ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Constrained center and range joint model for interval-valued symbolic data regression ☆

Junpeng Guo *, Peng Hao

College of Management and Economics, Tianjin University, Tianjin 300072, China

HIGHLIGHTS

- We introduce a constrained center and range joint model to fit linear regression to interval-valued symbolic data.
- We apply both the center and range of the interval to fit the linear regression model.
- The model avoids the negative value of the range of the predicted dependent interval variable.
- We adopt overlapping constraints to improve the model's prediction accuracy.

ARTICLE INFO

Article history:
Received 7 September 2016
Received in revised form 27 May 2017
Accepted 12 June 2017
Available online xxxx

Keywords:
Interval-valued data
Linear regression model
Constrained center and range joint model
Least squares estimation

ABSTRACT

A constrained center and range joint model to fit linear regression to interval-valued symbolic data is introduced. This new method applies both the center and range of the interval to fit a linear regression model, and avoids the negative value of the range of the predicted dependent interval variable by adding nonnegative constraints. To improve prediction accuracy it adopts overlapping constraints. Using a little algebra, it is constructed as a special case of the least squares with inequality (LSI) problem and is solved with a Matlab routine. The assessment of the proposed prediction method is based on an estimation of the average root mean square error and accuracy rate. In the framework of a Monte Carlo experiment, different data set configurations take into account the rich or lack of error, as well as the slope with respect to the dependent and independent variables. A statistical t-test compares the performance of the new model with that of four previously reported methods. Based on experiment results, it is outlined that the new model has better fitness. An analysis of outliers is performed to determine the effects of outliers on our proposal. The proposed method is illustrated by analyses of data from two real-life case studies to compare its performance with those of the other methods.

1. Introduction

The exploration of the relationship between dependent and independent variables is an important task in many contexts, including data analysis, pattern recognition, data mining, and machine learning. Regression analysis is a common method for analyzing the relationship of dependent and independent variables. The traditional regression model, which is mainly applied to traditional point data, is generally used to predict the behavior of a dependent variable Y as a function of other

☆ A software is attached in the appendix.

* Correspondence to: Tianjin University, No.92, Weijin Road, Nankai District, Tianjin, 300072, China. Fax: +86 22 87402183.
E-mail address: guojp@tju.edu.cn (J. Guo).

http://dx.doi.org/10.1016/j.csda.2017.06.005
0167-9473/

independent variables X that are responsible for the variability of Y. To fit this model, it is necessary to estimate a parameter vector $\boldsymbol{\beta}$, based on the vector $\boldsymbol{Y}$ and matrix $\boldsymbol{X}$.

However, most information cannot be represented as real point data, as in the cases of missing, censored, or ambiguous data. As such, the need arises for regression with interval-valued data. Interval-valued data, a kind of symbolic data, implements data dimension reduction by "data packaging", resulting in data analyses with less computational complexity. Symbolic data analysis (SDA) has been discussed by Bock and Diday (2000), Billard and Diday (2003), Billard (2007), Diday and Noirhomme-Fraiture (2008), Brito and Noirhomme-Fraiture (2011), Chiun-How et al. (2014), and Débora et al. (2016), among others.

Previous research has addressed interval-valued parametric regression whose parameter β is interval-valued, i.e., the coefficient β is interval-valued whereas variables Y and X are either real or interval-valued. Such instances have been addressed by Tanaka et al. (1982), Hojati et al. (2005), Savic and Pedrycz (1991), Tanaka et al. (1989), Tanaka and Ishibuchi (1991), Sakawa and Yano (1992), Chen and Hsueh (2007, 2009), and Hladík and Černý (2012). In SDA, however, regression model parameters of interval-valued variables are typically real numbers, i.e., the regression coefficient β is real whereas variables Y and X are interval-valued data.

Symbolic interval-valued regression involves parametric and nonparametric regression algorithms. Being applicable to both linear and nonlinear regression problems is the valuable advantage of nonparametric regression methods. Based on CRM (Lima Neto and De Carvalho, 2008) and a reverse fitting method (Buja et al., 1989; Friedman and Stuetzle, 1981), Lim (2016) developed the nonparametric additive model (CRAM) to estimate the midpoint and radius of the response variables. By simulation and example verification, this method has the better fitness than both CRM (Lima Neto and De Carvalho, 2008) and SCM (Xu, 2010). Roberta et al. (2014) applies kernel regression method, which employs the Gauss kernel function, to fit interval-valued data. In this paper, we determine the regression method that best obtains the model parameters.

Due to their salient performance in nonlinear situations, machine learning methods have also been applied to interval-valued data regression. Examples include interval multi-layer perceptrons (iMLP) (San Roque et al., 2007), exponential smoothing (Arroyo et al., 2007), the linear autoregressive integrated moving average—nonlinear artificial neural network (ARIMA-ANN) (Maia et al., 2008), Holt-MLP (Maia and De Carvalho, 2011), firefly algorithm—multidimensional support vector regression (FA-MSVR) (Xiong et al., 2014), and the MSVR-vector error correction model (VECM) (Xiong et al., 2015). The common feature of machine learning processes in interval regression is to estimate the upper and lower bounds (respectively, $\overline{y}$ and $\underline{y}$) without assuming any constraint of $\overline{y} \geq \underline{y}$.

As a rough classification, we can identify at least three kinds of linear interval-valued regression algorithms, including least square (LS) (e.g., Billard and Diday, 2000; Domingues et al., 2010; Lima Neto and De Carvalho, 2008, 2010), set arithmetic (e.g., Blanco-Fernandez et al., 2011), and probabilistic assumptions (e.g., Ahn et al., 2012; Xu, 2010). Set arithmetic ensures the existence of Hukuhara distance and adopts it as a criterion for building a fitting model. The probabilistic assumptions method takes into account the inner point distribution feature of an interval in a regression model. Essentially, these two kinds of algorithms are based to some degree on the LS method.

With respect to interval regression, since it was first introduced by Billard and Diday (2000), LS algorithms have been studied for a relatively long period of time from Billard and Diday (2000). The authors developed the center method (CM) model and utilized the LS algorithm for the first time to estimate the upper and lower boundaries (respectively,$\overline{y}$ and $\underline{y}$ with $\overline{y} \geq \underline{y}$) of the interval with the same coefficients. Their approach consists of fitting a linear regression model to the midpoint ($y^c = \frac{\overline{y}+\underline{y}}{2}$) of the interval values and applying this model to predict the lower and upper boundaries of the interval value of the dependent variable. However, having the same coefficients of the two models results in the models lacking flexibility and in an inability to characterize real-life situations. The MinMax model (Billard and Diday, 2002) improved the CM model by establishing two models to fit the lower- and upper-bound data series. Compared with end point (EP) of the interval, the expression of the midpoint–radius (MR) separates the uncertainty and the variation tendency, which are represented by the radius ($y^r = \frac{\overline{y}-\underline{y}}{2}$) and midpoint, respectively, and in many practical cases, MR demonstrates more natural results (Boukezzoula et al., 2011). From this perspective Lima Neto and De Carvalho (2008) proposed the center and range method (CRM) to fit a center and radius model by the LS method into a uniform model. But this model does not mathematically ensure that the radius is greater than zero, which induces the case in which the upper bound may be less than the lower. To solve this problem, Lima Neto and De Carvalho (2010) put forward the constrained center and range method (CCRM) to ensure the rationality of the predicted interval. After a nonnegative transformation of the radius parameters, however, the predicted radius may be very unlikely to recover its true original value (i.e., nonnegative predicted radius parameters are only a sufficient rather than necessary condition for a nonnegative predicted radius). Allowing for negative relationships in the radius model, Giordani (2015) introduced a more flexible linear regression method, Lasso-IR, which uses least absolute shrinkage and selection operator (Lasso) constraints (Tibshirani, 1996) in his proposed model. With an optimization solution provided by Gill et al. (1981), Lasso-IR determines the midpoints and radii parameters through an LS approach to improve prediction accuracy. The objective function is defined as the Euclidian distance between the observed and the estimated intervals, which are represented by the midpoints and radii. This kind of objective function also can be found in Trutschnig et al. (2009), Sinova et al. (2012), and Blanco-Fernandez et al. (2011, 2012). Interval regression methods based on midpoint–radius (MR) (Billard and Diday, 2000, 2002) and end point (EP) (Lima Neto and De Carvalho, 2008, 2010; Giordani, 2015) representation are to some degree applications of the LS approach (e.g., Billard and Diday, 2002; Lima Neto and De Carvalho, 2008) or its improvement, whose instances are radius nonnegative constraints (Lima Neto and De Carvalho, 2010; Giordani, 2015).

Please cite this article in press as: Guo, J., Hao, P., Constrained center and range joint model for interval-valued symbolic data regression. Computational Statistics and Data Analysis (2017), http://dx.doi.org/10.1016/j.csda.2017.06.005.