lambda_n|, \qquad \mathcal{D}_\infty(\tilde{M}, \tilde{N}) = |m - n| + |\lambda_m - \lambda_n|.$$

3 The proposed model

Assume that the observed data on n statistical units are denoted as $(\tilde{y}_1, \mathbf{x}_1), \ldots, (\tilde{y}_n, \mathbf{x}_n)$, where $\tilde{\mathbf{y}}_{n\times 1} = [\tilde{y}_1, \ldots, \tilde{y}_n]^t$ is a vector of symmetric triangular fuzzy numbers, i.e. $\tilde{y}_i = (y_i, s_i)_T$ $(i = 1, \ldots, n)$, which determines the fuzzy observed of the dependent variable, and $\mathbf{x}_i = [x_{i0}, x_{i1}, \ldots, x_{ik}] \in \mathbb{R}^{k+1}$ $(i = 1, \ldots, n; k < n; x_{i0} = 1)$, forms the $(k + 1)$-dimensional vector of crisp observed independent variables. Based on such a data set, we consider the following model between $\tilde{\mathbf{y}}_{n\times 1}$ and $\mathbf{X}_{n\times(k+1)}$

$$\begin{aligned}\tilde{\mathbf{y}}_{n\times 1} &= \mathbf{X}_{n\times(k+1)} \otimes \tilde{\beta}_{(k+1)\times 1} \\ \begin{bmatrix} \hat{\tilde{y}}_1 \\ \vdots \\ \hat{\tilde{y}}_n \end{bmatrix} &= \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} \otimes \begin{bmatrix} (\beta_0, \sigma_0)_T \\ \vdots \\ (\beta_k, \sigma_k)_T \end{bmatrix} = \begin{bmatrix} (\sum_{j=0}^k x_{1j}\beta_j, \sum_{j=0}^k |x_{1j}|\sigma_j)_T \\ \vdots \\ (\sum_{j=0}^k x_{nj}\beta_j, \sum_{j=0}^k |x_{nj}|\sigma_j)_T \end{bmatrix}.\end{aligned}$$

The procedure for estimating the parameter $\tilde{\beta}_{(k+1)\times 1}$ is based on minimizing the total difference between the observed values of the response variable, i.e. $\tilde{\mathbf{y}}_{n\times 1}$, and its theoretical counterpart, i.e. $\hat{\tilde{\mathbf{y}}}_{n\times 1} = \mathbf{X}_{n\times(k+1)} \otimes \hat{\tilde{\beta}}_{(k+1)\times 1}$, with respect to the distance $\mathcal{D}_1$. Thus, consider the least-absolutes optimization problem as follows

$$(1) \qquad \min_{\tilde{\beta}_{(k+1)\times 1}} \mathcal{D}_1(\tilde{\mathbf{y}}, \mathbf{X} \otimes \tilde{\beta}) = \min_{\tilde{\beta}_{(k+1)\times 1}} \sum_{i=1}^n |y_i - \sum_{j=0}^k x_{ij}\beta_j| + 0.5 \sum_{i=1}^n |s_i - \sum_{j=0}^k |x_{ij}|\sigma_j|,$$

which leads to a constrained non-linear programming problem. The minimization of $\mathcal{D}_1$ over $\mathbb{R}^{k+1} \times {\mathbb{R}^+}^{k+1}$ can separately be solved: once for all possible candidates for $\underline{\beta} = [\beta_0, \beta_1, \ldots, \beta_k] \in \mathbb{R}^{k+1}$, and then for all possible candidates for $\underline{\sigma} = [\sigma_0, \sigma_1, \ldots, \sigma_k] \in {\mathbb{R}^+}^{k+1}$, which are the center and spread values of the fuzzy coefficients $\tilde{\beta}$, respectively. Thus, the optimization problem (1) can be rewritten as the following two sub-optimization non-linear programming problems

$$\min_{\tilde{\beta}_{(k+1)\times 1}} \mathcal{D}_1(\tilde{\mathbf{y}}, \mathbf{X} \otimes \tilde{\beta}) \equiv \begin{cases} (A) & \min_{\underline{\beta}} \sum_{i=1}^{n} \left| y_i - \sum_{j=0}^{k} x_{ij}\beta_j \right|, \\ (B) & 0.5 \min_{\underline{\sigma}} \sum_{i=1}^{n} \left| s_i - \sum_{j=0}^{k} |x_{ij}|\sigma_j \right|. \end{cases}$$

In order to simplify the above optimization problem, we show how by introducing additional variables a linear programming can handle this optimization problem. First, we consider the sub-optimization problem (A). Let ε_i^+ and ε_i^-, $i = 1, \ldots, n$, represent two nonnegative variables such as

$$|y_i - \sum_{j=0}^{k} x_{ij}\beta_j| = \varepsilon_i^+ + \varepsilon_i^-, \qquad y_i - \sum_{j=0}^{k} x_{ij}\beta_j = \varepsilon_i^+ - \varepsilon_i^-.$$

Then we construct the following matrices

$$\begin{array}{rcl} \underline{\varepsilon}^+_{n\times 1} & = & [\varepsilon_1^+, \ldots, \varepsilon_n^+]^t, \\ \underline{\varepsilon}^-_{n\times 1} & = & [\varepsilon_1^-, \ldots, \varepsilon_n^-]^t, \\ \underline{e}_{(k+1+2n)\times 1} & = & [\underline{\beta}^t_{1\times(k+1)} \ \underline{\varepsilon}^{+t}_{n\times 1} \ \underline{\varepsilon}^{-t}_{n\times 1}]^t, \end{array} \qquad \begin{array}{rcl} \mathbf{H}_{n\times(k+1+2n)} & = & [\mathbf{X}_{n\times(k+1)} \ \mathbf{I}_{n\times n} \ -\mathbf{I}_{n\times n}], \\ \underline{h}_{(k+1+2n)\times 1} & = & [\underline{0}_{1\times(k+1)} \ \underline{J}_{1\times 2n}]^t, \\ & & . \end{array}$$

where $\mathbf{I}_{n\times n}$ is a diagonal identity matrix of order n and $\underline{J}$ denotes the $(n\times 1)$-vector of 1's. Finally, the non-linear optimization problem (A) becomes equivalent to the following linear optimization problem

$$\begin{array}{ll} \min\limits_{\underline{e}_{(k+1+2n)\times 1}} & \underline{h}^t_{(k+1+2n)\times 1}\underline{e}_{(k+1+2n)\times 1} \\ s.t. & \mathbf{H}_{n\times(k+1+2n)}\underline{e}_{(k+1+2n)\times 1} = \mathbf{y}_{n\times 1}, \\ & \underline{\varepsilon}^+_{n\times 1} \in {\mathbb{R}^+}^n, \ \underline{\varepsilon}^-_{n\times 1} \in {\mathbb{R}^+}^n, \ \underline{\beta}_{1\times(k+1)} \in \mathbb{R}^{k+1}, \end{array}$$

which can be solved by a common software such as Lingo [7]. The same method may be easily used to solve the other sub-optimization problem (B). In this case, we replace $\underline{\beta}_{1\times(k+1)} \in \mathbb{R}^{k+1}$ with $\underline{\sigma}_{1\times(k+1)} \in {\mathbb{R}^+}^{k+1}$, and all variables are assumed to be nonnegative.

4 Goodness of fit criteria

To evaluate the performance of a fuzzy regression model, Kim and Bishu [6] defined the following error of estimation (see also [1, 8])

$$E_1(i) = \frac{\int |\tilde{y}_i(x) - \widehat{\tilde{y}}_i(x)|\, dx}{\int \tilde{y}_i(x)\, dx}.$$

The following index is also used by some authors to evaluate fuzzy regression models [1, 5, 8]

$$E_2(i) = \int |\tilde{y}_i(x) - \widehat{\tilde{y}}_i(x)|\, dx.$$

The following similarity measure was also used for evaluating the performance of a fuzzy regression model (see [4, 8, 15]), such as

$$S(i) = \frac{\int \min\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}\, dx}{\int \max\{\tilde{y}_i(x), \widehat{\tilde{y}}_i(x)\}\, dx}.$$

The range of E_1 and E_2 is $[0, \infty)$, while that of S is $[0, 1]$. Thus, in order to compare the indices, we modify the range of E_1 and E_2 by taking into account $G_1(i) = \frac{1}{1+E_1(i)}$ and $G_2(i) = \frac{1}{1+E_2(i)}$. In practice, we use $G_1 = \frac{1}{n}\sum_{i=1}^{n} G_1(i)$, $G_2 = \frac{1}{n}\sum_{i=1}^{n} G_2(i)$, and $S = \frac{1}{n}\sum_{i=1}^{n} S(i)$ to evaluate the goodness of fit of the models.

Table 1: Performance of different models in Example 5.1

Proposed by	S	G_1	G_2
Ferraro et al. [3] $\widehat{\tilde{y}} = (4.95 + 1.71x, \exp(0.6098 + 0.0742x))_T$	0.4079	0.3620	0.5534
Modarres et al. [10] $\widehat{\tilde{y}} = (4.82 + 1.66x, 1.84 + 0.16x)_T$	0.4670	0.4387	0.6034
Nasrabadi and Nasrabadi [11] $\widehat{\tilde{y}} = (4.6812 + 1.7306x, 2.3221)_T$	0.4408	0.3958	0.5765
Xu and Li [18] $\widehat{\tilde{y}} = (4.95 + 1.71x, 1.84 + 0.16x)_T$	0.4087	0.3619	0.5533
The proposed least-absolutes Model $\widehat{\tilde{y}} = (6.444 + 1.3112x, 1.8 + 0.2x)$	0.5008	0.4480	0.6209

Table 2: Performance of different models in Example 5.2

Proposed by	S	G_1	G_2
Ferraro et al. [3] $\widehat{\tilde{y}} = (-672.731 + 0.0181x, \exp\{5.9244 + 0.000002482x\})_T$	0.2135	0.0010	0.4386
Modarres et al. [10] $\widehat{\tilde{y}} = (-188.8609 + 0.01605x, 194.6657 + 0.00348x)_T$	0.2435	0.0012	0.4305
Nasrabadi and Nasrabadi [11] $\widehat{\tilde{y}} = (-1714.8222 + 0.01813x, 1675.4064)_T$	0.1834	0.0005	0.3546
Xu and Li [18] $\widehat{\tilde{y}} = (-672.731 + 0.01807x, 185.7144 + 0.00347x)_T$	0.2431	0.0010	0.4357
The proposed least-absolutes Model $\widehat{\tilde{y}} = (-254.7958 + 0.01737x, 66.4543 + 0.00336x)$	0.2352	0.0032	0.4518

5 Illustrative examples

In this section a couple of examples are given to illustrate the efficiency of the proposed fuzzy regression model with respect to some other competitive techniques.

Example 5.1 ([13]). *Consider the following crisp input-fuzzy output data given by Tanaka et al. [13]*

$$(\tilde{y}; x) = ((8.0, 1.8)_T; 1), ((6.4, 2.2)_T; 2), ((9.5, 2.6)_T; 3), ((13.5, 2.6)_T; 4), ((13.0, 2.4)_T; 5).$$

By applying the proposed approach described in Section 3, the fuzzy regression model is derived as $\widehat{\tilde{y}} = (6.4440, 1.8)_T \oplus (1.3112, 0.2)_T x$*. A summary of the results of some other techniques, including their models as well as their performances, are given in Table 1. The results indicate that, conserving all three criteria our proposed method yields a better model than the other ones.*

Example 5.2 ([3]). *In this example we are interested in analyzing the dependence relationship of the Retail Trade Sales of the U.S. in 2002 by kind of business on the number of employees. The Retail Trade Sales has been in the period January 2002 through December 2002 (see the data set given in Table 2 by Ferraro et al. [3]). The results of fitting in our proposed model and in some other models are presented in Table 2. The model obtained from the least-absolutes method performs better than the*

Table 3: Performance of different models in Example 5.3

Proposed by	S	G_1	G_2
Ferraro et al. [3]	0.5830	0.4043	0.6826
$\widehat{y} = 70.5873 + 6.9253x_1 - 0.5628x_2 - 0.3958x_3$			
$\widehat{l} = \exp(1.7897 + 0.1538x_1 - 0.0120x_2 - 0.0028x_3)$			
Modarres et al. [10]	0.4412	0.2814	0.5838
$\widehat{y} = 69.3354 + 9.5098x_1 - 0.6193x_2 - 0.3972x_3$			
$\widehat{l} = 7.0587 + 0.6925x_1 - 0.0563x_2 - 0.0396x_3$			
Nasrabadi-Nasrabadi [11]	0.3663	0.2161	0.5283
$\widehat{y} = 74.0638 + 9.0493x_1 - 0.6855x_2 - 0.4406x_3$			
$\widehat{l} = 4.7053$			
Xu-Li [18], and Mohammadi-Taheri [9]	0.5727	0.3707	0.6696
$\widehat{y} = 70.5873 + 6.9253x_1 - 0.5628x_2 - 0.3958x_3$			
$\widehat{l} = 3.19688 + 1.0407x_1$			
The proposed least-absolutes Model	0.6014	0.4462	0.7031
$\widehat{y} = 72.0466 + 6.4557x_1 - 0.5523x_2 - 0.4278x_3$			
$\widehat{l} = 1.0240 + 0.7163x_1 + 0.0599x_3$			

models given in Table 2. The results show that, while the index S for models given in [10, 11] is close to that of the proposed model, the proposed model has, however, better performance concerning G_1 and G_1.

Example 5.3 ([9]). *One of the classical problems in soil sciences is the measurement of physical, chemical, and biological soil properties. Based on a study in a part of Silakhor plain (situated in the province of Lorestan, west of Iran), different soil physical and chemical properties were measured using standard procedures [9]. But, due to some impreciseness in experimental environment, the observed data were reported as fuzzy sets (see the data set given in Table 1 by Mohammadi and Tageri [9]). The data set show soils saturated by water* `(SP)` *$(\tilde{y})$, as symmetric triangular fuzzy observations of the dependent variable, organic matter content* `(OM)` *(x_1), sand content percentage* `(SAND)` *(x_2), and silt* `(SILT)` *(x_3) as the crisp observations of the independent variables. Based on such data set, we wish to model the relationship between the response variable* `SP` *and explanatory variables* `OM`, `SAND`, *and* `SILT` *by a fuzzy regression model. The results of the performance of different methods as well as their models are given in Table 3, which favor the proposed approach in comparison with the other methods.*

6 Conclusions

A least-absolutes fuzzy multiple linear regression model was developed by using the generalized Hausdorff-metric on the space of fuzzy numbers. Using this model we can deal with multiple linear regression problem with crisp input-fuzzy output observations. The solution of the proposed methodology relies on a non-linear optimization problem, which was also translated to a linear optimization problem, making the computations of the proposed method very simple. The number of variables in the optimization program is the number of regression coefficients to be determined.

In order to evaluate the goodness of fit of the proposed model, some indices were employed. Upon these indices, the results of numerical examples showed that the proposed method is able to determine the regression coefficients with better explanatory power.

Acknowledgment

The authors would like to thank the Research Affairs at Isfahan University of Technology for partially supporting of this work.

References

[1] S.-P. Chen, J.-F. Dang, A variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy, Inf. Sci. 178 (2008) 3973-3988.

[2] P. Diamond, P. Kloeden, Metric Spaces of Fuzzy Sets, World Scientific, Singapore, (1994).

[3] M.B. Ferraro, R. Coppi, G. González-Rodríguez, A. Colubi, A linear regression model for imprecise response, Int. J. Approx. Reason. 51 (2010) 759-770.

[4] M. Hojati, C.R. Bector, K. Smimou, A simple method for computation of fuzzy linear regression, European J. Ope. Res. 166 (2005) 172-184.

[5] C. Kao, C.-L. Chyu, Least-squares estimates in fuzzy regression analysis, European J. Oper. Res. 148 (2003) 426-435.

[6] B. Kim, R.R. Bishu, Evaluation of fuzzy linear regression models by comparison membership function, Fuzzy Set Syst. 100 (1998) 343-352.

[7] Lingo User's Guide, LINDO Systems Inc., Chicago, (1999).

[8] J. Lu, R. Wang, An enhanced fuzzy linear regression model with more flexible spreads, Fuzzy Set Syst. 160 (2009) 2505-2523.

[9] J. Mohammadi, S.M. Taheri, Pedomodels fitting with fuzzy least squares regression, Iran. J. Fuzzy Syst. 1 (2004) 45-62.

[10] M. Modarres, E. Nasrabadi, M.M. Nasrabadi, Fuzzy linear regression models with least square errors, Appl. Math. Comput. 163 (2005) 977-989.

[11] M.M. Nasrabadi, E. Nasrabadi, A mathematical-programming approach to fuzzy linear regression analysis, Appl. Math. Comput. 155 (2004) 873-881.

[12] S. Pourahmad, S.M.T. Ayatollahi and S.M. Taheri, Fuzzy logistic regression: A new possibilistic model and its application in clinical vague status, Iranian J. Fuzzy Syst. 8 (2011) 1-17.

[13] H. Tanaka, I. Hayashi, J. Watada, Possibilistic linear regression analysis for fuzzy data, European J. Oper. Res. 40 (1989) 389-396.

[14] S.M. Taheri, Trends in fuzzy statistics, Aust. J. Stat. 32 (2003) 239-257.

[15] S.M. Taheri, M. Kelkinnama, Fuzzy Least Absolutes Regression, Proc. 4th Int. IEEE Conf. Intell. Syst. 11 (2008) 55-58.

[16] W. Trutschnig, G. González-Rodríguez, A. Colubi, M. Ángeles Gil, A new family of metrics for compact, convex (fuzzy) sets based on a generalized concept of mid and spread, Inf. Sci. 179 (2009) 3964-3972.

[17] R. Viertl, Statistical Methods for Fuzzy Data, John Wiley and Sons, Chichester, (2011).

[18] R. Xu, C. Li, Multidimensional least-squares fitting with a fuzzy model, Fuzzy Set Syst. 119 (2001) 215-223.

[19] H.J. Zimmermann, Fuzzy Set Theory and Its Applications, 4th ed., Kluwer Nihoff, Boston, (2001).