Applied Soft Computing 13 (2013) 734–742

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

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

Outlier detection in fuzzy linear regression with crisp input–output by linguistic variable view

H. Shakouri G.*, R. Nadimi

School of Industrial and Systems Engineering, College of Engineering, University of Tehran, Tehran, Iran

ARTICLE INFO

Article history:
Received 4 December 2011
Accepted 2 July 2012
Available online 3 August 2012

Keywords:
Fuzzy regression
Ordinary regression
Mathematical programming
Outlier data
Linguistic variables

ABSTRACT

Existence of outlier data among the observation data leads to inaccurate results in modeling. Detection to omit or lessen the impact of such data has a significant effect to make corrections in a model. Either elimination or reduction of the outlier data influence is two ways to prevent their negative effect on the modeling. Both approaches of elimination and impact reduction are taken into account in dealing with the mentioned problem in fuzzy regression, where both the input and output data are non-fuzzy. The main idea is considered based on linguistic variables and possibility concept as well as ordinary regression to deal with the outlier data. Several examples as well as a case study are put into effect to show the capability of proposed approach.

1. Introduction

Statistical regression is a common way to find a crisp relationship between dependent variable (y) and independent variables ($\boldsymbol{x}$). An ordinary regression analysis is indeed an explanation for the variation of the former in terms of the latter in which probability distribution is used to find its parameters. However, the *possibility theory* [2] is applied to extract a fuzzy relationship between the input and output data, when fuzzy regression is considered. This relationship can lead to an inaccurate model with the existence of outlier data. Detection and omission of outlier data is an important process that may prevent from obtaining untrustworthy models.

Fuzzy Linear Regression (FLR) analysis is introduced by Tanaka et al. [1], who established his idea on the basis of the possibility theory while until yet. However, many revisions have been proposed on fuzzy regression models. Linear programming method [3–6], and the least-squares model [7–11] are the two classes of solutions that are currently known for fuzzy regression models. Nonetheless, Tanaka's approach is used yet because of its simplicity; but it has some problems that can be classified into two categories:

1. Influence of difference trend problems.
2. Outlier data problem.

Chang and Lee [12,13] considered the first set of problems. They demonstrated that fuzziness and uncertainty in the structure of a system are two essential factors that deeply affected on the trend of the centers and spreads. Investigation on outliers was carried out by Peters [14] to control bad influence of the training data on the estimated interval. For this purpose, he applied fuzzy linear programming with triangular membership the width of which depends on some adjusting parameters such as "goodness" of the solution, the tolerance interval, and the desired value of the objective function.

Chen [15] illustrated that Peter's model may result in error, particularly when data contain outliers. Indeed, his finding revealed that PFLR (Possibilistic Fuzzy Linear Regression) or UFLR (Unrestricted sign Fuzzy Linear Regression) model is led to wrong outcomes whenever the estimated confidence interval is too broad. He put an additional restriction (k-value, which is stated as a difference in the width between the spread of the estimated data and the spread of the dependent observation data) to keep influence of outliers away. Nonetheless his model was very sensitive to the value of k. Other investigators, comprising Ortiz et al. [16] indicated that robust regression may be an alternative tool for detection of outlier data. Tanaka and Lee [17] used linear programming with quadratic programming to handle outlier data based on combination of central tendency and possibility properties.

Because Chang and Lee [12,13], Ortiz et al. [16] and Chen [15] models consider fuzzy observation while the proposed model regards crisp data, thus we consider the results of Tanaka et al. [1] model and Peters [14] model.

This paper deals with outlier data problems for non-fuzzy input and non-fuzzy output models by applying linguistic variables. Outlier data are determined by applying ordinary regression along with possibility concept to omit or lessen their effects.

* Corresponding author.
E-mail address: h.shakouri@gmail.com (H. Shakouri G.).

1568-4946/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.asoc.2012.07.001

The organization of the remaining parts of the paper is as follows. In Section 2, preliminary definition of fuzzy numbers is considered. Proposed method will be introduced in Section 3. Numeric examples as well as a case study will be applied in order to demonstrate the ability of proposed approach in the Section 4. Conclusion of the paper will be pointed out in the last section.

2. Fuzzy numbers and fuzzy regression

Based on Dubois and Prade [18], $\tilde{A}$ is defined as a fuzzy number which satisfies the following criteria:

First: normality, $\exists x \in \boldsymbol{R}$ such that $\mu_{\tilde{A}}(x) = 1$

Second: convexity, $\forall x_1, x_2 \in \boldsymbol{R}$, $\forall h \in [0;1]$

$$\mu_{\tilde{A}}(hx_1 + (1-h)x_2) \geq \min(\mu_{\tilde{A}}(x_1), \mu_{\tilde{A}}(x_2))$$

$\tilde{A} = (c_L, a, c_R)_{LR}$ is a *LR*-type fuzzy number where a, c_L and c_R are the center, left spread and right spread of the fuzzy number, respectively (c_L and $c_R > 0$). When $c_L = c_R = c$, we have a symmetric triangular fuzzy number. Thus, $\tilde{A} = (a, c)_L$ is a symmetric triangular fuzzy number if:

$$\mu_{\tilde{A}}(x) = L\left(\frac{a-x}{c}\right) = 1 - \frac{|a-x|}{c}, \quad a - c \leq x \leq a + c \tag{1}$$

In this paper symmetric triangular fuzzy numbers is only considered for simplicity.

Fuzzy Linear Regression (FLR) model was introduced initially by Tanaka et al. [1] as:

$$\tilde{Y}_i^* = \tilde{A}_0 X_{i0} + \tilde{A}_1 X_{i1} + \cdots + \tilde{A}_{p-1} X_{i,p-1} = \tilde{\boldsymbol{A}} X_i^t \tag{2}$$

where $\tilde{Y}_i^*$; $i = 1, \ldots, n$, are the estimated fuzzy data, $\tilde{A}_j = (a_j, c_j)_L$; $j = 0,1, \ldots, p-1$ are the set of symmetric fuzzy coefficients, and $X_i = [X_{i0}, X_{i1}, \ldots, X_{i,p-1}]$ are vectors of the independent variables. The extension principle [2] which plays a basic role in the fuzzy set theory, provides a foundation for all manipulations on fuzzy sets. By applying the extension principal, membership functions of $\tilde{Y}_i^*$ in the fuzzy linear regression model (2) can be obtained as:

$$\mu_{\tilde{Y}_i^*}(y^*) = \max_{y^* = f(X_i, \tilde{A})} \min_j (\mu_{\tilde{A}_j}(a_j)) \tag{3}$$

3. The new approach

Outlier data are a subset of data that have great difference with the *majority* of data. We name the non-outlier data by *reliable* data. Emerged problems by such difference can be resolved by detection of outlier data and disregarding them by either omission or reduction of their effects. Indeed, we delete a data that is *far* enough from the majority of data, where the concept of "*far*" may create various senses in the mind. One may feel that a certain point should be included in outlier data, while others do not. Such different judgments motivate us to think of a fuzzy concept as a measure for membership in the *reliable* subset of the data set. To do so, a linguistic variable is used to describe the data position and then identify outlier data. We take the concept of "*far*" as a fuzzy value with different degrees, consequently, any data point that falls outside a bound defined by an interval of $[f_{min}, f_{max}]$ is considered to be an outlier data. Then we define the fuzzy set of "*far*" with its membership function illustrated in Fig. 1. By this definition we consider a reliable data point to be *farther* and thus less reliable than another one when it spots nearer to either f_{min} or f_{max}. On the other hand, it discovers as an outlier data when to be far from f_p, which will define in follow. All members of the data set with a f_p bigger than f_{max} or less than f_{min} is supposed as an *outlier* data where $\mu_p(f) = 0$. Interval $[f_{min}, f_{max}]$ is taken into account by coefficient $\mu_i(f_{\mathrm{Tr}})$ which index Tr makes a brief of *Triangular*.

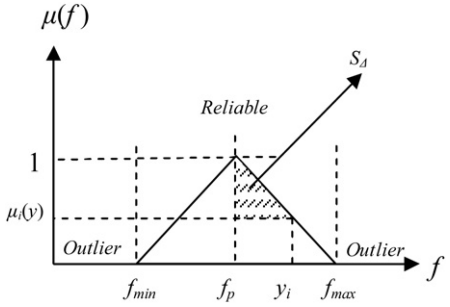

Fig. 1. Fuzzy membership function for *far*, *f*.

To find suitable f_p, here, ordinary regression as well as R-square criterion, R^2, which is introduced as a goodness of fitting method, is used. This statistic measures how successful the fit is in explaining the variation of the data. R-square is defined as the ratio of the sum of squares of the regression (SSR) and the total sum of squares (SST). Where SSR and SST are defined as follows:

$$SSR = \sum (\hat{Y} - \bar{y})^2$$

$$SST = \sum (y_i - \bar{y})^2$$

where $\hat{Y}$ and $\bar{y}$ are respectively the estimated data and the mean of the dependent data. After fitting an appropriate curve to the data, possibility concept will be defined to determine whether data position is outlier or not. Indeed, appropriate curve will be identified by a simple rule of thumb about R^2. If its amount gets greater than or equal to 0.8, it will be assessed as a proper curve. We will use the possibility concept provided that the amount of *R*-square to be identified greater than or equal 0.8, else the ordinary regression curve will be fitted to each set of data by deleting one observation in each step and saving the corresponding *R*-square amount. At the end, there is an array of *R*-square data, which their maximum will be deleted. In other words, we believe that by canceling an outlier data, the *R*-square amount will improve. Then, an ordinary regression is applied to new data set. The process will over if the amount of R^2 confirms the mentioned rule, otherwise, it will repeat until to reach a desired *R*-square. The quantity of f_i will be achieved by substituting its corresponding x_i amount in the final ordinary regression equation ($f_i = \hat{Y}(x_i)$).

By fitting suitable curve to the data, we simply define below the possibility of being reliable for each data by a triangular membership, $\mu_i(f)$. It is used to displace dependent data and create a new data set, namely y_i^*.

$$\text{If } y_i \leq f_{\min} \text{ or } y_i = f_{\max} \text{ then delete } y_i \tag{4}$$

$$\text{If } y_i \subseteq [f_{\min}, f_i] \text{ then } y_i^* = f_i - S_\Delta \tag{5}$$

$$\text{If } y_i \subseteq [f_i, f_{\max}] \text{ then } y_i^* = f_i + S_\Delta \tag{6}$$

where S_Δ, f_{min} and f_{max} are calculated as follows:

$$S_\Delta = \frac{\left|y_i - \hat{Y}(x_i)\right| (1 - \mu_i(y))}{2} \tag{7}$$

$$f_{\max} = f_i + \beta \times \sqrt{Var(y)} \tag{8}$$

$$f_{\min} = f_i - \beta \times \sqrt{Var(y)} \tag{9}$$

Var(.) stands for variance and β is a parameter the amount of which is supposed to equate (3). Therefore, by substituting the new data, y^*_i, instead of the original one, y_i, the proposed fuzzy regression model below is achieved.