Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 160 (2009) 2505–2523
FUZZY sets and systems
www.elsevier.com/locate/fss

An enhanced fuzzy linear regression model with more flexible spreads

Jingli Lu, Ruili Wang*

School of Engineering and Advance Technology, Massey University, Private Bag 11-222, Palmerston North, New Zealand

Received 15 May 2008; received in revised form 13 February 2009; accepted 18 February 2009
Available online 4 March 2009

Abstract

One of the deficiencies of previous fuzzy linear regression models is that with the increase of the magnitudes of independent variables, the spreads of estimated fuzzy dependent variables are increasing, even though the spreads of observed dependent variables actually decrease or remain unchanged. Some solutions have been proposed to solve this spreads increasing problem. However, those solutions still cannot model a decreasing trend in the spreads of the observed dependent variables as the magnitudes of the independent variables increase. In this paper we propose an enhanced fuzzy linear regression model (model FLR_{FS}), in which the spreads of the estimated dependent variables are able to fit the spreads of the observed dependent variables, no matter the spreads of the observed dependent variables are increased, decreased or unchanged as the magnitudes and spreads of the independent variables change. Four numerical examples are used to demonstrate the effectiveness of model FLR_{FS}.

Keywords: Fuzzy linear regression; Fuzzy number; Linear programming; Least-square method

1. Introduction

Fuzzy linear regression (FLR) was first proposed by Tanaka et al. [25] as an extension of the classical regression analysis, which is becoming a powerful tool to explore the vague relationship between dependent and independent variables [3]. In fuzzy regression, some elements of the regression models are represented by imprecise data.

General FLR models for crisp input–fuzzy output data [25] and fuzzy input–fuzzy output data [21] can be represented as follows, respectively:

$$\hat{\tilde{Y}}_i = \tilde{A}_0 + \tilde{A}_1 x_{i1} + \cdots + \tilde{A}_j x_{ij} + \cdots + \tilde{A}_m x_{im} \qquad (\mathrm{FLR}_{CF})$$

$$\hat{\tilde{Y}}_i = \tilde{A}_0 + \tilde{A}_1 \tilde{X}_{i1} + \cdots + \tilde{A}_j \tilde{X}_{ij} + \cdots + \tilde{A}_m \tilde{X}_{im} \qquad (\mathrm{FLR}_{FF})$$

where $\tilde{A}_j$ is the jth fuzzy regression coefficients, x_{ij} or $\tilde{X}_{ij}$ is the jth independent variable of the ith instance, $x_{i0}(\tilde{X}_{i0})$ is 1, $\hat{\tilde{Y}}_i$ is the ith estimated dependent variable, $i = 1, 2, \ldots, n$, $j = 0, 1, \ldots, m$. A tilde character ($\sim$) is placed above

* Corresponding author. Tel.: +64 21 077 2118.
E-mail address: r.wang@massey.ac.nz (R. Wang).

0165-0114/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2009.02.023

the name of a fuzzy variable to distinguish a fuzzy variable from a crisp variable. As crisp numbers are special fuzzy numbers, model FLR_{CF} can be treated as a special case of model FLR_{FF}.

The methods to estimate the fuzzy regression coefficients can be roughly categorized into two groups. One is the linear programming (LP) methods [18,23,25]; the other is the least-squares (LS) methods [1,4–6,16,27]. The LP methods minimize the total spread of the estimated dependent variables or that of the fuzzy regression coefficients, subject to the constraint that the estimated dependent variables include the observed dependent variables within a certain h-level. The advantage of the LP methods is low computational complexity. However, the LP methods have been criticized by Redden and Woodall [20] as (i) they are extremely sensitive to outliers [10]; (ii) they do not allow all observations to contribute to the estimation; and (iii) the estimated intervals become wider as more data are collected. Multi-objective fuzzy regression techniques are developed to overcome these deficiencies of the LP methods [18,19,22,26]. The LS methods minimize the total difference between the estimated dependent variables and their observed counterparts. Thus, compared with the estimations of the LP methods, the estimations of the LS methods have relatively small differences between the estimated dependent variables and the observed ones. However, the LS methods have relatively higher computational complexity. A comprehensive literature review of fuzzy regression can be found in [14].

As indicated in [2,11–13,17], a problem of model FLR_{FF} is that with the increase of the magnitudes of independent variables, the spreads of estimated dependent variables are increasing (refer to Section 2.3), even though the spreads of observed dependent variables are roughly constant or decreasing. We call it *spreads increasing problem* (refer to Section 2.3) in this paper. Some models [2,5,8,11–13,17], which address this problem, and their deficiencies are briefly discussed below. More details are given in Section 3.

FLR models presented in [2,5,8] can avoid the *spreads increasing problem* by modelling centres and spreads of dependent variables separately. However, the number of parameters to be estimated in model FLR_{CD08} [2] proportionally increases with the increase of the number of instances. Although more parameters involved in a regression model increase the model fitness, these also decrease the model generality [13]. Therefore, model FLR_{CD08} is unsuitable for large dataset regression (refer to Section 3.4). In models $FLR_{D'Urso03}$ [8] and $FLR_{Coppi06}$ [5], the spreads of estimated dependent variables are only determined by the centres of the estimated dependent variables. This limits the ability of $FLR_{D'Urso03}$and $FLR_{Coppi06}$ to model the spreads of the dependent variables by independent variable (refer to Section 3.3).

Although solutions proposed in [11,12,17] also alleviate *the spreads increasing problem*, these solutions still cannot model a *decreasing* trend in the spreads of the observed dependent variables, as the magnitudes of the independent variables increase. For example, in these models [11,12,17], if the independent variables are crisp, the spreads of the estimated dependent variables can only be a constant (refer to Section 3), even though the spreads of the observed dependent variables are decreasing with the increase of the magnitudes of the independent variables, as shown in Example 1.

Example 1. In Table 1, the independent variable is the height of the male candidates; and the fuzzy dependent variable measures how a candidate's height belongs to the concept *high*. L-type fuzzy numbers in the form of (m_y, α_y) are used to describe *high* (for a detailed description of L-type fuzzy number, refer to Section 2). m_y is the centre of a fuzzy number, which measures the possibility of a given candidate's height belonging to *high*. In this example, m_y is not greater than 1. α_y is the spread of a fuzzy number, which describes the vagueness of m_y. The taller a candidate's height is, closer the possibility of the candidate's height is to 1, and lessens the vagueness of the candidate's height belonging to *high*. However, it is difficult to model this relationship between the candidates' heights and *high* by model FLR_{FF}, because of the *spreads increasing problem* in model FLR_{FF}, which is that the estimated dependent variables can only increase with the magnitudes of the independent variables. Moreover, neither the models proposed in [11,12] nor the model proposed in [17] can capture the relationship between height and *high*, because in these models, when the independent variables are crisp, the spread of the estimated dependent variable can only be a constant (refer to Section 3), which is not true for dataset1 in Table 1.

Note that another problem of modelling the relationship between the candidates' heights and *high* by FLR_{FF} is that the estimated spreads of *high* maybe negative, since the relationship between the candidates' heights and *high* is not strictly linear. When the heights are greater than 2.1, the spreads of observed *high* stop decreasing and the spreads of estimated *high* are negative. Following the arguments in D'Urso [8] and Coppi et al. [5], negative predicted spreads can be interpreted as a lack of uncertainty and set to 0.