Knowledge-Based Systems 52 (2013) 302–310

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A two-stage approach for formulating fuzzy regression models

Liang-Hsuan Chen *, Chan-Ching Hsueh, Chia-Jung Chang

Department of Industrial and Information Management, National Cheng Kung University, Tainan, Taiwan, ROC

ARTICLE INFO

Article history:
Received 22 August 2012
Received in revised form 2 August 2013
Accepted 4 August 2013
Available online 15 August 2013

Keywords:
Fuzzy sets
Fuzzy regression model
Mathematical programming
Least-squares method
Two-stage approach

ABSTRACT

Fuzzy regression models have been widely applied to explain the relationship between explanatory variables and responses in fuzzy environments. This paper proposes a simple two-stage approach for constructing a fuzzy regression model based on the distance concept. Crisp numbers representing the fuzzy observations are obtained using the defuzzification method, and then the crisp regression coefficients in the fuzzy regression model are determined using the conventional least-squares method. Along with the crisp regression coefficients, the proposed fuzzy regression model contains a fuzzy adjustment variable so that the model can deal with the fuzziness from fuzzy observations in order to reduce the fuzzy estimation error. A mathematical programming model is formulated to determine the fuzzy adjustment term in the proposed fuzzy regression model to minimize the total estimation error based on the distance concept. Unlike existing approaches that only focus on positive coefficients, the problem of negative coefficients in the fuzzy regression model is taken into account and resolved in the solution procedure. Comparisons with previous studies show that the proposed fuzzy regression model has the highest explanatory power based on the total estimation error using various criteria. A real-life dataset is adopted to demonstrate the applicability of the proposed two-stage approach in handling a problem with negative coefficients in the fuzzy regression model and a large number of fuzzy observations.

1. Introduction

Regression analysis is usually adopted for investigating the relationships between independent (or input, explanatory) and dependent (or output, response) variables using a group of crisp observations in various research and application fields. The least-squares method is often applied to find regression coefficients to minimize the squared errors. In practice, observations are sometimes described in linguistic terms, such as "approximately equal to 100", due to fuzzy, inexact, or insufficient information. In order to describe such fuzzy data, fuzzy set theory [36] is frequently adopted. Fuzzy regression models are developed to formulate a relationship between input and output variables in a fuzzy environment.

Tanaka et al. [30] were the first to propose a linear programming technique for determining fuzzy regression coefficients in the case of fuzzy dependent variables and crisp independent ones. Since then, fuzzy regression has been used by many researchers for applications and the development of methodologies [8,10,11,15,17,22,26,35]. Although Tanaka [28], Tanaka et al. [29], and Tanaka and Watada [31] made some improvements to Tanaka et al.'s [30] original fuzzy regression model, the improved models still suffer some shortcomings, such as over-sensitivity to outliers, as pointed out by Redden and Woodall [23]. Wang and Tsaur [32] stated that Tanaka et al.'s [30] model possibly produces excessively wide ranges of estimation. Furthermore, Kao and Chyu [12,13] showed that fuzzy regression coefficients broaden the spread of estimated responses when the magnitude of the independent variables increases, even though the actual spreads of the observed responses are approximately constant or decreasing. These flaws lessen the explanatory power of the model.

To overcome these problems, Kao and Chyu [12] proposed a two-stage procedure for determining the crisp coefficients and fuzzy error terms in fuzzy regression models based on Kim and Bishu's [14] criterion. They also proposed a least-squares method to derive crisp regression coefficients [13]. Chen and Dang [5] improved Kao and Chyu's [13] model and achieved better performance in the reduction of the total estimation error based on Kim and Bishu's [14] criterion. Although several researchers [5,12,14] used Kim and Bishu's [14] criterion to develop fuzzy regression models, the development of models using this criterion is flawed due to the fact that if the observed and estimated fuzzy responses do not intersect, the estimation error remains constant irrespective of the distance between the observed and estimated responses. In such circumstances, the measure does not reflect the actual difference between the responses. Another drawback commonly found in the literature is that the fuzzy regression model is limited to fuzzy observations with symmetrical triangular fuzzy numbers [7,9,14,20,25,33]. To minimize the distance between the observed and estimated fuzzy responses, Chen and Hsueh [2] considered fuzzy adjustment variables in the fuzzy regression models, and

* Corresponding author.
E-mail address: lhchen@mail.ncku.edu.tw (L.-H. Chen).

0950-7051/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.knosys.2013.08.010

developed a mathematical programming method based on the distance criterion to determine the crisp regression coefficients and fuzzy adjustment variables simultaneously. In addition to dealing with non-symmetrical triangular fuzzy numbers, their approach has several advantages, including an effective reduction of the total estimation error, with the limitation that feasible solutions might not be found when the number of fuzzy observations is large.

In most existing approaches, fuzzy regression models are formulated with the assumption that the sign of regression coefficients is positive [2,7,9–13,20,25]. When the relationship of the fuzzy response and a fuzzy (or crisp) independent variable has the opposite change, the formulations may be incorrect based on fuzzy arithmetic. Recently, Bargiela et al. [1] developed an iterative algorithm that employs the gradient-descent optimization technique to determine possible negative coefficients for fuzzy multiple regression models to achieve the minimum sum of squared errors. In their approach, they used defuzzification methods to defuzzify the observed and the predicted fuzzy responses for demonstrating the performance of the proposed fuzzy regression model. Chen and Hsueh [3] proposed a least-squares method to construct fuzzy regression models by minimizing the total estimation error. The concept of correlation coefficients was adopted in their approach to find a possible opposite change between the fuzzy response and a fuzzy (or crisp) independent variable.

In the present study, a two-stage approach based on Kao and Chyu's [12] and Chen and Hsueh's [2] concepts is proposed to establish fuzzy regression models with crisp coefficients and resolve the problem of negative coefficients. Different from Chen and Hsueh's [3] approach in which each correlation coefficient between the fuzzy response and a fuzzy (or crisp) independent variable must be calculated separately to find possible negative coefficients and which may require some iterations to determine the coefficients, the proposed approach is efficient. In the proposed approach, the fuzzy observations are first defuzzified into crisp values to calculate the crisp regression coefficients using the classical least-squares method. This stage constructs a defuzzified version of the fuzzy regression model. To incorporate the fuzziness of the input and/or output variables into the model, fuzzy adjustment variables are added into the model in the second stage. With the crisp regression coefficients treated as the known parameters, the fuzzy adjustment variables in the fuzzy regression models are determined using a mathematical program to minimize the total estimation error based on the distance concept. It is noted that the formulations of fuzzy regression models in the mathematical programming should consider the sign of the regression coefficient for each input variable based on the principle of fuzzy arithmetic.

The rest of this paper is organized as follows. Section 2 formulates the fuzzy regression model with the fuzzy adjustment term. The solution approaches, namely the least-squares method and mathematical programming, are described in Section 3. In Section 4, three examples are used to illustrate applications, and the performance of the proposed models is compared with that of existing models. A fourth example is used to demonstrate the applicability of the proposed approach to a problem with negative coefficients in the fuzzy regression model. In addition, a real-life dataset is used as a fifth example to demonstrate the ability of the proposed two-stage approach to deal with a large number of observations, where the established fuzzy regression model has negative coefficients. Conclusions are provided in the final section.

2. Fuzzy linear regression model

The traditional regression model is frequently used to express the relationship between the response and one or more explanatory variables. If the observations are fuzzy, the fuzzy regression model is applied to establish the relationship between the fuzzy response and fuzzy input variables. Let $(\widetilde{X}_{i1}, \ldots, \widetilde{X}_{ip}, \widetilde{Y}_i)$, $i = 1, \ldots, n$, be n pairs of fuzzy observations, each with a fuzzy response and p fuzzy input variables, which are characterized by membership functions $\mu_{\widetilde{Y}_i}$ and $\mu_{\widetilde{X}_{ij}}$, respectively. In general, for computational efficiency, triangular fuzzy numbers, which are defined as normal and convex sets, are used [36]. The fuzzy number of $\widetilde{X}_{ij}$ can be expressed as a triple element set $(X_{ijl}, X_{ijm}, X_{iju})$ with the following membership function:

$$\mu_{\widetilde{X}_{ij}}(x) = \begin{cases} (x - X_{ijl})/(X_{ijm} - X_{ijl}), & X_{ijl} \leqslant x \leqslant X_{ijm} \\ (X_{iju} - x)/(X_{iju} - X_{ijm}), & X_{ijm} \leqslant x \leqslant X_{iju} \end{cases} \tag{1}$$

where X_{ijl}, X_{ijm}, and X_{iju} are the smallest, most possible, and largest values, respectively, $\mu_{\widetilde{X}_{ij}}(X_{ijm}) = 1$, and $\mu_{\widetilde{X}_{ij}}(X_{ijl}) = \mu_{\widetilde{X}_{ij}}(X_{iju}) = 0$. Let $b_0, b_1, \ldots, b_p$ denote the crisp estimates of the regression parameters. Referring to the traditional regression model, the fuzzy regression model can be formulated as:

$$\hat{\widetilde{Y}}_i = b_0 + b_1\widetilde{X}_{i1} + b_2\widetilde{X}_{i2} + \cdots + b_p\widetilde{X}_{ip}, \quad i = 1, \ldots, n \tag{2}$$

Unlike existing models which have fuzzy regression coefficients [9,14,20,22,25,31,33,35], the regression coefficients in (2) are crisp values, so that the problem of fuzzy responses with wide estimation spreads due to large explanatory variables is avoided. However, if the explanatory variables are crisp, the predicted or estimated responses in (2) are also crisp, leading to a large fuzzy error for a fuzzy response [2]. To deal with this problem, Chen and Hsueh [2] added a fuzzy adjustment term $\tilde{\delta}$, which denotes a fuzzy number with the membership function $\mu_{\tilde{\delta}}$. For generalization, this concept is also adopted in the present study because it effectively resolves the problem. The membership function of a fuzzy adjustment term is a triangular fuzzy number when $\widetilde{X}_{ij}$ and $\widetilde{Y}_i$ are triangular numbers. Let $\tilde{\delta}$ represent a triangular fuzzy number with the membership function $\tilde{\delta} = (\delta_l, \delta_m, \delta_u)$, which is defined as:

$$\mu_{\tilde{\delta}} = \begin{cases} (\delta - \delta_l)/(\delta_m - \delta_l), & \delta_l \leqslant \delta \leqslant \delta_m \\ (\delta_u - \delta)/(\delta_u - \delta_m), & \delta_m \leqslant \delta \leqslant \delta_u \end{cases} \tag{3}$$

The fuzzy adjustment term, $\tilde{\delta}$, is added to (2), so that the proposed fuzzy regression model is formulated as follows:

$$\hat{\widetilde{Y}}'_i = b_0 + b_1\widetilde{X}_{i1} + b_2\widetilde{X}_{i2} + \cdots + b_p\widetilde{X}_{ip} + \tilde{\delta} \tag{4}$$

That is, b_0 and $\tilde{\delta}$ can actually be combined in the formulation. Other than the regression coefficients, the three parameters of the membership function $\mu_{\tilde{\delta}}$, i.e., δ_l, δ_m, and δ_u, are regarded as variables to be determined. The coefficients and $\mu_{\tilde{\delta}}$ should be determined to maximize the explanatory power of $\hat{\widetilde{Y}}'_i$ for the fuzzy response $\widetilde{Y}_i$.

3. Solution approaches

Based on the fuzzy regression model (4), directly determining all coefficients and the three parameters of $\mu_{\tilde{\delta}}$ is difficult, since fuzzy observations represented as fuzzy numbers with membership functions are considered in the model formulation. To overcome this problem, a two-stage approach is used to find the optimal coefficients and parameters. In the first stage, the fuzzy observations $\widetilde{X}_{ij}$ and $\widetilde{Y}_i$ are defuzzified into crisp values. In fuzzy problems, defuzzification is usually adopted for simplification. Since the fuzzy regression model contains some fuzzy numbers, the problem can be simplified by using a crisp number to represent each fuzzy number. The crisp numbers can then be used to determine the relationship between the corresponding fuzzy numbers and the fuzzy response. Many defuzzification approaches have been developed, among which the centroid method is commonly used in solving fuzzy problems [6,24]. The defuzzified (crisp) values, X_{ijc} and