ELSEVIER

Fuzzy Sets and Systems 126 (2002) 401–409

FUZZY sets and systems

www.elsevier.com/locate/fss

A fuzzy linear regression model with better explanatory power

Chiang Kao *, Chin-Lu Chyu

Department of Industrial Management, National Cheng Kung University, Tainan, Taiwan 70101, ROC

Received 10 April 2000; received in revised form 24 November 2000; accepted 20 January 2001

Abstract

Previous studies on fuzzy linear regression analysis have a common characteristic of increasing spreads for the estimated fuzzy responses as the independent variable increases its magnitude, which is not suitable for general cases. This paper proposes a two-stage approach to construct the fuzzy linear regression model. In the first stage, the fuzzy observations are defuzzified so that the traditional least-squares method can be applied to find a crisp regression line showing the general trend of the data. In the second stage, the error term of the fuzzy regression model, which represents the fuzziness of the data in a general sense, is determined to give the regression model the best explanatory power for the data. The results from two examples, one with crisp data and the other with fuzzy data for the independent variable, indicate that the two-stage method proposed in this paper has better performance than the previous studies.

Keywords: Defuzzification; Fuzzy linear regression; Least-squares method; Mathematical programming

1. Introduction

Regression analysis is a powerful and comprehensive methodology for analyzing relationships between a response variable, called the dependent variable, and one or more explanatory variables called independent variables. Inferential problems associated with the regression model include the estimation of the model parameters and prediction of the response variable from knowledge of the explanatory variables. Many areas such as economics, engineering, biology, and physical sciences have wide applications of this methodology. By the classical statistical technique, the observations, either the response variable or the explanatory variables, are required to follow certain probability distributions, usually a normal distribution. However, in practice, there are cases that the observations are inherently fuzzy. For example, the observations are described by linguistic terms such as large, heavy, or approximately equal to five. How to estimate the regression coefficients and make the subsequent prediction under the fuzzy environment is a challenge to the classical regression analysis.

In a paper published in 1970, Bellman and Zadeh [1] proposed the concept of fuzzy set theory. Since then, several scholars have constructed different fuzzy regression models and proposed the associated solution methods. The article by Tanaka et al. [18] is probably the first research on this topic. In their study, a regression problem with fuzzy dependent variable

* Corresponding author. Tel.: +886-6-2753-396; fax: +886-6-2362-162.

E-mail address: ckao@mail.ncku.edu.tw (C. Kao).

0165-0114/02/$ - see front matter
PII: S0165-0114(01)00069-0

and crisp independent variable was formulated as a mathematical programming problem. The objective was to minimize the total spread of the fuzzy regression coefficients subject to the constraint that the regression model needed to satisfy a prespecified membership value in estimating the fuzzy responses. The main drawback of this approach is that it is scale dependent. Although this approach was later improved by Tanaka [15], Tanaka and Watada [16], and Tanaka et al. [17], it still suffered the problem of being extremely sensitive to outliers as pointed out by Redden and Woodall [12]. Moreover, it can produce infinite solutions and the spread of the estimated response becomes wider as more data are included in the model.

In addition to the Tanaka approach, there are two other approaches being proposed. One is the fuzzy least-squares approach [3], in that analogues of the conventional normal equations are derived with a suitable metric. The other is the fuzzy random variable approach [9,10]. A kind of linear estimation theory is developed to derive the well-known Aumann expectation [10] for fuzzy random variables.

Recently, Kim and Bishu [7] proposed another approach based on the criterion of minimizing the difference of the membership values between the observed and estimated fuzzy responses. The results from two examples favor this approach in comparison with that of Tanaka et al. [17]. There are several other studies [6,8,11,13,14,20] on fuzzy regression analysis. Most studies emphasized on the fuzziness of the response variable alone. Only a few [3,14] discussed the situation that both the response and explanatory variables are fuzzy.

A common characteristic of these studies is that the regression coefficients were treated as fuzzy numbers. A consequence is that, in estimation, the spread of the estimated responses becomes wider as the magnitude of the explanatory variables increases, even though the spreads of the observed responses are roughly constant or decreasing. This assumption has no theoretical base to support.

In this paper we propose a two-stage methodology to obtain the regression coefficients which are crisp in the first stage and to determine a fuzzy error term in the second stage to produce fuzzy estimations. The idea is to defuzzify the fuzzy observations to crisp values and apply the classical least-squares method to calculate the regression coefficients. The fuzzy error term is determined via a mathematical program by minimizing the errors in estimation. Since the regression coefficients are crisp, the problem that the spreads in estimation are increasing suffered by the previous studies can be avoided.

In the sections that follow, firstly, the defuzzification method for determining the crisp regression coefficients is introduced. Then the derivation of the fuzzy error term is discussed. Finally, two examples are solved to illustrate the advantages of the two-stage approach over those of the previous studies.

2. Regression coefficients

Let X_{ij} and Y_i be the jth independent variable and the response variable, respectively, in the ith case, β_j be the parameter associated with the jth independent variable, and ε_i be the error term associated with the ith observation. The classical regression model can be stated as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \varepsilon_i, \quad i = 1, \ldots, n. \tag{1}$$

The regression parameter β_j must usually be estimated from sample data. If some of the observations X_{ij} and Y_i are fuzzy, then it falls into the category of fuzzy regression analysis [6].

Without loss of generality, assume all observations are fuzzy numbers, since crisp values can be represented by degenerated fuzzy numbers. Consider the simplest case of one independent variable. The cases of multiple variables can be generalized from this case. The fuzzy linear regression model is

$$\tilde{Y}_i = \beta_0 + \beta_1 \tilde{X}_i + \tilde{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2}$$

where $\tilde{X}_i, \tilde{Y}_i$, and $\tilde{\varepsilon}_i$ are fuzzy numbers with the membership functions, $\mu_{\tilde{X}_i}$, $\mu_{\tilde{Y}_i}$ and $\mu_{\tilde{\varepsilon}_i}$, respectively. The problem is to find the estimates for β_0, β_1, and $\tilde{\varepsilon}_i$ which provide the best explanation for the relationship between the independent and dependent variables.

For easy explanation, assume all observations are triangular fuzzy numbers defined as $\tilde{X}_i =$