Applied Soft Computing 46 (2016) 104–114

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Error measures for fuzzy linear regression: Monte Carlo simulation approach

Duygu İçen[a,*], Haydar Demirhan[a,b]

[a] Hacettepe University, Faculty of Science, Department of Statistics, Ankara, Turkey
[b] RMIT University, School of Science, Mathematical and Geospatial Sciences, Melbourne, Australia

ARTICLE INFO

Article history:
Received 10 July 2015
Received in revised form 5 March 2016
Accepted 7 April 2016
Available online 27 April 2016

Keywords:
Monte Carlo
Fuzzy linear regression
Random vectors
Random fuzzy vectors

ABSTRACT

The focus of this study is to use Monte Carlo method in fuzzy linear regression. The purpose of the study is to figure out the appropriate error measures for the estimation of fuzzy linear regression model parameters with Monte Carlo method. Since model parameters are estimated without any mathematical programming or heavy fuzzy arithmetic operations in fuzzy linear regression with Monte Carlo method. In the literature, only two error measures (E_1 and E_2) are available for the estimation of fuzzy linear regression model parameters. Additionally, accuracy of available error measures under the Monte Carlo procedure has not been evaluated. In this article, mean square error, mean percentage error, mean absolute percentage error, and symmetric mean absolute percentage error are proposed for the estimation of fuzzy linear regression model parameters with Monte Carlo method. Moreover, estimation accuracies of existing and proposed error measures are explored. Error measures are compared to each other in terms of estimation accuracy; hence, this study demonstrates that the best error measures to estimate fuzzy linear regression model parameters with Monte Carlo method are proved to be E_1, E_2, and the mean square error. One the other hand, the worst one can be given as the mean percentage error. These results would be useful to enrich the studies that have already focused on fuzzy linear regression models.

1. Introduction

Regression analysis is a statistical tool that is used to figuring out the mathematical relation between two or more quantitative variables. In the literature, most of the available regression modelling approaches are rather restrictive and their applications to real life problems require various assumptions. Therefore, new techniques have been proposed to relax some of these assumptions. Fuzzy regression is one of these techniques that attracts more attention nowadays.

After introduced by Tanaka et al. [1,2], the fuzzy regression analysis has become very popular with the introduction of fuzziness into regression. Many regression models including crisp input and fuzzy parameters, as well as fuzzy input and crisp parameters have been studied. Diamond [3] implemented regression models for crisp input and fuzzy output and fuzzy input-output. In these models, distance between fuzzy numbers was used to measure the goodness-of-fit for models. Furthermore, Näther and Körner [4] extended estimators of Tanaka et al. [1,2] with a least squares approach in the linear regression with crisp and fuzzy input and fuzzy output cases. Hong et al. [5] adopted a regression model with fuzzy input and fuzzy parameters. Additionally, Bardossy et al. [6] defined a new class of distance measures on fuzzy numbers and considered the regression model involving fuzzy input and fuzzy parameters. Peters [7], Luczynski and Matloka [8], Tanaka et al. [9], and Yen et al. [10] are some of the authors who focused on crisp input and fuzzy output regression models. D'Urso [11] carried out fuzzy linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data. Moreover, Roh et al. [12] presented a new estimation approach based on Polynomial Neural Networks for fuzzy linear regression. Recently, a generalized maximum entropy estimation approach to fuzzy regression model is introduced by Ciavolinoa and Calcagni [13].

Application areas of fuzzy linear regression analysis have been considerably improved by different approaches in recent years. For instance, the relationship between dimensions of health related quality of life and health conditions are investigated under fuzzy linear regression [14]. A new fuzzy linear regression approach for dissolved oxygen prediction is suggested by Khan and Valeo [15]. Fuzzy linear regression is also used in electricity demand forecasting by Sarkar et al. [16] and in global solar radiation prediction by Ramedani et al. [17]. Estimation of relationship between forest fires and meteorological conditions are investigated within the framework of fuzzy linear regression by Akdemir and Tiryaki [18].

* Corresponding author.

http://dx.doi.org/10.1016/j.asoc.2016.04.013
1568-4946/

Abdalla and Buckley [19–21] are the first practitioners of Monte Carlo (MC) method within the fuzzy linear regression context. In the MC method, a number of regression coefficient vectors are randomly generated; then values of dependent variables are estimated by using each generated vector, and the vector that gives minimum value to an error measure is taken as the best estimate of regression parameters. Abdalla and Buckley [19,21] used two error measures and tackled problems that are defined on the positive side of the real line. However, they did not mention which definition of the absolute value of a fuzzy number is used in the calculations for the error measures.

In this study, MC method for fuzzy linear regression analysis introduced by Abdalla and Buckley [19,21] is taken as a focal point. The definition of AbuAarqob et al. [22] is used for the absolute value of a fuzzy number and the MC method is applied to the data set of Abdalla and Buckley [19,21]. Six error terms are used in this study. Two of them have already been used by Abdalla and Buckley [19,21] and the other four different error measures that have not been previously calculated for MC method in fuzzy linear regression are used. These error measures are mean square error, mean percentage error, mean absolute percentage error, and symmetric mean absolute percentage error. In order to evaluate estimation accuracy of the new and existing error measures in fuzzy linear regression modeling, an extended simulation study is conducted over the whole real line. In the design of simulation study, two cases that fuzzy input-fuzzy output and crisp input-fuzzy output are taken into consideration. The performance of four error measures along with those used by Abdalla and Buckley [19,21] are evaluated and compared with each other. The best error measure and the one that should not be used for the estimation of fuzzy linear regression parameters are identified by MC method without using any mathematical programming or heavy fuzzy arithmetic operations.

The rest of the paper is organized as follows: some preliminaries for fuzzy numbers, random crisp vectors, and random fuzzy vectors are presented in Section 2. Brief information about fuzzy linear regression models with MC method is given in Section 3.1. Error measures proposed for the MC method are given in Section 3.2. The simulation study that compares the performances of error measures is conducted in Section 4. After the decision of the best and the worst error measures in MC method for fuzzy linear regression models, values of the error measures are calculated for the real data sets used by Abdalla and Buckley [19,21] in Section 5. Concluding remarks and some possible future perspectives are addressed Section 6.

2. Preliminaries

This section contains various definitions of fuzzy numbers and random fuzzy vectors that are defined by Abdalla and Buckley [19,21], Dubois and Prade [23], and AbuAarqob et al. [22].

Definition 2.1. A fuzzy number $\overline{A}$ is a fuzzy subset of the real line $\mathfrak{R}$. Its membership function $\mu_A(x)$ satisfies the following criteria [23]:

- α-cut set of $\mu_A(x)$ is a closed interval,
- $\exists$x such that $\mu_A(x)$=1, and
- convexity such that

$$\mu_A(\lambda x_1 + (1-\lambda)x_2) \geq \min(\mu_A(x_1), \quad \mu_A(x_2)) \text{ for } \lambda \in [0, 1],$$

where α-cut set contains all x elements that have a membership grade $\mu_A(x) \geq \alpha$.

Definition 2.2. A triangular shaped fuzzy number $\overline{A}$ is a fuzzy number whose membership function defined by three values, $a_1 < a_2 < a_3$, where the base of triangular is the interval $[a_1, a_3]$ and the vertex is $x = a_2$ [23].

Definition 2.3. The α-cut of a fuzzy number $\overline{A}$ is a non-fuzzy set defined as $\overline{A}(\alpha) = \{x \in \mathfrak{R}, \mu_A(x) \geq \alpha\}$. Hence $\overline{A}(\alpha) = [A^L(\alpha), A^U(\alpha)]$ where $A^L(\alpha) = \inf\{x \in \mathfrak{R}, \mu_A(x) \geq \alpha\}$ and $A^U(\alpha) = \sup\{x \in \mathfrak{R}, \mu_A(x) \geq \alpha\}$ [23].

Definition 2.4. The absolute value of a fuzzy number $\overline{A} \in \mathfrak{R}_F$ is a function $F : \mathfrak{R}_F \to \mathfrak{R}_F$ denoted by $F(\overline{A}) := |\overline{A}|$ with α-cut $\overline{A}(\alpha)$. From the interval analysis [24], it is known that if $I = [I^-, I^+]$, then $|I| = [\max(I^-, -I^+, 0), \max(-I^-, I^+)]$; and hence, the α-cut of $|\overline{A}|$ is given by

$$(|\overline{A}|)_\alpha = [\max(\overline{A}^-(\alpha), -\overline{A}^+(\alpha), 0), \max(-\overline{A}^-(\alpha), \overline{A}^+(\alpha))]. \quad (1)$$

From Eq. (1), the absolute value of a triangular fuzzy number is given as follows (for more information see [22,24]):

$$(|\overline{A}|)_\alpha = \begin{cases} \overline{A}(\alpha) & \text{if } \overline{A} \geq 0 \\ -\overline{A}(\alpha) & \text{if } \overline{A} \leq 0 \\ \{0, \max(-\overline{A}^-(\alpha), \overline{A}^+(\alpha))\} & \text{if } x \in (\overline{A}^-(0), \overline{A}^+(0)). \end{cases} \quad (2)$$

Definition 2.5. Random crisp vectors are defined as $\mathbf{v}_k = (v_{0k}, \ldots, v_{mk})$, elements of which are all real numbers in intervals I_i, $i = 0, 1, \ldots, m$. To obtain $\mathbf{v}_k$, firstly randomly crisp vectors $v_k = (x_{1k}, x_{2k}, \ldots, x_{mk})$ with all x_{ik} in [0, 1], $k = 1, 2, \ldots, N$ are needed to be generated. Since each x_{ik} starts out in [0,1], it is possible to put them into $I_i = [c_i, d_i]$ by $v_{ik} = c_i + (d_i - c_i)x_{ik}$, $i = 0, 1, \ldots, m$ [19,25].

Definition 2.6. Random fuzzy vectors are defined as $\overline{\mathbf{V}}_k = (\overline{V}_{0k}, \ldots, \overline{V}_{mk})$, $k = 1, 2, \ldots N$, where each $\overline{V}_{ik}$ is a triangular fuzzy number. Firstly crisp vectors $v_k = (x_{1k}, \ldots, x_{3m+3,k})$ with $x_{ik} \in$ [0, 1], $k = 1, \ldots, N$ are needed to be generated. Then first three numbers in v_k are chosen and ordered from smallest to largest. If it is assumed that $x_{3k} < x_{1k} < x_{2k}$, the first triangular fuzzy number is $\overline{V}_{0k} = (x_{3k}/x_{1k}/x_{2k})$. It is possible to continue with the next three numbers in v_k to form $\overline{V}_{ik}$, $i = 1, 2, \ldots, m$. In order to obtain $\overline{V}_{ik}$ within certain intervals, it is supposed to be in interval $I_i = [a_i, b_i]$, $i = 0, 1, 2, \ldots, m$. Since each $\overline{V}_{ik}$ starts out in [0, 1], it is possible to put it into $[a_i, b_i]$ by computing $a_i + (b_i - a_i)x_{ik}$, $i = 1, 2, \ldots, m$ (for more information see [21,25]).

3. Fuzzy linear regression

Fuzzy regression model is classified into three cases according to the type of independent and dependent variables by Choi and Buckley [26] as the following:

(I) Input and output data are both crisp.
(II) Input data is crisp and output data is fuzzy.
(III) Input and output data are both fuzzy.

The first category is considered as an ordinary regression model. Hence, Case-I is not taken into consideration in this paper. Fuzzy regression model for the second (Case-II) and third (Case-III) cases are considered.

The fuzzy linear regression model for Case-II is given as follows:

$$\overline{Y}_l = \overline{A}_0 + \overline{A}_1 x_{1l} + \overline{A}_2 x_{2l} + \cdots + \overline{A}_m x_{ml} \quad (3)$$

where $x_{1l}, \ldots, x_{ml}$ for $l = 1, \ldots, n$ are crisp numbers and $\overline{A}_0, \ldots, \overline{A}_m$ and $\overline{Y}_l$ are all triangular fuzzy numbers. Given the data, the objective is to find a combination of $\overline{A}_j, j = 1, \ldots, m$ values that makes the overall difference between estimated and observed values of