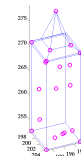
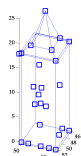


UNIVERZITA PALACKÉHO V OLOMOUCI
PŘÍRODOVĚDECKÁ FAKULA
KATEDRA MATEMATICKÉ ANALÝZY A APLIKACÍ MATEMATIKY

„ODAM 2006“

SBORNÍK PRACÍ



Olomoucké dny aplikované matematiky

čtvrtek 7. 9. 2006 a pátek 8. 9. 2006

Olomouc – Hejčín, Tomkova 40

OBSAH

Barbora ARENDACKÁ	
O jednom intervalom odhade pre variančný komponent	5
Marek BRABEC a Jaroslav MRÁZ	
Statistický model pro korigovaný odhad toxikokinetiky MVH	13
Marie BUDÍKOVÁ a Jiří DAN	
Osobnostní a sociální determinanty hodnocení prestiže učitelského povolání u středoškoláků	25
Katarína CIMERMANOVÁ	
Štatistické charakteristiky vybraných plynných komponentov vydýchnutého vzduchu meraného pomocou PTR-MS	37
Veronika KAINZOVÁ, Renáta HOLUBOVÁ, Jaroslav MAREK, Marta KLEČKOVÁ	
Závěr statistického výzkumu prekonceptů žáků základních škol v ČR	49
Lubomír KUBÁČEK a Eva TESAŘÍKOVÁ	
Nápadně vybočující údaje v slabě nelineárních regresních modelech	77
Pavla KUNDEROVÁ	
Eliminating transformations for nuisance parameters in linear regression models with constraints of the type I	85
Jaroslav MAREK, Roman PAWIŃSKI	
Edikt o zeměpisné délce a první statistické algoritmy pro aproximaci dat	95
Pavel TUČEK, Jaroslav MAREK	
Rychlost proudění v řece	121
Blanka ZEMKOVÁ	
Fuzzy škály v psychologii	131



O jednom intervalom odhade pre variančný komponent^{*}

BARBORA ARENDACKÁ

Ústav merania, Slovenská akadémia vied,
Dúbravská cesta 9, 841 04 Bratislava
e-mail: barendacka@gmail.com

Abstrakt: V článku sa budeme zaoberať intervalovým odhadom pre variančný komponent zodpovedajúci náhodnému efektu v zmiešanom lineárnom modeli s dvomi variančnými komponentami. Konkrétne sa zameriame na obojstranné konfidenčné intervaly navrhnuté El-Bassiounim v [5], ktoré možno považovať jednak za zovšeobecnenie známeho Williamsovho–Tukeyho intervalu ([13, 11]), jednak za modifikáciu intervalu skúmaného napr. Hartungom a Knappom v [6]. Uvedieme priaznivé vlastnosti navrhnutých intervalov v niektorých limitných situáciách, ktoré podporujú doterajšie dobré simulačné výsledky.

Kľúčová slova: zmiešaný lineárny model, variančné komponenty, približné konfidenčné intervaly

1 Úvod

Zmiešaný lineárny model s dvomi variančnými komponentami, ktorým sa budeme v článku zaoberať, predpokladá, že vektor pozorovaní y pochádza z mnohorozmerného normálneho rozdelenia $N_n(X\beta, \sigma_1^2 V + \sigma^2 I)$, kde X, V sú známe matice, V je pozitívne semidefinitná a $\beta, (\sigma_1^2, \sigma^2)^T$ sú vektory neznámych parametrov. Využívajúc invariantnosť variančných komponentov σ_1^2, σ^2 vzhľadom na posunutie v strednej hodnote, je inferencia o týchto parametroch založená na minimálnej postačujúcej množine štatistík pre maximálny invariant $B^T y$, kde $BB^T = M = I - X(X^T X)^{-1} X^T$, $B^T B = I$. Spomínané štatistiky sú navzájom nezávislé kvadratické formy vektora y ,

$$U_i = y^T E_i y \sim (\lambda_i \sigma_1^2 + \sigma^2) \chi_{\nu_i}^2, i = 1, 2, \dots, r$$

^{*}Práca bola podporená grantom VEGA 1/3016/06 Vedeckej grantovej agentúry MŠ SR a SAV.

kde $\lambda_1 > \dots > \lambda_r \geq 0$ sú navzájom rôzne vlastné čísla matice $B^T V B$, ν_i , $i = 1, \dots, r$ sú ich násobnosti a E_i , $i = 1, \dots, r$ ($E_i^2 = E_i$, $E_i E_j = 0$, $i \neq j$) sú matice zo spektrálneho rozkladu $B^T V B = \sum_i \lambda_i E_i$, pozri tiež [7]. Typicky, $\lambda_r = 0$, a teda $U_r \sim \sigma^2 \chi_{\nu_r}^2$. Takúto situáciu budeme v ďalšom predpokladať. Zaujímáť nás budú obojstranné konfidenčné intervaly pre σ_1^2 , konštrukcia ktorých je sťažená prítomnosťou rušivého parametra σ^2 . V časti 2 najprv uvedieme dobre známe približné riešenie Williamsa a Tukeyho (pozri [13, 11]) pre prípad, keď $r = 2$, čo zodpovedá napríklad vyváženému modelu jednoduchého triedenia s náhodným efektom. V časti 3 sa budeme zaoberať riešením navrhnutým El-Bassiounim v [5], ktoré je rozšírením Williamsovho–Tukeyho riešenia pre prípad $r > 2$ a v časti 4 uvedieme niektoré vlastnosti tohto riešenia.

V ďalšom $F_{m,n;\alpha}$, $\chi_{m;\alpha}^2$ označujú α kvantily príslušného F , resp. χ^2 rozdelenia a symbolom s označíme $\sum_{i=1}^{r-1} \nu_i$. Ďalej, α_l , α_u sú z $(0, 1)$, $\alpha_l < \alpha_u$, $\alpha_u - \alpha_l = 1 - \alpha$, kde $1 - \alpha$ je zvolená konfidenčná úroveň.

2 Williamsov–Tukeyho interval

Asi najjednoduchším príkladom zmiešaného lineárneho modelu s dvomi variančnými komponentami je vyvážený model jednoduchého triedenia s náhodným efektom, t.j.

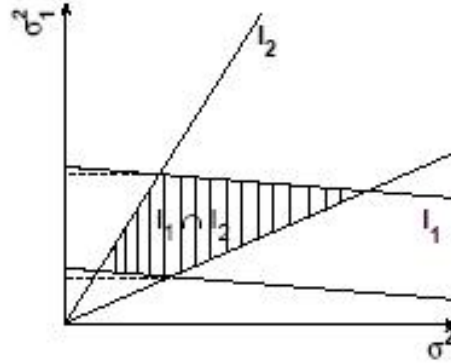
$$y_{jk} = \mu + \alpha_j + \epsilon_{jk}, j = 1, \dots, J, k = 1, \dots, K,$$

kde $\alpha \sim N_J(0, \sigma_1^2 I)$, $\epsilon \sim N_{JK}(0, \sigma^2 I)$, $\text{cov}(\alpha, \epsilon) = 0$. Prirodzene, práve v tomto modeli boli najprv hľadané konfidenčné intervaly pre σ_1^2 . Ide o situáciu, kedy sa počet postačujúcich štatistík redukuje na dve ($r = 2$), U_1 , U_2 , ktoré zodpovedajú sumám štvorcov rozdielov medzi súbormi a vo vnútri súborov známym z tabuľky analýzy rozptylu, a tieto štatistiky sú zároveň úplné. Avšak tento fakt vôbec nezjednodušuje situáciu pri konštrukcii konfidenčných intervalov pre σ_1^2 a presné riešenia nie sú známe. Približné riešenie dávajúce pomerne dobré výsledky bolo nezávisle navrhnuté Williamsom [13] a Tukeym [11].

Williams pri navrhovaní svojho riešenia využil, že v situácii s $U_1 \sim (\lambda_1 \sigma_1^2 + \sigma^2) \chi_{\nu_1}^2$, $U_2 \sim \sigma^2 \chi_{\nu_2}^2$, vieme pri známom σ^2 skonštruovať presný konfidenčný interval pre σ_1^2 , a to z presného konfidenčného intervalu

1. pre $\lambda_1 \sigma_1^2 + \sigma^2$ založeného na štatistike U_1 ,
2. pre σ_1^2 / σ^2 založeného na $U_1 / U_2 (\sim (\lambda_1 \sigma_1^2 / \sigma^2 + 1) \chi_{\nu_1}^2 / \chi_{\nu_2}^2)$.

Označme výsledné intervaly pri známom σ^2 , $I_1(\sigma^2)$, $I_2(\sigma^2)$. Potom, ako uvádza Williams, pravdepodobnosť, že pri danom σ^2 , pokryje prienik oboch intervalov skutočnú hodnotu σ_1^2 je aspoň $1 - 2\alpha$, pozri [13], a teda aj po preintegrovaní cez ľubovoľné rozdelenie parametra σ^2 , bude pravdepodobnosť pokrytia σ_1^2 prienikom oboch intervalov aspoň $1 - 2\alpha$. Spomínaný prienik sa vždy nachádza v intervale, ktorého hranice dostaneme ako prieniky dolných, resp. horných hraníc $I_1(\sigma^2)$, $I_2(\sigma^2)$, viď Obr. 1, a preto Williams navrhol zobrať práve tento interval (ktorý už od rušivého parametra σ^2 nezávisí) ako približný konfidenčný



Obrázek 1: Williamsom([13]) uvádzané grafické znázornenie navrhovaného intervalu.

interval pre σ_1^2 . Takýto interval bude mať totiž skutočnú pravdepodobnosť pokrytia určite väčšiu ako $1 - 2\alpha$. Hranice Williamsovho–Tukeyho intervalu majú tvar:

$$\left[\frac{1}{\lambda_1 \chi_{\nu_1; \alpha_u}^2} \left(U_1 - \frac{\nu_1 U_2 F_{\nu_1, \nu_2; \alpha_u}}{\nu_2} \right), \frac{1}{\lambda_1 \chi_{\nu_1; \alpha_l}^2} \left(U_1 - \frac{\nu_1 U_2 F_{\nu_1, \nu_2; \alpha_l}}{\nu_2} \right) \right]$$

V prípade, že by nejaká z hraníc vyšla záporná, kladie sa rovná nule (keďže σ_1^2 je nezáporný parameter).

Williamsov–Tukeyho interval je jedným z odporúčaných na použitie na základe rozsiahlej Boardmanovej simulačnej štúdie (pozri [3]), v ktorej boli porovnávané viaceré riešenia pre situáciu s $r = 2$. Z výsledkov tejto štúdie čiastočne vychádzal aj El-Bassiouni pri navrhovaní riešenia pre situáciu s $r > 2$, ako bude popísané ďalej. v simuláciách bola konfidenčná úroveň Williamsovho - Tukeyho intervalu blízka nominálnej hodnote a neskôr Wang [14] zdôvodnil, že za istých podmienok je jeho konfidenčná úroveň aspoň $1 - \alpha$.

3 Interval navrhnutý El-Bassiounim

Jedným z intervalov uvažovaných v porovnávacej Boardmanovej štúdii [3] ($r = 2$) bol aj interval uvádzaný napr. v práci Andersona a Bancrofta [1] (ako je uvedené v [5]) alebo Brossa [4]. Je skonštruovaný z presného konfidenčného intervalu pre podiel $\rho = \sigma_1^2 / \sigma^2$ variančných komponentov, $[\hat{\rho}_l, \hat{\rho}_u]$, prenasobením jeho hraníc nevychýleným odhadom $\hat{\sigma}^2 = U_2 / \nu_2$ pre σ^2 :

$$[\hat{\sigma}^2 \hat{\rho}_l, \hat{\sigma}^2 \hat{\rho}_u]. \quad (1)$$

Bross [4] aj Boardman [3] zhodne uvádzajú, že interval (1) má ale tendenciu byť mierne konzervatívny. Boardman uvádza, že je trochu širší ako Williamsov–

Tukeyho interval, ktorý v jeho simulačnej štúdii dosahoval pravdepodobnosť pokrytia blízku nominálnej hodnote $1 - \alpha$. Z toho a z faktu, že Williamsov–Tukeyho interval možno zapísať v tvare

$$[c_1 \hat{\sigma}^2 \hat{\rho}_l, c_2 \hat{\sigma}^2 \hat{\rho}_u], \quad (2)$$

kde $c_1 = \nu_1 F_{\nu_1, \nu_2; \alpha_u} / \chi_{\nu_1; \alpha_u}^2$, $c_2 = \nu_1 F_{\nu_1, \nu_2; \alpha_l} / \chi_{\nu_1; \alpha_l}^2$, El-Bassiouni uzatvára, že konštanty c_1 , c_2 možno považovať za korekčné faktory aplikované na interval (1), ktoré zohľadňujú, že namiesto skutočnej hodnoty σ^2 použijeme jej odhad. Keďže v zdanlivo komplikovanejšom prípade, keď $r > 2$, je tiež možné skonštruovať presný konfidenčný interval pre ρ , napr. ako bol odvodený Waldom [12], pozri tiež [8], El-Bassiouni navrhuje skonštruovať interval pre σ_1^2 analogický intervalu (2), kde $\hat{\rho}_l, \hat{\rho}_u$ sú hranice Waldovho presného konfidenčného intervalu pre ρ , t.j. sú riešeniami rovníc $G(\rho) = F_{s, \nu_r; \alpha_u}$, resp. $G(\rho) = F_{s, \nu_r; \alpha_l}$, kde

$$G(\rho) = \frac{\nu_r \sum_{i=1}^{r-1} U_i / (\lambda_i \rho + 1)}{s U_r},$$

(riešenia uvedených rovníc je možné získať napr. Newtonovou–Raphsonovou metódou, opäť, ak neexistuje nezáporné riešenie, príslušná hranica sa kladie rovná 0), $\hat{\sigma}^2 = U_r / \nu_r$ a konštanty c_1, c_2 sú tvaru $c_1 = s F_{s, \nu_r; \alpha_u} / \chi_{s; \alpha_u}^2$, $c_2 = s F_{s, \nu_r; \alpha_l} / \chi_{s; \alpha_l}^2$. Navrhovaný interval je teda vlastne zovšeobecnením Williamsovho–Tukeyho intervalu a ako ukážeme v ďalšej časti, je ho možné odvodiť aj priamo aplikovaním Williamsovho postupu. V ďalšom preto budeme tento interval označovať ako El-Bassiouniho–Williamsov–Tukeyho interval.

V [5] El-Bassiouni simulačne porovnal tento interval s ďalšími približnými konfidenčnými intervalmi pre σ_1^2 odvodenými pre $r > 2$. Ukázalo sa, že iba pravdepodobnosť pokrytia navrhovaného intervalu nebola mešia ako $1 - \alpha$ vo všetkých uvažovaných dizajnoch a pre všetky uvažované hodnoty variančných komponentov, čím tento interval predstihol ostatné riešenia napr. vo výrazne nevyvážených modeloch. Jeho trochu väčšia výpočtová náročnosť v porovnaní s inými riešeniami, napr. Thomasovým–Hultquistovým [10], sa teda vyplatí.

4 Niektoré vlastnosti uvedeného intervalu

V tejto časti spomenieme niektoré priaznivé vlastnosti El-Bassiouniho–Williamsovho–Tukeyho intervalu, ktoré podporujú výsledky, ktoré dosahoval v simuláciách. Vráťme sa ale najprv k intervalu (1) z predchádzajúcej časti. Pre situáciu s $r > 2$, t.j. $\hat{\rho}_l, \hat{\rho}_u$ zodpovedajúce hraniciam Waldovho intervalu pre ρ , $\hat{\sigma}^2 = U_r / \nu_r$, bol uvažovaný v [6] a tiež v [2]. Dá sa o ňom ukázať, že ak $\nu_r \rightarrow \infty$, je presný pre $\sigma_1^2 > 0$ (dôvod vylúčenia 0 bude zrejmy neskôr), ($\nu_r \rightarrow \infty$ je situácia, keď sa odhad $\hat{\sigma}^2$ stáva presným, t.j. rušivý parameter, ktorý komplikuje situáciu, sa stáva známym a interval (1) je pre známe σ^2 presný) a skutočnú nulovú hodnotu σ_1^2 pokrýva s pravdepodobnosťou $1 - \alpha_l > 1 - \alpha$, čo je však čiara, ktorú by mal aj presný konfidenčný interval pre σ_1^2 . (V prípade presných intervalov pre ρ na podobnú situáciu upozornili Seely a El-Bassiouni v [8].) Dôvodom

je, že prípadné záporné hranice intervalu sa kladú rovné 0, a teda ak považujeme aj degenerovaný interval v nule za interval, tak nula je pokrytá kedykoľvek je dolná hranica intervalu rovná nule, bez ohľadu na hodnotu hornej hranice. To, že sa interval (1) správa ako presný pre $\sigma_1^2 = 0$ a $\nu_r \rightarrow \infty$ sa odzrkadlilo aj v pomerne dobrých výsledkoch, čo sa hladiny významnosti týka, v simuláciách. Avšak zhodne s tvrdeniami Brossa a Boardmana, sa tento interval javil byť mierne konzervatívny s narastajúcou skutočnou hodnotou ρ . Dá sa ukázať, že skutočná pravdepodobnosť pokrytia tohto intervalu

$$\begin{aligned} P(\hat{\sigma}^2 \hat{\rho}_l \leq \sigma_1^2 \leq \hat{\sigma}^2 \hat{\rho}_u) &= P(G(\hat{\rho}_u) \leq G(\sigma_1^2/\hat{\sigma}^2) \leq G(\hat{\rho}_l)) = \\ &= P(sF_{s,\nu_r;\alpha_l} \leq \sum_{i=1}^{r-1} \frac{U_i}{\lambda_i \sigma_1^2 + U_r/\nu_r} \leq sF_{s,\nu_r;\alpha_u}) \end{aligned} \quad (3)$$

(využili sme klesajúcosť $G(\cdot)$ na $(-1/\lambda_1, \infty)$), konverguje pre $\sigma_1^2/\sigma^2 \rightarrow \infty$ k

$$P(sF_{s,\nu_r;\alpha_l} \leq \chi_s^2 \leq sF_{s,\nu_r;\alpha_u}). \quad (4)$$

(Keďže $(\frac{U_1}{\lambda_1 \sigma_1^2}, \dots, \frac{U_{r-1}}{\lambda_{r-1} \sigma_1^2}, \frac{U_r}{\sigma^2} \frac{\sigma^2}{\sigma_1^2})$ konverguje v distribúcii k $(Q_1, \dots, Q_{r-1}, 0)$, kde $Q_i \sim \chi_{\nu_i}^2$ sú navzájom nezávislé.) Pre obvykle volené $\alpha_l = \alpha/2$, $\alpha_u = 1 - \alpha/2$ je hodnota (4) väčšia ako $1 - \alpha$, pozri [2]. Výrazy (3) a (4) zároveň naznačujú, ako je možné uvedený interval modifikovať, aby bol výsledný interval presný pre $\sigma_1^2/\sigma^2 \rightarrow \infty$. Riešením, ktoré sa ponúka, je prenášobenie hraníc intervalu (1) konštantami $c_1 = sF_{s,\nu_r;\alpha_u}/\chi_{s;\alpha_u}^2$, $c_2 = sF_{s,\nu_r;\alpha_l}/\chi_{s;\alpha_l}^2$. Pre pravdepodobnosť pokrytia skutočnej hodnoty σ_1^2 novovzniknutým intervalom platí:

$$\begin{aligned} P(c_1 \hat{\sigma}^2 \hat{\rho}_l \leq \sigma_1^2 \leq c_2 \hat{\sigma}^2 \hat{\rho}_u) &\xrightarrow{\sigma_1^2/\sigma^2 \rightarrow \infty} \\ &P\left(sF_{s,\nu_r;\alpha_l} \leq \frac{sF_{s,\nu_r;\alpha_l}}{\chi_{s;\alpha_l}^2} \chi_s^2 \ \& \ \frac{sF_{s,\nu_r;\alpha_u}}{\chi_{s;\alpha_u}^2} \chi_s^2 \leq sF_{s,\nu_r;\alpha_u}\right) = \\ &= P(\chi_{s;\alpha_l}^2 \leq \chi_s^2 \leq \chi_{s;\alpha_u}^2) = 1 - \alpha. \end{aligned}$$

Navyše, uvedená modifikácia zachová vyššie spomínané dve vhodné vlastnosti pôvodného intervalu (1), a to, že sa správa ako presný pre $\sigma_1^2 = 0$ a $\nu_r \rightarrow \infty$ (pre $\sigma_1^2 > 0$). (Modifikovaný aj pôvodný interval totiž súčasne obsahujú, alebo neobsahujú nulu a pre $\nu_r \rightarrow \infty$, c_1 aj c_2 konvergujú k 1.) El-Bassiouniho–Williamsov–Tukeyho interval je teda presný v troch hraničných situáciách. Zároveň je zrejmé, ako konštanty c_1 , c_2 , ktoré El-Bassiouni označil za korekčné, zlepšujú vlastnosti pôvodného intervalu (1).

Pohľad na El-Bassiouniho–Williamsov–Tukeyho interval ako na modifikáciu intervalu (1), ktorý je prítomný aj v pôvodnom odvodení v [5], umožňuje vidieť, že pre bežne používané $\alpha (\leq 0.1)$, $s \geq 3$ a voľbu $\alpha_l = \alpha/2$, $\alpha_u = 1 - \alpha/2$, je nový interval kratší ako interval (1), keďže $c_1 \geq 1$ a $c_2 \leq 1$. (K podobnému záveru prišiel aj Boardman [3].) S narastajúcim ν_r sa rozdiel medzi oboma intervalmi, samozrejme, stiera.

Pozn. 1 Pre $s \leq 2$ (a voľbu $\alpha_l = \alpha/2$, $\alpha_u = 1 - \alpha/2$) je síce stále $c_1 \geq 1$, ale aj $c_2 \geq 1$, a teda výsledný interval môže byť dlhší ako v nemodifikovanom prípade. Rozdiel dĺžok oboch intervalov možno vyjadriť ako $(c_2 - 1)H_u - (c_1 - 1)H_l$, kde H_u , H_l označujú hornú, resp. dolnú hranicu nemodifikovaného intervalu, t.j. intervalu (1). Pre prípad $s = 2, \nu_r = 30$ je možné tento rozdiel ohraničiť zhora výrazom $(c_2 - 1)(H_u - H_l)$, z čoho plynie, že El-Bassiouniho–Williamsov–Tukeyho interval je dlhší maximálne o 0.18%, 0.09%, 0.02% dĺžky intervalu (1) pre 90%, 95%, 99%-né konfidenčné intervaly. Podobne pre $s = 1, \nu_r = 50$, je El-Bassiouniho–Williamsov–Tukeyho interval dlhší maximálne o 1.01% dĺžky intervalu (1) v prípade 90%, 95%, 99%-ných intervalov. Samozrejme, pre nižšie hodnoty ν_r môžu byť tieto rozdiely väčšie.

Pozn. 2 Dĺžku El-Bassiouniho–Williamsovho–Tukeyho intervalu môžeme ovplyvniť aj voľbou α_l, α_u . V [5] sú skúmané tzv. krátke intervaly, pre ktoré $\alpha_u = 1 - \alpha/2$, $\alpha_l = \alpha/2$ pre kvantily F rozdelenia a χ^2 kvantily sú z tabuľky 678 v [9], t.j. pracuje sa s $F_{s, \nu_r; \alpha/2}, F_{s, \nu_r; 1 - \alpha/2}, \chi_{s; \beta_l}^2, \chi_{s; \beta_u}^2$. Pri takejto voľbe zostáva výsledný interval presný pre $\sigma_1^2 = 0$ a $\sigma_1^2/\sigma^2 \rightarrow \infty$. (Konštanty c_1, c_2 pri tejto voľbe nekonvergujú vo všeobecnosti pre $\nu_r \rightarrow \infty$ k 1.)

Ak by sme El-Bassiouniho–Williamsov–Tukeyho interval vnímali len ako modifikáciu intervalu (1), mohla by vyvstať otázka, či po prenasobení dolnej hranice konštantou c_1 a hornej konštantou c_2 , bude výsledkom naozaj interval. Kladná odpoveď je však okamžitá, ak si uvedomíme, že El-Bassiouniho–Williamsov–Tukeyho interval je možné odvodiť doslovným zopakovaním Williamsovho postupu, a teda jeho hranice sú prienikmi dolných, resp. horných hraníc nejakých dvoch intervalov. (Tento výsledok nie je prekvapujúci, ale El-Bassiouni ho pri svojom odvodení explicitne nespomína.)

Postup, opísaný v časti 2, ktorý použil Williams pri odvodení približného konfidenčného intervalu pre σ_1^2 , spočíva v schopnosti zostrojiť presné konfidenčné intervaly pre tento parameter, pri známom σ^2 . Ako prvý interval nám posluží interval odvodený z presného konfidenčného intervalu pre $\rho, I_1(\sigma^2)$, ktorého dolná, resp. horná hranica sú tvaru:

$$H_l = \hat{\rho}_l \sigma^2 : \sum_{i=1}^{r-1} \frac{U_i}{\lambda_i \hat{\rho}_l + 1} = s F_{s, \nu_r; \alpha_u} U_r / \nu_r$$

$$H_u = \hat{\rho}_u \sigma^2 : \sum_{i=1}^{r-1} \frac{U_i}{\lambda_i \hat{\rho}_u + 1} = s F_{s, \nu_r; \alpha_l} U_r / \nu_r$$

a teda ich dostaneme ako riešenia nasledujúcich rovníc

$$\sum_{i=1}^{r-1} \frac{U_i}{\lambda_i H_l + \sigma^2} = \frac{1}{\sigma^2} s F_{s, \nu_r; \alpha_u} U_r / \nu_r$$

$$\sum_{i=1}^{r-1} \frac{U_i}{\lambda_i H_u + \sigma^2} = \frac{1}{\sigma^2} s F_{s, \nu_r; \alpha_l} U_r / \nu_r. \quad (5)$$

Pri konštrukcii druhého intervalu využijeme, že pri známom σ^2 , $\sum_{i=1}^{r-1} \frac{U_i}{\lambda_i \sigma_1^2 + \sigma^2} \sim \chi_s^2$. Hranice intervalu $I_2(\sigma^2)$ teda získame riešením rovníc

$$\sum_{i=1}^{r-1} \frac{U_i}{\lambda_i \tilde{H}_l + \sigma^2} = \chi_{s;\alpha_u}^2 \quad \sum_{i=1}^{r-1} \frac{U_i}{\lambda_i \tilde{H}_u + \sigma^2} = \chi_{s;\alpha_l}^2. \quad (6)$$

Obr. 1 zostáva v platnosti, akurát závislosť hraníc I_2 na σ^2 nie je lineárna. Porovnaním (5) a (6) vidíme, že dolné hranice I_1 a I_2 sa pretnú v bode $\sigma^2 = \frac{sF_{s,\nu_r;\alpha_u} U_r/\nu_r}{\chi_{s;\alpha_u}^2}$ a horné v bode $\sigma^2 = \frac{sF_{s,\nu_r;\alpha_l} U_r/\nu_r}{\chi_{s;\alpha_l}^2}$, pričom $\tilde{H}_l = H_l = \frac{sF_{s,\nu_r;\alpha_u} U_r/\nu_r \hat{\rho}_l}{\chi_{s;\alpha_u}^2}$ a $\tilde{H}_u = H_u = \frac{sF_{s,\nu_r;\alpha_l} U_r/\nu_r \hat{\rho}_u}{\chi_{s;\alpha_l}^2}$. Získané hranice sú práve hranice El-Bassiouniho–Williamsovho–Tukeyho intervalu.

5 Záver

V článku sme sa zaoberali obojstranným konfidenčným intervalom pre variančný komponent zodpovedajúci náhodnému efektu v zmiešanom lineárnom modeli s dvomi variančnými komponentami navrhnutým El-Bassiounim v [5]. Poukázali sme na priaznivé vlastnosti tohto intervalu v niektorých limitných situáciách, ktoré sa odzrkadľujú v doteraz uskutočnených simulačných štúdiách a ktoré podporujú používanie tohto približného konfidenčného intervalu. Taktiež sme bližšie rozobrali vzťah medzi intervalom navrhnutým El-Bassiounim a intervalom uvažovaným napr. v [6], za modifikáciu ktorého je možné El-Bassiouniho interval považovať. Poukázali sme tiež na to, že El-Bassiouniho interval možno odvodiť aj priamou aplikáciou Williamsovho postupu, ktorý použil na odvodenie známeho Williamsovho–Tukeyho intervalu.

References

- [1] Anderson, R. L., Bancroft, T.A. (1952) - Statistical Theory in Research, McGraw-Hill, New York.
- [2] Arendacká, B. (2006) - Approximate confidence intervals on the variance component in a general case of a two-component model, Proc. of ROBUST 2006, JČMF, Praha, 9–17.
- [3] Boardman, T.J. (1974) Confidence intervals for variance components - a comparative Monte Carlo study, Biometrics 30, 251–262.
- [4] Bross, I. (1950) - Fiducial intervals for variance components, Biometrics 6(2), 136–144.
- [5] El-Bassiouni, M.Y. (1994) - Short confidence intervals for variance components, COCOMMUNICATIONS IN STATISTICS Theory and Methods 23 (7), 1951–1933.

- [6] Hartung, J., Knapp, G. (2000) - Confidence intervals for the between group variance in the unbalanced one-way random effects model of analysis of variance, *Journal of Statistical Computation and Simulation* 65, 311–323
- [7] Olsen, A., Seely, J., Birkes, D. (1976) - Invariant quadratic unbiased estimation for two variance components, *The Annals of Statistics* 4 (5), 878–890.
- [8] Seely, J.F., El-Bassiouni, Y. (1983) - Applying Wald's variance component test, *The Annals of Statistics*, 11, 197–201.
- [9] Tate, R.F., Klett, G.W. (1959) - Optimal confidence intervals for the variance of a normal distribution, *JASA* Vol. 54, Num. 287, 674–682.
- [10] Thomas, J.D., Hultquist, R.A. (1978) - Interval estimation for the unbalanced case of the one-way random effects model, *The Annals of Statistics* 6, 582–587.
- [11] Tukey, J.W. (1951) - Components in regression, *Biometrics* 7, 33–69.
- [12] Wald, A. (1947) - A note on regression analysis, *The Annals of Mathematical Statistics* 18, 586–589.
- [13] Williams, J.S. (1962) - A confidence interval for variance components, *Biometrika* 49, 278–281.
- [14] Wang, C. M. (1990) - On the lower bound of confidence coefficients for a confidence interval on variance components, *Biometrics* 46, 187–192.



Statistický model pro korigovaný odhad toxikokinetiky MVH^{*}

MAREK BRABEC¹ A JAROSLAV MRÁZ²

¹ Státní zdravotní ústav, Šrobárova 48, Praha 10, 100 42
a Ústav informatiky ČAV, Pod Vodárenskou věží 2, Praha 8, 182 07
e-mail: mbrabec@szu.cz, mbrabec@cs.cas.cz

² Státní zdravotní ústav, Šrobárova 48, Praha 10, 100 42
e-mail:

Abstrakt: V tomto článku popíšeme statistický model pro odhad zátěže průmyslových pracovníků látkou DMF (*N,N*-dimethylformamidem). Model je vybudován jako (po logaritmické transformaci) lineární model se smíšenými efekty (efektivně tedy jde o multiplikativní smíšený lognormální model na původní škále). Jeho podstatným rysem je však specifická korekce na intergrovanou povahu dat plynoucí ze známých toxikokinetických vlastností DMF. Zohlednění skutečnosti, že pozorovaná data vznikají konvolucí procesu příchozího DMF (expozice) s jádrem známých tvarů (daných jejich biochemickými a kinetickými vlastnostmi) poskytuje možnost korigovaného odhadu expozice (efektivně je takovýto odhad dekonvolucí příslušných pozorování). Až po dekonvoluci pak lze srovnávat různé měřitelné veličiny nepřímou charakterizující expozici a sestavit z nich kompozitní index srovnatelný s dlouhodobým markerem expozice - látkou MVH. V článku budeme diskutovat jak strukturu statistického modelu, tak problém konvoluce v pozorovaných datech, jež je speciálním případem obecně zajímavé situace s přechodem od agregovaných k disagregovaným datům (a zpět). Ten se vyskytuje i ve zcela jiných kontextech (např. při odhadu derivací z dat jež jsou efektivně integrály zkoumaného procesu - při modelování růstových křivek, [1] i jinde, [2]). Chování modelu a jeho praktické použití ukážeme na reálných datech pocházejících z italské studie průmyslových pracovníků ve vybraných průmyslových závodech.

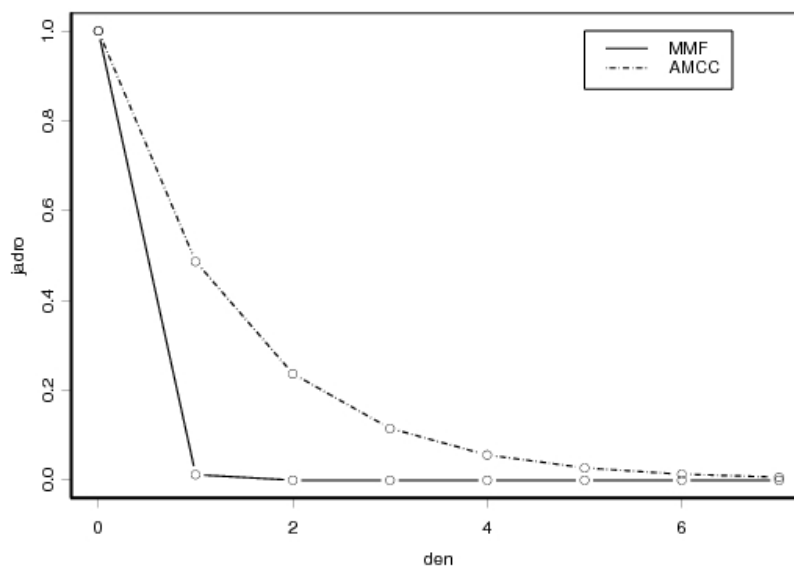
1 Úvod

Při kontrole pracovního prostředí relevantních průmyslových podniků se rutinně sleduje zátěž zaměstnanců toxickou látkou DMF (podrobnější informace

^{*}Studie, z níž byla data získána byla sponzorována Italian Institute of Prevention and Safety on Workplaces (ISPESL), project No. 28/DIL/01.

o chemické povaze i jiných aspektech této látky lze nalézt např. v článcích [6], [7]). Při monitoringu lze zátěž DMF měřit pomocí různých markerů. Například je možné měřit koncentraci DMF přímo ve vzduchu pracoviště s pomocí pasivního vzorkovacího zařízení připevněného na oděvu sledovaného pracovníka (dále ji budeme označovat jako DMFa) nebo měřit metabolické produkty její přeměny, a to v moči (látky AMCC, MMF) či krvi (látky MVH) exponovaných osob. Cílem je zjistit dávku DMF, které byl daný pracovník efektivně vystaven a porovnat ji se standardem tak, aby bylo možné posoudit zda expozice byla únosně nízká (eventuelně přijmout příslušná opatření). I když každé ze zmíněných měření dává nějakou informaci o dávce, nejsou jejich hodnoty zcela souměřitelné. Například DMFa charakterizuje okamžitou hodnotu procesu expozice v době měření. V dalším průběhu času se může DMFa podstatně měnit. Navíc DMFa charakterizuje jen potenciální výši expozice, neboť různí pracovníci do sebe vstřebávají velmi různé množství z DMF rozptýleného ve vzduchu (v závislosti na fyzické kondici, okolní teplotě, námáhavosti jejich práce, atd.). Měření metabolických produktů je z tohoto pohledu lepší (jde o realizovanou, nikoli potenciální expozici). Problémem je, že vlivem složité přeměny DMF uvnitř organismu je informace o DMF expozici (vstřebaném DMF) v měřeních metabolitů (markerů) obsažena v dosti komplikované podobě. Nově vstřebané DMF se na metabolity nepřemění ani ihned ani najednou. Expozice jediného dne se projeví v koncentracích metabolitů s různým zpožděním a různě distribuovaná i v několika dnech.

Není tedy těžké si uvědomit, že formálně je koncentrace toho kterého metabolitu $Y(t)$ v čase t konvolucí procesu expozice, $y(t)$ s nějakým jádrem w , tedy $Y(t) = \int_0^{\infty} y(t-x) \cdot w(x) dx$ (byť celý proces běží ve skutečnosti spojitěm čase, vzorkování je diskrétní - po dnech, proto pro další formulaci modelu uvažujeme diskrétní konvoluci $Y_t = \sum_{i=0}^{\infty} y_{t-i} \cdot w_i$ s diskrétním $t = 1, \dots, T$). Jádro w je dáno charakterem procesu přeměny dané látky uvnitř organismu (tedy biologicko-chemickou podstatou, nikoli tedy např. “volbou” jako v případě známého jádrového vyhlazování při neparаметrickém odhadu hustot či regresní funkce). Jeho tvar je (v zásadě) známý z předchozích (rozsáhlých) laboratorních měření. Různé látky mají různá jádra. Pro MVH jde o trojúhelníkové jádro (dané skutečností, že daný metabolit se váže na červené krvinky kde setrvává po celou dobu jejich životnosti, přičemž jejich přežití prakticky neovlivňuje): $w_i = (1 - \frac{i}{130}) \cdot I_{(0 \leq i \leq 130)}$. Jádra pro MMF a AMCC jsou dosti odlišná, v podstatě exponenciální a hlavně mnohem méně distribuovaná, viz Obrázek 2. Pro DMFa je jádro degenerované (jednotkový impuls v nule). Jinými slovy, různé markery lze společně vidět jako hodnoty vážených průměrů minulé expozice, ale s tím že vážení minulosti je pro různé látky velmi odlišné. Tyto odlišnosti je třeba vzít v úvahu při porovnání “kvality” informace poskytnuté různými látkami (relevantní např. při srovnání tradičních charakteristik DMFa, MMF, AMCC proti MVH jako markeru nové generace, charakterizujícímu dlouhodobý průběh expozice). Povšimněme si, že přímočarý srovnání typu “korelace naměřené DMFa s naměřeným MVH” jsou sice snadno proveditelná (a proto prakticky velmi oblíbená), ale v podstatě zcela nerelevantní (jde o srovnání typu “jablka s hruškami”). Nadto je výhodné zkonstruovat z několika, v praxi simultánně



Obrázek 2: MMF a AMCC jádra

měřených markerů (AMCC, MMF, DMFa) kompozitní index využívající kompletní informace v datech (krátko- i dlouho-dobější v závislosti na příslušných jádrech).

K těmto účelům v dalším odstavci článku zformulujeme statistický model s výhodou použitelný pro dekonvoluci pozorovaných dat (korekci na vliv biologicky a chemicky daného jádra) či konstrukci odhadu procesu expozice. Následně pak budeme ilustrovat jeho aplikaci v praxi na datech pocházejících z italské studie, poskytnuté pracovníky Public Health Laboratory z Florencie (jmenovitě děkujeme za pomoc při pořizování, kontrole dat a jejich přípravě pro analýzu zejména kolegům Baldasseroni, Bavazzano, Perico, Tanturli, Baglioni).

2 Model

Za podmínek typických v praxi je výsledek měření různých markerů expozice DMF ovlivněn různými systematickými odlišnostmi v podmínkách měření na něž je třeba korigovat. Měření jsou prováděna opakovaně na těchže pracovnících. Samozřejmě pak jde o analýzu typu “repeated measures”, [5] - neboť opakovaná měření s sebou přinášejí přítomnost korelace uvnitř subjektu (zatímco mezi subjekty jsou pozorování v podstatě nekorelovaná), se všemi důsledky pro odhad středních chyb, testy apod. Typicky jde v první řadě o hodnocení praco-

viště jako celku a až potom (pokud vůbec) o eventuelní porovnání individuálních odhadů expozice u jednotlivých osob. Proto je přirozené na odchylky expozice konkrétních individuí od typické hodnoty pracoviště nahlížet jako na náhodné efekty. Celkově je výhodné statistický model postavit jako model regresní, se smíšenými (tedy pevnými i náhodnými) efekty (koeficienty). Naše formulace základního modelu vychází z obecného lineárního smíšeného modelu (linear mixed model, LME), viz např. [4]:

$$y_{lij} = \exp \left(\beta_l + \sum_{s=0}^S \alpha_s \cdot t_{ij}^s + b_i + \epsilon_{lij} \right) \quad (1)$$

$$b_i \sim N(0, \sigma_B^2)$$

$$\epsilon_{lij} \sim N(0, \sigma^2 \cdot \sigma_l^2)$$

Kde:

- y_{lij} je hodnota látky l (v našem případě půjde o látky AMCC, DMFa, MMF), jež je výsledkem DMF vstřebaného u osoby i ve dni t_{ij} (elaborované značení dovolující individuálně-specifický rozvrh měření je nutné, u reálných dat máme totiž typicky co činit s více či méně nevyváženým designem daným buď už při sestavování pokusného plánu a/nebo vlivem chybějících pozorování).
- β_l je (pevný) efekt látky l .
- b_i je (náhodný) efekt pracovníka i . Povšimněme si, že mezi-individuální odchylky od “typického” stavu jsou efektivně multiplikativní a lognormálně rozdělené.
- ϵ_{lij} je normálně rozdělená chyba (pohlcuje převážně měřicí chybu, nepřesnosti modelu apod.). Zatímco typ rozdělení je společný pro všechny látky, rozptyl nikoli (měřicí chyby jsou pro různá měření obecně dosti odlišné). Rozptyl pro látku l je dán součinem obecné konstanty σ^2 a specifického činitele σ_l^2 .
- člen $\sum_{s=0}^S \alpha_s$ popisuje trend expozice v čase. Předpokládáme trend polynomického tvaru (nízkého stupně S), který ovšem nemusí být konstantní (odpovídajícímu $S = 0$).

Model je (po logaritmické transformaci) lineárním smíšeným modelem za normality. Na původní škále (před transformací) jde tedy o lognormální, smíšený, plně multiplikativní model. Podstatnou vlastností modelu (1) je separace: i) společného trendu expozice v čase, ii) společného činitele specifického pro danou látku, iii) individuálně-specifického (náhodného) činitele (stejného pro různé látky), iv) lognormálně rozdělené “měřicí chyby” (zohledňující heteroskedasticitu danou různou přesností různých chemických měření). Aby model (1) byl identifikovatelný, je samozřejmě nutné zavést (obdobně jako např. v typických modelech analýzy rozptylu) příslušné restriktce, a to jak pro parametry ovládající

střední hodnoty, tak pro parametry spojené s rozptylem. Volíme je (i s ohledem na snadnou interpretovatelnost) jako restriktce “baseline typu”: $\beta_1 = 0$, $\sigma_1^2 = 1$, pro $l = 1$ odpovídající referenční látce DMFa.

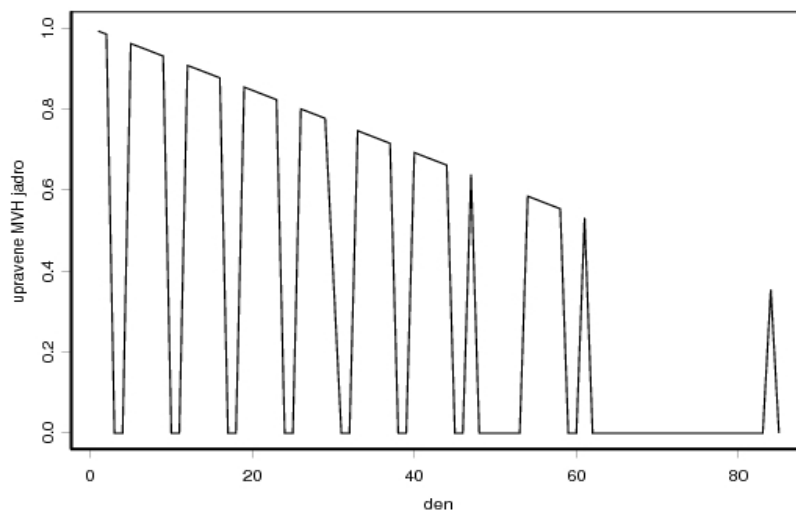
Povšimněme si, že y_{lij} je (s výjimkou DMFa) nepozorovatelnou veličinou, proto se základní model (1) na měřených datech nedá přímo použít. Jak jsme diskutovali v odstavci 1, pozorovatelné jsou jen konvoluce veličin y_{lij} s příslušnými jádry, tedy veličiny které značíme Y_{lij} . Jde o součty součinů $y_{li}(t_{ij} - s) \cdot w_s$ (tedy vážené součty minulých expozic za mnoho dnů zpět, pro jedince i , v časech $t_{ij} - s$). Skutečně proložitelný model dostáváme až po napočtení příslušných konvolucí:

$$Y_{lij} = \sum_{q=0}^{\infty} w_{i,t_{ij}-q}^P \cdot w_{l,q}^K y_{li}(t_{ij} - q) \quad (2)$$

Kde:

- Y_{lij} je hodnotota koncentrace látky l pro člověka i , pozorovatelná ve dni t_{ij} . Na rozdíl od y_{lij} jde o přímo měřitelnou veličinu.
- $w_{l,\cdot}^K$ je (konvoluční) jádro pro látku l . Je dáno kinetikou dané látky v organismu.
- Reálná situace je ještě mírně komplikovanější oproti kineticky motivované konvoluci zmíněné v odstavci 1. Důležitou roli totiž hraje i absence v práci, která pro daný den vylučuje expozici (jde o látku jejichž příspěvek z jiných než pracovních zdrojů se považuje za zanedbatelný). $w_{i,t}^P$ je váha příspěvku daného dne t pro jedince i daná prezencí/absencí v práci. Typicky jde o nula/jedničkový indikátor ($w_{i,t}^P = 0$ pro t odpovídající víkendům, dnům dovolené apod. a $w_{i,t}^P = 1$ pro pracovní dny), ale v případě necelých odpracovaných směn může $w_{i,t}^P$ nabývat hodnot mezi nulou a jedničkou (ty se vyskytují např. i v našich ilustrativních datech analyzovaných v sekci 3). Vliv modulace dané $w_{i,t}^P$ není nepodstatný. Například Obrázek 3 ukazuje součin $w_{i,q}^P \cdot w_{MVH,q}^K$ jako funkci q pro jednoho konkrétního jedince (jeho pracovní rozvrh). Průběh je viditelně podstatně odlišný od původního trojúhelníkového jádra pro MVH (výrazně odlišná je i plocha pod křivkou).

Kompletní model je specifikován rovnicemi (1) a (2). Povšimněme si, že konvoluce v (2) s sebou přináší komplikace, zejména z výpočetního pohledu. Pokud předpokládáme “konstantní trend” expozice (tedy $S = 0$), nepředstavuje (2) větší problém a celý model zůstává lineárním smíšeným modelem, jež lze pohodlně proložit s pomocí monoha statistických balíků. Ovšem s tím že musíme design matici pro pevné s časem spojené efekty (tedy polynomický trend v čase) příslušným způsobem naintegrovat ještě před použitím rutiny pro lineární smíšený model. Pokud ovšem je $S > 0$, je situace obtížnější v tom, že celkový model se stává nelineárním modelem se smíšenými efekty. I takový lze ale relativně snadno prokládat s použitím vhodného SW.



Obrázek 3: Upravené MVH jádro

3 Ilustrativní příklad na reálných datech

Chování modelu budeme ilustrovat na datech z toskánské studie provedené ve čtyřech vybraných závodech (označených 1, 3, 4, 5) na celkem 40 subjektech s kompletními výsledky MVH měření (na počátku a na konci studie). Kromě toho u nich byly měřeny DMFa, MMF a AMCC, a to v obecně různém počtu opakování (maximálně 6) s přibližně dvoutýdenním odstupem. Délka studie byla 85 dnů.

Byl použit model specifikovaný rovnicemi (1) a (2) v odstavci 2 s nekonztantním trendem expozice (na základě Akaikeho kritéria jsme zvolili $S = 5$). Jde tedy o nelineární smíšený model (díky konvolučním součtům nelze nelinearitě odstranit transformací obou stran, či TBS metodou, [3]). Při výpočetní implementaci se nám osvědčila nlme knihovna v R či S-plus, [8]. Prokládání se realizuje pomocí REML (restricted maximum likelihood), tedy s penalizací rozsahu pevné části modelu. Povšimněme si nejprve, že korelace mezi hodnotami na témže subjektu je nezanedbatelná (pokud by nebyla v modelu zohledněna, došlo by k výrazné deformaci odhadu variability, k inflaci chyby I. druhu chyby u testů např. pro látkové efekty). Tuto korelaci lze vcelku snadno spočítat z REML odhadů komponent rozptylu a udává ji následující tabulka:

Továrna	$\hat{\sigma}_B^2$	$\hat{\sigma}^2$	$\hat{\rho}$
1	0.4338	0.3808	0.5648
3	0.4254	0.7295	0.2537
4	0.3029	0.3897	0.3766
5	0.4115	0.5665	0.3455

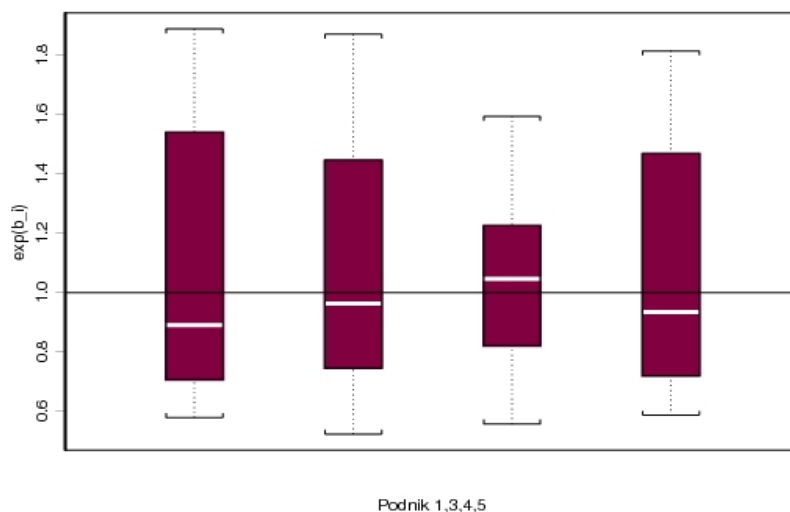
Obrázek 4 srovnává krabicové grafy rozdělení multiplikatивních náhodných efektů $\exp(b_i)$ (po exponencování, tedy na původní škále) mezi jednotlivými podniky. Je patrný jejich poměrně výrazný asymetrický charakter i to že rozdělení individuálních efektů se mezi podniky příliš neliší - na rozdíl od pevných efektů podniku, charakterizujících odlišnosti v typické hladině expozice během studie (průměr přes čas, medián přes individua), jež jsou vysoce významně odlišné.

Model i přes svou parsimonii (počet parametrů, zejména pevných je relativně nízký, vzhledem k velikosti dat a tedy k potenciální velikosti saturovaného modelu) prokládá empirická data vcelku dobře. Typický diagnostický graf, často používaný v regresi je na Obrázku 5. Je třeba si však uvědomit, že díky své povaze (modelu se smíšenými efekty), model 1+(2) umožňuje odhad na dvou úrovních hierarchie: i) průměry přes individua (odpovídající pevným efektům), ii) odhady uvnitř jednotlivých individuí. Ukázka odhadu AMCC pro jednoho konkrétního jedince je na Obrázku 6. Vidíme odtud, jak komplikovanou dynamiku AMCC konvoluční jádro implikuje. Koncentrace prudce klesá o víkend (v obrázku jsou soboty a neděle vyznačeny krátkými svislými čarami na vodorovné ose) - kdy je "přívod DMF" vypojen. Během pracovní části týdne se (vlivem) kumulace dané relativně distribuovaným tvarem jádra koncentrace zvyšuje až do pátku. Kromě typického chodu jsou patrné odchylky dané absencí v práci (pro daného jedince ve 4. týdnu studie). Diagnostiku modelu lze provést i na úrovni jednotlivého subjektu. Například Obrázek 7 srovnává odhadnutý průběh individuální trajektorie s pozorovanými daty (měřeními v diskretních bodech).

Kromě rozdílných úrovní hierarchie lze odhady (a inference obecně) z modelu počítat jak pro "konvoluovaný proces" Y (tj. srovnatelný s pozorovanými daty) jako na Obrázku 7, tak i pro "dekonvoluovaný proces" y (tj. proces čistých přírůstků za jednotlivé dny). Obrázek 8 srovnává odhady obou veličin pro jednoho konkrétního pracovníka a MMF (v obou případech se započtením členu $w_{i,t}^P$ který zohledňuje prezenci/absenci) a demonstruje že jde o zcela různé věci - výrazně se odlišují nejen teoreticky (hladší integrovaný průběh pro Y), ale i z mnoha praktických hledisek.

S pomocí modelu 1+(2) lze také snadno zkonstruovat kompozitní indikátor DMF expozice založený na souhrnné informaci ze simultánního měření DMFa, MMF, AMCC. Jde o vyhodnocení individuálního časového trendu (na úrovni čistých denních přírůstků expozice, y) pro arbitrárně zvolenou látku (např. DMFa). Takovýto indikátor lze pak srovnávat s (příslušně dekonvoluovaným) rozdílem měření MVH na konci a počátku studie. Připomeňme, že přímá srovnání MVH diference s kteroukoli z látek DMFa, MMF, AMCC nejsou příliš relevantní, neboť jejich jádra mají výrazně odlišný nosič. Obrázek 9 srovnává kompozitní indikátor s MVH diferencí a ukazuje, že shoda mezi oběma typy informace (pokud je

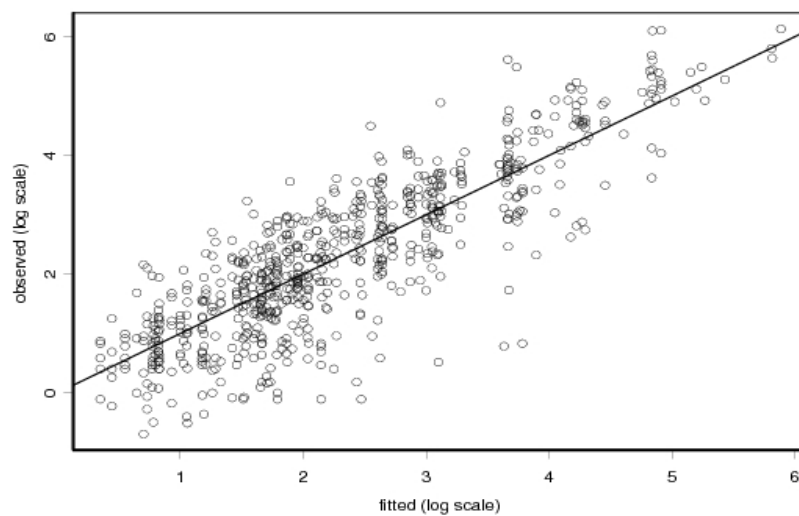
použita ve správné podobě) je velmi dobrá (přibližně lineární ale se směrnicemi i absolutními členy odlišnými v různých závodech). Nelinearita v grafu je předmětem diskusí expertů s tím, že není zcela jasné nakolik je dána odlišnostmi mezi jednotlivými závody (v grafu jsou shrnuty odhady ze všech 4 závodů jež se mohou poněkud lišit např. ve způsobu expozice) a nakolik může být ovlivněna i nelinearitou některých částí MVH kinetiky.



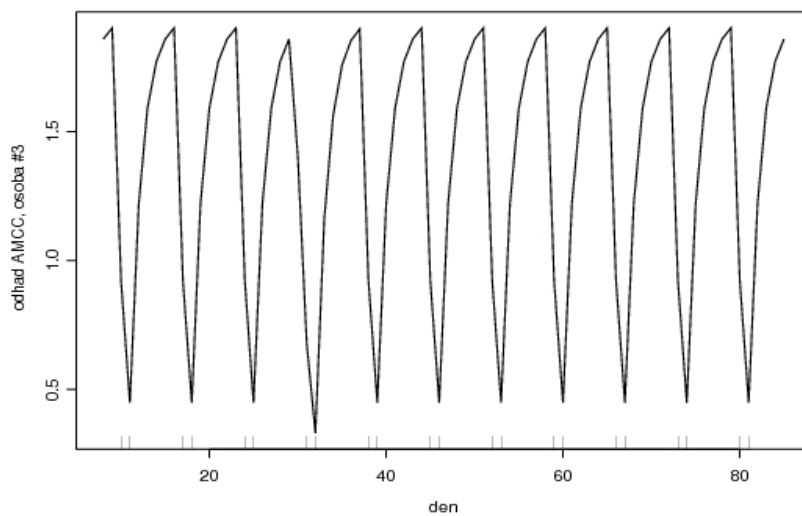
Obrázek 4: Náhodné (multiplikativní) efekty

4 Závěr

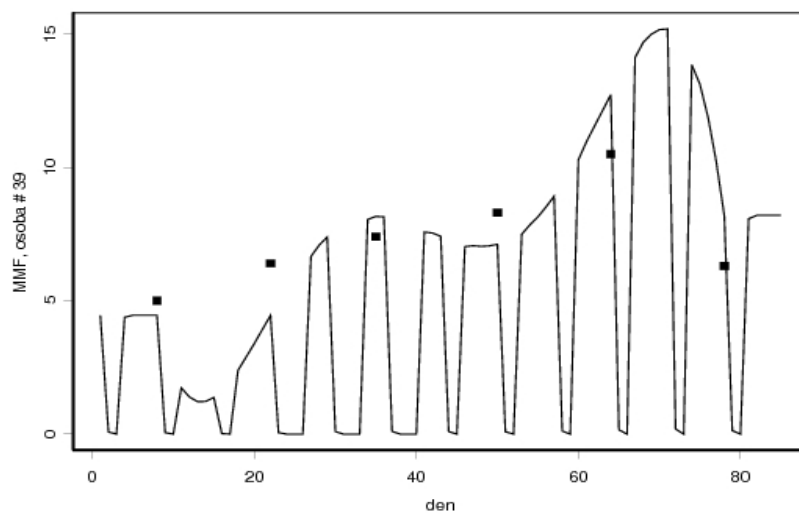
V článku je zformulován statistický model pro opakovaná měření různých markerů expozice DMF. Zohledňuje jak korelaci mezi opakovanými měřeními na stejném jedinci (ve stylu standardního lineárního smíšeného modelu), tak nestandardní komplikace dané integrovanou (konvoluční) povahou měření. Pozorovaná data odpovídají konvoluci základního expozičního procesu se známými jádry (odlišnými pro různé látky). Právě odlišnost tvaru a nosiče jader pro různé látky je podstatnou vlastností komplikující srovnatelnost jednotlivých indikátorů expozice. Srovnání lze bezpečně provést až po korekci na agregovaný charakter dat (efektivně dekonvoluci).



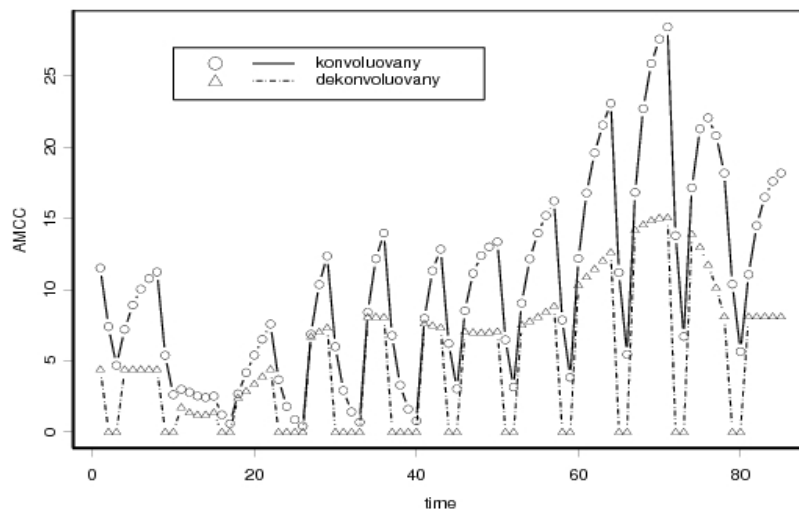
Obrázek 5: Proložení modelu daty



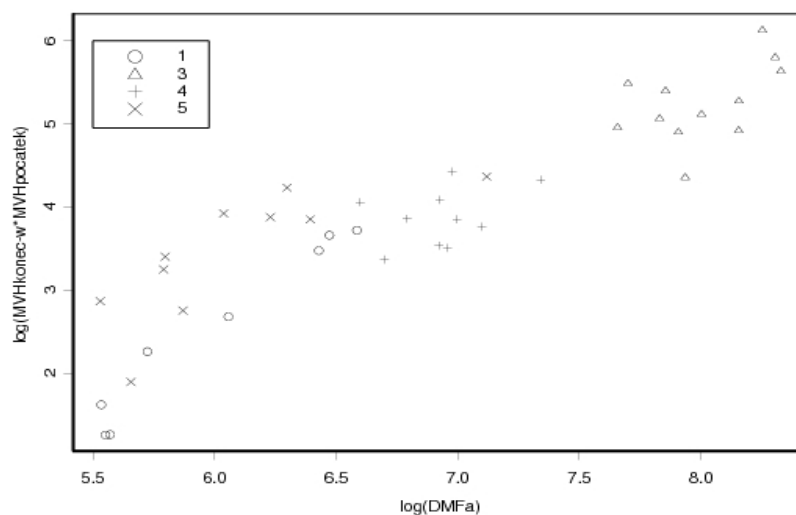
Obrázek 6: Individuální odhad AMCC



Obrázek 7: Individuální odhad MMF



Obrázek 8: Odhad původní a dekonvoluované veličiny



Obrázek 9: Kompozitní indikátor a MVH

References

- [1] Brabec,M.(2000): Modelování růstových křivek. Sborník prací 11. letní školy JČMF. ISBN 80-7015-792-5
- [2] Brabec,M.-Brabec,T.-Čermáková,J.-Konár,O.-Malý,M.-Naxerová,O.-Matějovic,J.-Pelikán,E.-Šimůnek,M.-Vondráček,J.(2005): Matematické modelování spotřeby zemního plynu zákazníků bez průběhového měření. *Plyn*, 2, 34-37
- [3] Carroll,R.J.-Ruppert,D.(1988): Transformation and weighting in regression. Chapman and Hall. New York.
- [4] Davidian,M.-Giltinan,D.M.(1995): Nonlinear models for repeated measurement data. Chapman and Hall. London.
- [5] Diggle,P.J.-Liang,K.-Zeger,S.L. (1994): Analysis of longitudinal data. Oxford University Press. Oxford.
- [6] Mráz,J.-Nohová,H.(1992): Absorption, metabolism and elimination of N,N-dimethyl-formamide in humans. *Int. Arch. Occup. Environ. Health*, 64, 85-92.
- [7] Mráz,J.-Dušková,Š.-Gálová,E.-Nohová,H.-Brabec,M.(2002): Biological monitoring of N,N-dimethylformamide. Reference value for N-methylcarbamoyl adduct at the N-terminal valine of globin as biomarker

of chronic occupational exposure. *Int. Arch. Occup. Environ. Health* 75 (Suppl.), S93-S96.

- [8] Pinheiro, J.C.-Bates, D.M. (2000): *Mixed-effects models in S and S-Plus*. Springer. New York.



Osobnostní a sociální determinanty hodnocení prestiže učitelského povolání u středoškoláků

MARIE BUDÍKOVÁ¹ A JIŘÍ DAN²

¹ *Katedra aplikované matematiky, Přírodovědecká fakulta Masarykovy univerzity, Janáčkovo náměstí 2a, 602 00 Brno
e-mail: budikova@math.muni.cz*

² *Akademické psychodiagnostické centrum Masarykovy univerzity, Žerotínovo náměstí 9, 601 77 Brno
e-mail: dan@rect.muni.cz*

Abstrakt: V příspěvku jsou zkoumány faktory, které ovlivňují hodnocení prestiže učitelského povolání studenty středních škol. Rozdíly ve středních hodnotách hodnocení prestiže učitelského povolání v jednotlivých skupinách studentů jsou posuzovány pomocí dvouvýběrového t -testu (s uvedením síly testu) a velikost vlivu skupiny na hodnocení prestiže je vyjádřena Cohenovým koeficientem.

Klíčová slova: Prestiž učitelského povolání, síla testu, Cohenův koeficient

1 Úvod

Od roku 2002 se na některých fakultách MU v Brně používá v přijímacím řízení TSP. Tento test zkoumá schopnosti uchazeče úspěšně studovat na MU. Je tvořen 80 položkami, které jsou seřazeny do 8 subtestů po 10 položkách. Každý subtest je zaměřen na jinou kvalitu lidského myšlení, a to na tzv. myšlení verbální, numerické, symbolické, analytické, kritické a vědecké a vedle toho i na prostorovou představivost a schopnost úsudků.

Při tvorbě TSP je zapotřebí ověřit kvalitu jednotlivých položek pomocí pilotní studie. Položky, které nevyhovují psychometrickým kritériím, nejsou do finální verze TSP zařazeny. Na přelomu let 2004 - 2005 proběhla pilotní studie, jíž se zúčastnilo 1100 studentů třetích ročníků čtyřletých gymnázií, septim osmiletých gymnázií a kvint šestiletých gymnázií. V rámci pilotního projektu studenti nejenom řešili TSP, ale vyplnili též škálu prestiže povolání zjišťující, jak respondent hodnotí prestiž učitelského povolání a dotazník týkající se různých osobnostních a sociálních charakteristik.

Statistická analýza datového souboru, který vznikl na základě uvedeného pilotního projektu, umožní posoudit, jak souvisí sledované faktory s hodnocením prestiže učitelského povolání. Samozřejmě by bylo možné rovněž zkoumat vztah mezi výkony v TSP resp. výkony v jednotlivých subtestech TSP a hodnocením prestiže, avšak touto problematikou se zde zabývat nebudeme.

2 Popis výzkumu

2.1 Informace o pilotním projektu

Pilotní projekt proběhl v prosinci 2004 a lednu 2005 ve 22 třídách 10 státních gymnázií v kraji Vysočina, kraji Pardubickém, Královehradeckém, Zlínském, Moravskoslezském a Jihomoravském a zúčastnilo se jej 1100 studentů. Pro účely tohoto sdělení jsme zpracovali datový soubor s údaji o 483 studentech, u nichž jsme mj. zjišťovali postavení studentů z hlediska prospěchu v rámci jedné třídy. 187 studentů navštěvovalo 3. ročník čtyřletého gymnázia, 277 septimu osmiletého gymnázia a 19 kvintu šestiletého gymnázia.

Hodnocení prestiže učitelského povolání jednotlivým studentem bylo zkoumáno pomocí škály, která obsahuje 26 dvojic povolání, přičemž jedním členem dvojice byl vždy učitel a druhým členem pak další povolání, např. programátor, ministr, soukromý zemědělec atd. Student v každé dvojici volí to povolání, které považuje za prestižnější. Pokud zvolí učitele, je příslušné dvojici povolání přiřazena jednička, jinak nula. Prestiž učitelského povolání tedy může nabývat hodnot od 0 do 26. Seznam povolání byl převzat z Centra pro výzkum veřejného mínění, které se výzkumy prestiže povolání zabývá již dlouhou řadu let.

Anamnestický dotazník obsahoval otázky, které se týkaly pohlaví studenta, jeho věku, prospěchu, velikosti sídla trvalého bydliště, zda student pochází z rodiny, kde aspoň jeden z rodičů učí na základní či střední škole a zda má zájem o získání vzdělání pro učitelské povolání.

2.2 Popis proměnných

V datovém souboru budeme pracovat s následujícími proměnnými:

SEX - specifikuje pohlaví studenta (varianta 1 - muž, varianta 2 - žena).

SÍDLO - udává velikost sídla trvalého bydliště (varianta 1 - sídlo do 8000 obyvatel, varianta 2 - sídlo nad 8000 obyvatel).

PROSPĚCH - odlišuje "výborné" studenty od ostatních studentů (varianta 1 - student je jinými studenty označen za výborného studenta a sám s tímto zařazením souhlasí, přičemž souhlasili všichni. V jednotlivých třídách bylo takových studentů od dvou do pěti. Varianta 2 - ostatní studenti. V dalším textu zkracujeme na označení "jedničkář", "není jedničkář").

UČ-RODINA - specifikuje, zda student pochází z učitelské rodiny (varianta 1 - aspoň jeden z rodičů je učitelem na ZŠ nebo SŠ, varianta 2 - žádný z rodičů není učitelem na ZŠ nebo SŠ).

UČ-POVOLÁNÍ - informuje o snaze studenta získat vzdělání pro učitelské po-

volání (varianta 1 - chce se stát učitelem, varianta 2 - nechce se stát učitelem).
PRESTIŽ - prestiž učitelského povolání (hodnoty 0 až 26).

2.3 Prestiž povolání

Prestiž povolání je jednou z dimenzí sociální stratifikace společnosti a lze ji chápat jako vyjádření vážnosti či úcty, které se ve společnosti těší příslušníci různých profesí. Prestiž se váže k pozici, kterou daná osoba ve společnosti zastává, tudíž hodnocení prestiže povolání se nevztahuje k dané osobě, ale k její pozici ve společnosti. Seznam povolání, která se ve výzkumech používají, není přesně daný a během doby se obměňuje (tak, aby v něm byla zachycena povolání odpovídající dané době). Centrum pro výzkum veřejného mínění (dříve Institut pro výzkum veřejného mínění) opakovaně provádí výzkumy na toto téma.

3 Základní statistické zpracování datového souboru

3.1 Četnostní tabulky alternativních proměnných

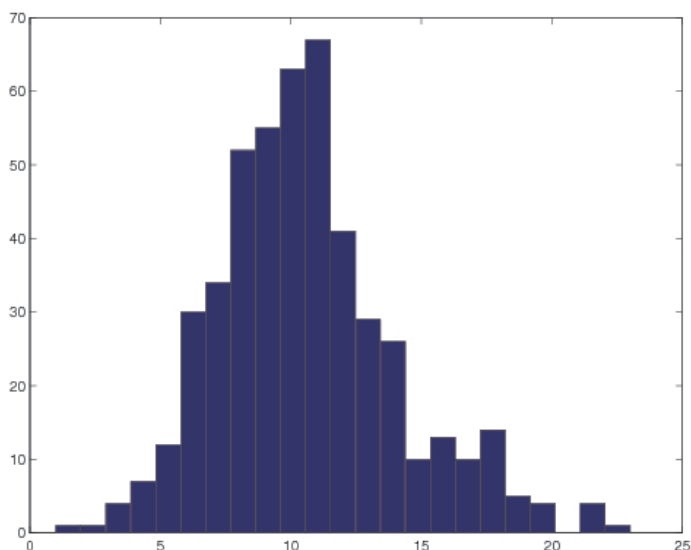
První informace nám poskytnou tabulky četností jednotlivých alternativních proměnných získané pomocí programu STATISTICA z výše popsaného datového souboru.

SEX	absolutní četnost	relativní četnost
muž	175	36,2
žena	308	63,8
SÍDLO	absolutní četnost	relativní četnost
do 8000	220	45,6
nad 8000	263	54,4
PROSPĚCH	absolutní četnost	relativní četnost
je jedničkář	83	17,2
není jedničkář	400	82,8
UČ-RODINA	absolutní četnost	relativní četnost
ano	86	17,8
ne	397	82,2
UČ-POVOLÁNÍ	absolutní četnost	relativní četnost
ano	132	27,3
ne	351	72,7

Tabulka 1: Četnosti alternativních proměnných

3.2 Histogram a číselné charakteristiky proměnné PRESTIŽ

Vlastnosti proměnné PRESTIŽ popíšeme pomocí histogramu a číselných charakteristik:



Obr. 1: Histogram proměnné PRESTIŽ

Proměnná	Popisné statistiky (prestiz-TSP)						
	Průměr	Medián	Min.	Max.	Sm.odch.	Šířkost	Špičatost
PRESTIŽ	10,55	10	1	23	3,63	0,64	0,63

Tabulka 2: Číselné charakteristiky proměnné PRESTIŽ

Bodové vyjádření hodnocení prestiže učitelského povolání studenty se pohybuje v rozmezí 1 až 23 s průměrem 10,6 a směrodatnou odchylkou 3,6. Rozložení proměnné PRESTIŽ je poněkud odlišné od normálního rozložení, je kladně sešikmené a má větší špičatost než normální rozložení. Shapirův - Wilksův test normality poskytuje testovou statistiku $W = 0,9665$ a p-hodnotu blízkou nule.

3.3 Závislost alternativních proměnných

K dispozici je 5 alternativních proměnných, a to SEX, SÍDLO, PROSPĚCH, UČ-RODINA, UČ-POVOLÁNÍ. Na hladině významnosti 0,05 testujeme pomocí jednostranného Fisherova exaktního testu hypotézu, že libovolné dvě proměnné jsou stochasticky nezávislé. Po provedení těchto 10 testů se ukázalo, že hypotéza o nezávislosti se zamítá u dvojic proměnných SEX a PROSPĚCH, SEX a UČ-POVOLÁNÍ a SÍDLO a UČ-POVOLÁNÍ.

SEX a PROSPĚCH: mezi muži je 9,7% jedničkářů (17 ze 175), mezi ženami je 21,4% jedničkářek (66 z 308). Fisherův test poskytl p-hodnotu 0,00102.

SEX a UČ-POVOLÁNÍ: mezi muži je 16,6% těch, kteří chtějí být učiteli (29 ze 175), mezi ženami je 33,4% těch, které chtějí být učitelkami (103 z 308). Fisherův test poskytl p-hodnotu 0,00005.

SÍDLO a UČ-POVOLÁNÍ: mezi studenty ze sídel do 8000 obyvatel je 34,6% těch, kteří chtějí být učiteli (76 z 220), mezi studenty ze sídel nad 8000 obyvatel je 21,3% těch, kteří chtějí být učiteli (56 z 263). Fisherův test poskytl p-hodnotu 0,00144.

4 Vliv jednotlivých alternativních proměnných na hodnocení prestiže učitelského povolání

Každá z uvažovaných pěti alternativních proměnných rozdělí zkoumaný soubor studentů na dvě nezávislé skupiny, např. na muže a ženy, na studenty z malých a větších sídel či na studenty z učitelských a neučitelských rodin. Na hladině významnosti 0,05 budeme testovat hypotézu, že střední hodnota proměnné PRESTIŽ je v daných dvou skupinách stejná.

K tomuto účelu použijeme dvouvýběrové t-testy, i když proměnná PRESTIŽ vykazuje určité odchylky od normálního rozložení. Vzhledem k značně velkým rozsahům skupin studentů však lze použití parametrického testu považovat za adekvátní.

Kromě obvyklého výpočtu realizace testové statistiky a příslušné p-hodnoty zjistíme též sílu dvouvýběrového t-testu a vliv skupiny na variabilitu hodnot závisle proměnné veličiny posoudíme pomocí Cohenova koeficientu.

4.1 Síla dvouvýběrového t-testu

Nechť X_{11}, \dots, X_{1n_1} je náhodný výběr z rozložení $N(\mu_1, \sigma^2)$ a X_{21}, \dots, X_{2n_2} je na něm nezávislý náhodný výběr rozložení $N(\mu_2, \sigma^2)$, přičemž $n_1 \geq 2$ a $n_2 \geq 2$ a σ^2 neznáme. Označme

$$M_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i}, M_2 = \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i} \text{ výběrové průměry,}$$

$$S_1^2 = \frac{1}{n_1-1} \sum_{i=1}^{n_1} (X_{1i} - M_1)^2, S_2^2 = \frac{1}{n_2-1} \sum_{i=1}^{n_2} (X_{2i} - M_2)^2 \text{ výběrové rozptyly a}$$

$$S_*^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} \text{ vážený průměr výběrových rozptylů.}$$

Nechť c je konstanta.

Test $H_0 : \mu_1 - \mu_2 = c$ proti $H_1 : \mu_1 - \mu_2 \neq c$ se nazývá dvouvýběrový t-test.

Při provádění tohoto testu se používá statistika

$$T = \frac{(M_1 - M_2) - (\mu_1 - \mu_2)}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

kteřá se řídí Studentovým rozložením $t(n_1 + n_2 - 2)$.

Test na hladině významnosti α lze provést buď pomocí $100(1 - \alpha)\%$ intervalu

spolehlivosti pro parametrickou funkci $\mu_1 - \mu_2$ nebo pomocí kritického oboru nebo výpočtem p-hodnoty a jejím porovnáním s hladinou významnosti α .

$100(1 - \alpha)\%$ interval spolehlivosti pro parametrickou funkci $\mu_1 - \mu_2$ má meze:

$$D = M_1 - M_2 - S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2),$$

$$H = M_1 - M_2 + S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{1-\alpha/2}(n_1 + n_2 - 2).$$

Silofunkce tohoto testu je dána vztahem:

$$\forall c \in R : \gamma(c) = P(c \notin (D, H)),$$

tedy pro dané reálné číslo c vyjadřuje pravděpodobnost, s jakou test vypoví, že nulová hypotéza neplatí.

Po určitých úpravách dospějeme k vyjádření:

$$\begin{aligned} \forall c \in R : \gamma(c) &= \\ &= 2 - \Phi\left(t_{1-\alpha/2}(n_1 + n_2 - 2) - \frac{c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right) - \\ &\quad - \Phi\left(t_{1-\alpha/2}(n_1 + n_2 - 2) + \frac{c}{S_* \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}\right), \end{aligned}$$

kde $\Phi(x)$ je distribuční funkce Studentova rozložení $t(n_1 + n_2 - 2)$. Při konkrétním výpočtu síly testu nahradíme číslo c rozdílem realizací výběrových průměrů. Síla testu by se měla pohybovat nad 0,8.

4.2 Cohenův koeficient

Předpokládáme, že sledujeme hodnoty nějaké náhodné veličiny ve dvou nezávislých skupinách objektů o rozsazích n_1 a n_2 . Označme m_1, m_2 realizace výběrových průměrů hodnot dané veličiny v těchto dvou skupinách, s_1^2, s_2^2 realizace výběrových rozptylů a

$$s_*^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

realizaci váženého průměru výběrových rozptylů.

Cohenův koeficient d pak vypočteme podle vzorce:

$$d = \frac{|m_1 - m_2|}{s_*}.$$

Tento koeficient slouží k posouzení velikosti rozdílu průměrů, který je standardizován pomocí odmocniny z váženého průměru výběrových rozptylů. Jedná se o posouzení vlivu zařazení do skupiny na rozdílnost hodnot sledované náhodné

veličiny v daných dvou skupinách. Velikost tohoto vlivu hodnotíme podle tabulky 3:

Hodnota d	vliv
aspoň 0,8	velký
mezi 0,5 až 0,8	střední
mezi 0,2 až 0,5	malý
pod 0,2	zanedbatelný

Tabulka 3: Význam Cohenova koeficientu

4.3 Výsledky dvouvýběrových t-testů

Po provedení pěti dvouvýběrových t-testů se ukázalo, že nulová hypotéza o shodě středních hodnot proměnné PRESTIŽ je zamítnuta v případě alternativních proměnných SEX a UČ-POVOLÁNÍ.

Výsledky dvouvýběrového t-testu pro muže a ženy jsou uvedeny v tabulce 4:

	n	m	s	F	p	t	sv	p
muž	175	9,40	3,58	1,0179	0,9050	-2,9992	481	0,0028
žena	308	10,93	3,61					

Tabulka 4: Dvouvýběrový t-test pro PRESTIŽ ve skupinách mužů a žen

Význam symbolů:

n... rozsah souboru

m... průměr

s... směrodatná odchylka

F... hodnota testové statistiky pro test shody rozptylů

p... p-hodnota

t... hodnota testové statistiky pro test shody středních hodnot

sv... počet stupňů volnosti

Lze tedy konstatovat, že s rizikem omylu maximálně 5% je učitelské povolání pro ženy prestižnější než pro muže. Síla tohoto dvouvýběrového t-testu je 0,85, tudíž neplatnou nulovou hypotézu test zamítá s pravděpodobností aspoň 0,85. Cohenův koeficient nabývá hodnoty $d = 0,29$, což svědčí o malém vlivu pohlaví na hodnocení prestiže učitelského povolání.

Výsledky dvouvýběrového t-testu pro studenty, kteří se chtějí resp. nechtějí stát učiteli, jsou uvedeny tabulce 5:

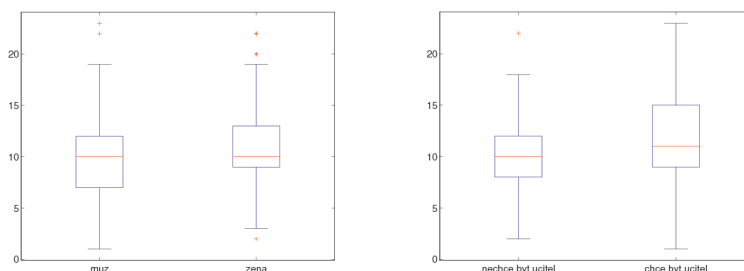
učitel	n	m	s	F	p	t	sv	p
chce být	351	9,94	3,10	1,9844	0,0000	-5,3974	182,9	0,0000
nechce být	132	12,18	4,37					

Tabulka 5: Dvouvýběrový t-test pro PRESTIŽ (se separovanými odhady rozptylů) ve skupinách studentů, kteří chtějí resp. nechtějí být učiteli

S rizikem omylu maximálně 5% můžeme říci, že učitelské povolání je pro studenty, kteří se chtějí stát učiteli, prestižnější než pro ostatní studenty. Síla tohoto dvouvýběrového t-testu je 0,99, tudíž neplatnou nulovou hypotézu test zamítá s pravděpodobností aspoň 0,99.

Cohenův koeficient nabývá hodnoty $d = 0,51$, což svědčí o středně silném vlivu snahy o získání vzdělání pro učitelské povolání na hodnocení prestiže učitelského povolání.

Výpočty ještě doplníme krabicovými diagramy proměnné PRESTIŽ (je použita varianta krabicového diagramu s průměrem, směrodatnou odchylkou, minimem a maximem).



Obr. 2: Krabicové diagramy proměnné PRESTIŽ

4.4 Vliv dvojic alternativních proměnných na hodnocení prestiže učitelského povolání

Doposud jsme se zabývali problémem, zda existuje rozdíl v hodnocení prestiže učitelského povolání u dvou skupin studentů, kteří byli rozděleni podle pohlaví, podle velikosti sídla trvalého bydliště, podle prospěchu, podle původu z učitelské či neučitelské rodiny a podle snahy o získání vzdělání pro učitelské povolání. Nyní budeme zkoumat současný vliv dvojice alternativních proměnných na hodnocení prestiže. K řešení tohoto problému použijeme analýzu rozptylu dvojného třídění s interakcí. Jedná se tedy o provedení deseti analýz, přičemž ve všech případech mají oba zkoumané faktory dvě úrovně. Průměry proměnné PRESTIŽ a rozsahy uvažovaných skupin jsou uvedeny v tabulce 6 spolu s p-hodnotami pro test hypotézy o nevýznamnosti interakcí dvojic faktorů:

SEX	SÍDLO	PRESTIŽ Průměr	N
muž	do 8000	9,64	70
muž	nad 8000	10,08	105
žena	do 8000	11,04	150
žena	nad 8000	10,82	158

$$p = 0,3428$$

Osobnostní a sociální determinanty hodnocení prestiže učitelského povolání 33

SEX	PROSPĚCH	PRESTIŽ Půměr	N
muž	je jedničkář	9,29	17
muž	není jedničkář	9,97	158
žena	je jedničkář	11,03	66
žena	není jedničkář	10,90	242

p = 0,4412

SEX	UČ-RODINA	PRESTIŽ Půměr	N
muž	je z učitelské rodiny	9,60	35
muž	není z učitelské rodiny	9,98	140
žena	je z učitelské rodiny	11,41	51
žena	není z učitelské rodiny	10,83	257

p = 0,2734

SEX	UČ-POVOLÁNÍ	PRESTIŽ Půměr	N
muž	nechce být učitelem	9,55	146
muž	chce být učitelem	11,66	29
žena	nechce být učitelem	10,22	205
žena	chce být učitelem	12,33	103

p = 1

SÍDLO	PROSPĚCH	PRESTIŽ Půměr	N
do 8000	je jedničkář	10,93	43
do 8000	není jedničkář	10,51	177
nad 8000	je jedničkář	10,40	40
nad 8000	není jedničkář	10,54	223

p = 0,5256

SÍDLO	UČ-RODINA	PRESTIŽ Půměr	N
do 8000	je z učitelské rodiny	10,76	38
do 8000	není z učitelské rodiny	10,56	182
nad 8000	je z učitelské rodiny	10,60	48
nad 8000	není z učitelské rodiny	10,50	215

p = 0,9079

SÍDLO	UČ-POVOLÁNÍ	PRESTIŽ Půměr	N
do 8000	nechce být učitelem	9,97	144
do 8000	chce být učitelem	11,79	76
nad 8000	nechce být učitelem	9,93	207
nad 8000	chce být učitelem	12,71	56

$p = 0,1835$

PROSPĚCH	UČ-RODINA	PRESTIŽ Půměr	N
je jedničkář	je z učitelské rodiny	11,82	11
je jedničkář	není z učitelské rodiny	10,50	72
není jedničkář	je z učitelské rodiny	10,51	75
není jedničkář	není z učitelské rodiny	10,54	325

$p = 0,2880$

PROSPĚCH	UČ-POVOLÁNÍ	PRESTIŽ Půměr	N
je jedničkář	nechce být učitel	9,93	61
je jedničkář	chce být učitel	12,73	22
není jedničkář	nechce být učitel	9,94	290
není jedničkář	chce být učitel	12,07	110

$p = 0,4864$

UČ-RODINA	UČ-POVOLÁNÍ	PRESTIŽ Půměr	N
je z učitelské rodiny	nechce být učitel	10,02	59
je z učitelské rodiny	chce být učitel	12,11	27
není z učitelské rodiny	nechce být učitel	9,93	292
není z učitelské rodiny	chce být učitel	12,20	105

$p = 0,8445$

Tabulka 6: Průměry proměnné PRESTIŽ a p-hodnoty pro test nevýznamnosti interakcí dvojic faktorů

Analýza rozptylu dvojného třídění nepřinesla žádné nové závažné výsledky. Interakce mezi faktory se ani v jednom případě neprokázaly významné na hladině významnosti 0,05. Z tabulky 6 vyplývá, že v průměru nejvyšší prestiž má učitelské povolání u výborných studentů, kteří se chtějí stát učiteli (těch je však pouze 22), naopak nejnižší u mužů s výborným prospěchem (takových je 17). Tento výsledek samozřejmě vede k závěru, že rozhodující vliv na vysokém hodnocení prestiže učitelského povolání budou mít ženy s výborným prospěchem,

které se chtějí stát učitelkami. Pokud tedy roztřídíme soubor podle tří faktorů - SEX, PROSPĚCH, UČ-POVOLÁNÍ a v uvažovaných osmi skupinách spočteme průměry, dostaneme výsledky uvedené v tabulce 7:

SEX	PROSPĚCH	UČ-POVOLÁNÍ	PRESTIŽ Průměr	N
muž	je jedničkář	nechce být učitel	9,27	15
muž	je jedničkář	chce být učitel	9,50	2
muž	není jedničkář	nechce být učitel	9,59	131
muž	není jedničkář	chce být učitel	11,81	158
žena	je jedničkář	nechce být učitel	10,15	46
žena	je jedničkář	chce být učitel	13,05	20
žena	není jedničkář	nechce být učitel	10,24	159
žena	není jedničkář	chce být učitel	12,16	83

Tabulka 7: Průměry proměnné PRESTIŽ ve skupinách studentů roztříděných podle faktorů SEX, PROSPĚCH, UČ-POVOLÁNÍ

Vidíme tedy, že muži s výborným prospěchem, kteří se nechtějí stát učiteli, hodnotí učitelské povolání velmi nízko, průměrné hodnocení prestiže je pouze 9,27. Naopak ženy - jedničkářky, které chtějí být učitelkami, dosahují v průměru hodnocení 13,05.

5 Závěr

Zabývali jsme se zpracováním údajů o 483 studentech předposledních ročníků 10 státních gymnázií, kteří se v prosinci 2004 a lednu 2005 zúčastnili pilotní studie nutné pro sestavení TSP. V rámci tohoto pilotního projektu studenti jednak řešili TSP a jednak hodnotili prestiž učitelského povolání ve srovnání s 26 jinými profesemi. Zajímaly nás vlivy osobnostních a sociálních faktorů na hodnocení prestiže učitelského povolání.

Sledovány byly tyto faktory: pohlaví studenta, velikost sídla trvalého bydliště (do 8000 obyvatel a nad 8000 obyvatel), prospěch (studenti - jedničkáři a ostatní studenti), původ z učitelské rodiny (tj. aspoň jeden z rodičů je učitelem na základní či střední škole) a snaha o získání vzdělání pro učitelské povolání.

Nejprve jsme zkoumali vzájemnou závislost sledovaných pěti faktorů. Jednostranným Fisherovým exaktním testem jsme na hladině významnosti 0,05 prokázali, že

- podíl výborných studentů mezi ženami je vyšší než mezi muži (21,4% oproti 9,7%)
- podíl studentů, kteří chtějí být učiteli, je mezi ženami vyšší než mezi muži (33,4% oproti 16,6%)
- podíl studentů, kteří chtějí být učiteli, je mezi studenty ze sídel do 8000 obyvatel vyšší než mezi studenty ze sídel nad 8000 obyvatel (34,6% oproti 21,3%).

Vliv jednotlivých faktorů na hodnocení prestiže učitelského povolání byl proveden pomocí dvouvýběrových t-testů s následným zjištěním síly testu a velikost vlivu skupiny byla posouzena Cohenovým koeficientem d . Na hladině významnosti 0,05 jsme prokázali, že

- a) střední hodnota prestiže učitelského povolání je různá pro muže a pro ženy (9,90 oproti 10,93), síla testu je vyhovující, nabývá hodnoty 0,85, avšak vliv pohlaví na hodnocení prestiže je malý ($d = 0,29$)
- b) střední hodnota prestiže učitelského povolání je různá pro studenty, kteří se chtějí resp. nechtějí stát učiteli (12,18 oproti 9,94), síla testu je vyhovující, nabývá hodnoty 0,99. Vliv snahy o získání vzdělání pro učitelské povolání na hodnocení prestiže je středně silný ($d = 0,51$).

Zkoumáme-li pomocí analýzy rozptylu dvojného třídění vliv jakýchkoliv kombinací dvou faktorů na hodnocení prestiže, nedostaneme již žádné psychologicky zajímavé výsledky, mezi dvojicemi faktorů nejsou žádné interakce. Za pozornost však stojí, že vůbec nejnižší prestiž má učitelské povolání u mužů s výborným prospěchem, kteří se nechtějí stát učiteli (průměrné hodnocení je 9,27), zatímco nejvyšší hodnocení se vyskytuje u žen s výborným prospěchem, které se chtějí stát učitelkami (průměrné hodnocení je 13,05).

References

- [1] Hendl, Jan. Přehled statistických metod zpracování dat. Analýza a metaanalýza dat. 1. vydání. 2004.
- [2] Osecký, Pavel. Pravděpodobnost a statistika. Univerzita J. E. Purkyně v Brně 1988.
- [3] STATISTICA for Windows. Statsoft, Inc. 2000

Analýza vydychovaných plynov

Štatistické charakteristiky vybraných plynných komponentov vydýchnutého vzduchu meraného pomocou PTR-MS

KATARÍNA CIMERMANOVÁ

*Ústav merania, Oddelenie teoretických metód, Slovenská akadémia vied,
Dúbravská cesta 9, 841 04 Bratislava
e-mail: katarina.cimermanova@gmail.com*

Abstrakt: V článku sa zaoberáme testovaním štatistických hypotéz, či zložky náhodného vektora

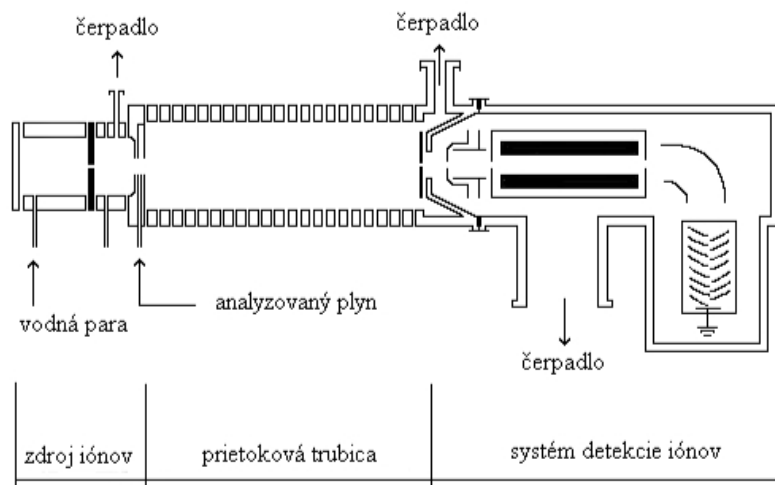
$$\mathbf{X}_i = (x_1, x_2, \dots, x_n),$$

kde $i = 1, \dots, N$, N je počet subjektov v databáze a n je počet zložiek vektora, pochádzajú z normálneho jednorozmerného rozdelenia. Ďalej testujeme hypotézu, či náhodný vektor \mathbf{X}_i pochádza z n -rozmerného normálneho rozdelenia. Nakoľko náhodná veličina po transformovaní môže mať viac normálne rozdelenie testujeme hypotézy aj pre náhodný vektor

$$\mathbf{Y}_i = \log_{10}(\mathbf{X}_i).$$

1 Formulácia problému

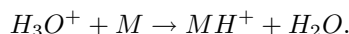
Rakovina je jednou z najčastejších príčin úmrtí v Európe a v západnom svete. Príčinou je neskorá diagnóza tohto ochorenia, nakoľko dostupné diagnostické metódy nie sú dostatočne rýchle a presné. Horúcim tipom na detekciu niektorých typov rakoviny je analýza vydychovaných plynov [1]. Vo vydychovanom plyne sa dá detekovať 300 až 3000 rôznych chemických látok. Význam pri diagnóze rakoviny majú však len niektoré molekulové hmotnosti [1], [8], [9] a [10], ktoré boli vybrané na základe dlhodobých pozorovaní a znalostí metabolizmu. Na meranie plynných zložiek vzduchu sa používajú nové experimentálne metódy ako sú plynová chromatografia s hmotnostnou spektrometriou (GC-MS), hmotnostná spektrometria s protónovou prenosovou reakciou (PTR-MS), hmotnostná spektrometria s tokom selektívnych iónov v tube (SIFT-MS), laserová



Obrázek 1: Principiálna schéma metódy hmotnostnej spektrometrie s protónovou prenosovou reakciou (PTR-MS).

spektrometria a spektrometria pohyblivosti iónov [1], [2], [5]. Princíp metód je založený na detekovaní zložiek plynu a vyjadrení ich koncentrácie.

Spracovávané dáta boli získané metódou hmotnostnej spektrometrie s protónovou prenosovou reakciou (PTR-MS, *Proton transfer reaction mass spectrometry*). Na obrázoku ?? sa nachádza principiálna schéma metódy PTR-MS [2]. Metóda sa používa na priame monitorovanie prchavých organických zložiek vzduchu. Ide o chemickú reakciu zložky M (napr. formaldehyd, metanol, acetón, izoprén, butanón, benzén, toluén, xyléne, o-toluidín, heptanón, ...) s protonizovanou vodou H_3O^+ ako prekursorom v prietokovej trubici (*drift chamber*)



Produkty MH^+ (prekursor odovzdá protón H^+ molekule plynu M) sa separujú na základe molekulových hmotností ($Mr(M) + 1$) v elektromagnetickom poli. V niektorých prípadoch dochádza k tomu, že dve rôzne molekuly majú tú istú hmotnosť. V takomto prípade sú tieto molekuly detekované ako jedna zložka - masa. Relatívne početnosti prchavých organických zložiek získaných PTR-MS sú ďalej kvantifikované na základe kinetiky protónovej reakcie

$$c(M) = |MH^+|/|H_3O^+| * k_M * t_{drift},$$

kde $|MH^+|$ je počet detekovaných iónov meranej zložky, $|H_3O^+|$ je množstvo molekúl protonizovanej vody. Kritickými parametrami sú k_M koeficient rýchlosti reakcie zložky M s H_3O^+ a čas t_{drift} , za ktorý pretekajú ióny prietokovou trubicou. Ďalšou zložitosťou procesu merania je, že niektoré zložky

môžu byť rozpustné vo vode, takže časť molekúl sa naviaže na molekuly vody (H_2O , $M_r(H_2O) = 18$) a nameraná hmotnosť takýchto iónov je iná, ako očakávaná hmotnosť ($M_r(M) + 1$). Zdokonaľovanie procesu merania a analýzy dát pomocou PTR-MS (kvalita a spoľahlivosť získaných meraní) je predmetom ďalšieho výskumu v tejto oblasti.

Štatistické aplikácie na analýzu prchavých organických zložiek plynu sa postupne vyvíjajú a zhrňajú pre vytvorenie BAMOD štatistického balíku (*BAMOD Statistical Toolbox*) v Matlabovom prostredí [6]. Ide o mnohé zatiaľ nepredprogramovateľné algoritmy, ktoré spracovávajú, filtrujú a separujú dáta. V ďalšej časti balíku sú algoritmy na popis databázy, ako základné štatistické charakteristiky, testovanie zhody rozdelenia, porovnávanie rozdelení a štatistických charakteristík a klasifikačné metódy. Cieľom je navrhnúť matematickú alebo štatistickú metódu na rozoznanie potenciálnych pacientov s rakovinou pľúc prípadne pažeráka.

Spracovávaná databáza vychádzajúca z pilotnej štúdie sa skladá z 368 meraní pre 106 subjektov. Subjekty sa delia do dvoch skupín a to skupina 17-tich pacientov s diagnostikovanou rakovinou pľúc alebo pažeráka ($N_1 = 33$ pozorovaní) a skupina 89-tich zdravých dobrovoľníkov ($N_2 = 335$ pozorovaní). Pre každý subjekt sú v databáze dodatočné informácie o pohlaví, veku a fajčiarskych návykoch. Na ich základe sa subjekty delia do podskupín (viď tabuľka 1).

Pre každý objekt je meraných 12 hodnôt, ktoré reprezentujú koncentráciu prchavých organických zložiek. Ide o ióny s molekulovou hmotnosťou m31, m33, m42, m59, m63, m69, m73, m79, m93, m107, m108 a m115. Koncentrácia jednotlivých zložiek je veľmi rozdielná a veľmi premenlivá medzi subjektmi aj v závislosti od ďalších faktorov (napr. od veku, prípadného fajčenia, vplyvu vonkajšieho prostredia, atď.) [11]. Hodnoty koncentrácie prchavých organických zložiek plynu sú udávané v jednotke ppb (*particles-per-billion*, počet molekúl meraného plynu na miliardu molekúl ako celok).

Mnohé z týchto prchavých organických zložiek sa vyskytujú v dychu v takých nízkych úrovniach, že ich nie je možné kvantifikovať dostatočne presne ani pomocou súčasných techník. Predpokladaný detekčný limit meracieho zariadenia je rádovo $10^{-1} - 10^{-2}$ ppb. V databáze sa nachádzajú hodnoty koncentrácie organických zložiek plynu, ktoré sú nižšie ako predpokladaný detekčný limit meracieho zariadenia. Preto sa pri spracovávaní tieto hodnoty vyradia, alebo sa nahradia výrazom *NaN* (*Not-a-Number*, hodnota, ktorú nie je možné vyčísliť). Takéto hodnoty sa vyskytli pri meraní prchavých organických zložiek vydychovaného plynu s molekulovou hmotnosťou m79 (1 meranie), m107 (12 meraní), m108 (13 meraní) a m115 (16 meraní).

Podľa skupiny vedcov z Innsbrucku sú z diagnostického hľadiska významné len tie koncentrácie prchavých organických zložiek vydychovaného plynu, pre ktoré platí, že pomer vdychovaného a vydychovaného vzduchu je mimo rozmedzia intervalu $\langle 0.8, 1.25 \rangle$ [1] (pre zlogaritmované dáta je to interval $\langle -0.0969, 0.0969 \rangle$). V tabuľke 2 je množstvo diagnosticky významných hodnôt pre jednotlivé prchavé organické zložky vydychovaného plynu z databázy.

Na obrázku 2 sú znázornené krabicové diagramy (*boxplot*, poskytuje informáciu o maximálnej, minimálnej hodnote, mediáne a kvartiloch z výberu náhodnej

<i>pohlavie</i>	<i>fajč. návyky</i>	<i>pacienti</i>				<i>všetci</i>
		<i>vek < 21</i>	$21 \leq \text{vek} < 45$	$45 \leq 60$	$60 \leq \text{vek}$	
<i>všetci</i>	<i>všetci</i>	0	1	14	18	33
	<i>fajčiari</i>	0	1	6	8	15
	<i>nefajčiari</i>	0	0	8	10	18
<i>ženy</i>	<i>všetky</i>	0	0	5	4	9
	<i>fajčiarky</i>	0	0	0	0	0
	<i>nefajčiarky</i>	0	0	5	4	9
<i>muži</i>	<i>všetci</i>	0	1	9	14	24
	<i>fajčiari</i>	0	1	6	8	15
	<i>nefajčiari</i>	0	0	3	6	9

<i>pohlavie</i>	<i>fajč. návyky</i>	<i>zdraví dobrovoľníci</i>				<i>všetci</i>
		<i>vek < 21</i>	$21 \leq \text{vek} < 45$	$45 \leq 60$	$60 \leq \text{vek}$	
<i>všetci</i>	<i>všetci</i>	28	201	59	47	335
	<i>fajčiari</i>	0	60	5	4	69
	<i>nefajčiari</i>	28	131	49	38	246
<i>ženy</i>	<i>všetky</i>	18	107	32	35	192
	<i>fajčiarky</i>	0	29	5	0	34
	<i>nefajčiarky</i>	18	78	27	30	153
<i>muži</i>	<i>všetci</i>	10	94	27	12	143
	<i>fajčiari</i>	0	31	0	4	35
	<i>nefajčiari</i>	10	53	22	8	93

Tabulka 1: Počet meraní v jednotlivých podskupinách rozdelených podľa zdravotného stavu, pohlavia a fajčiarskych návykov.

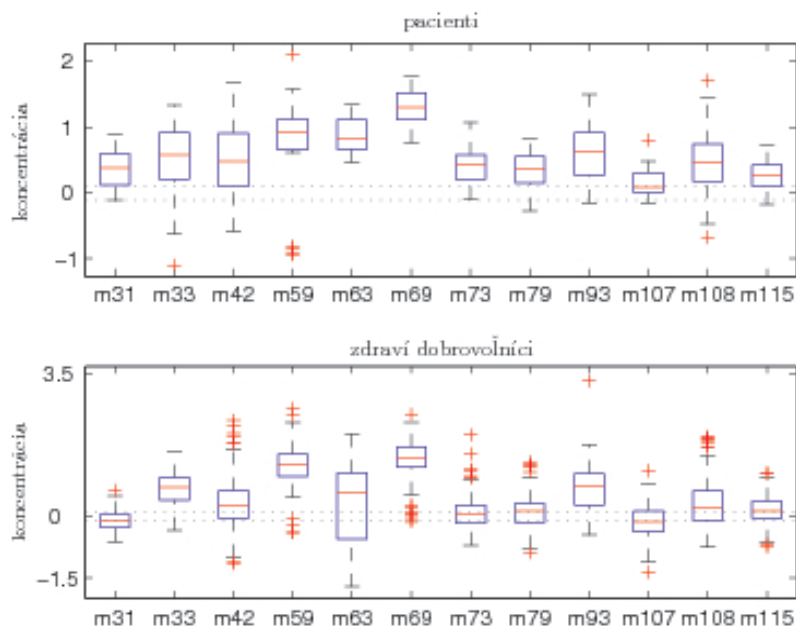
veličiny) pre pomer koncentrácií prchavých organických zložiek vdychovaného a vydychovaného plynu s vyznačenými hranicami významnosti hodnôt.

<i>prchavé organické zložky plynu</i>											
<i>m31</i>	<i>m33</i>	<i>m42</i>	<i>m59</i>	<i>m63</i>	<i>m69</i>	<i>m73</i>	<i>m79</i>	<i>m93</i>	<i>m107</i>	<i>m108</i>	<i>m115</i>
250	328	304	342	336	340	263	284	322	271	291	255

Tabulka 2: Počet diagnosticky významných hodnôt pre jednotlivé prchavé organické zložky z celkového počtu $N = 335$.

2 Testy dobrej zhody

Pri mnohých štatistických aplikáciách sa požaduje vedieť, či dáta pochádzajú z požadovaného rozdelenia. V tejto kapitole testujeme hypotézu, či náhodná



Obrázek 2: Krabicové diagramy pre pomer prchavých organických zložiek zlogaritmovaného vdychovaného a vydychovaného plynu s vyznačenými intervalmi významnosti hodnôt z diagnostického hľadiska podľa vedcov z Innsbrucku, kde mimo intervalu sa nachádzajú významné hodnoty.

veličina X a náhodný vektor \mathbf{X} prípadne náhodná veličina $Y = \log_{10}(X)$ a náhodný vektor $\mathbf{Y} = \log_{10}(\mathbf{X})$ pochádzajú z normálneho rozdelenia [3], [4].

2.1 Test na overenie normality jednorozmerného rozdelenia

Základným popisom rozdelenia náhodnej veličiny X je distribučná funkcia

$$F(x) = P\{X \leq x\},$$

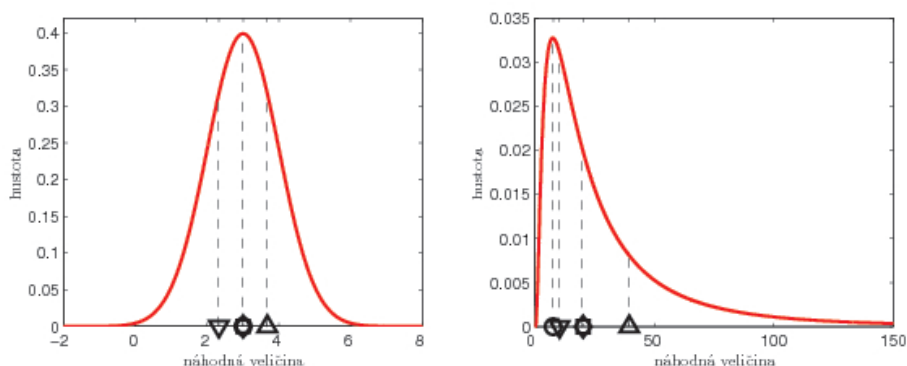
kde $P\{X \leq x\}$ znamená pravdepodobnosť udalosti, že náhodná veličina X nadobúda hodnoty menšie alebo rovné ako reálne číslo x (iná bežne používaná definícia distribučnej funkcie $F(x) = P\{X < x\}$ [3]). Medzi distribučnou funkciou a rozdelením pravdepodobnosti náhodnej veličiny je jednoznačný vzťah. Ak pre spojitú veličinu existuje nezáporná integrovateľná funkcia f , kde pre všetky $x \in \mathbb{R}$ platí

$$F(x) = \int_{-\infty}^x f(t)dt,$$

potom f je hustota rozdelenia. Normálne rozdelenie náhodnej veličiny je častý predpoklad pre mnohé štatistické metódy. Funkcia hustoty $f(x)$ pre náhodnú veličinu tohto rozdelenia je vyjadrená vzťahom

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x - \mu_n)^2}{2\sigma_n^2}\right\}$$

pre $x \in \langle -\infty, \infty \rangle$ s parametrami μ a σ^2 , ozn. $N(\mu, \sigma^2)$ (viď obrázok 3 vľavo). V prípade, že náhodná veličina X má logaritmicko-normálne (lognormálne) rozdelenie s parametrami μ a σ^2 , potom jej transformáciou $Y = \log_{10}(X)$ dostaneme veličinu s normálnym rozdelením $Y \sim N(\mu/\ln(10), \sigma^2/\ln^2(10))$. Pre náhodnú veličinu z lognormálneho rozdelenia sa stredná hodnota $E(X)$ vypočíta $E(X) = \exp\left\{\mu + \frac{1}{2}\sigma^2\right\}$ a rozptyl $Var(X) = E(X)^2(\exp\{\sigma^2\} - 1)$ (viď obrázok 3 vpravo).



Obrázok 3: Hustota normálneho rozdelenia $N(3, 1)$ a hustota lognormálneho rozdelenia s parametrami 3 a 1 (ozn. stredná hodnota \diamond , medián \square , modus \circ , dolný kvartil ∇ a horný kvartil \triangle).

Na testovanie normality jednorozmerného výberového rozdelenia sa najčastejšie používajú Kolmogorov-Smirnovov test, jeho modifikácia Lillieforsov test a Jarqueov-Berov test. Pri overovaní či náhodná veličina má rozdelenie $N(\mu, \sigma^2)$, kde μ a σ^2 sú známe parametre, sa štandardne používa Kolmogorov-Smirnovov test. Jeho modifikácia Lillieforsov test sa používa pri testovaní rozdelenia, pre ktoré je nutné parametre odhadnúť.

Lillieforsov test porovnáva empirickú distribučnú funkciu $S(x)$ s distribučnou funkciou normálneho rozdelenia $F(x)$ náhodnej veličiny X , kde $S(x) = j/N$ a j je počet pozorovaní náhodnej veličiny X , ktoré majú nižšiu alebo rovnakú hodnotu ako x . $F(x)$ sa vypočíta pomocou odhadov výberových štatistík parametrov μ a σ^2 pre výber náhodnej veličiny X .

Pri testovaní sa najprv pozorovania zoradia do neklesajúcej postupnosti a vypočíta sa maximálne pozitívna hodnota K_{max}^+ , maximálna negatívna hodnota

K_{max}^- a maximálna absolútna hodnota K_{max} pomocou vzťahov

$$K_{max}^+ = N^{1/2} \max_{1 \leq j \leq N} (j/N - F(x))$$

$$K_{max}^- = N^{1/2} \max_{1 \leq j \leq N} (F(x) - j/N)$$

$$K_{max} = \max(K_{max}^+, K_{max}^-).$$

K_{max} sa upraví pre rozsah výberu náhodnej veličiny N ako $K = N^{1/2} + 0.12 + \frac{0.11}{N^{1/2}} K_{max}$ [7] a potom sa porovnáva s kritickou hodnotou k pre dané α ($\alpha = 0.05$). Kritické hodnoty k_N pre štandardné normálne rozdelenie $N(0, 1)$ sú tabuľkovo známe. Pre rozdelenie $N(\mu, \sigma^2)$ je $k = \frac{k_N - \mu}{\sigma}$. V prípade, že $K < k$ sa hypotéza, že náhodná veličina X má normálne rozdelenie nedá zamietnuť.

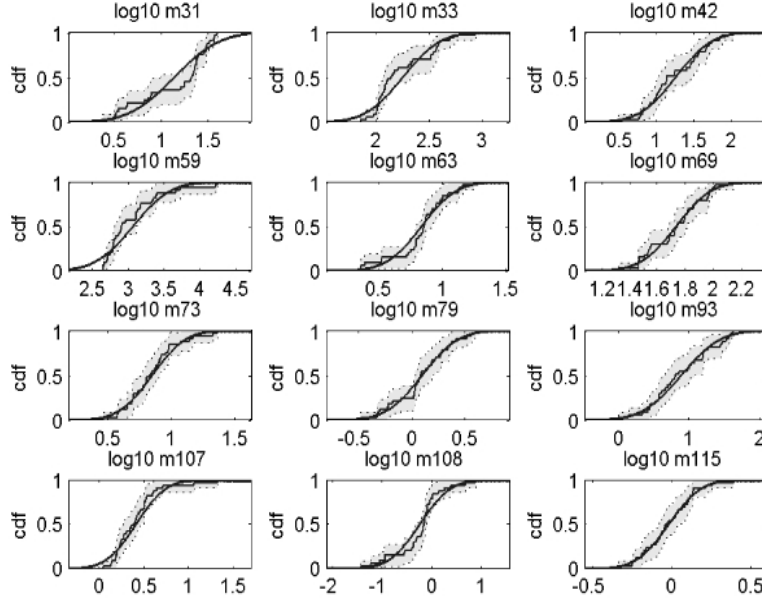
V tabuľke 3 sú výsledky testovania hypotézy pomocou Lillieforsovho testu. Testujeme či náhodná veličina X_i a $Y_i = \log_{10}(X_i)$, pre $i = 1, \dots, n$ a n je počet prchavých organických zložiek vydychovaného vzduchu (náhodných veličín), pochádza z normálneho rozdelenia. V prípade platnosti hypotézy, že výber z náhodnej veličiny pochádza z normálneho rozdelenia je hodnota výsledku 0, v prípade ak túto hypotézu zamietneme je hodnota výsledku 1. V prípade, že náhodná veličina Y_i pochádza z normálneho rozdelenia môžeme tvrdiť, že náhodná veličina X_i má lognormálne rozdelenie.

Na obrázkoch 4 a 5 je grafické znázornenie empirickej distribučnej funkcie (ECDF, *empirical cumulative distribution function*) pre zlogaritmované dáta s 95% konfidenčnými intervalmi a odhadnutá distribučná funkcia (FCDF, *fitted cumulative distribution function*) pre normálne rozdelenie s parametrami μ_i a σ_i^2 , kde μ_i a σ_i^2 sme vypočítali ako výberový priemer a výberový rozptyl z nameraných hodnôt náhodnej veličiny Y_i . V prípade, že FCDF leží medzi hornou a dolnou hranicou konfidenčných intervalov ECDF, môžeme tvrdiť, že náhodná veličina Y_i pochádza z normálneho rozdelenia $N(\mu_i, \sigma_i^2)$.

normálne rozdelenie												
skup.	m31	m33	m42	m59	m63	m69	m73	m79	m93	m107	m108	m115
pac.	1	1	1	1	1	0	1	1	0	1	1	0
zdr.	1	1	1	1	1	1	1	1	1	1	1	1

lognormálne rozdelenie												
skup.	m31	m33	m42	m59	m63	m69	m73	m79	m93	m107	m108	m115
pac.	1	1	0	1	1	0	0	0	0	0	1	0
zdr.	0	1	1	1	1	1	1	1	0	0	1	0

Tabuľka 3: Testovanie dobrej zhody rozdelenia pomocou Lillieforsovho testu koncentrácie prchavých organických zložiek vydychovaného plynu pre skupinu pacientov a skupinu zdravých dobrovoľníkov, kde v prípade platnosti nulovej hypotézy je výsledok 0, ak nemôžeme tvrdiť, že náhodná veličina pochádza z daného rozdelenia je výsledok 1.



Obrázek 4: Empirická distribučná funkcia (ECDF) s 95%-nými konfidenčnými intervalmi a odhadnutá normálna distribučná funkcia (FCDF) pre zlogaritmované hodnoty koncentrácie prchavých organických zložiek vydychovaného plynu pacientov.

2.2 Test na overenie normality viacrozmerného rozdelenia

V prípade, že jednotlivé zložky (náhodné veličiny) majú normálne rozdelenie a sú nezávislé, môžeme povedať, že aj viacrozmerné rozdelenie zložené z týchto zložiek je normálne. Keď vieme, že pozorovania majú viacrozmerné normálne rozdelenie, potom aj každá zložka má normálne rozdelenie.

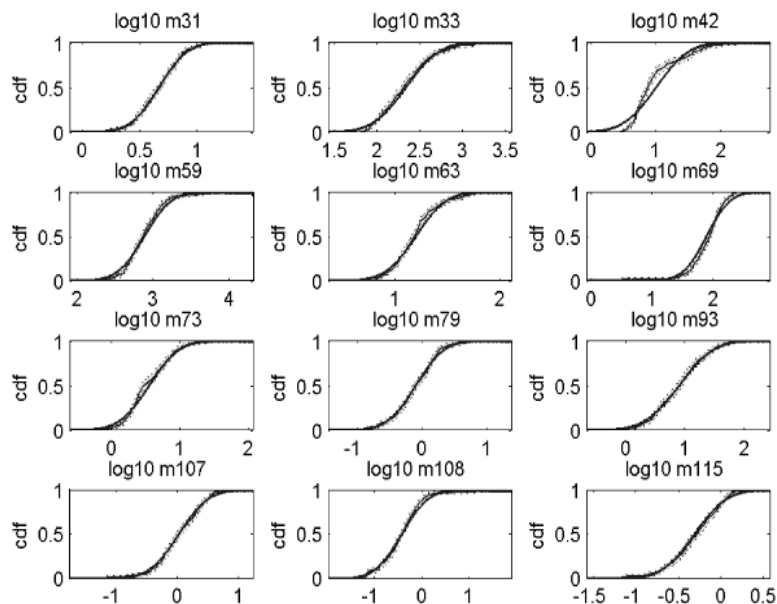
Náhodný vektor $\mathbf{X} = (X_1, \dots, X_n)'$ má n -rozmerné normálne rozdelenie s parametrami $\boldsymbol{\mu}$ a $\boldsymbol{\Sigma}$ s hustotou

$$f(x_1, \dots, x_n) = \frac{1}{(2\pi)^{\frac{n}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

kde $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ je vektor stredných hodnôt a $\boldsymbol{\Sigma} = (\sigma_{ij})$ je kovariančná matica vektora \mathbf{X} , pre ktorú platí $\boldsymbol{\Sigma} = E \left\{ (\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' \right\} \geq 0$, ozn. $\mathbf{X} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Vlastnosť, že koeficient šikmosti (asymetrie, *skewness*)

$$\beta_{1n} = E \left[(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right]^3 = 0$$



Obrázek 5: Empirická distribučná funkcia (ECDF) s 95%-nými konfidenčnými intervalmi a odhadnutá normálna distribučná funkcia (FCDF) pre zlogaritmované hodnoty koncentrácie prchavých organických zložiek vydychovaného plynu zdravých dobrovoľníkov.

a koeficient špicatosti (exces, *kurtosis*)

$$\beta_{2n} = E \left[(\mathbf{X} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right]^2 = n(n+2)$$

sa využíva na testovanie hypotézy, že výber má n -rozmerné normálne rozdelenie [4].

Hypotézu o normalite n -rozmerného rozdelenia [4] zamietneme, ak pre odhad koeficientu špicatosti z náhodného výberu $\mathbf{X}_1, \dots, \mathbf{X}_N$ platí

$$b_1 = \frac{1}{6N} \sum_{i=1}^N \sum_{j=1}^N \left[(\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_j - \bar{\mathbf{X}}) \right]^3 > \chi^2 \left[\frac{(n(n+1)(n+2))}{6}, \alpha \right]$$

kde α je hladina významnosti ($\alpha = 0.05$), \mathbf{S} je výberová kovariančná matica, ktorá sa vypočíta pomocou vzorca $\mathbf{S} = \frac{1}{N-1} \sum_{i=1}^N (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$ a $\chi^2 \left[\frac{(n(n+1)(n+2))}{6}, \alpha \right]$ je kritická hodnota χ -kvadrát rozdelenia.

Hypotézu o normalite n -rozmerného rozdelenia zamietneme, ak pre odhad ko-

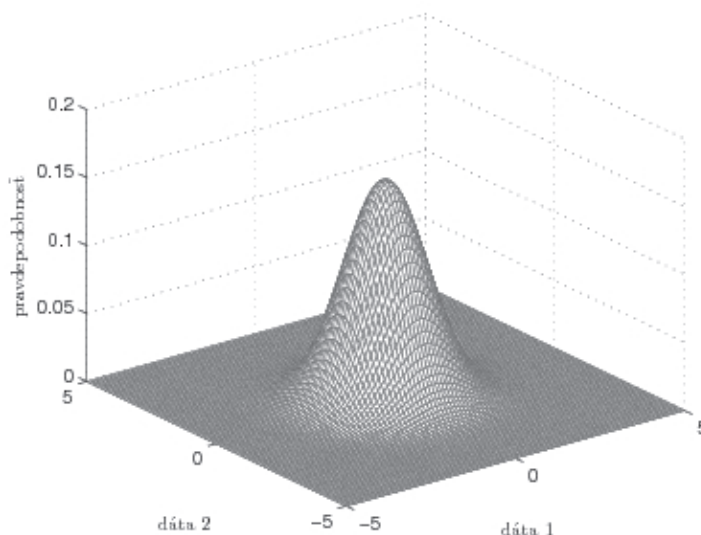
test	originálne dáta		zlogaritmované dáta	
	zdraví dobrovoľníci	pacienti	zdraví dobrovoľníci	pacienti
asymetria	1	1	1	1
exces	0	0	1	0

Tabuľka 4: Testovanie dobrej zhody n -rozmerného normálneho rozdelenia ($n = 12$) pre skupinu zdravých dobrovoľníkov a pacientov pomocou koeficientu šikmosti (asymetria) a koeficientu špicatosti (exces).

eficientu šikmosti platí

$$b_2 = \sqrt{\frac{N}{8n(n+2)}} \left[\frac{1}{N} \sum_{i=1}^N \left\{ (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}) \right\}^2 - n(n+2) \right] > Z_{(\alpha/2)}$$

kde $Z_{(\alpha/2)}$ je kritická hodnota rozdelenia $N(0, 1)$. V tabuľke 4 sú výsledky testovania hypotézy, že náhodné vektory \mathbf{X} a $\mathbf{Y} = \log_{10}(\mathbf{X})$ pochádzajú z n -rozmerného normálneho rozdelenia. Na testovanie sme použili koeficient šikmosti a koeficient špicatosti. Pre testy platí, že v prípade výsledku 1 sa hypotéza, že dáta pochádzajú z n -rozmerného normálneho rozdelenia zamietla. Naopak v prípade výsledku 0 môžeme predpokladať, že dáta pochádzajú z n -rozmerného normálneho rozdelenia.



Obrázek 6: Hustota dvojrozmerného normálneho rozdelenia $N_2(\mathbf{0}, \mathbf{1})$ pre korelačný koeficient zložiek $\rho = 0$.

3 Záver

V článku sme pozorovali, či náhodná veličina (prchavá organická zložka plynu) pochádza z normálneho rozdelenia popřípadě má lognormálne rozdelenie. V ďalšej časti sme testovali normalitu rozdelenia viacrozmerného náhodného vektora (zložený z prchavých organických zložiek plynu) pre nespracované dáta a dáta po logaritmickej transformácii.

Výsledkov vyplýva, že pre zlogaritmované dáta je normalita viac pravdepodobná ako pre surové netransformované dáta. Preto v ďalších štúdiách budeme pracovať s dátami po logaritmickej transformácii.

Normalita pre viacrozmerné rozdelenie bola v oboch prípadoch zamietnutá.

References

- [1] Amann, A., Smith, D. (2005), *Breath analysis for clinical diagnosis and therapeutic monitoring*, World Scientific, Singapore
- [2] Amann, A., Smith, D. (2006), *Breath analysis; the approach towards clinical applications* Mini-Reviews in Medicinal Chemistry, to appear
- [3] Anděl, J. (1985), *Matematická statistika*, SNTL - nakladatelství technické literatury, Praha
- [4] Lamoš, F., Potocký, R. (1998), *Pravdepodobnosť a metamatická štatistika - Štatistické analýzy*, Vydavateľstvo UK, Bratislava
- [5] Španěl, P., Smith, D. (2005), *Selected ion flow tube mass spectroscopy, SIFT-MS, for on-line trace gas analysis of breath*. In: Amann and Smith (Eds.). *Breath Analysis for Clinical and Therapeutic Monitoring*. World Scientific Publishing Co. Pte. Ltd., Singapore, 3–34.
- [6] Štolc S., Arendacká B., Cimermanová K., Hornišová K., Rosipal R., Rublík F., Witkovský V., *Bamod Statistical Toolbox 2006-11-30 BETA; Technical Documentation*, BAMOD PROJECT: BREATH-GAS ANALYSIS FOR MOLECULAR-ORIENTED DETECTION OF MINIMAL DISEASES, LSHC-CT-2005-019031
- [7] Teofilo, G., *An Efficient Algorithm for the Kolmogorov-Smirnov and Lilliefors tests.*, ACM Transactions on Mathematical Software, Vol. 3, No. 1, (1977), 64.
- [8] Turner, C. (2006), *A longitudinal study of ammonia, acetone and propanol in the exhaled breath of 30 subjects using selected ion flow tube mass spectrometry, SIFT-MS*, Institute of physics publishing, *Physiological Measurement* 27, 321–337.
- [9] Turner, C. (2006), *A longitudinal study of breath isoprene in healthy volunteers using selected flow tube mass spectrometry (SIFT-MS)*, Institute of physics publishing, *Physiological Measurement* 27, 13–22.

- [10] Turner, C. (2006), *A longitudinal study of ethanol and acetaldehyde in the exhaled breath of healthy volunteers using selected-ion flow-tube mass spectrometry*, Rapid communications in mass spectrometry 20, 61–22.
- [11] Witkovsky, V., Cimermanova, K. (2006), *Testovanie zhodnosti stredných hodnôt vzoriek dychu pre detekciu rakoviny pľúc*. In: J. Hartmann, J. Michálek, editors, XVII. letní škola biometriky. Biometrické metody a modely v současné vědě a výzkumu. Lednice, ČR, 433–442.



Závěr statistického výzkumu prekonceptů žáků základních škol v ČR

VERONIKA KAINZOVÁ

*Ústav merania, Oddelenie teoretických metód, Slovenská akadémia vied,
Dúbravská cesta 9, 841 04 Bratislava
e-mail: katarina.cimermanova@gmail.com*

Abstrakt: Článek je věnován analýze dotazníkového šetření 954 respondentů žáků základních škol z míst s různým počtem obyvatel.

Cílem šetření bylo prozkoumání dětského pojetí některých pojmů – znalost prekonceptů – v oblasti přírodovědného vzdělávání.

Otázky v dotaznících byly přitom voleny tak, aby umožnili odhalit nejzávažnější miskoncepty a aby bylo možno v budoucnosti diferencovaně přistupovat k vytváření vzdělávacích modulů v oblasti přírodních věd. Při vyhodnocení byly porovnány odpovědi dívek a chlapců a byly také srovnávány správnosti odpovědí dle lokality (obce do 5000 obyvatel, do 20 000 obyvatel a nad 20 000 obyvatel).

Klíčová slova: dotazníkové šetření, dětské pojetí vybraných fyzikálních, chemických a biologických pojmů, prekoncepty v přírodních vědách, miskoncepty,

1 Úvod

Statistický výzkum prekonceptů (intuitivních představ) žáků se konal v několika dílčích etapách vuplynulém školním roce na vybraných základních školách v ČR. Jako metoda výzkumu byl zvolen dotazník, který byl sestaven tak, že první část obsahovala základní informace a ve druhé části následovaly otázky. Jednotlivé úlohy byly kombinované otázky svýběrem odpovědi a otázky otevřené, kde žák odpovědi tvoří či doplňuje. I když úlohy svýběrem odpovědi nejsou při zjišťování prekonceptů jedinou možnou alternativou, někdy nejsou považovány za nejvhodnější pro tento typ výzkumu, většina otázek byla tohoto typu. Možnosti odpovědi byly voleny tak, aby ukázaly na základní nedostatky vporozumění vybraným pojmům. Základními kritérii při tvorbě dotazníků byla srozumitelnost, jednoduchost, malá časová náročnost a vhodná úroveň otázek a výběr pojmů odpovídající mentální úrovni žáka základní školy. Celý výzkum proběhl v rámci projektu GA ČR „Konstruktivismus a jeho aplikace vintegrovaném pojetí přírodovědného vzdělávání“, který navrhl tým oborových didaktiků fyziky, chemie

abiologie Přírodovědecké fakulty Univerzity Palackého v Olomouci a Pedagogické fakulty Univerzity Hradec Králové a pedagogů Univerzity Tomáše Bati ve Zlíně. Základním cílem projektu je rozvinout konstruktivistický přístup do oblasti teorie výuky fyziky, chemie a biologie a zvýšit vědecko-výzkumný potenciál příslušných oborových didaktik. Během listopadu 2005 byla sestavena první verze dotazníku, který obsahoval celkem deset otázek. Prvních devět otázek bylo zaměřeno na výzkum představ týkajících se pojmů látka, energie (její přenos, možnosti využití) a hustota. Desátá otázka byla tvorba kognitivní (pojmové) mapy k základnímu pojmu Slunce pomocí zadaných vedlejších pojmů. Na základě vyhodnocení tohoto předvýzkumu, který proběhl v Olomouci (ZŠ Holečkova, ZŠ Hálkova – škola s rozšířenou výukou cizích jazyků), byly dotazníky upraveny do stávající podoby. Předvýzkumu se zúčastnilo 75 respondentů (42 chlapců, 33 dívek). Následovala 1. etapa celorepublikového výzkumu prekonceptů. Dotazníky byly rozeslány na základní školy v rámci celé republiky (Čechy i Morava). Průzkum byl opět realizován v patřících třídách základních škol. Výběr škol byl zcela náhodný, snahou bylo pokrýt všechny oblasti České republiky i různé typy základních škol. V první etapě výzkumu bylo zpracováno 418 vyplněných dotazníků. Z celkového počtu 418 respondentů bylo 196 chlapců a 222 dívek. Žáci měli na vyplnění dotazníku jednu vyučovací hodinu. Druhé etapy výzkumu se zúčastnilo 512 respondentů, kteří vyplňovali stejný dotazník a za stejných podmínek jako žáci v etapě první. Druhé etapy výzkumu se zúčastnily základní školy převážně ze Zlínského kraje a Olomouckého kraje. Opět byly zastoupeny obce s různým počtem obyvatel, takže bylo možné provést vyhodnocení úspěšnosti i z hlediska geografického. Navíc oproti první etapě výzkumu bylo u otázek, které obsahovaly podúkoly, provedeno vyhodnocení i jednotlivých dílčích otázek.

2 Výsledky výzkumu

2.1 Způsob hodnocení

Každá otázka byla oklasifikována škálou 1, 2, . . . , 5 (1 = nejlepší, 5 = nejhorší). Pokud šlo o otázky typu „správná – špatná odpověď“, byly oklasifikovány 1 (správná) nebo 5 (špatná).

Otázka č. 1 Slova *slon, křemen, cihla, voda, žula, muchomůrka, skála, lípa, televize, mravenec, špendlík, jehla, bakterie, vzduch, automobil*, která označují různé objekty rozděl do skupin a zapiš do připravené tabulky.

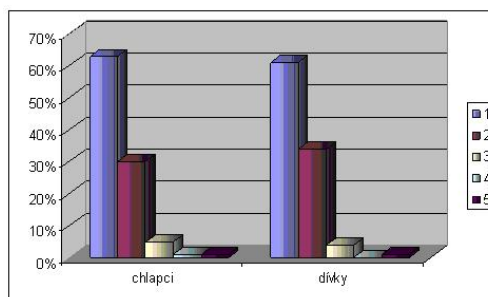
název skupiny		
<i>živá přírodnína</i>	<i>neživá přírodnína</i>	<i>lidský výrobek</i>

Vyhodnocení otázky č. 1 Otázka byla hodnocena škálou 1 až 5 (1 nejlepší, 5 nejhorší).

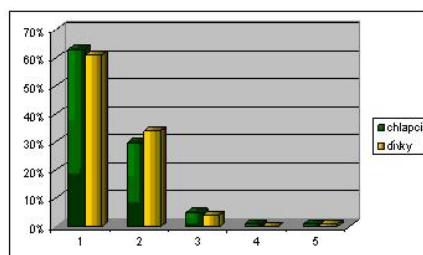
Procentní zastoupení odpovědí hodnocených známkou 1 je jak u chlapců tak u dívek přibližně stejné (63 % chlapci, 61 % dívky), viz graf č. 1. Četnost odpovědí klasifikovaných známkou 2 je 30% chlapci a 34 % dívky. Nejnižší je podíl odpovědí klasifikovaných známkami 4 a 5 (1% chlapci, 1 % dívky).

Nejčastěji se vyskytovaly chyby vpřirazení slov: vzduch lidský výrobek (chybná odpověď), bakterie lidský výrobek (chybná odpověď), záměna slova bakterie za slovo baterie, voda lidský výrobek (chybná odpověď), muchomůrka neživá přírodnina (chybná odpověď).

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	63%	61%
známka 2	30%	34%
známka 3	5%	4%
známka 4	1%	0%
známka 5	1%	1%



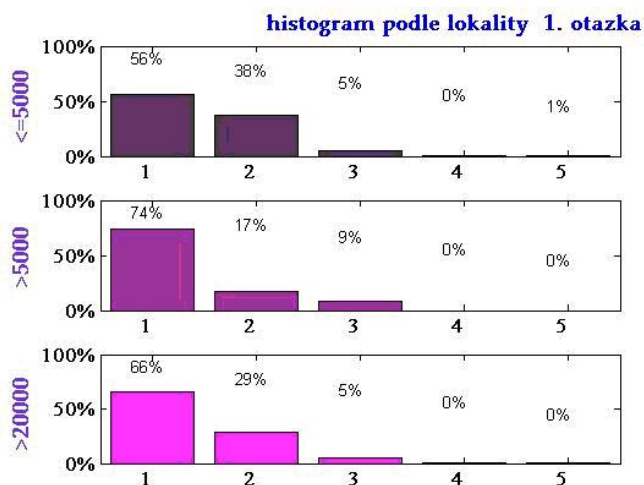
Obrázek 1: Srovnání úspěšnosti chlapců a dívek



Obrázek 2: Histogram počtu známek u dívek a chlapců

Vyhodnocení z hlediska geografického viz graf na Obr. 3

klasifikace	počet obyvatel		
	< 5000	5000 – 20 000	> 20 000
známka 1	56%	74%	66%
známka 2	38%	17%	29%
známka 3	5%	9%	5%
známka 4	0%	0%	0%
známka 5	1%	0%	0%



Obrázek 3: Histogram podle lokality

Otázka č. 2a Vyhledej slova (označují látky nebo předměty), která mají společnou vlastnost - stejné skupenství nebo obsahují stejnou látku. Slova vepiš do tabulek (některá slova se mohou opakovat vícekrát!).

Voda, železo, dřevo, papír, zmrzlina, mléko, jablko, automobil, moře, sníh, lavice, kniha, déšť, strom, kámen, vzduch, mlha, sešit.

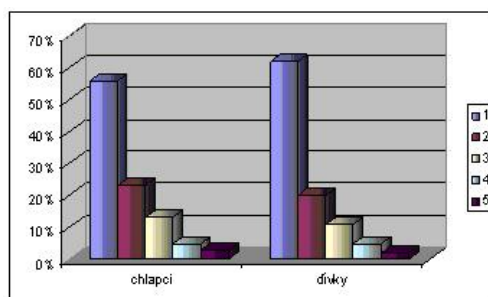
Skupenství		
<i>pevná látka</i>	<i>kapalina</i>	<i>plyn</i>

Vyhodnocení otázky č. 2a Otázka byla ohodnocena škálou 1 až 5 (1 nejlepší, 5 nejhorší).

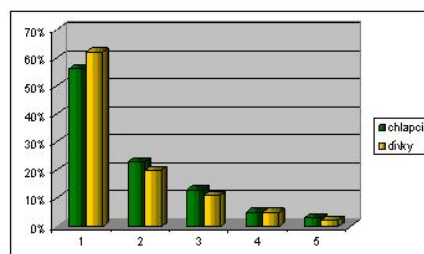
Graf č. 1 srovnává vzájemnou úspěšnost chlapců a dívek. Největší rozdíl můžeme vidět u klasifikace známkou 1, kdy dívky (62 %) byly o málo úspěšnější než chlapci (56 %). Z grafu č. 2 je vidět procentní zastoupení klasifikační škály (1 až 5) u chlapců a dívek. Největší je podíl odpovědí hodnocených známkou 1, jak u chlapců (56 %), tak u dívek (62 %). Procentní zastoupení odpovědí klasifikovaných známkou 2 je přibližně stejné u obou pohlaví (chlapci 23 %, dívky 20 %). Nejnižší je podíl odpovědí klasifikovaných známkou 5 (dívky 2 %, chlapci 3 %).

Nejčastější chybné odpovědi: sníh - kapalina (chybná odpověď), zmrzlina - kapalina (chybná odpověď), jablko - kapalina (chybná odpověď), automobil plyn (chybná odpověď), mlha - kapalina (chybná odpověď).

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	56%	62%
známka 2	23%	20%
známka 3	13%	11%
známka 4	5%	5%
známka 5	3%	2%



Obrázek 4: Srovnání úspěšnosti chlapců a dívek



Obrázek 5: Histogram počtu známek u dívek a chlapců

Otázka č. 2b Vyhledej slova (označují látky nebo předměty), která mají společnou vlastnost - stejné skupenství nebo obsahují stejnou látku. Slova vepiš do tabulek (některá slova se mohou opakovat vícekrát!).

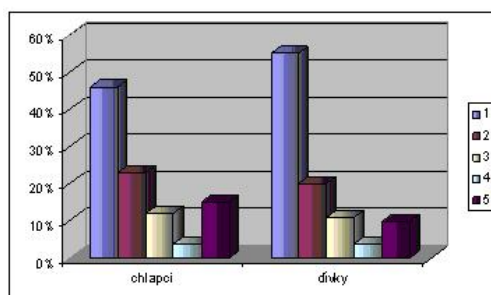
Voda, železo, dřevo, papír, zmrzlina, mléko, jablko, automobil, moře, sníh, lavice, kniha, déšť, strom, kámen, vzduch, mlha, sešit.

Obsahují stejnou látku		
<i>voda</i>	<i>dřevo</i>	<i>kov</i>

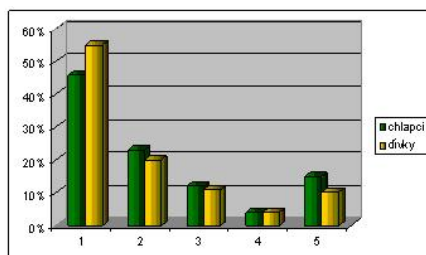
Vyhodnocení otázky č. 2b Způsob hodnocení je identický s otázkou 2a. Klasifikační škála 1-5 (1 nejlepší, 5 nejhorší).

Opět se zde vyskytuje největší podíl odpovědí klasifikovaných známkou 1 (chlapci 46 %, dívky 55 %). Četnost odpovědí klasifikovaných známkou 2 je vyšší u chlapců (23 %) než u dívek (20 %). Zajímavé je vyšší procento odpovědí klasifikovaných známkou 5. U dívek je tomu v 10% případů, u chlapců v 15% případů. Je to způsobeno především tím, že někteří žáci danou otázku vůbec nevyplnili. Graf č. 2 ukazuje vyšší úspěšnost dívek. Podíl odpovědí hodnocených známkou 1 je vyšší než u chlapců a zároveň podíl odpovědí hodnocených známkou 5 je nižší než u chlapců.

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	46%	55%
známka 2	23%	20%
známka 3	12%	11%
známka 4	4%	4%
známka 5	15%	10%



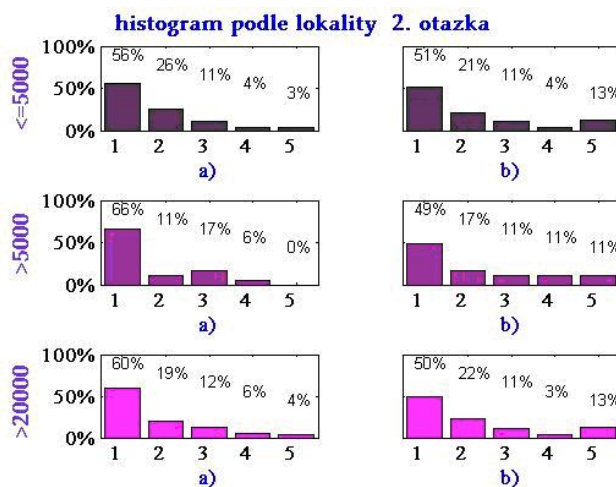
Obrázek 6: Srovnání úspěšnosti chlapců a dívek



Obrázek 7: Histogram počtu známek u dívek a chlapců

Vyhodnocení [první hodnota pro část otázky a), druhá hodnota pro část otázky b)] z hlediska geografického, viz graf na Obr. 8

klasifikace	počet obyvatel		
	< 5000	5000 – 20 000	> 20 000
známka 1	56%, 51%	74%	60%, 50%
známka 2	26%, 21%	17%	19%, 22%
známka 3	11%, 11%	9%	12%, 11%
známka 4	4%, 4%	0%	6%, 3%
známka 5	3%, 13%	0%	4%, 13%



Obrázek 8: Histogram podle lokality

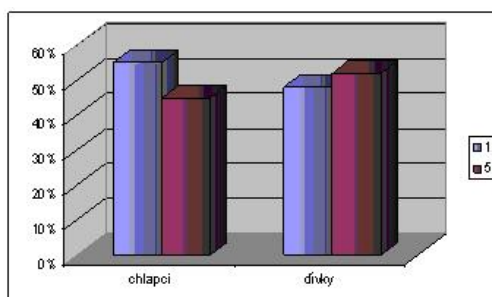
Otázka č. 3 Představ si, že máš malý hrníček svodou a budeš jej zahřívát plamenem jedné svíčky. Voda začne vřít (víš, že vroucí voda má teplotu 100 oC). Co se stane, když místo plamene jedné svíčky použiješ svíčky dvě? Vybranou odpověď zakroužkuj.

- a) *Voda bude vřít při teplotě 200° C.*
- b) *Voda začne vřít při teplotě 100° C.*
- c) *Voda bude vřít při teplotě 50° C.*
- d) *Nevím.*

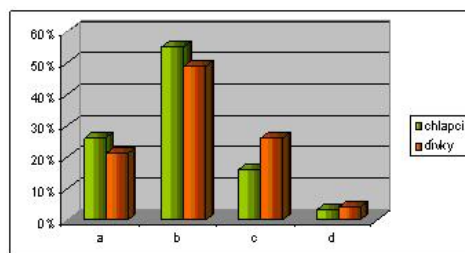
Vyhodnocení otázky č. 3 Otázka č. 3 je klasifikována jen dvěma známkami 1 a 5 (1 správná odpověď, 5 chybná odpověď).

Z grafu č. 1 můžeme porovnat úspěšnost odpovědi chlapců a dívek. Je zřejmé, že podíl správných odpovědí je u chlapců vyšší než u dívek. (55 % - chlapci, 48 % - dívky). U nich převažují odpovědi klasifikované známkou 5 (52 % dívky, 45 % chlapci). Graf č. 2 ukazuje procentní zastoupení jednotlivých odpovědí (a - d) u chlapců a dívek, bez klasifikace. Největší podíl odpovědí je u možnosti b) Voda začne vřít při teplotě 100°C - správná odpověď (55 % chlapci, 49 % dívky). Větší procento dívek než chlapců volilo možnost c) Voda bude vřít při teplotě 50°C - chybná odpověď (26 % dívky, 16 % chlapci). Naopak tomu bylo u možnosti a) Voda bude vřít při teplotě 200°C - chybná odpověď, kterou volilo větší procento chlapců než dívek (26 % chlapci, 21 % dívky).

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	55%	48%
známka 5	45%	52%



Obrázek 9: Srovnání úspěšnosti chlapců a dívek



Obrázek 10: Histogram počtu známek u dívek a chlapců

varianta	<i>chlapani</i>	<i>dívky</i>
a)	21%	21%
b)	55%	49%
c)	16%	26%
d)	3%	4%

Vyhodnocení z hlediska geografického viz graf na Obr. 11

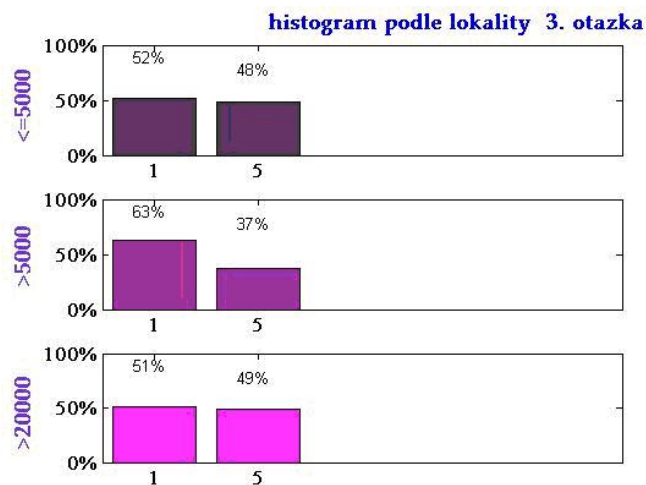
klasifikace	<i>počet obyvatel</i>		
	<i>< 5000</i>	<i>5000 – 20 000</i>	<i>> 20 000</i>
známka 1	52%	63%	51%
známka 5	48%	37%	49%

Otázka č. 4 *Vzimě Tě příjemně hřeje kožíšek nebo prošívaná bunda. Co se stane skouskem zmrzliny, když jej zabalíš do kožichu? (výběr–zakroužkuj)*

- Zmrzlina bude tát pomaleji.*
- Zmrzlina roztaje dříve.*
- Zmrzlina bude tát stejně rychle.*
- Zmrzlina vůbec neroztaje.*
- Nevím.*

Vyhodnocení otázky č. 4 Stejně jako u otázky č. 3 je zde použita klasifikace dvěma známkami 1 a 5 (1 správná odpověď, 5 chybná odpověď).

Z grafu č. 1 můžeme porovnat úspěšnost odpovědí obou pohlaví. Lze si všimnout, že podíl špatných odpovědí (klasifikovaných známkou 5, 87 % chlapci, 83 % dívky), je jak u chlapců tak u dívek mnohem vyšší než podíl odpovědí správných (13 % chlapci, 17 % dívky). Graf č. 2 zachycuje procenta zastoupení jednotlivých odpovědí (a až e), u chlapců a dívek, bez klasifikace.



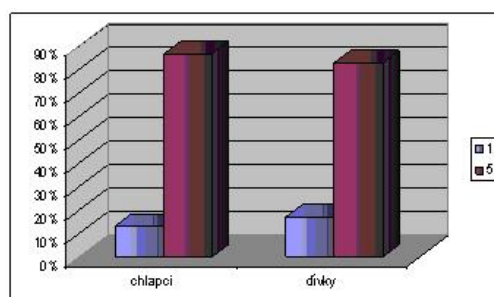
Obrázek 11: Histogram podle lokality

Nejčastěji volenou odpovědí byla možnost b) Zmrzlina roztaje dříve - chybná odpověď (52 % - chlapci, 41 % - dívky). Možnost a) Zmrzlina bude tát pomaleji - správná odpověď, volilo o málo větší procento dívek než chlapců (13 % chlapci, 17 % dívky). U odpovědi c) Zmrzlina bude tát stejně rychle, jako kdyby ležela volně na stole - chybná odpověď, tomu bylo naopak (21 % chlapci, 13 % dívky). Odpověď e) Nevím, volilo jen malé procento chlapců (4 %), narozdíl od dívek (19 %).

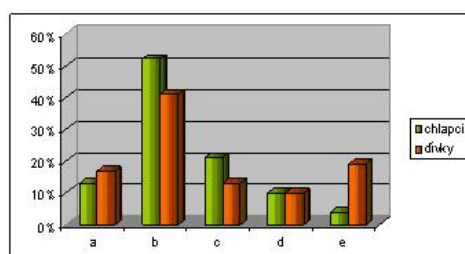
klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	13%	17%
známka 5	87%	83%

varianta	<i>chlapci</i>	<i>dívky</i>
a)	13%	17%
b)	52%	41%
c)	21%	13%
d)	10%	10%
e)	4%	19%

Vyhodnocení z hlediska geografického viz graf na Obr. 14



Obrázek 12: Srovnání úspěšnosti chlapců a dívek

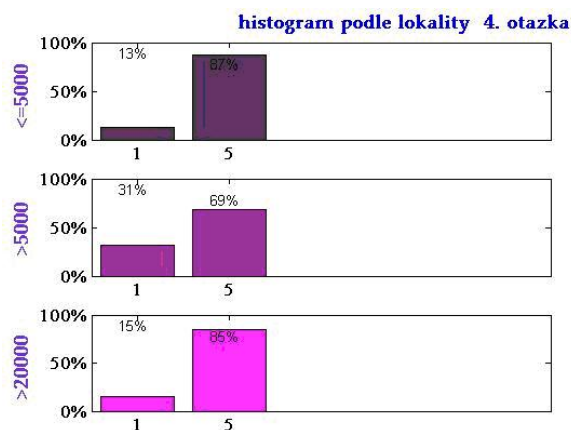


Obrázek 13: Histogram počtu známek u dívek a chlapců





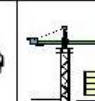
klasifikace	počet obyvatel		
	< 5000	5000 – 20 000	> 20 000
známka 1	13%	31%	15%
známka 5	87%	69%	85%

Otázka č. 5 V prvním sloupci tabulky máš uvedené látky, tělesa a živočichy. Ti mají v sobě skrytou energii. Vyber podle obrázků u každé látky, tělesa a živočicha, jak se může jejich **energie využít**. Svoji volbu označ do příslušného políčka křížkem.

Vyhodnocení otázky č. 5 V prvním sloupci tabulky máš uvedené látky, tělesa a živočichy. Ti mají v sobě skrytou energii. Vyber podle obrázků u každé látky, tělesa a živočichovi, jak se může jejich **energie využít**? Svoji volbu označ do příslušného políčka **křížkem**.



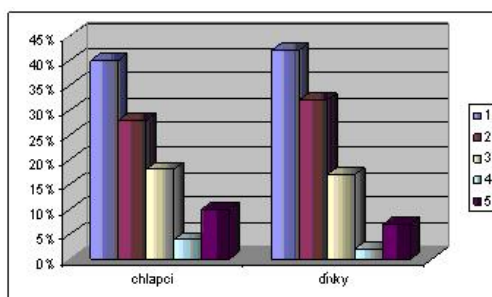
Obrázek 14: Histogram podle lokality

					
uhlí					
voda					
pára					
slunce					
berzán					
kůň					
člověk					

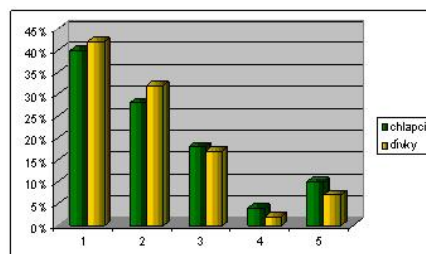
Otázka č. 5 je klasifikována škálou 1 5 (1 nejlepší, 5 nejhorší).

Graf č. 1 vzájemně srovnává podíly klasifikovaných odpovědí u obou pohlaví. Z grafu č. 2 můžeme vidět, že procentní zastoupení odpovědí klasifikovaných známkami 1 až 5 je u dívek zhruba stejné jako u chlapců. Četnost odpovědí klasifikovaných známkou 1 a 2 je u chlapců vyšší než u dívek. Nejvyšší četnost vykazují klasifikace známkou 1 (40 % chlapci, 42 % dívky), dále pak klasifikace známkou 2 (25 % chlapci, 32 % dívky). Vyšší počet odpovědí klasifikovaných známkou 5 (10 % chlapci, 7 % dívky) je způsobena opět tím, že někteří žáci otázku vůbec nevyplnili.

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	40%	42%
známka 2	28%	32%
známka 3	18%	17%
známka 4	4%	2%
známka 5	10%	7%



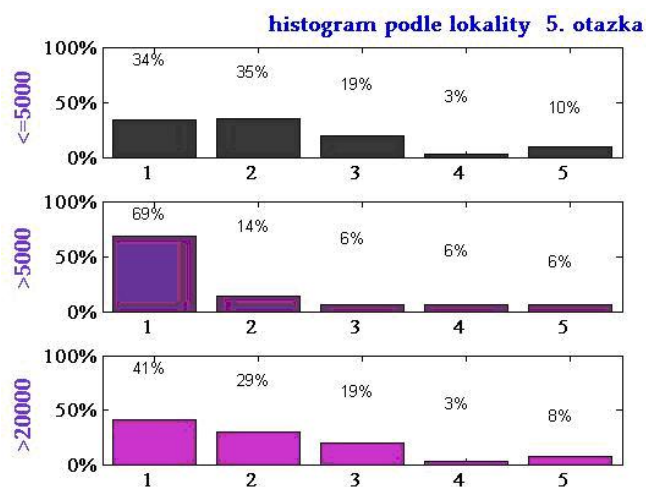
Obrázek 15: Srovnání úspěšnosti chlapců a dívek



Obrázek 16: Histogram počtu známek u dívek a chlapců

Vyhodnocení z hlediska geografického viz graf na Obr. 17.

klasifikace	<i>počet obyvatel</i>		
	<i>< 5000</i>	<i>5000 – 20 000</i>	<i>> 20 000</i>
známka 1	34%	69%	41%
známka 2	35%	14%	29%
známka 3	19%	6%	19%
známka 4	3%	6%	3%
známka 5	10%	6%	8%



Obrázek 17: Histogram podle lokality

Otázka č. 6 Označ správné odpovědi (zakroužkuj ANO - NE):

- a) *Světlušky vyzařují světelnou energii.* ANO – NE
- b) *Kolem sněženek rychleji taje sníh (protože vydávají teplo).* ANO – NE
- c) *Velryby vydechují ve vodě vodní páru.* ANO – NE
- d) *Stěny kvasné nádoby, ve které kvasinky způsobují kvašení, se ohřívají.* ANO – NE
- e) *Tvé tělo vydává teplo.* ANO – NE

Vyhodnocení otázky č. 6 Každá z možností (a – e) byla ohodnocena známkami 1 (správná odpověď) nebo 5 (chybná odpověď). Správné možnosti byly ve všech případech možnosti ANO - klasifikace známkou 1.

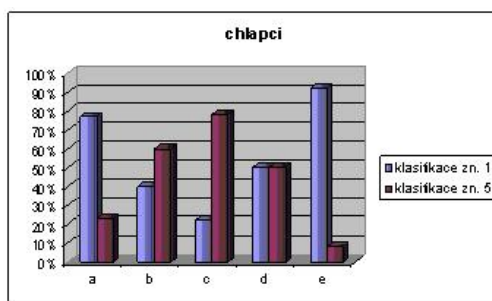
Z grafu č. 1 a 2 je zřejmé, že nejlépe zodpovězené možnosti byly: a) Světlušky vyzařují světelnou energii – (ANO) – (77 % chlapci, 69 % dívky) e) Tvé tělo vydává teplo – (ANO) – (92 % chlapci, 88 % dívky)

jak u chlapců, tak u dívek převažují v obou případech správné odpovědi.

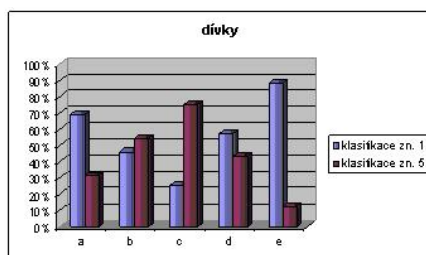
Nejhůře zodpovězené byly možnosti: b) Kolem sněženek rychleji taje sníh (protože vydávají teplo) – (ANO) – (40 % chlapci, 46 % dívky) c) Velryby vydechují ve vodě vodní páru. – (ANO) – (22 % chlapci, 25 % dívky)

Graf č. 19 srovnává úspěšnost dívek a chlapců dohromady.

klasifikace	<i>chlapci</i>	<i>dívky</i>
a) známka 1	77%	69%
a) známka 5	23%	31%
b) známka 1	40%	46%
b) známka 5	60%	54%
c) známka 1	22%	25%
c) známka 5	78%	75%
d) známka 1	50%	57%
d) známka 5	50%	43%
e) známka 1	92%	88%
e) známka 5	8%	12%



Obrázek 18: Histogram počtu známek chlapců



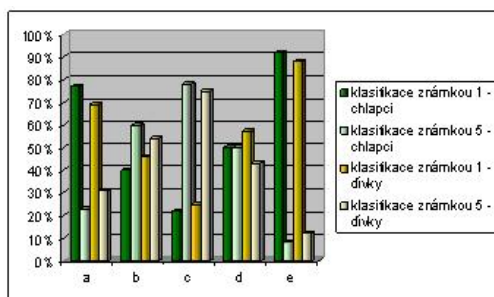
Obrázek 19: Histogram počtu známek dívek

Vyhodnocení z hlediska geografického:

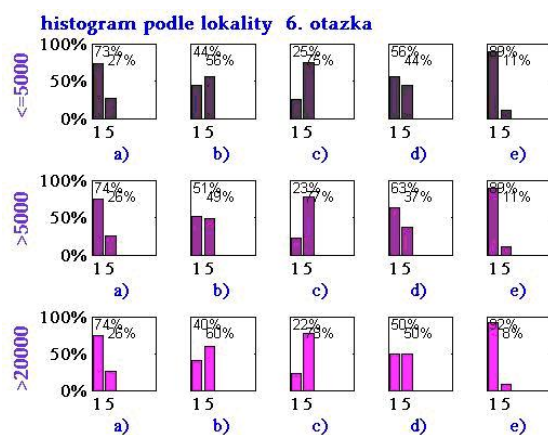
Otázka č. 7 Označ, (zakroužkuj) předměty a látky, které **plavou ve vodě**.

a) **Kulička ze železa,**

b) **kulička ze dřeva,**



Obrázek 20: Srovnání úspěšnosti chlapců a dívek



Obrázek 21: Histogram podle lokality

- c) *skleněná kulička,*
- d) *korková zátka,*
- e) *polystyren,*
- f) *brambor,*
- g) *benzín,*
- h) *olej.*

Vyhodnocení otázky č. 7 Správné odpovědi v otázce č. 7 byly: b) kulička ze dřeva, d) korková zátka, e) polystyren, g) benzín a h) olej.

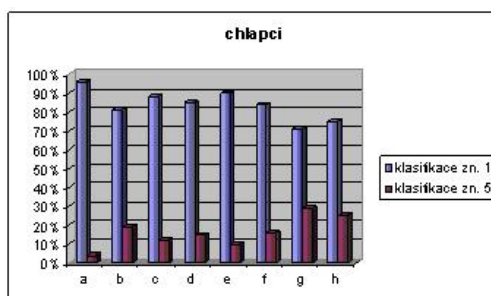
Každá z možností (a až h) byla ohodnocena známkami 1 (správná volba) nebo 5 (špatná volba).

Grafy 1 a 2 ukazují, že podíl odpovědí správných je ve všech případech vyšší než podíl špatných odpovědí jak u chlapců, tak u dívek.

Chybné odpovědi se nejčastěji vyskytují v případech b) kulička ze dřeva (19 % chlapci, 21 % dívky), g) benzín (29 % chlapci, 41 % dívky), h) olej (25 % chlapci, 25 % dívky). U dívek je chybná odpověď v případě g) benzín, čtenější než u chlapců (29 % chlapci, 41 % dívky).

Nejvyšší podíl správných odpovědí je v případě a) kulička ze železa (96% chlapci, 93% dívky).

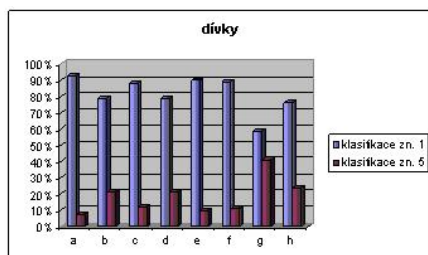
klasifikace	<i>chlapci</i>	<i>dívky</i>
a) známka 1	96%	93%
a) známka 5	4%	7%
b) známka 1	81%	79%
b) známka 5	19%	21%
c) známka 1	88%	88%
c) známka 5	12%	12%
d) známka 1	85%	79%
d) známka 5	15%	21%
e) známka 1	90%	90%
e) známka 5	10%	10%
f) známka 1	84%	89%
f) známka 5	16%	11%
g) známka 1	71%	59%
g) známka 5	29%	41%
h) známka 1	75%	76%
h) známka 5	25%	24%



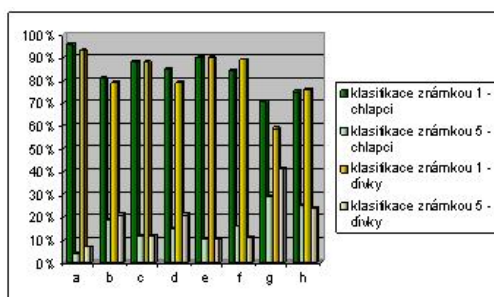
Obrázek 22: Histogram počtu známek chlapců

Vyhodnocení z hlediska geografického:

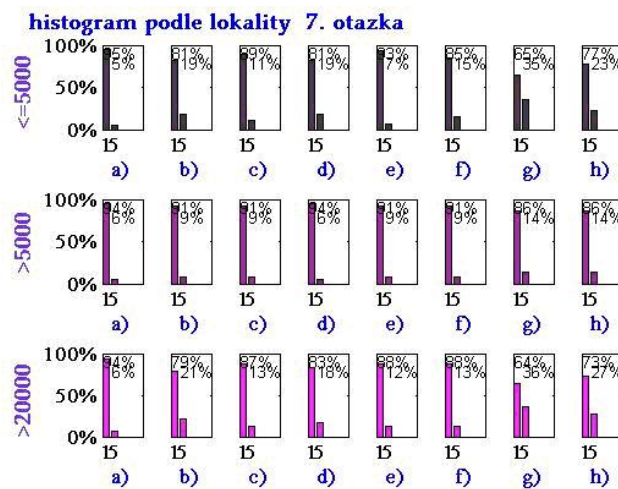
Otázka č. 8 V každé dvojici označ (zakroužkuj) látku nebo předmět, který má **větší hustotu**.



Obrázek 23: Histogram počtu známek dívek



Obrázek 24: Srovnání úspěšnosti chlapců a dívek



Obrázek 25: Histogram podle lokality

- a) **Voda — sirup,**
- b) **dřevo — železo,**
- c) **voda — polystyren,**
- d) **dřevo — voda,**
- e) **voda — vzduch,**
- f) **olej — voda.**

Vyhodnocení otázky č. 8 Správné řešení otázky č. 8: a) sirup, b) železo, c) voda, d) voda, e) voda, f) voda.

Otázka byla klasifikována stejně jako ot. č. 7, tj. správně zvolená varianta v případech a) f) byla ohodnocena známkou 1, chybná varianta známkou 5.

Z grafu č. 1 je zřejmé, že nejhůře zodpovězenými možnostmi u chlapců byly d) dřevo chybně (67 %), f) olej chybně (82 %), c) polystyren chybně (46 %).

Obdobné výsledky byly i v případě dívek: d) dřevo chybně (71 %), f) olej chybně (86 %), c) polystyren chybně (53 %).

U možností a) (88 % chlapci správně, 91 % dívky - správně), b) (79 % chlapci správně, 80 % dívky správně), e) (73 % chlapci správně, 63 % dívky - správně), převažují správné odpovědi. Z grafu č. 3 vidíme vzájemné srovnání úspěšnosti chlapců a dívek. U varianty a), byly dívky o málo úspěšnější než chlapci (88 % chlapci, 91 % dívky). Variantu e) volilo správně více chlapců než dívek (73 % chlapci, 63 % dívky).

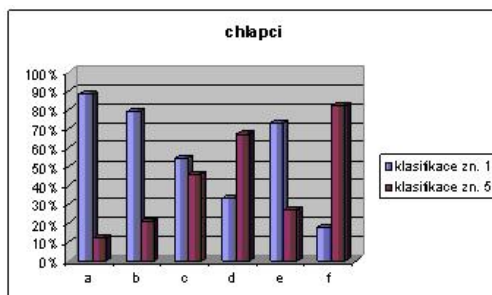
Na základě vyhodnocení otázek č. 7 a 8 vidíme, že pojem hustota je pro věkovou skupinu žáků v 4. 5. třídě obtížný. To, které látky plavou, určili žáci z uvedených 8 látek celkem bez problémů správně. Nejvíce chybovali u benzínu (chybně odpovědělo 35 % žáků) a u oleje (chybně 25 %). Je překvapivé, že 20 % žáků se domnívá, že dřevo neplave.

Při výběru správné látky s větší hustotou (žáci volili vždy 1 látku z 6 dvojic) se vyskytlo výrazně více chybných odpovědí, 68 % žáků se domnívá, že dřevo má větší hustotu než voda; 84 % žáků označilo olej a resp. 47 % polystyren jako látku s větší hustotou než voda. Žáci si zřejmě zaměňují hustotu s viskozitou.

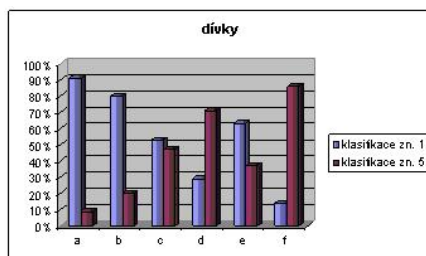
Potvrdilo se, že pojmu hustota je třeba věnovat zvýšenou pozornost a názornými ukázkami, demonstracemi a především experimenty konkretizovat tento pojem, aby se případné miskoncepty (nesprávné představy) žáků „pře-psaly“ správnými představami. Pojem hustota je součástí učiva fyziky, chemie, biologie, nabízí se tedy více možností objasnit tento pojem v několika tématech i využít odlišná hlediska.

Podíváme-li se na vyhodnocení správných odpovědí dle lokality resp. velikosti města je v otázkách 7, 8 patrný pouze jeden výraznější rozdíl v odpovědích, a to při posouzení hustoty vody a polystyrenu či dřeva. Žáci z malých obcí chybovali nejméně, polystyren má větší hustotu než voda zvolilo jen 44 % respondentů (z větších měst označilo kolem 55 % žáků), dřevo má větší hustotu než voda vybralo pouze 62 % respondentů (z větších měst kolem 75 % žáků).

klasifikace	<i>chlapci</i>	<i>dívky</i>
a) známka 1	88%	91%
a) známka 5	12%	9%
b) známka 1	79%	80%
b) známka 5	21%	20%
c) známka 1	54%	53%
c) známka 5	46%	47%
d) známka 1	33%	29%
d) známka 5	67%	71%
e) známka 1	73%	63%
e) známka 5	27%	37%
f) známka 1	18%	14%
f) známka 5	22%	86%



Obrázek 26: Histogram počtu známek chlapců

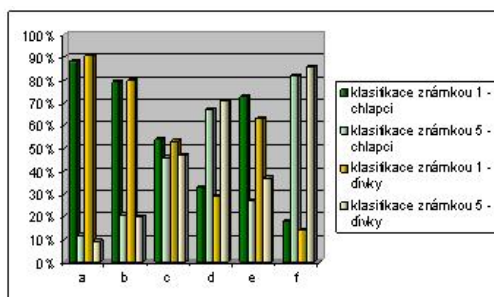


Obrázek 27: Histogram počtu známek dívek

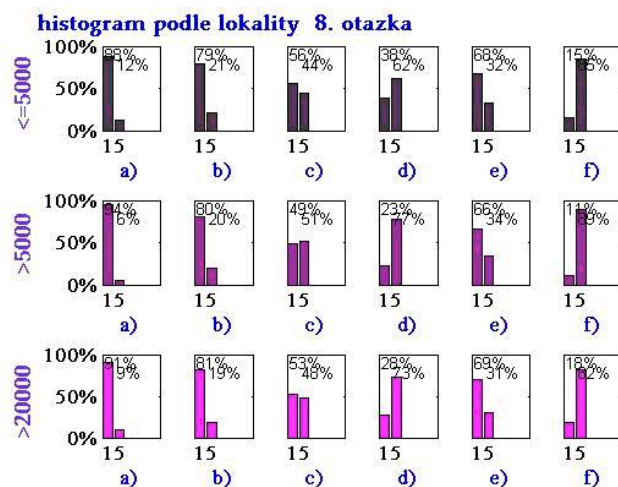
Vyhodnocení z hlediska geografického:

Otázka č. 9 Vyber a zakroužkuj, co potřebují všechny **živé organismy**.

a) **vodu**,



Obrázek 28: Srovnání úspěšnosti chlapců a dívek



Obrázek 29: Histogram podle lokality

- b) *energii,*
- c) *vzduch,*
- d) *led,*
- e) *pohyb.*

Vyhodnocení otázky č. 9 Otázka č. 9 byla hodnocena škálou 1-5 (1 nejlepší, 5 nejhorší).

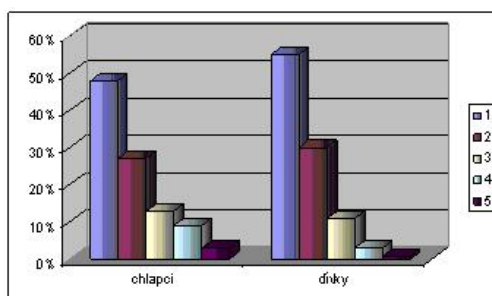
Z grafu č. 1 je zřejmé, že nejčastěji zastoupená je klasifikace známkou 1, jak u chlapců (48 %) tak u dívek (55 %). Rozložení četností známek 2 až 5 je u chlapců a dívek přibližně stejné.

Z grafu č. 2 vidíme vyšší úspěšnost dívek při klasifikaci známkami 1 (55 % dívky, 48 % chlapci) a 2 (30 % dívky, 27 % chlapci). Graf č. 3 ukazuje četnost jednotlivých zvolených odpovědí (a až e).

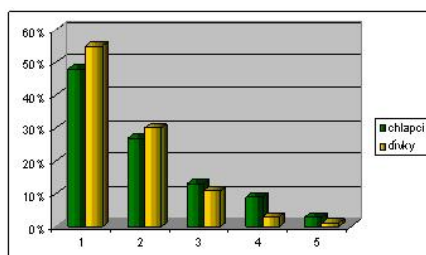
Svislá osa není vyjádřena v procentech, ale přímo jsou zde vyneseny počty odpovědí. Některý žák mohl označit 2 odpovědi, jiný např. 4 odpovědi atd., proto procentní vyjádření není v tomto případě možné.

Je zde vidět vyšší počet volených odpovědí v případech a) voda, b) energie, c) vzduch. U chlapců převažují případy a), b), u dívek varianta c) vzduch. Nejnižší počet odpovědí vidíme v případě d) led - jak u chlapců, tak u dívek. Volba varianty e) pohyb, je čtenější u chlapců.

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	48%	55%
známka 2	27%	30%
známka 3	13%	11%
známka 4	9%	3%
známka 5	3%	1%



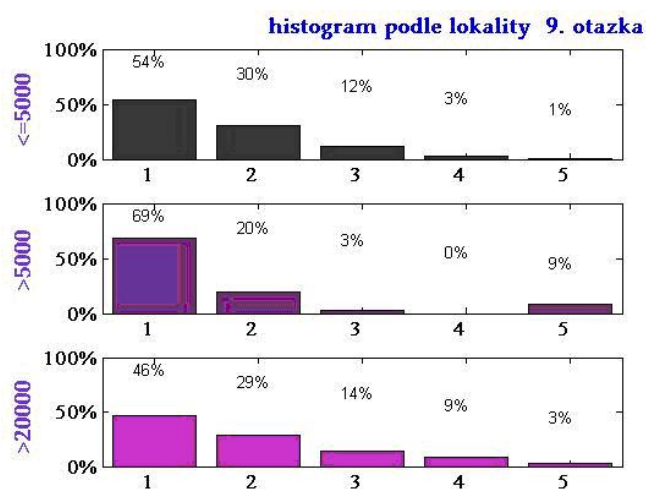
Obrázek 30: Srovnání úspěšnosti chlapců a dívek



Obrázek 31: Histogram počtu známek u dívek a chlapců

Vyhodnocení z hlediska geografického viz graf na Obr. 32.

klasifikace	počet obyvatel		
	< 5000	5000 – 20 000	> 20 000
známka 1	54%	69%	46%
známka 2	30%	23%	29%
známka 3	12%	3%	14%
známka 4	3%	0%	9%
známka 5	1%	9%	3%



Obrázek 32: Histogram podle lokality

Otázka č. 10 Ke každému obrázku **ČARAMI** přiřaď **JEDEN** název (slovo) z pravého sloupce.

- a) **vodu,**
- b) **energii,**
- c) **vzduch,**
- d) **led,**
- e) **pohyb.**

Otázka č. 10 Ke každému obrázku **ČARAMI** přiřaď **JEDEN** název (slovo) z pravého sloupce.



Vyhodnocení otázky č. 10 Stejně jako ot. č. 9 byla ot. č. 10 ohodnocena škálou 1 až 5 (1 nejlepší, 5 nejhorší).

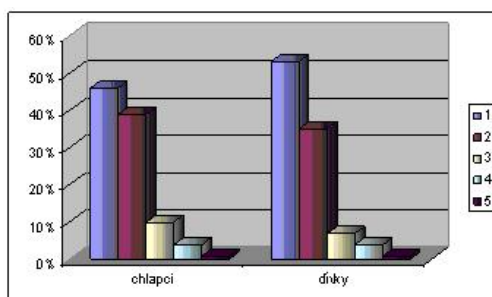
Graf č. 1 ukazuje rozložení četností klasifikační škály (1–5) srovnatelně u chlapců a dívek.

U dívek vidíme větší rozdíly v zastoupení klasifikace známkou 1 (53 %) a 2 (35 %) než u chlapců (1 — 46 %, 2 — 39 %). U dívek výrazněji převažuje hodnocení známkou 1 oproti hodnocení známkou 2. U chlapců jsou poměry obou klasifikací přibližně stejné.

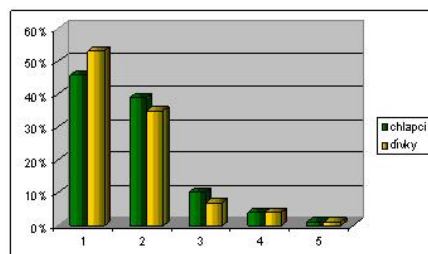
Z grafu č. 1 můžeme rovněž porovnat úspěšnost obou pohlaví. Klasifikace známkou 1 je u dívek zastoupená četněji než u chlapců. Klasifikace známkou 4 a 5 se vyskytuje minimálně (4 — 4 % chlapci, 4 % dívky), (5 — 1% chlapci, 1% dívky) .

Nejčastější chybná přiřazení byla: kyslík — továrna (chybně), železo — továrna (chybně), železo — magnet (správně), planeta — strom (chybně), kyslík — strom (správně), teplo — teploměr (chybně), teplo — slunce (správně), rtuť — kost (chybně), rtuť — teploměr (správně).

klasifikace	<i>chlapci</i>	<i>dívky</i>
známka 1	46%	53%
známka 2	39%	35%
známka 3	10%	7%
známka 4	4%	4%
známka 5	1%	1%



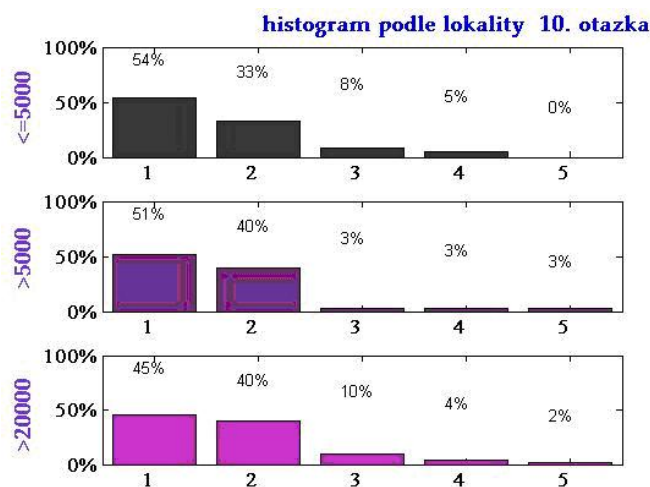
Obrázek 33: Srovnání úspěšnosti chlapců a dívek



Obrázek 34: Histogram počtu známek u dívek a chlapců

Vyhodnocení z hlediska geografického viz graf na Obr. 35.

klasifikace	<i>počet obyvatel</i>		
	<i>< 5000</i>	<i>5000 – 20 000</i>	<i>> 20 000</i>
známka 1	54%	51%	45%
známka 2	33%	40%	40%
známka 3	8%	3%	10%
známka 4	5%	3%	4%
známka 5	0%	3%	2%



Obrázek 35: Histogram podle lokality

3 Závěr

Výsledky druhé etapy výzkumu odpovídají přibližně výsledkům z etapy první.

Je zřejmé, že nejlépe zodpovězenými otázkami, jak u chlapců tak u dívek, byly ot. č. 1 (živá, neživá přírodnina, lidský výrobek), ot. č. 2 (pevná látka, kapalina, plyn), ot. č. 9 z oblasti biologie, a ot. č. 10 (přiřazování obrázek — slovo).

Nejhůře zodpovězenými otázkami byly ot. č. 3, kde úspěšnost dívek byla nižší než u chlapců, nejčastější chybnou odpovědí bylo a) voda bude vřít při teplotě 200°C. Dále pak ot. č. 4 (více jak 80% chlapců i dívek odpovídalo nesprávně — b) zmrzlina roztaje dříve).

Co se týká vyhodnocení z hlediska geografického, tak u problematické otázky č. 4 nejhůře dopadly obce do 5 000 obyvatel a nad 20 000, kde více jak 80% žáků odpovídalo chybně.

Jak ukazují výsledky uvedeného výzkumu prekonceptů, největší nedostatky se objevují u pojmů teplo, teplota, skupenství a hustota, stejně tak u přenosu tepla a energie v živých systémech. Ukazuje se, že pojmy se vytvářejí izolovaně a jejich začlenění do integrovaného systému přírodovědných poznatků má ještě značné mezery. Na závěr je třeba připomenout základní otázku — s jakými vědomostmi a představami vstupují žáci do 6. ročníku základní školy a jak tedy vytvořit odpovídající didaktický model učiva na 2. stupni základní školy a v nižších ročnících víceletých gymnázií.

Prekoncepty (intuitivní představy) jsou individuálními charakteristikami učícího se jedince a jsou utvářeny všemi dosavadními vlivy a zkušenostmi, které na něho působily. Při jejich utváření hraje roli celá řada aspektů. Jsou to jednak

exogenní faktory (sociální, ekonomické, etnické, kulturní aj.), a faktory endogenní, vycházející z psychologických a psychosociálních charakteristik každého žáka.

Podle J. Lova má každý jedinec své pojetí přírodovědných pojmů utvářeno třemi vlivy. Jde o primitivní (elementární) vědu, která souvisí s individuálními zkušenostmi jedince. Druhým faktorem je vliv laické vědy, kterou ovlivňují informace přinášené médii a mimoškolním prostředím. Úlohu třetího faktoru plní školní věda, tzv. formální školní vzdělávání. Škola tedy nemůže být jediným zdrojem informací. Typickým rysem intuitivních představ je jejich trvalost a odolnost vůči změnám.

Cílem těchto statistických výzkumů je snaha o „rozbití“ chybných intuitivních představ a samozřejmě povědomí učitelů i žáků o nejčastějších chybách ve struktuře prekonceptů.

Vyučování by tedy mělo směřovat k harmonickému sblížení tzv. „primitivní vědy“ každého žáka s tím, co prezentuje škola („školní věda“). A toho učitel bez znalosti dětských pojetí nemůže být schopen.

Výzkum prekonceptů a změna těchto prekonceptů (concept change) je základem v konstruktivistickém pojetí přírodovědného vzdělávání.



Nápadně vybočující údaje v slabě nelineárních regresních modelech

LUBOMÍR KUBÁČEK¹ AND EVA TESAŘÍKOVÁ²

¹ *Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
Tomkova 40, 77900 Olomouc, Czech Republic
e-mail: kubacekl@aix.upol.cz*

² *Department of Algebra and Geometry
Faculty of Science, Palacký University
Tomkova 40, 77900 Olomouc, Czech Republic
e-mail: tesariko@aix.upol.cz*

Abstrakt: *Nápadně vybočující data v slabě nelineárním modelu můžeme za určitých předpokladů identifikovat podobně jako v modelech lineárních. Je však potřebné mít k dispozici kritérium, pomocí kterého rozhodneme jestli linearizace je možná, nebo ne.*

Existuje více postupů, pomocí kterých identifikujeme outlier ve statistickém souboru, viz např. „Barnett, V., Lewis, T.: Outliers in Statistical Data. John Wiley, New York 1994.“

V tomto příspěvku procedury uvedené v „Zvára, K.: Regresní analýza. Academia, Praha 1989“ budou použity ve slabě nelineárních regresních modelech. Protože tyto postupy jsou zpracovány pro lineární modely, musíme hledat kritérium, pomocí kterého rozhodneme, zda procedura smí být použita.

Model nazveme slabě nelineárním, jestliže kritérium umožňuje lineární proceduru použít (viz „Kubáček, L., Kubáčková, L.: Regression models with a weak nonlinearity. Technical Report Nr.1998, Universitát Stuttgart, 1998,1–67“ a „Kubáček, L., Kubáčková, L.: Statistika a metrologie. Vydavatelství Palackého univerzity, Olomouc 2000“).

1 Úvod

Nelineární regresní model zapisujeme v kvadratickém tvaru

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n \left[\mathbf{F}\delta\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}), \boldsymbol{\Sigma} \right],$$

protože hledaná kritéria vychází z druhých derivací. Zde

\mathbf{Y} ... n -rozměrný náhodný vektor (observační vektor),
 $\mathbf{f}(\boldsymbol{\beta})$... střední hodnota observačního vektoru \mathbf{Y} ,
 $\boldsymbol{\beta}$... k -rozměrný neznámý vektorový parametr $\boldsymbol{\beta} \in R^k$ (k -rozměrný reálný vektorový prostor),
 $\boldsymbol{\beta}^*$... skutečná hodnota parametru $\boldsymbol{\beta}$,
 $\mathbf{F} = \partial \mathbf{f}(\mathbf{u}) / \partial \mathbf{u}'|_{u=\boldsymbol{\beta}_0}$, kde $\boldsymbol{\beta}_0$ je přibližná hodnota skutečné hodnoty $\boldsymbol{\beta}^*$ parametru $\boldsymbol{\beta}$ (předpokládáme, že hodnota matice \mathbf{F} je $r(\mathbf{F}) = k < n$),
 $\mathbf{f}(\cdot)$... n -rozměrná funkce se spojitými druhými derivacemi,

$$\begin{aligned}
 \mathbf{f}(\cdot) &= (f_1(\cdot), \dots, f_n(\cdot))', \\
 \mathbf{f}_0 &= \mathbf{f}(\boldsymbol{\beta}_0), \\
 \mathbf{F}_i &= \partial^2 f_i(\mathbf{u}) / \partial \mathbf{u} \partial \mathbf{u}'|_{u=\boldsymbol{\beta}_0}, i = 1, \dots, n, \\
 \kappa_i(\delta\boldsymbol{\beta}) &= \delta\boldsymbol{\beta}' \mathbf{F}_i \delta\boldsymbol{\beta}, i = 1, \dots, n, \\
 \delta\boldsymbol{\beta} &= \boldsymbol{\beta} - \boldsymbol{\beta}_0, \\
 \boldsymbol{\kappa}(\delta\boldsymbol{\beta}) &= (\kappa_1(\delta\boldsymbol{\beta}), \dots, \kappa_n(\delta\boldsymbol{\beta}))',
 \end{aligned}$$

$\boldsymbol{\Sigma}$... kovarianční matice observačního vektoru \mathbf{Y} (předpokládáme, že kovarianční matice $\boldsymbol{\Sigma}$ je pozitivně definitní),

Předpokládáme, že kvadratická aproximace modelu $\mathbf{Y} \sim N_n[\mathbf{f}(\boldsymbol{\beta}), \boldsymbol{\Sigma}]$, i.e. $\mathbf{Y} - \mathbf{f}_0 \sim N_n[\mathbf{F}\delta\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}), \boldsymbol{\Sigma}]$, je pro naše účely postačující. Zde $\delta\boldsymbol{\beta} = \boldsymbol{\beta} - \boldsymbol{\beta}_0$.

2 Pomocná tvrzení

V prvním kroku uvažujme lineární model

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n(\mathbf{F}\delta\boldsymbol{\beta}, \boldsymbol{\Sigma}). \quad (1)$$

Nechť

$$\begin{aligned}
 \mathbf{v} &= \mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}(\mathbf{Y} - \mathbf{f}_0) = \mathbf{Y} - \mathbf{f}_0 - \mathbf{F}\widehat{\delta\boldsymbol{\beta}}, \\
 \widehat{\delta\boldsymbol{\beta}} &= \mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{f}_0), \quad \mathbf{C} = \mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{F}.
 \end{aligned}$$

Jestliže nulová hypotéze je $H_0 : E(\mathbf{v}) = \mathbf{0}$ a alternativní je $H_1 : E(\mathbf{v}) \neq \mathbf{0}$, potom testová statistika je (viz Rao, C.R.: Linear Statistical Inference and Its Applications. J. Wiley, New York-London-Sydney 1965, p. 155)

$$T = \mathbf{v}'\boldsymbol{\Sigma}^{-1}\mathbf{v} \sim \chi_{n-k}^2(\delta), \quad (2)$$

kde $\delta = E(\mathbf{v})'\boldsymbol{\Sigma}^{-1}E(\mathbf{v})$, tzn. jestliže $E(\mathbf{v}) = \mathbf{0}$, potom též $\delta = 0$. Jestliže $\delta \neq 0$, potom (1) některá měření nápadně vybočují.

V modelu

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n\left(\mathbf{F}\delta\boldsymbol{\beta} + \frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}), \boldsymbol{\Sigma}\right) \quad (3)$$

platí $E(\mathbf{v}) = \mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}\frac{1}{2}\boldsymbol{\kappa}(\delta\boldsymbol{\beta})$ a tedy $\mathbf{v}'\boldsymbol{\Sigma}^{-1}\mathbf{v} \sim \chi_f^2(\delta)$, kde

$$\delta = \frac{1}{4}\boldsymbol{\kappa}'(\delta\boldsymbol{\beta})\left(\mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}\right)'\boldsymbol{\Sigma}^{-1}\mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\kappa}(\delta\boldsymbol{\beta})$$

(v případě, že outliery se nevyskytují).

Nechť $\delta_{n-k,max}$ je řešením rovnice

$$P\{\chi_{n-k}^2(\delta_{n-k,max}) \geq \chi_{n-k}^2(0; 1 - \alpha)\} = \alpha + \varepsilon,$$

kde $\varepsilon > 0$ je dostatečně malé reálné číslo.

Jestliže $\frac{1}{4}\boldsymbol{\kappa}'(\delta\boldsymbol{\beta})\left(\mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}\right)'\boldsymbol{\Sigma}^{-1}\mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}\boldsymbol{\kappa}(\delta\boldsymbol{\beta}) \leq \delta_{n-k,max}$, potom test (2) můžeme použít v modelu (3) s malým zvýšením ($\varepsilon > 0$) rizika α .

Definice 2.1 The Batesova a Wattsova míra vnitřní křivosti $K^{(int)}(\boldsymbol{\beta}_0)$ (viz [2]) modelu (3) v bodě $\boldsymbol{\beta}_0$ je

$$K^{(int)}(\boldsymbol{\beta}_0) = \sup \left\{ \frac{\sqrt{\boldsymbol{\kappa}'(\delta\boldsymbol{\beta})(\mathbf{M}_F\boldsymbol{\Sigma}\mathbf{M}_F)'\boldsymbol{\kappa}(\delta\boldsymbol{\beta})}}{\delta\boldsymbol{\beta}'\mathbf{C}\delta\boldsymbol{\beta}} : \delta\boldsymbol{\beta} \in R^k \right\}.$$

Jestliže uvážíme definici 2.1 a předchozí výpočty, lze vyslovit lemma

Lemma 2.2 Jestliže

$$\delta\boldsymbol{\beta}'\mathbf{C}\delta\boldsymbol{\beta} \leq 2\frac{\sqrt{\delta_{n-k,max}}}{K^{(int)}(\boldsymbol{\beta}_0)},$$

potom v případě absentujících outlierů

$$P\{\mathbf{v}'\boldsymbol{\Sigma}^{-1}\mathbf{v} \geq \chi_{n-k}^2(0; 1 - \alpha)\} \leq \alpha + \varepsilon.$$

Vzhledem k Schefféově větě [8]

$$\begin{aligned} \mathbf{v}'\boldsymbol{\Sigma}^{-1}\mathbf{v} &\leq \chi_{n-k}^2(0; 1 - \alpha) \\ \Leftrightarrow \forall\{\mathbf{h} \in R^n\}|\mathbf{h}'\mathbf{v}| &\leq \sqrt{\chi_{n-k}^2(0; 1 - \alpha)} \\ &\times \sqrt{\mathbf{h}'(\boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}')\mathbf{h}}, \end{aligned}$$

z čehož nerovnost

$$|(\mathbf{e}_i^{(n)})'\mathbf{v}| > \sqrt{\chi_{n-k}^2(0; 1 - \alpha)}\sqrt{\{\boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}'\}_{i,i}}$$

implikuje, že anebo *ité* měření je outlier, anebo test ovlivňuje nelinearita. (Dlužno poznamenat, že uvedená formulace Schefféovy věty je důsledkem dvou faktů, tzn. $P\{\mathbf{v} \in \mathcal{M}[\text{Var}(\mathbf{v})] = \mathcal{M}(\boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}')\} = 1$ a $\boldsymbol{\Sigma}^{-1}$ je jednou z verzí zobecněné inverze (cf. [7]) matice $\boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}'$.)

3 Linearizační oblast pro test T

Množinu

$$\mathcal{L}_{H_0} = \left\{ \delta\boldsymbol{\beta} : \delta\boldsymbol{\beta}'\mathbf{C}\delta\boldsymbol{\beta} \leq 2\frac{\sqrt{\delta_{n-k,max}}}{K^{(int)}(\boldsymbol{\beta}_0)} \right\}$$

nazveme linearizační oblastí pro nulovou hypotézu H_0 v modelu (3).

Jestliže s praktickou jistotou víme, že $\delta\beta^* \in \mathcal{L}_{H_0}$, kde $\delta\beta^* = \beta^* - \beta_0$, potom test T můžeme použít s malým zvětšením rizika α v nelineárním modelu (3). Avšak zůstává ještě vyřešit problém jak je nelinearitou ovlivněna silofunkce testu T .

Let $E(Y_i) = f_i(\beta) + \Delta_i$, kde $\Delta_i \neq 0$. Potom parametr necentrality $\delta_{\Delta_i}^{(n-k)}$ testové statistiky T v modelu (3) je

$$\begin{aligned} \delta_{\Delta_i}^{(n-k)} &= \\ &= \left[\frac{1}{2} \kappa(\delta\beta) + \mathbf{e}_i^{(n)} \Delta_i \right]' (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \\ &\quad \times \left[\frac{1}{2} \kappa(\delta\beta) + \mathbf{e}_i^{(n)} \Delta_i \right] \\ &= \frac{1}{4} \kappa'(\delta\beta) (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \kappa(\delta\beta) \\ &\quad + \Delta_i \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i} \kappa(\delta\beta) \\ &\quad + \Delta_i^2 \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i}. \end{aligned}$$

Jestliže $\delta_{\Delta_i, \min}^{(n-k)}$ je řešením rovnice

$$\begin{aligned} P \left\{ \chi_{n-k}^2(\delta_{\Delta_i, \min}^{(n-k)}) \geq \chi_{n-k}^2(0; 1 - \alpha) \right\} &= \\ = P \left\{ \chi_{n-k}^2 \left(\Delta_i^2 \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i} \right) \geq \chi_{n-k}^2(0; 1 - \alpha) \right\} &- \varepsilon, \end{aligned}$$

potom linearizační oblast pro silofunkci testu T v alternativní hypotéze $\Delta_i \neq 0$ je

$$\begin{aligned} \left\{ \delta\beta : \Delta_i^2 \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i} + \frac{1}{4} \kappa'(\delta\beta) (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \kappa(\delta\beta) \right. \\ \left. + \Delta_i \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i} \kappa(\delta\beta) \geq \delta_{\Delta_i, \min}^{(n-k)} \right\}. \end{aligned}$$

V následujícím textu použijeme označení

$$\mathbf{A}_i = \sum_{j=1}^n \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,j} \mathbf{F}_j.$$

Potom $\{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i} \kappa(\delta\beta) = \delta\beta' \mathbf{A}_i \delta\beta$.

Protože $\forall \{ \delta\beta \in R^k \} \frac{1}{4} \kappa'(\delta\beta) (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \kappa(\delta\beta) \geq 0$, platí následující tvrzení.

Věta 3.1 Jestliže

$$\delta\beta' \mathbf{A}_i \delta\beta \geq \frac{1}{\Delta_i} \left[\delta_{\Delta_i, \min}^{(n-k)} - \Delta_i^2 \{ (\mathbf{M}_F \Sigma \mathbf{M}_F)^+ \}_{i,i} \right],$$

potom efekt nelinearity modelu (3) na hodnotu silofunkce v alternativní hypotéze $\Delta_i \neq 0$ je menší než ε .

Poznámka 3.2 Z hlediska praktického užití je rozumné definovat linerizační oblast pro silofunkci v alternativě $\Delta_i \neq 0$ následovně

$$\mathcal{L}_{\Delta_i} = \left\{ \delta\beta : \delta\beta' \mathbf{A}_i \delta\beta \geq \frac{1}{\Delta_i} \left[\delta_{\Delta_i, \min}^{(n-k)} - \Delta_i^2 \{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i} \right] \right\}.$$

Protože hodnota Δ_i je předem neznámá, je užitečné vyšetřit funkci

$$g_i(\Delta) = \frac{1}{\Delta} \left[\delta_{\Delta, \min}^{(n-k)} - \Delta^2 \{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i} \right], \quad \Delta \in R^1.$$

S ohledem na její průběh můžeme posoudit, zda existuje dostatečně vysoká pravděpodobnost, že $\delta\beta^* \in \mathcal{L}_{\Delta_i}$ (viz též numerické příklady v další sekci).

Jestliže $\delta\beta^* \in \mathcal{L}_{H_0} \cap \mathcal{L}_{\Delta_i}$ a současně

$$|\{\mathbf{v}\}_i| \geq \sqrt{\chi_{n-k}^2(0; 1 - \alpha)} \sqrt{\{\boldsymbol{\Sigma} - \mathbf{F}\mathbf{C}^{-1}\mathbf{F}'\}_{i,i}},$$

potom hypotézu $E(\mathbf{v}) = \mathbf{0}$ odmítáme kvůli *itému* měření. Tedy předpokládáme $\{E(\mathbf{Y}) - \mathbf{f}_0\}_i = \{\mathbf{F}\delta\beta + \frac{1}{2}\boldsymbol{\kappa}(\delta\beta)\}_i + \Delta_i$, kde $\Delta_i \neq 0$.

V dalším uvážíme model

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n(\mathbf{F}\delta\beta + \mathbf{e}_i^{(n)} \Delta_i, \boldsymbol{\Sigma}) \quad (4)$$

místo modelu (1) a model

$$\mathbf{Y} - \mathbf{f}_0 \sim N_n\left(\mathbf{F}\delta\beta + \frac{1}{2}\boldsymbol{\kappa}(\delta\beta) + \mathbf{e}_i^{(n)} \Delta_i, \boldsymbol{\Sigma}\right) \quad (5)$$

místo modelu (3). Dále předpokládáme $r(\mathbf{F}, \mathbf{e}_i^{(n)}) = k + 1 < n$.

Lemma 3.3 Nejlepší odhady parametrů $\delta\beta$ a Δ_i v modelu (4) jsou

$$\begin{aligned} \widehat{\delta\beta}_{out} &= \widehat{\delta\beta} - \mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{e}_i^{(n)}\widehat{\Delta}_i \\ &\sim N_k\left(\delta\beta, \mathbf{C}^{-1} + \mathbf{C}^{-1}\mathbf{F}'\boldsymbol{\Sigma}^{-1}\mathbf{e}_i^{(n)}\left(\{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i}\right)^{-1}\right. \\ &\quad \left. \times (\mathbf{e}_i^{(n)})'\boldsymbol{\Sigma}^{-1}\mathbf{F}\mathbf{C}^{-1}\right), \\ \widehat{\Delta}_i &= \left(\{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i}\right)^{-1}\boldsymbol{\Sigma}^{i,\cdot}\mathbf{v} \\ &\sim N_1\left(\Delta_i, \frac{1}{\{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i}}\right), \\ \mathbf{v} &= \mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}}(\mathbf{Y} - \mathbf{f}_0), \quad \boldsymbol{\Sigma}^{i,\cdot} = \{\boldsymbol{\Sigma}^{-1}\}_{i,\cdot}. \end{aligned}$$

Lemma 3.4 Nechť nulová hupotéza v modelu (4) je $H_0 : \Delta_i = 0$ versus alternativní $H_a : \Delta_i \neq 0$. Testová statistika je v tomto případě

$$\begin{aligned} T_{out,i} &= \widehat{\Delta}_i^2 \{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i} \sim \chi_1^2(\delta), \\ \delta &= [E(\widehat{\Delta}_i)]^2 \{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i} = \Delta_i^2 \{(\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+\}_{i,i}. \end{aligned}$$

4 Linearizační oblast pro test $T_{out,i}$

Statistika $T_{out,i}$ z lemmatu 3.4 má rozdělení v modelu (5) $\chi_1^2(\delta_{out,\Delta_i})$, kde

$$\begin{aligned}
\delta_{out,\Delta_i} &= \left\{ E \left(\left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \right)^{-1} \boldsymbol{\Sigma}^{i,i} \mathbf{M}_F^{\boldsymbol{\Sigma}^{-1}} (\mathbf{Y} - \mathbf{f}_0) \right\}^2 \\
&\quad \times \left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \\
&= \left[\Delta_i + \left(\left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \right)^{-1} \left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \right. \\
&\quad \left. \times \frac{1}{2} \boldsymbol{\kappa}(\delta\boldsymbol{\beta}) \right]^2 \left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \\
&= \left[\Delta_i + \frac{1}{2} \left(\left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \delta\boldsymbol{\beta}' \mathbf{A}_i \delta\boldsymbol{\beta} \right) \right]^2 \\
&\quad \times \left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i}. \tag{6}
\end{aligned}$$

Definice 4.1 Míra nelinearity pro test $T_{out,i}$ je

$$\begin{aligned}
C_i^{(out)}(\boldsymbol{\beta}_0) &= \sup \left\{ \frac{\sqrt{Q(\delta\boldsymbol{\beta})}}{\delta\boldsymbol{\beta}' \mathbf{C} \delta\boldsymbol{\beta}} : \delta\boldsymbol{\beta} \in R^k \right\} \\
Q(\delta\boldsymbol{\beta}) &= \boldsymbol{\kappa}'(\delta\boldsymbol{\beta}) \left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \cdot \left(\left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \right)^{-1} \\
&\quad \times \left\{ (\mathbf{M}_F \boldsymbol{\Sigma} \mathbf{M}_F)^+ \right\}_{i,i} \boldsymbol{\kappa}(\delta\boldsymbol{\beta})
\end{aligned}$$

Platí následující tvrzení.

Věta 4.2 Jestliže $\Delta_i = 0$ a $\delta_{1,max}$ je řešení rovnice

$$P\{\chi_i^2(\delta_{1,max}) \geq \chi_1^2(0; 1 - \alpha)\} = \alpha + \varepsilon,$$

potom

$$\begin{aligned}
\delta\boldsymbol{\beta}^* &\in \left\{ \delta\boldsymbol{\beta} : \delta\boldsymbol{\beta}' \mathbf{C} \delta\boldsymbol{\beta} \leq 2 \frac{\sqrt{\delta_{1,max}}}{C_i^{(out)}(\boldsymbol{\beta}_0)} \right\} \\
&\Rightarrow P\{T_{out,i} \geq \chi_1^2(0; 1 - \alpha)\} \leq \alpha + \varepsilon.
\end{aligned}$$

Poznámka 4.3 Vzhledem k větě 4.2 linearizační oblast pro nulovou hypotézu $\Delta_i = 0$ v modelu (5) je

$$\mathcal{L}_{H_0,out,i} = \left\{ \delta\boldsymbol{\beta} : \delta\boldsymbol{\beta}' \mathbf{C} \delta\boldsymbol{\beta} \leq 2 \frac{\sqrt{\delta_{1,max}}}{C_i^{(out)}(\boldsymbol{\beta}_0)} \right\}.$$

Pokud se jedná o silofunkci testu $T_{out,i}$, linearizační oblast pro alternativu $\Delta_i \neq 0$ můžeme určit analogicky jako v sekci 3.

Jestliže $\Delta_i \neq 0$, potom testová statistika $T_{out,i}$ v modelu (5) má rozdělení $\chi_1^2(\delta_{(out,\Delta_i)})$, kde $\delta_{(out,\Delta_i)}$ je daná vztahem (6).

Nechť $\delta_{(\Delta_i, min)}$ je řešením rovnice

$$\begin{aligned} & P\{\chi_1^2(\delta_{(\Delta_i, min)}) \geq \chi_1^2(0; 1 - \alpha)\} = \\ & = P\left\{\chi_1^2\left(\Delta_i^2 \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i}\right) \geq \chi_1^2(0; 1 - \alpha)\right\} - \varepsilon. \end{aligned}$$

Protože

$$\begin{aligned} \forall \{\delta\beta \in R^k\} \frac{1}{4} \kappa'(\delta\beta) \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \cdot \left(\left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i}\right)^{-1} \\ \times \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \cdot \kappa(\delta\beta) \geq 0 \end{aligned}$$

a

$$\begin{aligned} \delta_{(out, \Delta_i)} & = \\ & = \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \Delta_i^2 + \Delta_i \left(\left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i}\right)^{-1} \\ & \quad \times \delta\beta' \mathbf{A}_i \delta\beta + \frac{1}{4} \kappa'(\delta\beta) \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \\ & \quad \times \left(\left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i}\right)^{-1} \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \cdot \kappa(\delta\beta), \end{aligned}$$

můžeme vyslovit následující tvrzení.

Věta 4.4 Jestliže

$$\begin{aligned} \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \Delta_i^2 + \Delta_i \left(\left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i}\right)^{-1} \\ \times \delta\beta' \mathbf{A}_i \delta\beta \geq \delta_{\Delta_i, min}, \end{aligned}$$

potom

$$\begin{aligned} & P\{\chi_1^2(\delta_{(out, \Delta_i)}) \geq \chi_1^2(0; 1 - \alpha)\} \geq \\ & \geq P\left\{\chi_1^2\left(\left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \Delta_i^2\right) \geq \chi_1^2(0; 1 - \alpha)\right\} - \varepsilon. \end{aligned}$$

Linearizční oblast pro test $T_{(out, i)}$ v alternativě $\Delta_i \neq 0$ definujeme následovně

$$\begin{aligned} \mathcal{L}_{(\Delta_i, out)} & = \left\{ \delta\beta : \delta\beta' \mathbf{A}_i \delta\beta \geq \frac{1}{\Delta_i} \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \right. \\ & \quad \left. \times \left[\delta_{(\Delta_i, min)} - \Delta_i^2 \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \right] \right\}. \end{aligned}$$

Protože hodnotu Δ_i předem neznáme, je rozumné vyšetřit funkci

$$\begin{aligned} g_{(i, out)}(\Delta) & = \\ & = \frac{1}{\Delta} \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \left[\delta_{(\Delta_i, min)} - \Delta^2 \left\{(\mathbf{M}_F \Sigma \mathbf{M}_F)^+\right\}_{i,i} \right], \\ & \Delta \in R^1 \end{aligned}$$

v dostatečně velkém okolí odhadu $\widehat{\Delta}_i$.

Vzhledem k lemmatu 3.3. odhad $\widehat{\Delta}_i$ and jeho konfidenční oblast umožní posoudit, jestli $\delta\beta^* \in \mathcal{L}_{(\Delta_i, out)} \cap \mathcal{L}_{(H_0, out)}$ s dostatečně vysokou pravděpodobností. Jestliže to platí, potom statistiku $T_{(out, i)}$ můžeme použít v nelineárním modelu (5).

References

- [1] Barnett, V., Lewis, T.: *Outliers in Statistical Data*. John Wiley, New York 1994.
- [2] Bates, D.M., Watts, D.G.: Relative curvature measures of nonlinearity. *J. Roy. Stat. Soc. B* 42, 1980, 1–25.
- [3] Kubáček, L., Kubáčková, L.: *Regression models with a weak nonlinearity*. Technical Report Nr. 1998, Universität Stuttgart, 1998, 1–67.
- [4] Kubáček, L., Kubáčková, L.: *Statistika a metrologie*. Vydavatelství Univerzity Palackého, Olomouc 2000.
- [5] Kubáček, L., Tesaříková, E.: *Variance components and nonlinearity* (submitted to *Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica* 45).
- [6] Rao, C.R.: *Linear Statistical Inference and Its Applications*. J. Wiley, New York-London-Sydney 1965.
- [7] Rao, C.R., Mitra, S.K.: *Generalized Inverse of the Matrix and Its Application*. John Wiley, New York 1971.
- [8] Scheffé, H.: *The Analysis of Variance*. John Wiley, New York-London-Sydney 1967 (fifth printing).
- [9] Zvára, K.: *Regresní analýza*. Academia, Praha 1989.



Eliminating transformations for nuisance parameters in linear regression models with constraints of the type I

PAVLA KUNDEROVÁ

*Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
Tomkova 40, 77900 Olomouc, Czech Republic
e-mail: kunderovl@inf.upol.cz*

Abstrakt: *The linear regression model in which the vector of the first order parameter is divided into two parts: to the vector of the useful parameters and to the vector of the nuisance parameters is considered. Constraints of the type I are given on the useful parameters. We examine eliminating transformations which eliminate the nuisance parameters without loss of information on the useful parameters.*

1991 Mathematics Subject Classification: 62J05

Key words: regular linear regression model, nuisance parameters, BLUE, constraints

1 Introduction, notations

The following notation will be used throughout the paper:

R^n	the space of all n -dimensional real vectors;
$u_p, A_{m,n}$	the real column p -dimensional vector, the real $m \times n$ matrix;
$A', r(A)$	the transpose, the rank of the matrix A ;
$\mathcal{M}(A), Ker(A)$	the range, the null space of the matrix A ;
A^-	a generalized inverse of a matrix A (satisfying $AA^-A = A$);
A^+	the Moore-Penrose generalized inverse of a matrix A (satisfying $AA^+A = A, A^+AA^+ = A^+, (AA^+)' = AA^+, (A^+A)' = A^+A$);
P_A	the orthogonal projector in the Euclidean norm onto $\mathcal{M}(A)$;
$M_A = I - P_A$	the orthogonal projector in the Euclidean norm onto $\mathcal{M}^\perp(A)$;
I_k	the $k \times k$ identity matrix;
$O_{m,n}$	the $m \times n$ null matrix.

If $\mathcal{M}(\mathbf{A}) \subset \mathcal{M}(\mathbf{S})$, \mathbf{S} p.s.d., then the symbol $\mathbf{P}_A^{S^-}$ denotes the projector projecting vectors in $\mathcal{M}(\mathbf{S})$ onto $\mathcal{M}(\mathbf{A})$ along $\mathcal{M}(\mathbf{S}\mathbf{A}^\perp)$. A general representation of all such projectors $\mathbf{P}_A^{S^-}$ is given by $\mathbf{A}(\mathbf{A}'\mathbf{S}^-\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}^- + \mathbf{B}(\mathbf{I} - \mathbf{S}\mathbf{S}^-)$, where \mathbf{B} is arbitrary, (see [6], (2.14)). $\mathbf{M}_A^{S^-} = \mathbf{I} - \mathbf{P}_A^{S^-}$.

Let $\mathbf{N}_{n,n}$ is p.d. (p.s.d.) matrix and $\mathbf{A}_{m,n}$ an arbitrary matrix, then the symbol $\mathbf{A}_{m(N)}^-$ denotes the matrix satisfying $\mathbf{A}\mathbf{A}_{m(N)}^-\mathbf{A} = \mathbf{A}$ and $\mathbf{N}\mathbf{A}_{m(N)}^-\mathbf{A} = (\mathbf{N}\mathbf{A}_{m(N)}^-\mathbf{A})'$. [$\mathbf{A}_{m(N)}^- \mathbf{y}$ is any solution of the consistent system $\mathbf{A}\mathbf{x} = \mathbf{y}$ whose \mathbf{N} -seminorm is minimal, see [3], p.151].

Assertion 1 (see [2], Lemma 10.1.35)

Let \mathbf{X} be any $n \times k$ matrix and Σ an $n \times n$ p.s.d. matrix.

(i) If Σ is p.d., then

$$(\mathbf{M}_X \Sigma \mathbf{M}_X)^+ = \Sigma^{-1} - \Sigma^{-1} \mathbf{X} (\mathbf{X}' \Sigma^{-1} \mathbf{X})^{-1} \mathbf{X}' \Sigma^{-1} = \Sigma^{-1} \mathbf{M}_X^{\Sigma^{-1}}.$$

(ii)

$$(\mathbf{M}_X \Sigma \mathbf{M}_X)^+ = \mathbf{M}_X (\mathbf{M}_X \Sigma \mathbf{M}_X)^+ = (\mathbf{M}_X \Sigma \mathbf{M}_X)^+ \mathbf{M}_X = \mathbf{M}_X (\mathbf{M}_X \Sigma \mathbf{M}_X)^+ \mathbf{M}_X.$$

Let us consider following linear model with nuisance parameters

$$\mathbf{Y} \sim [(\mathbf{X}, \mathbf{S}) \begin{pmatrix} \beta \\ \kappa \end{pmatrix}, \Sigma], \quad \Sigma \text{ known matrix}, \quad (1)$$

where $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)'$ is a random observation vector; $\beta \in R^k$ is a vector of the useful parameters; $\kappa \in R^l$ is a vector of the nuisance parameters; $\mathbf{X}_{n,k}$ is a design matrix belonging to the vector β ; $\mathbf{S}_{n,l}$ is a design matrix belonging to the vector κ .

We suppose that

1. $E(\mathbf{Y}) = \mathbf{X}\beta + \mathbf{S}\kappa$, $\forall \beta \in R^k$, $\forall \kappa \in R^l$,
2. $var(\mathbf{Y}) = \Sigma$ is not a function of the vector $(\beta', \kappa)'$.

If matrix Σ is positive definite and $r(\mathbf{X}, \mathbf{S}) = k + l < n$, the model is said to be *regular*, (see [2], p.13).

Parametric function $f'\beta$ is unbiasedly estimable in model (1) iff $f \in \mathcal{M}(\mathbf{X}'\mathbf{M}_S)$, see [5], Remark 2.

There are situations in the practice that auxiliary information on the vector of useful regression coefficients β is known, it means that the parametric space for β is not R^k but its subset only,

$$\beta \in \{\mathbf{u} \in R^k : \mathbf{b} + \mathbf{B}\mathbf{u} = \mathbf{o}\}, \quad (2)$$

where \mathbf{B} is a $q \times k$ known matrix. Since no assumption on the $r(\mathbf{B})$ is considered, it must be assumed that a given q -dimensional vector \mathbf{b} satisfies $\mathbf{b} \in \mathcal{M}(\mathbf{B})$. This constraints on the useful parameters will be called *constraints of the type I*.

In the literature there are investigated properties of estimators of the parameters β, κ in model (1) under constraints (2), see for example [1], [4]. In cases when we are interested on useful parameters only it is possible to simplify model (1) by the propriate eliminating transformation, see [2], [3], [5].

In this paper we unite both of the procedures mentioned. Firstly we use eliminating transformation and then we add constraints to the transformed model.

2 Constraints of the type I in the transformed

Our task will be to eliminate the matrix S belonging to the vector of nuisance parameters, i.e. we consider the following class of eliminating matrices

$$\mathcal{T} = \{T : TS = O\},$$

where T is matrix of the type $r \times n$.

That leads us to linear models

$$TY \sim [TX\beta, T\Sigma_{\vartheta}T']. \quad (3)$$

If we now add constraints (2) to the model (3), we get model

$$\begin{pmatrix} TY \\ -b \end{pmatrix} \sim \left[\begin{pmatrix} TX \\ B \end{pmatrix} \beta, \begin{pmatrix} T\Sigma_{\vartheta}T' & O \\ O & O \end{pmatrix} \right]. \quad (4)$$

Remark 1 Linear function $f'\beta + a, f \in R^k, a \in R$ is unbiasedly estimable in model (4), iff

$$f \in \mathcal{M}(X'T', B').$$

It can be proved as follows: $f'\beta$ is in model (4) unbiasedly estimable iff there exists statistic $g'TY + c, g \in R^r, c \in R$ such that

$$E(g'TY + c) = g'TX\beta + c = f'\beta + a, \forall \beta \Leftrightarrow (g'TX - f')\beta + (c - a) = 0$$

$$\Leftrightarrow \text{there exists vector } k \in R^q \text{ such that } k'B = g'TX - f' \wedge k'b = c - a.$$

Because c can be chosen arbitrarily, the necessary and sufficient condition for unbiasedly estimable function is

$$g'TX - k'B = f' \Leftrightarrow f = X'T'g - B'k \Leftrightarrow f \in \mathcal{M}(X'T', B').$$

Theorem 1 For the BLUE in the model (4) holds

$$\widehat{TX}\beta = P_{TXM_{B'}}^{[T(\Sigma + XM_{B'}X')T']^+} TY - M_{TXM_{B'}}^{[T(\Sigma + XM_{B'}X')T']^+} TXB'(BB')^{-1}b.$$

Proof. According to Theorem 3.1.3. in [2]

$$\begin{pmatrix} \widehat{TX} \\ B \end{pmatrix} \beta = \begin{pmatrix} TX \\ B \end{pmatrix} \left[(X'T', B')_{m(T\Sigma T', O)}^- \right]' \begin{pmatrix} TY \\ -b \end{pmatrix}$$

$$= \begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix} \left[(\text{X}'\text{T}', \text{B}') \left\{ \begin{pmatrix} \text{T}\Sigma\text{T}', & \text{O} \\ \text{O}, & \text{O} \end{pmatrix} + \begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix} (\text{X}'\text{T}', \text{B}') \right\}^{-} \begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix} \right]^{-} \\ \times (\text{X}'\text{T}', \text{B}') \left\{ \begin{pmatrix} \text{T}\Sigma\text{T}', & \text{O} \\ \text{O}, & \text{O} \end{pmatrix} + \begin{pmatrix} \text{TX}\text{X}'\text{T}', & \text{TX}\text{B}' \\ \text{BX}'\text{T}', & \text{BB}' \end{pmatrix} \right\}^{-} \begin{pmatrix} \text{TY} \\ -\text{b} \end{pmatrix},$$

where the formula

$$\text{A}_{m(N)}^{-} = (\text{N} + \text{A}'\text{A})^{-} \text{A}' [\text{A}(\text{N} + \text{A}'\text{A})^{-} \text{A}']^{-},$$

was used. By the help of the Rohde's formula for g-inverse of the p.s.d. partitioned matrix (see [2], Theorem 4.1.40) we can write

$$\begin{pmatrix} \text{T}[\Sigma + \text{X}\text{X}']\text{T}', & \text{TX}\text{B}' \\ \text{BX}'\text{T}', & \text{BB}' \end{pmatrix}^{-} = \begin{pmatrix} \boxed{11}, & \boxed{12} \\ \boxed{21}, & \boxed{22} \end{pmatrix},$$

where

$$\begin{aligned} \boxed{11} &= [\text{T}(\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}')\text{T}']^{-}, \\ \boxed{12} &= -[\text{T}(\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}')\text{T}']^{-} \text{TX}\text{B}'(\text{BB}')^{-} \\ \boxed{21} &= -(\text{BB}')^{-} \text{BX}'\text{T}'[\text{T}(\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}')\text{T}']^{-} \\ \boxed{22} &= (\text{BB}')^{-} + (\text{BB}')^{-} \text{BX}'\text{T}'[\text{T}(\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}')\text{T}']^{-} \text{TX}\text{B}'(\text{BB}')^{-}. \end{aligned}$$

Then (we use Moore-Penrose g-inverse matrix for the sake of simplicity)

$$\begin{aligned} &\left[(\text{X}'\text{T}', \text{B}') \begin{pmatrix} \boxed{11}, & \boxed{12} \\ \boxed{21}, & \boxed{22} \end{pmatrix} \begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix} \right]^{+} \\ &= [\text{M}_{\text{B}'}\text{X}'\text{T}'(\text{T}[\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}']\text{T}')^{-} \text{TX}\text{M}_{\text{B}'} + \text{P}_{\text{B}'}]^{+}, \end{aligned}$$

thus

$$\begin{aligned} &\widehat{\begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix}} \beta \\ &= \begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix} \{ [\text{M}_{\text{B}'}\text{X}'\text{T}'(\text{T}[\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}']\text{T}')^{-} \text{TX}\text{M}_{\text{B}'}]^{+} + \text{P}_{\text{B}'} \} (\text{X}'\text{T}', \text{B}') \\ &\quad \times \begin{pmatrix} \boxed{11}, & \boxed{12} \\ \boxed{21}, & \boxed{22} \end{pmatrix} \begin{pmatrix} \text{TY} \\ -\text{b} \end{pmatrix}. \end{aligned}$$

After some calculations we get

$$\widehat{\begin{pmatrix} \text{TX} \\ \text{B} \end{pmatrix}} \beta = \begin{pmatrix} \text{P}_{\text{TX}\text{M}_{\text{B}'}}^{[\text{T}(\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}')\text{T}']^{+}} \text{TY} - \text{M}_{\text{TX}\text{M}_{\text{B}'}}^{[\text{T}(\Sigma + \text{X}\text{M}_{\text{B}'}\text{X}')\text{T}']^{+}} \text{TX}\text{B}'(\text{BB}')^{-}\text{b} \\ -\text{b} \end{pmatrix}.$$

In the course of the proof following assertion has been used

$$\text{A}'\text{B} = \text{O} \wedge \text{BA}' = \text{O} \Rightarrow (\text{A} + \text{B})^{+} = \text{A}^{+} + \text{B}^{+}.$$

□

Theorem 2 The covariance matrix of the estimator $\widehat{\text{TX}}\beta$ in model (4) is

$$\text{var}[\widehat{\text{TX}}\beta] = \text{TX}\{[\text{M}_{B'}\text{X}'\text{T}'(\text{T}[\Sigma + \text{XM}_{B'}\text{X}']\text{T}')^{-1}\text{TXM}_{B'}]^+ - \text{M}_{B'}\}\text{X}'\text{T}'.$$

Proof

$$\begin{aligned} \text{var}[\widehat{\text{TX}}\beta] &= \text{P}_{\text{TXM}_{B'}}^{[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+} [\text{T}\Sigma\text{T}' + \text{TXM}_{B'}\text{X}'\text{T}' - \text{TXM}_{B'}\text{X}'\text{T}'] (\text{P}_{\text{TXM}_{B'}}^{[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+})' \\ &= \text{TXM}_{B'} (\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'})^+ \text{M}_{B'}\text{X}'\text{T}' [\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \\ &\quad \times [\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}'] [\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'} \\ &\quad \times (\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{X}'\text{T}'\text{M}_{B'})^+ \text{M}_{B'}\text{X}'\text{T}' \\ &\quad - \text{TXM}_{B'} (\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'})^+ \text{M}_{B'}\text{X}'\text{T}' [\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'} \\ &\quad \times \text{M}_{B'}\text{X}'\text{T}' [\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'} (\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'})^+ \text{M}_{B'}\text{X}'\text{T}' \\ &= \text{TXM}_{B'} (\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'})^+ \text{M}_{B'}\text{X}'\text{T}' - \text{TXM}_{B'}\text{X}'\text{T}' \\ &= \text{TX} \left\{ (\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ \text{TXM}_{B'})^+ - \text{M}_{B'} \right\} \text{X}'\text{T}'. \end{aligned}$$

In the course of the proof we have used Assertion 1, (ii) and following statement

$$\mathcal{M}(\text{TXM}_{B'}) \subset \mathcal{M}(\text{T}\Sigma\text{T}' + \text{TXM}_{B'}\text{X}'\text{T}') \Leftrightarrow$$

$$\text{M}_{B'}\text{X}'\text{T}'[\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}']^+ [\text{T}(\Sigma + \text{XM}_{B'}\text{X}')\text{T}'] = \text{M}_{B'}\text{X}'\text{T}'.$$

□

Remark 2 If we change the ordering of the procedures described at the beginning of this section, we get the same model. Indeed by joining linear model (1) with constraints (2), we can write

$$\begin{pmatrix} \text{Y} \\ -b \end{pmatrix} \sim \left[\begin{pmatrix} \text{X} & \text{S} \\ \text{B} & \text{O} \end{pmatrix} \begin{pmatrix} \beta \\ \kappa \end{pmatrix}, \begin{pmatrix} \Sigma & \text{O} \\ \text{O} & \text{O} \end{pmatrix} \right].$$

The transformation by the matrix $\begin{pmatrix} \text{T} & \text{O} \\ \text{O} & \text{I} \end{pmatrix}$, such that $\text{TS} = \text{O}$, leads to the model (5).

3 Examples of the transformation matrices

The general solution of the matrix equation $\text{TS} = \text{O}$ is of the form

$$\text{T} = \text{A}(\text{I} - \text{SS}^-),$$

where A is an arbitrary matrix of the corresponding type, S^- is some version of generalized inverse of the matrix S .

If we choose $S^- = (S^-WS)^-S'W$, where W is an arbitrary p.s.d. matrix such that

$$\mathcal{M}(S') = \mathcal{M}(S'WS),$$

then $T = AM_S^W$, A arbitrary, where M_S^W is given uniquely.

First we confine us to the transformation matrix

a) $T = M_S^W$,

i.e. we consider transformed linear model

$$M_S^W Y \sim [M_S^W X \beta, M_S^W \Sigma (M_S^W)'] \quad (5)$$

Thus model with constraints of the type I is following

$$\begin{pmatrix} M_S^W Y \\ -b \end{pmatrix} \sim \left[\begin{pmatrix} M_S^W X \\ B \end{pmatrix} \beta, \begin{pmatrix} M_S^W \Sigma \vartheta (M_S^W)', & 0 \\ 0, & 0 \end{pmatrix} \right] \quad (6)$$

It can be proved (see [5], chapter 3) that

$$\mathcal{M}(M_S) = \mathcal{M}((M_S^W)'),$$

thus

$$\mathcal{M}(X'M_S, B') = \mathcal{M}(X'(M_S^W)', B'),$$

i.e. the classes of the estimable functions $g'\beta$ in model (1) with constraints (2) and in model (6) are identical.

According to Theorem 1, Theorem 2

$$\begin{aligned} \widehat{M_S^W X \beta} &= P_{M_S^W X M_{B'}}^{[M_S^W (\Sigma + X M_{B'} X') (M_S^W)']^+} M_S^W [Y + X B' (B B')^{-1} b] - M_S^W X B' (B B')^{-1} b \\ &= M_S^W X M_{B'} \left[M_{B'} X' (M_S^W)' \left(M_S^W (\Sigma + X M_{B'} X') (M_S^W)' \right)^+ M_S^W X M_{B'} \right]^+ \\ &\times X' (M_S^W)' \left(M_S^W (\Sigma + X M_{B'} X') (M_S^W)' \right)^+ M_S^W [Y + X B' (B B')^{-1} b] - M_S^W X B' (B B')^{-1} b. \\ &\quad \text{var}[\widehat{M_S^W X \beta}] \\ &= M_S^W X \left\{ \left[M_{B'} X' (M_S^W)' \left(M_S^W [\Sigma + X M_{B'} X'] (M_S^W)' \right)^- M_S^W X M_{B'} \right]^+ - M_{B'} \right\} X' (M_S^W)'. \end{aligned}$$

Remark 3 If the matrix $\Sigma + X M_{B'} X'$ is regular or if

$$\mathcal{M}(S) \subset \mathcal{M}(\Sigma + X M_{B'} X'),$$

it can be proved that (see [5], Lemma 1)

$$(M_S^W)' \left[M_S^W (\Sigma + X M_{B'} X') (M_S^W)' \right]^+ M_S^W = [M_S (\Sigma + X M_{B'} X') M_S]^+.$$

Then

$$\begin{aligned} \widehat{\mathbf{M}}_S^W \mathbf{X} \beta &= \mathbf{M}_S^W \mathbf{X} (\mathbf{M}_{B'} \mathbf{X}' [\mathbf{M}_S (\boldsymbol{\Sigma} + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}') \mathbf{M}_S]^+ \mathbf{X} \mathbf{M}_{B'})^+ \mathbf{X}' [\mathbf{M}_S (\boldsymbol{\Sigma} + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}') \mathbf{M}_S]^+ \\ &\quad \times (\mathbf{Y} + \mathbf{X} \mathbf{B}' (\mathbf{B} \mathbf{B}')^{-1} \mathbf{b}) - \mathbf{M}_S^W \mathbf{X} \mathbf{B}' (\mathbf{B} \mathbf{B}')^{-1} \mathbf{b}. \end{aligned}$$

$$\text{var}[\widehat{\mathbf{M}}_S^W \mathbf{X} \beta] = \mathbf{M}_S^W \mathbf{X} \{ \mathbf{M}_{B'} \mathbf{X}' [\mathbf{M}_S (\boldsymbol{\Sigma} + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}') \mathbf{M}_S]^+ \mathbf{X} \mathbf{M}_{B'} \}^+ - \mathbf{M}_{B'}' \} \mathbf{X}' (\mathbf{M}_S^W)'$$

When we choose transformation matrix

$$\mathbf{b) \quad T} = \mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+},$$

we get the model with constraints

$$\begin{pmatrix} \mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \mathbf{Y} \\ -\mathbf{b} \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{X} \\ \mathbf{B} \end{pmatrix} \beta, \begin{pmatrix} \mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \boldsymbol{\Sigma} (\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+})', \mathbf{0} \\ \mathbf{0}, \mathbf{0} \end{pmatrix} \right],$$

because it is

$$\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \mathbf{S} = \mathbf{0}, \quad \mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \mathbf{X} = \mathbf{X}.$$

In this model

$$\begin{aligned} \widehat{\mathbf{X}} \beta &= \mathbf{P}_{\mathbf{X} \mathbf{M}_{B'}}^{[\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \boldsymbol{\Sigma} (\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+})' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}']^+} \mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \mathbf{Y} \\ &\quad - \mathbf{M}_{\mathbf{X} \mathbf{M}_{B'}}^{[\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \boldsymbol{\Sigma} (\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+})' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}']^+} \mathbf{X} \mathbf{B}' (\mathbf{B} \mathbf{B}')^{-1} \mathbf{b}, \end{aligned}$$

$$\text{var}[\widehat{\mathbf{X}} \beta] = \mathbf{X} \left\{ \left[\mathbf{M}_{B'} \mathbf{X}' \left(\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+} \boldsymbol{\Sigma} (\mathbf{M}_S^{(M_X \boldsymbol{\Sigma} M_X)^+})' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}' \right)^- \mathbf{X} \mathbf{M}_{B'} \right]^+ - \mathbf{M}_{B'} \right\} \mathbf{X}'.$$

If we suppose, that

$$\mathcal{M}(\mathbf{X}') \subset \mathcal{M}(\mathbf{X}' [\mathbf{M}_S \boldsymbol{\Sigma} \mathbf{M}_S]^+ \mathbf{X}), \quad (7)$$

we can use transformation matrix

$$\mathbf{c) \quad T} = \mathbf{P}_X^{(M_S \boldsymbol{\Sigma} M_S)^+}$$

that leads to the model

$$\begin{pmatrix} \mathbf{P}_X^{(M_S \boldsymbol{\Sigma} M_S)^+} \mathbf{Y} \\ -\mathbf{b} \end{pmatrix} \sim \left[\begin{pmatrix} \mathbf{X} \\ \mathbf{B} \end{pmatrix} \beta, \begin{pmatrix} \mathbf{X} (\mathbf{X}' [\mathbf{M}_S \boldsymbol{\Sigma} \mathbf{M}_S]^+ \mathbf{X})^{-1} \mathbf{X}', \mathbf{0} \\ \mathbf{0}, \mathbf{0} \end{pmatrix} \right],$$

because under assumption (7) it is

$$\mathbf{P}_X^{(M_S \boldsymbol{\Sigma} M_S)^+} \mathbf{X} = \mathbf{X}, \quad \mathbf{P}_X^{(M_S \boldsymbol{\Sigma} M_S)^+} \mathbf{S} = \mathbf{0},$$

$$\begin{aligned}
& \mathbf{P}_X^{(M_S \Sigma M_S)^+} \Sigma (\mathbf{P}_S^{(M_S \Sigma M_S)^+})' = \mathbf{X}(\mathbf{X}'[\mathbf{M}_S \Sigma \mathbf{M}_S]^+ \mathbf{X})^{-} \mathbf{X}'. \\
& \widehat{\mathbf{X}}\beta = \mathbf{P}_{\mathbf{X} \mathbf{M}_{B'}}^{[\mathbf{X}(\mathbf{X}'[\mathbf{M}_S \Sigma \mathbf{M}_S]^+ \mathbf{X})^{-} \mathbf{X}' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}']^+} \mathbf{P}_X^{(M_S \Sigma M_S)^+} \mathbf{Y} \\
& \quad - \mathbf{M}_{\mathbf{X} \mathbf{M}_{B'}}^{[\mathbf{X}(\mathbf{X}'[\mathbf{M}_S \Sigma \mathbf{M}_S]^+ \mathbf{X})^{-} \mathbf{X}' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}']^+} \mathbf{X} \mathbf{B}' (\mathbf{B} \mathbf{B}')^{-} \mathbf{b} \\
& = \left\{ \mathbf{X} \mathbf{M}_{B'} \left[\mathbf{M}_{B'} \mathbf{X}' (\mathbf{X}[\mathbf{X}'(\mathbf{M}_S \Sigma \mathbf{M}_S)^+ \mathbf{X}]^{-} \mathbf{X}' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}')^+ \mathbf{X} \mathbf{M}_{B'} \right]^{-} \mathbf{M}_{B'} \mathbf{X}' \right. \\
& \quad \left. - \mathbf{X} \mathbf{M}_{B'} \mathbf{X}' \right\} (\mathbf{M}_S \Sigma \mathbf{M}_S)^+ [\mathbf{Y} + \mathbf{X} \mathbf{B}' (\mathbf{B} \mathbf{B}')^{-} \mathbf{b}] - \mathbf{X} \mathbf{B}' (\mathbf{B} \mathbf{B}')^{-} \mathbf{b}. \\
& \quad \quad \quad \text{var}[\widehat{\mathbf{X}}\beta] \\
& = \left\{ \mathbf{X} \left[\mathbf{M}_{B'} \mathbf{X}' (\mathbf{X}[\mathbf{X}'(\mathbf{M}_S \Sigma \mathbf{M}_S)^+ \mathbf{X}]^{-} \mathbf{X}' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}')^+ \mathbf{X} \mathbf{M}_{B'} \right]^+ \mathbf{X}' - \mathbf{X} \mathbf{M}_{B'} \mathbf{X}' \right\} (\mathbf{M}_S \Sigma \mathbf{M}_S)^+ \\
& \quad \times \left(\left\{ \mathbf{X} \left[\mathbf{M}_{B'} \mathbf{X}' (\mathbf{X}[\mathbf{X}'(\mathbf{M}_S \Sigma \mathbf{M}_S)^+ \mathbf{X}]^{-} \mathbf{X}' + \mathbf{X} \mathbf{M}_{B'} \mathbf{X}')^+ \mathbf{X} \mathbf{M}_{B'} \right]^+ \mathbf{X}' - \mathbf{X} \mathbf{M}_{B'} \mathbf{X}' \right\}' \right)'.
\end{aligned}$$

Remark 4 For the classes of unbiasedly estimable functions in model

$$\mathbf{P}_X^{(M_S \Sigma M_S)^+} \mathbf{Y} \sim [\mathbf{P}_X^{(M_S \Sigma M_S)^+} \mathbf{X}\beta, \mathbf{P}_X^{(M_S \Sigma M_S)^+} \Sigma (\mathbf{P}_X^{(M_S \Sigma M_S)^+})']',$$

and in model (1) is valid

$$\mathcal{M}[\mathbf{X}'(\mathbf{P}_X^{(M_S \Sigma M_S)^+})'] \subset \mathcal{M}(\mathbf{X}'\mathbf{M}_S),$$

because if $\mathbf{y} \in \mathcal{M}[(\mathbf{P}_X^{(M_S \Sigma M_S)^+})']$, there exists $\mathbf{u} \in R^n$ such that $\mathbf{y} = (\mathbf{P}_X^{(M_S \Sigma M_S)^+})'\mathbf{u}$. Then $\mathbf{y}'\mathbf{S} = \mathbf{u}'\mathbf{P}_X^{(M_S \Sigma M_S)^+} \mathbf{S} = \mathbf{o}$, i.e. $\mathbf{y} \in \mathcal{M}(\mathbf{M}_S)$ and thus $\mathcal{M}[(\mathbf{P}_X^{(M_S \Sigma M_S)^+})'] \subset \mathcal{M}(\mathbf{M}_S)$.

4 Another transformations of regression models with nuisance parameters

Let us consider model (1)

$$\mathbf{Y} \sim [(\mathbf{X}, \mathbf{S}) \begin{pmatrix} \beta \\ \kappa \end{pmatrix}, \Sigma], \quad \Sigma \text{ known matrix.}$$

I. Let the covariance matrix Σ of the observation vector is p.s.d. matrix and let $\Sigma = \mathbf{U}\mathbf{U}'$, where \mathbf{U} is a matrix of the type $n \times r(\Sigma)$.

Suppose that we transform model (1) by the matrix \mathbf{T} such that

$$\mathbf{T}\mathbf{X} = \mathbf{X}, \quad \mathbf{T}\mathbf{S} = \mathbf{O}, \quad \mathbf{T}\mathbf{U} = \mathbf{U},$$

then we obtain regression model without nuisance parameters and with the same covariance matrix

$$TY \sim [TX\beta, \Sigma].$$

How to find the matrix T ?

The equation $T(U, S) = (U, O)$ is solvable if and only if (see [2], Theorem 10.1.6, Theorem 10.1.13)

$$(U, O)(U, S)^-(U, S) = (U, O) \Leftrightarrow \mathcal{M} \begin{pmatrix} U' \\ O \end{pmatrix} \subset \mathcal{M} \begin{pmatrix} U' \\ S' \end{pmatrix},$$

i.e. if there exists matrix E such that

$$\begin{pmatrix} U' \\ O \end{pmatrix} = \begin{pmatrix} U'E \\ S'E \end{pmatrix},$$

it means that

$$S'E = O \Leftrightarrow \mathcal{M}(E) \subset Ker(S') = \mathcal{M}(M_S),$$

$$U'M_S A = U' \text{ for some } A \Rightarrow \mathcal{M}(U') \subset \mathcal{M}(U'M_S) \Rightarrow \mathcal{M}(U') = \mathcal{M}(U'M_S).$$

It follows from the last equality (see [2], Theorem 10.1.34) that

$$\mathcal{M}(U) \cap \mathcal{M}(S) = \{o\}.$$

If we will use this transformation, we can follow the statements in [2], Chapter 4, 4.2. by substitution $X \rightarrow U$.

II. Let us consider the situation described in **I**. Transformation by the matrix T satisfying

$$TX = X, \quad TS = O \quad TU = U,$$

gives the model

$$TY \sim [X\beta, \Sigma].$$

The equation $T(X, U, S) = (X, U, O)$ is solvable iff

$$(X, U, O)(XU, S)^-(X, U, S) = (X, U, O) \Leftrightarrow \mathcal{M} \begin{pmatrix} X' \\ U' \\ O \end{pmatrix} \subset \mathcal{M} \begin{pmatrix} X' \\ U' \\ S' \end{pmatrix},$$

i.e. iff there exists matrix F such that

$$X' = X'F, \quad U' = U'F, \quad S'F = O,$$

it is iff

$$\begin{aligned} X' &= X'M_S B \text{ for some } B, \quad U' = U'M_S B \\ \Leftrightarrow \mathcal{M}(X') &\subset \mathcal{M}(X'M_S), \quad \mathcal{M}(U') \subset \mathcal{M}(U'M_S) \\ \Leftrightarrow \mathcal{M}(X) \cap \mathcal{M}(S) &= \{o\}, \quad \mathcal{M}(\Sigma) \cap \mathcal{M}(S) = \{o\}. \end{aligned}$$

In situations satisfying this conditions, we can use theory in [2], Chapter 4, 4.2. by substituting $X \rightarrow (X, U)$. Examples of the useful transformation matrices are

$$T = M_S^{[M_{(X,U)} W M_{(X,U)}]^+} \quad \text{or} \quad T = P_{(X,U)}^{[M_S W M_S]^+}, \quad W \text{ arbitrary p.d. matrix.}$$

References

- [1] Fišerová, E., Kubáček, L., Kunderová, P.: *Linear statistical models: regularity and singularities*. Accepted in Academia, Praha.
- [2] Kubáček, L., Kubáčková, L., Volaufová, J.: *Statistical models with linear structures*. Veda, Publishing House of the Slovak Academy of Sciences, Bratislava 1995.
- [3] Kubáček, L., Kubáčková, L.: *Statistika a metrologie*. Nakladatelství Palackého univerzity, 2000.
- [4] Kunderová, P.: *Regular linear model with the nuisance parameters with constraints of the type I*. Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica **40** (2001), 151-159
- [5] Kunderová, P.: *Eliminating transformations for nuisance parameters in linear model*. Acta Univ. Palacki. Olomuc., Fac. rer. nat., Mathematica **42** (2003), 59 - 68.
- [6] Nordström, K., Fellman, J.: *Characterizations and Dispersion-Matrix Robustness of Efficiently Estimable Parametric Functionals in Linear Models with Nuisance Parameters*. Linear Algebra and its Applications **127** (1990), 341–361.



Edikt o zeměpisné délce a první statistické algoritmy pro aproximaci dat

JAROSLAV MAREK¹ A ROMAN PAWIŃSKI²

¹ *Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta Univerzity Palackého
Tomkova 40, 779 00 Olomouc
e-mail: marek@inf.upol.cz*

² *Nejvyšší kontrolní úřad, Jakovcova 1518/2, 170 00 Praha
e-mail: romario_@email.cz*

Abstrakt: Cílem článku je uvést některé historické postupy pro zpracování dat a uvést úlohy, které přispěly k vytvoření těchto algoritmů.

V příspěvku budou prezentovány první pokusy o řešení nekonzistentních rovnic a první algoritmy pro aproximaci dat přímkou. Na úloze měření délky polodníku bude provedeno vzájemné porovnání těchto algoritmů a výsledky budou také porovnány s odhadem získaným metodou nejmenších čtverců.

Velkým impulsem pro hledání těchto algoritmů byla astronomická měření shromážděná v souvislosti se Zeměpisnou soutěží anglické královny Anny (viz [2]).

Klíčová slova: Mayerova metoda průměrů, Boškovičova metoda, Laplaceova metoda, Lambertova metoda, metoda nejmenších čtverců, lineární regrese.

1 Soutěž o zeměpisné délce a historie aproximace dat přímkou

Koncem 18. století se nahromadila astronomická pozorování planet, podobně se nahromadil bohatý materiál ze stupňových měření k určení rozměrů Země čekající na vyrovnání. Metoda Cotesova, medián ani přibližný způsob skupinového vyrovnání, které navrhl Mayer (17. 2. 1723 – 20. 2. 1762), nemohly uspokojit teorii ani geodetickou praxi, které žádaly také odhad přesnosti a stupně spolehlivosti výsledků měření a vyrovnání.

Další metody počet navrhl Joseph Louis Lagrange, Roger Josip Boškovič, Johann Heinrich Lambert a Pierre Simone Laplace.

V další kapitole se budeme některými algoritmy zabývat podrobněji (Mayer – str. 100, Boškovič – str. 103, Laplace – str. 110, Lambert – str. 108).

Popsaný stav byl také mocným podnětem k hledání nevhodnější metody vyrovnání a vzniku MNČ.

Velkým impulsem pro hledání těchto algoritmů byla také astronomická měření shromážděná při Zeměpisné soutěži anglické královny Anny (viz [2]).

2 Vypsání soutěže

Soutěž byla vypsána poté, co anglická flotila se Sirem Clowdisleym na lodi Association mylně stanovila zeměpisnou délku a ztroskotala u ostrovů Scilly cca dvacet mil od jihozápadního cípu Anglie. Oné mlhavé noci 22. října 1707 se tyto ostrůvky staly náhrobním kamenem bez nápisů pro dva tisíce mužů z vojska Sira Shovella Clowdisleyho. Jen dva muži se dostali na břeh živí. Jedním z nich byl sám Sir Clowdisley. Jakmile však omdlel vyčerpáním na pobřežním písku, údajně našla jeho tělo jakási místní žena, která prohledávala pláž. Zalíbil se jí smaragdový prsten na jeho ruce. Její chtíč a jeho vyčerpání byly pomocníky, díky nimž admirála bez obtíží zavraždila. Pláž se dnes nazývá Land's End.

Ztráta flotily jen korunovala dlouhou ságu mořeplaveckých strastí, které námořníky doprovázely předtím, než dokázali stanovit polohu podle zeměpisné délky. Tato ztráta, k níž došlo v bezprostřední blízkosti námořních center Anglie, katapultovala problém stanovení zeměpisné délky na čelné místo žebříčku státního zájmu.

Stín duší námořníků Sira Clowdisleyho uspišil vydání slavného Ediktu o zeměpisné délce z roku 1714, v němž byla přislíbena odměna 20 000 liber sterlingů¹ za vyřešení problému zeměpisné délky s přesností na půl stupně hlavní kružnice, 15 000 liber sterlingů za metodu s přesností na dvě třetiny stupně a 10 000 liber za metodu s přesností na jeden stupeň.

8. července 1714 ve Westminsterském paláci na zasedání vytvořené parlamentní komise čte Edmond Halley expertní posudek Sira Isaaca Newtona. V referátu Newton shrnul existující prostředky pro měření zeměpisné délky a prohlásil o nich, že veškeré jsou teoreticky správné ale obtížně proveditelné. Jedním (postupem) je přesné měření času hodinami, avšak z důvodu pohybu lodí, teplotních odchylek, změn vlhkosti a rozdílu v zemské přitažlivosti na rozdílných zeměpisných šířkách nebyly dosud takové hodiny vyrobeny a ani s největší pravděpodobností nebudou. Aby si jejich strůjce zasloužil cenu 20 000 liber, nemohli by se od přesného času odchýlit o více než 3 sekundy za 24 hodin.

Podle tehdejších znalostí a přesvědčení mohla odpověď přijít pouze z oblohy, tedy z časomíry vesmíru, a nikoliv z obyčejných kyvadlových hodin.

Byla ustanovena Rada pro zeměpisnou délku. V té se sešli vědci, námořní důstojníci a vládní úředníci — jejich úkolem bylo dohlížet na udílení ceny. Členem Rady se stává i člen profesorského sboru Cambridge *Isaac Newton (4. 1. 1643 – 31. 3. 1727)*. Tato rada podle Ediktu mohla udílet odměny pro financování nadějných nápadů vedoucích k řešení.

¹Odpovídá dnešní hodnotě několika milionů dolarů.

Naděje na řešení je vkládána do astronomie a do metody lunárních vzdáleností. Další navržené metody určení zeměpisné délky byly založeny na vzájemné poloze magnetického pólu a Polárky — *Edmont Haley* (8. 11. 1656 – 14. 1. 1742), na poloze hvězd noční oblohy – *Samuel Fyler*, či zatmění Jupiterových satelitů – *Isaac Newton*.

Za nejnadějnější je považována metoda lunárních vzdáleností. Díky Newtonem formulovanému univerzálnímu zákonu gravitace byly pohyby měsíce lépe pochopeny a snáze se daly předvídat.

S vynaloženým úsilím se nakonec astronomům podařilo vytvořit jeden z pilířů metody lunárních vzdáleností: stanovili pozice hvězd a studovali pohyb Měsíce.

Pro zdokonalení metody už jen zbývalo vytvořit podrobné lunární tabulky, které by převedly zjištěné hodnoty do hodnoty zeměpisné délky. To se ale ukázalo být tím nejobtížnějším problémem. Složitost oběžné dráhy Měsíce mařila pokrok v určování vzdáleností mezi Měsícem a Sluncem, rovněž mezi Měsícem a hvězdami.

Úkolu se zhostil německý kartograf *Tobias Mayer* (17. 2. 1723 – 20. 2. 1762), kterému se budeme věnovat podrobněji. Vytvořil první soustavu lunárních tabulek pro umístění Měsíce ve dvanáctihodinových intervalech. Neocenitelnou pomoc mu poskytla spolupráce s *Leonhardem Eulerem* (15. 4. 1707 – 18. 9. 1783), který zjednodušil vzájemné pohyby Slunce, Země a Měsíce do soustavy rovnic. Mayer navrhl jako první metodu pro řešení nekonzistentní soustavy lineárních rovnic. Mayer se nikdy nezmyslil v úhlové vzdálenosti více než o 1,5 minuty. Mayer se ocenění nedožil, ale jeho žena obdržela odměnu ve výši 3000 liber. Dalších 300 liber obdržel Euler za své základní teorémy.

Těžkopádná metoda lunárních vzdáleností však vyžadovala příliš mnoho astronomických pozorování, konzultací s efemeridami a opravných výpočtů, což představovalo příliš mnoho kroků, během nichž mohlo dojít k pochybení.

První svazek Námořního almanachu a astronomických efemerid dokončuje *Nevil Maskelyne* (5. 10. 1732 – 9. 2. 1811) v roce 1766. Almanach samotný je vydáván dodnes.

Stane se ale to, co Newton považoval za nemožné. *John Harrison* komisi předkládá chronograf s požadovanou přesností. Vítězství nakonec nepatří hvězdám, ale času. A to přestože členové komise měnili pravidla soutěže, kdykoli to uznali za vhodné, aby tak upřednostnili astronomy před mechaniky.

Průběh celé soutěže je poznamenán několika skandály a dlouhé snažení vědců je zesměšňováno. Důvody nyní uvedeme.

Např. dne 10. 12. 1717 list *Englishman* a *Guardian* informují o velkolepém ohňostroji a otiskují informaci, že byl financován Radou pro zeměpisnou délku. Ta vyplatila zálohu 10 000 liber profesorům matematiky z Cambridge *Williamu Whistonovi* a *Humphrey Dittonovi*, kteří navrhli tzv. „Novou metodu určování zeměpisné délky na moři i na souši“. Metoda spočívala na výstřelech — viditelných na vzdálenost 100 mil — z děl na lodích strategicky zakotvených na signálních stanovištích. V průběhu ohňostroje bylo zjištěno, že výstřely z děla dosáhnou do výšky 6 440 stop. Dále se zjišťuje, že za 1000 liber příbuzný prvního ze jmenovaných vypracoval legislativní podporu signálních stanovišť. Po zjištění

uvedených skutečností jsou oba vyloučeni z profesorského sboru Cambridge.

Jako úspěšná metoda je vyhodnocen „Prášek souznění“ *Sira Kenelma Digbyho*. Vše spočívalo v nalodění zraněného psa na palubu lodi, na břehu musela zůstat o úspěšnosti metody přesvědčená osoba, která denně v poledne namáčela obvaz ze psí rány do roztoku prášku. Pes v reakci na tento úkon měl štěkat a tak poskytnout kapitánovi klíč k určování času. Psí nářek znamenal, že v Londýně je Slunce na nultém poledníku. Tak mohl kapitán porovnat čas na lodi s časem v domovském přístavu a tudíž stanovit i přesnou zeměpisnou délku. Noviny píší, že se Anglii směje nejen Francie ale celá Evropa.

Mapováním oblohy se zabývá Královská observatoř v čele s *Flamsteedem* (19. 8. 1646 – 31. 12. 1719). Ten ale výsledky čtyřicetiletého měření stále neuvádí. Newton a Halley tajně krádeží získávají Flamsteedovy záznamy a pirátsky je zveřejňují jako hvězdný katalog v r. 1717. Flamsteed shromažďuje 300 ze 400 výtisků a tyto pálí.

Před vydáním Maskelynova almanachu noviny sarkasticky uvádí, že problém zeměpisné délky je již vyřešen a jeho autor, známý malíř *William Hogarth* (1697 – 1764), ho znázornil na stěnu své cely č. 55 ústavu pro duševně choré Bedlam Asylum (dnes Bethlem Royal Hospital v Londýně).



Obrázek 1.: The Rake's Progress No. 8

Je také otištěna báseň, která komentuje průběh soutěže (překlad viz [2])

Hvězdný závod

Dva měsíce minuly, čas plyne
a deset mužů vrhlo se hrdinně
do zkoušky svého umění i sil,
by Flamsteedův se vrchol přiblížil . . .

Však opatrně, pane Maskelyne,
vy prohnáný své vědy harlekýne,
nemyslete, že zvítězíte klamem . . .

Vždyť velkým soudcem, jehož zatím
cena,
je spravedlivá příroda vznešená.

Odkaz na soutěž lze vidět i v Gulliverových cestách, kde se kapitán Lemuel Gulliver představí slovy, že je nemajetný, nesmrtelný a živořící stařec odkázaný na almužny, vidí klady takovéto možnosti v potěšení vidět návrat různých komet, vidět, jak mocné řeky vysychají a stávají se z nich mělké potůčky, a být svědkem *objevu zeměpisné šířky a délky, objevu nekonečného pohybu, univerzální medicíny a mnoha dalších velkých vynálezů, dovedených k nemyslitelné dokonalosti.*

Po smrti Newtona se stává předsedou Rady pro zeměpisnou délku Nevil Maskelyne, pátý královský astronom. Ten neváhá učinit cokoliv, aby prosadil svého chráněnce Tobiae Mayera a zabránil vítězství Johna Harrisona. Noviny otiskují jeho dopis příteli: „Kdyby nebylo toho zpropadeného mechanika, už dávno jsme si mohli s panem Mayerem rozdělit hlavní cenu.“

Nastává soudní spor Rady s vynálezci chronografu.

Zjišťuje se, že při testovacích plavbách byly hodiny úmyslně Maskelynem poškozeny (např. uplacení nosiči hodiny upustili z prudkého schodiště u budovy admirality, na loď byly přepraveny po neodpružené káře, na lodi byly umístěny na slunci, došlo k poškození hodin při natahování).

Nejvyšší odměnu v soutěži nakonec přece jen získává John Harrison. Jeho chronometry H-1, H-2, H-3 a H-4 byly schopny dosáhnout velké přesnosti měření času. Harrison uspěl přes všechny překážky s použitím čtvrtého rozměru – časového – ke spojení bodů na trojrozměrném glóbu. Vyrval tajemství orientace hvězdám a uzamkl je do hodin. Model H-4 dokázal určit zeměpisnou délku s přesností na deset mil — třikrát přesněji než požadovaly stanovy Ediktu. Po jedenaosmdesáti dnech na moři se zpozdil o pouhých pět sekund. Při dalších zkouškách ale model neuspěl. Někteří se ani netajili názorem, že Maskelyne poškodil hodiny hrubým zacházením při natahování nebo je uhranula jeho zlá vůle nebo že záměrně výsledky dalších zkoušek zkreslil. Vyplacení odměny bylo vázáno na splnění dalších různých podmínek. Harrison získal od Rady 8 750 liber teprve červnu 1773 po přimluvě krále Jiřího III., který osobně prováděl kontrolu funkce modelu H-5. Po deseti týdnech pozorování mohl konstatovat, že H-5 se ukázal být schopným měřit s odchylkou jedné třetiny sekundy za den.

Zajímavá je i historie chronometrů Harrisonova následovníka Kendalla. Jeho model K-1 měl s sebou na své třetí výpravě Cook. Podle pověsti ve stejném okamžiku, kdy byl kapitán Cook v roce 1779 na Havajských ostrovech zavražděn, se model K-1 zastavil. Model K-2 se dostal na palubu lodi Bountu, v okamžiku vzpoury pak model zůstává na ostrově Pitcairn.

Při testování dalšího Mudgeho modelu dokonce královský astronom Nevil Maskelyne náhodou zlomil hlavní pružinu. Mudgeův syn, právník Thomas Mudge, nakonec na Radě vymohl částku 3 000 liber.

V r. 1791 Východoindická společnost vydává pro své kapitány nový formulář palubního deníku, kde byly speciálně předtištěny stránky s kolonkou „zeměpisná délka podle chronometru“. V roce 1828 je Výnos o zeměpisné délce odvolán. Paradoxně se členové rozpuštěné Rady stávají členy komise pro testování a schvalování chronometrů pro lodě královského veličenstva.

Pravomoc k disponování s finančním fondem udělala z Rady pro zeměpisnou délku snad první oficiální agenturu pro výzkum a rozvoj. Radu pro zeměpisnou délku je možno považovat za první grantovou agenturu v historii (se všemi současnými nešvary). Ačkoliv to nikdo při jejím vzniku nemohl předvídat, Rada ve své základní podobě vydržela přes sto let. Do svého definitivního rozpuštění v roce 1828 rozdělila prostředky ve výši přes 100 000 liber.

Paradoxně všichni členové Rady pro zeměpisnou délku — zarytí odpůrci určování zeměpisné délky pomocí měření času — se stávají rozhodnutím parlamentu členy komise nové; s názvem Komise pro testování a vývoj chronografů pro lodě Jeho veličenstva.

3 První statistické metody

3.1 Mayerova metoda průměrů

Tobias Mayer (17. 2. 1723 – 20. 2. 1762) se při vývoji metody lunárních vzdáleností potýká s problémem řešení soustav rovnic, když má k dispozici větší počet rovnic než je počet neznámých.

V literatuře [3] je uvedeno 27 Mayerových rovnic sestavených z pozorování kráteru Manilius na měsíci.

Číslo rovnice	Tvar rovnice	Skupina
1	$\beta - 13^{\circ} 10' = 0,8836\alpha - 0,4682\alpha \sin \theta$	I
2	$\beta - 13^{\circ} 8' = 0,9996\alpha - 0,0282\alpha \sin \theta$	I
3	$\beta - 13^{\circ} 12' = 0,9899\alpha + 0,1421\alpha \sin \theta$	I
4	$\beta - 14^{\circ} 15' = 0,2221\alpha + 0,9750\alpha \sin \theta$	III
5	$\beta - 14^{\circ} 42' = 0,0006\alpha + 1,0000\alpha \sin \theta$	III

Číslo rovnice	Tvar rovnice	Skupina
6	$\beta - 13^{\circ}1' = 0,9308\alpha - 0,3654\alpha \sin \theta$	I
7	$\beta - 14^{\circ}31' = 0,0602\alpha + 0,9982\alpha \sin \theta$	III
8	$\beta - 14^{\circ}57' = -0,1570\alpha + 0,9876\alpha \sin \theta$	II
9	$\beta - 13^{\circ}5' = 0,9097\alpha - 0,4152\alpha \sin \theta$	I
10	$\beta - 13^{\circ}2' = 1,0000\alpha + 0,0055\alpha \sin \theta$	I
11	$\beta - 13^{\circ}12' = 0,9689\alpha + 0,2476\alpha \sin \theta$	I
12	$\beta - 13^{\circ}11' = 0,8878\alpha + 0,4602\alpha \sin \theta$	I
13	$\beta - 13^{\circ}34' = 0,7549\alpha + 0,6558\alpha \sin \theta$	III
14	$\beta - 13^{\circ}53' = 0,5755\alpha + 0,8178\alpha \sin \theta$	III
15	$\beta - 13^{\circ}58' = 0,3608\alpha + 0,9326\alpha \sin \theta$	III
16	$\beta - 14^{\circ}14' = 0,1302\alpha + 0,9915\alpha \sin \theta$	III
17	$\beta - 14^{\circ}56' = -0,1068\alpha + 0,9943\alpha \sin \theta$	III
18	$\beta - 14^{\circ}47' = -0,3363\alpha + 0,9418\alpha \sin \theta$	II
19	$\beta - 15^{\circ}56' = -0,8560\alpha + 0,5170\alpha \sin \theta$	II
20	$\beta - 13^{\circ}29' = 0,8002\alpha + 0,5997\alpha \sin \theta$	III
21	$\beta - 15^{\circ}55' = -0,9952\alpha - 0,0982\alpha \sin \theta$	II
22	$\beta - 15^{\circ}39' = -0,8409\alpha + 0,5412\alpha \sin \theta$	II
23	$\beta - 16^{\circ}9' = -0,9429\alpha + 0,3330\alpha \sin \theta$	II
24	$\beta - 16^{\circ}22' = -0,9768\alpha + 0,2141\alpha \sin \theta$	II
25	$\beta - 15^{\circ}38' = -0,6262\alpha - 0,7797\alpha \sin \theta$	II
26	$\beta - 14^{\circ}54' = -0,4091\alpha - 0,9125\alpha \sin \theta$	II
27	$\beta - 13^{\circ}7' = 0,9284\alpha - 0,3716\alpha \sin \theta$	I

Tabulka 1: Mayerovo měření

Pro řešení soustavy nejprve Mayer užil metodu selekce bodů. Vybral tři rovnice z 27 rovnic a to takovým způsobem, aby se hodnoty koeficientů lišily co nejvíce. To by mělo zajistit dobré výsledky neznámých. Volbou rovnic číslo 9, 16, 19 dostal následující soustavu rovnic:

$$\begin{aligned}\beta - 13^{\circ}5' &= 0,9097\alpha - 0,4152\alpha \sin \theta, \\ \beta - 14^{\circ}14' &= 0,1302\alpha + 0,9915\alpha \sin \theta, \\ \beta - 15^{\circ}56' &= -0,8560\alpha + 0,5170\alpha \sin \theta.\end{aligned}$$

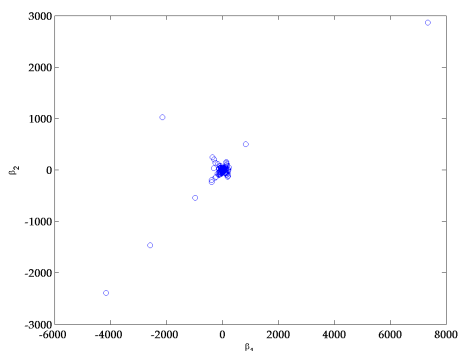
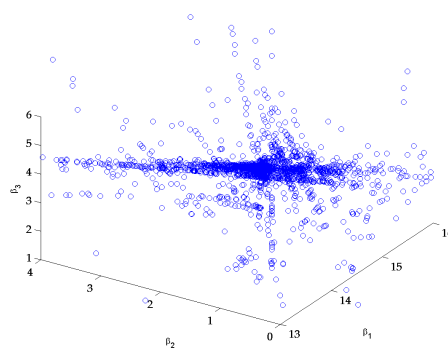
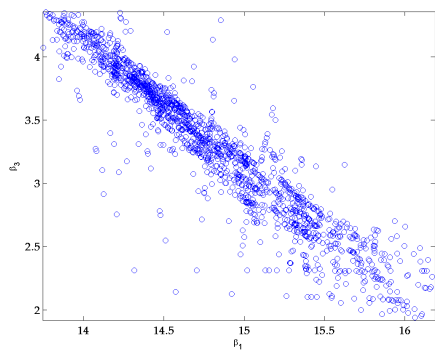
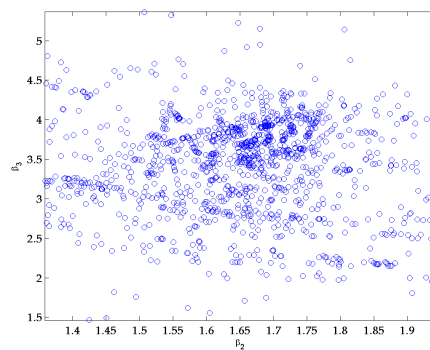
Tuto soustavu vyřešil postupnou eliminací neznámých a zjistil tyto výsledky: $\hat{\beta} = 14^{\circ}33'$, $\hat{\alpha} = 1^{\circ}40'$, $\hat{\theta} = -3^{\circ}43'$.

Mayer ovšem dospěl k názoru, že tato metoda je nevyhovující, protože výběr jiných tří rovnic vede k jiným odhadům. Proto by měla být zahrnuta všechna pozorování. Ideální by bylo použít všechny možné kombinace těchto trojic rovnic a zprůměrovat výsledky; nicméně k tomu by bylo zapotřebí vyřešit $\binom{27}{3}=2925$ systémů rovnic. Není tedy divu, že Mayer vzdává tento postup díky jeho přílišné pracnosti. Namísto toho navrhl podobný postup, spočívající v rozdělení 27 rovnic do tří skupin po devíti (rozdělení rovnic do skupin je uvedeno v tabulce 1). V každé ze skupin pak rovnice sečetl a vyřešil takto vzniklé tři rovnice (viz následující tabulka).

Skupina	Tvar rovnice
I	$\beta - 118^\circ 8' = 8,4987\alpha - 0,7932\alpha \sin \theta$
II	$\beta - 140^\circ 17' = -6,1404\alpha + 1,7443\alpha \sin \theta$
III	$\beta - 127^\circ 32' = 2,7977\alpha + 7,9649\alpha \sin \theta$

Tabulka 2: Rovnice vzniklé součty v jednotlivých skupinách

Soustavu řešil opět postupnou eliminací a dospěl k těmto výsledkům:
 $\hat{\beta} = 14^\circ 32'$, $\hat{\alpha} = 1^\circ 29'$, $\hat{\theta} = 3^\circ 49'$.

projekce všech řešení do roviny (β_1, β_2) výřez řešení v prostoru $(\beta_1, \beta_2, \beta_3)$ výřez řešení v rovině (β_1, β_3) výřez řešení v rovině (β_2, β_3)

Obrázek 2.: Řešení pro různé trojice rovnic.

Metoda průměru pro odhadování parametrů se stala spolu s dalšími algoritmy, které uvedeme v další části, velmi populární a používanou až do té doby, než byla nahrazena metodou nejmenších čtverců. Důvod, proč byla tato metoda tak úspěšná, je bezesporu její koncepční i numerická jednoduchost.

Pro určování zeměpisné polohy bylo třeba také určit tvar Země. Tímto problémem se zabývá Boškovič. Při výpočtu potřebuje najít aproximující přímku.

3.2 Boškovičova metoda nejmenších absolutních odchylek (1757, 1760)

Při zpracování dat z měření délky poledníku Roger Josef Boškovič (1711 – 1787) navrhuje jako první způsob výpočtu aproximující přímky.

Boškovič, od roku 1740 profesor na Collegiu Romanu v roce 1740, je v roce 1750 papežem Benediktem XIV. povolán spolu s anglickým jezuitou Christopherem Mairem, rektora anglické jezuitské koleje v Římě, aby změřili poledník a zkonstruovali novou mapu papežského státu. Jejich zpráva, psána v latině, vyšla v roce 1755.

Za svého života se Boškovič účastnil různých expedic zorganizovaných za účelem měření délky oblouku poledníku (například v Peru nebo na Mysu dobré naděje). Tato měření získaná na různých zeměpisných šířkách pak byla porovnána s podobnými měřeními blízko Francie. Cílem měření bylo potvrdit či vyvrátit předpoklad o tvaru Země jako rotačního elipsoidu.

Prozkoumejme nyní postup Boškoviče při hledání aproximující přímky, viz [1].

Pro malé oblouky je vztah mezi obloukovou délkou a šířkou dán přibližně vztahem $y = \alpha + \beta x$, kde y znamená délku jednostupňového oblouku poledníku a $x = \sin^2 L$, kde L je zeměpisná šířka středu oblouku. K nalezení elipticity, která může být aproximována podílem $\beta/3\alpha$, je nezbytné odhadnout parametry α a β . Boškovič seřadil pozorování $(x_1, y_1), \dots, (x_n, y_n)$ vzestupně podle velikosti x . Protože $x_1 = 0$, označil $\alpha = y_1$ a zpozoroval, že hodnota β může být nalezena z každé dvojice pozorování rovnice pro místo i a j tak, že se položí:

$$\beta_{ij} = \frac{y_j - y_i}{x_j - x_i}, \quad i = 1, \dots, n, j = 1, \dots, n, \quad j > i.$$

Boškovič má k dispozici 15 pozorování, z nich je ale 11 pořízeno ve Francii. Boškovič se obává, že měření by mohli být ovlivněny stejnou chybou. Proto z Francie ponechal pouze jedno měření a pro výpočet použije data tvořená pěti měření.

místo	zeměpisná šířka	$x = \sin^2 L$	y =délka oblouku
1 (a) Quito	0°0'	0,0000	56751
2 (b) Mys dobré naděje	33°18'	0,2987	57037
3 (c) Řím	42°59'	0,4648	56979
4 (d) Paříž	49°23'	0,5762	57074
5 (e) Laplace	66°19'	0,8386	57422

Tabulka 3: Boškovičova data (délky měřeny v toisech²)

K odhadu parametrů použil Boškovič nejprve stejně jako Mayer metodu selekce bodů. Výběrem prvního a posledního měření v místech (a) a (e), kde byl největší možný rozdíl měřených hodnot, dostal

$$\beta_{1,5} = \frac{y_5 - y_1}{x_5 - x_1} = \frac{57422 - 56751}{0,8386 - 0} = 800.$$

²1 toise \cong 1,949 m, jednotka platná ve Francii do roku 1799.

Dopracoval se tedy k přímce tvaru

$$\tilde{y} = 56751 + 800x$$

s rezidui v místech (b), (c) a (d) o velikosti $-46, 144, 138$. Protože ale shledal tato rezidua příliš velká, začal hledat jiné řešení a po vzoru Mayera se snažil brát v úvahu všechny možné kombinace rovnic, což obnášelo spočítat 10 hodnot β_{ij} pohybujících se v rozmezí mezi -350 a 1327 . Výsledky jsou uvedeny v tabulce:

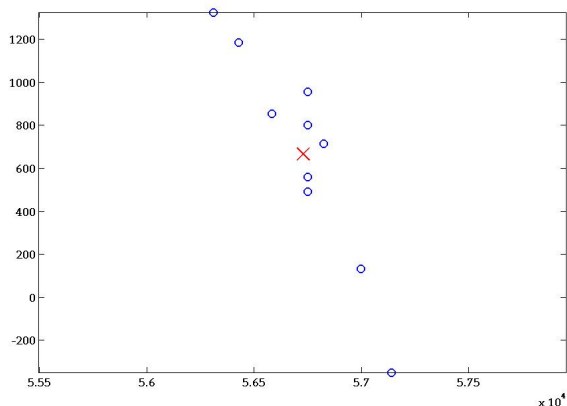
i	j	$\beta_{i,j}$	$\alpha_{i,j}$	tvar přímky	elipticita	rezidua pro ostatní místa
1	2	957	56751	$\tilde{y} = 56751 + 957x$	0,0056	217, 229, 132
1	3	491	56751	$\tilde{y} = 56751 + 491x$	0,0029	-139, -40, -260
1	4	561	56751	$\tilde{y} = 56751 + 561x$	0,0033	-119, 33, -201
1	5	800	56751	$\tilde{y} = 56751 + 800x$	0,0047	-46, 143, 138
2	3	-349	57141	$\tilde{y} = 57141 - 349x$	-0,0020	390, -133, -574
2	4	133	56997	$\tilde{y} = 56997 + 133x$	0,0008	246, 80, -313
2	5	713	56824	$\tilde{y} = 56824 + 713x$	0,0042	73, 176, 161
3	4	853	56583	$\tilde{y} = 56583 + 853x$	0,0050	-168, -200, -124
3	5	1185	56428	$\tilde{y} = 56428 + 1185x$	0,0070	-323, -255, 37
4	5	1326	56310	$\tilde{y} = 56310 + 1326x$	0,0079	-441, -331, -53

Tabulka 4: Výsledky pro všechny možné kombinace dvojic rovnic

Z uvedených výsledků spočítal průměry $\bar{\beta} = 667$, $\bar{\alpha} = 56729$ a získal průměrnou elipticitu rovnou 0,0039. Hledaná rovnice přímky potom byla ve tvaru

$$\tilde{y} = 56729 + 667x$$

a rezidua pro všechna místa o velikostech $-22, -109, 60, 39, -134$. Získané hodnoty ještě zakreslíme do grafu:



Obrázek 3.: Znázornění výsledků pro různé dvojice rovnic a průměrné hodnoty

Poznámka: Skutečná zploštělost Země je rovna $f = \frac{a-b}{a}$, kde $a = 6378245$ m je poloměr Země v rovině určené rovníkem a $b = 6356863$ m je vzdálenost od středu rovníku k pólu. Skutečná zploštělost Země je rovna $f = 0,003352861$. Jak je vidět z předchozího, Boškovičovy výsledky se tedy moc neliší.

Boškovič se k problémům sladování nekonzistentních rovnic stále vrací a v roce 1755 formuluje novou metodu řešení, známou jako *Boškovičova metoda nejmenších absolutních odchylek*.

Požadavky na tuto metodu formuloval následovně: Mějme určitý počet pozorování. K získání oprav, které musí být zhotoveny ke každému z nich, je nutno splnit tyto podmínky:

- součet kladných oprav by měl být roven součtu záporných oprav (bez ohledu na znaménka),
- součet všech oprav, ať už kladných či záporných, by měl být nejmenší možný.

Záměnou oprav za rezidua tyto dvě podmínky v podstatě znamenají, že

$$\sum_{i=1}^n (y_i - a - bx_i) = 0,$$

kde

$$\sum_{i=1}^n |(y_i - a - bx_i)|$$

by měla být minimalizována.

Jeho motivací pro první podmínku byla symetrie rozdělení chyb. Druhá podmínka je zapotřebí, aby bylo možno aproximovat pozorování tak přesně, jak jen to bude možné. Boškovič jako první formuloval kritérium pro aproximaci dat přímkou založené na minimalizaci funkce reziduí. Z první podmínky vyplývá, že

$$\bar{y} = a + b\bar{x},$$

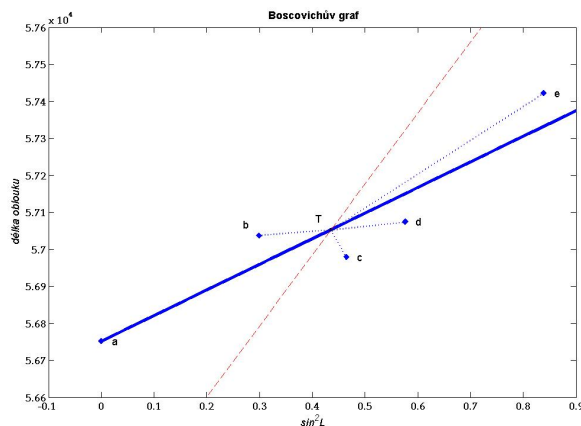
což znamená, že aproximační přímka prochází těžištěm pozorovaných bodů. Užitím toho výsledku lze eliminovat a z druhé podmínky a obdržíme

$$S(b) = \sum_{i=1}^n |y_i - \bar{y} - b(x_i - \bar{x})|,$$

což by mělo být minimalizováno vzhledem k b .

Boškovič se ale nezajímal o analytické vysvětlení a uvedl geometrickou formulaci a řešení problému.

Boškovičův pozoruhodný graf ukazuje současně data, přímky se směrnici $b_i = (y_i - \bar{y}) / (x_i - \bar{x})$ a demonstuje, jak nalézt aproximační přímku se směrnici \bar{b} , která bude minimalizovat $S(b)$.



Obrázek 4.: Boškovičova konstrukce aproximující přímky.

Aby zjistil, jak $S(b)$ závisí na b , proložil Boškovič přímku procházející těžištěm — v grafu znázorněno čárkovanou čarou, kterou pohyboval ve směru hodinových ručiček ze svislé polohy do vodorovné a zpátky do svislé. To způsobovalo změnu b „z plus nekonečna do minus nekonečna“. A současně se plynule $S(b)$ zmenšovalo „z nekonečna“ až na jeho minimum a opět vystoupalo „na nekonečno“. Pohybující se přímka prochází pozorováními v pořadí e, a, d, b, c . Boškovič podle toho přeuspořádal pozorování a měření vztáhl k těžišti jako

$$X_i = x_i - \bar{x} \quad \text{a} \quad Y_i = y_i - \bar{y}.$$

Získal tyto hodnoty:

i	X_i	Y_i	b_i	$\sum X_i $	$S(b_i)$
1(e)	0,40294	369,4	917	0,40294	416
2(a)	-0,43566	-301,6	692	0,83860	340
3(d)	0,14054	21,4	152	0,97914	627
4(b)	-0,13696	-15,6	114	1,11610	658
5(c)	0,02914	-73,6	-2526	1,14524	3527

Tabulka 5: Boškovičovy přeuspořádané hodnoty

Pak pro směrnice přímek (na obrázku tečkovaně) $b_i = Y_i/X_i$ platí vztah

$$S(b) = \sum_{i=1}^n |Y_i - bX_i| = \sum_{i=1}^n |X_i| \times |b_i - b|,$$

kde se index i nyní již vztahuje k novému uspořádání pozorování.

Dokážeme, že minima $S(b)$ je dosaženo pro $b = b_k$, kde k dostaneme z nerovnosti

$$\sum_{k+1}^n |X_i| - \sum_1^k |X_i| \leq 0 < \sum_k^n |X_i| - \sum_1^{k-1} |X_i|, \quad (1)$$

popřípadě ekvivalentně z

$$\sum_1^{k-1} |X_i| < \frac{1}{2} \sum_1^n |X_i| \leq \sum_1^k |X_i|, \quad (2)$$

což také Boškovič používal. Z čísel spočtených v tabulce 5 zjistil Boškovič po dosazení do (1) následující:

Pro $k = 2$

$$|X_1| = 0,40294 < 0,57262 < 0,83860 = |X_1| + |X_2|,$$

Pro $k = 3$

$$|X_1| + |X_2| = 0,83860 > 0,57262 < |X_1| + |X_2| + |X_3| = 0,97914.$$

Je vidět, že nerovnost (2) již není splněna pro $k = 3$. A tedy $\tilde{b} = b_2 = 692$.

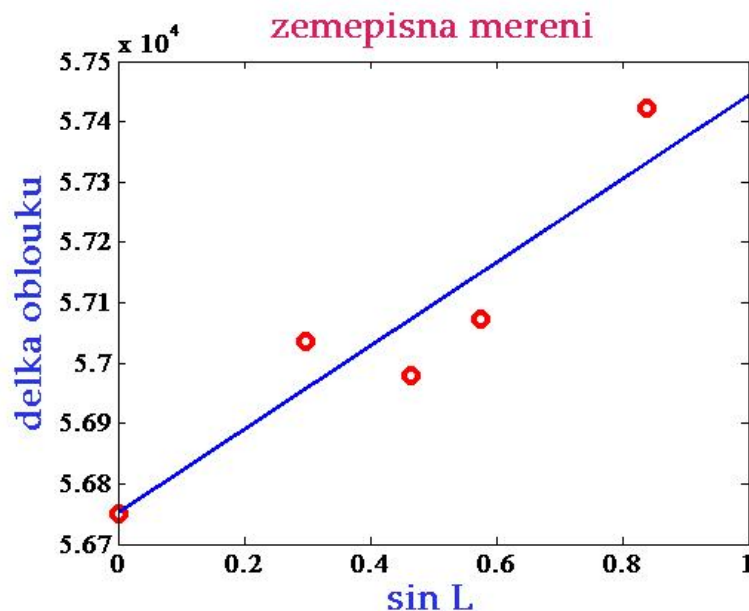
Rovnice pro nejlepší aproximující přímku je

$$\tilde{y} = \bar{y} + \tilde{b}(x - x_k), \quad \tilde{b} = b_k.$$

Dosazením konkrétních hodnot (viz tabulka 5) obdržíme

$$\tilde{y} = 56751 + 692x,$$

což je také přímka znázorněná na obrázku.



Obrázek 5.: Boškovičova přímka

Nyní ještě doplníme Boškovičův geometrický důkaz důkazem algebraickým. Ze vztahu

$$S(b_j) = \sum_{i=1}^{j-1} (b_i - b_j) |X_i| + \sum_{i=j+1}^n (b_j - b_i) |X_i|$$

plyne, že směrnice $S(b)$ mezi b_{j+1} a b_j je rovna

$$\begin{aligned} \frac{S(b_j) - S(b_{j+1})}{b_j - b_{j+1}} &= \sum_{i=j+1}^n |X_i| - \sum_{i=1}^j |X_i| \\ &= \sum_{i=1}^n |X_i| - 2 \sum_{i=1}^j |X_i|. \end{aligned}$$

Boškovičova analýza těchto dat je prvním úspěšným statistickým postupem a prvním závěrem získaným z nekonsistentních měření.

Další jeho myšlenkou bylo minimalizovat vážený součet absolutních hodnot odchylek měření od hledané hodnoty; neboť věděl, že se jednotlivá měření liší svou přesností. Měření ve všech zemích byly koncipovány jako dvouleté. Měření v Peru ale probíhalo 12 let, když se účastníci měření nemohli vzhledem k válce se Španělskem vrátit zpět do Francie. V Laponsku měření zahájili na jaře, avšak po oteplení se ocitli v bažinatém terénu a navíc byli napadeni obrovskými hejny komárů, a již po dvou měsících ze země uprchli.

3.3 Lambertova metoda (1765)

Johann Heinrich Lambert se narodil roku 1728 v Mülhausenu a zemřel roku 1777 v Berlíně. Působil jako matematik, fyzik a astronom.

Lambert při návrhu metody pro aproximaci dat předpokládal, že vztahy jsou lineární, popřípadě byly zlinearizovány transformací proměnných. Při zakreslování pozorování $(x_1, y_1), \dots, (x_n, y_n)$ do souřadného systému požadoval, stejně jako u vlastnosti aritmetického průměru, že aproximující přímka $y = a + bx$ by měla procházet těžištěm (\bar{x}, \bar{y}) , protože sumy odchylek na obou stranách takové přímky jsou si rovny pro každou hodnotu b . Tedy

$$\sum_{i=1}^n (y_i - \bar{y} - b(x_i - \bar{x})) = 0.$$

Aby odhadl směrnici přímky, rozdělil n pozorování na dvě přibližně stejně početné skupiny, kde v první skupině se nacházely menší hodnoty x a ve druhé větší hodnoty. Poté spočítal těžiště pro každou skupinu, (\bar{x}_1, \bar{y}_1) a (\bar{x}_2, \bar{y}_2) a vyjádřil b jako

$$b = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}.$$

Proto aproximující přímka prochází body (\bar{x}_1, \bar{y}_1) , (\bar{x}, \bar{y}) a (\bar{x}_2, \bar{y}_2) . Rovnost zapsal jako

$$\bar{y} = \bar{y}_1 - b\bar{x}_1 + bx.$$

Lambertova metoda je v knize [1] uvedena včetně zpracování jeho analýzy vztahů mezi barometrickými záznamy a nadmořskou výškou 12 francouzských hor.

Graf, který při zpracování našich dat získáme, je uveden na následujícím obrázku.

Nejprve spočítáme těžiště ze všech pozorování jako

$$(\bar{x}, \bar{y}) = \left(\frac{\sum_{i=1}^n x_i}{n}, \frac{\sum_{i=1}^n y_i}{n} \right) = (0,43925; 25660).$$

Nyní si pozorování rozdělíme na dvě skupiny. Řekněme, že v první skupině budou čtyři pozorování a ve druhé skupině tři. Stejným způsobem spočítáme těžiště pro každou ze skupin a dostaneme

$$(\bar{x}_1, \bar{y}_1) = (0,29161; 25611) \text{ a } (\bar{x}_2, \bar{y}_2) = (0,6361; 25725).$$

Dosazením do

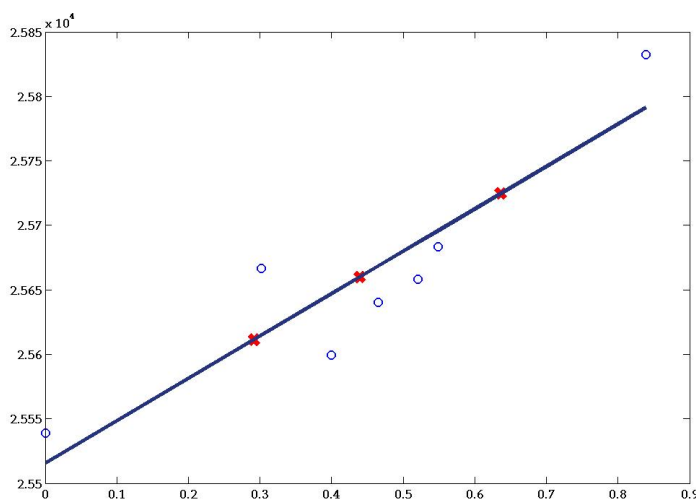
$$\tilde{y} = \bar{y}_1 - b\bar{x}_1 + bx,$$

kde

$$b = \frac{\bar{y}_2 - \bar{y}_1}{\bar{x}_2 - \bar{x}_1}$$

obdržíme

$$\tilde{y} = 25516 + 329x.$$



Obrázek 6.: Lambertova přímka

3.4 Laplaceova metoda minimalizující největší absolutní residuum (1786, 1793, 1799)

Pierre – Simon Laplace se narodil v roce 1749 a zemřel roku 1827.

Ve své analýze pohybu Saturnu a Jupitera navázal Laplace na Eulera a při řešení nekonzistentních lineárních rovnic i na Mayerovu metodu. Ovšem zatímco Mayer pouze sčítal rovnice uvnitř určitých skupin, Laplace vzájemně kombinoval tytéž rovnice různými způsoby a tím se dostal blíže k obecnému statistickému řešení problému. Své závěry publikoval v r. 1787 ve 127 stránkovém memoáru. Laplace navázal na Boškoviče — druhý díl jeho knihy obsahuje studii tvaru Země včetně diskuse dat získaných z několika expedic, na jejichž analýzu Laplace aplikoval svou teorii chyb pozorování. Pozdější vydání jeho další knihy též pokrývá aplikace pravděpodobnosti na určení hmotnosti Jupitera, Saturnu a Uranu a na problémy geodézie se speciálním zaměřením na měření francouzského poledníku.

O nalezení hodnot a a b v lineární rovnici

$$y_i = a + bx_i + e_i, \quad i = 1, \dots, n,$$

metodou minimalizace největších absolutních reziduí se zmiňují již mnozí autoři před Laplacem, ale nikdo nedokázal nalézt řešení. V jedné se svých knih Laplace formuluje problém a dává metodu řešení, kterou užil pro odhad a a b ve vztahu mezi délkou oblouku a $\sin^2 L$, prostřednictvím čtyř pozorování z Quita, Mysu dobré naděje, Paříže a Laponsko. Chyba modelu je dána součtem chyb měření a odchylkou od eliptického tvaru. Jestliže minimum největších absolutních reziduí je podstatně větší než chyba měření, může být zamítnuta hypotéza o eliptickém tvaru Země. Laplace navíc připustil, že tato procedura je použitelná ve všech případech, kdy je potřeba porovnat, zda výsledky hypotéz jsou v souladu s chybami měření. Z toho důvodu považoval Laplace tuto metodu zejména jako prostředek k testování hypotéz. Z naměřených dat zjistil, že

$$\tilde{y} = 56753 + 684x$$

s reziduí rovnými $-75, -75, 2, 75$. Usoudil, že hypotéza o eliptickém tvaru Země neodpovídá měřením, protože chyba měření o velikosti 75 toisů není příliš uspokojivá. Nicméně neuvedl, jakou chybu měření by považoval za přijatelnou. V monografii publikované v roce 1793 se k problému vrací a představuje podobnou metodu měření, kterou ilustruje provedením devíti měření, nalezením rovnice tvaru

$$\tilde{y} = 56723 + 685x,$$

což je velmi podobné předcházejícím výsledkům, ale s největšími reziduí, v tomto případě rovnými ± 108 . Jeho závěr byl shodný se závěrem z roku 1786. Ve své další knize opakuje teorii, kterou použil v monografii z roku 1793, kde zkoumá sedm pozorování: šest použil z předešlých prací a dvě měření konaná ve Francii nahradil jediným precizním. Měření z Holandska vyřadil jako nespolehlivá. Eliptičnost vyšla $1/248$, což je podobné výsledku, ke kterému došel také Boškovič. Data jsou dána v následující tabulce:

místo	zeměpisná šířka	$x = \sin^2 L$	počet stupňů	y = délka (na jeden stupeň)
1 Peru	0,0000	0,00000	3,4633	25538,85
2 Mys dobré naděje	37,0093	0,30156	1,3527	25666,65
3 Pensylvánie	43,5556	0,39946	1,6435	25599,60
4 Itálie	47,7963	0,46541	2,4034	25640,55
5 Francie	51,3327	0,52093	10,7487	25658,28
6 Rakousko	53,0926	0,54850	3,2734	25683,30
7 Laponsko	73,7037	0,83887	1,0644	25832,25

Tabulka 6: Laplaceova data

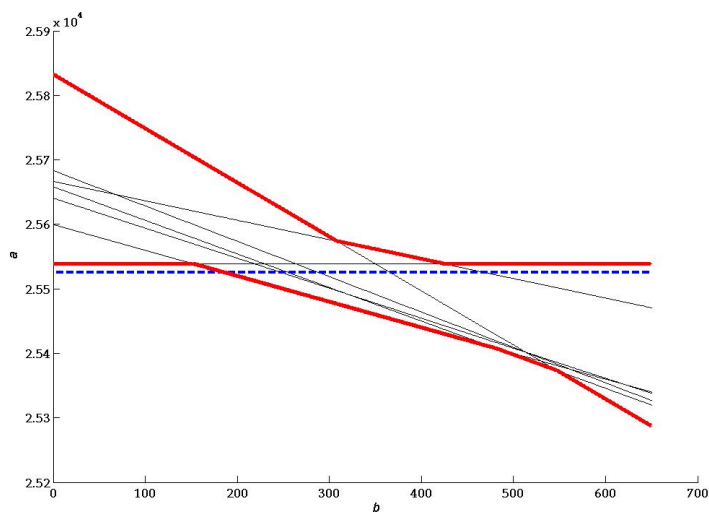
Problém spočívá v nalezení hodnot a a b minimalizováním

$$\max |e_i| = \max |y_i - a - bx_i|, \quad i = 1, \dots, n.$$

Předpokládáme, že pozorování byla seřazena následovně: $x_1 < x_2 < \dots < x_n$. Laplaceova algebraická teorie řešení je mnohem jednodušší na pochopení, pokud se vše znázorní do grafu. Proto je v obrázku zakresleno sedm přímk odpovídajících Laplaceovým datům tvaru:

$$e_i + a = y_i - bx_i, \quad i = 1, \dots, 7,$$

jako funkce b .



Obrázek 7.: Laplaceova metoda

Z grafu je vidět, že $\max(e_i + a)$ a $\min(e_i + a)$ jsou po částech lineární funkce b určené průsečíky přímk. Maximum bylo nalezeno pro $i = 7, 2, 1$ s průsečíky v bodech $b_1 = 308$, $b_2 = 424$ a minimum pro $i = 1, 3, 5, 7$ a $b_3 = 152$, $b_4 = 483$,

$b_5 = 547$. Nejmenší rozdíl, $e_7 - e_3 = 97$ je dosažen pro hodnotu $b = 308$. Výběrem a ležícím mezi dvěma extrémy $e_7 + a$ a $e_3 + a$ dostaneme $a = 25525$, největší reziduum je minimalizováno a $e_7 = -e_3 = 49$. Řešení a jeho vlastnosti jsou jednoduše dosažitelné z grafu.

Laplace začal jeho algebraické řešení důkazem následujícího lemmatu:

Lemma 3.1 *Jestliže a a b minimalizují $\max |e_i|$, potom existují rezidua e_h, e_k, e_m taková, že $|e_h| = |e_k| = |e_m| = \max |e_i|$, přičemž jedno z nich se liší znaménkem od ostatních. Předpokládáme-li například, že se jedná o e_k , potom číslo k leží mezi h a m .*

Důkaz: Důkaz je nepřímý. Předpokládejme, že a a b byly vybrány tak, aby se nejvzdálenější rezidua, řekněme, že se jedná o e_k a e_m , rovnala až na znaménko. Tedy $-e_k = e_m > 0$. Zvětší-li se a a b , všechna kladná rezidua se sníží, absolutní hodnota záporných reziduí se zvýší, tedy maximální reziduum se také zvýší. Proto odchylky mezi a a b vytvářející rovnovážný stav mezi nejvzdálenějšími rezidui povedou k většímu nejvzdálenějšímu reziduu.

Aby vypočetl průsečíky přímek, odečetl Laplace první podmínkovou rovnici od ostatních, což vedlo na rovnice

$$y_i - y_1 - b(x_i - x_1) = 0, \quad i = 2, \dots, n,$$

s řešením

$$b_{i1} = \frac{y_i - y_1}{x_i - x_1}.$$

Řekněme, že největší hodnota b_{i1} je b_{r1} ; pokud je na výběr z více možností, největší hodnota i je vybírána pro r_1 . Uvažujme, že $e_i + a = y_i - bx_i$. Je jednoduché dokázat, že

$$e_1 > e_i \text{ pro } b > b_{r1}, \text{ kde } i = 2, \dots, n,$$

a

$$e_1 < e_{r1} \text{ pro } b < b_{r1}, \text{ kde } i = 1, \dots, r_1 - 1.$$

□

Podobné vztahy s nerovnostmi změnilly postavení b ve vztahu k nejmenší hodnotě b_{i1} , kterou budeme označovat b_{s1} .

Zjištěním, že $\max(e_i + a) = e_1 + a$ pro $b_{r1} < b < \infty$ a $\min(e_i + a) = e_1 + a$ pro $-\infty < b < b_{s1}$, Laplace pokračoval stejnou procedurou od r_1 a s_1 počínaje. K nalezení $\max(e_i + a)$ pro $b < b_{r1}$ spočítal:

$$b_{i2} = \frac{y_{r1+i} - y_{r1}}{x_{r1+i} - x_{r1}}, \quad i = 1, \dots, n - r_1.$$

Největší hodnotu b_{i2} označíme jako b_{r2} . Pak vyplývá, že e_{r1} bude největší kladné reziduum pro $b_{r2} < b < b_{r1}$. Záměnou s_1 za r_1 a označením nejmenší hodnoty b_{i2} jako b_{s2} nám vyplyne, že absolutní hodnota e_{s1} bude největší absolutní hodnotou záporných reziduí pro $b_{s1} < b < b_{s2}$, atd.

Výsledky, které jsme dostali užitím algoritmu, jsou uvedeny v tabulce 7. Z této tabulky dostaneme dvě skupiny nejbližších reziduí, jak je ukázáno v tabulce 8. V každém průsečíku jsou si dvě rezidua rovna. S odkazem na lemma 3.1 Laplace usoudil, že b dostaneme právě z jedné z dvou skupin

$$b_{r_1}, b_{r_2}, \dots \text{ nebo } b_{s_1}, b_{s_2}, \dots$$

i	b_{i1}	$r_1 = 2$	$s_1 = 3$	$s_2 = 5$
		b_{i2}	b_{i2}	b_{i3}
2	<u>424</u>			
3	<u>152</u>	-685		
4	219	-159	621	
5	229	-38	<u>483</u>	
6	263	67	562	908
7	350	<u>308</u>	529	<u>547</u>

Tabulka 7: Laplaceův algoritmus určování průsečíků

b	e_r	b	e_s
$-\infty - 308$	e_7	$-\infty - 152$	e_1
$308 - 424$	e_2	$152 - 483$	e_3
$424 - \infty$	e_1	$483 - 547$	e_5
		$547 - \infty$	e_7

Tabulka 8: Vztahy mezi b a maximálními rezidui

Nakonec se hodnota a spočítá vyřešením rovnosti $e_m = -e_k$, což dává řešení

$$a = \frac{1}{2}(y_k + y_m) - \frac{1}{2}b(x_k + x_m),$$

tedy

$$-e_k = e_h = e_m = \frac{1}{2}(y_m - y_k) - \frac{1}{2}b(x_m - x_k).$$

Z kladných reziduí v tabulce 8 plyne, že $b = 424$ musíme vyřadit, protože neexistuje hodnota i mezi 1 a 2. Pro $b = 308$ dostaneme $e_7 = e_2$. Podíváme-li se do skupiny záporných reziduí, zjistíme, že 308 leží v intervalu (152, 483) pro hodnotu e_3 . Proto pro $b = 308$ máme $e_7 = e_2 = -e_3$, $a = 25525$ a absolutní hodnota tří nejbližších reziduí je rovna 49.

Přesnějšími výpočty Laplace zjistil $a = 25525$, $b = 308,202$ a $\max |e_i| = 48,60$. Převodem těchto výsledků do tisíců, dostaneme

$$\tilde{y} = 56722 + 685x,$$

pro kterou jsou největší rezidua rovna ± 108 , což je stejný výsledek, jako dosažený z devíti pozorování v roce 1793.

Laplaceův závěr byl proto stejný: „Z předcházejících měření patrně plyne, že délky stupňů zemského poledníku se liší.“

3.5 Laplaceova modifikace Boškovičovy metody (1799)

Laplace bezprostředně pokračoval v diskuzích týkajících se minimalizace největších absolutních reziduí a začal počítat odhady a a b jinou metodou. Takovou, která pouze optimalizuje absolutní odchylky. Inspirován vynalézavou Boškovičovou metodou uvedl Laplace podmínky tak, abych dostal analytický tvar řešení a uspokojivá rezidua.

Funkci, která má být minimalizována, Laplace přepsal do tvaru

$$S(b) = \sum_{i=1}^n |X_i| \times |b_i - b|,$$

kde $b_1 > b_2 > \dots > b_n$. Za předpokladu, že $b_k > b > b_{k+1}$, dostal

$$S(b) = \sum_{i=1}^k |X_i|(b_i - b) - \sum_{i=k+1}^n |X_i|(b_i - b),$$

tedy

$$S'(b) = -\sum_{i=1}^k |X_i| + \sum_{i=k+1}^n |X_i|.$$

Minima $S(b)$ je proto dosaženo pro $b = b_k$. Řekněme, že index k se získá řešením nerovnice

$$\sum_{i=1}^{k-1} |X_i| < \frac{1}{2} \sum_{i=1}^n |X_i| \leq \sum_{i=1}^k |X_i|,$$

což je Boškovičův výsledek.

Dále navrhl užít Boškovičových dvou podmínek. k měření celých oblouků namísto jejich délek ve stupních. Což znamená, že místo y_i uvažuje $w_i y_i$, kde w_i je počet stupňů oblouků (viz tabulka 7). Počet stupňů se pohyboval v rozmezí od 1 do 11. Motivací pro tuto změnu bylo, že pokud v tomto duchu uvážíme celé oblouky namísto stupňů, které z nich počítáme, měli bychom dostat ke všem těmto stupňům mnohem větší vliv při počítání elipticity Země.

Z obecnějšího úhlu pohledu, o kterém se Laplace nezmiňuje, jeho modifikace Boškovičovy metody spočívá v zavedení vážených reziduí. Stejně jako třeba Daniel Bernouilli nebo Euler užili vážených průměrů místo aritmetických. Nicméně Laplace vůbec neuvažoval, jak zvolit váhy. Pro něj existovaly v podstatě jen dvě možnosti. Buď použít délku ve stupních, jako je tomu v memoáru z roku 1793, nebo úplné délky, jako tomu bylo v roce 1799.

Laplace nyní zopakoval důkaz Boškovičových výsledků z roku 1793 touto modifikací. Obdržel rovnice

$$w_i e_i = w_i y_i - a w_i - b w_i x_i, \quad i = 1, \dots, n.$$

Pořád diskutované dvě podmínky nyní vypadají tak, že $\sum w_i e_i = 0$ a $\sum w_i |e_i|$ by měla být minimalizována. Dále pak $w_i > 0$. Z první podmínky vyplývá

$$a = \bar{y}_w - b \bar{x}_w, \quad \bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}, \quad \bar{y}_w = \frac{\sum w_i y_i}{\sum w_i},$$

což znamená, že aproximující přímka prochází váženým těžištěm.

Odstraněním a dostaneme

$$e_i = y_i - \bar{y}_w - b(x_i - \bar{x}_w) = Y_i - bX_i,$$

kde Y_i a X_i znamenají odchylky od váženého průměru. Označíme-li $b_i = Y_i/X_i$, musíme minimalizovat výraz

$$S(b) = \sum w_i |e_i| = \sum w_i |X_i| \times |b_i - b|.$$

Tímto jsme se dostali k řešení stejného problému jako výše. Pouze s tím rozdílem, že $|X_i|$ je nahrazen $w_i |X_i|$. Přechíslováním hodnot b_i tak, aby $b_1 > b_2 > \dots > b_n$, dostaneme $b = b_k$, kde k je opět spočítáno z nerovnice

$$\sum_{i=1}^{k-1} w_i |X_i| < \frac{1}{2} \sum_{i=1}^n w_i |X_i| \leq \sum_{i=1}^k w_i |X_i|.$$

Aplikací tohoto algoritmu na data z tabulky 7 Laplace zjistil, že pozorování by se měla brát v pořadí 3, 7, 6, 1, 5, 2, 4 a že platí (již vzhledem k novému pořadí)

$$\sum_1^3 w_i |X_i| = 0,85398 < 1,76013 < 2,36803 = \sum_1^4 w_i |X_i|.$$

Tedy $b = b_4 = 246,93$ a $a = 25538,85$. Dostal také elipticitu odpovídající aproximující přímce jako $1/312$. Což je menší hodnota oproti předchozím nalezeným a podobná hodnotě získané při fyzikálních pokusech s kyvadlem. Výsledek, ke kterému dospěl, zakreslíme společně s Boškovičovou aproximační přímkou do jednoho obrázku, abychom viděli, jak se obě metody liší.

4 Porovnání odhadů jednotlivých algoritmů z konce 18. století s MNČ

V této části se budeme zabývat opět jednotlivými metodami, přičemž na konkrétních datech metody porovnáme a výsledky zakreslíme do grafu včetně pásu spolehlivosti pro MNČ. Pro tyto účely vybereme data naměřená Laplacem (viz tabulka 7).

1. Boškovičovou metodou dospějeme k těmto výsledkům:

i	X_i	Y_i	b_i	$\sum X_i $	$S(b_i)$
1	-0,43925	-121,1	276	0,43925	214
2	-0,13769	6,7	-49	0,57694	446
3	-0,03979	-60,3	1516	0,61673	1521
4	0,02616	-19,4	-741	0,64289	1263
5	0,08168	-1,6	-20	0,72457	420
6	0,10925	23,4	214	0,83382	246
7	0,39962	172,3	431	1,23344	269

Tabulka 9: Data získána Boškovičovým algoritmem

Pozorování přeuspořádáme sestupně dle hodnoty b_i

i	X_i	Y_i	b_i	$\sum X_i $	$S(b_i)$
3	-0,03979	-60,3	1516	0,03979	1521
7	0,39962	172,3	431	0,43941	269
1	-0,43925	-121,1	276	0,87866	214
6	0,10925	23,4	214	0,98791	246
5	0,08168	-1,6	-20	1,06959	420
2	-0,13769	6,7	-49	1,20728	446
4	0,02616	-19,4	-741	1,23344	1263

Tabulka 10: Přeuspořádaná data

Nyní již budeme hledat index k splňující nerovnost (2) ze str. 107.

Tedy pro $k = 2$ platí

$$|X_1| = 0,03979 < 0,61672 > 0,43941 = |X_1| + |X_2|.$$

Je vidět, že nerovnost není splněna pro $k = 2$.

Pro $k = 3$ platí

$$|X_1| + |X_2| = 0,43941 < 0,57262 < |X_1| + |X_2| + |X_3| = 0,87866.$$

Pro $k = 4$ platí

$$|X_1| + |X_2| + |X_3| = 0,87866 > 0,57262 < |X_1| + |X_2| + |X_3| + |X_4| = 0,98791.$$

Z výše uvedeného plyne, že nerovnost je splněna pouze pro $k = 3$. Jak již bylo uvedeno v první kapitole, rovnice pro nejlepší aproximující přímku je

$$\tilde{y} = \bar{y} + \tilde{b}(x - x_k), \quad \tilde{b} = b_k.$$

Dosazením konkrétních hodnot obdržíme

$$\tilde{y} = 25539 + 276x.$$

2. Laplaceovy výsledky máme k dispozici z první kapitoly. Jen připomeneme, že Laplace došel k aproximující přímce tvaru

$$\tilde{y} = 25525 + 308x.$$

3. Mějme obecně daný model přímky

$$Y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, \dots, n,$$

kde e_i je chyba měření β_0, β_1 jsou neznámé parametry. Na základě změřených dat odhadneme tyto parametry a označíme je b_0, b_1 . Tím získáme regresní funkci $\tilde{y} = b_0 + b_1 x$. Parametry odhadneme pomocí metody nejmenších čtverců. Tedy

$$g(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \longrightarrow \min$$

$$\frac{\partial g}{\partial b_0} = \sum_{i=1}^n 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial g}{\partial b_1} = \sum_{i=1}^n 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

$$\begin{aligned} \sum_{i=1}^n y_i - b_0 n - b_1 \sum_{i=1}^n x_i &= 0 \\ \sum_{i=1}^n y_i x_i - b_0 \sum_{i=1}^n x_i - b_1 \sum_{i=1}^n x_i^2 &= 0 \end{aligned}$$

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

Po vyřešení soustavy rovnic vyjádříme odhady b_0, b_1 jako

$$b_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}, \quad b_1 = \frac{n \sum x_i y_i - \sum y_i \sum x_i}{n \sum x_i^2 - (\sum x_i)^2}.$$

Metodu aplikujeme na naše data. Užitím výše uvedených vzorců dostaneme aproximační přímku tvaru

$$\tilde{y} = 25520 + 320x.$$

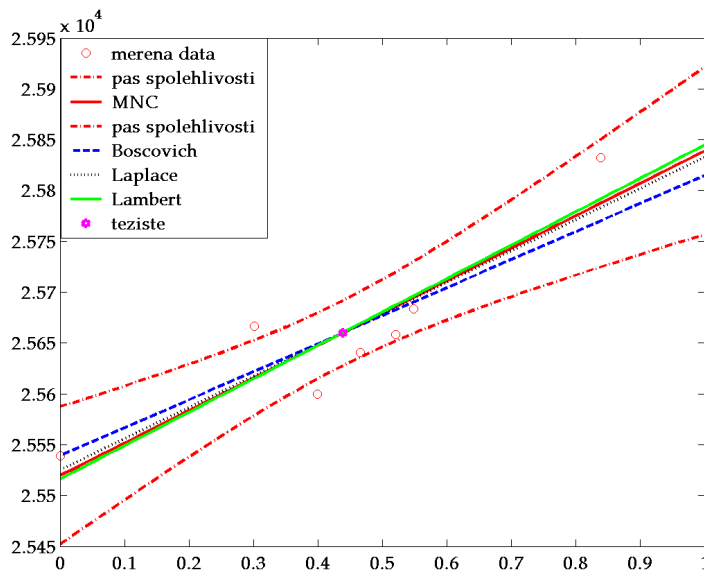
Pro větší názornost zakreslíme všechny aproximační přímky do jednoho obrázku a to včetně pásu spolehlivosti pro MNČ s koeficientem spolehlivosti $1 - \alpha$, který zkonstruujeme podle následujícího vzorce

$$b_0 + b_1 x \pm t_{n-2}(\alpha) s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}},$$

kde

$$s = \sqrt{\frac{\sum Y_i^2 - b_0 \sum Y_i - b_1 \sum x_i Y_i}{n - 2}}$$

V našem případě zvolíme $\alpha = 0,05$ a příslušný kvantil $t_5(0,95) = 2,015$ najdeme v tabulkách.



Obrázek 8.: Porovnání metod

Na závěr ještě uvedeme tabulku reziduí pro jednotlivá místa za použití konkrétních metod.

	Boškovič	Lambert	Laplace	MNČ
Peru	0,0000	-22,8500	-13,8500	-18,8500
Mys dobré naděje	-44,4194	-51,4368	-48,7695	-50,1508
Pensylvánie	49,6510	47,8223	48,4337	48,2272
Itálie	26,9032	28,5699	27,7963	28,3812
Francie	24,4967	29,1060	27,1664	28,4176
Rakousko	7,0860	13,1565	10,6380	12,2200
Laponsko	-61,7219	-40,2618	-48,8780	-43,8116

Tabulka 11: Rezidua pro jednotlivá místa

V tabulce reziduí lze zpozorovat, že Bošovičova metoda provádí největší korekci u měření v Laponsku, o kterém je známo, že bylo nejméně přesné (viz str. 108).

References

- [1] Hald, Anders: *A History of Mathematical Statistics (from 1750 to 1930)*, John Wiley & sons, New York, 1998
- [2] Sobelová, Dava: *Osamělý génius*, Alpress, Český Těšín, 1997
- [3] Stigler, Stephen M.: *The history statistics – The measurement of uncertainty before 1900*, The Belknap press of Harvard university, Cambridge, Massachusetts and London, 1986



Rychlost proudění v řece

PAVEL TUČEK¹ A JAROSLAV MAREK²

¹ *Katedra geoinformatiky
Přírodovědecká fakulta Univerzity Palackého
Tř. Svobody 16, 779 00 Olomouc, Czech Republic
e-mail: tucekp@inf.upol.cz*

² *Katedra matematické analýzy a aplikací matematiky
Přírodovědecká fakulta Univerzity Palackého
Tomkova 40, 779 00 Olomouc
e-mail: marek@inf.upol.cz*

Abstrakt: Cílem článku je statisticky popsat vodní tok. K nejdůležitějším úkolům hydrologie se řadí určení rychlostí proudění v říčním korytu. K určení rychlosti slouží většinou Pitotova trubice. Měření rychlosti se provádí na několika místech říčního koryta, u kterých budeme mít k dispozici i polohové souřadnice. Cílem měření je najít odhad říční čáry, kde je nejvyšší rychlost a parametry, které popisují pokles rychlosti v závislosti na vzdálenosti od této říční čáry ve směru osy x , y a z . Tento model může být studován pomocí teorie lineárních modelů.

AMS Classification: 62J05; 62F10

Klíčová slova: lineární regresní model s podmínkou typu II, hydrologie, proudění

1 Úvod

Hydrometrie představuje nejstarší vědu o vodě. Měření vodních stavů a později i průtoků se prováděla ještě před tím, než hydraulika a hydrologie určily sféru svého zkoumání. Předbílání rozvoje hydrometrie bylo přirozené v období začátků věd, které se obvykle spojují s nahromaděním údajů o přírodních procesech a jevech. Osvojení si metod měření na vodních útvarech je nejen nezbytným předpokladem činnosti v operativní hydrologii, ale je potřebné i při plnění úkolů hydrologického výzkumu a v neposlední době také pro ochranu a tvorbu životního prostředí, jak dokazují například nedokonalé předpovědi průtoků řeky Moravy při povodních v roce 1997 nebo 2005. Při hydrometrickém výzkumu se setkáváme s problémy určování délky a šířky vodního koryta, což představuje polohopisné a výškové měření, jehož výsledkem je stanovení příčného a

podélného profilu vodního útvaru. Nejpodstatnějším úkolem ale bezpochyby je stanovení rychlosti proudění vody v příslušném korytu.

Podstatou tohoto příspěvku je nalezení metody pro stanovení průtoku vody v říčním korytu na základě měření rychlostí průtoku v říčním korytu. Taková měření, jak bude ukázáno, jsou velice nákladná a tudíž se snažíme za použití co nejmenšího počtu měření dosáhnout uspokojivého výsledku, který by měl poté sloužit dalším subjektům k možnému rozhodování, čímž například myslíme vyhlášení stavů ohrožení atd.

K využití modelování průtoku lze samozřejmě, za využití fyzikálních vlastností proudění kapalin, použít různé metody modelování. Přístup, ke kterému jsme se rozhodli v našem případě, vychází ze statistického pohledu na danou problematiku. K dispozici máme údaje získané z hydrometrického měření průtoku vody v korytu a naším cílem je nalezení odhadu rychlosti proudění v bodech, ve kterých měření nebylo provedeno.

Základní charakteristikou pohybu vody v povrchových vodních útvarech je rychlost a směr proudění vody. Pohyb vody je buď laminární (jednotlivé molekuly vody se pohybují rovnoběžně stejným směrem), nebo turbulentní (jednotlivé částice se pohybují různými směry různou rychlostí) V povrchových vodních útvarech patří právě modelování vlastností průtoku vody říčním korytem k nejsložitější části hydrometrického výzkumu. Vzhledem ke stálým změnám rychlosti vody v jednotlivých bodech průtočného profilu se v hydrometrické praxi používá průměrná hodnota rychlosti v bodě, v příčném průřezu, po dráze apod. K měření rychlosti proudění vody se nejčastěji používá hydrometrické vrtule, která je výhodná zejména v otevřených oblastech, nebo Pitotovy trubice, jejíž použití je limitováno na laboratorní podmínky.



Obr. 1: Hydrometrická vrtule

Při měření rychlosti proudění vody hydrometrickou vrtulí se využívá závislosti mezi počtem otáček vrtule – rotoru a rychlostí proudění vody. Počet otáček je přímo úměrný rychlosti tekoucí vody. Tato závislost se určuje pro každou vrtuli tzv. tárováním a znázorňuje se graficky nebo se vyjadřuje analytickým výrazem. V praxi se jedná o křivku, která je nahrazena jednou nebo několika přímkami.

Konkrétní měření v terénu se provádí buď pomocí hydrometrické vrtule připevněné k tyči, nebo upevněné pomocí závěsu umístěných na měrných lanovkách, nebo člunech.

Pro stanovení konkrétního průtoku, respektive rychlosti proudění vody v říčním korytu se používá průměrné hodnoty naměřených rychlostí v měrné svislici, což ale značně zneprůstňuje konečný výsledek. Ve všech doposud známých metodách stanovení výsledné rychlosti průtoku v říčním korytu je používáno pouze získaných naměřených hodnot, potažmo přímého výpočtu z těchto naměřených hodnot. Při značné finanční náročnosti měření se tak často setkáváme s odhadem stanoveným na základě velmi malého počtu měření v malém počtu svislic. V případě letošních povodní to pak mělo za následek nepřesné stanovení hodnot množství protékající vody říčním korytem. Protože, jak je obecně známo, množství protékající vody je závislé na rychlosti, s jakou voda korytem teče. Vyhodnocením průtoku je veličina

$$Q = \int_0^B \int_0^H v_i dH dB,$$

kde B a H jsou hloubkou a šířkou daného říčního koryta. To se vzhledem k neznámému rozložení rychlostí proudění v průtočné ploše a ani velikosti této plochy, řeší po částech a dílčí hodnoty se sčítají.

Jak je vidět, postupy stanovení a měření rychlostí proudění jsou známé a korektní. Nelze však zpochybnit fakt, že nedostatečný počet uskutečněných měření zkresluje reálnou představu o chování říčního toku, což ve finální fázi vede k chybným úsudkům o chování toku popřípadě a možnosti ohrožení v důsledku velkého průtoku vody říčním korytem. Položme si tedy otázku, zda v tomto postupu stanovení popisných charakteristik říčního toku není podceněn přechod od naměřených hodnot k výsledným charakteristikám. Proč nemodelovat průtok vody říčním korytem jako celek a nebrat pouze střední hodnoty jednotlivých charakteristik? Ty se dají přece z výsledného celkového pohledu také vypočítat.

Mějme tedy k dispozici naměřené hodnoty v říčním korytu a z nich zkonstruujeme celkový popis a rozložení rychlostí proudění vody v říčním korytu. Výsledkem budou izotachy v příčném a podélném profilu říčního koryta a popisné charakteristiky získané jednak z namodelovaných hodnot tak z naměřených hodnot. Teprve tyto informace, nechť slouží odborníkům pro další rozhodování.

2 Model

Model nepřímého měření s podmínkou II. typu na parametry 1. řádu má tvar

$$\mathbf{Y} \sim (\mathbf{F}\boldsymbol{\Theta}, \boldsymbol{\Sigma}), \quad \mathbf{b} + \mathbf{C}\boldsymbol{\Theta} + \mathbf{B}\boldsymbol{\beta} = \mathbf{0}, \quad (1)$$

kde $\mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_n)'$ je náhodný observační vektor; $\boldsymbol{\theta} \in R^{k_1}$ je vektor měřených parametrů; $\boldsymbol{\beta} \in R^{k_2}$ je vektor dodatečných neměřených parametrů.

Jestliže $h(\mathbf{F}_{n,k}) = k_1 < n$, $h(\mathbf{C}_{(q,k_1)}, \mathbf{B}_{(q,k_2)}) = q < k_1 + k_2$, $h(\mathbf{B}) = k_2 < q$, a $\boldsymbol{\Sigma}$ je pozitivně definitní, pak se model nazývá regulární.

V tomto příspěvku budeme vždy uvažovat regulární model.

Věta 1 BLUE (Best Linear Unbiased Estimator) vektoru $\begin{pmatrix} \Theta \\ \beta \end{pmatrix}$ je

$$\begin{pmatrix} \hat{\Theta} \\ \hat{\beta} \end{pmatrix} = - \begin{pmatrix} (\mathbf{F}'\Sigma^{-1}\mathbf{F})^{-1}\mathbf{B}'\mathbf{Q}_{1,1} \\ \mathbf{Q}_{2,1} \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{I} - (\mathbf{F}'\Sigma^{-1}\mathbf{F})^{-1}\mathbf{C}'\mathbf{Q}_{1,1}\mathbf{C} \\ -\mathbf{Q}_{2,1}\mathbf{B}_1 \end{pmatrix} \hat{\Theta}, \quad (2)$$

kde $\hat{\Theta} = (\mathbf{F}'\Sigma^{-1}\mathbf{F})^{-1}\mathbf{F}'\Sigma^{-1}\mathbf{Y}$ (odhad nerespektující podmínku týkající se parametrů Θ, β); jeho kovarianční matice je

$$\text{var} \begin{pmatrix} \hat{\Theta} \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \text{var}(\hat{\Theta}), & \text{cov}(\hat{\Theta}, \hat{\beta}) \\ \text{cov}(\hat{\beta}, \hat{\Theta}), & \text{var}(\hat{\beta}) \end{pmatrix}, \quad (3)$$

kde

$$\text{var}(\hat{\Theta}) = (\mathbf{X}'\Sigma^{-1}\mathbf{F})^{-1} - (\mathbf{X}'\Sigma^{-1}\mathbf{F})^{-1}\mathbf{C}'\mathbf{Q}_{1,1}\mathbf{C}(\mathbf{F}'\Sigma^{-1}\mathbf{F})^{-1},$$

$$\text{cov}(\hat{\Theta}, \hat{\beta}) = -(\mathbf{F}'\Sigma^{-1}\mathbf{F})^{-1}\mathbf{C}'\mathbf{Q}_{1,2},$$

$$\text{var}(\hat{\beta}_2) = -\mathbf{Q}_{2,2}$$

a

$$\begin{pmatrix} \mathbf{Q}_{1,1} & \mathbf{Q}_{1,2} \\ \mathbf{Q}_{2,1} & \mathbf{Q}_{2,2} \end{pmatrix} = \begin{pmatrix} \mathbf{C}(\mathbf{F}'\Sigma^{-1}\mathbf{F})^{-1}\mathbf{C}' & \mathbf{B} \\ \mathbf{B}' & \mathbf{0} \end{pmatrix}^{-1}.$$

Důkaz Viz [2] str. 129-131. □

Model proudění řeky

Mějme naměřeny hodnoty $X^\times, Y^\times, Z^\times$ a V^\times parametrů X, Y, Z a V , které reprezentují v bodě o souřadnicích (x, y, z) rychlost proudění řeky.

Cílem měření je nalézt odhady skutečných hodnot $v_0, \kappa_1, \kappa_2, \kappa_3, \gamma_1, \gamma_2, \gamma_3, \gamma_4, \delta_1, \delta_2, \delta_3, \delta_4$, které reprezentují maximální rychlost na říční čáře, parametr κ_1 popisující úbytek této rychlosti v závislosti na vzdálenosti na ose x od říční čáry, parametr κ_2 popisující úbytek této rychlosti v závislosti na vzdálenosti na ose y od říční čáry, parametr κ_3 popisující úbytek této rychlosti v závislosti na vzdálenosti na ose z od říční čáry, parametry $\gamma_i, i = 1, 2, 3, 4$ popisující y -ovou souřadnici říční čáry a parametry $\delta_i, i = 1, 2, 3, 4$ popisující z -ovou souřadnici říční čáry.

Říční čáru budeme uvažovat ve tvaru

$$\begin{pmatrix} x \\ \beta_1(x) \\ \beta_2(x) \end{pmatrix} = \begin{pmatrix} x \\ \gamma_1 + \gamma_2 x + \gamma_3 x^2 + \gamma_4 x^3 \\ \delta_1 + \delta_2 x + \delta_3 x^2 + \delta_4 x^3 \end{pmatrix}. \quad (4)$$

Výsledek měření je reprezentován vektorem měřených souřadnic a rychlostí

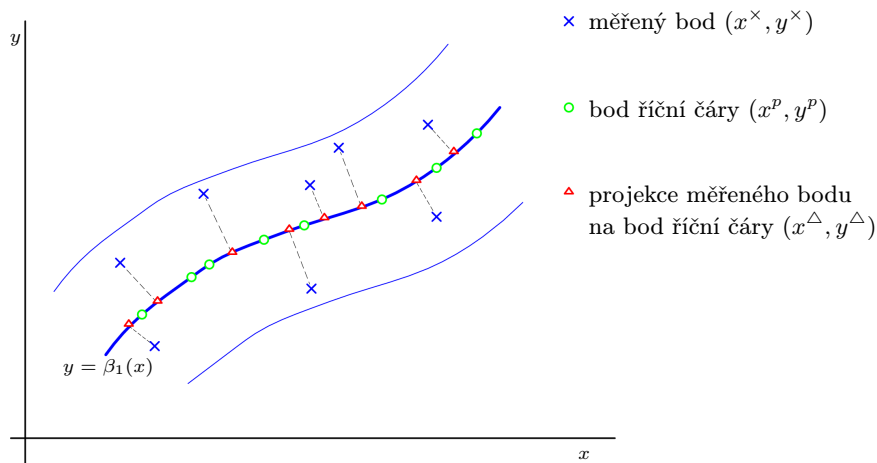
$$\mathbf{Y}_{4 \cdot I} \sim N_{4 \cdot I} \begin{bmatrix} X_1^\times \\ Y_1^\times \\ Z_1^\times \\ V_1^\times \\ \vdots \\ X_I^\times \\ Y_I^\times \\ Z_I^\times \\ V_I^\times \end{bmatrix} + \epsilon, \text{ kde } \text{Var}(\epsilon) = \Sigma. \quad (5)$$

Předpokládáme, že náhodný vektor má normální rozdělení s kovarianční maticí Σ , která je určena nejistotou měření rychlosti Pitotovou trubicí a nejistotou měření rovinných souřadnic.

Rychlost je dána rovnicí

$$V_i^\times = V_0 - (X_i^\times - X_i^\Delta)^2 \kappa_1 - (Y_i^\times - Y_i^\Delta)^2 \kappa_2 - (Z_i^\times - Z_i^\Delta)^2 \kappa_3, \quad (6)$$

kde symbol Δ označuje projekci na říční čáru (viz obr. 2).



Obr. 2: Měření, projekce a známé body na říční čáře

V našem případě

$$\Theta = [X_1^\times, Y_1^\times, Z_1^\times, V_1^\times, \dots, X_I^\times, Y_I^\times, Z_I^\times, V_I^\times, V_0, \kappa_1, \kappa_2, \kappa_3]', \quad (7)$$

$$\beta = [\gamma_1, \gamma_2, \gamma_3, \gamma_4, \delta_1, \delta_2, \delta_3, \delta_4]. \quad (8)$$

Parametry Θ a β musí splnit následující podmínky:

$$\begin{aligned} g_1^i &= X_i^\times - X_i^\Delta + (Y_i^\times - Y_i^\Delta)(\gamma_2 + \gamma_3 X_i^\Delta + \gamma_4 (X_i^\Delta)^2) + \\ &\quad + (Z_i^\times - Z_i^\Delta)(\delta_2 + \delta_3 X_i^\Delta + \delta_4 (X_i^\Delta)^2), \quad i = 1, \dots, I, \\ g_2^i &= \gamma_1 + \gamma_2 X_i^\Delta + \gamma_3 (X_i^\Delta)^2 + \gamma_4 (X_i^\Delta)^3 - Y_i^\Delta = 0, \quad i = 1, \dots, I, \\ g_3^i &= \delta_1 + \delta_2 X_i^\Delta + \delta_3 (X_i^\Delta)^2 + \delta_4 (X_i^\Delta)^3 - Z_i^\Delta = 0, \quad i = 1, \dots, I. \end{aligned} \quad (9)$$

Pomocí teorie nelineárních modelů zkonstruujeme lineární verzi modelu. Nechť $\mathbf{F} = \frac{\partial E\mathbf{Y}}{\partial \Theta}$, kde $E\mathbf{Y} = [X_1^\times, Y_1^\times, Z_1^\times, V_1^\times, \dots, X_I^\times, Y_I^\times, Z_I^\times, V_I^\times]$.

Matice \mathbf{F} má tuto strukturu

$$\mathbf{F} = \begin{pmatrix} \mathbf{F}_1^\times, \mathbf{F}_1^\Delta, \mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \mathbf{0}, \mathbf{F}_1 \\ \mathbf{0}, \mathbf{0}, \mathbf{F}_2^\times, \mathbf{F}_2^\Delta, \dots, \mathbf{0}, \mathbf{0}, \mathbf{F}_2 \\ \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \dots, \mathbf{F}_I^\times, \mathbf{F}_I^\Delta, \mathbf{F}_I \end{pmatrix}, \quad \text{kde} \quad (10)$$

$$\mathbf{F}_i^\times = \begin{pmatrix} 1, & 0, & 0 \\ 0, & 1, & 0 \\ 0, & 0, & 1 \\ -2(X^\times - X^\Delta)\kappa_1, & -2(Y^\times - Y^\Delta)\kappa_2, & -2(Z^\times - Z^\Delta)\kappa_3 \end{pmatrix},$$

$$\mathbf{F}_i^\Delta = \begin{pmatrix} 0, & 0, & 0 \\ 0, & 0, & 0 \\ 0, & 0, & 0 \\ 2(X^\times - X^\Delta)\kappa_1, & 2(Y^\times - Y^\Delta)\kappa_2, & 2(Z^\times - Z^\Delta)\kappa_3 \end{pmatrix},$$

$$\mathbf{F}_i = \begin{pmatrix} 0, & 0, & 0, & 0 \\ 0, & 0, & 0, & 0 \\ 0, & 0, & 0, & 0 \\ 1, & -(X^\times - X^\Delta)^2, & -(Y^\times - Y^\Delta)^2, & -(Z^\times - Z^\Delta)^2 \end{pmatrix}.$$

Matice \mathbf{C} má tvar

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_1^\times, \mathbf{C}_1^\Delta, \mathbf{0}, \mathbf{0}, \dots, \mathbf{0}, \mathbf{0}, \mathbf{C}_1 \\ \mathbf{0}, \mathbf{0}, \mathbf{C}_2^\times, \mathbf{C}_2^\Delta, \dots, \mathbf{0}, \mathbf{0}, \mathbf{C}_2 \\ \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \dots, \mathbf{C}_I^\times, \mathbf{C}_I^\Delta, \mathbf{C}_I \end{pmatrix}, \quad (11)$$

$$\mathbf{C}_i^\times = \begin{pmatrix} 1, \gamma_2 + \gamma_3 X^\Delta + \gamma_4 (X^\Delta)^2, \delta_2 + \delta_3 X^\Delta + \delta_4 (X^\Delta)^2 \\ 0, & 0, & 0 \\ 0, & 0, & 0 \\ 0, & 0, & 0 \end{pmatrix},$$

$$\mathbf{C}_i^\Delta = \begin{pmatrix} c_{11}, c_{12}, c_{13}, \\ c_{21}, c_{22}, c_{23} \\ c_{31}, c_{32}, c_{33} \end{pmatrix},$$

kde $c_{11} = -1 + (Y^\times - Y^\Delta)(\gamma_3 + 2\gamma_4 X^\Delta) + (Z^\times - Z^\Delta)(\delta_3 + 2\delta_4 X^\Delta)$, $c_{12} = -\gamma_2 - \gamma_3 X^\Delta - \gamma^4 (X^\Delta)^2$, $c_{13} = -\delta_2 - \delta_3 X^\Delta - \delta^4 (X^\Delta)^2$, $c_{21} = \gamma_2 + 2\gamma_3 X^\Delta + 3\gamma^4 (X^\Delta)^2$, $c_{31} = \delta_2 + 2\delta_3 X^\Delta + 3\delta^4 (X^\Delta)^2$ a $c_{22} = c_{23} = c_{32} = c_{33} = 0$,

$$\mathbf{C}_i = \mathbf{0}_{3,4}.$$

Matice \mathbf{B} je ve tvaru

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1^\gamma, & \mathbf{B}_1^\delta \\ \mathbf{B}_2^\gamma, & \mathbf{B}_2^\delta \\ \dots & \\ \mathbf{B}_I^\gamma, & \mathbf{B}_I^\delta \end{pmatrix}, \quad (12)$$

$$\mathbf{B}_i^\gamma = \begin{pmatrix} 0, & Y^\times - Y^\Delta, & (Y^\times - Y^\Delta)X^\Delta, & (Y^\times - Y^\Delta)(X^\Delta)^2 \\ 1, & X^\Delta, & (X^\Delta)^2, & (X^\Delta)^3 \\ 0, & 0, & 0, & 0 \end{pmatrix},$$

$$\mathbf{B}_i^\delta = \begin{pmatrix} 0, & Z^\times - Z^\Delta, & (Z^\times - Z^\Delta)X^\Delta, & (Z^\times - Z^\Delta)(X^\Delta)^2 \\ 0, & 0, & 0, & 0 \\ 1, & X^\Delta, & (X^\Delta)^2, & (X^\Delta)^3 \end{pmatrix}.$$

3 Projekce známých bodů na říční čáru

Nechť je říční čára v rovině (X, Y) dána explicitně zadanou funkcí $y = \beta_1(x)$ (viz obr. 2). Analogicky uvažujme funkci $z = \beta_2(x)$ v rovině (X, Z) .

Ve skutečnosti máme k dispozici měření rychlosti a rovinných souřadnic na I místech říčního koryta. Tyto body budeme označovat v grafu symbolem \times and jejich souřadnice symboly $x^\times, y^\times, z^\times$, rychlost v těchto bodech označíme v^\times . Observační vektor Y má pak délku $4 \cdot I$.

V naší numerické studii číslo I je přibližně 2000, tato měření byla získána v korytě Dunaje při výzkumném projektu Universität für Bodan Kultur (BOKU) v Rakousku.

Vektory parametrů θ a β uvedené ve vzorcích (7) a (8) mají délku $4 \cdot I + 4$, resp. délku 8.

Dále předpokládejme, že máme již k dispozici tvar říční čáry ve tvaru (4).

Parametry γ_i , $i = 2, 3, 4$ popisují y -ovou souřadnici říční čáry a parametry δ_i , $i = 2, 3, 4$ popisují z -ovou souřadnici říční čáry.

V našem případě numerická studie prokázala, že můžeme uvažovat funkce $\beta_1(x)$ a $\beta_2(x)$ ve tvaru polynomů třetího stupně. Stupeň polynomů závisí na tvaru říčního koryta.

Budeme uvažovat, že říční tok splňuje požadavky na laminární proudění. Z tohoto předpokladu plyne, že rychlost klesá s rostoucí vzdáleností od říční čáry.

Poznámka: Nultou aproximaci tvaru říční čáry můžeme nalézt pomocí:

1. známých bodů na levém a pravém břehu řeky,
2. bodů, kde byla naměřena největší rychlost (nespolehlivé, vzhledem k různému spádu).

Dále při vytváření modelu budeme uvažovat maximální rychlost w a parametr κ_1 , který určuje, jak rychle se snižuje rychlost na říční čáře v bodech ležících na normále k ní (ve vodorovném směru).

Parametr κ_2 bude určovat, jak rychle se snižuje rychlost na říční čáře v bodech ležících na normále k ní (ve svislém směru).

Body na říční čáře označíme $x_p, \beta_1(x_p)$, respektive v obrázku 1 symbolem o .

Naším cílem je nyní najít projekci libovolného měřeného bodu $x^\times, y^\times, z^\times$ na bod na říční čáře, ten označíme $x^\Delta, y^\Delta, z^\Delta$.

Uřídíme směrnice tečny k říční čáře (tu nyní uvažujeme, že je popsána explicitně $y_p = \beta_1(x_p)$). Derivace je

$$\beta_1'(x_p) = \frac{y^\Delta - \beta_1(x_p)}{x^\Delta - x_p}, \quad (13)$$

$$\beta_1'(x_p) = \frac{x - x^\Delta}{\beta_1(x_p) - y^\Delta}. \quad (14)$$

Z rovnice (13) a (14) dostáváme po úpravě

$$\beta_1'(x_p) \cdot x^\Delta - y^\Delta = \beta_1'(x_p) \cdot x_p - \beta_1(x_p), \quad (15)$$

$$-x^\Delta + \beta_1'(x_p) \cdot y^\Delta = \beta_1'(x_p) \cdot \beta_1(x_p) - x_p. \quad (16)$$

Souřadnice projekce zvoleného bodu na říční čáru pak najdeme řešením této soustavy rovnic

$$\begin{pmatrix} \beta_1'(x_p), & -1 \\ -1, & \beta_1'(x_p) \end{pmatrix} \begin{pmatrix} x^\Delta \\ y^\Delta \end{pmatrix} = \begin{pmatrix} \beta_1'(x_p) \cdot x_p - \beta_1(x_p) \\ \beta_1'(x_p) \cdot \beta_1(x_p) - x_p \end{pmatrix},$$

nebo po úpravě

$$\begin{pmatrix} \beta_1'(x_p), & -1 \\ \frac{1}{\beta_1'(x_p)}, & -1 \end{pmatrix} \begin{pmatrix} x^\Delta \\ y^\Delta \end{pmatrix} = \begin{pmatrix} \beta_1'(x_p) \cdot x_p - \beta_1(x_p) \\ -\beta_1(x_p) + \frac{x_p}{\beta_1'(x_p)} \end{pmatrix}.$$

Nejprve určíme inverzi k matici na levé straně:

$$A^{-1} = \frac{1}{(\beta_1'(x_p))^2 - 1} \cdot \begin{pmatrix} \beta_1'(x_p), & -\beta_1'(x_p) \\ 1, & -(\beta_1'(x_p))^2 \end{pmatrix}.$$

Získanou matici přenásobíme zprava vektorem pravé strany:

$$A^{-1} \cdot b = \frac{1}{(\beta_1'(x_p))^2 - 1} \cdot \begin{pmatrix} (\beta_1'(x_p))^2 \cdot x_p - \beta_1'(x_p) \cdot \beta_1(x_p) + \beta_1'(x_p) \cdot y^\times - x^\times \\ \beta_1'(x_p) \cdot x_p - \beta_1(x_p) + (\beta_1'(x_p))^2 \cdot y^\times - \beta_1'(x_p) \cdot x^\times \end{pmatrix}$$

Po úpravě dostáváme nové souřadnice

$$\begin{pmatrix} x^\Delta \\ y^\Delta \end{pmatrix} = \frac{\beta_1'(x_p)}{1 - (\beta_1'(x_p))^2} \cdot \begin{pmatrix} (-\beta_1'(x_p))^2 \cdot x_p + \beta_1(x_p) - y^\times + \frac{x^\times}{\beta_1'(x_p)} \\ -x_p + \frac{\beta_1(x_p)}{\beta_1'(x_p)} - \beta_1'(x_p) \cdot y^\times + x^\times \end{pmatrix}. \quad (17)$$

Závěr

V článku jsme upozornili na způsob, jak modelovat regresním modelem proudění řeky. V této práci jsme určovali rychlost říčního toku na základě měření v říčním korytě a na základě modelu, který jsme zkonstruovali. Možnost sestavit tento regresní model vyšla z předpokladu, že rychlost v určovaném bodě v říčním korytu je určena rychlostí na říční čáře a je funkcí pouze rovinných souřadnic tohoto bodu.

Příspěvek tvoří pouze úvod do prezentované úlohy a bude na něj navazovat rozsáhlá numerická studie. Dále je zvažováno použití dalších metod jako je Kriging nebo IDW.

References

- [1] Kubáček, L., Kubáčková, L.: Statistics and Metrology (in Czech), Publishing House of Palacký University, Olomouc, 2000
- [2] Kubáček, L., Kubáčková, L., L., Volaufová, J.: Statistical models with linear structures. Veda, Publishing House of the Slovak Academy of Sciences, Bratislava, 1995.



Fuzzy škály v psychologii

BLANKA ZEMKOVÁ

AXIOM SW, s.r.o.,
Pod Vrškem 5360, 760 01 Zlín
e-mail: blanka.zemkova@seznam.cz

Abstrakt: V článku je ukázána možnost využití fuzzy škál v psychometrii, konkrétně v oblasti hodnocení pracovníků. První část je věnována definici škály a základním typům škál. Dále je zaveden pojem jazykové škály. Škálovými stupni jsou hodnoty jazykové proměnné, tedy výrazy přirozeného jazyka, jednotlivé škálové stupně jsou modelovány pomocí fuzzy čísel. Jazyková fuzzy škála je kardinální škálou. Je ukázáno několik speciálních struktur odvozených na základě jazykových fuzzy škál. Další část článku je věnována konkrétním škálám využívaným v hodnocení pracovníků. Jsou uvedeny příklady škály třídílné, čtyřdílné, pětídílné a dvou různých šestídílných škál. Pro každou uvedenou škálu je dále navržen způsob, jak je možno tuto škálu modelovat pomocí jazykových fuzzy škál, který typ škály je vhodný pro modelování dané škály využít.

1 Škála

Pojem škály je jedním ze základních pojmů, kterému se v teorii měření přisuzuje velký význam. Pod tímto pojmem se nechápe jenom výsledek měření, ale i měřicí prostředek nebo jen stupnice nějakého měřícího přístroje.

Původně byla škála chápána jako měřicí přístroj, resp. jeho specifická součást (stupnice). Jakmile došlo k extenzi měření do společenskovedních oborů, začaly se měřit i takové nemetrické veličiny, jako např. inteligence, pocity, postoje, pro něž se nedá zkonstruovat nějaký měřicí přístroj. Pojem škály tehdy nabyl mnohostrannější funkce jak z hlediska teoretického, tak i metodologického.

V následujícím textu bude pod pojmem škála chápána určitá posloupnost hodnot představujících stupně metrických i nemetrických veličin.

Ve formální teorii měření je za škálu považována uspořádaná trojice $\langle E, N, \Phi \rangle$, kde E je empirický relační systém, N je numerický relační systém a Φ je izomorfní nebo homomorfní zobrazení z E do N .

1.1 Typy škál

Z hlediska metodologického je možno rozlišovat škály materiální a konceptuální. Uvažujeme-li nějaký měřicí prostředek, vzhledem k němuž se objektivně srovnává a uspořádává, můžeme každé dvojici <veličina, měřicí přístroj> přiřadit dvě různé škály: konceptuální a materiální. Tato diferenciaci se vztahuje především na škály metrických veličin.

1.1.1 Konceptuální škála

Konceptuální škála je charakterizována určitým uspořádaným intervalem (dle [2] numerických) hodnot (tzv. škálových hodnot), které lze teoreticky přisuzovat měřeným veličinám v závislosti na jejich kvalitativních a kvantitativních aspektech.

Dle [2] škálové hodnoty nemohou být jazykovými výrazy, ale pouze mimojazykovými entitami (číslly) z toho důvodu, aby bylo možno určit uspořádání nebo velikosti měřených veličin.

Jak bude ukázáno dále, jiné pojetí představuje možnost vyjádření škálových hodnot pomocí hodnot jazykové proměnné modelovaných fuzzy čísly. Taková škála má pak podobné vlastnosti, jako např. již zmíněnou možnost uspořádání hodnot.

Forma konceptuální škály je vymezena třemi aspekty: charakterem škálových hodnot, počátkem škály (škálovou nulou) a povahou distance mezi libovolnými dvěma bezprostředně sousedícími škálovými hodnotami v závislosti na měrné jednotce.

1.1.2 Materiální škála

Materiální škála je určena uspořádanou množinou znaků na měřicím přístroji, většinou číslic, jejichž čtení umožňuje přiřazovat numerické hodnoty k velikostem měřených veličin. Z důvodu lepšího rozlišování bývá zvykem používat pojem škála pouze pro konceptuální škály a materiální škály označovat pojmem stupnice. V dalším textu budou uvažovány pouze škály konceptuální.

1.2 Klasifikace škálových typů

V závislosti na chápání pojmu měření jsou rozlišovány širší a užší klasifikační systémy. V širších systémech jsou za škály měření považovány všechny škály, metrické a nemetrické, mezi nemetrické se řadí i ty škály, ke kterým dospíváme pouhým číslicovým přiřazováním. V užších klasifikačních systémech je respektován rozdíl mezi očíslováním, škálováním a měřením. Škálové typy jsou tříděny v závislosti na diferenciaci topologických a metrických podmínek měření. Pokud příslušné veličiny splňují pouze podmínky topologizace, a jejich škálovými hodnotami jsou tedy pouze ordinální čísla, odpovídají jim nemetrické škály (topologické), splňují-li podmínky metrizace, a jejich škálovými hodnotami jsou tedy kardinální čísla, korespondují jim metrické (aditivní) škály.

Širší klasifikační systémy vycházejí ze Stevensova systému a najdeme je v každé učebnici psychologického škálování. Vychází ze tří základních kritérií: empirické operace, matematické struktury škál a použitelné statistiky při zpracování získaných numerických údajů.

Ve skutečnosti jsou pro posouzení jeho koncepce relevantní pouze první dvě, třetí kritérium je již záležitostí praktické povahy a nesouvisí bezprostředně s konstrukcí škálových typů.

Kritérium empirické operace sleduje, jaké operace je možno se škálovými hodnotami provádět, zda je lze uspořádat, zda se dá v případě uspořádání něco říct o vzdálenostech mezi jednotlivými hodnotami apod.

Druhé kritérium sleduje, jaká matematická struktura je tvořena škálovými hodnotami (grupová struktura škálových typů podmiňuje přípustné transformace, za nichž se forma škály nemění).

Na základě těchto kritérií jsou rozlišovány následující typy škál: nominální škála, ordinální škála, intervalová škála a poměrová škála.

1.2.1 Nominální škála

Jedná se o nejelementárnější škálový typ, je založena na libovolném přiřazování číslic ve funkci pouhého pojmenování, rozřídění objektů podle určité vlastnosti a označení příslušnosti ke stejné třídě. Základní empirickou operací je operace rovnosti. Lze stanovit pouze to, zda dva prvky patří do téže třídy nebo nepatří. Matematická struktura je vymezena permutační grupou $x' = f(x)$, kde x a x' označují libovolné číslice a $f(x)$ vzájemně jednoznačnou substituci. Tento škálový typ zůstává invariantní vzhledem ke každé změně číslicového přiřazení, které zachovává identitu třídy.

1.2.2 Ordinální škála

U ordinálních škál se předpokládá přirozené uspořádání objektů vzhledem k nějaké vlastnosti. Je vymezena dvěma empirickými operacemi: určením rovnosti a určením vztahu více nebo méně. Matematická struktura je charakterizována izotonickou (uspořádání zachovávající) grupou $x' = f(x)$, přičemž $f(x)$ je v tomto případě jakákoli monotónně vzrůstající nebo klesající funkce. Forma ordinálních škál zůstává invariantní vzhledem ke každé monotónní transformaci.

1.2.3 Intervalová škála

Intervalová škála vyžaduje stanovení měrné jednotky a arbitrárního počátku (nějaký bod zvolený konvenčně jako nejvýhodnější počátek). Rozdíl sousedních bodů na stupnici je stále stejný, není důležité, zda se všechny body stupnice v reálném životě vyskytnou. Intervalové škály nemají přesně definovaný absolutní nulový bod (např. nelze stanovit nulovou inteligenci). Intervalové škály jsou invariantní vůči libovolné lineární transformaci $y = ax + b$, intervalová škála je charakterizována obecnou lineární grupou $x' = ax + b$, $a > 0$.

1.2.4 Poměrová škála

U poměrové škály je poměr intervalů mezi libovolnými dvěma sousedními body stejný jako poměr mezi kterýmikoli jinými sousedními body. Musí mít pevný počátek, volba měrné jednotky je libovolná, hodnoty na škále mají pouze kladná znaménka. Poměrové škály lze měnit libovolnou transformací typu $y = ax$. Struktura poměrových škál je vyjádřena podobnostní grupou $x' = ax$, $a > 0$, kde a označuje měrnou jednotku.

1.3 Jazyková škála

Poněkud širší pohled představuje koncepce jazykové škály (viz [1]). Škálovými hodnotami jazykové škály jsou hodnoty jazykové proměnné (tedy výrazy přirozeného jazyka), které navíc splňují další podmínky (viz dále). Jednotlivé škálové hodnoty můžeme modelovat pomocí fuzzy čísel. Taková jazyková fuzzy škála je kardinální škálou, s jejími hodnotami můžeme dále pracovat (porovnávat je, měřit jejich vzdálenosti) stejně jako u kardinálních škál.

Jazykovou proměnnou je možno definovat jako pěticí $(\nu, T(\nu), X, G, M)$, kde ν je jméno jazykové proměnné, $T(\nu)$ množina jejích jazykových hodnot, X je univerzum, na němž jsou definována fuzzy čísla představující významy jazykových hodnot, G je syntaktické pravidlo pro generování jazykových hodnot a M je sémantické pravidlo, které každé jazykové hodnotě přiřadí její význam, který je fuzzy množinou na X .

Předpokládáme, že významy hodnot jazykové proměnné jsou definovány fuzzy čísly na intervalu reálných čísel, speciálně na některém uzavřeném intervalu $\langle A, B \rangle$.

Předpokládáme, že T_1, \dots, T_n jsou fuzzy čísla definovaná na $\langle A, B \rangle$, očíslovaná ve shodě s jejich uspořádáním, která představují fuzzy rozklad intervalu $\langle A, B \rangle$, tj.

$$\forall x \in \langle A, B \rangle : \sum_{i=1}^n T_i(x) = 1. \quad (1)$$

Pak řekneme, že fuzzy čísla T_1, \dots, T_n tvoří na intervalu $\langle A, B \rangle$ fuzzy škálu.

Dále předpokládáme, že $(\nu, T(\nu), X, G, M)$ je jazyková proměnná, T_1, \dots, T_n jsou fuzzy čísla, která tvoří na intervalu $\langle A, B \rangle$ fuzzy škálu. Nechť pro hodnoty jazykové proměnné T_i platí $T_i = M(T_i)$. Pak řekneme, že hodnoty jazykové proměnné T_i tvoří jazykovou fuzzy škálu.

Jazykové škály splňují další vlastnosti přirozeně očekávané od neurčitých jazykově definovaných škál. Významy jazykových hodnot jsou lineárně uspořádány, každý bod intervalu buď plně náleží některé hodnotě jazykové škály, nebo je jeho náležení rozděleno mezi dvě po sobě jdoucí hodnoty jazykové škály (tam, kde končí jádro jednoho fuzzy čísla začíná nosič fuzzy čísla následujícího a tam, kde začíná jádro tohoto fuzzy čísla, končí nosič fuzzy čísla předcházejícího), a konečně významy jazykových hodnot jsou modelovány fuzzy čísly se spojitou funkcí příslušnosti.

Pro modelování jazykových hodnot jsou vhodná lineární nebo kvadratická fuzzy čísla zadaná čtveřicemi bodů.

V praxi jsou většinou používány rovnoměrné jazykové škály, kdy velikosti jader a nosičů všech fuzzy čísel představujících významy jazykových hodnot dané jazykové proměnné jsou stejné (kromě nosičů krajních fuzzy čísel). Není to však jediná možnost, jak bude ukázáno dále.

Jazykové hodnoty jsou definovány expertně, což vychází z možností přirozeného jazyka, podobně jako jejich významy. Počet škálových hodnot závisí na podrobnosti znalostí o hodnotách dané proměnné. Nedochází zde ke vzniku takových nepřirozených složenin, které se v přirozeném jazyce nepoužívají, nebo jejichž významy nekorespondují s významy v přirozeném jazyce.

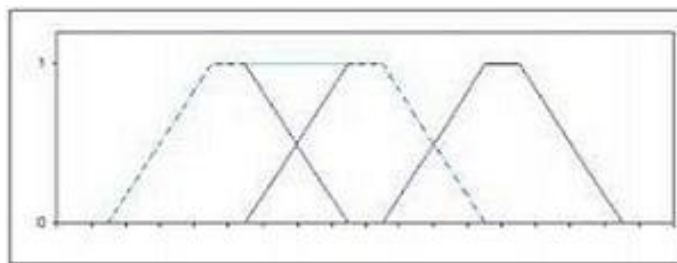
2 Speciální fuzzy škály a struktury na jejich základě

Jsou odvozeny ze základní jazykové škály (viz výše) při dodržení základních pravidel: a) množina elementárních termů tvoří jazykovou škálu; b) odvozené termy musí představovat jazykové hodnoty, které jsou obvyklé a mají jasně vymezený význam; c) odvozené termy jsou zaváděny systematicky na celé základní škále; d) pro modelování významů odvozených termů jsou používána pravidla, která jsou zvolena jednoduše pro celý model.

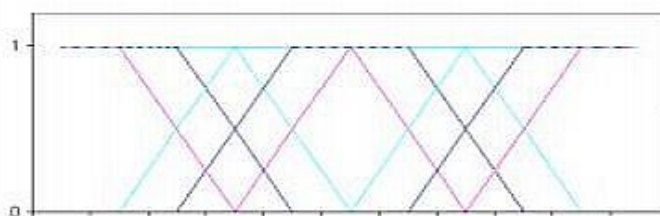
2.1 Rozvinutá jazyková škála

Rozvinutá jazyková škála vychází ze základní škály, kromě množiny základních termů obsahuje dále termy „ T_i až T_j “ představující sjednocení sousedních fuzzy čísel definované pomocí Lukasiewiczovy disjunkce; maximální neurčitost (tj. každá hodnota intervalu $\langle A, B \rangle$ je stejně možná) je přitom vyjádřena pojmem „neurčeno“ (je možno využít například pro modelování chybějících dat).

Pak platí množina fuzzy čísel modelujících významy jazykových hodnot rozvinuté jazykové škály tvoří vzhledem k relaci uspořádání fuzzy čísel úplný svaz. Množina fuzzy čísel modelujících významy jazykových hodnot rozvinuté jazykové škály tvoří vzhledem k relaci inkluze fuzzy množin úplný horní polosvaz.



Obrázek 36: Rozvinutá jazyková škála



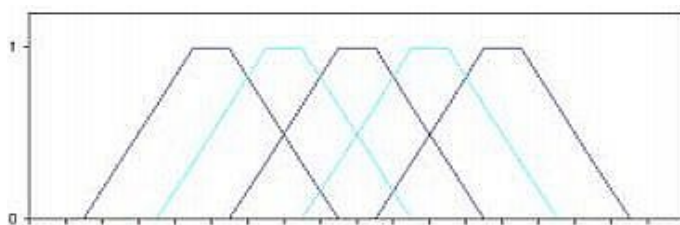
Obrázek 37: Obohacená jazyková škála

2.2 Obohacená jazyková škála

Obohacená jazyková škála vychází rovněž ze základní škály, kromě množiny základních termů obsahuje dále termy „určitě T_i “, „víceméně T_i “, operátor „určitě“ značí snížení neurčitosti, operátor „víceméně“ zvýšení neurčitosti. Posloupnosti fuzzy čísel představujících významy termů odvozených střídavou aplikací jazykových operátorů „určitě“ a „víceméně“ tvoří fuzzy škály na intervalu $\langle A, B \rangle$, tj. ve stejné míře, v jaké daná reálná hodnota přestává „určitě“ odpovídat dané jazykové hodnotě škály, začíná „víceméně“ korespondovat s hodnotou sousední.

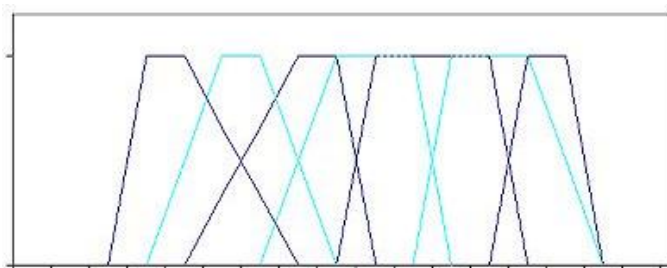
2.3 Škála s mezihodnotami

Myšlenka škály s mezihodnotami vychází z často se opakující situace, kdy se proband zdráhá zvolit některou z variant a kloní se spíše k možnosti „něco mezi“, krajním hodnotám odpovídají termy „ještě méně než T_1 “, „ještě více než T_n “. Tato škála obsahuje kromě základních termů termy „mezi T_i a T_{i+1} “, „ještě méně než T_1 “, „ještě více než T_n “.



Obrázek 38: Škála s mezihodnotami

Postup konstrukce „mezihodnot“ by mohl být následující: vycházet můžeme buď ze škály rovnoměrné nebo nerovnoměrné, u nerovnoměrné škály předpokládáme, že její hodnoty mají obecně různé velké zóny neurčitosti a velikosti jejich jader se taky mohou měnit (v tomto případě zmenšují směrem ke krajním hodnotám). Pokud vycházíme z nerovnoměrné škály, měla by být nerovnoměr-

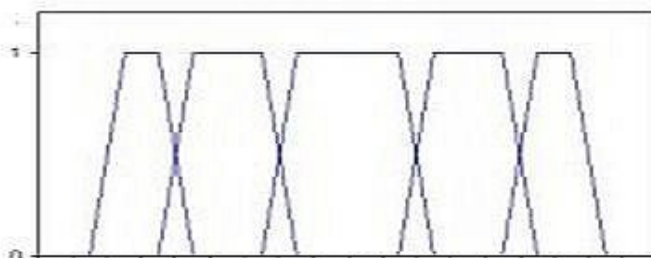


Obrázek 39: Škála s mezihodnotami vycházející z nerovnoměrné škály

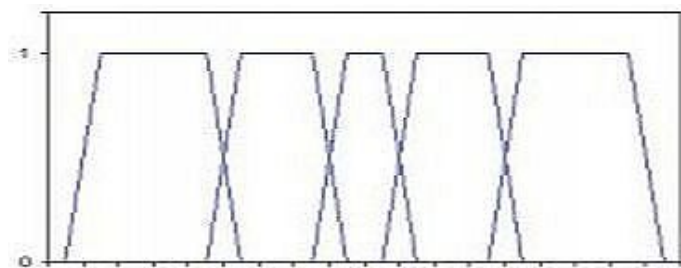
nost zachována i v případě mezihodnot. Velikosti jader mezihodnot by měly mít velikost průměru velikostí jader, mezi nimiž jsou umístěny (tj. v případě rovnoměrné škály stejně velké jádro jako hodnoty základní škály), levá (pravá) zóna neurčitosti by měla mít velikost průměru levých (pravých) zón neurčitosti hodnot základní škály, mezi nimiž je mezihodnota umístěna (u rovnoměrných i nerovnoměrných škál). jádra mezihodnot by byla umístěna tak, aby střed jádra mezihodnoty byl umístěn uprostřed mezi středy jader hodnot základní škály, mezi nimiž je mezihodnota umístěna.

2.4 Nerovnoměrná škála

Varianta 1. Koncepce vychází z předpokladu, že lidé mají sklony volit „zlatou střední cestu“ a vyhýbají se extrémním hodnotám. Pokud tedy proband zvolí krajní hodnotu, předpokládá se, že si je svou volbou více jistý než proband, který volí středovou hodnotu, a tato hodnota má tedy větší vypovídací hodnotu, než hodnota středová. Nejmenší neurčitost tedy mají krajní hodnoty, s přibližováním se ke středové hodnotě neurčitost roste a u středové hodnoty je maximální. Tato varianta bude připadat v úvahu v psychologických aplikacích, tj. v případech, kde se při volbě škálové hodnoty projevuje subjektivní vnímání probanda.



Obrázek 40: Nerovnoměrná jazyková škála - varianta 1



Obrázek 41: Nerovnoměrná jazyková škála - varianta 2

Varianta 2. Opačný přístup je zpracován v [5] (na příkladu výnosu akcií), zde je naopak za nejméně neurčitou považována středová hodnota škály a směrem ke krajním hodnotám neurčitost vzrůstá. Tato varianta bude připadat v úvahu v takových aplikacích, kde je nahuštění kolem středových hodnot dáno objektivními požadavky na hodnocení, tj. je žádoucí větší diferenciaci kolem nejpravděpodobnějších hodnot.

Kromě výše uvedených speciálních škál je možno uvažovat i o jiných možnostech, např. vycházet z nerovnoměrné jazykové škály a ke konstrukci odvozených termů použít myšlenky obohacené jazykové škály. Tím dosáhneme toho, že budou brány v úvahu dva typy neurčitosti: jednak neurčitost, kterou proband vědomě zvyšuje či zmenšuje použitím odvozených termů „určitě T_i “ a „více méně T_i “, jednak neurčitost spontánní, která plyne automaticky z tendence vyhýbat se krajním hodnotám a volit spíše hodnoty středové.

Typ jazykové škály, kterou je vhodné v daném modelu použít, zcela závisí na povaze modelu. To je zřejmé např. na výše uvedeném příkladu nerovnoměrných škál, kdy v hodnocení výsledků psychodiagnostických testů se lze přiklonit spíše k první variantě (největší neurčitost má středová hodnota, krajní hodnoty mají větší vypovídací schopnost), zatímco při modelování ekonomických ukazatelů (objektivní ukazatele, nezávislé na vůli hodnotitele) je podrobněji diferencována oblast středových hodnot a krajní hodnoty mají naopak velkou neurčitost. Z tohoto důvodu by volba typu škály a její vlastní konstrukce měla být v rukou experta pro danou oblast.

3 Použití fuzzy škál v psychometrii

3.1 Škály v psychologickém hodnocení

Psychologické procesy jsou velmi složité a poznání jejich zákonitostí s sebou nese řadu problémů. Většinu psychologických vlastností a procesů nelze měřit klasickým fyzikálním způsobem. Je těžké měřit například odolnost vůči stresu, intenzitu strachu a podobné jevy. Proto se začaly postupem času vytvářet nové metody měření, které vhodnějším způsobem zpracovávaly ty druhy dat, které má psychologie k dispozici. Tyto kvantifikační metody procházejí v posledních

letech prudkým rozvojem. Jejich výhodou je rozmanitost co do složitosti i rozsahu aplikovatelnosti. Jednou z forem kvantifikace psychologických jevů je psychologické škálování.

Data získaná z psychologických výzkumů jsou většinou spíše povahy nominální nebo ordinální, snahou bylo přiblížit získaná data kvantitativním stupnicím. Z tohoto důvodu se postupem času vyvinulo velké množství škálovacích technik (podrobně v [3]).

Pro ilustraci je dále uvedeno několik příkladů konkrétních škál používaných v psychologickém hodnocení. Podkladem pro zpracování byla především publikace [4], která je věnována metodám hodnocení zaměstnanců. Důvodem je skutečnost, že se jedná o publikaci novou a metody hodnocení, které jsou zde popsány, se skutečně v praxi ve velkých firmách používají.

Třídílná škála obsahuje hodnoty podprůměr, průměr a nadprůměr. Praxe ukazuje, že nejvyšší hodnotu získá v průměru 20 procent a nejnižší hodnotu 10 procent hodnocených.

Čtyřdílná škála nemá středovou hodnotu, resp. má středové hodnoty dvě (lepší a horší střed). Prostředním dvěma hodnotám v praxi odpovídá 65 až 75 procent, nejvyšší hodnotě 15 až 25 procent, nejnižší hodnotě 10 procent. Tato škála tedy v podstatě poskytuje větší diferenciaci ve středové oblasti.

Pětídílná škála je dána bodovými hodnotami 0-4 (resp. 1-5). Těmto hodnotám je přiřazen slovní popis, který dané hodnoty charakterizuje. Přestože tato škála má středovou hodnotu, podle slovního popisu tato středová hodnota reprezentuje spíše kvalitu mírně pod středem. Škála má následující hodnoty: 0 nedostatečná, ohrožující úroveň; 1 podprůměrná, limitující úroveň, 2 postačující, minimální úroveň, 3 optimální úroveň, 4 excelentní úroveň, ideální stav.

Šestídílná škála V publikaci věnované hodnocení pracovníků jsou představeny dvě odlišné šestídílné škály: jednak škála, z níž je patrná velká snaha vyhnout se středovým hodnotám a největší rozestup je právě mezi 3. a 4. stupněm. Tato škála má jednotlivé škálové stupně charakterizovány následovně: 1 mizivá úroveň, 2 nízká úroveň, 3 spíše nízká úroveň, 4 spíše vysoká úroveň, 5 vysoká úroveň, 6 velmi vysoká úroveň. Dále je představena škála, má středovou hodnotu, která je, podobně jako u pětídílné škály, posunuta mírně dolů. Jednotlivé stupně jsou charakterizovány následovně: E excelentní kvalita, G velmi dobrá kvalita, A+ lepší průměr, A průměr, A- podprůměr, P nízká kvalita.

Desetídílná škála poskytuje dostatek stupňů diferenciaci a umožňuje snadnější přizpůsobení normálnímu rozdělení, je však pravděpodobné, že každý bude přisuzovat jednotlivým stupňům jiný obsah. Je proto nutný slovní komentář. Není však realistické mít dostatečně diferencující slovní pojmenování pro všechny stupně.

Je rozlišováno několik typů škál, neverbální, verbální a kombinované. Mezi neverbální škály patří škály grafické a numerické. Tyto škály jsou jednoduché, nejčastěji používané, časově nenáročné. Hodnoceny jsou jimi jednotlivé atributy práce (rysy zaměstnance), pro každou položku hodnotitel „oznámkuje“ zaměstnance. Zatímco na grafické stupnici se hodnotitel rozhoduje o tom, co jednotlivá hodnocení znamenají, u verbální stupnice je každý stupeň pojmenován. U kombinovaných škál jsou uplatňovány grafické, numerické i verbální vyjádření ke každé charakteristice.

K celkovému hodnocení podle většího počtu kritérií jsou používány různé jednoduché metody, např. grafická hodnotící stupnice, která v podstatě není žádnou agregací, ale pouze grafickým vyjádřením hodnocení podle jednotlivých kritérií (hodnocení podle jednotlivých kritérií je vyjádřeno graficky body a jejich spojením vznikne profil hodnoceného probanda). V uvedené publikaci zabývající se hodnocením pracovníků jiná metoda není uvedena, pouze je konstatováno, že celkové hodnocení lze jen těžko získat sečtením nebo zprůměrnováním pořadí podle jednotlivých kritérií. Spíše je naznačena možnost použití více metod hodnocení (různé kontrolní metody) a srovnání výsledků.

3.2 Fuzzy škály psychometrii

V psychometrii se setkáváme především škálami ordinálními, což souvisí s tím, v jaké formě požadujeme odpověď na dané otázky. Získaná data ve většině případů nelze přesně měřit (fyzikálními přístroji). Jak vstupní hodnoty, tak i vztahy mezi veličinami mají do jisté míry neurčitý charakter. A právě teorie fuzzy množin je matematický aparát, který s takovou neurčitostí pracuje.

K modelování vstupních dat, tedy hodnocení varianty podle daného kritéria (hodnocení určitého znaku či vlastnosti u daného probanda) je možno použít místo běžně užívaných škál jazykové fuzzy škály. Výhodou je jednak možnost zadávat hodnocení dané vlastnosti jazykově pomocí běžně užívaných výrazů, což je pro většinu hodnotitelů přirozenější a srozumitelnější. Další výhodou je možnost adekvátního zpracování získaných dat. Metody, pomocí nichž agregujeme dílčí hodnocení v hodnocení celkové, s neurčitostí vstupních dat pracují, je tedy vhodnější zadávat neurčitá data neurčitě, tedy pomocí fuzzy škál, než prostřednictvím běžných škál, kdy vstupní data svou neurčitost ztrácí.

Jak je vidět z výše uvedených příkladů škál používaných v psychometrii, vyskytují se škály různých typů. Některé jsou rovnoměrné, některé nerovnoměrné, některé mají středovou hodnotu, u některých je zřejmá snaha se středové hodnotě vyhnout nebo ji posunout směrem s vyšším nebo nižším hodnotám.

Z těchto důvodů je třeba ke každé tvorbě škály přistupovat jednotlivě, se zřetelem k tomu, jakou charakteristiku hodnotíme a jaká data od hodnocení požadujeme. K tomu by mělo být využíváno znalostí expertů z daného oboru, pokud možno většího počtu, aby se eliminovaly rozdíly v chápání pojmů označujících jednotlivé škálové hodnoty u jednotlivých expertů. Samozřejmým předpokladem je, že hodnoty jazykových škál musí být modelovány v souladu s významem používaných jazykových hodnot v přirozeném jazyce.

Třídílná škála Pokud vyjdeme z příkladů škál uvedených v předcházející podkapitole, za nejjednodušší škálu je možno považovat škálu se třemi hodnotami. Jedná se o škálu rovnoměrnou, se středovou hodnotou, která i významem odpovídá středu. Tuto škálu by bylo možno v hodnocení modelovat základní rovnoměrnou fuzzy škálou.

Čtyřdílná škála Tato škála v podstatě kopíruje škálu třídílnou, s tím rozdílem, že středová hodnota je rozdělena na dvě části, takže umožňuje zjemnění ve středové oblasti. Hodnota vyjadřující skutečný střed chybí. Jak bylo uvedeno výše, v praxi se ukazuje podobné rozložení, jako u škály třídílné. Nejhorší hodnotě odpovídá asi 10 procent a nejlepší asi 20 procent hodnocení. Prostředním dvěma hodnotám čtyřdílné škály odpovídá přibližně stejný počet hodnocení jako prostřední hodnotě škály třídílné.

Pokud bychom tuto čtyřdílnou škálu modelovali pomocí rovnoměrné škály se čtyřmi hodnotami, došlo by pravděpodobně k posunu významu škálových hodnot směrem ke krajním hodnotám. Aby k tomuto jevu nedošlo, bylo by vhodnější použít v tomto případě fuzzy škálu nerovnoměrnou, kde prostřední dvě hodnoty budou mít menší neurčitost, než hodnoty krajní, a jejich vzdálenost bude menší.

Pětidílná škála Uvedená pětidílná škála nemá významově středovou hodnotu, prostřední škálový stupeň odpovídá spíše významu mírně pod středem.

Hodnota, která by odpovídala středu, by byla někde mezi druhým a třetím škálovým stupněm. Ze slovních popisů jednotlivých stupňů je zřejmé, že vzdálenost mezi druhým a třetím stupněm je větší, než mezi ostatními stupni, škála je více diferencující v nízkých hodnotách a méně ve vyšších. Nicméně v komentáři k této škále je uvedeno, že rozložení víceméně odpovídá Gaussovu rozdělení.

U této škály by bylo vhodné provést analýzu, zda by z hlediska hodnocení nebyla vhodná větší diferenciace mezi druhým a třetím škálovým stupněm. Mezi postačující minimální a optimální úrovní je z psychologického hlediska dost velký rozestup a dá se předpokládat, že v mnoha případech nebude pro hodnotitele snadné se pro jeden z těchto stupňů rozhodnout.

Pokud bychom chtěli zachovat stávajících pět škálových stupňů, aby nedošlo k posunu významů, bylo by možno škálu modelovat pomocí rovnoměrné fuzzy škály. Velký rozestup mezi druhým a třetím škálovým stupněm by bylo možno překonat doplněním „mezihodnot“.

Šestidílná škála U první zmíněné šestidílné škály je patrná snaha vyhnout se středové hodnotě, tomu odpovídá i velký rozestup mezi třetím a čtvrtým stupněm. Pokud bychom použili fuzzy škálu s „mezihodnotami“, je pravděpodobné, že bychom popřeli záměr hodnocení (rozdělit hodnoty na „lepší než střed“ a „horší než střed“). Vzhledem k tomu, že vzdálenost mezi třetím a čtvrtým stupněm je větší, než mezi ostatními sousedními stupni, byla by spíše vhodnější fuzzy škála nerovnoměrná.

Pokud bychom chtěli hodnotiteli umožnit větší diferenciaci mezi třetím a čtvrtým škálovým stupněm, a přitom se vyhnout středové hodnotě, bylo by možno doplnit stupně dva, „lepší střed“ a „horší střed“. V takovém případě by ale bylo nutno posoudit rovnoměrnost škály zejména v oblasti nově přidávaných stupňů.

Posledním uvedeným příkladem je škála šestidílná, která má středovou hodnotu posunutou směrem k horším hodnotám. Přesto se lze podle slovního popisu domnívat, že významový rozestup mezi jednotlivými stupni je přibližně stejný, posunutí středové hodnoty je způsobeno spíše privací stupně vyjadřujícího „velmi nízkou kvalitu“.

Škálu by bylo možno modelovat pomocí rovnoměrné škály, s doplněním posledního škálového stupně. V tomto případě by bylo možno využít možností většiny zmíněných speciálních škál.

Pokud bychom chtěli zachovat stávající počet stupňů, bylo by vhodné poslední stupeň „rozšířit“, čímž by byla porušena rovnoměrnost. Škálu by nebylo možno zařadit ani mezi výše uvedené typy nerovnoměrných škál, nerovnoměrnost by byla pouze jednostranná.

Jak je vidět z výše uvedených příkladů konkrétních škál, při tvorbě modelů je nutno ke každému hodnocenému znaku přistupovat individuálně a škálu konstruovat podle požadavků a cíle hodnocení. Zcela nutnou podmínkou je spolupráce expertů z dané oblasti. Jazykové hodnoty jednotlivých škálových stupňů musí být jednoznačné a musí odpovídat skutečnému významu v přirozeném jazyce. Vzdálenosti mezi jednotlivými stupni modelu musí odpovídat vzdálenostem jejich významů v přirozeném jazyce. Vytvořená jazyková fuzzy škála by pak měla být ověřena na konkrétních datech.

Výše uvedené příklady škál s nástinem možností jejich modelování pomocí jazykových fuzzy škál byly jen nástinem možného postupu a inspirací pro další podrobnou analýzu.

References

- [1] Talašová, J., *Fuzzy metody víceriteriálního hodnocení a rozhodování*, VUP, Olomouc 2003.
- [2] Berka, K., *Měření - pojmy/teorie/problémy*, Academia, Praha 1977.
- [3] Břicháček, V., *Úvod do psychologického škálování*, Psychodiagnostické a didaktické testy, n.p., Bratislava 1978.
- [4] Hroník, F., *Hodnocení pracovníků*, Grada Publishing, a.s., 2006.
- [5] Talašová, J., Pavlačka, O., *Fuzzy Probability Spaces and Their Applications in Decision Making*.