Beyond the non-probabilistic symbolic regression models for interval variables

Eufrásio de A. Lima Neto[1], Gauss M. Cordeiro[2], Francisco de A.T. de Carvalho[3]

[1] Departamento de Estatística, Universidade Federal da Paraíba - Cidade Universitária s/n - CEP 58051-900 - João Pessoa (PB) - Brazil
[2] Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco - Dois Irmãos - CEP 52171-900 - Recife (PE) - Brazil
[3] Centro de Informática, Universidade Federal de Pernambuco, Av. Prof. Luiz Freire - s/n - Cidade Universitária - CEP 50740-540 - Recife (PE) - Brazil

Abstract This paper presents a overview about the symbolic regression models to interval-valued data. The major symbolic regression methods proposed in literature visualized the problem like a optimization point of view. Lima Neto et. al. (2009) proposed a new symbolic regression model for interval variables, called bivariate generalized linear model (BGLM), which are based on bivariate exponential family of distributions [5], making possible the use of statistical inference techniques and goodness-of-fit measures over symbolic regression models.

Keywords: Symbolic Regression Models, Bivariate Generalized Linear Models, Interval-valued Data

1 Introduction

In regression analysis of quantitative data, the items are usually represented as a vector of quantitative measurements [7]. However, due to recent advances in information technologies, it is now common to record interval-valued data. In the Symbolic Data Analysis (SDA) framework [1, 3, 4], interval-valued data appear when the observed values of the variables are intervals from the set of real numbers $\Re$. Moreover, interval-valued data arise in practical situations such as recording monthly interval temperatures in meteorological stations, daily interval stock prices, among others. Another source of interval data is the aggregation of huge data-bases into a reduced number of groups, the properties of which are described by symbolic interval variables.

Billard and Diday [2] presented the first approach to fit a linear regression model to a symbolic interval-valued data set. Their approach consists of fitting a linear regression model to the midpoint of the interval values assumed by the variables in the learning set and then applies this model to the lower and upper limits of the interval values of the explanatory variables to predict, respectively, the lower and upper limits of the interval values of the dependent variable. Lima Neto and De Carvalho [8] improved the former 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 bounds of the interval-values of the dependent variable in a more efficient way.

Despite recent contributions to symbolic regression models, current approaches view the problem from an optimization point of view and do not consider the probabilistic

aspects related to regression models. This make it impossible to use inference techniques over the parameters estimates, such as hypothesis tests or confidence intervals.

Generalized linear models represent a major synthesis of regression models by allowing a wide range of types of response data and explanatory variables to be handled in a single unifying framework. These models are based on the exponential family of distributions and represent a very important regression tool due to their flexibility and applicability in practical situations [6]. Iwasaki and Tsubaki [5] introduced a class of bivariate generalized linear models (BGLMs) based on the bivariate exponential family of distributions with an application to meteorological data analysis.

Lima Neto et. al. (2009) considered the BGLM as an important tool for solving problems related to SDA and presented a model based on bivariate Gaussian distribution. They also presented an alternative way to estimate the dispersion parameter ϕ and the coefficient of correlation ρ. The latter is based on the log-likelihood profile method. Additionally, the goodness-of-fit measures, which are not addressed by Iwasaki and Tsubaki, were considered by them. Application to a real interval data sets demonstrated that the BGLM method presented a better fit when compared with the non-probabilistic symbolic regression methods proposed by [2] and [8]. However, the authors recommend a simulated study in future works for a more consistent conclusion about the BGLM method.

References

[1] Book, H.H., Diday, E.: Analysis of Symbolic Data, Exploratory methods for extracting statistical information from complex data. Springer-Verlag (2000).

[2] Billard, L., Diday, E.: Regression Analysis for Interval-Valued Data. Proceedings of the Seventh Conference of the International Federation of Classification Societies. Springer-Verlag (2000) 369-374.

[3] Billard, L., Diday, E.: Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley (2006).

[4] Diday, E., Fraiture-Noirhomme, M.: Symbolic Data Analysis and the SODAS Software. Wiley-Interscience (2008).

[5] Iwasaki, M., Tsubaki, H.: A bivariate generalized linear model with an application to meteorological data analysis. Statistical Methodology **2** (2005) 175-179.

[6] McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman & Hall (1989).

[7] Montgomery, D.C., Peck, E.A.: Introduction to Linear Regression Analysis. John Wiley (1982).

[8] Lima Neto, E.A., De Carvalho, F.A.T.: Centre and range method to fitting a linear regression model on symbolic interval data. Computational Statistics and Data Analysis **52** 1500–1515.

[9] Lima Neto, E.A., De Carvalho, F.A.T., Cordeiro, G.M, Anjos, U.U., Costa, A.G.: Bivariate Generalized Linear Model for Interval-Valued Variables. Proceedings of the 2009 IEEE International Joint Conferences on Neural Networks. IEEE (2009) accepted for publication.