iRegression: a regression library to symbolic interval-valued variables

Lima Neto Eufrásio[1*], Souza Filho Cláudio[1], Anjos Ulisses[1]

1. Departamento de Estatística, Universidade Federal da Paraíba, Cidade Universitária s/n, João Pessoa, PB, Brazil
*Contact author: eufrasio@de.ufpb.br

Keywords: Regression, interval variable, symbolic data analysis.

Symbolic data analysis (SDA) has been introduced as a domain related to multivariate analysis, pattern recognition and artificial intelligence in order to introduce new methods and to extend classical data analysis techniques and statistical methods to symbolic data (Billard and Diday 2006). In SDA, a variable can assume as a value an interval from, a set of real numbers, a set of categories, an ordered list of categories or even a histogram. These new variables take into account the variability and/or uncertainty presented in the data. Interval variables have been studied in the area of SDA, where very often an object represents a group of individuals and the variables used to describe it need to assume a value which express the variability inherent to the description of a group. Moreover, interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Nowadays, different approaches have been introduced to analyze interval-valued variables. In the field of SDA, approaches to fit a regression model to interval-valued data have been discussed in the literature. However, the access to such methods still is restricted, being necessary to request to the authors. Billard and Diday (2000) were first to propose an approach to fit a linear regression model to symbolic interval-valued data sets. Lima Neto and De Carvalho (2008) improved the previous approach presenting a new method based on two linear regression models, the first regression model over the midpoints of the intervals and the second one over the ranges, which reconstruct the boundaries of the interval-values of the dependent variable in a more efficient way when compared with the Billard and Diday's method. The aim of this work is to development a R library, called iRegression, that includes some regression methods for interval-valued variables. This new library will be the first one developed to treat symbolic data in the regression context. Thus, some regression methods for symbolic interval-valued variables will be accessible to students, teachers and professionals.

Acknowledgments : The authors would like to thank CNPq (Brazilian Agency) for their financial support.

References

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, John Wiley, New York.

Lima Neto, E.A. and De Carvalho, F.A.T.. (2008), Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis 52*, pp. 1500-1515