Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 274 (2015) 109–123

FUZZY sets and systems

www.elsevier.com/locate/fss

The normalized interval regression model with outlier detection and its real-world application to house pricing problems

Chao Wang [a], Jing Li [b], Peijun Guo [a,*]

[a] *Faculty of Business Administration, Yokohama National University, 79-4 Tokiwadai, Hodogaya-ku, Yokohama 240-8501, Japan*
[b] *Business School, East China Normal University, 500 Dongchuan Rd., Shanghai 200241, China*

Received 1 October 2013; received in revised form 9 June 2014; accepted 11 June 2014

Available online 1 July 2014

Abstract

Tanaka and his colleagues initially proposed fuzzy linear regression models in 1982. From then on, fuzzy regression analysis has been widely studied by many researchers. Since Tanaka's model uses the inclusion relationship between the given data and the estimated data, it was pointed out that Tanaka's model is sensitive to outliers. How to deal with the outlier problem has been a very important issue and attracted much attention from fuzzy regression researchers. Given crisp multiple independent variables and single interval dependent variable, we introduce the normalized upper and lower interval regression models and propose two outlier detection approaches for them. The dual problem of the normalized upper interval regression model is used to find out potential outliers. For the normalized lower interval regression model, the existence of the outliers makes it infeasible. The relaxation approach is proposed to find out the outliers. By deleting the outliers, we can find out the possibilistic functional relationship of the major data. A case study involving a house pricing problem is analyzed in detail. Household size, loan ratio and annual household income are identified as the independent variables and acceptable purchase price is the interval dependent variable. With the data collected from the survey in Shanghai, the normalized upper and lower interval regression models are built with deleting the outliers. The proposed models provide managerial insights into the real estate market and important policy implications in regulating urban land development.

Keywords: Interval linear regression model; Outlier detection; Linear programming; Dual problem; Duality approach; Relaxation approach

1. Introduction

Fuzzy regression analysis is a method to obtain a fuzzy functional relationship between the dependent and independent variables. Different from conventional statistical regression models assuming that the deviations between the observed and estimated data are due to random errors, the uncertainty in the fuzzy regression model is thought to be caused by fuzziness, such as indefiniteness of a system, imprecise observation, vagueness in real-world problems, etc.

* Corresponding author.

http://dx.doi.org/10.1016/j.fss.2014.06.009

Generally speaking, there are two kinds of approaches for fuzzy regression analysis. One is the least square approach which aims at minimizing the distances between the estimated and observed fuzzy outputs [2,5,34–36]. The other is the possibilistic regression model where inclusion relationships between given outputs and estimated outputs play a critical role [9]. The possibilistic regression model was initiated and developed by Tanaka and his colleagues [26–29]. In order to deal with the interactive fuzzy coefficients, the quadratic and exponential membership functions of fuzzy parameters were considered [30,31]. Based on Tanaka's possibilistic regression model, an approach with considering the central tendency was proposed [32]. The method was extended to the non-symmetrical case [6] and utilized in the CAPM beta estimation problem [17]. Further, the possibilistic regression model has been extended to the real time fuzzy regression analysis [24] and fuzzy autocorrelation models [37]. To deal with the hybrid uncertain data, the confidence–interval-based fuzzy random regression model (CI-FRRM) was introduced [38]. For interval regression analysis, an approach based on the Minkowski difference was investigated [14]. Also, support vector machines were extensively used in the interval regression analysis [4,10,13,15]. Recently, tolerance analysis approaches were adapted for interval regression analysis [11].

Due to the inclusion relationship between the given data and the estimated data in the possibilistic regression model, the possibilistic regression model is sensitive to outliers. How to deal with the outlier problem has been a very important issue and attracted much attention from fuzzy regression researchers. Generally speaking, the existing outlier detection approaches can be divided into three categories. First, auxiliary variables are introduced into the original model [20,21] to relax the inclusion relationship between the given output and the estimated one. However, introducing new variables makes the computing process complex and their methods lack theoretical support base. Second, the restriction in sign is dropped [3]. It seems that an unrestricted spread makes the regression model fit the observed data better. However, there is no reasonable explanation for a negative spread and more seriously, the methods violate the extension principle of fuzzy set theory. Third, the omission methods are utilized [8,12]. In these approaches, the number of the linear programming problems is that of the observed data. Obviously, it requires much more computational efforts when the number of samples is large.

In this paper, we consider the interval regression model with multiple crisp inputs and an interval output. We introduce the normalized interval regression models and propose two outlier detection approaches corresponding to upper and lower interval regression models, respectively. The dual problem of the normalized upper interval regression model is used to find out the potential outliers. For the normalized lower interval regression model, the existence of the outliers makes it infeasible. The relaxation approach is proposed to find out the outliers. By deleting the outliers, we can find out the possibilistic functional relationship of the major data. A case study involving a house pricing problem is analyzed in detail. Household size, loan ratio and annual household income are identified as the inputs and acceptable purchase price is the interval output. With the data collected from the survey which was conducted in Shanghai, the normalized upper and lower interval regression models are built with deleting the outliers and the interesting analysis results are obtained.

This paper extends the existing literature in four important dimensions. First, the normalized interval regression model is given. Thus, it becomes possible to analyze how the uncertainty of the output is allocated to the coefficients of input variables and which input variable is more deterministic. This extension makes the interval regression model not only be able to predict but also to gain managerial insights from regression analysis. Second, the dual problem of the normalized upper interval regression model is used to detect the outliers. This method takes advantage of the characteristic of this model, i.e. the linear programming problem. It makes the proposed method theoretically reasonable and practically efficient. Third, it is the first time that the outlier detection approach for the lower regression model is proposed. Fourth, a real application of interval regression analysis with multiple inputs is studied. The output is acceptable purchase price which is intrinsically of interval but not artificially created. To the best of our knowledge, our research is the first real application which uses the data that are intrinsically interval. The dataset used in this paper will be greatly useful for the future research of fuzzy regression community.

The paper is organized as follows: In Section 2, we introduce the classic interval regression model. In Section 3, we make a brief review of the existing outlier detection approaches. In Section 4, we propose the normalized upper and lower interval regression models with two outlier detection approaches. In Section 5, a real application of a house pricing problem is studied. In Section 6, some concluding remarks are made.