Engineering Applications of Artificial Intelligence 26 (2013) 564–573

Contents lists available at SciVerse ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Robust regression with application to symbolic interval data

Roberta A.A. Fagundes [a], Renata M.C.R. de Souza [a,*], Francisco José A. Cysneiros [b]

[a] Centro de Informática, Universidade Federal de Pernambuco, Av. Jornalista Anibal Fernandes, s/n, Cidade Universitária, CEP 50740-560 Recife (PE), Brazil
[b] Departamento de Estatística, Centro de Ciências Exatas, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire, s/n, Cidade Universitária, CEP 50740-540 Recife (PE), Brazil

ARTICLE INFO

Article history:
Received 3 June 2011
Received in revised form
21 December 2011
Accepted 1 May 2012
Available online 7 June 2012

Keywords:
Robust regression
Symbolic data analysis
Interval-valued data
Outliers

ABSTRACT

This paper presents a robust regression model that deals with cases that have interval-valued outliers in the input data set. Each interval of the input data is represented by its range and midpoint and the fitting to interval-valued data is not sensible in the presence of midpoint and/or range outliers on the interval response. The predictions of the lower and upper bounds of new intervals are performed and simulation studies are carried out to validate these predictions. Two applications with real-life interval data sets are considered. The prediction quality is assessed by a mean magnitude of relative error calculated from a test data set.

1. Introduction

Regression analysis is one of the most widely used techniques in engineering, management and many other fields. The widespread availability of regression software has greatly expanded its application in recent years. A problem that is frequently encountered in the application of regression is the presence of one or more outliers in the data. The outliers give valuable information about the fit of the model on data quality and they are indicative for atypical phenomena. Outlier observations are unusual observations in a data set that substantially differ from the rest. Such data may have a strong influence on the statistical analysis, particularly in regression models based on least square estimators. In view of such potential impact on the fitted model, identifying outlying observations is an important concern of the regression model building process. That is, occasionally certain observations will have a disproportionate effect on the precision of the parameter estimates, and/or the overall predictive ability of the model.

Robust regression is an important technique for analyzing data that are contaminated with outliers. A robust estimation technique is essentially a method which tolerates the presence of data atypical points. This technique has been developed as an alternative to least squares estimation in the presence of outliers. The primary purpose of robust regression techniques is to fit a model that describes the information in the majority of the data. This general definition implies that this technique should perform well on both messy data (with outliers) and on clean data (without outliers).

The statistical treatment of interval data has been considered in the context of *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 a comprehensive way to summarize data sets by means of symbolic data resulting in a smaller and more manageable data set which preserves the essential information, and its subsequent analysis by means of the generalization of the exploratory data analysis and data mining techniques to symbolic data. Symbolic data allow multiple values for each variable. Those new variables (set-valued, interval-valued and histogram-valued) make it possible to hold data intrinsic variability and/or uncertainty from the original data set as shown in Diday and Noirhomme-Fraiture (2008).

The process of obtaining symbolic data starts with the extraction of knowledge from data sets as in data mining process in order to provide symbolic descriptions. In practice, symbolic descriptions are mathematically modeled by a generalization process applied to a set of individuals described by classical data (categorical or quantitative values). According to Diday and Noirhomme-Fraiture (2008), overgeneralization problems can arise when extreme values are presented in classical descriptions and these values are in fact outliers or when the set of individuals to generalize is in fact composed of subsets of different distributions. In classical data analysis, sometimes specialists after the identification of point outliers prefer to discard outliers before computing the line that best fits the data under investigation. In symbolic data analysis, a single interval outlier may represent an aggregation of a group of measurements that contain valuable

* Corresponding author. Tel.: +55 81 21268430; fax: +55 81 21268438.
E-mail address: rmcrs@cin.ufpe.br (R.M.C.R. de Souza).

0952-1976/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.engappai.2012.05.004

information about the process being analyzed. Therefore, it is not recommendable to discard interval outliers because these observations can cause great loss of information.

This paper introduces a robust regression for estimation and prediction in the presence of atypical interval data. The outline of this work is as follows: Section 2 shows the motivation and related works for linear regression model with interval symbolic data. Section 3 describes the robust regression for interval data proposed in this paper. Section 4 carries out a simulation study and an analysis with two real-life interval data sets to show the performance of the introduced approach in comparison with a linear regression method for interval data of the literature of symbolic data analysis. Section 5 concludes the work.

2. Motivation and related works

Regarding that an interval can be represented by its center (midpoint) and range, interval outliers can be identified by investigating if there are point outliers in the respective midpoint and range data sets. Fig. 1 displays mushroom and football data sets in which interval-valued data outliers are presented. The mushroom data set (Fig. 1(a)) consists of 23 species described by two predictor interval variables that are stipe length and stipe thickness and the response interval variable that is pileus cap width. From this data set, we can observe that there are intervals that have unusual pileus cap coordinate. These unusual intervals are generated by the presence of point outliers in their midpoints and ranges. The football data set (Fig. 1(b)) provides information about professional football players in 20 French teams. Each player is described by the height and age independent interval variables and the weight dependent interval variable. Again, from this interval data set, we can verify that there are midpoint outliers that have unusual weight coordinate.

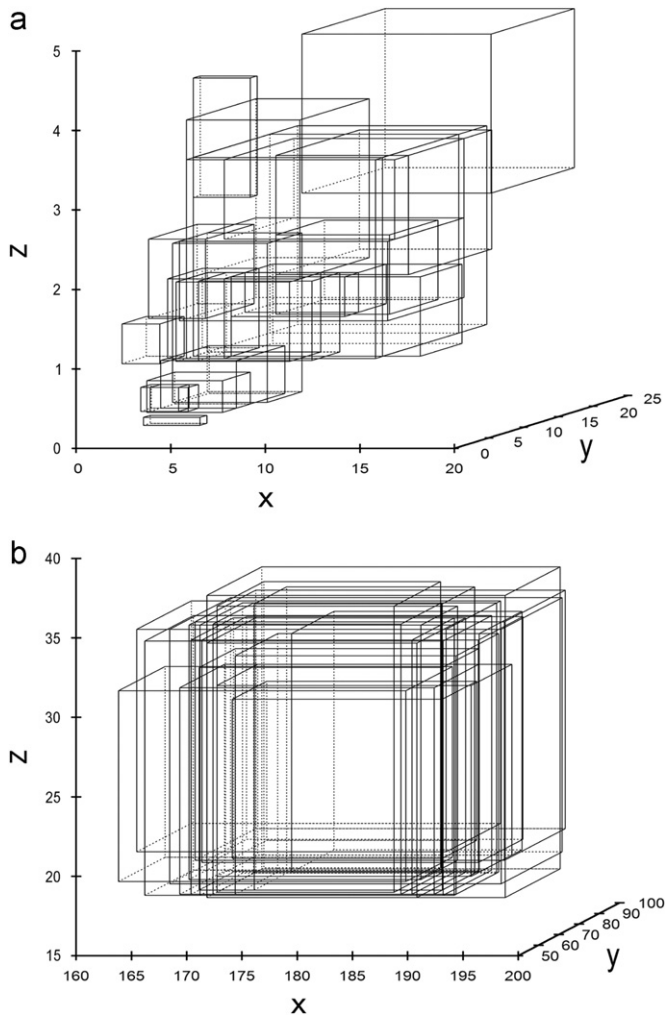

Fig. 1. Mushroom (a) and football (b) interval-valued data sets. (a) 3D scatter plot: stipe length (X), pileus cap width (Y) and stipe thickness (Z). (b) 3D scatter plot: height (X), weight (Y) and age (Z).

In the framework of regression models for symbolic interval data, several models have been introduced. Most of these models consider that their parameters are estimated by the minimization squared error criterion. Billard and Diday (2000) presented an approach to extend the classic linear regression model (*CLRM*) to symbolic interval data by fitting the method of least squares to the midpoints of the interval valued data assumed by the interval variables. Billard and Diday (2002) proposed another approach that fits two independent *CLRM* on the lower and upper bounds of the intervals. Billard and Diday (2006) also included explanatory variables as well as hierarchical variable structure into symbolic regression framework. Maia and De Carvalho (2008) showed a least absolute deviation regression model suitable for manage interval-valued data and modeling based on regression $L1$. Lima Neto and De Carvalho (2008) proposed the midpoint and range method for fitting the *CLRM* to interval valued data as an improvement in comparison with the methods presented in Billard and Diday (2000, 2002). Lima Neto and De Carvalho (2010) proposed an approach that fits a constrained linear regression model on the midpoint and range of the interval values in order to ensure mathematical coherence between the predicted values of the lower and upper boundaries of the interval. Concerning interval-valued time series, Maia et al. (2008) presented approaches to interval-valued time series forecasting.

It is well known in the literature that a least squares model (as in Lima Neto and De Carvalho, 2008) is sensitive to outliers. In the previous work (Fagundes et al., 2009), we presented a robust prediction method for symbolic interval data based on the simple linear robust regression methodology for intervals. Robust regression uses the reweighted least squares method, where the weights are estimated iteratively along the process. In this case, the observations with large values residuals have small weights. Therefore, these values produce small influence in the estimation of parameters. The method is validated with a simulation scenario regarding interval outliers on midpoint.

Here, we propose a generalization of the previous work to the multiple linear regression methodology for intervals. In addition, definitions of interval outliers are given and simulation studies considering several scenarios of outliers on the midpoint and/or range are presented. This model consists of fitting two linear robust regression models to, respectively, the midpoint and range of the intervals. The prediction of an interval is based on a combination between the fitted models. The proposed method is compared with the model given by Lima Neto and De Carvalho (2008) since both methods do not assume distribution probability for errors. This fact allows to construct models more flexible to real data applications.

In the context of regression models for interval data that assume probability distributions for the errors, Domingues et al. (2010) proposed a methodology of analysis for interval data based on a symmetrical linear regression. In this model the prediction of the lower and upper bounds of the intervals is not damaged in the presence of interval outliers defined by midpoint outliers. Lima Neto et al. (2011) introduce the bivariate symbolic regression model for interval data based on generalized linear model theory. Souza et al. (2011) introduced multi-class logistic linear regression models for the lower and upper bounds of the intervals conjointly and separately.