INFORMAČNÍ BULLETIN

České statistické společnosti

Ročník 23, číslo 2, červen 2012

VÝZNAM VÁHOVÉ FUNKCE V LINEARIZOVATELNÝCH REGRESNÍCH MODELECH

Bohumil Maroš, Marie Budíková

Adresa: ÚM FSI VUT, Technická 2896/2, 616 69 Brno;
ÚMS PřF MU, Kotlářská 2, 611 37 Brno

E-mail: maros@fme.vutbr.cz, budikova@math.muni.cz

Abstrakt: V článku se zabýváme jednoduchým nelineárním regresním modelem, který může být vhodnou transformací převeden na model lineární v parametrech. Popisujeme dva problémy vznikající při této transformaci a ukazujeme, jak je eliminovat pomocí váhové funkce. Doporučujeme opakované použití váhové funkce, přičemž její vliv na kvalitu predikovaných hodnot posuzujeme pomocí relativního zlepšení reziduálního součtu čtverců. Použití váhové funkce ilustrujeme na dvou příkladech z biologické praxe.

Abstract: In this paper we deal with a simple non-linear regression model, which can be moved to a linear domain by a suitable transformation. We describe two problems originating from this transformation and show how it is eliminated by the weight function. We recommend to use the weight function repeatedly. Effect of weight function is assessed by a relative improvement in the residual sum of squares. The using of the weight function is illustrated by two examples of biological practices.

1. Úvod

Uvažme regresní model s jednou vysvětlující proměnnou

$$Y = f(x; \beta_0, \beta_1, \ldots, \beta_k) + \varepsilon,$$

kde $y = f(x; \beta_0, \beta_1, \ldots, \beta_k)$ je funkce nelineární v parametrech $\beta_0, \beta_1, \ldots, \beta_k$ a ε je náhodná odchylka. Pořídíme n dvojic pozorování (x_i, Y_i), $i = 1, \ldots, n$ a předpokládáme, že pro všechna $i = 1, \ldots, n$ platí

$$Y_i = f(x_i; \beta_0, \beta_1, \ldots, \beta_k) + \varepsilon_i,$$

přičemž $\varepsilon_i \sim N(0, \sigma^2)$ a $C(\varepsilon_i, \varepsilon_j) = 0$ pro $i \neq j$. Pokud můžeme funkci $y = f(x; \beta_0, \beta_1, \ldots, \beta_k)$ převést vhodnou transformační funkcí $F(y)$ na funkci

$$F(y) = \beta_0 f_0(x) + \beta_1 f_1(x) + \ldots + \beta_k f_k(x),$$

která je lineární v parametrech $\beta_0, \beta_1, \ldots, \beta_k$, říkáme, že jde o linearizovatelný regresní model.

Např. funkce a) $y = \sqrt{\beta_0 + \beta_1 x}$, b) $y = \beta_0 \beta_1^x$, c) $y = \frac{1}{\beta_0 + \beta_1 x}$ linearizujeme na funkce a) $y^2 = \beta_0 + \beta_1 x$, tj. $F(y) = y^2$, b) $\ln y = \ln \beta_0 + \ln \beta_1 x$, tj. $F(y) = \ln y$, c) $\frac{1}{y} = \beta_0 + \beta_1 x$, tj. $F(y) = \frac{1}{y}$. Odhad parametrů β_j, $j = 0, \ldots, k$, lze provést metodou nejmenších čtverců:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{b}^{(0)} = \begin{pmatrix} b_0^{(0)} & b_1^{(0)} & \cdots & b_k^{(0)} \end{pmatrix}^T = \left(\boldsymbol{X}^T \cdot \boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}^T \cdot F(\boldsymbol{Y}),$$

kde

$$\boldsymbol{X} = \begin{pmatrix} f_0(x_1) & f_1(x_1) & \cdots & f_k(x_1) \\ f_0(x_2) & f_1(x_2) & \cdots & f_k(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ f_0(x_n) & f_1(x_n) & \cdots & f_k(x_n) \end{pmatrix}$$

a $F(\boldsymbol{Y}) = \begin{pmatrix} F(Y_1) & F(Y_2) & \cdots & F(Y_n) \end{pmatrix}^T$. Reziduální součet čtverců S_0 pro nelineární funkci $y = f(x; \beta_0, \beta_1, \ldots, \beta_k)$ je pak roven

$$S_0 = \sum_{i=1}^{n} \left(Y_i - \widehat{y}_i^{(0)}\right)^2,$$

kde $\widehat{y}_i^{(0)} = f\left(x_i; b_0^{(0)}, b_1^{(0)}, \ldots, b_k^{(0)}\right)$, $i = 1, \ldots, n$.

2. Problémy při linearizaci a jejich eliminace

Při použití transformační funkce $F(y)$ dojde k porušení normality náhodných odchylek $\varepsilon_1, \ldots, \varepsilon_n$ a k porušení homogenity jejich rozptylů. Abychom tyto problémy aspoň částečně eliminovali, aplikujeme na linearizovanou regresní funkci váhovou funkci $w(y)$, která se vyjádří pomocí vzorce $w(y) = \left[\frac{\mathrm{d}F(y)}{\mathrm{d}y}\right]^{-2}$. Pro funkce z bodů a), b), c) dostáváme váhové funkce tvaru:

a) $w(y) = \left(\frac{\mathrm{d}y^2}{\mathrm{d}y}\right)^{-2} = (2y)^{-2}$, protože konstantu lze zanedbat, tak $w(y) = \frac{1}{y^2}$;

b) $w(y) = \left(\frac{\mathrm{d}}{\mathrm{d}y} \ln y\right)^{-2} = y^2$;

c) $w(y) = \left(\frac{\mathrm{d}}{\mathrm{d}y} \frac{1}{y}\right)^{-2} = y^4$.

Při použití váhové funkce získáme odhad neznámých regresních parametrů váženou metodou nejmenších čtverců:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{b}^{(1)} = \begin{pmatrix} b_0^{(1)} & b_1^{(1)} & \dots & b_k^{(1)} \end{pmatrix}^T = \left(\boldsymbol{X}^T \cdot \boldsymbol{W} \cdot \boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}^T \cdot \boldsymbol{W} \cdot F(\boldsymbol{Y}),$$

kde diagonální matice $\boldsymbol{W}$ má prvky $w_{ii}^{(1)} = w\left(y_i^{(0)}\right)$. Vypočteme reziduální součet čtverců S_1 s opravenými odhady parametrů β_j, $j = 0, \dots, k$, tedy

$$S_1 = \sum_{i=1}^{n} \left(Y_i - \widehat{y}_i^{(1)}\right)^2,$$

kde $\widehat{y}_i^{(1)} = f\left(x_i; b_0^{(1)}, b_1^{(1)}, \dots, b_k^{(1)}\right)$. Mnohdy zjistíme, že hodnota reziduálního součtu čtverců s použitím váhové funkce poklesla, tzn. $S_1 < S_0$. Iteračním způsobem budeme pokračovat:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{b}^{(r)} = \begin{pmatrix} b_0^{(r)} & b_1^{(r)} & \dots & b_k^{(r)} \end{pmatrix}^T = \left(\boldsymbol{X}^T \cdot \boldsymbol{W} \cdot \boldsymbol{X}\right)^{-1} \cdot \boldsymbol{X}^T \cdot \boldsymbol{W} \cdot F(\boldsymbol{Y}),$$

$$w_{ii}^{(r)} = w\left(y_i^{(r-1)}\right), \quad S_r = \sum_{i=1}^{n} \left(Y_i - \widehat{y}_i^{(r)}\right)^2,$$

kde $\widehat{y}_i^{(r)} = f\left(x_i; b_0^{(r)}, b_1^{(r)}, \dots, b_k^{(r)}\right)$, přičemž zjistíme, že reziduální součty čtverců S_r, $r = 0, 1, 2, \dots, t$, vcelku rychle konvergují k nejmenší hodnotě S_t. Vliv váhové funkce posuzujeme pomocí podílu $RI_r = \frac{S_r}{S_0} 100\,\%$, $r = 0, 1, 2, \dots, t$, který nazveme relativní zlepšení reziduálního součtu čtverců.

Vedle bodového odhadu regresních parametrů je samozřejmě důležitý též intervalový odhad (v našem případě pouze přibližný, neboť přesný nelze určit), při jehož stanovení potřebujeme znát odhad rozptylu náhodných odchylek. Ten je pro poslední iteraci dán vzorcem $s_t^2 = \frac{S_t}{n-k-1}$ (a pro Levenbergovu-Marquardtovu nelineární metodu odhadu regresních parametrů vzorcem $s_E^2 = \frac{S_E}{n-k-1}$, viz dále).

Při konstrukci přibližného $100(1-\alpha)\%$ intervalu spolehlivosti pro regresní parametr β_j, $j = 0, 1, \dots, k$, vycházíme ze simulace založené na znalosti odhadu rozptylu náhodných odchylek s_t^2 resp. s_E^2 a znalosti regresního parametru β_j, za nějž pro účely simulace považujeme výsledek nelineární metody. Vzhledem k tomu, že tento interval nemusí být při použití váhové funkce (či nelineární metody) symetrický kolem bodového odhadu, hledáme pomocí simulací zvlášť jeho dolní a horní mez, přičemž se při simulacích používá riziko $\frac{\alpha}{2}$.

3. Ilustrace použití váhové funkce

Na dvou příkladech z biologické praxe ukážeme, jaký vliv má opakované použití váhové funkce na linearizovatelný regresní model.

Příklad 1: Množení bakterií v tekutém prostředí
Byl sledován počet bakterií (závisle proměnná veličina y, v tisících v krychlovém centimetru) v závislosti na čase (nezávisle proměnná veličina x, v hodinách):

x	2,5	2,8	5,4	6,5	9,2	9,5	11	13,3	14,6	16,4
y	3,03	6,213	13,91	19,305	27,037	27,381	49,845	55,069	55,453	75,943

Je známo, že závislost počtu bakterií na čase lze popsat modelem $y = \exp(\beta_0 + \beta_1 x)$. Odhadněte parametry modelu bez váhové funkce a poté s váhovou funkcí (iteračním postupem) a dosažené výsledky porovnejte. Zjistěte, za jak dlouho dojde ke zdvojnásobení počtu bakterií při použití modelu bez váhové funkce a toho modelu s váhovou funkcí, který je nejkvalitnější z hlediska relativního zlepšení reziduálního součtu čtverců.

Řešení: Nejprve získáme výsledky bez použití váhové funkce. Model linearizujeme: $\ln y = \beta_0 + \beta_1 x$. Metodou nejmenších čtverců získáme odhady $b_0 = 1{,}273$ a $b_1 = 0{,}205$ regresních parametrů β_0, β_1. Reziduální součet čtverců je $S_0 = 1272$.

Odhad počtu bakterií v okamžiku $x_0 = 0$ je $N_0 = \exp(b_0) = \exp(1{,}273)$. Ke zdvojnásobení počtu bakterií dojde v okamžiku x_2, tedy

$$2N_0 = N_0 \exp(b_1 x_2) \Rightarrow \ln 2 = b_1 x_2$$

$$\Rightarrow x_2 = \frac{\ln 2}{b_1} = \frac{\ln 2}{0{,}205} = 3{,}38\,\text{h} = 3\,\text{h}\ 23\,\text{min}.$$

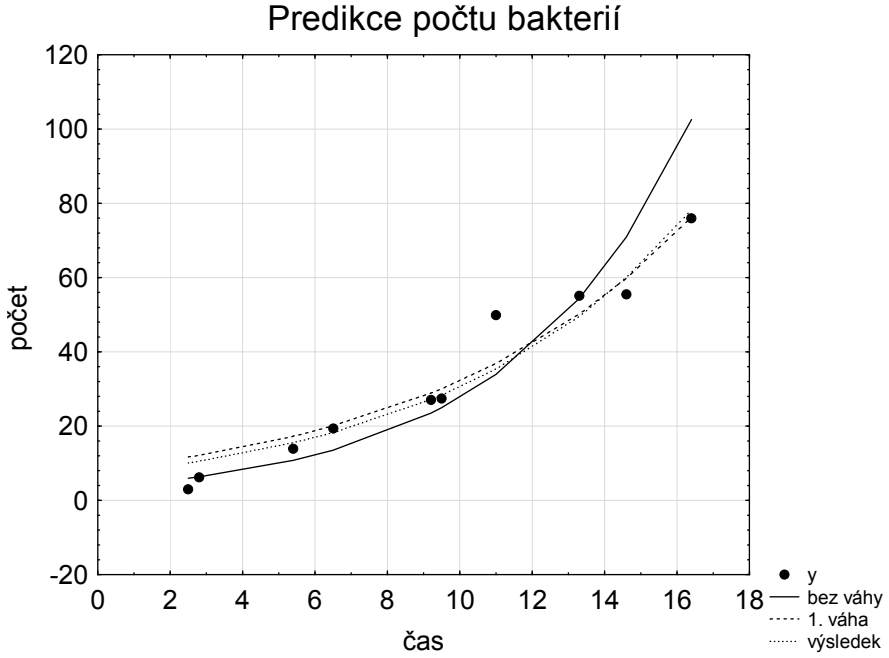
Nyní použijeme váhovou funkci $w(y) = y^2$. Váženou metodou nejmenších čtverců získáme odhady $b_0^{(1)} = 2{,}123$ a $b_1^{(1)} = 0{,}135$ regresních parametrů β_0, β_1. Reziduální součet čtverců je $S_1 = 346{,}375$. Relativní zlepšení reziduálního součtu čtverců činí $RI_1 = \frac{S_1}{S_0} 100\,\% = \frac{346}{1272} 100\,\% = 27,23\,\%$.

Při druhém použití váhové funkce dostaneme $b_0^{(2)} = 1{,}874$, $b_1^{(2)} = 0{,}152$, $S_2 = 352{,}708$, tedy $RI_2 = \frac{S_2}{S_0} 100\,\% = \frac{352{,}708}{1272} 100\,\% = 27{,}74\ \%$. Vidíme, že výsledek je nepatrně horší než při 1. iteraci. 3. iterace poskytne hodnoty: $b_0^{(3)} = 1{,}946$, $b_1^{(3)} = 0{,}147$, $S_3 = 340{,}55$, tedy $RI_3 = \frac{S_3}{S_0} 100\,\% = \frac{340{,}55}{1272} 100\,\% = 26{,}78\,\%$.

Při dalších iteracích už dochází ke zvětšování reziduálního součtu čtverců. 3. iteraci budeme tudíž považovat za konečnou. Ke zdvojnásobení počtu bakterií dojde v okamžiku

$$x_2 = \frac{\ln 2}{b_1^{(3)}} = \frac{\ln 2}{0{,}147} = 4{,}72\,\text{h} = 4\,\text{h}\ 43\,\text{min}.$$

Na obrázku 1 vidíme data s proloženými exponenciálami pro model bez váhové funkce, při prvním použití váhové funkce a při třetím (tj. konečném) použití váhové funkce.

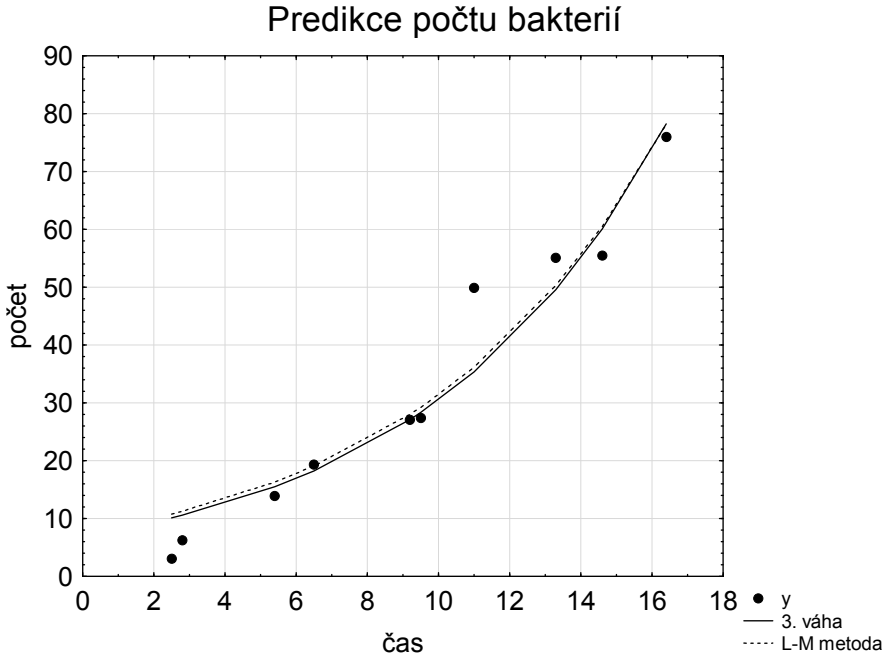


Obrázek 1: Data s proloženými exponenciálami

Pro úplnost ještě odhadneme regresní parametry Levenbergovou-Marquardtovou metodou s počátečními aproximacemi $b_0 = b_1 = 0{,}1$. Po 10 iteracích dostaneme odhady $b_0 = 2{,}020$ a $b_1 = 0{,}143$. Reziduální součet čtverců je $S_E = 335{,}08$, tudíž relativní zlepšení činí $RI = \frac{S_E}{S_0}100\,\% = \frac{335{,}08}{1272}100\,\% = 26{,}34\,\%$. Ke zdvojnásobení počtu bakterií dojde v okamžiku $x_2 = \frac{\ln 2}{b_1} = \frac{\ln 2}{0{,}143} = 4{,}85\,\text{h} = 4\,\text{h}\ 51\,\text{min}$. Obrázek 2 ukazuje výsledky exponenciálního modelu při 3. použití váhové funkce a při Levenbergově-Marquardtově metodě.

Tabulka 1 obsahuje přehledné shrnutí výsledků pro exponenciální model bez použití váhové funkce, s 1., 2. a 3. použitím váhové funkce a pro Levenbergovu-Marquardtovu metodu.

V tabulce 2 jsou uvedeny meze 95% přibližných intervalů spolehlivosti obou regresních parametrů, a to pro 3. použití váhové funkce a pro Levenbergovu-Marquardtovu metodu.



Obrázek 2: Srovnání výsledků při 3. použití váhové funkce a při L-M metodě

Tabulka 1: Shrnutí výsledků pro všechny uvažované modely množení bakterií

Model $y = e^{\beta_0+\beta_1 x}$	Bez váhy	1. váha	2. váha	3. váha	L-M metoda
Odhad b_0	1,273	2,123	1,874	1,946	2,020
Odhad b_1	0,205	0,135	0,152	0,147	0,143
Rez. součet čtverců	1272	346,38	352,71	340,55	335,08
Rel. zlepšení (%)	100	27,24	27,74	26,78	26,34
Odhad N_0	3,57	8,36	6,51	7	7,54
Čas x_2	3h 23min	5h 8min	4h 34min	4h 43min	4h 51min

Tabulka 2: Meze 95% přibližných intervalů spolehlivosti pro regresní parametry

Parametr	3. váha		L-M metoda	
	dolní mez	horní mez	dolní mez	horní mez
β_0	1,238	2,456	1,445	2,490
β_1	0,110	0,194	0,108	0,182

Příklad 2: Hynutí bakterií v tekutém prostředí
V živném roztoku jsou pěstovány bakterie a zjišťován jejich počet. Po přidání antibiotika je již z dřívějších pozorování známo, že pokles počtu bakterií je dán hyperbolickým modelem 2. stupně $y = \frac{1}{\beta_0+\beta_1 x^2}$, kde y je počet bakterií (v tisících v krychlovém centimetru roztoku) za x hodin po přidání antibiotika:

x	0	0,5	0,7	1	1,3	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6
y	51,3	31,74	23	15,99	8	4,81	4,25	2,19	0,25	2,23	0,2	1,19	0,18	0,31	0,25

Odhadněte parametry modelu bez váhové funkce a poté s váhovou funkcí (iteračním postupem) a dosažené výsledky porovnejte. Zjistěte, za jak dlouho dojde k poklesu počtu bakterií na polovinu při použití modelu bez váhové funkce a toho modelu s váhovou funkcí, který je nejkvalitnější z hlediska relativního zlepšení reziduálního součtu čtverců.

Řešení: Nejdříve se budeme věnovat modelu bez použití váhové funkce. Model linearizujeme: $\frac{1}{y} = \beta_0 + \beta_1 x^2$. Metodou nejmenších čtverců získáme odhady $b_0 = 0{,}197$ a $b_1 = 0{,}129$ regresních parametrů β_0, β_1. Reziduální součet čtverců je $S_0 = 3472{,}64$.

Odhad počtu bakterií v okamžiku $x_0 = 0$ je $N_0 = \frac{1}{b_0} = \frac{1}{0{,}129}$. K poklesu počtu bakterií na polovinu dojde v okamžiku x_2, tedy

$$\frac{N_0}{2} = \frac{1}{2b_0} = \frac{1}{b_0 + b_1 x_2^2} \Rightarrow b_0 + b_1 x_2^2 = 2b_0$$

$$\Rightarrow x_2 = \sqrt{\frac{b_0}{b_1}} = \sqrt{\frac{0{,}197}{0{,}129}} = 1{,}24\,\mathrm{h} = 1\mathrm{h}\,14\,\mathrm{min}\,.$$

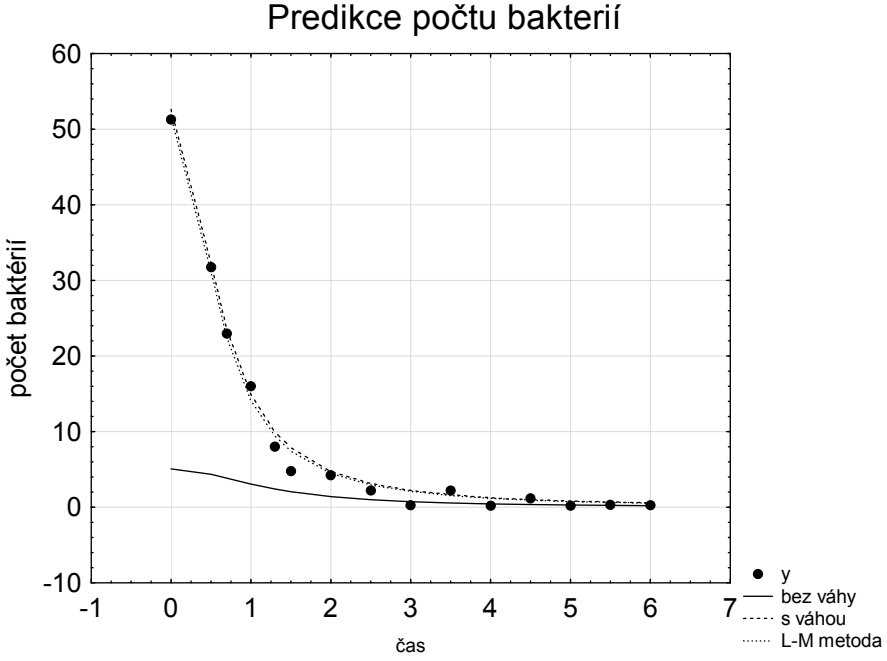
Dále se budeme věnovat modelu, v němž aplikujeme váhovou funkci $w(y) = y^4$. Při prvním použití váhové funkce dostaneme odhady $b_0^{(1)} = 0{,}019$ a $b_1^{(1)} = 0{,}048$ regresních parametrů β_0, β_1.

Reziduální součet čtverců je $S_1 = 21{,}48$. Relativní zlepšení reziduálního součtu čtverců činí $RI_1 = \frac{S_1}{S_0} 100\,\% = \frac{21{,}48}{3472{,}64} 100\,\% = 0{,}62\,\%$.

Při dalších aplikacích váhové funkce již nedochází ke zmenšení reziduálního součtu čtverců. První iterace je tedy hned výsledná. K poklesu počtu bakterií na polovinu dojde v okamžiku $x_2 = \sqrt{\frac{b_0}{b_1}} = \sqrt{\frac{0{,}019}{0{,}048}} = 0{,}63\,\mathrm{h} = 38\,\mathrm{min}$. Odhadneme-li parametry Levenbergovou–Marquardtovou metodou (počáteční aproximace byly $b_0 = b_1 = 0{,}1$, proběhlo 7 iterací), získáme výsledky: $b_0 = 0{,}019$ a $b_1 = 0{,}051$.

Reziduální součet čtverců je $S_E = 19{,}07$, tudíž relativní zlepšení činí $RI = \frac{S_E}{S_0} 100\,\% = \frac{19{,}07}{3472{,}64} 100\,\% = 0{,}57\,\%$. K poklesu počtu bakterií na polovinu dojde v okamžiku $x_2 = \sqrt{\frac{b_0}{b_1}} = \sqrt{\frac{0{,}019}{0{,}051}} = 0{,}61\,\text{h} = 37\,\text{min}$.

Na obrázku 3 jsou znázorněna naměřená data s proloženými hyperbolami 2. stupně pro model bez použití váhové funkce, s prvním (a v tomto případě konečným) použitím váhové funkce a pro Levenbergovu–Marquardtovu metodu.



Obrázek 3: Data s proloženými hyperbolami 2. stupně

Tabulka 3: Shrnutí výsledků pro všechny uvažované modely hynutí bakterií

Model $y = \frac{1}{\beta_0 + \beta_1 x^2}$	Bez váhy	1. váha	L-M metoda
Odhad b_0	1,197	0,019	0,019
Odhad b_1	0,129	0,048	0,051
Reziduální součet čtverců	3472,64	21,48	19,07
Relativní zlepšení (%)	100	0,62	0,57
Odhad N_0	5,08	52,63	52,63
Čas x_2	1 h 14 min	38 min	37 min

Tabulka 4: Meze 95% přibližných intervalů spolehlivosti pro regresní parametry

Parametr	3. váha		L-M metoda	
	dolní mez	horní mez	dolní mez	horní mez
β_0	0,018	0,169	0,018	0,020
β_1	−0,011	0,083	0,045	0,058

V tabulce 3 na předchozí straně vidíme výsledky pro hyperbolický model 2. stupně bez použití váhové funkce, s prvním použitím váhové funkce a pro Levenbergovu-Marquardtovu metodu.

Tabulka 4 obsahuje meze 95% přibližných intervalů spolehlivosti obou regresních parametrů, a to pro 1. použití váhové funkce a pro Levenbergovu--Marquardtovu metodu. Z tabulky 4 je vidět, že 95% přibližné intervaly spolehlivosti parametrů β_0, β_1 pro 1. váhu jsou příliš široké a navíc je dolní mez pro β_1 záporná, což pro větší hodnoty nezávisle proměnné vede k záporným hodnotám počtu bakterií.

Závěr

1. Váhová funkce má v řadě případů velký vliv na odhady parametrů linearizovatelného regresního modelu.
2. První použití váhové funkce nemusí výsledky zlepšit, doporučuje se použití iteračního postupu.
3. Výsledky iteračního postupu pro bodové odhady jsou srovnatelné s výsledky Levenbergovy-Marquardtovy metody, odpadají však problémy s počátečními aproximacemi parametrů a s konvergencí metody.
4. Nevýhodou vážené metody nejmenších čtverců je skutečnost, že není přímo implementována v některých profesionálních statistických systémech, např. ve STATISTICE, kde ji však lze pro některé speciální případy najít v modulu GLM.

Literatura

[1] Maroš B. (2001): *Empirické modely I.* CERM, Brno. 112 stran. ISBN 80-214-1984-9.

[2] Mathcad software. Dostupné na `http://www.dtn.mathsoft.cz/`

[3] Minitab software. Dostupné na `http://www.minitab.com/`

[4] Zvára K. (1989): *Regresní analýza.* 1. vyd. Praha: Academia. 245 stran. ISBN 80-200-0125-5.

CLUSTERING ALGORITHMS BASED ON SAMPLING

Marta Žambochová

Adresa: University of J. E. Purkyně, Faculty of Social and Economic Studies, Department of Mathematics and Informatics, Moskevská 54, Ústí nad Labem

E-mail: marta.zambochova@ujep.cz

Abstrakt: Příspěvek se zabývá popisem a srovnáním vybraných algoritmů pro shlukování ve velkých datových souborech. Společným rysem sledovaných algoritmů je snaha o pokles časové náročnosti shlukování snižováním počtu průchodů celým datovým souborem a dále použitím různých způsobů vzorkování. V některých případech je vzorkování provedeno jednorázovým průchodem datového souboru, při němž jsou konstruovány stromy různých typů (R*-stromy, CF-stromy). Pomocí takto vzniklých stromů je vytvořen soubor reprezentující strukturu původního datového souboru, ovšem s mnohem menším počtem objektů. V jiných případech je vzorek dat vybrán náhodně. Vlastní shlukování již probíhá pouze v rámci tohoto výběrového souboru. V dalším postupu jsou vytvořené shluky modifikovány v rámci jednoho či několika málo průchodů původním souborem.

Klíčová slova: Shlukování, velký soubor dat, čas zpracování, vzorkování, stromy.

Abstract: This paper deals with the description and the comparison of selected algorithms for clustering large datasets. A common feature of the observed algorithms is attempting to decrease the time-consuming process of clustering by reducing the number of passages through the data file and by using different sampling methods. In some cases, sampling is done by means of one-time passage of a data file in which there are constructed trees of various types (R*-trees, CF-trees). By application of the resulting tree, a file representing the structure of the original data file is created, but with a much smaller number of objects. In other cases, the data sample is chosen at random. Custom clustering has already taken place only within this sample file. When further steps are created clusters are modified within one or a few passages through the original file.

Keywords: Clustering, Large Data Set, Processing Time, Sampling, Trees.

1. Introduction

Clustering is an important technique for detecting the data structure. It deals mainly with the similarities of data objects. The set of data objects is divided into several groups (clusters), so that objects within the individual clusters as

well as those of different clusters were as similar as possible. With established clusters, we can better investigate the structure of the data and can then work with each cluster as an individual object.

It is often necessary to process very large data files when analyzing data. The processing of such files is often very difficult, because at first it is a time-consuming process, but also due to the fact that the data file does not fit into memory. The size of the data file is discussed from two different viewpoints. The first viewpoint is the number of variables that we examine for each object. Another viewpoint for determining the size of the data file is the number of data objects. The time-consuming problem is solved by various authors significantly differently, ranging from algorithmically minor, almost cosmetic changes, to some more fundamental changes. Another approach to the process may be variety in the sampling method, in which a representative part is selected from the entire file. This sub-file contains only the amount of objects that can be clustered within a reasonable time. At first, a set of clusters is created in the selected subset of objects, and then assigned to the remaining objects within the already formed clusters.

Another approach is to attempt to minimize the data passed through the file. This article will discuss both of these approaches. They will be described as selected algorithms using trees for sampling, such as CLARANS for large data files (Clustering Large Application based on Randomized Search), or BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). Secondly, two algorithms will be described that do not use the trees for sampling, such as the BIRCH k-means or FEKM (Fast and Exact K-Means).

2. Algorithm CLARANS (for large files)

One of the algorithms for the clustering of large data sets is the algorithm called CLARANS (Clustering Large Application based on Randomized Search) described in [1]. The basis of this algorithm is the method known as k-medoids. The algorithm CLARANS solves the problem of the clustering of a data file for a predetermined number of clusters k and for a defined measure of distance between two objects. The algorithm searches for a set of selected objects (called medoids) so that the average distance for all objects of the investigation files from their closest medoid is minimal. Thus we can maintain that the medoid is the representative object of the cluster whose average distance to all the objects in the cluster is minimal. This means that the medoid represents some sort of "center" of its cluster. Medoids are similar to means or centroids. The output of an algorithm is a set of k clusters that minimizes the sum of the distances of the objects to their nearest medoid.

CLARANS requires two parameters. The first parameter "p1" specifies the number of locally optimal clusterings to be considered and the second parameter "p2" that of the number of exchanges between one medoid and one non-medoid to be tried in order to control the heuristic search strategy. The authors suggest the setting of the value of p1 to 2 and to set the value of p2 to $\max\{1.25 * k(n-k); 250\}$, whereby k is the number of clusters and n is number of objects. The steps of the algorithm are as follows:

1. To create randomly an initial set of k medoids, in order to calculate the average distance of all objects to their closest medoid;
2. to select randomly one of the k medoids and one of the $(n-k)$ non-medoids;
3. to delete the selected medoid from the set of all medoids and replace it with the selected non-medoid object, in order to calculate the difference in the average distance implied by exchanging the two selected objects;
4. if the difference is less than 0, to exchange the selected medoid and non-medoid;
5. if the number of exchanges is less than parameter p2, to move to step 2;
6. to calculate the average distance within the current clustering;
7. if this distance is less than the distance in the best clustering, to specify the current clustering as the best clustering;
8. if the number of selections of the k medoids is less than the parameter p2, move to step 1.

The most problematic part of the algorithm (from the point of view of efficiency for large data sets) is to evaluate the improvement of the average distance. In the algorithm concerning large spatial databases, this problem is solved in that a relatively small number of representatives from the file and process of algorithm CLARANS is applied only to those representatives. This sampling method is used in database systems, which are described by the R*-trees (the term will be explained at a later state). The R*-tree is created out of all the objects within the examined set.

The algorithm selects an appropriate number of representatives (depending on the file size and on the desired number of clusters) by using this tree. Then, the algorithm CLARANS is used to find k medoids from the representative file. Finally, the algorithm assigns all individual data objects to medoids and thereby creates the desired clusters.

R*-trees are a variant of R-trees. Both R-trees and R*-trees are data structures used for the indexing spatial information. The records in the leaf nodes of the tree contain pointers to data objects representing spatial objects. Each node is a multidimensional rectangle that contains all the multidimensional rectangles belonging to the child nodes. Some individual rectangles may overlap. The main difference between R-trees and R*-trees is the algorithm insertion into and deletion of objects from tree structures. Examples of R-tree respective R*-tree are shown in Figures 1, 2 and 3. A more detailed description is in [4] or [6].

The advantages of this algorithm include the fact that it can be used not only for clustering objects for which it is possible to define the averages, but also for objects for which the distance is defined as the degree of similarity between two objects. The advantage is its robustness towards outliers, which is a property of all k-medoids algorithms. A third positive feature of the algorithm is its efficiency in processing the large data sets.

3. Algorithm BIRCH

Another method is the algorithm BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies). This algorithm has been described in [7] and [8]. The data are assigned to clusters when entering into the process. The algorithm produces a CF-tree, into which are gradually entered data classifications. The role of the internal nodes of the CF-tree is to enable the discovery of the correct node for the inclusion of new objects. Each leaf node represents the cluster.

CF-trees (see Figure 4) are highly balanced trees with two parameters. The first parameter is the branching factor (F, L) and the second parameter is the threshold P. Each internal node of the CF-tree contains the number of child nodes maximum F, each leaf node contains no more than L objects, and the range (radius) of each node is less than the threshold P. It is obvious that the size of the tree is a function of the threshold P. It is true that the larger the value of the threshold P, the smaller the size of the tree and vice versa. A similar argument applies to the branching factor. The CF-trees use the so-called Clustering Feature CF. This characteristic is an ordered triple $CF = (N, LS, SS)$, whereby N is the number of data points in the cluster, LS is the vector of sums of the coordinates of all data points in the cluster, and SS is the vector of sums of the squares of the coordinates of data points.

The additivity theorem describes an important feature of this characteristic, which says that the CF-characteristic of a cluster formed by the merger of two disjoint clusters is equal to the sum of the CF-characteristics of the

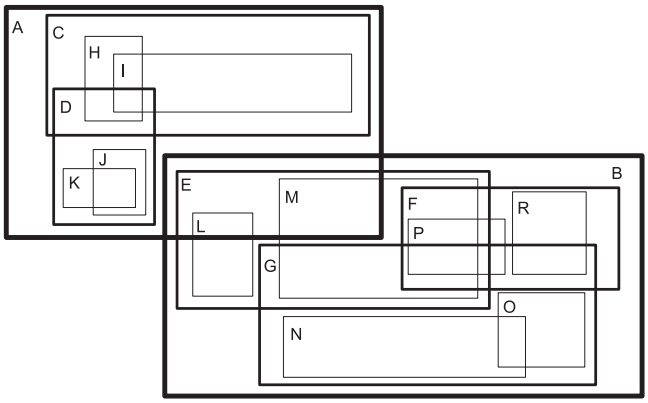


Figure 1: Structure of the R-tree

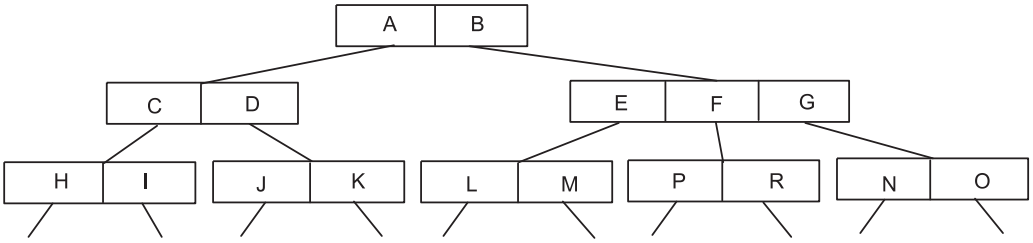


Figure 2: Example of the R-tree

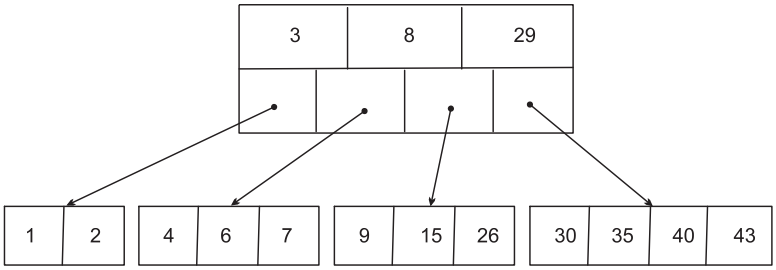


Figure 3: Example of the R*-tree

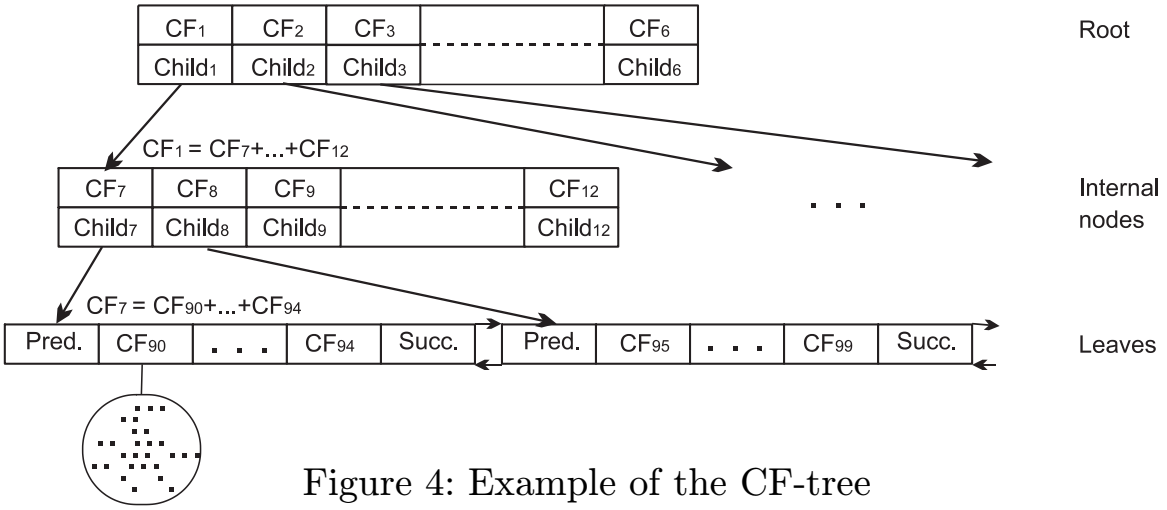


Figure 4: Example of the CF-tree

original clusters. The information contained in the CF-characteristics is sufficient to calculate the centroids, the measure of the distance between two clusters and the degree of the compactness of clusters, which are the characteristics necessary for the formation of the tree.

As already noted, the CF-tree is built dynamically, whereby new objects are gradually inserted. The insertion process begins in the root of the tree. The branch leading to the child node closest to it is found. This will continue until it reaches the leaf. We verify the validity threshold rules for the sub cluster represented in this leaf. If this rule applies, we will include a new data point into the cluster and adjust the relevant CF-characteristics of the leaf and the CF-characteristics of the nodes on the path from the leaf to the root of the CF-tree. If the threshold for the inclusion of a new object does not apply, we have to create a new cluster for this object (with an appropriate leaf). Subsequently, it should recalculate all CF-characteristics (i.e. characteristics pertaining to the nodes on the path from the leaf to the root of the tree). When creating a new node comes into conflict with the branching factor (i.e. the number of branches of some node is greater than F, respectively L), we split the node, again with a recalculation of the CF-characteristics. Similarly, the entire tree is rebuilt.

The creation of the CF-tree is the first phase of the algorithm BIRCH. The second phase of this algorithm condenses the created CF-tree and optimizes its size by adjusting the threshold value P. The third phase of the algorithm BIRCH minimizes the impact of sensitivity on the order of the input data. The algorithm clusters here the leaf nodes using a hierarchical algorithm. The algorithm redistributes the reference points to their closest centres, and obtains a new composition of clusters in the fourth, optional phase. This phase enables the elimination of one of the negative consequences of the CF-tree which occurs when one reference point formally entering at two different times can be assigned to two different clusters. These two instances of the object fall into a common cluster after redistribution. Unfortunately, it is the second phase, in which it is necessary to examine the entire file again, object by object.

4. Algorithm BIRCH *k*-means

The author of the book [5] describes the clustering method, which is a variant of the algorithm k-means influenced by the BIRCH algorithm. The main idea of this algorithm enhancing method is to divide objects into groups prior to processing with the Lloyd's classical algorithm. This pre-processing is advantageous for work at a later date.

The algorithm BIRCH k-means possesses two parameters – the first one is the released number of objects in the cluster and the second is the maximum permitted value of the "cluster's radius" (i.e. limits of variability). This algorithm does not work with CF-trees as opposed to the classical algorithm BIRCH. In addition to this, the CF-characteristic of the classical approach is replaced by another equivalent indication. The information contained in both the characteristics is sufficient to calculate the centre of any cluster, such as the distance between two different clusters and the measure of the compactness of clusters. This is the information, which is necessary and sufficient for the implementation of the algorithm. An important feature of both of the characteristics is the additivity theorem. The main characteristic of the algorithm BIRCH k-means is the ordered triplet (m, q, c), whereby m is the size of the cluster, q the quality of the cluster (calculated as the sum of the squared distances of all objects of a given cluster from the centre of a cluster) and c the centre of the cluster. This characteristic is maintained for each cluster.

The algorithm operates in three phases. A set of objects includes all monitored objects whereby the set of clusters is empty at the beginning of the first phase. Thereafter in the cycle, an object from a set of objects is selected and the algorithm endeavours to find the "nearest" cluster of the set of clusters in which the addition of an object is not limited by the number of elements in a cluster or by variants in the cluster. If such a cluster does not exist, a new cluster is created for this object. This cluster contains only this object. The object is deleted from the set of all objects after the inclusion of the object into the cluster. The cycle is performed until ultimately the set of objects is empty. In the second phase, the algorithm clusters the centres of all clusters that were created in the initial phase. This clustering can be done using any clustering method. The authors have used a special variant of the quadratic method k-means in [5].

The original objects are all classified into clusters in the third phase. Each of the objects is assigned to the nearest centre, which results from the second phase. This phase is the final phase in [5].

In our experiments, however, we added a fourth phase. The division into clusters formed in first three phases was taken to form the initial clustering. The classical Lloyd's algorithm followed. The results of clustering were significantly improved. However, the increase in the computation time is the disadvantage of this modification.

5. Algorithm FEKM

The algorithm FEKM (Fast and Exact K-Means) is described in [2]. It is one of the variants of the algorithm k-means enabling the processing of very large data sets that do not otherwise fit into memory. The efficiency of this variant has been improved by minimizing the number of passes through the data file. The main idea of the algorithm is the initial creation of an appropriately large sample from the original data set. The clusters are created in this sampling set using the classical algorithm k-means. All centres and their descriptive statistics are recorded in each iteration of this clustering of the sample set. All of these centres and their statistics are used to calculate the target clustering. This clustering of a whole set is created with a minimum of passage within the entire data file.

The input of the algorithm is a data file which is necessary for the cluster, the initialization centres and stopping criterion as in the classical algorithm k-means. The output is partitioned into target clusters.

First of all, a random sample data set is created. The algorithm clusters the sample set employing the classical methods k-means in the first phase. All centres and their descriptive statistics are recorded in each iteration. In the second phase, the algorithm goes through the entire data set. For each data object and each iteration from the first phase, a centre is established that is nearest to this object. This assigns the data object into a cluster. The target of the second phase is to identify all objects that are suspected of altering the resulting clustering of the whole file as opposed to a clustering of the sample file. This problem mainly refers to objects that lie on the outskirts of clusters. It is obvious that objects deep inside the clusters are from this perspective, unproblematic. The already identified as suspicious points are saved. If the object is not identified as a suspect peripheral point, we can assume that the object would be assigned by the clustering of the whole set to the same cluster to which it is now assigned as an additional assignment. The characteristics of this cluster are recalculated in this case. In the third phase, the algorithm deals with some suspect peripheral objects which were stored by the previous phase. The algorithm recalculates the centres of all clusters. It works on the assumption that all suspicious objects were assigned to the nearest centres. If there is a recalculated centre, which is more distant from the original centre than the pre-specified critical value (i.e. radius of confidence), the algorithm returns to the second phase and is re-passed to the entire data file.

It is evident that the passage of just one entire data file can only occur in an exceptionally good case. In the worst case scenario, the same number

of passes as in the classical algorithm k-means is essential. The authors recommend that the number of suspected peripheral points should not exceed 20 % of the size of the entire data set. In addition, these points should fit into memory. If a violation of any of these conditions has occurred in the course of the run of the algorithm, it is necessary to reduce the appropriate parameter algorithm, and hence to establish a different radius of confidence, which will inevitably result in a reduction in the number of suspect peripheral points. The authors have demonstrated the accuracy and effectiveness of the proposed algorithm in their paper [3].

6. The results of the experiments

Individual algorithms were programmed in the MATLAB system. The behaviour of the algorithms was verified in several data sets. The IRIS file with real data (see [9]) is one of them. Generated data sets were also employed.

The significant disadvantage of most of the algorithms was the dependency on the input parameters. In addition, there is no recommendation for a suitable choice of these parameters in some algorithms (notably both algorithms BIRCH). They can be estimated only with a thorough knowledge of the structure of the processed data, which is contrary to the main target of the clustering, i.e., the identification of the data structure. The algorithm CLARANS achieved the best results in this regard. Its authors have proposed some recommended parameter values. Moreover, small deviations in the values of the parameters do not affect the results of clustering very much.

Another disadvantage of the majority of these algorithms is that it provides worse quality sampling clustering. The quality of clustering of both algorithms BIRCH after proceeding through all the phases described by the authors of the algorithms was by far the worst. The quality of clustering can be (easily) improved by adding of a phase in which we would run a few iterations of the classical algorithm k-means. We found that two to four iterations were sufficient to add, after which the resulting clustering quality improved significantly. It was outperforming many other comparative algorithms. Unfortunately, this quality improvement is at the expense of the processing time. The resulting quality of the clustering algorithm FEKM was highly dependent on the generated sample file (from very poor to excellent quality). The CLARANS algorithm achieved average quality in most attempts. The quality of the resulting clustering in this case was obviously better if higher parameters of the algorithm were chosen. The algorithm FEKM has revealed one more particular disadvantage. Its authors declared a small number of passages through the whole file required to run the algorithm. This claim

was confirmed only in exceptional cases. The number of passages is strongly dependent on the initial selection of sample data. Only the algorithm FEKM creates sample data randomly. Other referred to algorithms that create samples with the help of certain data structures (trees). The resulting sample thus gives a better picture of the distribution thereof. It seems that the algorithm FEKM would be appropriate to combine with any of the other referred to algorithms to achieve better results. This might be a subject for some further investigation.

References

[1] Ester, M.; Kriegel, H.-P.; Xu, X. (1995): A database interface for clustering in large spatial databases. *Proceedings of 1st International Conference on Knowledge Discovery and Data Mining.*

[2] Goswami, A.; Ruoming, J.; Agrawal, G. (2004): Fast and exact out-of-core k-means clustering Data Mining. *ICDM'04. Fourth IEEE International Conference on Volume.*

[3] Goswami, A.; Ruoming, J.; Agrawal, G. (2006): Fast and exact out-of-core and distributed k-means clustering. *Knowledge and Information Systems* **10**, 17–40.

[4] Guttman, A.; Stonebraker, M. (1983): A Dynamic Index Structure for Spatial Searching. *EECS Department University of California, Berkeley,* No. UCB/ERL M83/64.

[5] Kogan, J. (2007): *Introduction to Clustering Large and High-Dimensional Data.* Cambridge University Press, New York.

[6] Pokorný, J. (2000): Prostorové datové struktury a jejich použití k indexaci prostorových objektů. *GIS Ostrava 2000,* 146–160.

[7] Zhang, T.; Ramakrishnan, R.; Livny, M. (1996): An efficient data clustering method for very large databases. *ACM SIGMOD Record* **25**, 103–114.

[8] Zhang, T.; Ramakrishnan, R.; Livny, M. (1997): BIRCH: A new data clustering algorithms and its applications. *Journal of Data Mining and Knowledge Discovery* **1**, 141–182.

[9] `http://archive.ics.uci.edu/ml/datasets/`

NEŠŤASTNÁ *P*-HODNOTA

Karel Zvára

Adresa: KPMS MFF UK, Sokolovská 83, 186 00, Praha 8

E-mail: zvara@karlin.mff.cuni.cz

Abstrakt: Článek se pokouší reagovat na nesprávnou interpretaci p-hodnoty jako na pravděpodobnost nulové hypotézy. Ukazuje, že podobnou interpretaci nedá ani bayesovský model.

Abstract: The paper is a respond to the incorrect interpretation of p-value as the probability of the null hypothesis. It shows that even a simple Bayesian model does not lead to a similar interpretation.

1. Úvod

Nedávno se mi dostala do rukou kniha Flegr [4]. Věnoval mi ji sám autor, profesor biologie. V knize je velice často řeč o statistice. Jsou tu užitečné postřehy o použití statistiky, jsou tu bohužel také některé mírně řečeno nepřesnosti. Věřím, že v případném dalším vydání bude jejich výskyt řidší, nebudou-li dokonce zcela vymýceny. Tím hlavním bludem je opakovaná interpretace p-hodnoty jako pravděpodobnosti pravdivosti nulové hypotézy. Diskuse, která se následně rozvinula mezi mojí maličkostí a objevitelem podivuhodných následků toxoplasmózy, mne přivedla k napsání následujících řádků.

Nejsem jistě první, kdo na ten problém narazil. Jak interpretovat pojem, kterému kolega Havránek [5, str. 104] říkal *dosažená hladina významnosti*, který ale téměř všichni označují jako p-hodnotu. Podobné tříslovné označení (dosažená hladina testu) lze najít i v moderní učebnici Anděl [1, str. 72], v níž jsou uvedeny také anglické ekvivalenty P-value a significance level.

Podle přehledového článku [3] se v anglicky psané literatuře tento pojem vyskytl prvně jako P value v knize [2]. Dalšími používanými anglickými označeními jsou podle zmíněného článku probability level, sample level of significance, observed significance level, significance probability, descriptive level of significance, critical level, significance level, prob-value, associated probability. O p-hodnotě, především o interpretaci tohoto pojmu, existuje rozsáhlá literatura, kterou jsem samozřejmě nedokázal celou prostudovat. Nicméně uvedu aspoň něco málo.

Nechci probírat rozdíl v chápání statistického rozhodování mezi R. A. Fisherem a pojetím Neymana a Pearsona, natož abych šel do hloubky v případných bayesovských úvahách. Sám ve své práci a zejména v tom, jak se

snažím něco naroubovat do hlav studentů, se přidržuji spíše Neymannova a Pearsonova statistického rozhodování.

2. Jednoduchá představa

V nejjednodušším případě máme dvojici hypotéz (nulovou a alternativní) charakterizovaných dvojicí hodnot θ_0 a θ_1. Rozhodujeme pomocí nějaké statistiky T se spojitým rozdělením, která podle toho, která z hypotéz platí, má rozdělení s distribuční funkcí $F_j(t)$, resp. s hustotou $f_j(t), j = 0, 1$. Pro danou hladinu významnosti α je kritický obor dán nerovností $T \geq t_\alpha$, přičemž kritickou hodnotu t_α určuje požadavek $\mathsf{P}(T \leq t_\alpha | H_0) = F_0(t_\alpha) = 1 - \alpha$. Alternativní vyjádření kritického oboru je dáno nerovností $p \leq \alpha$, kde $p = p(t) = \mathsf{P}(T \geq t | H_0) = 1 - F_0(t)$ a t je hodnota statistiky T realizovaná v daném pokusu.

Pravděpodobnost jsem tu zapsal poněkud nedůsledně jako podmíněnou pravděpodobnost, přestože nulovou hypotézu, resp. její pravdivost za náhodný jev nepovažuji. Jak známo (viz například Anděl [1, Věta 1.4]), za platnosti nulové hypotézy má p-hodnota rovnoměrné rozdělení $\mathsf{R}(0, 1)$. Proto lze kritický obor zapsat jako $W = \{t : p(t) \leq \alpha\}$ a p-hodnotu není třeba nějak interpretovat.

Jenže v aplikacích se bez interpretace neobejdeme. Vysvětlení, že p-hodnota je vlastně za platnosti nulové hypotézy spočítaná pravděpodobnost toho, že pokus mohl dopadnout pro nulovou hypotézu ještě hůř, připadá zákazníkům poněkud komplikované. Samotná p-hodnota přímo vypovídá o výsledku pokusu, resp. o vztahu výsledku pokusu k nulové hypotéze a nikoliv o samotné hypotéze, která zajímá zákazníka mnohem víc.

Pokusme se nad představou, že p-hodnota je pravděpodobnost nulové hypotézy, trochu zamyslet. Abychom mohli o takové pravděpodobnosti něco tvrdit, musíme připustit, že nulová hypotéza je náhodný jev. Do jisté míry bude inspirací článek Sellke et al. [7], zejména jeho původní verze Sellke et al. [6]. Ve zmiňované původní verzi se autoři pomocí simulací snaží odhadnout podmíněnou pravděpodobnost $\mathsf{P}(H_0 | p \approx \alpha)$, kde přibližnost chápou v případě $\alpha = 0{,}05$ tak, že má být $p \in (0{,}0455; 0{,}0500)$.

3. Aposteriorní pravděpodobnost nulové hypotézy

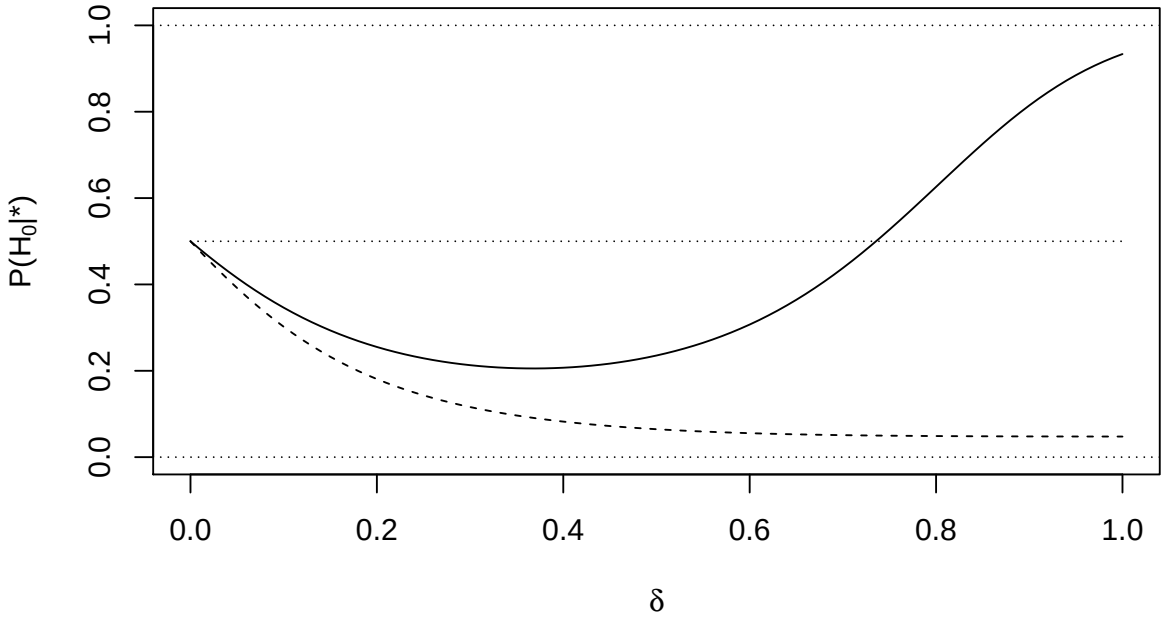
Jednou ze situací uvažovaných v Sellke et al. [6] je model, kdy se předpokládá pouze jednoduchá alternativní hypotéza. V následujícím se pokusíme bez simulací o výpočet aposteriorní pravděpodobnosti nulové hypotézy za podmínky, že p-hodnota je blízká číslu α.

Nechť $O(\alpha) = (\alpha - \varepsilon, \alpha + \varepsilon)$ je malé okolí hladiny α. Označme apriorní pravděpodobnost nulové hypotézy jako $\pi_0 = \mathsf{P}(H_0)$. Pak pro $H_j, j = 0, 1$, a odpovídající malé ε^* můžeme psát

$$\begin{aligned}\mathsf{P}\left(p \in O(\alpha)|H_j\right) &= \mathsf{P}(t_\alpha - \varepsilon^* < t < t_\alpha + \varepsilon^*|H_j) \\ &\doteq 2\varepsilon^* f_j(t_\alpha).\end{aligned}$$

Použijeme-li nyní Bayesův vzorec, dostaneme

$$\begin{aligned}\mathsf{P}\left(H_0|p \in O(\alpha)\right) &= \frac{\mathsf{P}(p \in O(\alpha)|H_0)\,\mathsf{P}(H_0)}{\mathsf{P}(p \in O(\alpha)|H_0)\,\mathsf{P}(H_0) + \mathsf{P}(p \in O(\alpha)|H_1)\,\mathsf{P}(H_1)} \\ &\doteq \frac{2\pi_0\varepsilon^* f_0(t_\alpha)}{2\pi_0\varepsilon^* f_0(t_\alpha) + 2(1-\pi_0)\varepsilon^* f_1(t_\alpha)} \\ &= \frac{f_0(t_\alpha)\pi_0}{f_0(t_\alpha)\pi_0 + f_1(t_\alpha)(1-\pi_0)}. \qquad (1)\end{aligned}$$



Obrázek 1: Závislost $\mathsf{P}\left(H_0|p \in O(\alpha)\right)$ (—), resp. $\mathsf{P}(H_0|W)$(− − −) na alternativě δ při $\pi_0 = 0{,}5$

V citovaném článku Sellke et al. [7] autoři zdůrazňují, že v žádném případě neklesla pravděpodobnost nulové hypotézy při $\alpha \doteq 0{,}05$ pod 22 %. Tomu odpovídá i následující příklad.

Uvažujme podobně jako autoři konkrétní situaci, kdy na základě n nezávislých náhodných veličin $X_1, \ldots, X_n$ s rozdělením $\mathsf{N}\left(\theta, \sigma^2\right)$ testujeme hypotézu $\mathsf{H}_0 : \theta = 0$ proti jednostranné alternativě $\mathsf{H}_1 : \theta = \delta$, kde je $\delta > 0$. Ve zmíněném článku pracovali autoři s alternativou oboustrannou. Rozptyl $\sigma^2 > 0$ je

známá konstanta. Statistika $T = \sqrt{n}\,\overline{X}/\sigma$ založená na výběrovém průměru $\overline{X}$ má rozdělení $\mathsf{N}\big(\sqrt{n}\theta/\sigma, 1\big)$. Proto je v naší konkrétní situaci $t_\alpha = \Phi^{-1}(1-\alpha)$, kde $\Phi^{-1}(u)$ je kvantilová funkce $\mathsf{N}(0,1)$.

Nyní vyšetříme, při které střední hodnotě za alternativy bude podmíněná pravděpodobnost $\mathsf{P}\big(H_0|p \in O(\alpha)\big)$ minimální. Z (1) plyne, že to bude pro takové δ, pro které je hustota statistiky T v tomto kritickém bodě za platnosti alternativy, totiž $f_1\big(\Phi^{-1}(1-\alpha)\big)$, jako funkce δ maximální. K tomu musí být střední hodnota T rovna kritické hodnotě $\Phi^{-1}(1-\alpha)$. Podmíněná pravděpodobnost $\mathsf{P}\big(H_0|p \in O(\alpha)\big)$ bude tedy minimální pro $\delta = \Phi^{-1}(1-\alpha)\sigma/\sqrt{n}$. Při $\sigma = 1$ a $n = 20$ je to pro $\delta \doteq 0{,}368$ a minimum $\mathsf{P}\big(H_0|p \in O(\alpha)\big)$ je pak rovno

$$\frac{\varphi(t_\alpha)\pi_0}{\varphi(t_\alpha)\pi_0 + \varphi(0)(1-\pi_0)}, \tag{2}$$

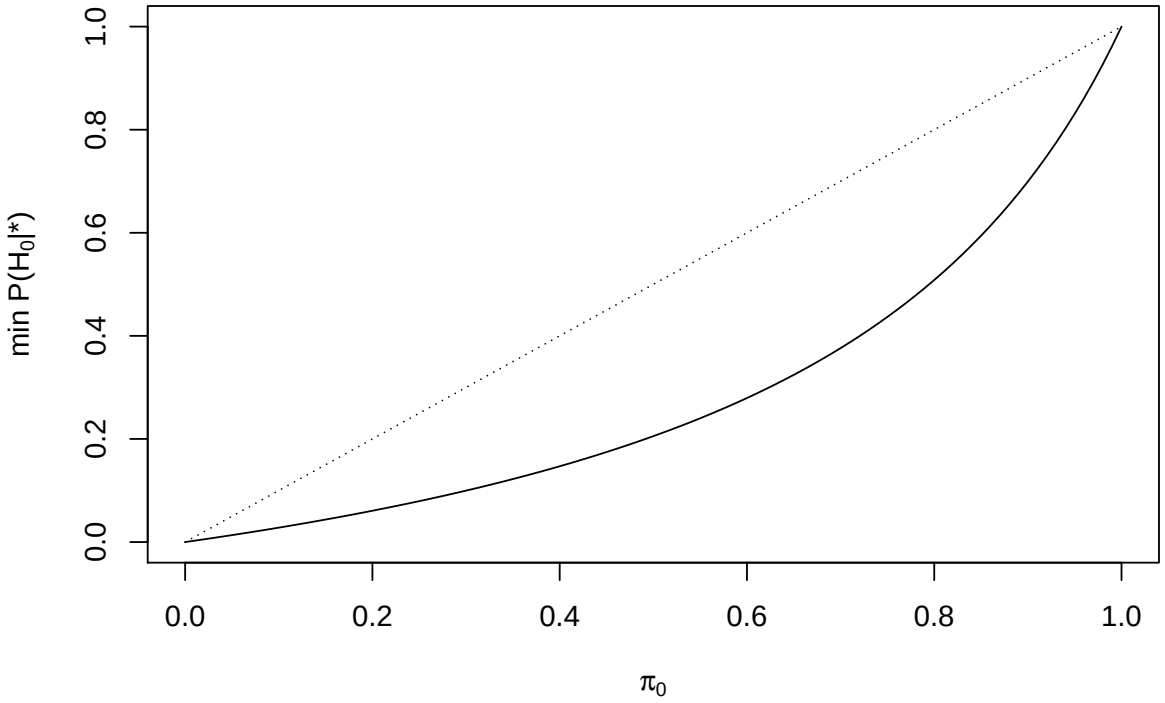
kde $\varphi(t)$ je hustota $\mathsf{N}(0,1)$. Při $\pi_0 = 0{,}5$ je minimum rovno zhruba hodnotě 0,205, jak je také patrné z obrázku 1. Došli jsme tak k podobné hodnotě jako autoři Sellke et al. [7], u nich v simulačním pokusu při oboustranné alternativě a poněkud jinak určené podmínce vycházela dolní mez 0,23.

Uvedené hodnoty ovšem silně závisí na apriorní pravděpodobnosti nulové hypotézy. Závislost na π_0 pro minimální aposteriorní pravděpodobnost nulové hypotézy za podmínky, že $p \in O(\alpha)$, je patrná z obrázku 2.

4. Jiná podmínka

Pokusme se o poněkud jinou úvahu. Místo podmínky $p \in O(\alpha)$ použijme podmínku $p \in W$, kde kritický obor W je určen nerovností $p \le \alpha$. Nulové hypotéze přiřadíme nějakou apriorní pravděpodobnost π_0 a budeme hledat aposteriorní pravděpodobnost nulové hypotézy za podmínky, že jsme nulovou hypotézu na zvolené hladině zamítli. Opět budeme předpokládat, že $\theta_0 = 0$ a $\theta_1 = \delta$. Hladinu významnosti označíme obligátním α, sílu testu, která závisí na δ, jako $1 - \beta(\delta)$.

$$\begin{aligned}
\mathsf{P}(H_0|W) &= \frac{\mathsf{P}(W|H_0)\,\mathsf{P}(H_0)}{\mathsf{P}(W|H_0)\,\mathsf{P}(H_0) + \mathsf{P}(W|H_1)\,\mathsf{P}(H_0)} \\
&= \frac{\alpha\pi_0}{\alpha\pi_0 + \big(1-\beta(\delta)\big)(1-\pi_0)} \\
&= \frac{\alpha}{\alpha + \big(1-\beta(\delta)\big)(1-\pi_0)/\pi_0}.
\end{aligned} \tag{3}$$



Obrázek 2: Závislost minimální hodnoty
$\mathsf{P}\left(H_0|p \in O(\alpha)\right)$ na apriorní pravděpodobnosti π_0

Ukázku toho, jak aposteriorní pravděpodobnost $\mathsf{P}(H_0|W)$ závisí na alternativě $\theta = \delta$, najdeme na obrázku 1. Protože je v našem příkladu síla testu rostoucí funkcí δ, aposteriorní pravděpodobnost $\mathsf{P}(H_0|W)$ je klesající funkcí parametru δ. Protože je $\lim_{\delta\to\infty}\left(1-\beta(\delta)\right) = 1$, zmíněná aposteriorní pravděpodobnost konverguje k hodnotě

$$\frac{\alpha}{\alpha + (1-\pi_0)/\pi_0}.$$

Pro $\pi_0 = 0{,}5$ a $\alpha = 0{,}05$ je touto dolní mezí hodnota nepatrně menší, než α, totiž zhruba 0,0476.

Vraťme se k obecné apriorní pravděpodobnosti π_0. Pro každý nestranný test, což je i náš případ, platí $1-\beta(\delta) \geq \alpha$. Když použijeme tuto nerovnost ve vztahu (3), dostaneme omezení $\mathsf{P}(H_0|W) \leq \pi_0$, takže aposteriorní pravděpodobnost, že platí nulová hypotéza, v případě, že jsme hypotézu zamítli, nemůže být větší, než její apriorní pravděpodobnost. Na druhé straně je vždy $1-\beta(\delta) \leq 1$ a velice hrubě také $\alpha < 1$, takže hrubým dolním odhadem pro $\mathsf{P}(H_0|W)$ je $\alpha\pi_0$.

5. Závěr?

Co mám poradit svým zákazníkům, jak označoval Josef Machek „badáky“, s nimiž statistiku provozujeme? Kde to je možné, raději pracovat s intervalem spolehlivosti a zjednodušenému rozhodování se vyhnout. Ne vždy to je možné nebo aspoň snadné. Ne vždy je k disposici jednorozměrný parametr, jehož hodnota by určovala nulovou hypotézu. Už třeba jednoduchá hypotéza chí-kvadrát testu dobré shody určuje celý vektor pravděpodobností. Tam bych použití p-hodnoty omezil na jednoduchý popis kritického oboru.

Platí-li $p \leq \alpha$, nulovou hypotézu zamítni, protože data se příliš liší od hypotézou očekávaných hodnot. Aspoň tak špatná data mohla za platnosti testované hypotézy vyjít jen s pravděpodobností p. Jen si nejsem jist, nakolik pochodím. Spíš bych měl být zvědav, s jakou tvořivou interpretací se setkám příště.

Literatura

[1] J. Anděl (2005). *Základy matematické statistiky.* Matfyzpress, Praha.

[2] K. A. Brownlee (1960). *Statistical Theory and Methodology in Science and Engineering.* Wiley.

[3] H. A. David (1995). First (?) occurence of common terms in mathematical statistics. *The American Statistician*, 49, 121–133.

[4] J. Flegr (2011). *Pozor Toxo (Tajná učebnice praktické metodologie vědy).* Academia, Praha.

[5] T. Havránek (1993). *Statistika pro biologické a lékařské vědy.* Academia, Praha.

[6] Thomas Sellke, M. J. Bayarri, James O. Berger (1999). Calibration of P-values for testing precise null hypotheses. Technical Report 1999-38, Duke University. Dostupné na webové stránce:
`http://www.stat.duke.edu/~berger/papers/99-13.html`

[7] Thomas Sellke, M. J. Bayarri, James O. Berger (2001). Calibration of P-values for testing precise null hypotheses. *The American Statistician*, 55, 62–71.

ZAJÍMAVÉ USPOŘÁDÁNÍ POKUSU

Karel Zvára

Adresa: KPMS MFF UK, Sokolovská 83, 186 00, Praha 8

E-mail: zvara@karlin.mff.cuni.cz

Abstrakt: Článek je věnován hledání plánu pokusu v úloze z praxe, která vede na model s pevnými i náhodnými efekty.

Abstract: The article is devoted to searching the design of experiment in practical task, which is described by a model with fixed and random effects.

1. Zadání úlohy

Cílem pokusu je zjistit reakci na různé koncentrace jisté látky. K dispozici je celkem $M = 20$ různých koncentrací růstového regulátoru benzyladeninu BA, každá v $P = 5$ různých nádobkách. Vedle toho máme po $P = 5$ vzorcích (řízcích) pořízených vždy na jednom z $M = 20$ různých míst lodyhy. Je třeba navrhnout způsob jak rozmístit vzorky po jednom do každé nádobky tak, aby test prokazující závislost nějaké měřené vlastnosti vzorku na koncentraci měl maximální možnou sílu. Je třeba přihlédnout k tomu, že chování vzorku může ovlivnit nejen koncentrace v dané nádobce, ale také místo na lodyze, z něhož byl vzorek odebrán.

Lze předpokládat, že v pětici nádobek se stejnou koncentrací opravdu je stejná a **známá** experimentátorem zvolená koncentrace. Půjde tedy o **pevný efekt**.

Jinak je to s příslušností vzorku k místu na lodyze. Vzorky z jednoho místa mají stejné předpoklady k hodnotě cílové veličiny, vzorky z různých míst mohou mít tyto předpoklady nestejné, což je dáno nejen například jinou vzdáleností od kořene (opět pevný efekt), ale také přirozenou biologickou variabilitou vlastností lodyhy. Tento poslední zdroj variability experimentátor neurčuje, dokonce jeho vliv ani nezná, nastavuje jej příroda. Proto příslušnost k místu na lodyze budeme chápat jako realizaci **náhodného efektu**.

Dosavadním úvahám odpovídá představa, že výsledná hodnota odezvy je součtem deterministické závislosti na koncentraci látky v nádobce, deterministické závislosti na umístění vzorku na lodyze a náhodné složky modelující vliv biologické variability. Půjde tedy o **smíšený lineární model**.

Dříve, než se pokusíme zapsat uvedené předpoklady do modelu, popišme ještě možné rozmístění vzorků do nádobek. Zaveďme matici $\mathbf{K}$, jejíž ij-tý prvek k_{ij} bude udávat místo odebrání vzorku, který dáme do j-té nádobky s i-tou koncentrací. Matice $\mathbf{K}$ má tedy M řádků a P sloupců.

Uvedené předpoklady vedou k modelu ($j = 1, \dots, P,\ i = 1, \dots, M$)

$$Y_{ij} = \mu + \beta \cdot x_i + \gamma \cdot v_{k_{ij}} + A_{k_{ij}} + E_{ij}, \tag{1}$$

$$= \mu^* + \beta(x_i - \overline{x}) + \gamma(v_{k_{ij}} - \overline{v}) + A_{k_{ij}} + E_{ij}, \tag{2}$$

kde β je pevný koeficient určující závislost na známé koncentraci x_i, γ je podobný koeficient popisující závislost na tom, odkud byl pořízen vzorek, v_k je známá konstanta, která zachycuje vliv umístění k-tého odběrového místa (např. vzdálenost místa od kořene), $A_k \sim \mathsf{N}\big(0, \sigma_A^2\big)$ je náhodný efekt k-tého místa a $E_{ij} \sim \mathsf{N}\big(0, \sigma^2\big)$ je náhodná složka zahrnující vedle přirozené biologické variability všechny zatím nesledované vlivy. O náhodných veličinách E_{ij} i A_k předpokládáme, že jsou vesměs nezávislé. Symbolem Y_{ij} je označena hodnota vysvětlované veličiny změřená na vzorku umístěném v j-tém exempláři i-té koncentrace. Všude dál budeme předpokládat, že každý z regresorů X, V nabývá aspoň dvou různých hodnot.

Klíčová je volba matice permutací $\mathbf{K}$. Různé matice se budeme snažit porovnat podle přesnosti, s jakou lze odhadnout parametr β, který je klíčový při prokazování závislosti na koncentracích x_i. V našem smíšeném modelu je rozptyl odhadu parametrů u pevných efektů poměrně složitou funkcí i v situaci, kdy známe varianční matice všech zúčastněných náhodných veličin.

Zaveďme vektor $\mathbf{Y}$ o $M \cdot P$ složkách a matici $\mathbf{X}$ typu $M \cdot P \times 2$

$$\mathbf{Y} = \begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1P} \\ \vdots \\ Y_{M1} \\ \vdots \\ Y_{MP} \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_1 - \overline{x} & v_{k_{11}} - \overline{v} \\ x_1 - \overline{x} & v_{k_{12}} - \overline{v} \\ \vdots & \vdots \\ x_1 - \overline{x} & v_{k_{1P}} - \overline{v} \\ \vdots & \vdots \\ x_M - \overline{x} & v_{k_{M1}} - \overline{v} \\ \vdots & \vdots \\ x_M - \overline{x} & v_{k_{MP}} - \overline{v} \end{pmatrix}$$

2. Model bez náhodných efektů

Pro začátek předpokládejme, že je $\sigma_A^2 = 0$, takže můžeme náhodné efekty A_k pominout. V takovém případě jsou náhodné veličiny $Y_{11}, \dots, Y_{MP}$ nezávislé a odhad $\widehat{\boldsymbol{\beta}}$ vektoru $\boldsymbol{\beta} = (\beta, \gamma)^T$ má varianční matici $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$. Pokusíme se najít závislost rozptylu odhadu parametru β na matici $\mathbf{K}$. K tomu stačí zjistit chování prvku matice $(\mathbf{X}^T\mathbf{X})^{-1}$ na místě $(1,1)$.

Matice $\mathbf{X}^T\mathbf{X}$ je nutně regulární a má tvar

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} P\sum_{i=1}^{M}(x_i - \overline{x})^2 & c(\mathbf{K}) \\ c(\mathbf{K}) & P\sum_{i=1}^{M}(v_i - \overline{v})^2 \end{pmatrix},$$

kde jsme pro mimodiagonální prvky zavedli označení

$$c(\mathbf{K}) = \sum_{i=1}^{M}\sum_{j=1}^{P}(x_i - \overline{x})(v_{k_{ij}} - \overline{v}) \tag{3}$$
$$= \sum_{i=1}^{M}(x_i - \overline{x})\sum_{j=1}^{P}(v_{k_{ij}} - \overline{v}).$$

Označíme-li dále průměr hodnot veličiny V při i-té koncentraci

$$q_i = \frac{1}{P}\sum_{j=1}^{P} v_{k_{ij}}, \qquad (\text{zřejmě je } \overline{q} = \overline{v})$$

můžeme psát také

$$c(\mathbf{K}) = P\sum_{i=1}^{M}(x_i - \overline{x})(q_i - \overline{q}). \tag{4}$$

Prvek inverzní matice, který nás zajímá, dostaneme jako podíl dvou výrazů. V čitateli je prvek na místě (2,2) matice $\mathbf{X}^T\mathbf{X}$, který ale nijak nezávisí na $c(\mathbf{K})$, tedy ani na matici $\mathbf{K}$. Ve jmenovateli je determinant matice $\mathbf{X}^T\mathbf{X}$, který je roven

$$\det(\mathbf{X}^T\mathbf{X}) = P^2\sum_{i=1}^{M}(x_i - \overline{x})^2\sum_{i=1}^{M}(q_i - \overline{q})^2 - c(\mathbf{K})^2.$$

Minimalizace rozptylu odhadu parametru β přes různé matice $\mathbf{K}$ je tedy ekvivalentní s maximalizací determinantu $\det(\mathbf{X}^T\mathbf{X})$, což je dále ekvivalentní s minimalizací absolutní hodnoty $c(\mathbf{K})$. Z vyjádření v (4) je patrné, že absolutní minimum dostaneme, pokud vektor průměrů $\mathbf{q} = (q_1, q_2, \ldots, q_M)^T$ má všechny prvky stejné nebo pokud je výběrová kovariance vektorů $\mathbf{x}$ a $\mathbf{q}$ nulová. Nemůže-li být $c(\mathbf{K})$ rovno nule, budeme hledat matici $\mathbf{K}$ takovou, aby byl výraz $c(\mathbf{K})$ aspoň blízký nule. Nepřehlédněme přitom, že nulovost kovariance vektorů $\mathbf{x}$ a $\mathbf{q}$ je totožná s nulovostí výběrové kovariance (resp. ortogonalitou) dvou sloupců matice $\mathbf{X}$.

Uvažme speciální případ, kdy je $v_k = k$ pro všechna k. Pak je zřejmě $\overline{q} = \overline{v} = (M+1)/2$, takže podle (3) můžeme psát

$$c(\mathbf{K}) = \sum_{i=1}^{M}(x_i - \overline{x})\left(\sum_{i=1}^{P} k_{ij} - \frac{P(M+1)}{2}\right).$$

Součet $\sum_{i=1}^{P} k_{ij}$ je samozřejmě vždy celočíselný, kdežto výraz $P(M+1)/2$ pro M sudé a P liché celočíselný není. V takovém případě odpadá první možnost dosažení nulové hodnoty $A(\mathbf{K})$ (všechny průměry q_i jsou stejné).

3. Model s náhodnými efekty

Abychom mohli úlohu zapsat pomocí běžně používaného označení, potřebujeme ještě matici $\mathbf{Z}$, která popisuje vliv náhodných efektů na jednotlivé složky vektoru $\mathbf{Y}$. Počet řádků je dán počtem složek $\mathbf{Y}$, tedy $M \cdot P$, počet sloupců je dán počtem náhodných efektů A_k, těch je M. Matici $\mathbf{Z}$ dostaneme z matice $\mathbf{K}$ tak, že do řádku, který odpovídá Y_{ij} vložíme do sloupce k_{ij} jedničku, všude jinde budou nuly. Řádek matice $\mathbf{Z}$ tedy identifikuje nádobku a sloupec identifikuje místo, odkud pochází vzorek do nádobky vložený. V každém řádku bude takto jediná jednička, v každém sloupci bude celkem P jedniček, neboť jedna realizace náhodného efektu A_k ovlivní P vzorků pořízených ze stejného místa. Model pak můžeme zapsat standardně jako

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{A} + \mathbf{E}, \tag{5}$$

kde je $\boldsymbol{\beta} = (\beta, \gamma)^T$, $\mathbf{A} = (A_1, \ldots, A_M)^T$ a vektor $\mathbf{E}$ vznikne z E_{ij} stejným uspořádáním, jako vznikl vektor $\mathbf{Y}$. Předpokládáme, že je $\mathbf{A} \sim \mathsf{N}\big(\mathbf{0}, \sigma_A^2 \mathbf{I}_M\big)$ a $\mathbf{E} \sim \mathsf{N}\big(\mathbf{0}, \sigma^2 \mathbf{I}_{M \cdot P}\big)$, přičemž náhodné vektory $\mathbf{A}, \mathbf{E}$ jsou nezávislé. Uvedené předpoklady vedou k tomu, že střední hodnota vektoru $\mathbf{Y}$ je rovna $\mathbf{X}\boldsymbol{\beta}$ a jeho varianční matice je $\mathbf{V} = \sigma_A^2 \mathbf{Z}\mathbf{Z}^T + \sigma^2 \mathbf{I}_{M \cdot P}$. Platí tedy $\mathbf{Y} \sim \mathsf{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. Maximálně věrohodný odhad vektoru $\boldsymbol{\beta}$ je tedy

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{Y}, \tag{6}$$

přičemž platí $\mathsf{var}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$. Zdůrazňuji, že uvedené vztahy platí v případě, že oba rozptyly σ^2, σ_A^2 skryté v matici $\mathbf{V}$ známe. O hodnotách rozptylů můžeme mít představu z předchozích nebo přípravných pokusů, případně z literatury. Bohužel, v mém případě jsem se nedostal dál, než k datům z nešikovně provedeného pokusu, kdy experimentátor nerozlišoval vzorky z různých míst stonku. Za rozumné bych považoval zjistit citlivost výsledku na volbě obou rozptylů.

Vraťme se ke zvláštní struktuře matice $\mathbf{V}$. Protože je v každém řádku matice $\mathbf{Z}$ kromě nul jediná jednička, jsou sloupce této matice navzájem ortogonální. Protože v každém sloupci je právě P jedniček, platí $\mathbf{Z}^T\mathbf{Z} = P\mathbf{I}$. Použijeme-li známý vzoreček (např. výsledek cvičení 11 na str. 73 knihy Anděl [1])

$$\big(\mathbf{A} + \mathbf{U}\mathbf{W}^T\big)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\mathbf{R}^{-1}\mathbf{W}^T\mathbf{A}^{-1},$$

kde je $\mathbf{R} = \mathbf{I} + \mathbf{W}^T\mathbf{A}^{-1}\mathbf{U}$ a všechny invertované matice jsou regulární, dostaneme

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2}\mathbf{I} - \frac{\sigma_A^2}{\sigma^2(\sigma^2 + P\sigma_A^2)}\mathbf{Z}\mathbf{Z}^T.$$

Varianční matici odhadu můžeme tedy zapsat jako

$$\begin{aligned}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} &= \left(\mathbf{X}^T\left(\frac{1}{\sigma^2}\mathbf{I} - \frac{\sigma_A^2}{\sigma^2(\sigma^2+P\sigma_A^2)}\mathbf{Z}\mathbf{Z}^T\right)\mathbf{X}\right)^{-1}\\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} + \sigma^2\sigma_A^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\mathbf{R}^{*-1}\mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}, \qquad (7)\end{aligned}$$

kde je (použijeme rovnost $\mathbf{Z}^T\mathbf{Z} = P\mathbf{I}$)

$$\begin{aligned}\mathbf{R}^* &= \sigma^2\mathbf{I} + \sigma_A^2\big(\mathbf{Z}^T\mathbf{Z} - \mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\big)\\ &= \sigma^2\mathbf{I} + \sigma_A^2\mathbf{Z}^T\big(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\big)\mathbf{Z}.\end{aligned}$$

Z posledního vyjádření je patrné, že $\mathbf{R}^*$ je pozitivně definitní, takže varianční matice odhadu regresních koeficientů počítaná v (7) je rovna varianční matici modelu bez náhodných efektů zvětšené o kladný násobek pozitivně semidefinitní matice

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}\mathbf{R}^{*-1}\mathbf{Z}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}. \qquad (8)$$

Při interpretace posledního výrazu začněme maticí $\mathbf{Z}^T\mathbf{X}$. Označme M-tice výchozích hodnot dvou regresorů jako $\mathbf{x}$ a $\mathbf{v}$, jejich průměry jako $\overline{x}$ a $\overline{v}$. Matici $\mathbf{X}$ pak můžeme pomocí Kroneckerova součinu zapsat jako

$$\mathbf{X} = \big((\mathbf{x} - \overline{x}) \otimes \mathbf{1}_P, \mathbf{Z}(\mathbf{v} - \overline{v})\big),$$

takže s ohledem na $\mathbf{Z}^T\mathbf{Z} = P\mathbf{I}_M$ dostaneme

$$\mathbf{Z}^T\mathbf{X} = \begin{pmatrix}\mathbf{Z}^T((\mathbf{x} - \overline{x}\mathbf{1}_M) \otimes \mathbf{1}_P) & P(\mathbf{v} - \overline{v}\mathbf{1}_P)\end{pmatrix}.$$

Druhý sloupec této matice udává ke každému místu na stonku (těm odpovídají sloupce matice $\mathbf{Z}$) P-násobek hodnoty centrovaného regresoru V, který charakterizuje deterministický vliv tohoto místa na měřenou hodnotu. Na tvaru matice $\mathbf{K}$ druhý sloupec nijak nezávisí, na rozdíl od prvního sloupce, kde jsou podobně uvedeny součty hodnot centrovaného regresoru X, který zprostředkovává vliv koncentrace, opět vždy pro vzorky ze stejného místa na stonku. Tady bychom se mohli snažit volbou matice $\mathbf{K}$ minimalizovat velikost jednotlivých složek.

Když si představíme lineární modely závislosti jednotlivých sloupců matice $\mathbf{Z}$ na matici $\mathbf{X}$, pak odhady regresních koeficientů poskytne matice $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$. Matice $\mathbf{C}$ pak jakýmsi způsobem měří „velikost“ této matice odhadů regresních koeficientů s ohledem na „vzdálenost“ danou maticí $\mathbf{R}^*$. Přitom je matice $\mathbf{R}^*$ rovna kladnému násobku jednotkové matice zvětšenému o kladný násobek matice „reziduálních součtů součinů“ zmíněné závislosti sloupců $\mathbf{Z}$ na $\mathbf{X}$ (viz Rao [2, Kapitola 8c.2]).

4. Co jsem nakonec poradil

Ani když jsem předpokládal speciální hodnoty obojích regresorů, totiž $x_i = i$ a $v_k = k$, tak jsem ve zjednodušování výrazu pro rozptyl odhadu parametru β příliš nepokročil. Proto mi nezbylo, než použít počítač. Celkem jednoduchý program v R umožnil spočítat směrodatnou odchylku odhadu parametru β pro několik rozmístění vzorků popsaných maticemi typu $\mathbf{K}$ a několik hodnot σ a σ_A.

Pokud se v některém řádku matice $\mathbf{K}$ opakovaly stejné hodnoty (vzorky ze stejného místa se dostaly do nádobek se stejnou koncentrací), znamenalo to vždy zhoršení přesnosti odhadu směrnice β. Zkoušel jsem tedy zejména taková rozmístění, kdy druhý a další sloupce matice $\mathbf{K}$ vznikl pokaždé jiným cyklickým posunutím prvního sloupce. V případě $M = 20$ a $P = 5$ bylo nejlepší posunovat hodnoty vždy o čtyři jednotky. Jako nejvýhodnější se v tomto případě jevila matice

$$\mathbf{K}^T = \begin{pmatrix} 1 & 3 & 5 & 7 & 9 & 11 & 13 & 15 & 17 & 19 & 20 & 18 & 16 & 14 & 12 & 10 & 8 & 6 & 4 & 2 \\ 8 & 6 & 4 & 2 & 1 & 3 & 5 & 7 & 9 & 11 & 13 & 15 & 17 & 19 & 20 & 18 & 16 & 14 & 12 & 10 \\ 16 & 14 & 12 & 10 & 8 & 6 & 4 & 2 & 1 & 3 & 5 & 7 & 9 & 11 & 13 & 15 & 17 & 19 & 20 & 18 \\ 17 & 19 & 20 & 18 & 16 & 14 & 12 & 10 & 8 & 6 & 4 & 2 & 1 & 3 & 5 & 7 & 9 & 11 & 13 & 15 \\ 9 & 11 & 13 & 15 & 17 & 19 & 20 & 18 & 16 & 14 & 12 & 10 & 8 & 6 & 4 & 2 & 1 & 3 & 5 & 7 \end{pmatrix} \quad (9)$$

Nicméně rozdíl mezi touto maticí a maticí, která měla v prvním sloupci aritmetickou posloupnost $1, 2, \ldots, 20$ a další sloupce byly jen vždy o 4 řádky cyklicky posunuté, byl vcelku malý, poměr mezi směrodatnými odchylkami činil pouze 1,00012.

Literatura

[1] J. Anděl (1978). *Matematická statistika.* SNTL, Praha.

[2] C. R. Rao (1978). *Lineární metody statistické indukce a jejich aplikace.* Academia, Praha.

PRAVIDELNÝ SEMINÁŘ: ANALÝZA DAT

Karel Kupka, Marcela Salfická

E-mail: `info@trilobyte.cz`

Seminář Analýza dat, nabízený již od roku 1992 a oslavuje 20. výročí, je jedinečnou příležitostí seznámit se s posledním vývojem metod analýzy dat, modelování, statistiky a výpočetních metod a zůstat v kontaktu se současnou úrovní vědomostí a možností v této oblasti. Přednášky probíhají v češtině. Přednáší (o tom co vám ve škole neřekli) pozvaní odborníci z Akademie věd ČR, Karlovy a Masarykovy univerzity, VUT Brno, ČVUT Praha, Výzkumného centra CQR a dalších předních akademických i aplikačních pracovišť a jsou bohatě ilustrovány ukázkami řešení reálných i ad-hoc simulovaných úloh a studií na datech z technologie, managementu kvality a výzkumu se softwarem QCExpert 4, DARWin a S-Plus.

Během celého semináře je k dispozici výstava odborné literatury a mezinárodních časopisů (Quality Progress, Journal of Quality Technology, Computational Statistics & Data Analysis, Data Mining and Knowledge Discovery aj.) a počítače se softwarem. Je možné přinést i vlastní data. Po celou dobu jsou přítomní přednášející a pracovníci TriloByte jsou připravení odpovídat na vaše případné dotazy i mimo tématiku semináře.

Více informací o pravidelném semináři hledejte na webových stránkách společnosti TriloByte Statistical Software, viz `http://www.trilobyte.cz/`.

Osnova

Informační Bulletin České statistické společnosti vychází čtyřikrát do roka v českém vydání. Příležitostně i mimořádné české a anglické číslo.

Časopis je zařazen do seznamu Rady pro výzkum, vývoj a inovace, více viz server `http://www.vyzkum.cz/`.

The Bulletin of the Czech Statistical Society is published quarterly. Most of the contributions are published in Czech and Slovak languages.

DOI: 10.5300/IB, http://dx.doi.org/10.5300/IB
ISSN 1210–8022 (Print), ISSN 1804–8617 (Online)