Regression with Compositional Response

Eva Fišerová

Palacký University Olomouc
Czech Republic

LinStat2014, August 24 - 28, 2014, Linköping

joint work with Karel Hron and Sandra Donevska

Objectives of the Contribution

- **Explain what are the compositional data**
- **Show the methodology for regression with compositional response**
- **The application of the methodology on modeling real-world geochemical data**

What is it Compositional Data (Compositions)?

- **Multivariate data where the variables represent parts of some whole carrying only relative information**

- **Examples**: Religious or national composition of the population, representation of political parties according to the election results, household expenses on final consummation (food, accommodation, clothes)

- **Usual units of measurement**: proportions, percentages, mg/kg

- **Sample space** of a *D*-part composition: **simplex**

$$\mathcal{S}^D = \left\{ \mathbf{y} = (y_1, \ldots, y_D),\ y_i > 0,\ \sum_{i=1}^{D} y_i = \kappa \right\}$$

  only ratios of the parts are informative $\Rightarrow$ κ arbitrary proper representation of compositions

- **The data are in contrary to assumptions for almost statistical methods** (**not follow the Euclidean geometry in real space**)

Log-ratio Transformations

Aitchison, 1986

Performing standard statistical methods on the simplex is impossible:

→ find new methods

→ find proper representations of the compositions in a real space

- **Additive** log-ratio (alr) transformation
- **Centred** log-ratio (clr) transformation
- **Isometric Log-ratio (Ilr) Transformation** (Egozcue, et al., 2003)
  - **standard multivariate statistical methods can be applied**
  - preserve distance
  - associated with an orthogonal coordinate system with respect to a basis on simplex
  - results are interpreted either in coordinates, or on simplex
  - interpretation of ilr coordinates directly in the sense of original parts is impossible
  - canonical orthonormal basis in the simplex does not exist - chosen according to particular situations

Special Choice of Ilr Coordinates - SBP

Egozcue and Pawlowsky-Glahn, 2005

Sequential Binary Partition (SBP)

- often used because they enable interpretation in terms of grouped parts of the composition

Example of construction: Structure of Causes of Death

y_1 lung cancer (LC)	y_2 colorectal cancer (CC)
y_3 circulatory disease (CD)	y_4 heart disease (HD)
y_5 respiratory disease (RD)	

	RD	HD	CD	CC	LC
z_1	+	+	+	-	-
z_2	+	-	-	0	0
z_3	0	+	-	0	0
z_4	0	0	0	+	-

$$z_i = \sqrt{\frac{rs}{r+s}} \ln \frac{\sqrt[r]{\prod_{i=1,i\in+}^{r} y_i}}{\sqrt[s]{\prod_{j=1,j\in-}^{s} y_j}}, \qquad r \text{ number of } +, \; s \text{ number of } -$$

Other Special Choice of Ilr Coordinates

Fišerová and Hron, 2011

$$z_k = \sqrt{\frac{D-k}{D-k+1}} \ln \frac{y_k}{\sqrt[D-k]{\prod_{j=k+1}^{D} y_j}}, \; k = 1, \ldots, D-1.$$

The 1st ilr coordinate z_1 include all relative information about the compositional part y_1

The permutation of the parts $y_2, \ldots, y_n$ doesn't change the interpretation of $z_1 \Rightarrow$ We construct D different ilr transformations

$$(y_l, y_1, \ldots, y_{l-1}, y_{l+1}, \ldots, y_D) =: (y_1^{(l)}, y_2^{(l)}, \ldots, y_{l-1}^{(l)}, y_l^{(l)}, y_{l+1}^{(l)}, \ldots, y_D^{(l)})$$

Formally

$$z_1^{(l)} = \sqrt{\frac{D-1}{D}} \ln \frac{y_1^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} y_j^{(l)}}}, \; l = 1, 2, \ldots, D$$

$\mathbf{z}^{(l)} = \mathbf{z}\mathbf{V}\mathbf{P}^{(l)}\mathbf{V}'$ $\quad$ $\mathbf{P}^{(l)}$ permutation matrix, $\mathbf{V}$ orthonormal basis on simplex

Regression with Compositional Response

Simplex sample space D-part compositions $(y_1, y_2, \ldots, y_D)$

$\Downarrow$ isometric log-ratio transformation

Euclidean real space $\mathbb{R}^{D-1}$ $(z_1, z_2, \ldots, z_{D-1})$

Multivariate regression model

- Respect the association between outcomes, and thus, in general, procedures are more efficient.
- Can evaluate the joint influence of predictors on all outcomes.
- Avoid the issue of multiple testing.

Egozcue et al. (2012) proposed univariate approach with series of $(D-1)$ submodels

Test of the Decomposition of Multivariate Model into $D-1$ Univariate Submodels

Normality assumption

Bartlett's test of sphericity

Hypothesis

$$H_0 : \rho_{Z_i,Z_j} = 0,\ i,j = 1,2,\ldots,D-1,\ i \neq j$$

Test statistic

$$V = -\left(n - \frac{2\,(D-1)+1}{6}\right) \ln\left[\det\left(\mathbf{R}_{\underline{\mathbf{Z}}}\right)\right] \overset{H_0}{\sim}_{as} \chi^2_{\frac{(D-1)(D-2)}{2}}$$

$\mathbf{R}_{\underline{\mathbf{Z}}}$ is the sample correlation matrix of $\underline{\mathbf{Z}}$

- $n \geq 2(D-1)-2$
- suitable for $n > 30$

Independence allows to analyse each submodel via univariate approach

Multivariate Linear Model

$$(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{D-1}) = \boldsymbol{X}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{D-1}) + (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{D-1})$$

Equivalently, in the matrix form

$$\underline{\boldsymbol{Z}}_{(n\times(D-1))} = \boldsymbol{X}_{(n\times k)}\boldsymbol{B}_{(k\times(D-1))} + \underline{\varepsilon}_{(n\times(D-1))}$$

Assumption: multivariate responses $\underline{\boldsymbol{Z}}_{i.}$ are independent with the same variance-covariance matrix

$$\mathrm{cov}(\underline{\boldsymbol{Z}}_{\mathrm{i.}}, \underline{\boldsymbol{Z}}_{\mathrm{j.}}) = \mathbf{0}_{((\mathrm{D}-1)\times(\mathrm{D}-1))}, \ \mathrm{i} \neq \mathrm{j}$$
$$\mathrm{var}(\underline{\boldsymbol{Z}}_{\mathrm{i.}}) = \boldsymbol{\Sigma}_{((\mathrm{D}-1)\times(\mathrm{D}-1))}, \ \mathrm{i} = 1, \ldots \mathrm{n}$$

Estimation in Multivariate Model

The BLUE of the parameter matrix $\boldsymbol{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_{D-1})$

$$\widehat{\boldsymbol{B}} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'(\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_{D-1})$$ invariant to $\boldsymbol{\Sigma}$

$$\widehat{\boldsymbol{\beta}}_i = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Z}_i,\ i = 1, 2, \ldots, D-1$$ univariate approach

The variance-covariance matrix of the vector $\mathrm{vec}(\widehat{\boldsymbol{B}}) = (\widehat{\boldsymbol{\beta}}_1', \widehat{\boldsymbol{\beta}}_2', \ldots, \widehat{\boldsymbol{\beta}}_{D-1}')'$

$$\mathrm{var}\left[\mathrm{vec}(\widehat{\boldsymbol{B}})\right] = \boldsymbol{\Sigma} \otimes \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$ $\otimes$ Kronecker product

$$\mathrm{var}(\widehat{\boldsymbol{\beta}}_i) = \sigma_{ii}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$$ univariate approach

If $\boldsymbol{\Sigma}$ is known, the estimators of β_i using multivariate and univariate approaches are equivalent.

Estimation of Covariance Matrix

Unbiased estimator of $\boldsymbol{\Sigma}$

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n-k}\underline{\boldsymbol{Z}}'\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\underline{\boldsymbol{Z}}$$

Under normality, $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Sigma}}$ are statistically independent. If moreover $n-k > (D-1)$, then

$$(n-k)\widehat{\boldsymbol{\Sigma}} \sim W_{D-1}[n-k, \boldsymbol{\Sigma}]$$ Wishart distribution

$$\begin{aligned}\widehat{\sigma}_{ii} &= \frac{1}{n-k}\boldsymbol{Z}_i{}'\left[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right]\boldsymbol{Z}_i \\ &= \frac{1}{n-k}(\boldsymbol{Z}_i - \boldsymbol{X}\widehat{\beta}_i)'(\boldsymbol{Z}_i - \boldsymbol{X}\widehat{\beta}_i), \quad i = 1, 2, \ldots, D-1\end{aligned}$$

univariate approach

Also for unknown $\boldsymbol{\Sigma}$ the estimators of β_i using multivariate and univariate approaches are equivalent.

Hypothesis Testing in Multivariate CoDa Model

- Test that regression functions for ilr coordinates z_j are significant
- Verify that the regressor x_i contribute to explanation of the overall variability in $\boldsymbol{Z}$
- General hypothesis on regression parameters

Hypothesis Testing on Significance of the *j*th Regression Function

Hypothesis

$$H_0 : \ \boldsymbol{\beta}_j = \mathbf{0}$$

Test statistic

$$F_{ilr} = \frac{(n-k)\left[\widehat{\boldsymbol{\beta}}_j' \left(\mathbf{X}'\mathbf{X}\right) \widehat{\boldsymbol{\beta}}_j\right]}{k\widehat{\sigma}_{jj}} \overset{H_0}{\sim} F_{k,n-k}$$

- Multivariate and univariate approaches are equivalent when each regression function is tested individually
- Test for several regression functions simultaneously - in general, multivariate approach is necessary
- In case of special choice of ilr transformation, it is sufficient to test only $z_1^{(l)}$, $l = 1, 2, \ldots, D$,

Tests for $D - 1$ Regression Functions Simultaneously

Hypothesis

$$H_{0j}: \ \beta_j = \mathbf{0}, \ j = 1, 2, \ldots, D-1, \quad \text{simultaneously}$$

The most popular test procedures

- Wilks's Lambda (Wilks, 1932)
- Hotelling-Lawley trace
- Pillai-Bartlett trace - the most powerful and robust test
- Roy's largest root

Properties of tests

- Each of test statistics is associated with its own F-statistic
- In some cases, the F statistic is exact and in other cases it is approximate
- In some cases, the test statistics generate identical F statistic and identical probabilities. In other cases they differ.
- As the sample sizes increase the values produced by Pillai-Bartlett trace, Hotelling-Lawley trace and Roy's largest root become similar

Hypothesis Testing on Significance of the *i*th Regressor

Hypothesis

$$H_0 : \ \boldsymbol{B}_{i.} = \left(\beta_{i1}, \beta_{i2}, \ldots, \beta_{i(D-1)}\right) = \mathbf{0}$$

Test statistic

$$F_{pred} = \frac{(n-D-k+2)\left[\hat{\boldsymbol{B}}_{i.}\left(\underline{\mathbf{Z}}'\mathbf{M}_{\mathbf{X}}\underline{\mathbf{Z}}\right)^{-1}\hat{\boldsymbol{B}}_{i.}'\right]}{(D-1)\left\{(\mathbf{X}'\mathbf{X})^{-1}\right\}_{ii}} \overset{H_0}{\sim} F_{D-1,n-D-k+2}$$

General Hypothesis on Regression Parameters

Hypothesis

$$H_0:\ \mathbf{N}\boldsymbol{B} + \boldsymbol{B}_0 = \mathbf{0}$$

Test statistic based on Wilks's Lambda

$$X = -\left[n - k + \mathrm{rank}(\mathbf{N}) - \frac{D + \mathrm{rank}(\mathbf{N})}{2}\right]\ln\Lambda \overset{H_0}{\sim} \chi^2_{(D-1)\mathrm{rank}(\mathbf{N})}$$

$$\Lambda = \frac{\det\left(\underline{\mathbf{Z}}'\mathbf{M}_{\mathbf{X}}\underline{\mathbf{Z}}\right)}{\det\left\{\underline{\mathbf{Z}}'\mathbf{M}_{\mathbf{X}}\underline{\mathbf{Z}} + \left(\mathbf{N}\widehat{\boldsymbol{B}} + \boldsymbol{B}_0\right)'\left[\mathbf{N}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{N}\right]^{-1}\left(\mathbf{N}\widehat{\boldsymbol{B}} + \boldsymbol{B}_0\right)\right\}}.$$

Study of Sediments in Czech Republic

Experiment
Sediment cores were extracted from several reservoirs in CZ. Samples from the cores were air dried, manually ground in agate mortars and subjected to laboratory analyses without further treatment.

Element analysis has been performed by the EDXRF Spectrometer (Energy Dispersive X-ray Fluorescence)

Fifteen selected elements: Al, Si, P, Ti, K, Ca, Fe, Cr, Mn, Ni,Cu, Zn, Zr, Rb, Pb

Units of measurement: c.p.s (counts per seconds)

Problem
Do counts per seconds of these 15 chemical elements depend on the layer depth from which samples were taken?

Brno Reservoir (South Moravia, CZ)

- Capacity - 7.6 mil m^3
- Area - 259 ha
- Max. depth - 23.5 m
- Used for relaxation, hydropower

Nové Mlýny Reservoirs (South Moravia, CZ)

- **Mušov Reservoir** - the upper reservoir, 528 ha, 7.49 mil m^3 (left panel)
  - relaxation
- **Věstonice Reservoir** - the middle reservoir, 1031 ha, 32.062 mil m^3
  - nature reserve with islands for nesting birds
- **Nové Mlýny Reservoir** - the lower reservoir, 1668 ha, 23.8 mil m^3 (right panel)
  - irrigation,small hydropower, relaxation, fishing
- Reservoirs are shallow - depth often does not exceed 2m

Used Models for the Analysis

- **Standard approach:** logarithm of the response variables, linear trend

$$\log(Y_l) = \beta_0^l + \beta_1^l \mathrm{depth} + \varepsilon_l, \quad l = 1, 2, \ldots, 15$$

  - univariate approach to modeling each of the chemical element
  - change the scale of the response variable
  - each chemical element is modeled with its own absolute information

- **Compositional approach:** ilr transformation of the compositional response, linear trend

$$Z_1^{(l)} = \beta_0^l + \beta_1^l \mathrm{depth} + \varepsilon_l, \quad l = 1, 2, \ldots, 15$$

  - univariate approach to modeling each of the 1st ilr coordinate
  - each chemical element is modeled relatively subject to the others elements

Results for Brno Reservoir

Compositional approach:
proportions of 4 elements significantly depend on depth

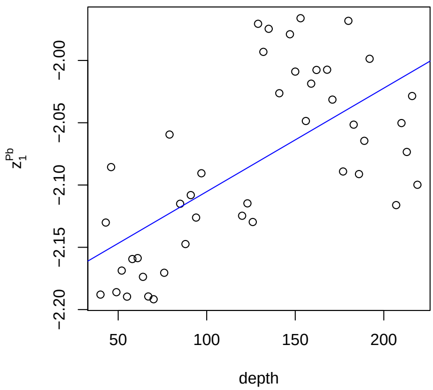
Pb	($\uparrow$, $R^2 = 0.41$)
Mn	($\downarrow$, $R^2 = 0.47$)
Si	($\uparrow$, $R^2 = 0.43$)
K	($\uparrow$, $R^2 = 0.43$)

Standard approach:
2 element significantly depends on depth

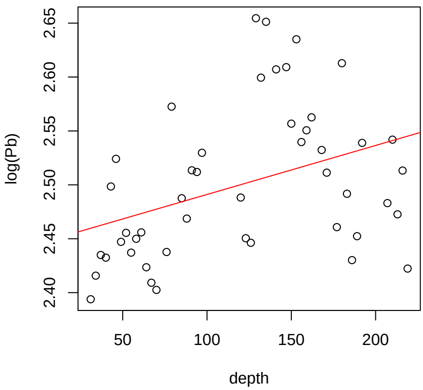
P	($\downarrow$, $R^2 = 0.48$)
Mn	($\downarrow$, $R^2 = 0.43$)

Brno Reservoir - Plumbum

Pb − compositional approach



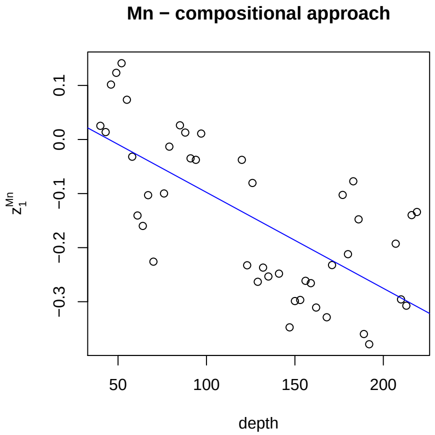
Pb − standard approach

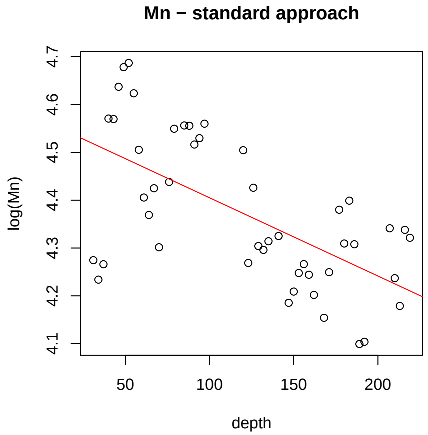


significant ($R^2 = 0.41$)

not significant

Brno Reservoir - Manganum

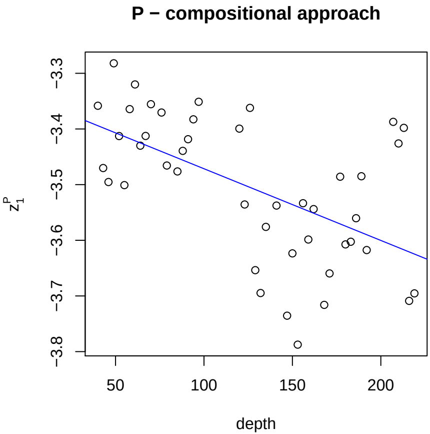




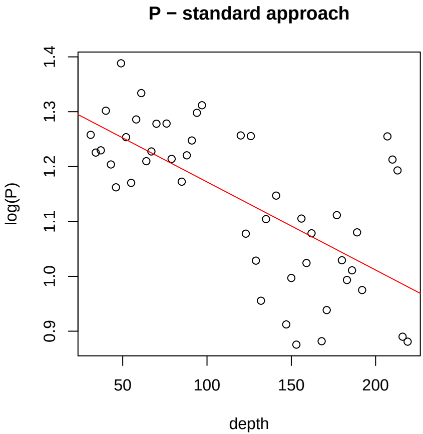
significant ($R^2 = 0.47$)

significant ($R^2 = 0.43$)

Brno Reservoir - Phosphorus



slightly significant ($R^2 = 0.30$)



significant ($R^2 = 0.48$)

Results for Nové Mlýny Reservoirs

Compositional approach:
proportion of 7 elements significantly depends on depth

Ni	$(\uparrow, R^2 = 0.48)$,	P	$(\downarrow, R^2 = 0.61)$,	Cr	$(\uparrow, R^2 = 0.68)$
Fe	$(\uparrow, R^2 = 0.73)$,	K	$(\uparrow, R^2 = 0.42)$,	Ca	$(\downarrow, R^2 = 0.45)$
Al	$(\uparrow, R^2 = 0.57)$				

Standard approach:
7 elements significantly depend on depth

Ni	$(\uparrow, R^2 = 0.63)$,	Mn	$(\downarrow, R^2 = 0.66)$,	Cr	$(\uparrow, R^2 = 0.76)$
Fe	$(\uparrow, R^2 = 0.72)$,	Ti	$(\uparrow, R^2 = 0.82)$,	Si	$(\downarrow, R^2 = 0.63)$
Al	$(\uparrow, R^2 = 0.66)$				

Summary

- Regression with compositional response can be analyzed in a standard way after using isometric log-ratio tranformation.
- Althought regression with coda response leads naturally to multivariate modeling, univariate approach to estimation give the same results.
- When performing inference, in some cases multivariate and univariate approaches equaivalent, in others multivariate approach is necessary.

Key References

Anderson, T.W. (2003) An Introduction to Multivariate Statistical Analysis, 3rd Ed. Wiley.

Aitchison, J. (1986) The statistical analysis of compositional data. Chapman and Hall, London.

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barceló-Vidal (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3), 279–300.

Egozcue, J.J., J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron and P. Filzmoser (2012) Simplicial Regression. The Normal model. Journal of Applied Probability and Statistics Vol. 6, No. 1&2, pp. 87-108.

Egozcue, J.J., V. Pawlowsky-Glahn (2005) Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7), 795–828.

Fišerová, E., K. Hron (2011) On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences 43(4),