Regression Analysis on Compositional Data

Eva Fišerová

Palacký University in Olomouc, Czech republic

April 8, 2013

Abstract

The analysis of the relationship between variables belongs to fundamental problems of statistical inference. If the variables carry absolute information, the existing methods behave properly and the obtained results can be directly interpreted. Another situation occurs when the data are not absolute, but rather relative. As an example, let us imagine that the goal is to analyze the structure of causes of death in the European population, divided into five main causes (lung cancer, colorectal cancer, circulatory disease, heart disease and respiratory disease) in European countries. Taking the raw data would be fully uninformative because different states have different total number of inhabitants. Thus, it seems intuitive to express the data in proportions in order to see the relative contributions of groups of causes of death to the overall European population. More generally, percentages or proportions can be expressed as representations of data carrying quantitative descriptions of parts of a whole, conveying exclusively relative information, so called compositional data (Aitchison, 1986; Pawlowsky-Glahn et al., 2007). They are characterized by the simplex sample space with the Aitchison geometry that forms the Euclidean structure of the sample space. Using proper log-ratio transformations, the data could be moved isometrically to the real Euclidean space. Once the compositional data are expressed in coordinates, it is possible to use any statistical method like outlier detection, principal component analysis, imputation of missing values, etc. In order to achieve unicity of the results on the simplex, it is necessary that used statistical procedure is invariant to rotation of the coordinates.

In the lecture, we will show the basic methodology for compositional data processing. We will focus on regression between parts of compositional data. Orthogonal regression is a popular regression tool, especially when errors naturally occur in variables. This modeling technique is invariant to rotation of coordinates and thus it is convenient for regression analysis between parts of compositional data, performed after isometric log-ratio transformation. We will present an iterative algorithm for estimation and some statistical inference, together with the corresponding interpretation for compositional data (Fišerová and Hron, 2010, 2012).

Keywords

Compositional data, log-ratio transformation, orthogonal regression.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman and Hall.

Pawlowsky-Glahn, V., Egozcue, J.J., and Tolosana-Delgado, J. (2007). *Lecture Notes on Compositional Data Analysis.*

Fišerová, E., and Hron, K. (2010). Total least squares solution for compositional data using linear models. *Journal of Applied Statistics 37*, 1137–1152.

Fišerová, E., and Hron, K. (2012). Statistical Inference in Orthogonal Regression for Three-Part Compositional Data Using a Linear Model with Type-II Constraints. *Communications in Statistics - Theory and Methods 41*, 2367–2385.