Chapter 4

Linear regression models (LRM)

Contents

Objectives

- ✓ To estimate and interpret an LRM when the response is compositional.
- ✓ To estimate and interpret an LRM when the predictor is compositional.
- ✓ To introduce some extensions for an LRM

4.1. LRM for a compositional response and scalar predictor

In this section we are dealing with the Linear Regression Models (LRM) where the compositional variables are the response variables of the model [TB11].

Let $\mathbf{X}$ be a data set in $\mathcal{S}^D$ formed by n observations $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$. The i-th observation $\mathbf{x}_i$ is associated with r external variables or covariates ($r \geq 1$) grouped in the real vector $\mathbf{t}_i = [t_{i0}, t_{i1}, \ldots, t_{ir}]$, where $t_{i0} = 1$, for $i = 1, 2, \ldots, n$.

The goal is to estimate the coefficients $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_r$ of a linear surface into $\mathcal{S}^D$ whose equation is

$$\hat{\mathbf{x}}(\mathbf{t}) = \boldsymbol{\beta}_0 \oplus (t_1 \odot \boldsymbol{\beta}_1) \oplus \cdots \oplus (t_r \odot \boldsymbol{\beta}_r) = \bigoplus_{j=0}^{r} (t_j \odot \boldsymbol{\beta}_j) \ ,$$

where $\mathbf{t} = [t_0, t_1, \ldots, t_r]$ are real covariates and are identified as the parameters of the linear surface; the first parameter is defined as the constant $t_0 = 1$; and $\hat{\mathbf{x}}(\cdot)$ are the expected value of the CoDa-response variable. The compositional coefficients of the model, $\boldsymbol{\beta}_j \in \mathcal{S}^D$, are to be estimated from the data. The most popular fitting method is the least-square deviation criterion which minimizes the sum of squared errors. Because this model is presented as a least-squares problem in the simplex, it could be formulated in terms of orthonormal log-ratio coordinates (olr). In other words,

(1) we select a *olr*-basis in $\mathcal{S}^D$, for example, according to an SBP.

(2) we represent the responses in coordinates: $\mathbf{x}_i^* = \mathrm{olr}\,(\mathbf{x}_i) \in \mathbb{R}^{D-1}$.

(3) we solve $D-1$ ordinary-least-squares regression problems in coordinates to obtain the *olr*-coordinates $\boldsymbol{\beta}_j^*$ vectors of the $\boldsymbol{\beta}_j$ coefficients ($j = 1, 2, \ldots, r$). That is, for the coordinates $k = 1, 2, \ldots, D-1$, find $\boldsymbol{\beta}_j^*$ minimizing the usual sum of squared errors:

$$\mathrm{SSE}_k = \sum_{i=1}^{n} |\hat{x}_k^*(\mathbf{t}_i) - x_{ik}^*|^2 \ , \ k = 1, 2, \ldots, D-1 \ ,$$

where

$$\hat{x}_k^*(\mathbf{t}) = \beta_{0k}^* + \beta_{1k}^*\ t_1 + \cdots + \beta_{rk}^*\ t_r, \text{and}$$

(4) back-transform the coefficients $\boldsymbol{\beta}_j^*$ to $\boldsymbol{\beta}_j \in \mathcal{S}^D$ using $\boldsymbol{\beta}_j = \mathrm{olr}^{-1}(\boldsymbol{\beta}_j^*)$.

Interpretation can thus alternatively be made in coordinates or in the simplex. Coefficients β_{jk}^*, j= 1,..., r and k=1,..., $D-1$, can be interpreted as the effect of an increase in t_j by one unit (keeping the other t_j constant) on the *olr*-coordinate x_k^*. Thus, the values of a coefficient β_{jk}^* and its interpretation depend on the chosen *olr*-basis. The coefficient $\boldsymbol{\beta}_j$ is the perturbation vector which is applied to