Chapter 4

Linear regression models (LRM)

Contents

4.1 LRM for a compositional response and scalar predictor
4.2 LRM for a scalar response and compositional predictor
4.3 LRM extensions
   4.3.1 Extensions of an LRM with a compositional predictor
   4.3.2 Compositions as both predictor and response
4. Linear regression models (LRM)

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

The goal is to estimate the coefficients \( \beta_0, \beta_1, \ldots, \beta_r \) of a linear surface into \( S^D \) whose equation is

\[
\hat{x}(t) = \beta_0 \oplus (t_1 \odot \beta_1) \oplus \cdots \oplus (t_r \odot \beta_r) = \bigoplus_{j=0}^{r} (t_j \odot \beta_j),
\]

where \( t = [t_0, t_1, \ldots, t_r] \) 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 \( \hat{x}() \) are the expected value of the CoDa-response variable. The compositional coefficients of the model, \( \beta_j \in S^D \), 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 \( S^D \), for example, according to an SBP.
2. we represent the responses in coordinates: \( x^*_i = \text{olr}(x_i) \in \mathbb{R}^{D-1} \).
3. we solve \( D-1 \) ordinary-least-squares regression problems in coordinates to obtain the olr-coordinates \( \beta^*_j \) vectors of the \( \beta_j \) coefficients \( (j = 1, 2, \ldots, r) \). That is, for the coordinates \( k = 1, 2, \ldots, D-1 \), find \( \beta^*_j \) minimizing the usual sum of squared errors:

\[
\text{SSE}_k = \sum_{i=1}^{n} |\hat{x}^*_k(t_i) - x^*_{ik}|^2 , \quad k = 1, 2, \ldots, D-1 ,
\]

where \( \hat{x}^*_k(t) = \beta^*_0 t_0 + \beta^*_1 t_1 + \cdots + \beta^*_r t_r \), and
4. back-transform the coefficients \( \beta^*_j \) to \( \beta_j \in S^D \) using \( \beta_j = \text{olr}^{-1}(\beta^*_j) \).

Interpretation can thus alternatively be made in coordinates or in the simplex. Coefficients \( \beta^*_j k, j = 1, \ldots, r \) and \( k = 1, \ldots, D-1 \), 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 \( \beta^*_j k \) and its interpretation depend on the chosen olr-basis. The coefficient \( \beta_j \) is the perturbation vector which is applied to