A Study on Linear Regression of Interval-valued Data

Spyridon Siouras[1], Stavros Adam[2]

[1] Department of Computer Science and Engineering, University of Ioannina
GR45110 Ioannina, Greece
ssiouras@cs.uoi.gr
[2] Department of Informatics and Telecommunications, University of Ioannina
GR47150 Arta, Greece
adamsp@uoi.gr

Keywords: Interval Methods, Interval-valued Data, Linear Regression

Introduction and Motivation

Intervals were proposed and used in computations to delimit truncation errors. Interval methods [9] constitute a framework for numerical computation which provide guaranteed bounds on the range of a quantity or function [14]. Considered from another point of view, intervals and interval computing are extensively used to model uncertainty pertaining data obtained from various processes considered as random experiments, regardless their distribution. In real life, processes which provide interval-valued data refer to studies and research in various fields such as Medicine, Economics, Engineering, Social Sciences etc. Examples of interval data include measurements with noise or fluctuation in some specific range, such as the temperature range measured on patients during some epidemiological investigation, the fluctuation of quotations on the stock exchange, the values of a model parameters in a bounded error context, etc. Compared to detailed data records, the significance of such interval data explains the relevance of grouping data values in the field of symbolic data analysis [3].

Typically, in all the above situations, one is given a sequence of interval-valued data that need to be statistically analyzed either for descriptive or inferential purposes in the uncertainty framework represented by intervals. As it usually happens in a real setting, where, one needs to consider for the data at hand some statistics and carry out some statistical analysis, a number of different approaches have been introduced to analyze interval-valued data. These approaches include both usual statistics, such as statistics of central tendency, as well as, statistical procedures such as linear regression, principal component analysis, clustering, etc.

We present the results of our work in progress on linear regression of interval-valued data. Following a literature review we focused on studying specific methods and applying them on a real life problem in order to assess the potential of linear regression on interval-valued data and investigate the potential of interval-valued data regression as a possible research direction.

Related work

Analyzing interval-valued data and considering statistics on these data has been the objective of a certain number of research reported in the interval data analysis related literature. Among these research efforts, one may cite the works of Bertrand and Goupil

[1], Chavent and LeChevallier[4], Chavent and Saracco [6], Billard and Diday [3], Chavent et al.[5], Palumbo and Lauro [16], Kreinovich [10], Ogasawara and Kon [15] etc.

Concerning linear regression a number of different approaches are proposed from several researchers. One should first mention the Center Method (CM) proposed by Billard and Diday [2], which is based on fitting a linear regression model on the center points of the intervals and the resulting model is then applied to the lower and upper bounds of the independent variables. This initial approach was improved by the works of Lima Neto et al. [12] and de Carvalho et al. [7] who proposed Center and Range Methods (CRM) utilizing the ranges of intervals in building two separate regression models. In addition to these efforts Billard and Diday [3] proposed the Bivariate Center and Range Method (BCRM), which included both the center points and the ranges of intervals. However, these methods proved to be defective in several cases as they resulted in predicting lower bounds which exceed the predicted upper bounds. In order to cope with this problem Lima Neto and Carvalho [13] proposed the Constrained Method which results in a regression model restricting all model parameters to positive numbers. To satisfy this restriction the least squares method is applied in conjunction with an algorithm introduced by Lawson and Hanson [11]. This algorithm, which is proven to converge, recognizes parameter values that do not respect the constraints and converts these values to non-negatives using a specific recalculation procedure. While this method has the advantage that it can be applied in conjunction with CM and CRM methods, its inconvenience is that, often, the constraint imposed on the values of the parameters (to be only non negative real numbers) provides models which do not represent the relation between dependent and independent variables.

Finally, we need to mention the method Lasso-IR proposed by Paolo Giordani [8] (based on the Lasso regression technique) which constructs two regression models, one for the centers and another one for the radii of the intervals. The main characteristic of this method is that the parameters of the centers regression model are computed to be as close as possible to the parameters of the radii model. This is achieved by setting a threshold representing the maximum level of distribution for the values of the parameters of the two regression models.

Experiments and results

For this study after experimenting with linear regression on the centers and on the bounds of the intervals separately, we applied the Lasso-IR technique on a dataset available by the Spanish Electric Company concerning the amount of energy demanded by the consumption on the electric power distribution network. The dataset adopted provides for each day of a month the minimum and the maximum amount of electric energy really demanded. Min and max values of energy are also provided for the predicted load as well as the scheduled production. For our experiments we retained the load values corresponding to the real demand which is analyzed against the predicted load. The initial data were converted to intervals providing summary information for each day of November 2021 and the Lasso-IR technique was applied to provide a linear regression model for the real demand of electric energy depending on the predicted one. Figure 1, hereafter, provides an illustration of the results obtained on the above mentioned dataset.

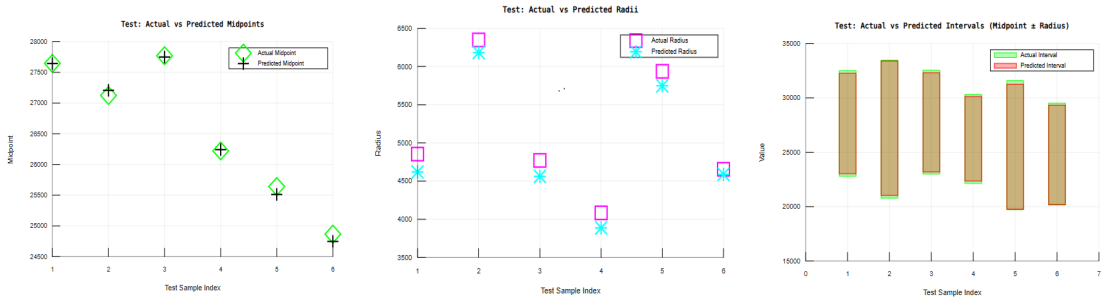

Figure 1: Regression analysis results on a test set consisting of six samples

References

[1] P. Bertrand and F. Goupil. Descriptive Statistics for Symbolic Data. In Hans-Hermann Bock and Edwin Diday, editors, *Analysis of Symbolic Data*, pages 106–124, Berlin, Heidelberg, 2000. Springer.

[2] L. Billard and E. Diday. Regression analysis for interval-valued data. In Henk A. L. Kiers, Jean-Paul Rasson, Patrick J. F. Groenen, and Martin Schader, editors, *Data Analysis, Classification, and Related Methods*, pages 369–374, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.

[3] L. Billard and E. Diday. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley Series in Computational Statistics. Wiley, 2012.

[4] M. Chavent and Y. Lechevallier. Dynamical Clustering of Interval Data: Optimization of an Adequacy Criterion Based on Hausdorff Distance. In Krzysztof Jajuga, Andrzej Sokołowski, and Hans-Hermann Bock, editors, *Classification, Clustering, and Data Analysis*, pages 53–60, Berlin, Heidelberg, 2002. Springer.

[5] M. Chavent, Y. Lechevallier, F. de A. T. de Carvalho, and R. Verde. New clustering methods for interval data. *Computational Statistics*, 21(2), 2006.

[6] M. Chavent and J Saracco. On central tendency and dispersion measures for intervals and hypercubes. *Communications in Statistics - Theory and Methods*, 37(9):1471–1482, 2008.

[7] F.A.T. de Carvalho, E.A. Lima Neto, and C.P. Tenorio. A New Method to Fit a Linear Regression Model for Interval-Valued Data. In Susanne Biundo, Thom Frühwirth, and Günther Palm, editors, *KI 2004: Advances in Artificial Intelligence*, pages 295–306. Springer Berlin Heidelberg, 2004.

[8] P. Giordani. Lasso-constrained regression analysis for interval-valued data. *Advances in Data Analysis and Classification*, 9(1):1862–5355, 2015.

[9] R.B. Kearfott. Interval computations: Introduction, uses, and resources. *Euromath Bulletin*, 2(1):95–112, 1996.

[10] V. Kreinovich. *Statistical Data Processing under Interval Uncertainty: Algorithms and Computational Complexity*, pages 11–26. Springer, Berlin, Heidelberg, 2006.

[11] C. Lawson and R. Hanson. *Solving Least Squares Problems*. Prentice-Hall, 1974.

[12] E.A. Lima Neto, F.A.T. de Carvalho, and C P. Tenorio. Univariate and multivariate linear regression methods to predict interval-valued features. In Geoffrey I. Webb and Xinghuo Yu, editors, *AI 2004: Advances in Artificial Intelligence*, pages 526–537. Springer Berlin Heidelberg, 2005.

[13] E.A. Lima Neto and de Carvalho F.A.T. Constrained Linear Regression Models for Symbolic Interval-valued Variables. *Computational Statistics & Data Analysis*, 54(2):333–347, 2010.

[14] R.E. Moore, R.B. Kearfott, and M.J. Cloud. *Introduction to Interval Analysis*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009.

[15] Y. Ogasawara and M. Kon. Two clustering methods based on the ward's method and dendrograms with interval-valued dissimilarities for interval-valued data. *International Journal of Approximate Reasoning*, 129:103–121, 2021.

[16] F. Palumbo and C.N. Lauro. A PCA for interval-valued data based on midpoints and radii. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, and J. J. Meulman, editors, *New Developments in Psychometrics*, pages 641–648, Tokyo, 2003. Springer Japan.