Available online at www.sciencedirect.com
ScienceDirect
Fuzzy Sets and Systems 157 (2006) 2627 – 2647
FUZZY sets and systems
www.elsevier.com/locate/fss

Goodness of fit and variable selection in the fuzzy multiple linear regression

Pierpaolo D'Urso*, Adriana Santoro

Dipartimento di Scienze Economiche, Gestionali e Sociali, Università degli Studi del Molise, Via De Sanctis, 86100 Campobasso, Italy

Available online 5 June 2006

Abstract

In performing a fuzzy multiple linear regression model, important topics are: to measure the fitting quality of the model and to find the "best" set of input variables that explain the variation in the observed system responses. In this paper, by considering an exploratory approach, to express the quality of fit of a fuzzy linear regression model, a coefficient of multiple determination R^2 for symmetrical fuzzy variable has been suggested. Furthermore, for overcoming the inconveniences of R^2 an adjusted version of R^2 (denoted by $\overline{R}^2$) has been defined. For measuring the fitting performances of the estimated model, a fuzzy extension of another goodness of fit measure, the so-called Mallows index (C_p), has been considered. All the proposed fitting measures have been utilized for selecting suitably the input variables of a fuzzy linear regression model. To this purpose, some variable selection procedures based on R^2, $\overline{R}^2$ and C_p have been suitably extended in a fuzzy framework. To explain the efficacy of the goodness of fit measures and the variable selection criteria some examples are also shown.

Keywords: Fuzzy linear regression analysis; Crisp input variables; Symmetrical fuzzy output variable; Least-squares approach; Exploratory approach; Fitting measure; Deviance decomposition; Coefficient of multiple determination; Adjusted coefficient of multiple determination; Mallows measure; Variable selection criteria

1. Introduction

Starting from a fuzzy approach and adopting an exploratory formalization, we focus our interest on the goodness of fit measurement and on the variable selection criteria in the fuzzy linear regression with symmetrical membership functions. Then, we assume that the regression model is formalized in a fuzzy manner. For this reason, the fitting measures and the selection variable criteria are formalized by assuming that the output variable of the regression model is only fuzzy (i.e. the only source of uncertainty of the model is the fuzziness of the output variable (i.e. the uncertainty connected to the imprecision of the output data)).

In the body of fuzzy literature, there are different works on goodness of fit measures. For instance, in a fuzzy-random framework, Gil et al. [14] suggested a goodness of fit test with fuzzy observations. To evaluate the goodness of fit of a fuzzy regression model, Toyoura and Watada [29] proposed two fitting indices. Chang [3] suggested a hybrid reliability measure for the model with fuzzy coefficients. Xu and Li [31] defined a performance index for a Gaussian fuzzy regression model. D'Urso and Gastaldi [10] proposed a fitting measure for the polynomial regression model.

* Corresponding author.
E-mail addresses: durso@unimol.it, pierpaolo.durso@uniroma1.it (P. D'Urso), santoro@unimol.it (A. Santoro).

0165-0114/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.fss.2005.03.015

Sadeghpour Gildeh and Gien [27] provided a goodness of fit index to evaluate the fitness between the observed values and the estimated values in a fuzzy regression model. Coppi and D'Urso [8] suggested two absolute fitting measures for comparing, respectively, the observed and estimated centers and the observed and estimated spreads of the fuzzy output variable. D'Urso and Giordani [11] proposed a fitting measure for linear regression analysis with fuzzy multivariate response. Another fitting measure has been suggested by Hong et al. [15] in a ridge regression learning framework (i.e. ridge regression based on support vector machine (SVM) [30]).

The aim of this paper is twofold. On one side, we want to measure the goodness of fit of a multiple linear regression model with symmetrical fuzzy output variable and crisp input variables by means of the coefficient of multiple determination R^2, the adjusted coefficient of multiple determination ($\overline{R}^2$) and the Mallows measure (C_p). On the other side, we utilize the previous fitting measures for building, in a fuzzy framework, some variable selection procedures.

The paper is structured as follows. In Section 2, we define mathematically the symmetric fuzzy variables and a distance measure for this kind of variables. Successively, in Section 3, we explain the fuzzy multiple linear regression model suggested by Coppi and D'Urso [4]. In Section 4, we define suitable goodness of fit measures with respect to previous fuzzy regression and proof some theoretical results connected to fitting measures. In Section 5, we show some variable selection criteria for the previous fuzzy regression model. Applicative examples are discussed in Section 6 and final remarks and perspectives of research are presented in Section 7.

2. Fuzzy variables: mathematical formalization and distance measures

In this section, we introduce the concept of symmetrical fuzzy variable, by assuming that the only source of uncertainty is the fuzziness (for the case in which we have a double source of uncertainty, randomness and fuzziness, i.e. for fuzzy random variable, see, e.g., [20,21,26,19,14,24,25,16]).

2.1. Symmetrical fuzzy variables: definition

In several substantive applications, the most utilized class of fuzzy variable is the so-called *symmetrical* fuzzy variable. Usually, a symmetrical fuzzy variable is denoted by $\widetilde{Y} = (m, l)$, where m denotes the *center* and l the *left* and *right spreads* (the spreads are symmetrical) with the following *membership function*:

$$\mu(\alpha) = L\left(\frac{m-\alpha}{l}\right), \quad m - l \leqslant \alpha \leqslant m + l \;\; (l > 0), \tag{2.1}$$

where L is a decreasing "shape" function from $\Re^+$ to [0,1] with $L(0) = 1$; $L(z) < 1$ for all $z > 0$; $L(z) > 0$ for all $z < 1$; $L(1) = 0$ (or $L(z) > 0$ for all z and $L(+\infty) = 0)L(z) = L(-z)$; furthermore, the shape function L is symmetric [34].

Particular cases of (2.1) are

$$\begin{aligned} L\left(\frac{m-\alpha}{l}\right) &= \max\left\{0, 1 - \left|\frac{m-\alpha}{l}\right|^q\right\}, \quad l > 0, \;\; q > 0, \\ L\left(\frac{m-\alpha}{l}\right) &= \exp\left(-\left|\frac{m-\alpha}{l}\right|^q\right), \\ L\left(\frac{m-\alpha}{l}\right) &= 1 \Big/ \left(1 + \left|\frac{m-\alpha}{l}\right|^q\right). \end{aligned} \tag{2.2}$$

From (2.1), we can define various typologies of symmetrical fuzzy variables: e.g., the symmetric triangular, normal, parabolic and square root fuzzy variables. Each case takes into account a different level of uncertainty around the centers of the fuzzy variable. In Fig. 1, for each case, we show the geometric representation of this kind of fuzzy variables.

2.2. Symmetrical fuzzy variables: distance measure

In the body of literature, several distance measures between fuzzy data have been suggested (see, e.g., [1,4,25,32,33]).