EKONOMICKÁ APLIKACE KOMPOZIČNÍHO REGRESNÍHO MODELU

Klára Hrůzová[1,2], Karel Hron[1,2]

[1] Katedra matematické analýzy a aplikací matematiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci
[2] Katedra geoinformatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci

Robust 2014
19. – 24. ledna 2014, Jetřichovice

Obsah

Dvousložková kompozice

- $\mathbf{x} = (x, c - x)'$, kde c je konstanta součtu
- základní operace Aitchisonovy geometrie speciálně pro takto nadefinované kompozice:
  - Perturbace: $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(xy, (c - x)(c - y))$;
  - Mocninná transformace: $\alpha \odot \mathbf{x} = \mathcal{C}(x^\alpha, (c - x)^\alpha)$;
  - Skalární součin: $||\mathbf{x}||_A = \frac{1}{\sqrt{2}} \left| \ln \frac{x}{c-x} \right|$;
  - Vzdálenost: $d_A(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{2}} \left| \ln \frac{x}{c-x} - \ln \frac{y}{c-y} \right|$,
    kde $\mathbf{x} = (x, c - x)'$, $\mathbf{y} = (y, c - y)'$, α je reálná konstanta a $\mathcal{C}$ označuje operaci uzávěru.

Regresní model

Pro kompoziční data můžeme zavést regresní model (resp. v nejjednodušším případě analogii regresní přímky) užitím Aitchisonovy geometrie:

$$\mathbf{y}_i = \boldsymbol{\beta}_0 \oplus \beta_1 \odot \mathbf{x}_i \oplus \boldsymbol{\varepsilon}_i, \;\; i = 1, \ldots, r, \tag{1}$$

s kompozičním regresním parametrem $\boldsymbol{\beta}_0$, skalárním parametrem β_1 a kompoziční chybou $\boldsymbol{\varepsilon}_i$.

Izometrická logratio transformace

Pro dvousložkovou kompozici definujeme ilr transformaci ve tvaru:

$$x^* = \mathrm{ilr}(\mathbf{x}) = \frac{1}{\sqrt{2}} \ln \frac{x}{c - x}. \qquad (2)$$

$\rightarrow$ transformace je proporcionální k logitové transformaci
$\rightarrow$ metodika logratio souřadnic umožňuje aplikovat standardní statistické metody a předpokládat normalitu souřadnic [Egozcue et al.2011]

Regresní přímka

$$y_i^* = \beta_0^* + \beta_1 x_i^* + \varepsilon_i^*, \quad i = 1, \dots, r,$$

kde neznámé parametry β_0^*, β_1 odhadujeme metodou nejmenších čtverců.

Za předpokladu normality závisle proměnné $y^* \equiv y^*(x^*)$ je konfidenční interval pro střední hodnotu y^* v x^* definován jako

$$\widehat{y^*(x^*)} \pm t_{1-\alpha/2,r-2}\sqrt{s^2\left[\frac{1}{r} + \frac{(x^* - \bar{x}^*)^2}{\sum_{i=1}^{r}(x_i^* - \bar{x}^*)^2}\right]}$$

se spolehlivostí $(1-\alpha)$.
Predikční interval pro y^*

$$\widehat{y^*(x^*)} \pm t_{1-\alpha/2,r-2}\sqrt{s^2\left[1 + \frac{1}{r} + \frac{(x^* - \bar{x}^*)^2}{\sum_{i=1}^{r}(x_i^* - \bar{x}^*)^2}\right]}.$$

Fitované hodnoty

Fitované hodnoty pro původní kompozici y_i získáme aplikací inverzní ilr transformace

$$\widehat{y}_i = \mathrm{ilr}^{-1}(\widehat{y}_i^*) = \frac{c \exp\left(\sqrt{2}\widehat{y}_i^*\right)}{1 + \exp\left(\sqrt{2}\widehat{y}_i^*\right)}. \tag{3}$$

Concentration-response models

- odhad ekologického rizika z chemického znečištění
- na základě koncentrace toxické látky x_i (v mg/l) měříme proporci odezvy p_i, kde $(0 < p_i < 1)$
- logaritmická transformace koncentrace (x_i)

Modely používané nyní

- model v základním tvaru:

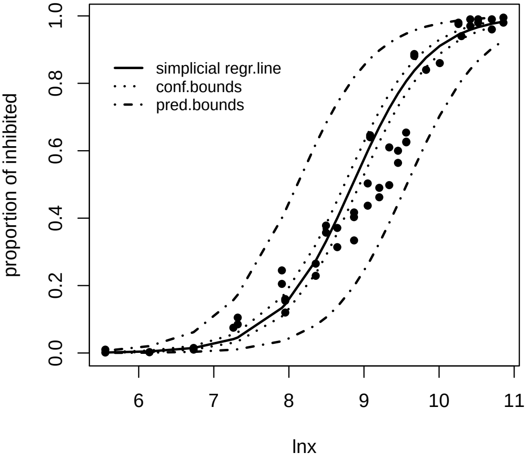
$$y_i = f\left(x_i, \boldsymbol{\beta}\right) + \varepsilon_i, \;\; i = 1, \ldots, r, \tag{4}$$

  kde funkce $f(x_i, \boldsymbol{\beta})$ reprezentuje průměr reakcí
- nejužívanější regresní funkce f

Model (RM)	**Regression function** $f(x_i^*, \boldsymbol{\beta})$
Logit (L)	$\frac{\exp\left(\beta_0+\beta_1 x_i^*\right)}{1+\exp\left(\beta_0+\beta_1 x_i^*\right)}$
Probit (P)	$\Phi\left(\beta_0 + \beta_1 x_i^*\right)$
Generalized Logit (GL)	$\left(\frac{\exp\left(\beta_0+\beta_1 x_i^*\right)}{1+\exp\left(\beta_0+\beta_1 x_i^*\right)}\right)^{\beta_2}$
Weibull (W)	$\exp\left(-\exp\left(\beta_0 + \beta_1 x_i^*\right)\right)$

Užití kompozičního modelu

- $x_i^* = \frac{1}{\sqrt{2}} \ln \frac{x_i}{10^6 - x_i}$;
- $p_i^* = \frac{1}{\sqrt{2}} \ln \frac{p_i}{1 - p_i}$;
- aplikace regresní přímky
- užití inverzní ilr transformace pro zobrazení dat v původním prostoru


simplicial regr.line
conf.bounds
pred.bounds
proportion of inhibited
0.0
0.2
0.4
0.6
0.8
1.0
6
7
8
9
10
11
lnx

Odhad efektivní koncentrace EC_P

- odhad míry koncentrace, při které bychom dosáhli $P = 100 \cdot p\%$-ního efektu
- pro $P = 5\%$ spočítáme ilr souřadnice $p_5^* = \frac{1}{\sqrt{2}} \log \frac{0.05}{0.95}$
- odpovídající odhad ilr koncentrace EC_5 získáme aplikací fitované regresní přímky

$$\widehat{EC_5}^* = \frac{1}{\widehat{\beta}_1}(p_5^* - \widehat{\beta}_0^*), \tag{5}$$

- výsledná koncentrace EC_5 je získána užitím inverzní ilr transformace

$$\widehat{EC_5} = \mathrm{ilr}^{-1}(\widehat{EC_5}^*) = \frac{1 \exp\left(\sqrt{2}\widehat{EC_5}^*\right)}{1 + \exp\left(\sqrt{2}\widehat{EC_5}^*\right)} \tag{6}$$

- konfidenční interval můžeme získat užitím teorie kalibrace v lineárních regresních modelech:

$$\overline{x}^* + d_1 \leq EC_5^* \leq \overline{x}^* + d_2,$$

kde d_1 a d_2 jsou kořeny kvadratické rovnice

$$d^2 \left[\widehat{\beta}_1^2 - \frac{t^2_{1-\alpha/2,r-2} s^2}{\sum_{i=1}^r (x_i^* - \overline{x}^*)^2} \right] - 2d\widehat{\beta}_1 (p_5^* - \overline{p}^*) +$$

$$+ \left[(p_5^* - \overline{p}^*)^2 - t^2_{1-\alpha/2,r-2} s^2 \left(1 + \frac{1}{r}\right) \right] = 0.$$

- po nějakém počítání dostaneme konfidenční intervaly pro EC_5^* ve formě

$$EC_5^* \in \overline{x}^* + \Big\{ (p_5^* - \overline{p}^*)\widehat{\beta}_1 \pm$$

$$\pm t_{1-\alpha/2, r-2} s \left[\frac{(p_5^* - \overline{p}^*)^2}{\sum_{i=1}^r \left(x_i^* - \overline{x}^*\right)^2} + \left(1 + \frac{1}{r}\right) H \right]^{1/2} \Bigg\} / H,$$

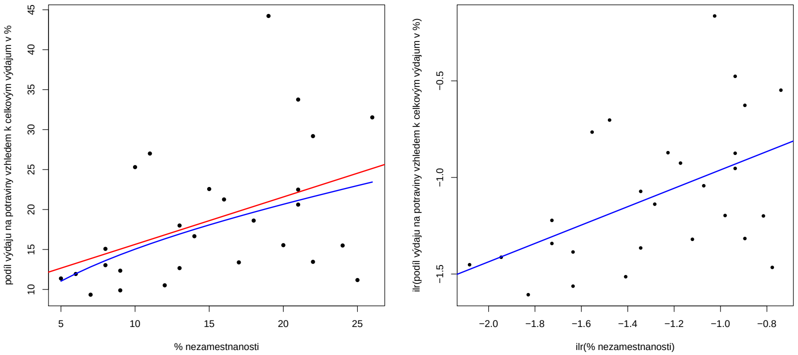
kde $H = \widehat{\beta}_1^2 - \frac{t_{1-\alpha/2, r-2} s^2}{\sum_{i=1}^r \left(x_i^* - \overline{x}^*\right)^2}$.

Ekonomická aplikace

- data různorodá a charakterizovaná odlišným způsobem jejich získání
- důležitý je trend ve vztahu mezi oběma proměnnými a signifikantnost parametru $\widehat{\beta_1}$

Ekonomický příklad

- závislost relativních výdajů na potraviny (v % na celkových rodinných výdajích) na výši nezaměstnanosti
- můžeme hovořit o existenci rostoucího trendu
- hypotézu o normalitě pro standardizovaná rezidua nebylo možné zamítnout na hladině 0,05 žádným z užitých testů (Shapiro-Wilk, Anderson-Darling, Kolmogorov-Smirnov)
  $\Rightarrow$ test nulovosti koeficientu u lineární složky regresní funkce aplikací standardní T_1 statistiky (pro ilr souřadnice)
- p-hodnota (0,0089) svědčí ve prospěch alternativy na standardní hladině 0,05
  $\Rightarrow$ se zvyšující se nezaměstnaností roste relativní podíl výdajů na potraviny



Obr.3: Obrázek znázorňuje závislost relativních podílů výdajů na potraviny na nezaměstnanosti pro původní data (vlevo) a ilr souřadnice (vpravo).

Výhody kompozičního regresního modelu

- zachovává kompoziční charakter závisle i nezávisle proměnných (vyjádřených v procentech, proporcích,atd.)
- jednoduchý model s dobrou interpretací výsledků
- z regresní přímky můžeme odvodit odpovídající statistické inference (konfidenční a predikční interval)
- dobré interpolační vlastnosti modelu
- logratio metoda umožňuje zavést předpoklad normality

Reference

Egozcue, J. J., J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron and P. Filzmoser (2011).
Simplicial regression. The normal model.
Journal of Applied Probability and Statistics 6 (1-2), pp. 87–108.

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

Monti, G.S., S. Migliorati, K. Hron, K. Hrůzová and E. Fišerová (2013).
Log-ratio approach in curve fitting for concentration-response experiments.
Environmental and Ecological Statistics (in approve).