DSS Statistics Seminar

December 3, 2021, 12:00

https://uniroma1.zoom.us/j/86881977368?pwd=SWRFcVFjMDZTa0lXZk05TE1zNm5adz09

Passcode: 432940

Marco Riani

Università degli studi di Parma

Data rarely follow the simple models of mathematical statistics. Often, there will be distinct subsets of observations so that more than one model may be appropriate. Further, parameters may gradually change over time. In addition, there are often dispersed or grouped outliers which, in the context of international trade data, may correspond to fraudulent behavior. All these issues are present in the datasets that are analyzed on a daily basis by the Joint Research Centre of the European Commission and can only tackled by using methods which are robust to deviations to model assumptions (see for example [6]). This distance between mathematical theory and data reality has led, over the last sixty years, to the development of a large body of work on robust statistics. In the seventies of last century, it was expected that in the near future any author of an applied article who did not use the robust alterative would be asked by the referee for an explanation [9]. Now, a further forty years on, there does not seem to have been the foreseen breakthrough into the wider scientific universe. In this talk, we initially sketch what we see as some of the reasons for this failure, suggest a system of interrogating robust analyses, which we call monitoring [5] and describe a series of robust and efficient methods to detect model deviations, groups of homogeneous observations [10], multiple outliers and/or sudden level shifts in time series ([8]). Particular attention will be given to robust and efficient methods (known as forward search) which enables to use a flexible level of trimming and understand the effect that each unit (outlier or not) exerts on the model (see for example [1], [2] [7]). Finally, we discuss the extension of the above methods to transformations and to the big data context. The Box-Cox power transformation family for non-negative responses in linear models has a long and interesting history in both statistical practice and theory. The Yeo-Johnson transformation extends the family to observations that can be positive or negative. In this talk, we describe an extended Yeo-Johnson transformation that allows positive and negative responses to have different power transformations ([4] or [3]). As an illustration of the suggested procedure, we analyse data on the performance of investment funds, 99 out of 309 of which report a loss. The problem is to use regression to predict medium term performance from two short term

indicators. It is clear from scatterplots of the data that the negative responses have a lower variance than the positive ones and a different relationship with the explanatory variables. Tests and graphical methods from our robust analysis allow the detection of outliers, the testing of the values of transformation parameters and the building of a simple regression model. All the methods described in the talk have been included in the FSDA Matlab toolbox freely donwloadable as a toolbox from Mathworks file exchange or from github at the web address https://uniprjrc.github.io/FSDA/

References

[1] Atkinson, A. C. and Riani, M. (2000). *Robust Diagnostic Regression Analysis*. Springer–Verlag, New York.

[2] Atkinson, A. C., Riani, M., and Cerioli, A. (2004). *Exploring Multivariate Data with the Forward Search*. Springer–Verlag, New York.

[3] Atkinson, A. C., Riani, M., and Corbellini, A. (2020). The analysis of transformations for profit-and-loss data. Journal of the *Royal Statistical Society: Series C (Applied Statistics)*, 69(2), 251–275.

[4] Atkinson, A. C., Riani, M., and Corbellini, A. (2021). The BoxCox Transformation: Review and Extensions. *Statistical Science*, 36(2), 239 – 255.

[5] Cerioli, A., Riani, M., Atkinson, A. C., and Corbellini, A. (2018). The power of monitoring: How to make the most of a contaminated multivariate sample (with discussion). *Statistical Methods and Applications*, 27, 559–666. https://doi.org/10.1007/s10260–017–0409–8.

[6] Perrotta, D., Torti, F., Cerasa, A., and Riani, M. (2020). The robust estimation of monthly prices of goods traded by the European Union. *Technical Report EUR 30188 EN, JRC120407, European Commission, Joint Research Centre, Publications Office of the European Union, Luxembourg*. ISBN 978-92-76-18351-8, doi:10.2760/635844.

[7] Riani, M., Atkinson, A. C., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, Series B*, 71, 447–466.

[8] Rousseeuw, P., Perrotta, D., Riani, M., and Hubert, M. (2019). Robust monitoring of time series with application to fraud detection. *Econometrics and Sttaistics*, 9, 108–121.

[9] Stigler, S. M. (2010). The changing history of robustness. *The American Statistician*, 64, 277–281.

[10] Torti, F., Perrotta, D., Riani, M., and Cerioli, A. (2018). Assessing trimming methodologies for clustering linear regression data. *Advances in Data Analysis and Classification*, 13, 227–257.