

Multivariate Analysis for Data Scientists

Kimmo Vehkalahti¹

¹*University of Helsinki, Finland*

Abstract

The ongoing “Data revolution” [1] sets more requirements for the researchers on all fields of science. One could say – without exaggerating too much – that *we should all be data scientists*.

A special pressure is put on the fields of social and behavioral sciences, where the phenomena, the measurements, and the data are affected by endless sources of uncertainty and may often be much more complex than in many applications of, say, natural sciences.

Hence, the question is: *What should be included in a data scientist’s “toolbox” in social and behavioral sciences?*

Our suggestion would be a good combination of classical and modern skills that are covered, for example, by a recent textbook on multivariate analysis [3]: A wide range of methods for visualizing data, linear and generalized linear mixed (and fixed) models, various methods of multivariate analysis (both exploratory and confirmatory), a bit (or a byte) of matrices behind the methods (even without a maths background), programming and using statistical software (preferably R [2]), algorithmic thinking in general, as well as documenting and sharing the code and data on open platforms such as GitHub. In this talk, we discuss some of these topics in more detail.

Keywords

Multivariate analysis, Linear models, Data science, Matrices, Statistics.

References

- [1] Kitchin, R. (2014). *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. London: SAGE.
- [2] R Core Team (2018). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>
- [3] Vehkalahti, K., and Everitt, B. S. (2019). *Multivariate Analysis for the Behavioral Sciences*, 2nd edition. Boca Raton, Florida: Chapman and Hall/CRC. <https://github.com/KimmoVehkalahti/MABS/>