Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Interval-valued time series models: Estimation based on order statistics exploring the Agriculture Marketing Service data

Wei Lin[a], Gloria González-Rivera[b,*]

[a] International School of Economics and Management, Capital University of Economics and Business, Beijing, 100070, China

[b] Department of Economics, University of California, Riverside, CA 92521, United States

ARTICLE INFO

Article history:
Received 13 September 2014
Received in revised form 14 July 2015
Accepted 15 July 2015
Available online xxxx

Keywords:
Interval-valued data
Order statistics
Intensity
Maximum likelihood estimation

ABSTRACT

The current regression models for interval-valued data ignore the extreme nature of the lower and upper bounds of intervals. A new estimation approach is proposed; it considers the bounds of the interval as realizations of the max/min order statistics coming from a sample of n_t random draws from the conditional density of an underlying stochastic process $\{Y_t\}$. This approach is important for data sets for which the relevant information is only available in interval format, e.g., low/high prices. The interest is on the characterization of the latent process as well as in the modeling of the bounds themselves. A dynamic model is estimated for the conditional mean and conditional variance of the latent process, which is assumed to be normally distributed, and for the conditional intensity of the discrete process $\{n_t\}$, which follows a negative binomial density function. Under these assumptions, together with the densities of order statistics, maximum likelihood estimates of the parameters of the model are obtained. They are needed to estimate the expected value of the bounds of the interval. This approach is implemented with the time series of livestock prices, of which only low/high prices are recorded making the price process itself a latent process. It is found that the proposed model provides an excellent fit of the intervals of low/high returns with an average coverage rate of 83%. In addition, a comparison with current models for interval-valued data is offered.

1. Introduction

Since the work on symbolic data by Billard and Diday (2003, 2006), a variety of regression models have been proposed to fit interval-valued data, see the survey article by Arroyo et al. (2010) for an extensive review. A first approach proposed by Billard and Diday was to regress the centers of the intervals of the dependent variable on the centers of the intervals of the regressors. Subsequent approaches considered two separate regressions, one for the lower bound and another for the upper bound of intervals (Brito, 2007), or one regression for the center and another for the range of the interval (Lima Neto and de Carvalho, 2010). None of these approaches guarantees that the fitted values from the regressions will satisfy the natural order of an interval, i.e., $\widehat{y}_l \leq \widehat{y}_u$, for all observations in the sample. A solution came from Lima Neto and de Carvalho (2010) who modified the previous regression models by imposing non-negative constraints on the regression coefficients of the model for the range. González-Rivera and Lin (2013) argued that these *ad hoc* constraints limit the usefulness of the model

[*] Corresponding author. Tel.: +1 951 827 1590; fax: +1 951 827 5685.
E-mail address: gloria.gonzalez@ucr.edu (G. González-Rivera).

http://dx.doi.org/10.1016/j.csda.2015.07.008
0167-9473/© 2015 Elsevier B.V. All rights reserved.

and proposed a constrained regression model that generalizes the previous regression models for lower/upper bounds or center/radius of intervals, and naturally guarantees that the proper order of the fitted intervals is satisfied.

A common thread to these approaches is that they consider the lower and upper bound as distinct stochastic processes. In this paper we propose an alternative approach and argue that there is only one stochastic process, say $\{Y_t\}$, that generates the upper and lower bounds of the interval. When we analyze interval-valued data, we only observe the bounds and these are extreme realizations of a latent random variable. This is our conceptual setup. At a fixed time t, we consider a random variable Y_t with a given conditional density function from which we draw randomly n_t realizations. The lower and upper bounds of the interval, i.e. (y_{lt} and y_{ut}) are the realized minimum and maximum values coming from the set of realizations associated with the n_t draws. As such, our interest moves towards the analysis of these two order statistics and their probability density functions. As an example, consider a time series of daily prices. In a given day t, from opening to closing time, there are n_t transactions, each one generating a market price. If we consider the daily number of trades as the n_t random draws, their corresponding intra daily prices are the realizations of the random variable daily price Y_t, and the highest/lowest prices are the realizations of the max/min order statistics of Y_t. Observe that we are not interested in the dynamics of the intra daily prices, only the lowest/highest prices carry information on the daily market activity. To start the modeling exercise, we require a set of assumptions regarding the density of the underlying stochastic process and the density of the number of draws. We will assume that the first process is continuous and it follows a conditional normal density function, and that the second process is naturally discrete and it follows a negative binomial density. Under these assumptions, we will obtain the expected values of the lower and upper bounds of the interval.

However, this modeling approach will also provide information on the latent process because we will be able to model its conditional mean and conditional variance. This is an advantage in those instances in which there are no records of opening or closing prices, typically the object of analysis, or when those prices are not very representative of the state of the market. In this paper, we model such a time series: agricultural and livestock prices provided by the US Department of Agriculture. We look into beef sales prices; the daily information provided is low price, high price, weighted average price, number of trades and total pounds traded. We could model the weighted average price but this is not very informative for potential sellers and buyers. Instead, we construct the daily interval-valued time series of low/high beef prices, which we manually dig from several archives provided by the US Department of Agriculture, and implement our approach to discover the characteristics of the latent price as well as the expected values of the low and high prices.

The paper is organized as follows. In Section 2, we discuss the key ideas of our modeling approach and its implementation under a set of assumptions. In Section 3, we use Monte Carlo simulation to investigate the properties of the proposed maximum likelihood estimator. In Section 4, we model the dynamics of the daily beef sales and prices. In Section 5, we compare our proposed model with some existing approaches on modeling interval-valued time series using both simulated and real data. Finally, in Section 6, we conclude by summarizing our findings.

2. General framework

We assume that there is an underlying stochastic process for the interval-valued time series, and in a given time t, e.g. day, month, etc. the high/low values of intervals are the realized highest and lowest order statistics based on the random draws from the conditional densities of the underlying stochastic process. Formally,

Assumption 1 (*DGP*)**.** Let $\{Y_t : t = 1, \dots, T\}$ be the underlying stochastic process. The latent random variable Y_t at time t has a conditional probability density function $f(y_t|\mathfrak{F}_t)$. At each time t, from the conditional density of Y_t we draw n_t observations. The number of draws has a discrete density function $h(n_t|\mathfrak{F}_t)$. Let y_{lt} and y_{ut} be the smallest and largest values of the random sample $\mathcal{S}_t \equiv \{y_{it} : i = 1, 2, \dots, n_t\}$:

$$y_{lt} \equiv \min_i \mathcal{S}_t = \min_{1\le i\le n_t} \{y_{it}\},$$
$$y_{ut} \equiv \max_i \mathcal{S}_t = \max_{1\le i\le n_t} \{y_{it}\}.$$

Then, $\{(y_{lt}, y_{ut}, n_t) : t = 1, \dots, T\}$ forms the observed interval time series and number of random draws, and $\mathfrak{F}_t \equiv \{(y_{ls}, y_{us}, n_s) : s = 1, \dots, t-1\}$ is the information set available at time t.

At time t, the low and high observations (y_{lt} and y_{ut}) are the lowest and highest ranked order statistics of the random sample $\mathcal{S}_t$ formed by the n_t draws or trades. The joint conditional probability density of $(y_{lt},\ y_{ut})$ given n_t and information set $\mathfrak{F}_t$ is

$$g(y_{lt}, y_{ut}|n_t, \mathfrak{F}_t) = n_t(n_t-1)\,[F(y_{ut}|\mathfrak{F}_t) - F(y_{lt}|\mathfrak{F}_t)]^{n_t-2} \times f(y_{lt}|\mathfrak{F}_t)f(y_{ut}|\mathfrak{F}_t),$$

where $F(\cdot|\mathfrak{F}_t)$ is the cumulative distribution function corresponding to the conditional density $f(\cdot|\mathfrak{F}_t)$. Then, the joint probability density of (y_{lt}, y_{ut}, n_t) conditional on information set $\mathfrak{F}_t$ is,

$$p(y_{lt}, y_{ut}, n_t|\mathfrak{F}_t) = g(y_{lt}, y_{ut}|n_t, \mathfrak{F}_t)h(n_t|\mathfrak{F}_t).$$

We still need to specify the conditional densities $f(y_t|\mathfrak{F}_t)$ and $h(n_t|\mathfrak{F}_t)$ and their dependence on the information set. Therefore, we have Assumptions 2 and 3.

Please cite this article in press as: Lin, W., González-Rivera, G., Interval-valued time series models: Estimation based on order statistics exploring the Agriculture Marketing Service data. Computational Statistics and Data Analysis (2015), http://dx.doi.org/10.1016/j.csda.2015.07.008