Regresní analýza kompozičních dat a její interpretace

Eva Fišerová

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

Robust 2016

ve spolupráci s Ivo Müllerem a Karlem Hronem

Cíle přednášky

- **Představit kompoziční data jako mnohorozměrná pozorování nesoucí relativní informaci**
- **Ukázat metodiku regresní analýzy s kompozičními daty a interpretaci výsledků**
- **Problematiku demonstrovat na reálné úloze analýzy struktury časové alokace**

Motivace

Psychometrická studie na Katedře Psychologie na UP v Olomouci s cílem zkoumat časovou alokaci a charakterové rysy studentů UP.

- dotazníky vyplnilo celkem 414 respondentů (347 žen, 67 mužů)
- osobní údaje - pohlaví, věk, studovaný obor
- časové alokace- zjišťována struktura, absolutní i relativní podíl
  - volný čas; práce/studium; cesta do práce/školy; spánek; domácí povinnosti; hygiena a oblékání; strava
- charakterové rysy
  - sebehodnocení - desetibodový Rosenbergův sebehodnotící dotazník $\Rightarrow$ z-skóre
  - vyhledávání výzev („možnosti překonání sama sebe“)- 4 kategorie $\Rightarrow$ 2 kategorie (1 „vždy“, „skoro vždy“; 0 „nikdy“, „skoro nikdy“)

Původní data časové alokace: modelování absolutních časů věnovaných jednotlivým činnostem

Data vyjádřená v podílech: modelování struktury využití času, vyjadřují relativní příspěvek jednotlivých činností na celkové časové alokaci $=$ **kompoziční data**

Kompoziční data

- **Mnohorozměrná data nesoucí pouze relativní informaci, proměnné reprezentují relativní příspěvky částí na daném celku**
- **Výběrový prostor** D-složkové kompozice: **simplex**

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

pouze podíly jsou informativní $\Rightarrow$ κ libovolná reprezentace kompozic (1 relativní podíly, 100 procenta)

- **data jsou v rozporu s předpoklady většiny statistických metod** (**neřídí se Euklidovskou geometrií v reálném prostoru**)
  - $\rightarrow$ nové metody zpracování na simplexu
  - $\rightarrow$ **vhodná reprezentace kompozic v reálném prostoru**

Ortonormální souřadnice kompozičních dat

Transformace kompozice $\mathbf{y} = (y_1, \ldots, y_D) \in \mathcal{S}^D$ na mnohorozměrné pozorování $\mathbf{z} = (z_1, \ldots, z_{D-1}) \in \mathbb{R}^{D-1}$ pomocí **izometrické transformace logaritmu podílů**

- ortonormální souřadnice se řídí euklidovskou geometrií
- lze užít standardní statistické metody
- výsledky možno interpretovat v souřadnicích nebo zpětně transformovat na simplex
- na simplexu neexistuje kanonická báze - volba dle konkrétní situace

Pivotové ortonormální souřadnice kompozice $\mathbf{y} = (y_1, \ldots, y_D)'$

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

Pivotová souřadnice z_1 zahrnuje všechnu relativní informaci o složce y_1 kompozice y

$$\Downarrow$$

D systémů pivotových ortonormálních souřadnic kompozic $\mathbf{y}^{(l)}$

$$z_k^{(l)} = \sqrt{\frac{D-k}{D-k+1}} \ln \frac{y_k^{(l)}}{\sqrt[D-k]{\prod_{j=k+1}^{D} y_j^{(l)}}}, \quad k = 1, \ldots, D-1$$

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

Pivotové ortogonální souřadnice kompozice $\mathbf{y}^{(l)} = (y_1^{(l)}, \ldots, y_D^{(l)})'$

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{y_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} y_j^{(l)}}}$$

$$\Downarrow$$

$$z_i^{(l)*} = \log_2 \frac{y_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} y_j^{(l)}}}$$

Neovlivní regresní závislost, vede k intuitivnější interpretaci regresních parametrů.

Základní regresní úlohy s kompozičními daty

- Regrese s kompozičními vysvětlujícími proměnnými
- Regrese s kompoziční vysvětlovanou proměnnou
- Regrese mezi složkami kompozice

Regrese s kompozičními regresory - příklad

Úloha: Jak závisí vyhledávání výzev na časové alokaci, sebehodnocení, věku a pohlaví.

Proměnné:

- závisle proměnná: výzvy (ano/ne)
- vysvětlující proměnné:
  - kompoziční: časová alokace (volný čas, práce/studium, spánek, domácí povinnosti, cesta do práce/školy)
  - nekompoziční: věk, pohlaví, sebehodnocení (z-skór)

Regrese s kompozičními regresory - metodika

- kompoziční regresory $\mathbf{x}_c = (x_1, x_2, \ldots, x_D)'$
- nekompoziční regresory $\mathbf{x}_s = (x_{D+1}, \ldots, x_{D+k})'$
- vysvětlovaná proměnná y

Metodika

1. nahrazení $\mathbf{x}_c$ pivotovými ortonormálními souřadnicemi $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ (permutace první složky $\mathbf{x}_c$)
2. mnohonásobná regrese

$$\Longrightarrow \quad D \textbf{ různých modelů mnohonásobné regrese}$$

$$Y_i = \beta_0 + \beta_1^{(l)} z_{i,1}^{(l)} + \cdots + \beta_{D-1}^{(l)} z_{i,D-1}^{(l)} + \beta_D x_{i,D+1} + \cdots + \beta_{D+k-1} x_{i,D+k} + \varepsilon_i^{(l)}$$

$$i = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, D$$

Model s kompozičními regresory - pivotové ortogonální souřadnice

Modely s pivotovými ortonormálními souřadnicemi

$$\mathbf{Y} = (\mathbf{1}, \mathbf{Z}^{(l)}, \mathbf{X}_s)\boldsymbol{\beta}^{(l)} + \boldsymbol{\varepsilon}^{(l)}, \; l = 1, 2, \ldots, D \tag{1}$$

Modely s pivotovými ortogonálními souřadnicemi

$$\mathbf{Y} = (\mathbf{1}, \mathbf{Z}^{(l)*}, \mathbf{X}_s)\boldsymbol{\beta}^{(l)*} + \boldsymbol{\varepsilon}^{(l)}, \; l = 1, 2, \ldots, D \tag{2}$$

Vztahy mezi regresními parametry

$$\beta_0^* = \beta_0, \quad \mathrm{var}(\widehat{\beta}_0^*) = \mathrm{var}(\widehat{\beta}_0)$$

$$\beta_i^{(l)*} = \ln(2)\sqrt{\frac{D-i}{D-i+1}}\beta_i^{(l)}, \; \mathrm{var}(\widehat{\beta}_i^{(l)*}) = [\ln(2)]^2 \frac{D-i}{D-i+1}\mathrm{var}(\widehat{\beta}_i^{(l)}),$$

$$i = 1, \ldots, D-1$$

$$\beta_j^* = \beta_j, \quad \mathrm{var}(\widehat{\beta}_j^*) = \mathrm{var}(\widehat{\beta}_j), \quad j = D, \ldots, D+k-1$$

Kompoziční regresory: $\mathbf{Y} = (\mathbf{1}, \mathbf{Z}^{(l)*}, \mathbf{X}_s)\boldsymbol{\beta}^{(l)*} + \boldsymbol{\varepsilon}^{(l)}$ Interpretace regresních parametrů

Standardní interpretace absolutního členu a parametrů nekompozičních regresorů

Jednotkový přírůstek $z_1^{(l)*} \Rightarrow$ průměrný přírůstek y o $\beta_1^{(l)*}$

$$\Delta z_1^{(l)*} = \log_2 \frac{2x_1^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(l)}}} - \log_2 \frac{x_1^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(l)}}} = 1$$

Dvojnásobný nárůst relativní dominance složky x_l způsobí přírůstek závisle proměnné y v průměru o $\beta_1^{(l)*}$

Model s kompozičními regresory - příklad

Logistická regrese: Závislost vyhledávání výzev na časové alokaci, sebehodnocení, věku a pohlaví.

5 modelů mnohonásobné logistické regrese

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.69708	1.24154	-0.561	0.57448
study/work	0.42200	0.14164	2.979	0.00289*
commuting	-0.06723	0.10961	-0.613	0.53959
sleep	-0.20460	0.17476	-1.171	0.24168
household duties	-0.02904	0.11187	-0.260	0.79519
leisure time	-0.13142	0.12714	-1.034	0.30129
self-esteem	0.45187	0.11105	4.069	4.72e-05*
age	0.04298	0.05516	0.779	0.43586
gender	0.19698	0.15494	1.271	0.20360

Šance na vyhledávání výzev se zvyšuje

- $e^{0.452} = 1.57$x při zvýšení z-skóru sebehodnocení o 1
- $e^{0.422} = 1.53$x s 2-násobným růstem relativní dominance práce/studium

Regrese s kompoziční závisle proměnnou - příklad

Úloha: Co a jakým způsobem ovlivňuje strukturu časové alokace studentů.

Proměnné:

- kompoziční závisle proměnná: časové alokace - 5 složek (volný čas, práce/studium, spánek, domácí povinnosti, cesta do práce/školy)
- vysvětlující proměnné: věk, pohlaví, sebehodnocení (z-skór), přijímání výzev (ano/ne)

Regrese s kompoziční závisle proměnnou - metodika

- nekompoziční regresory $\mathbf{x} = (x_1, \ldots, x_k)'$
- kompoziční vysvětlovaná proměnná $\mathbf{y} = (y_1, \ldots, y_D)'$

Metodika

1. nahrazení $\mathbf{y}$ pivotovými ortonormálními souřadnicemi $\mathbf{z}^{(l)} = (z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ (permutace první složky $\mathbf{y}$)
2. mnohonásobná regrese

$\Longrightarrow$ **D různých modelů mnohonásobné regrese**

$$z_{i,1}^{(l)} = \beta_0^{(l)} + \beta_1^{(l)} x_{i,1} + \cdots + \beta_k^{(l)} x_{i,k} + \varepsilon_i^{(l)}$$

$$i = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, D$$

Model s kompoziční závisle proměnnou - pivotové ortogonální souřadnice

Modely s pivotovými ortonormálními souřadnicemi

$$\mathbf{Z}_1^{(l)} = (\mathbf{1}, \mathbf{X})\beta^{(l)} + \varepsilon^{(l)}, \; l = 1, 2, \ldots, D \tag{3}$$

Modely s pivotovými ortogonálními souřadnicemi

$$\mathbf{Z}_1^{(l)*} = (\mathbf{1}, \mathbf{X})\beta^{(l)*} + \varepsilon^{(l)}, \; l = 1, 2, \ldots, D \tag{4}$$

Vztahy mezi regresními parametry

$$\beta_i^{(l)*} = \log_2(\mathrm{e})\sqrt{\frac{D}{D-1}}\beta_i^{(l)}, \quad i = 0, 1, \ldots, k$$

$$\mathrm{var}(\widehat{\beta}_i^{(l)*}) = [\log_2(\mathrm{e})]^2 \frac{D}{D-1}\mathrm{var}(\widehat{\beta}_i^{(l)})$$

Kompoziční závisle proměnná: $\mathbf{Z}_1^{(I)*} = (\mathbf{1}, \mathbf{X})\boldsymbol{\beta}^{(I)*} + \boldsymbol{\varepsilon}^{(I)}$
Interpretace regresních parametrů

Jednotkový přírůstek regresoru x_j $\Rightarrow$ průměrný přírůstek $Z_1^{(I)*}$ o $\beta_j^{(I)*}$

$$\Delta Z_1^{(I)*} = \log_2 \frac{Y_1^{(I)} \delta_j^{(I)}}{\sqrt[D-1]{\prod_{i=2}^{D} Y_i^{(I)}}} - \log_2 \frac{Y_1^{(I)}}{\sqrt[D-1]{\prod_{i=2}^{D} Y_i^{(I)}}} = \log_2 \delta_j^{(I)} = \beta_j^{(I)*}$$

$$\delta_j^{(I)} = 2^{\beta_j^{(I)*}}$$

Jednotková změna regresoru x_j způsobí zvýšení relativní dominance složky y_I v průměru $2^{\beta_j^{(I)*}}$ krát.

Regrese s kompoziční závisle proměnnou - příklad Jak je ovlivněna struktura časové alokace studentů

5 regresních modelů s pivotovými ortogonálními souřadnicemi korespondujícími se složkami časové alokace jako závisle proměnnými

	study/work	commuting	sleep	household	leisure time
(Intercept)	0.60673	-1.17704	1.50068*	-1.27186*	0.96194*
challenge	0.23723*	-0.02216	-0.01922	-0.07174	-0.13237
self-esteem	-0.02201	0.02367	0.03083	-0.01959	0.08817*
age	0.04117*	-0.00789	0.00186	0.02723	-0.01801
gender	-0.03217	-0.05116	0.14381*	-0.04412	0.22449*

Vyhledávání výzev **zvyšuje relativní dominanci práce/studium** o 18% ($2^{0.237} = 1.18$), 1 rok o 3% ($2^{0.041} = 1.03$).

Relativní dominance spánku je u mužů $2^{2*0.144} = 1.22$ krát větší než u žen.

Nárůst z-skóru sebehodnocení o 1 způsobuje nárůst **relativní dominance volného času** o 6% ($2^{0.088} = 1.06$), u mužů je $2^{2*0.224} = 1.37$ krát větší než u žen.

Regrese mezi složkami kompozice - příklad

Úloha: Jak je ovlivněn volný čas ostatními složkami časové alokace studentů dohromady s nekompozičními proměnnými.

Kompoziční proměnná: časová alokace - 5 složek (volný čas, práce/studium, spánek, domácí povinnosti, cesta do práce/školy)

Nekompoziční proměnné: sebehodnocení, výzvy, věk a pohlaví

Regrese mezi složkami kompozice

- nekompoziční regresory $\mathbf{x}_s = (x_{D+1}, \ldots, x_{D+k})'$
- kompozice $\mathbf{x}_c = (x_0, x_1, \ldots, x_D)'$

Metodika

1. regrese kompoziční závisle proměnné x_0 s kompozičními regresory $(x_1, \ldots, x_D)'$
2. nahrazení $\mathbf{x}_c$ pivotovými ortonormálními souřadnicemi $\mathbf{z}^{(l)} = (z_0, z_1^{(l)}, \ldots, z_{D-1}^{(l)})'$ (permutace složky x_1)

$$\Longrightarrow \quad D \textbf{ různých modelů mnohonásobné regrese}$$

$$Z_{i,0} = \beta_0 + \beta_1^{(l)} z_{i,1}^{(l)} + \cdots + \beta_{D-1}^{(l)} z_{i,D-1}^{(l)} + \beta_D x_{i,D+1} + \cdots + \beta_{D+k-1} x_{D+k} + \varepsilon_i^{(l)}$$

$$i = 1, 2, \ldots, n, \quad l = 1, 2, \ldots, D$$

Podrobněji viz K. Hrůzová „Klasická a ortogonální regrese mezi složkami kompozice“ (poster + úterý 11.20-11.30)

Regrese mezi složkami kompozice - pivotové ortogonální souřadnice

Modely s pivotovými ortonormálními souřadnicemi

$$\mathbf{Z}_0 = (\mathbf{1}, \mathbf{Z}^{(l)}, X_s)\beta^{(l)} + \varepsilon^{(l)}, \; l = 1, 2, \ldots, D \tag{5}$$

Modely s pivotovými ortogonálními souřadnicemi

$$\mathbf{Z}_0^* = (\mathbf{1}, \mathbf{Z}^{(l)*}, X_s)\beta^{(l)*} + \varepsilon^{(l)}, \; l = 1, 2, \ldots, D \tag{6}$$

Vztahy mezi regresními parametry

nekompoziční regresory

$$\beta_j^* = \log_2(\mathrm{e})\sqrt{\frac{D+1}{D}}\beta_j, \quad \mathrm{var}(\widehat{\beta}_j^*) = [\log_2(\mathrm{e})]^2 \frac{D+1}{D}\mathrm{var}(\widehat{\beta}_j^{(l)})$$

kompoziční regresory

$$\beta_j^{(l)*} = \sqrt{\frac{(D+1)(D-j)}{D(D-j+1)}}\beta_j^{(l)}, \; \mathrm{var}(\widehat{\beta}_j^{(l)*}) = \frac{(D+1)(D-j)}{D(D-j+1)}\mathrm{var}(\widehat{\beta}_j^{(l)})$$

Regrese mezi složkami kompozice - interpretace

$$\mathbf{Z}_0^* = (\mathbf{1}, \mathbf{Z}^{(l)*}, \mathbf{X}_s)\beta^{(l)*} + \varepsilon^{(l)},\ l = 1, \ldots, D$$

Jednotkový přírůstek regresoru $z_1^{(l)*}$ $\Rightarrow$ průměrný přírůstek Z_0^* o $\beta_1^{(l)*}$

$$\Delta z_1^{(l)*} = \log_2 \frac{2x_1^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(l)}}} - \log_2 \frac{x_1^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(l)}}} = 1$$

$$\Delta Z_0^* = \log_2 \frac{x_0 \delta^{(l)}}{\sqrt[D]{\prod_{i=1}^{D} x_i^{(l)}}} - \log_2 \frac{x_0}{\sqrt[D]{\prod_{i=1}^{D} x_i^{(l)}}} = \log_2 \delta^{(l)} = \beta_1^{(l)*}, \quad \delta^{(l)} = 2^{\beta_1^{(l)*}}$$

Dvojnásobný nárůst relativní dominance složky x_l způsobí nárůst relativní dominance závislé složky x_0 v průměru $2^{\beta_1^{(l)*}}$ krát.

Nárůst nekompozičního regresoru x_j o jednotku zvyšuje relativní dominanci x_0 v průměru $2^{\beta_{j-1}^{(l)*}}$ krát.

Regrese mezi složkami kompozice - příklad

Úloha: Jak je volný čas studentů ovlivněn zbývajícími složkami časové alokace, sebehodnocením, výzvami, pohlavím a věkem.

	Estimate	Std. Error	Pr(>|t|)	Interpret
(Intercept)	0.47988	0.45492	0.29211	
study/work	-0.15646	0.05576	0.00526*	0.90
commuting	-0.20414	0.04377	4.22e-06*	0.87
sleep	0.33976	0.06374	1.64e-07*	1.27
household duties	0.05734	0.04304	0.18351	-
challenge	-0.09358	0.08937	0.29568	-
self-esteem	0.08353	0.04330	0.05442*	1.06
age	-0.01908	0.01991	0.33852	-
gender	0.16760	0.05951	0.00509*	1.26

Zvýšení relativní dominance volného času o 27% při 2x vyšší dominanci spánku, o 6% při zvýšení z-skóru sebehodnocení o 1

Snížení dominance volného času o 10% při 2x vyšší r. dominanci práce/studium, o 13% při 2x vyšší dominanci dojíždění

Dominance volného času je 1.26x větší pro muže než pro ženy

Závěr

- **Kompoziční data jsou mnohorozměrná pozorování, kde jednotlivé proměnné představují relativní příspěvky částí na nějakém celku**

- Obecná metodika práce s kompozičními daty - **převést data do souřadnic pomocí vhodné transformace logaritmu podílů** a následně použít standardní metody
  - pivotové ortonormální souřadnice
  - **pivotové ortogonální souřadnice** - intuitivní interpretace regresních parametrů

Základní literatura

- J.J. Egozcue, J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron, P. Filzmoser (2012) Simplicial Regression. The Normal model. Journal of Applied Probability and Statistics Vol. 6, No. 1&2, pp. 87-108.
- E. Fišerová, K. Hron (2011) On interpretation of orthonormal coordinates for compositional data. Mathematical Geosciences 43(4), 455-468.
- K. Hron, P. Filzmoser, K. Thomson (2012) Linear regression with compositional explanatory variables. Journal of Applied Statistics, 39(5), 1115-1128.
- K. Hrůzová, V. Todorov, K. Hron, P. Filzmoser (2016) Classical and robust orthogonal regression between parts of compositional data. Statistics.
- I. Müller, K. Hron, E. Fišerová, J.Šmahaj, P. Cakirpaloglu, J. Vančáková (2016) Interpretation of Compositional Regression with application to time budget analysis. Journal of Applied Statistics, under review.