Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 52 (2008) 1500–1515
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Centre and Range method for fitting a linear regression model to symbolic interval data

Eufrásio de A. Lima Neto, Francisco de A.T. de Carvalho*

Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n - Cidade Universitária - CEP 50740-540-Recife (PE), Brazil

Received 2 June 2006; received in revised form 20 April 2007; accepted 20 April 2007
Available online 27 April 2007

Abstract

This paper introduces a new approach to fitting a linear regression model to symbolic interval data. Each example of the learning set is described by a feature vector, for which each feature value is an interval. The new method fits a linear regression model on the mid-points and ranges of the interval values assumed by the variables in the learning set. The prediction of the lower and upper bounds of the interval value of the dependent variable is accomplished from its mid-point and range, which are estimated from the fitted linear regression model applied to the mid-point and range of each interval value of the independent variables. The assessment of the proposed prediction method is based on the estimation of the average behaviour of both the *root mean square error* and the *square of the correlation coefficient* in the framework of a Monte Carlo experiment. Finally, the approaches presented in this paper are applied to a real data set and their performance is compared.

Keywords: Symbolic data analysis; Regression models; Symbolic interval data

1. Introduction

Predicting the behaviour of a (dependent) variable in relation to other (independent) variables that are thought be responsible for the variability of the former is an important task in data analysis, pattern recognition, data mining, machine learning, etc. The classical regression model is used to predict the values of a dependent quantitative variable in relation to the values of independent quantitative variables. However, to fit this model to the data, it is necessary to estimate a vector $\boldsymbol{\beta}$, of parameters from the data vector $\mathbf{Y}$ and the model matrix $\mathbf{X}$, assumed with complete rank p. The estimation using the *least square method* does not require any probabilistic hypothesis on the variable Y. This method consists of minimising the sum of the square of residuals. A detailed study on linear regression models for usual quantitative data can be found in Draper and Smith (1981), Montgomery and Peck (1982), Scheffé (1959), as well as others.

In regression analysis of usual data, the items are usually represented as a vector of quantitative measurements for which each column represents a variable. In practice, however, this model is too restrictive to represent complex data. In order to take into account the variability and/or uncertainty inherent to the data, variables must assume sets of categories

* Corresponding author. Tel.: +55 81 21268430; fax: +55 81 21268438.
E-mail addresses: ealn@cin.ufpe.br (E. de A. Lima Neto), fatc@cin.ufpe.br (F. de A.T. de Carvalho).

0167-9473/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2007.04.014

or intervals, possibly even with frequencies or weights. Such type of data have been mainly studied in *symbolic data analysis* (SDA), which is a domain in the area of knowledge discovery and data management related to multivariate analysis, pattern recognition and artificial intelligence. The aim of SDA is to provide suitable methods (clustering, factorial techniques, decision trees, etc.) for managing aggregated data described by multi-valued variables, for which the cells of the data table contain sets of categories, intervals or weight (probability) distributions (Bock and Diday, 2000).

As mentioned above, the items are usually represented as a vector of quantitative measurements. However, due to recent advances in information technologies, it is now common to record interval data. In the framework of SDA, interval data appear when the observed values of the variables are intervals from the set of real numbers $\Re$. Interval data arise in situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, etc. Another source of interval data is the aggregation of huge data-bases into a reduced number of groups, the properties of which are described by symbolic interval variables. Therefore, tools for symbolic interval data analysis are very much required.

Currently, different approaches have been introduced to analyse symbolic interval data. Regarding univariate statistics, Bertrand and Goupil (2000) and Billard and Diday (2003) introduced central tendency and dispersion measures suitable for symbolic interval data. DeCarvalho (1995) proposed histograms for symbolic interval data. Factorial methods for analysing symbolic interval data have also been considered in SDA. Cazes et al. (1997) and Lauro and Palumbo (2000) introduced principal component analysis methods suitable for symbolic interval data. Palumbo and Verde (2000) and Lauro et al. (2000) generalised factorial discriminant analysis (FDA) to symbolic interval data. Regarding classification, Ichino et al. (1996) introduced a symbolic classifier as a region-oriented approach for symbolic interval data. Rasson and Lissoir (2000) presented a symbolic kernel classifier based on dissimilarity functions suitable for symbolic interval data. Périnel and Lechevallier (2000) proposed a tree-growing algorithm for classifying symbolic interval data.

SDA provides a number of clustering methods for symbolic data. These methods differ with regard to the type of symbolic data considered, their cluster structures and/or the clustering criteria considered. With hierarchical clustering methods, an agglomerative approach has been introduced that forms composite symbolic objects using a join operator whenever mutual pairs of symbolic objects are selected for agglomeration based on minimum dissimilarity (Gowda and Diday, 1991) or maximum similarity (Gowda and Diday, 1992). Ichino and Yaguchi (1994) defined generalised Minkowski metrics for mixed feature variables and present dendrograms obtained from the application of standard linkage methods for data sets containing numeric and symbolic feature values. Chavent (1998) proposed a divisive clustering method for symbolic data that simultaneously furnishes a hierarchy of the symbolic data set and a monothetic characterisation of each cluster in the hierarchy. Guru et al. (2004) introduced agglomerative clustering algorithms based on similarity functions that are multi-valued and non-symmetric.

Regarding partitioning clustering algorithms for symbolic interval data, Bock (2002) proposed several clustering algorithms for symbolic data described by interval variables, and presented a sequential clustering and updating strategy for constructing a self-organising map (SOM) to visualise symbolic interval data. Chavent and Lechevallier (2002) proposed a dynamic clustering algorithm for interval data where the class representatives are defined by an optimality criterion based on a modified Hausdorff distance. Souza and De Carvalho (2004) presented partitioning clustering methods for interval data based on (adaptive and non-adaptive) city-block distances. More recently, De Carvalho et al. (2006) proposed an algorithm using an adequacy criterion based on adaptive Hausdorff distances.

This paper addresses linear regression models for predicting symbolic interval data. Billard and Diday (2000) presented the first approach to fitting a linear regression model to symbolic interval data sets from an SDA perspective. Their approach consists of fitting a linear regression model to the mid-points of the interval values assumed by the symbolic interval variables in the learning set and applies this model to the lower and upper bounds of the interval values of the independent symbolic interval variables to be predicted the lower and upper bounds of the interval value of the dependent variable, respectively.

This paper introduces a Centre and Range approach to fitting a linear regression model to symbolic interval data. The probabilistic assumptions that involve the linear regression model theory for classical data will not be considered in the case of symbolic data (symbolic interval variables), as this remains an open research topic. Thus, the problem will be investigated as an optimisation problem, in which we seek to minimise a predefined criterion.

In Table 1, we show the criteria and models that represent the three approaches presented in this paper.

The first method (Billard and Diday, 2000) is based on the minimisation of the mid-point error, since $(\varepsilon_{Li}+\varepsilon_{Ui})/2=\varepsilon_i^c$.