Ortogonální regrese pro 3-složkové kompoziční data využitím lineárních modelů

Eva Fišerová a Karel Hron

Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta Univerzity Palackého v Olomouci

Robust 2012

Cíle přednášky

- **ukázat principy modelování lineárního vztahu mezi složkami kompozic**
- **ukázat základní statistické inference pro odhadnutou přímku**

Obsah

Motivace: Analýza věkové struktury populace zemí OSN

Věkové kategorie:

- x_1 mladší 15 let
- x_2 15–60 let
- x_3 starší 60 let

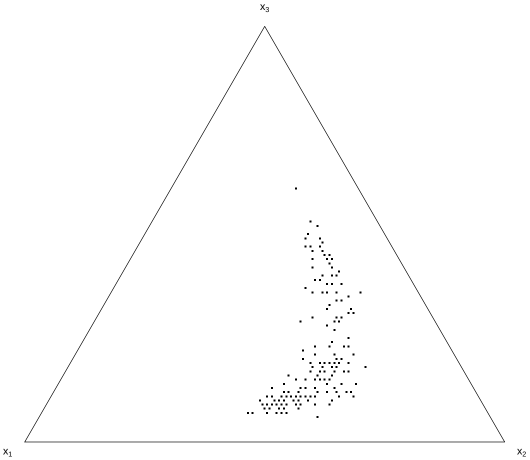
Data ze statistického oddělení OSN obsahují populační strukturu 196 členských zemí

Hrubá data zcela zavádějící, neboť jednotlivé státy mají různý počet obyvatel $\Rightarrow$ převedení dat na poměry - lze vidět relativní příspěvek jednotlivých věkových skupin na celkový počet obyvatel $=$ kompoziční data

Vizualizace 3-složkových kompozičních dat: ternární diagram

- x_1 mladší 15 let
- x_2 15–60 let
- x_3 starší 60 let

- **Konstrukce:** rovnoramenný trojúhelník $X_1X_2X_3$ $\boldsymbol{x} = (x_1, x_2, x_3)'$ zobrazena ve vzdálenosti x_1 od strany protilehlé k vrcholu X_1, atd.



Metodologie při regresní analýze složek kompozičních dat

Výběrový prostor simplex S^3 3-složkové kompozice

$\Downarrow$ ilr transformace

Euklidovský reálný prostor $\mathbb{R}^2$ regrese pro 2 náhodné proměnné

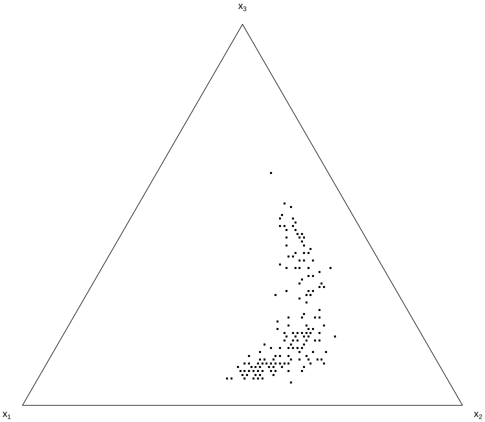
$\Downarrow$ zpětná transformace

Simplex

Ternární diagram

- Původní 3-složkové kompozice vyjádříme v ortonormálních souřadnicích, např.

$$\begin{aligned} z_1 &= \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \\ z_2 &= \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}. \end{aligned}$$

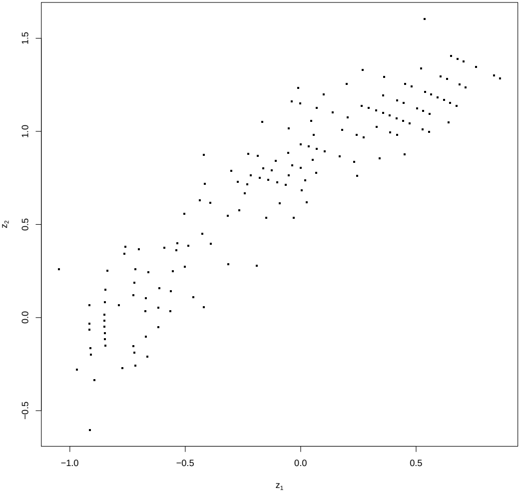


Ilr transformace: $S^3 \mapsto \mathbb{R}^2$

- Původní 3-složkové kompozice vyjádříme v ortonormálních souřadnicích, např.

$$
\begin{aligned}
z_1 &= \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \\
z_2 &= \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.
\end{aligned}
$$

- Nyní lze provést regresi

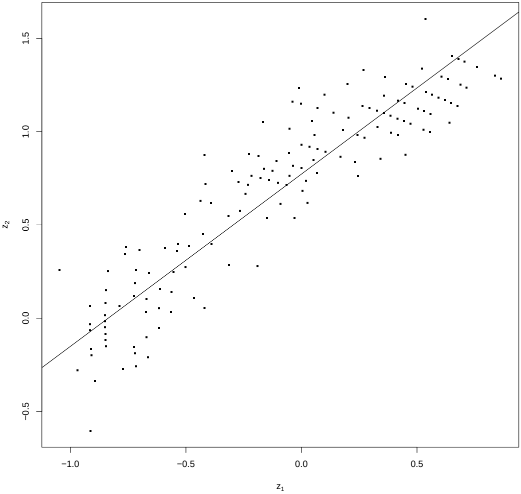


Ilr transformace: $S^3 \mapsto \mathbb{R}^2$

- Původní 3-složkové kompozice vyjádříme v ortonormálních souřadnicích, např.

$$\begin{aligned} z_1 &= \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}}, \\ z_2 &= \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}. \end{aligned}$$

- Nyní lze provést regresi



Regresní analýza mezi ilr souřadnicemi (z_1, z_2)

Chyby měření v obou proměnných

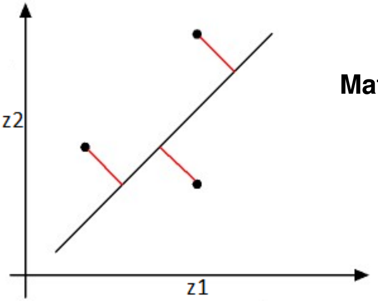
$$\Downarrow$$

Nelze užít metodu nejmenších čtverců

Řešení: Ortogonální regrese (metoda úplných nejmeších čtverců (TLS), kalibrační problém, modelování s chybami v proměnných)

Nejjednodušší úloha ortogonální regrese

Princip: najít přímku $z_2 = \beta_1 + \beta_2 z_1$ tak, aby součet čtverců vzdáleností napozorovaných bodů od této přímky byl minimální.



Matematicky:

$$\min_{\beta_1,\beta_2} \frac{\sum_{i=1}^{n} (z_{2i} - \beta_1 - \beta_2 z_{1i})^2}{\beta_2^2 + 1}$$

Ortogonální regrese: chyby měřeny kolmo k hledané přímce
Metoda nejmenších čtverců: chyby měřeny rovnoběžně s osou z_2

Standardní řešení pro ortogonální regresi - ML

Metoda maximální věrohodnosti (Kendall a Stuart, 1967)

$$\text{absolutní člen:} \quad \widehat{\beta}_1 = \overline{z}_2 - \widehat{\beta}_2 \overline{z}_1,$$

$$\text{směrnice:} \quad \widehat{\beta}_2 = \frac{s^2_{z_2} - s^2_{z_1} + \sqrt{(s^2_{z_2} - s^2_{z_1})^2 + 4 s^2_{z_1 z_2}}}{2 s_{z_1 z_2}}.$$

- $\overline{z}_1$ výběrový průměr, $s^2_{z_1}$ výběrový rozptyl, $s_{z_1 z_2}$ výběrová kovariance

$\rightarrow$ Odhady stejné jako metodou **ortogonálních nejmenších čtverců**

$\rightarrow$ Odhad směrnice je totožný se směrem první hlavní komponenty **metody PCA** (Jackson a Dunlevy, 1988)

Standardní řešení pro ortogonální regresi - SVD

Singulární rozklad matice $(\mathbf{1}_n, \boldsymbol{z}_1, \boldsymbol{z}_2)$

$$\underbrace{(\overbrace{\mathbf{1}_n, \boldsymbol{z}_1}^{\boldsymbol{X}}, \boldsymbol{z}_2)}_{\boldsymbol{T}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}'$$

- $\boldsymbol{U}$ ortonormální matice vl. vektorů $\boldsymbol{T}\boldsymbol{T}'$,
- $\boldsymbol{V}$ ortonormální matice vl. vektorů $\boldsymbol{T}'\boldsymbol{T}$,
- $\boldsymbol{D}$ diagonální matice singulárních hodnot - odmocněná vl. čísel matice $\boldsymbol{T}'\boldsymbol{T}$.

Odhad

$$\widehat{\beta} = (\boldsymbol{X}'\boldsymbol{X} - \lambda_3^2 \boldsymbol{I}_2)^{-1}\boldsymbol{X}'\boldsymbol{z}_2,$$

- λ_3 je nejmenší singulární hodnota ze singulárního rozkladu matice $(\mathbf{1}_n, \boldsymbol{z}_1, \boldsymbol{z}_2)$
- formule je často numericky nestabilní (Markovsky a Huffel, 2007)

Proložení dat přímkou pomocí lineárního modelu s podmínkami typu II

- **Podmínky typu II** (Kubáček et al., 1995)
  - omezení na regresní parametry
  - omezení na další neznámé parametry

nelineární podmínky typu II

- **Statistický model**

$$\begin{pmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\nu} \end{pmatrix} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\nu} = \beta_1 \mathbf{1} + \beta_2 \boldsymbol{\mu}, \quad \mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}.$$

$\boldsymbol{\mu}$ . . . . . . neznámá bezchybná hodnota souřadnice z_1,
$\boldsymbol{\nu}$ . . . . . . neznámá bezchybná hodnota souřadnice z_2,
β_1, β_2 . . . neznámý úsek a směrnice ortogonální regresní přímky.

Odhad ortogonální regresní přímky

MNČ v linearizovaném modelu

- nejlepší nestranný lineární odhad (BLUE)
  - úsek β_1
  - směrnice β_2
  - střední hodnoty $\boldsymbol{\mu}$ a $\boldsymbol{\nu}$
- varianční matice odhadů $\widehat{\boldsymbol{\mu}}$, $\widehat{\boldsymbol{\nu}}$ a $(\widehat{\beta}_1, \widehat{\beta}_2)'$
- kovarianční matice odhadů $\widehat{\boldsymbol{\mu}}$ a $\widehat{\boldsymbol{\nu}}$
- kovarianční matice odhadů $(\widehat{\boldsymbol{\mu}}', \widehat{\boldsymbol{\nu}}')'$ a $(\widehat{\beta}_1, \widehat{\beta}_2)'$
- nestranný odhad σ^2

Všechny výsledky závisí na volbě přibližných hodnot pro linearizaci
$\Longrightarrow$ nutno řešit iteračně

BLUE parametrů μ, ν, β_1 a β_2

$$\widehat{\boldsymbol{\mu}} = \mathbf{Z}_1 + \frac{\beta_2^{(0)}}{\left[\beta_2^{(0)}\right]^2 + 1} \mathbf{M}^{(0)} \left[\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2^{(0)} \left(\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)}\right)\right], \tag{1}$$

$$\widehat{\boldsymbol{\nu}} = \mathbf{Z}_2 - \frac{1}{\left[\beta_2^{(0)}\right]^2 + 1} \mathbf{M}^{(0)} \left[\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2^{(0)} \left(\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)}\right)\right], \tag{2}$$

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \beta_1^{(0)} \\ \beta_2^{(0)} \end{pmatrix} + \begin{pmatrix} n, & \mathbf{1}'\boldsymbol{\mu}^{(0)} \\ \left[\boldsymbol{\mu}^{(0)}\right]'\mathbf{1}, & \left[\boldsymbol{\mu}^{(0)}\right]'\boldsymbol{\mu}^{(0)} \end{pmatrix}^{-1} \times \begin{pmatrix} \mathbf{1}'\left[\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2^{(0)}\left(\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)}\right)\right] \\ \left[\boldsymbol{\mu}^{(0)}\right]'\left[\mathbf{Z}_2 - \boldsymbol{\nu}^{(0)} - \beta_2^{(0)}\left(\mathbf{Z}_1 - \boldsymbol{\mu}^{(0)}\right)\right] \end{pmatrix}, \tag{3}$$

- $\beta_1^{(0)}$, $\beta_2^{(0)}$, $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$ přibližné hodnoty

Iterační algoritmus pro odhad ortogonální regresní přímky

1. Stanovení počátečních hodnot: parametry $\beta_1^{(0)}$, $\beta_2^{(0)}$ ortogonální regresní přímky a bezchybné údaje $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$ tak, že

$$\boldsymbol{\nu}^{(0)} = \beta_1^{(0)}\mathbf{1} + \beta_2^{(0)}\boldsymbol{\mu}^{(0)}.$$

2. Výpočet odhadů $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\boldsymbol{\mu}}$ a $\widehat{\boldsymbol{\nu}}$ pro body $(z_{1k}, z_{2k})'$, $k = 1, \ldots, n$.

3. Stanovení nových počátečních hodnot podle schématu:

$$\boldsymbol{\nu}^{(0)} = \widehat{\boldsymbol{\nu}} + (\widehat{\beta}_2 - \beta_2^{(0)})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(0)}),\ \boldsymbol{\mu}^{(0)} = \widehat{\boldsymbol{\mu}},\ \beta_1^{(0)} = \widehat{\beta}_1,\ \beta_2^{(0)} = \widehat{\beta}_2.$$

4. Kroky 2-4 opakujeme dokud posloupnost odhadů konverguje.

Iterační algoritmus pro odhad ortogonální regresní přímky

1. Stanovení počátečních hodnot: parametry $\beta_1^{(0)}$, $\beta_2^{(0)}$ ortogonální regresní přímky a bezchybné údaje $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$ tak, že
$$\boldsymbol{\nu}^{(0)} = \beta_1^{(0)}\mathbf{1} + \beta_2^{(0)}\boldsymbol{\mu}^{(0)}.$$
2. Výpočet odhadů $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\boldsymbol{\mu}}$ a $\widehat{\boldsymbol{\nu}}$ pro body $(z_{1k}, z_{2k})'$, $k = 1, \ldots, n$.
3. Stanovení nových počátečních hodnot podle schématu:
$$\boldsymbol{\nu}^{(0)} = \widehat{\boldsymbol{\nu}} + (\widehat{\beta}_2 - \beta_2^{(0)})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(0)}),\ \boldsymbol{\mu}^{(0)} = \widehat{\boldsymbol{\mu}},\ \beta_1^{(0)} = \widehat{\beta}_1,\ \beta_2^{(0)} = \widehat{\beta}_2.$$
4. Kroky 2-4 opakujeme dokud posloupnost odhadů konverguje.

Iterační algoritmus pro odhad ortogonální regresní přímky

1. Stanovení počátečních hodnot: parametry $\beta_1^{(0)}$, $\beta_2^{(0)}$ ortogonální regresní přímky a bezchybné údaje $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$ tak, že

$$\boldsymbol{\nu}^{(0)} = \beta_1^{(0)}\mathbf{1} + \beta_2^{(0)}\boldsymbol{\mu}^{(0)}.$$

2. Výpočet odhadů $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\boldsymbol{\mu}}$ a $\widehat{\boldsymbol{\nu}}$ pro body $(z_{1k}, z_{2k})'$, $k = 1, \ldots, n$.

3. Stanovení nových počátečních hodnot podle schématu:

$$\boldsymbol{\nu}^{(0)} = \widehat{\boldsymbol{\nu}} + (\widehat{\beta}_2 - \beta_2^{(0)})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(0)}),\ \boldsymbol{\mu}^{(0)} = \widehat{\boldsymbol{\mu}},\ \beta_1^{(0)} = \widehat{\beta}_1,\ \beta_2^{(0)} = \widehat{\beta}_2.$$

4. Kroky 2-4 opakujeme dokud posloupnost odhadů konverguje.

Iterační algoritmus pro odhad ortogonální regresní přímky

1. Stanovení počátečních hodnot: parametry $\beta_1^{(0)}$, $\beta_2^{(0)}$ ortogonální regresní přímky a bezchybné údaje $\boldsymbol{\mu}^{(0)}$, $\boldsymbol{\nu}^{(0)}$ tak, že
$$\boldsymbol{\nu}^{(0)} = \beta_1^{(0)}\mathbf{1} + \beta_2^{(0)}\boldsymbol{\mu}^{(0)}.$$
2. Výpočet odhadů $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\boldsymbol{\mu}}$ a $\widehat{\boldsymbol{\nu}}$ pro body $(z_{1k}, z_{2k})'$, $k = 1, \ldots, n$.
3. Stanovení nových počátečních hodnot podle schématu:
$$\boldsymbol{\nu}^{(0)} = \widehat{\boldsymbol{\nu}} + (\widehat{\beta}_2 - \beta_2^{(0)})(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}^{(0)}),\ \boldsymbol{\mu}^{(0)} = \widehat{\boldsymbol{\mu}},\ \beta_1^{(0)} = \widehat{\beta}_1,\ \beta_2^{(0)} = \widehat{\beta}_2.$$
4. Kroky 2-4 opakujeme dokud posloupnost odhadů konverguje.

Vlastnosti iteračního algoritmu

- konverguje velmi rychle po několika málo iteracích
- Jestliže iterační algoritmus konverguje, konverguje k maximálně věrohodným odhadů ortogonální regresní přímky (Donevska et. al, 2011)
- Iterační procedura zaručuje, že výsledné odhady splňují podmínku

$$\widehat{\boldsymbol{\nu}} = \widehat{\beta}_1 \mathbf{1} + \widehat{\beta}_2 \widehat{\boldsymbol{\mu}}$$

- Numericky nestabilní, jestliže přímka má tendenci být kolmá k ose z_1 $\longrightarrow$ rotace proměnných (z_1, z_2)

Statistické inference pro ortogonální regresní přímku

Předpoklad: normální rozdělení

$$(\boldsymbol{Z}_1', \boldsymbol{Z}_2')' \sim N_{2n}\left[(\boldsymbol{\mu}', \boldsymbol{\nu}')', \sigma^2 \boldsymbol{I}\right]$$

- ekvivalentní s požadavkem, aby 3-složková kompozice měla normální nebo lognormální rozdělení na simplexu (Aitchison a Shen, 1980; Mateu-Figueras a Pawlowsky-Glahn, 2008).

Statistické inference pro ortogonální regresní přímku

Obvykle prováděny pro směrnici přímky

- pro velké výběry nalezeny intervaly spolehlivosti a testová statistika - založeny na ML (Kendall a Stuart, 1967)
- intervaly spolehlivosti při velkých i malých výběrech - založeny na PCA (Jolicoeur, 1968)
- testová statistika při velkém výběru - založena na PCA (Jackson and Dunlevy, 1988)

Statistické inference užitím lineárního modelu

Lze provést libovolné standardní inference pro regresní přímku

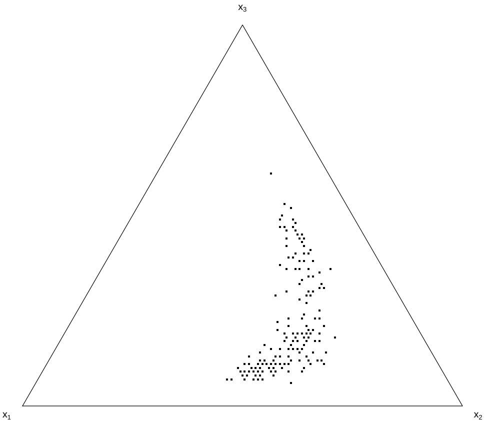
- intervaly spolehlivosti pro úsek a směrnici
- testování hypotéz o úseku a směrnici
- pás spolehlivosti pro ortogonální regresní přímku
- elipsy spolehlivosti pro bezchybné hodnoty

$$\left.\begin{array}{c}\text{Odhad ortogonální regresní přímky}\\ \text{Elipsy spolehlivosti pro bezchybné hodnoty}\\ \text{Pás spolehlivosti pro přímku}\end{array}\right\} \begin{array}{c}\text{transformace}\\ \text{podle rotace } (z_1, z_2)\end{array}$$

$\Rightarrow$ jednoznačná zpětná transformace na simplex

Analýza věkové struktury populace zemí OSN

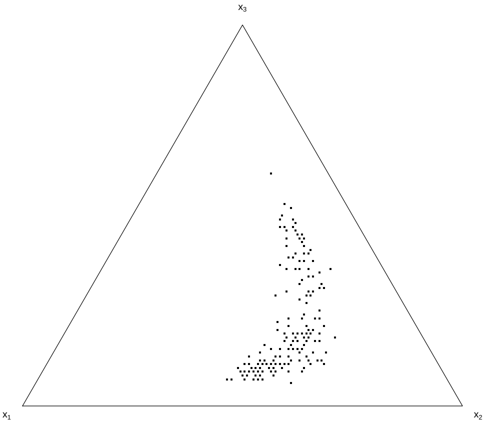
- **Složky kompozic**
  - x_1 mladší 15 let
  - x_2 15–60 let
  - x_3 starší 60 let



Analýza věkové struktury populace zemí OSN

- **Složky kompozic**
  - x_1 mladší 15 let
  - x_2 15–60 let
  - x_3 starší 60 let
- **Ilr transformace** do ortonormálních souřadnic

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}},$$
$$z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.$$



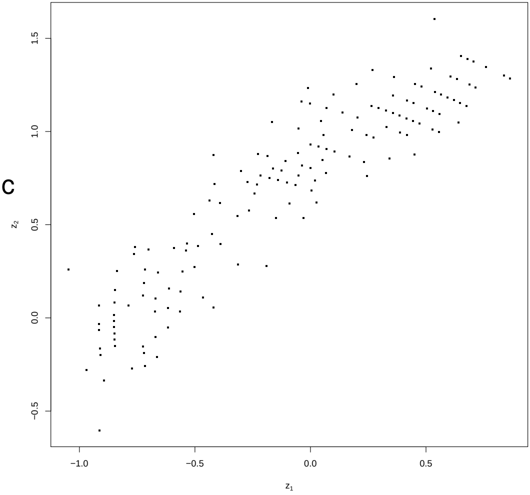
Analýza věkové struktury populace zemí OSN

- **Složky kompozic**
  - x_1 mladší 15 let
  - x_2 15–60 let
  - x_3 starší 60 let

- **Ilr transformace** do ortonormálních souřadnic

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}},$$

$$z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.$$



Analýza věkové struktury populace zemí OSN

- **Složky kompozic**
  - x_1 mladší 15 let
  - x_2 15–60 let
  - x_3 starší 60 let

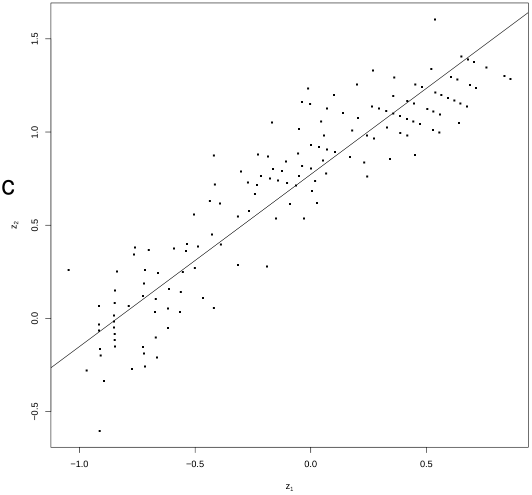
- **Ilr transformace** do ortonormálních souřadnic

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 x_3}},$$

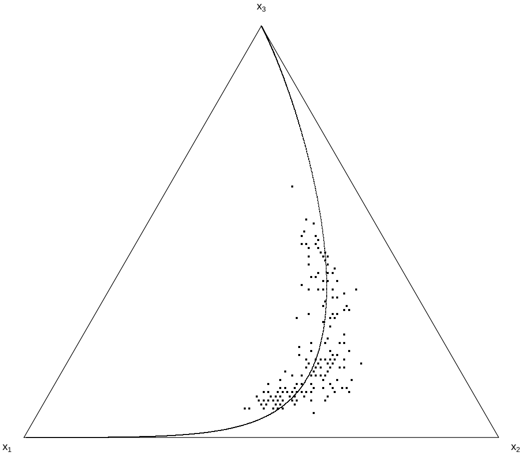
$$z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3}.$$

- Ortogonální regresní přímka

$$z_2 = 0.773 + 0.924 z_1$$

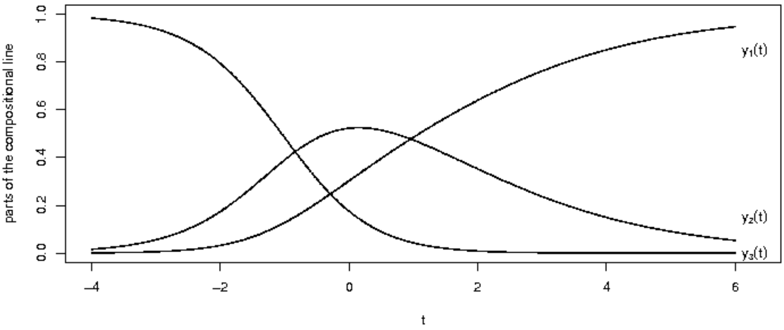


Kompoziční regresní přímka



Složky kompoziční regresní přímky

y_1 mladiství y_2 střední generace y_3 senioři



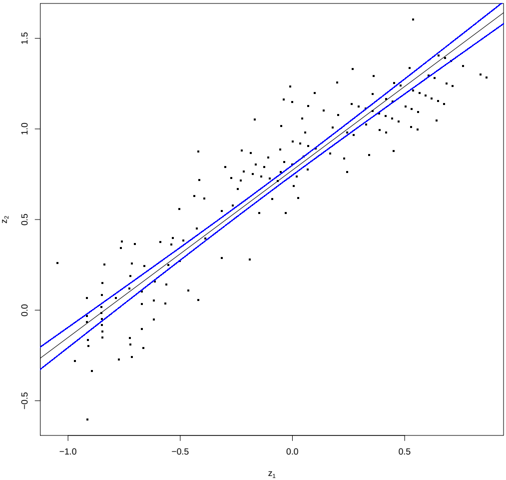
Očekávané podíly věkových skupin při přechodu od států s převažující mladistvou populací k zemím s převažujícími seniory

Odhad ortogonální regresní přímky: iterační algoritmus

- Kritérium konvergence: $\| \widehat{\beta}^{i} - \widehat{\beta}^{i-1} \|_E^2 < 10^{-9} \quad \rightarrow \quad$ 13 iterací
- Odhad regresní přímky: $z_2 = 0.773 + 0.924 z_1$
- Směrodatné odchylky odhadů: 0.014, 0.027
- Testy významnosti regresních parametrů: p-value$<<$ 0.0001
- Shapiro-Wilkův test normality: p-value=0.8452

95% pásy spolehlivosti pro ortogonální regresní přímku

— každý bod samostatně

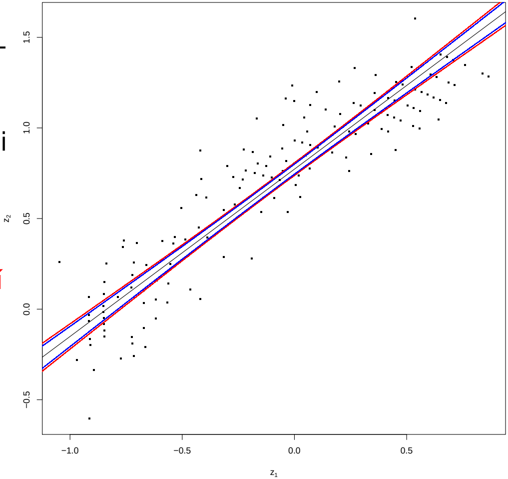


95% pásy spolehlivosti pro ortogonální regresní přímku

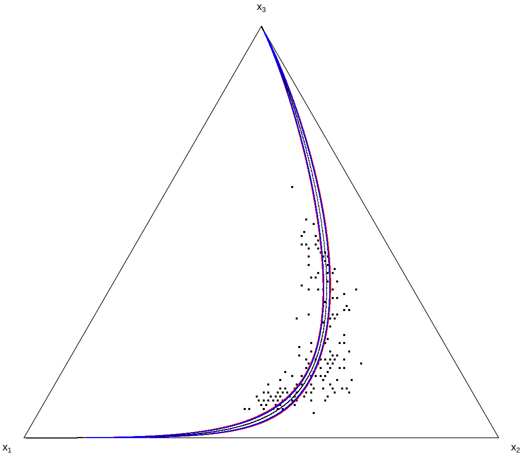
— každý bod samostatně
— sdružený pás spolehlivosti

Pásy spolehlivosti velmi úzké a skoro splývají

⟹ Vysoká přesnost určení ortogonální regresní přímky.



95% pásy spolehlivosti pro kompoziční regresní přímku



Srovnání různých přístupů k ortogonální regresi

Metoda	Interval spolehlivosti	Délka intervalu
Linearní model	$(0.8695, 0.9774)$	0.1079
Ortogonální regrese (ML)	$(0.8703, 0.9797)$	0.1094
Ortogonální regrese (PCA)	$(0.8698, 0.9802)$	0.1105

- Intervaly skoro splývají - velký výběr (196 pozorování)

Závěr

- Ortogonální regrese je vhodná technika pro analýzu vztahu mezi složkami kompozičních dat
  postup: ilr transformace + ortogonální regrese
- Modelovat ve smyslu ortogonální regrese lze i využitím teorie lineárních modelů.

Základní literatura

- Fišerová, E., Hron, K. (2010) Total least squares solution for compositional data using linear models. Journal of Applied Statistics, 37:7 1137–1152.
- Fišerová, E., Hron, K. (2012) Statistical inference in orthogonal regression for three-part compositional data using linear models using a linear model with type-II constraints. Communications in Statistics - Theory and Methods. 41 (13-14), 2367-2385.
- Kendall, M.G., Stuart, A. (1967) The advanced theory of statistics, vol 2. Charles Griffin, London.
- Kubáček, L., Kubáčková, L., Volaufová, J. (1995) Statistical models with linear structures. Veda, Bratislava.
- Jolicoeur, P. (1968) Interval estimation of the slope of the major axis of a bivariate normal distribution in the case of a small sample. *Biometrics*, **24**, 679–682.
- Jackson, J.D. and Dunlevy, J.A. (1988) Orthogonal least squares and the interchangeability of alternative proxy variables in the social sciences. *Journal of the Royal Statistical Society. Series D* **37**, No. 1, 7–14.