Quality Measures for Clusterwise Regression

Paula Brito, Sonia Dias, Ndèye Niang

Quality Measures for Clusterwise Regression

Paula Brito, Sónia Dias, and Ndèye Niang

Abstract We focus on interval-valued variables, whose observations are intervals of real numbers. The Interval Distributional (ID) regression model [1] considers intervals represented by the corresponding quantile functions. The error between predicted and observed intervals, for each unit, is evaluated by the Mallows Distance. However, sometimes a single regression model is not appropriate, and it may be necessary to cluster the units and fit a regression model in each cluster. We apply a Clusterwise Regression model, for interval-valued variables, that finds the best partition of the data in clusters and simultaneously provides a linear regression model for each cluster. The algorithm [4], combines the dynamical clustering algorithm [2], and the ID regression model. The process is applied repeatedly varying the number of clusters K; for each fixed K, the algorithm considers different initial partitions and selects the solution with lowest Total Error. To select the best solution across different K, quality measures are proposed, that evaluate the fit between the clusters and their representing regression models. In particular, we extend the well-known Silhouette coefficient to clusterwise regression. The proposed model and measures are applied to a problem of pollution prediction in West Africa.

Keywords: clusterwise regression, interval data, Silhouette coefficient

Acknowledgements This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project UIDB/50014/2020.

References

1. Dias, S. and Brito, P.: Off the beaten track: a new linear model for interval data. European Journal of Operational Research, **258**(3), 1118–1130 (2017)
2. Diday, E. and Simon, J.C.: Clustering analysis. In Digital Pattern Recognition, pp. 47–94. Springer (1976)
3. Späth., H.: A fast algorithm for clusterwise linear regression. Computing, **29**(2), 175–181 (1982)
4. Suresh, N.: Clusterwise Linear Regression for Interval Data - An Extension of Interval Distributional Model. Master's thesis, Faculdade de Economia, Universidade do Porto (2020).

Paula Brito
Fac. Economia, Univ. Porto & LIAAD-INESC TEC, Portugal, e-mail: mpbrito@fep.up.pt

Sónia Dias
ESTG, Instituto Politécnico de Viana do Castelo & LIAAD-INESC TEC, Portugal, e-mail: sdias@estg.ipvc.pt

Ndèye Niang
Cédric-CNAM, Paris, France, e-mail: ndeye.niang keita@cnam.fr