Information Sciences 178 (2008) 3973–3988

Contents lists available at ScienceDirect

Information Sciences

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

A variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy

Shih-Pin Chen *, Jr-Fong Dang

Department of Business Administration, National Chung Cheng University, Chia-Yi, Taiwan

ARTICLE INFO

Article history:
Received 1 January 2008
Received in revised form 10 April 2008
Accepted 7 June 2008

Keywords:
Forecasting
Fuzzy sets
Fuzzy inference
Linear regression
Mathematical programming

ABSTRACT

Fuzzy regression models have been applied to operational research (OR) applications such as forecasting. Some of previous studies on fuzzy regression analysis obtain crisp regression coefficients for eliminating the problem of increasing spreads for the estimated fuzzy responses as the magnitude of the independent variable increases; however, they still cannot cope with the situation of decreasing or variable spreads. This paper proposes a three-phase method to construct the fuzzy regression model with variable spreads to resolve this problem. In the first phase, on the basis of the extension principle, the membership functions of the least-squares estimates of regression coefficients are constructed to conserve completely the fuzziness of observations. In the second phase, then they are defuzzified by the center of gravity method to obtain crisp regression coefficients. In the third phase, the error terms of the proposed model are determined by setting each estimated spread equals its corresponding observed spread. Furthermore, the Mamdani fuzzy inference system is adopted for improving the accuracy of its forecasts. Compared to the previous studies, the results from five examples and an application example of Japanese house prices show that the proposed fuzzy linear regression model has higher explanatory power and forecasting performance.

1. Introduction

Fuzzy regression analysis has been demonstrated to be a powerful methodology for analyzing the vague relationship between a dependent variable (also called response variable) and independent variables (also called explanatory variables) in complex systems involving human subjective judgement under incomplete data conditions [23]. It has been successfully applied to various applications. Recently, for example, fuzzy regression models have been applied to insurance [1,2], housing [4], thermal comfort forecasting [12], productivity and consumer satisfaction [17], product life cycle prediction [20], R&D project evaluation [22], reservoir operations [32], actuarial analysis [38], robotic welding process [39], and business cycle analysis [46].

Much research has been devoted to fuzzy regression analysis (recently, for example [7,14,16,18,19,21,28,45]), and the fuzzy linear regression (FLR) model is the most frequently investigated. Tanaka et al. [40] was probably the first study of the FLR problem with crisp explanatory variables and fuzzy response variables. They formulated the FLR problem as a linear programming model to determine the regression coefficients as fuzzy numbers, where the objective was to minimize the total spread of the fuzzy regression coefficients subject to the constraint that the support of the estimated values is needed to cover the support of their associated observed values for a certain prespecified level. Later, this approach was improved by Tanaka [41], Tanaka and Watada [42], and Tanaka et al. [43]. However, several studies have pointed out the drawbacks of

* Corresponding author. Tel.: +886 5 2720411; fax: +886 5 2720564.
E-mail address: chensp@ccu.edu.tw (S.-P. Chen).

0020-0255/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2008.06.005

these approaches. For example, Redden and Woodall [34] pointed out that the above approaches are still very sensitive to outliers; Wang and Tsaur [44] pointed out that Tanaka's model provides overly wide ranges in estimation. Moreover, recently, Kao and Lin [26] stated that the main drawback of the Tanaka approach and its variations is that more observations result in fuzzier estimations, which contradicts the general observation that more observations provide better estimations.

There are many other studies on fuzzy regression analysis. For example, Diamond [15] proposed a fuzzy least-squares approach to determine the regression coefficients. Kim and Bishu [29] proposed an approach based on the criterion of minimizing the difference of membership values between the observed and estimated fuzzy dependent variable. Sakawa and Yano [37] formulated three types of multiobjective programming approaches to investigate the fuzzy linear regression model with fuzzy explanatory variables and responses. Hong et al. [18] used shape preserving arithmetic operations on *L*–*R* fuzzy numbers for least-squares fitting to investigate a class of fuzzy linear regression analysis problem. A common characteristic of these studies is that the derived regression coefficients are fuzzy numbers. However, as noted by Kao and Chyu [24,25], since the regression coefficients derived based on Zadeh's extension principle [47,49] are fuzzy numbers, the spread of the estimated dependent variable becomes wider as the magnitudes of the independent variables increase, even if the spreads of the observed dependent variables are actually decreasing.

To avoid the problem of wider spreads for larger value of the explanatory variables in estimation, Kao and Chyu [24] proposed a two-stage approach to obtain crisp regression coefficients in the first stage and to determine a unique fuzzy error term in the second stage. Moreover, Kao and Chyu [25] proposed a least-squares method to derive regression coefficients that are also crisp. The results of these two studies show proposed models have better performances than the previous studies. However, as pointed out by Kao and Lin [26], these two methods still cannot cope with the situation of decreasing spread or variable spread. In fact, little research has been published regarding the development of methods that can deal with this problem.

Another problem to be noted is that the crisp regression coefficients may eliminate the problem of increasing spread, but they also can mislead the functional relationship between the dependent variable and independent variables in fuzzy environments. In particular, when the spreads of fuzzy observed response or independent variables are large, it is possible that the spreads of regression coefficients are also large. In this case, the values of regression coefficients are a wide range, even from negative to positive values. If the derived regression coefficients are crisp, some useful insights and valuable information may be lost. As stated in Bargiela et al. [4], 'Regression model based on fuzzy data shows a very beneficial characteristic of enhanced generalization of data patterns compared to the regression models that are based on numerical data only. This is because the membership function associated with fuzzy sets has a significant informative value in terms of capturing either a notion of accuracy of information or a notion of proximity of patterns in the data set used for the derivation of the regression model'. Accordingly, when the response or explanatory variables are fuzzy, the regression coefficients will be fuzzy as well, and they should be described by membership functions to completely conserve the fuzziness of response or explanatory variables. However, there has been little research on the problem of deriving the membership functions of fuzzy regression coefficients.

This study addresses the above two important problems of fuzzy linear regression that little research has been devoted to. The purposes of this paper are, firstly, to propose a procedure for constructing the membership function of fuzzy regression coefficient such that the fuzziness of input information can be completely conserved. Secondly, then to propose a variable spread fuzzy linear regression model with higher explanatory power and forecasting accuracy, which can resolve the problem of wider spreads of the estimated response for larger values of independent variables in fuzzy regression analysis, and also can cope with the situation of decreasing or variable spreads.

In this paper we propose a three-phase approach that is an improved method based on the concept of Kao and Chyu [24] to tackle the above problems. In the first phase, to completely conserve the fuzziness and obtain some useful insights and valuable information, the membership functions of least-squares estimations of fuzzy response and explanatory variables are derived based on Zadeh's extension principle [47,49]. To avoid the problem of wider spreads for larger value of the explanatory variables in estimation, the fuzzy regression coefficients are defuzzified to crisp values via a fuzzy ranking method in the second phase. Then the third phase uses a mathematical programming approach to determine the fuzzy error term for each pair of explanation variables and response, in that phase, the objective is to minimize the errors in estimation subject to the constraints including the spread of each estimated response equal to that of the associated observed response. Since the spreads of error terms coincide with those of their associated observed responses, spreads derived in this paper vary and follow the variable spreads, no matter how the spreads of observed responses change. Thus the problem of decreasing spread or variable spread affecting the previous studies can be avoided.

In the following sections, firstly, the fuzzy linear regression problem is briefly introduced. The derivation of the membership function of least-squares estimation of fuzzy regression coefficients is discussed next. Then the defuzzification method for deriving the crisp regression coefficients is described, and the derivation of the varying fuzzy error terms is also presented. Finally, the advantages of the proposed method over some other methods are illustrated by solving several examples and a practical application example of multiple fuzzy regression.

2. Fuzzy linear regression model

The linear regression model is the most frequently used form in regression analysis for expressing the relationship between one or more explanatory variables and response. Without loss of generality, consider the case of simple linear