Journal of the Korean Statistical Society 45 (2016) 358–370

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Interval-valued data regression using nonparametric additive models

Changwon Lim [*]

Department of Applied Statistics, Chung-Ang University, Seoul 156-756, Republic of Korea

ARTICLE INFO

Article history:
Received 10 June 2015
Accepted 27 December 2015
Available online 28 January 2016

AMS 2000 subject classifications:
primary 62G08
secondary 62J99

Keywords:
Interval-valued data
Nonparametric additive model
Symbolic data
Penalized regression spline
Generalized cross validation

ABSTRACT

Interval-valued data are observed as ranges instead of single values and frequently appear with advanced technologies in current data collection processes. Regression analysis of interval-valued data has been studied in the literature, but mostly focused on parametric linear regression models. In this paper, we study interval-valued data regression based on nonparametric additive models. By employing one of the current methods based on linear regression, we propose a nonparametric additive approach to properly analyze interval-valued data with a possibly nonlinear pattern. We demonstrate the proposed approach using a simulation study and a real data example, and also compare its performance with those of existing methods.

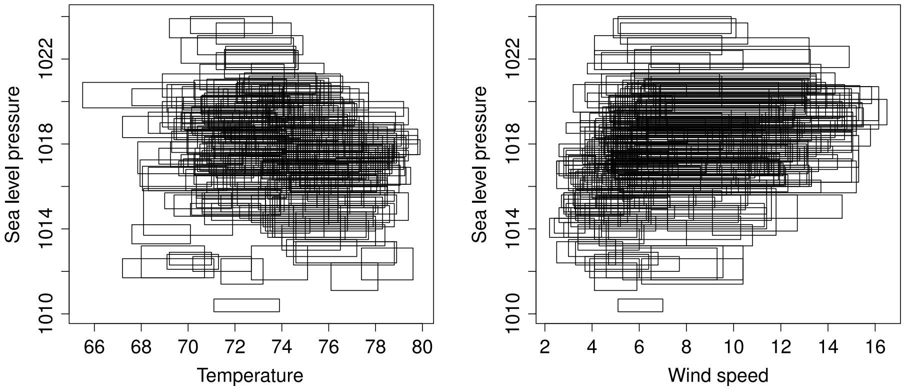
1. Introduction

Humans have been interested in weather forecasting through the ages. Reoccurring astronomical and meteorological events were used to record seasonal changes in the weather during early times. After developing instruments to measure the properties of the atmosphere, such as temperature, pressure, and humidity, efforts were made to understand the atmosphere using the measurements of the properties. Knowledge of the atmosphere has been considered a key factor for weather forecasting (Lutgens & Tarbuck, 2007). Statistical weather forecasting is a method of weather prediction using statistical models to describe relations among such meteorological variables. It was first studied during the mid-twentieth century by Wadsworth (1951) and Wadsworth, Bryan, and Gordon (1948). After that, statistical prediction of variability in weather or meteorological variables such as surface temperature and sea level pressure has been an important problem and studied by many researchers for decades. For example, see Barnett (1985), Davis (1976), Gillett, Zwiers, Weaver, and Stott (2003), Kutzbach (1967), and Min, Legutke, Hense, and Kwon (2005), among others. Stations were constructed world wide to observe weather and meteorological variables, and with development of science and technology the number of stations and the amount of data generated from them have increased exponentially.

Although computing power is highly advanced recently, sometimes it is not practical to analyze massive sized data sets. Consequently, such huge data sets are aggregated to intervals with lower and upper bounds or to histograms. Researchers sometimes encounter interval-valued data, which are either inherently observed as or processed to be intervals. Examples are blood pressure (Billard & Diday, 2000) and income level in survey data (Xu, 2010), among others. Interval-valued data belong to a broader category of data forms called symbolic data (Diday, 1987). It is difficult to analyze these types of data

[*] Tel.: +82 2 820 5547.
E-mail address: clim@cau.ac.kr.

http://dx.doi.org/10.1016/j.jkss.2015.12.003
1226-3192/



(a) Sea level pressure vs. temperature.

(b) Sea level pressure vs. wind speed.

Fig. 1. Plots of interval-valued data for Hawaiian climate data.

with classical methods. To illustrate this point, we consider a real interval-valued Hawaiian climate data set. There are three random variables: X_1 = the daily temperature in Hawaii, X_2 = the daily wind speed in Hawaii, and Y = the daily sea level pressure in Hawaii. The interval-valued data used in this illustration were converted from a total of 5408 single-valued observations which are collected in 2012 from 16 stations. The lower bound and the upper bound of the intervals are the Q1 and Q3 of the 16 stations, respectively. The sample size of the data is thus 366. The original data are publicly available from the National Climate Data Center at http://www.ncdc.noaa.gov/.

Fig. 1 shows the plots of the interval-valued data for Hawaiian climate data. We observe a decreasing pattern for the relationship between the sea level pressure and the temperature and an increasing pattern between the sea level pressure and the wind speed. However, the main difficulty is to take into account internal variation or structure within an observation, that is, an interval.

Researchers have studied adaptation of classical methods to the symbolic data extensively. For example, see Diday (1995), Diday and Emilion (1996, 1998), and Diday, Emilion, and Hillali (1996), among others. After the establishment of the adaptation, researchers have considered regression approaches to interval-valued data actively. Billard and Diday (2000) introduced a regression approach first, which fits a linear regression model on the center point of the intervals and applies the fitted model to the lower and the upper bounds of the predictor variables to obtain a prediction. Lima Neto, de Carvalho, and Tenorio (2004) extended this approach to the range of the intervals and proposed a regression method. This method fits two separate linear regression models on the center and the range of the intervals. Later, Billard and Diday (2007) employed this idea and proposed a bivariate approach which fits two regression models on both of the center and the range of the intervals simultaneously as the predictors. Recently, Lima Neto, Cordeiro, and de Carvalho (2011); Lima Neto, Cordeiro, Carvalho, Anjos, and Costa (2009) considered the bivariate generalized linear model by Iwasaki and Tsubaki (2005) to analyze interval-valued data, and Lima Neto and de Carvalho (2010) introduced a method for fitting a constrained linear regression model to interval-valued data. The proposed method fits a constrained linear regression model on the center point and range of the interval values. Xu (2010) proposed a symbolic covariance method based on the symbolic sample covariance introduced by Billard (2007, 2008). Research for analyzing interval-valued data has been a very active area and considered using various approaches. See, for example, Ahn, Peng, Park, and Jeon (2012), Blanco-Fernandez, Corral, and Gonzalez-Rodriguez (2011), Silva, Lima Neto, and Anjos (2011), and Yang, Jeng, Chuang, and Tao (2011) among others, and Blanco-Fernandez, Colubi, and Gonzalez-Rodriguez (2013) for a recent review.

Regression approaches for interval-valued data in the literature have been mostly developed based on linear regression models as described above. However, there might be cases where interval-valued data are not generated from a linear regression model, but some nonlinear regression model. One could check scatter plots of data to see if there are nonlinear patterns between the response variable and some of the explanatory variables. For such cases it may not be appropriate to use the existing regression approaches for interval-valued data.

In this paper we consider regression analysis of interval-valued data based on nonparametric additive models in order to provide a better prediction for interval-valued data with nonlinear patterns. In many practical applications, using linear regression models is too restrictive and may have a problem of misspecification. To avoid such limitations, researchers often prefer to use nonparametric regression models such as kernel regression, local polynomial, k-nearest neighbors, and so on. The reason is that nonparametric regression models make few assumptions about the regression function. However, because of the same reason they are very difficult to interpret compared to the classic linear models. Not only that, completely unstructured nonparametric regressions would not work well due to the curse of dimensionality (Friedman & Stuetzle, 1981). Nonparametric additive models (Buja, Hastie, & Tibshirani, 1989; Hastie & Tibshirani, 1990; Stone, 1985) provide a useful compromise between the restrictive linear model and the fully unstructured nonparametric model.

The additive model is a special case of the projection pursuit regression model proposed by Friedman and Stuetzle (1981). The alternating least squares model (van der Burg & de Leeuw, 1983) and the alternating conditional expectation model