Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 287–313
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Fuzzy clusterwise linear regression analysis with symmetrical fuzzy output variable

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 27 June 2006

Abstract

The traditional regression analysis is usually applied to homogeneous observations. However, there are several real situations where the observations are not homogeneous. In these cases, by utilizing the traditional regression, we have a loss of performance in fitting terms. Then, for improving the goodness of fit, it is more suitable to apply the so-called clusterwise regression analysis. The aim of clusterwise linear regression analysis is to embed the techniques of clustering into regression analysis. In this way, the clustering methods are utilized for overcoming the heterogeneity problem in regression analysis. Furthermore, by integrating cluster analysis into the regression framework, the regression parameters (regression analysis) and membership degrees (cluster analysis) can be estimated simultaneously by optimizing one single objective function. In this paper the clusterwise linear regression has been analyzed in a fuzzy framework. In particular, a fuzzy clusterwise linear regression model (FCWLR model) with symmetrical fuzzy output and crisp input variables for performing fuzzy cluster analysis within a fuzzy linear regression framework is suggested. For measuring the goodness of fit of the suggested FCWLR model with fuzzy output, a fitting index is proposed. In order to illustrate the usefulness of FCWLR model in practice, several applications to artificial and real datasets are shown.

Keywords: Heterogeneous data; Fuzzy clusterwise linear regression analysis; Crisp input; Symmetrical fuzzy output; Fuzzy clustering; Goodness of fit; Cluster validity

1. Introduction

In a statistical perspective, the regression analysis is utilized for studying the dependence relationship between a real phenomenon (dependent variable or output variable) and other (explanatory) real phenomena (explanatory variables or independent variables or input variables). The traditional regression analysis can be suitably utilized in the case of homogeneous observations. However, in many real cases, there are several situations where the observations are not homogeneous. In these cases, by utilizing the traditional regression, we have a loss of fitting performance of the regression model. In order to improve the goodness of fit, it is more suitable to utilize the so-called clusterwise regression analysis, in which we embed the techniques of clustering into regression analysis. In this way, the clustering methods are utilized for overcoming the heterogeneity problem in regression analysis. For explaining more clearly the aim and the real usefulness of the clusterwise regression analysis, we consider the following explicative example of clusterwise on a market segmentation problem in business, drawn by Lau et al. (1999): "The manager collects a sample of the sales and income data from a set of costumers. If the costumers have homogeneous income elasticity (i.e., the

* Corresponding author. Tel.: +39 0874 404407; fax: +39 0874 311124.
E-mail addresses: durso@unimol.it, pierpaolo.durso@uniroma1.it (P. D'Urso), santoro@unimol.it (A. Santoro).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.06.001

regression coefficient β), β can simply be estimated by regression of sales on income. In real business, costumers are heterogeneous and income elasticity will vary with customers of different clusters in the sample. The major tasks for the manager are: (i) use the income elasticity as the basis to divide customers into mutually exclusive segments, (ii) estimate the average income elasticity for each segment, (iii) identify the members of each segment. If we ignore the income elasticity differences among segments, the income elasticity estimated from the regression of sales on income will certainly be biased and inaccurate. In other words, if we want to model the parameter heterogeneity in the traditional regression, the appropriate statistical analysis will involve the simultaneous applications of the cluster analysis and regression model. One straightforward approach is the two stage method. In stage 1, we apply cluster analysis to the dataset to divide customers into segments. In stage 2, we perform regression for each segment to estimate the income elasticity. The problem is that the functions optimized in stages 1 and 2 are two different objective functions which are not necessarily related. A better formulation is to integrate the cluster analysis into regression framework, so that the income elasticities and segment membership parameters can be estimated simultaneously by optimizing one single objective function".

In the body of literature, there are many theoretical works on clusterwise regression analysis (see, for example, De Sarbo and Cron, 1988; De Sarbo et al., 1989; De Veaux, 1989; Hathaway and Bezdek, 1993; Hathaway et al., 1996; Hennig, 2000, 2003; Hong and Chao, 2002; Lau et al., 1999; Leşki, 2004; Preda and Saporta, 2005; Quandt and Ramsey, 1978; Shao and Wu, 2005; Spath, 1979; Yang and Ko, 1997; Van Aelest et al., 2006; Wedel and De Sarbo, 1995). Furthermore, the clusterwise regression analysis finds application in several fields, such as market segmentation and business, socio-economics, biology, engineering, and so on (see, for instance, Aurifeille and Quester, 2003; De Sarbo and Cron, 1988; Hosmer, 1974; Lau et al., 1999; Wedel and Steenkamp, 1991).

In this paper the clusterwise linear regression is analyzed in a fuzzy framework. In particular, we propose a fuzzy clusterwise linear regression model (FCWLR model) with symmetrical fuzzy output and crisp input variables for performing fuzzy cluster analysis within a fuzzy linear regression framework. We build our FCWLR model by considering, simultaneously, the Bezdek's approach to fuzzy cluster analysis (Bezdek, 1981) and the linear regression model with fuzzy output variable ($\tilde{Y}$) and crisp explanatory variables ($X_1, \ldots, X_k$) suggested by Coppi and D'Urso (2003):

$$\begin{cases} m_i = m_i^* + e_i, \quad m_i^* = \mathbf{x}_i'\mathbf{a}, \\ {}_{(-)}s_i = {}_{(-)}s_i^* + {}_{(-)}\varepsilon_i, \quad {}_{(-)}s_i = m_i - l_i, \quad {}_{(-)}s_i^* = m_i^* - l_i^*, \quad l_i^* = m_i^* b + d, \\ {}_{(+)}s_i = {}_{(+)}s_i^* + {}_{(+)}\varepsilon_i, \quad {}_{(+)}s_i = m_i + l_i, \quad {}_{(+)}s_i^* = m_i^* + l_i^*, \end{cases}$$

where $\mathbf{x}_i'$ is $(1 \times (k+1))$-vector containing the scalar 1 and the values of the k crisp input variables observed on the ith unit, m_i, m_i^* are, respectively, the ith *observed center* and the ith *interpolated center*, l_i, l_i^* are, respectively, the ith *observed spreads* and the ith *interpolated spreads*, $\mathbf{a}$ is $((k+1) \times 1)$-vector of regression parameters for m_i, b, d are the regression parameters for the other models, and e_i, ${}_{(-)}\varepsilon_i$, ${}_{(+)}\varepsilon_i$ are the residuals.

In matrix form, we can write the previous model as follows:

$$\begin{cases} \mathbf{m} = \mathbf{m}^* + \mathbf{e}, \quad \mathbf{m}^* = \mathbf{X}\mathbf{a}, \\ {}_{(-)}\mathbf{s} = {}_{(-)}\mathbf{s}^* + {}_{(-)}\boldsymbol{\epsilon}, \quad {}_{(-)}\mathbf{s} = \mathbf{m} - \mathbf{l}, \quad {}_{(-)}\mathbf{s}^* = \mathbf{m}^* - \mathbf{l}^*, \quad \mathbf{l}^* = \mathbf{m}^* b + \mathbf{1}d, \\ {}_{(+)}\mathbf{s} = {}_{(+)}\mathbf{s}^* + {}_{(+)}\boldsymbol{\epsilon}, \quad {}_{(+)}\mathbf{s} = \mathbf{m} + \mathbf{l}, \quad {}_{(+)}\mathbf{s}^* = \mathbf{m}^* + \mathbf{l}^*, \end{cases} \tag{1.1}$$

where $\mathbf{1}$ is $(n \times 1)$-vector of all 1's, $\mathbf{X}$ is $(n \times (k+1))$-matrix containing the vector $\mathbf{1}$ concatenated to k crisp input variables, $\mathbf{m}, \mathbf{m}^*$ are, respectively, $(n \times 1)$-vectors of *observed centers* and *interpolated centers*, $\mathbf{l}, \mathbf{l}^*$ are, respectively, $(n \times 1)$-vectors of *observed spreads* and *interpolated spreads*, $\mathbf{a}$ is $((k+1) \times 1)$-vector of regression parameters for $\mathbf{m}$, b, d are, respectively, the regression parameters for the other models, and $\mathbf{e}$, ${}_{(-)}\boldsymbol{\epsilon}$, ${}_{(+)}\boldsymbol{\epsilon}$ are, respectively, $(n \times 1)$-vectors of residuals.

Notice that, the above fuzzy regression model is based on three linear models. The first one interpolates the centers of the fuzzy observations, the second and third ones yield the *lower* and *upper bounds* (centers $\pm$ spreads), by building other linear models over the first one. The model is hence capable to take into account possible linear relations between the size of the spreads and the magnitude of the estimated centers. This is often the case in realistic applications, where dependence among centers and spreads is likely to occur (for instance, the uncertainty or fuzziness concerning a measurement may depend on its magnitude) (Coppi and D'Urso, 2003; D'Urso, 2003).

Furthermore, in order to test the performance of the proposed FCWLR we suggest a suitable fitting measure, i.e., the R^2 coefficient.