Klasická a robustní ortogonální regrese mezi složkami kompozice

K. Hrůzová, V. Todorov, K. Hron, P. Filzmoser

13. září 2016

Kompoziční data

- kladná reálná čísla nesoucí pouze relativní informaci, $\mathbf{x} = (x_1, \ldots, x_D)'$;
- výběrový prostor - simplex, $\mathcal{S}^D$, na němž je definována Aitchisonova geometrie, která respektuje základní principy analýzy kompozičních dat;
- standardní statistické metody jsou definovány pro celý reálný prostor, proto je nutné kompoziční data nejprve převést do souřadnic

Analýza vztahu mezi složkami kompozice

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \dots, D-1, \qquad (1)$$

- $\mathbf{x} = (x_1, x_2, x_3, x_4)' \ \rightarrow \ x_1 \sim x_2 + x_3 + x_4$;
- vysvětlovaná proměnná: $z_1^{(1)} = \sqrt{\frac{3}{4}} \ln \frac{x_1}{\sqrt[3]{x_2 x_3 x_4}}$;
- vysvětlující proměnné: $z_2^{(k)}, z_3^{(k)}, \ k = 2, 3, 4$, např. pro $k = 2$ dostaneme

$$z_2^{(2)} = \sqrt{\frac{2}{3}} \ln \frac{x_2}{\sqrt{x_3 x_4}}, \ z_3^{(2)} = \sqrt{\frac{1}{2}} \ln \frac{x_3}{x_4}$$

$\rightarrow$ dostaneme 3 regresní modely lišící se přidělenou vysvětlující kompoziční složkou,

$$z_1^{(1)} = \beta_0 + \beta_1 z_2^{(k)} + \beta_2 z_3^{(k)} + \varepsilon, \ k = 2, 3, 4. \qquad (2)$$

Ortogonální regrese

- vysvětlovaná i vysvětlující proměnné pochází z jedné kompozice $\rightarrow$ všechny proměnné jsou zatíženy chybou;
- patří mezi tzv. modely s chybou v proměnných (EIV), tvoří speciální případ metody totálních čtverců (TLS);
- při odhadu TLS se využívá singulárního rozkladu datové matice, který se (pro centrované proměnné) dá nahradit metodou hlavních komponent;
- odhady parametrů β získáme užitím hodnot normálového vektoru, $\mathbf{n} = (n_1, n_2, n_3)$, konkrétně

$$b_0 = \frac{\mathbf{t}'\mathbf{n}}{n_3}, \ b_1 = -\frac{n_1}{n_3}, \ b_2 = -\frac{n_2}{n_3}.$$

Robustní MM-odhady

- regresní modely jsou citlivé na odlehlé hodnoty;
- MM-odhady jsou velmi eficientní, jestliže chyby mají normální rozdělení, jejich bod zlomu je 0.5 a mají ohraničenou influenční funkci;
- mnohorozměrné MM-odhady jsou rozšířením S-odhadů, založeným na dvou ztrátových funkcích ρ_0 a ρ_1 splňujících dvě podmínky:
  1. ρ je symetrická a dvakrát spojitě diferencovatelná, s $\rho(0) = 0$;
  2. ρ je striktně rostoucí na intervalu $[0, k]$ a konstantní na $[k, \infty]$ pro konečné číslo k.

Robustní MM-odhady

- pro daný vektor souřadnic $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)' \in \mathbb{R}^{D-1}$ jsou MM-odhady definovány ve 2 krocích:
  1. S-odhady umístění a kovariance, $(\widetilde{\boldsymbol{\mu}}_n, \widetilde{\boldsymbol{\Sigma}}_n)$, minimalizují $|\mathbf{C}|$ vzhledem k

     $$\frac{1}{n}\sum_{i=1}^{n} \rho_0 \left([(\mathbf{z}_i - \mathbf{t})' \mathbf{C}^{-1} (\mathbf{z}_i - \mathbf{t})]^{1/2} \right) = b,$$

     pro všechna $(\mathbf{t}, \mathbf{C}) \in \mathbb{R}^{D-1}$; dále $\widehat{s} = |\widetilde{\boldsymbol{\Sigma}}_n|^{1/[2(D-1)]}$.
  2. MM-odhady pro umístění a tvar, $(\widehat{\boldsymbol{\mu}}_n, \ \widehat{\boldsymbol{\Gamma}}_n)$, minimalizují

     $$\frac{1}{n}\sum_{i=1}^{n} \rho_1 \left(\left[(\mathbf{z}_i - \mathbf{t})' \mathbf{S}^{-1} (\mathbf{z}_i - \mathbf{t}) \right]^{1/2} / \widehat{s} \right)$$

     pro všechna $\mathbf{t}$ a všechny symetrické, pozitivně definitní matice $\mathbf{S}$ s $|\mathbf{S}| = 1$;
     $\rightarrow$ MM-odhad kovarianční matice: $\widehat{\boldsymbol{\Sigma}}_n = \widehat{s}^2 \widehat{\boldsymbol{\Gamma}}_n$.

Neparametrický bootstrap

- neparametrický bootstrap nám umožňuje odhadovat výběrové rozdělení statistiky empiricky bez jakýchkoliv předpokladů o populaci a bez explicitního odvozování výběrového rozdělení;
- výpočet $p-$hodnoty pro testování oboustranné hypotézy o nulovosti regresních parametrů: $p_i = 2 \cdot \min\{l_i, h_i\}/R$, kde R je počet opakování, l_i (h_i) je počet odhadnutých parametrů menších (větších) než nula;
- percentilové intervaly: $(b^*_{i(l)}, b^*_{i(u)})$, $i = 1, \ldots, D-1$, kde $l = [(R+1)\alpha/2], u = [(R+1)(1-\alpha)/2]$.

„Rychlý a robustní“ bootstrap

- teorie robustních odhadů je omezená pro asymptotické výsledky;
- v případě robustních odhadů je obecný bootstrap nevhodný:
  $\rightarrow$ početní složitost robustních odhadů;
  $\rightarrow$ nestálost v případě odlehlých hodnot;
- robustní odhady mohou být vyjádřeny pomocí hladkých rovnic s fixními body, což nám umožňuje počítat pouze rychlou aproximaci odhadů v každém bootstrapovém výběru.

Vztah mezi aktivitami hrubé přidané hodnoty (GVA)

- VA obecně odráží příspěvek práce a kapitálu na výrobu;
- GVA: rozdíl mezi produkcí a spotřebou;
- GVA může být rozdělena na tyto činnosti:
  1. zemědělství;
  2. výroba;
  3. další průmysl;
  4. služby;

  $\rightarrow$ GVA může být vyjádřena jako součet těchto čtyř činností
- $\mathbf{Y}_{\text{man}} \sim \mathbf{X}_{\text{agr}} + \mathbf{X}_{\text{ind}} + \mathbf{X}_{\text{srv}}$;
- datový soubor pochází z databáze Světové banky (http://data.worldbank.org), obsahuje data ze 131 světových zemí z roku 2010.

Výsledky pro klasickou metodu

classical	par. estimate	perc. CI	p-value
intercept	-2.151	(-4.464, -1.559)	0.002
$b_1^{(2)}$ (agr)	-0.394	(-0.584, -0.115)	0.020
$b_1^{(3)}$ (ind)	-0.878	(-2.745, -0.498)	0.000
$b_1^{(4)}$ (srv)	1.272	(0.858, 2.978)	0.002
robust	par. estimate	perc. CI	p-value
intercept	-2.311	(-6.391, -1.666)	0.006
$b_1^{(2)}$ (agr)	-0.389	(-0.605, 0.180)	0.116
$b_1^{(3)}$ (ind)	-1.075	(-4.994, -0.556)	0.002
$b_1^{(4)}$ (srv)	1.464	(0.996, 5.184)	0.002

Tabulka : Souhrn výsledků regresní analýzy.

Závěr

- při práci s kompozičními daty je třeba nejprve data převést do reálného prostoru užitím vhodných souřadnic;
- jestliže jsou jak závisle, tak i nezávisle proměnné zatíženy chybou, standardní regrese by vedla k nekonzistentním odhadům $\rightarrow$ modely „chyby v proměnných";
- ortogonální regrese je obecně řešená singulárním rozkladem datové matice, což může být nahrazeno PCA (rozklad varianční matice na vlastní čísla);
- statistické inference byly získány užitím bootstrapu (v případě standardní metody) a FRB (pro případ robustní metody).

Reference

Fox J. (2002).

Bootstrapping Regression Models. Appendix to an R and S-PLUS Companion to Applied Regression.

http://statweb.stanford.edu/ tibs/sta305files/FoxOnBootingRegInR.pdf

Hron, K., P. Filzmoser and K. Thompson (2012).

Linear regression with compositional explanatory variables.

Journal of Applied Statistics 39(5), 1115–1128.

Hrůzová K., Todorov V., Hron K., Filzmoser P. (2016).

Classical and robust orthogonal regression between parts of compositional data.

Statistics, DOI: 10.1080/02331888.2016.1162164.

Markovsky, I. and S. Van Huffel (2007).

Overview of total least-squares methods.

ScienceDirect 87, pp. 2283–2302.

Pawlowsky-Glahn V., Egozcue J.J., Tolosana-Delgado R. (2015).
Modeling and Analysis of Compositional Data.
Wiley, Chichester.

Rousseeuw, P. and M. Hubert (2013).
High-Breakdown Estimators of Multivariate Location and Scatter.
In C. Becker, R. Fried and S. Kuhnt, editors, *Robustness and Complex Data Structure*.
Springer, Verlag Berlin Heidelberg (Germany), pp. 49–66.

Van Aelst S., Willems G. (2013).
Fast and robust bootstrap for multivariate inference: The R package FRB.
Journal of Statistical Software **53**(3).