Najlepší empirický lineárny prediktor a odhad jeho strednej kvadratickej chyby

Júlia Volaufová
Ústav merania SAV, Dúbravská 9
842 19 Bratislava

Abstrakt

V príspevku sa venujeme špeciálnemu typu mnohorozmerného lineárneho modelu s náhodným efektom. Odvodený je najlepší lineárny prediktor (BLUP) náhodného efektu, resp. empirická verzia BLUP, t.j. EBLUP, v ktorej neznáma kovariančná matica je nahradená optimálnym odhadom. Podobný problém bol študovaný pre jednorozmerný náhodný efekt napr. v prácach [4] a [1]. Odvodený je tiež odhad strednej kvadratickej chyby (MSE) lineárneho empirického prediktora.

1 Úvod

Uvažujme mnohorozmerný lineárny model s náhodným efektom v tvare

$$Y_i = \mu_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

kde Y_i sú p-rozmerné vektory pozorovaní, μ_i sú náhodné efekty, o ktorých predpokladáme, že sú normálne rozdelené so spoločnou strednou hodnotou μ a kovariančnou maticou Σ,

$$\mu_i \sim N_p(0, \Sigma), \quad i = 1, \ldots, n.$$

O chybových vektoroch ε_i predpokladáme, že majú nulovú strednú hodnotu a kovariančnú maticu závislú na indivíduách, t.j.:

$$\varepsilon_i \sim (0, \Sigma_i), \quad i = 1, \ldots, n.$$

Špecifikou uvažovaného modelu je predpoklad, že všetky matice Σ_i, $i = 1, \ldots, n$ sú známe (dajme tomu z predchádzajúceho experimentu) a neznáme sú vektor μ a kovariančná matica Σ.

Okrem predchádzajúceho predpokladáme, že náhodné efekty a chybové vektory sú navzájom nezávislé.

Takto uvedenú situáciu môžeme vyjadriť v tvare kombinovaného modelu, čo je možné uviesť nasledovne

$$\underline{Y} = (\underline{1} \otimes I)\mu + \underline{\xi} + \underline{\varepsilon}. \tag{2}$$

Symbol "$\otimes$" označuje Kroneckerov súčin matíc, resp. vektorov. Vektor $\underline{1} = (1, \ldots, 1)'$ je n-rozmerný vektor jednotiek a I je $p \times p$-rozmerná jednotková matica. Keďže uvažujeme

všetky pozorovania súčasne, vektor $\underline{Y}$ je np-rozmerný, vytvorený zo všetkých subvektorov Y_i, t.j. $\underline{Y} = (Y_1', \ldots, Y_n')'$. Podobne je vytvorený aj chybový vektor $\underline{\varepsilon}$. V tomto modeli vystupuje vektor $\underline{\xi}$, ktorý vznikol z náhodných efektov jednoduchým centrovaním, t.j. $\underline{\xi} = (\xi_1', \ldots, \xi_n')'$, kde $\xi_i = \mu_i - \mu$. Z toho vyplýva $\underline{\xi} \sim N_{np}(0, I \otimes \Sigma)$.

Kovariančná matica celého vektora $\underline{Y}$ má potom tvar

$$\text{cov}\,(\underline{Y}) = \Gamma = I \otimes \Sigma + \text{Diag}\,(\Sigma_i),$$

kde Diag (Σ_i) označuje blokovo diagonálnu maticu tvaru

$$\text{Diag}\,(\Sigma_i) = \begin{pmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ldots & \vdots \\ 0 & \cdots & \Sigma_n \end{pmatrix}.$$

Takto predstavený model v tvare (2) je prípad zmiešaného lineárneho modelu, kde neznámy vektor μ je parametrom fixného efektu, vektor $\underline{\xi}$ predstavuje náhodný efekt a $\underline{\varepsilon}$ je chybový vektor. Hlavným cieľom v tomto modeli, resp. v modeli (1) býva odhad (optimálny) vektora μ, s čím súvisí odhad kovariančnej matice Σ a veľmi často aj odhad náhodných efektov μ_i, $i = 1, \ldots, n$. V nasledujúcom paragrafe sa venujeme odhadu parametrov.

2 Odhad parametrov

Zmiešaný lineárny model v tvare (2), resp. v tvare (1) je v literatúre veľmi často študovaný a je viac možných prístupov ako odhadovať vektor μ a neznámu maticu Σ. Pre úplnosť uvedieme niektoré z nich.

2.1 Metóda maximálnej vierohodnosti (ML)

Predpokladajme, že náhodné efekty ξ_i, $i = 1, \ldots, n$ majú p-rozmerné normálne rozdelenie a tak isto aj chybové vektory ε_i, t.j. $\xi_i \sim N_p(0, \Sigma)$ a $\varepsilon_i \sim N_p(0, \Sigma_i)$, $i = 1, \ldots, n$. Potom logaritmicko-vierohodnostná funkcia vyzerá nasledovne:

$$l(\mu, \Sigma) = -\frac{np}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{n} \ln |\Sigma + \Sigma_i| - \frac{1}{2} (\underline{Y} - (\underline{1} \otimes I)\mu)' \,\text{Diag}\,(\Sigma + \Sigma_i)^{-1} (\underline{Y} - (\underline{1} \otimes I)\mu),$$

čo po zderivovaní podľa μ a Σ vedie k systému vierohodnostných rovníc:

$$\begin{aligned} \sum_{i=1}^{n} (\Sigma + \Sigma_i)^{-1} \mu &= \sum_{i=1}^{n} (\Sigma + \Sigma_i)^{-1} Y_i \\ \sum_{k=1}^{n} e_i' (\Sigma + \Sigma_k)^{-1} e_j &= \sum_{k=1}^{n} e_i' (\Sigma + \Sigma_k)^{-1} (Y_k - \mu)(Y_k - \mu)' (\Sigma + \Sigma_k)^{-1} e_j, \quad i \le j. \end{aligned}$$

Z numerického hľadiska je tento systém veľmi ťažko riešiteľný vzhľadom na μ a Σ a autorovi nie sú známe efektívne algoritmy vedúce k riešeniu.

Niekedy, keď sa obmedzíme len na odhad matice Σ je vhodné použiť tzv. reštringovanú metódu maximálnej vierohodnosti (REML). Princíp metódy spočíva v tom, že sa využije normalita vektora reziduí, ktorý vznikne na základe jednoduchého odhadu metódou najmenších štvorcov (OLS) vektora μ a z neho sa odvodí MLE pre maticu Σ. Týmto prístupom sa nebudeme bližšie zaoberať. (Pozri napr. [6]).

2.2 Najlepší lineárny nevychýlený odhad – dvojetapový odhad

Model (2) je možné vyjadriť aj v tvare

$$(\underline{Y}, (\underline{1} \otimes I)\mu, \Gamma), \tag{3}$$

kde $\Gamma = (I \otimes \Sigma) + \mathrm{Diag}\,(\Sigma_i) = \mathrm{Diag}\,(\Sigma + \Sigma_i)$ je kovariančná matica vektora $\underline{Y}$. Zo všeobecnej teórie lineárnych modelov je zrejmé, že najlepší (v závislosti na Σ) lineárny nevychýlený odhad (BLUE) vektora μ (podotýkame, že v modeli (3) je μ odhadnuteľný vektorový parameter) je daný vzťahom:

$$\hat{\mu} = \left[\sum_{i=1}^{n}(\Sigma + \Sigma_i)^{-1}\right]^{-1} \sum_{i=1}^{n}(\Sigma + \Sigma_i)^{-1} Y_i.$$

Vzhľadom na to, že matica Σ je neznáma, obyčajne je nahradená odhadom. Najčastejšie ju odhadujeme z tej istej realizácie vektora $\underline{Y}$, čo v konečnom dôsledku vedie k "nelineárnemu" odhadu vektora μ, nazývaného dvojetapovým odhadom μ. Ak označíme odhad (o jeho vlastnostiach zatiaľ nehovoríme) matice Σ ako $\hat{\Sigma}$, potom

$$\tilde{\mu} = \left[\sum_{i=1}^{n}(\hat{\Sigma} + \Sigma_i)^{-1}\right]^{-1} \sum_{i=1}^{n}(\hat{\Sigma} + \Sigma_i)^{-1} Y_i. \tag{4}$$

V prácach [3], [2] a [7] je problematike dvojetapových odhadov venovaná značná pozornosť. Za dostatočne všeobecných predpokladov:

1. rozdelenie vektorov $\underline{\xi}$ a $\underline{\varepsilon}$ je symetrické okolo 0,
2. $\hat{\Sigma}$ je párnou funkciou vektora $\underline{Y}$, t.j. $\hat{\Sigma}(\underline{Y}) = \hat{\Sigma}(-\underline{Y})$,
3. $\hat{\Sigma}$ je invariantnou štatistikou vzhľadom na grupu posunutí v strednej hodnote, t.j. $\hat{\Sigma}(\underline{Y}) = \hat{\Sigma}(\underline{Y} + (\underline{1} \otimes I)\alpha)$ pre všetky $\alpha \in R^p$

platí, že odhad $\tilde{\mu}$ je nevychýlený pre μ. Je treba zdôrazniť, že $\hat{\Sigma}$ nemusí byť nevychýleným odhadom matice Σ. Aký odhad môžeme vybrať za $\hat{\Sigma}$?

V literatúre je dostatočne podrobne študovaný tzv. MINQUE - odhad, t.j. kvadratický odhad, nevychýlený, invariantný vzhľadom na posun v strednej hodnote, taký, ktorý minimalizuje vhodne zvolenú euklidovskú normu matice kvadratickej formy odhadu. (Pozri napr. [5]). Je dôležité poznamenať, že za predpokladu normality rozdelenia, MINQUE odhad je lokálne (pre vopred zvolenú maticu Σ_0) najlepší nevychýlený invariantný odhad. Práve vzhľadom na to, že MINQUE odhad závisí od približnej hodnoty, resp. vopred zvolenej matice Σ_0, je vhodné uvažovať iné metódy odhadovania. Jeden z možných prístupov je uvedený v práci [8] a venujeme mu nasledujúci paragraf.

2.3 Prirodzený odhad

Uvažujme model (2). OLS odhad vektora μ, t.j. odhad získaný jednoduchou metódou najmenších štvorcov, je daný vzťahom

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Vzhľadom na to, že predpokladáme, že matica Σ je nezáporne definitná, aj 0 patrí do parametrického priestoru, a teda je možné uvažovať aj odhad

$$\mu^{**} = \left[\sum_{i=1}^{n} \Sigma_i^{-1}\right]^{-1} \sum_{i=1}^{n} \Sigma_i^{-1} Y_i,$$

ktorý dostaneme ako BLUE v $\Gamma = \text{Diag}\,(\Sigma_i)$.

V ďalšom budeme používať odhad μ^*, je však zrejmé, že ho môžeme nahradiť všade odhadom μ^{**}.

Prvotným odhadom kovariančnej matice Γ založenom na odhade μ^* bude odhad

$$\hat{\Gamma} = (\underline{Y} - (1 \otimes I)\mu^*)(\underline{Y} - (1 \otimes I)\mu^*)'.$$

V ďalších úvahách budeme používať nasledujúce vyjadrenie matice Σ.

$$\Sigma = \sum_{i \leq j} \sigma_{ij}(1 + \delta_{ij})^{-1}(e_i e_j' + e_j e_i'),$$

kde σ_{ij} sú prvky matice Σ, δ_{ij} označuje Kroneckerovo δ a vektor e_i, resp. e_j je p-rozmerný vektor s 1 na i-tom, resp. j-tom mieste a inde s nulami. Ak označíme $V_{ij} = (1 + \delta_{ij})^{-1}(e_i e_j' + e_j e_i')$, dostaneme

$$\Sigma = \sum_{i \leq j} \sigma_{ij} V_{ij}.$$

Uvažujme ďalej triedu matíc

$$\mathcal{D} = \{I \otimes D, D = D',\ D \text{ je typu } p \times p\} = \{V(d) : \sum_{i \leq j} d_{ij}(I \otimes V_{ij})\}.$$

Zrejme platí $\mathcal{D} \subset \mathcal{S}$, kde $\mathcal{S}$ je trieda všetkých symetrických matíc typu $np \times np$. So skalárnym súčinom $< A, B >= \text{tr}\, AB$, $A, B \in \mathcal{S}$, vytvára $\mathcal{S}$ euklidovský $np(np+1)/2$-rozmerný priestor a $\mathcal{D}$ je jeho vlastným podpriestorom. Označme

$$\hat{\hat{\Gamma}} = \hat{\Gamma} - \text{Diag}\,(\Sigma_i).$$

Potom, za "prirodzený" odhad matice $I \otimes \Sigma$ je možné považovať euklidovskú projekciu matice $\hat{\hat{\Gamma}}$ na podpriestor $\mathcal{D}$. Označme túto projekciu $P_{\mathcal{D}}(\hat{\hat{\Gamma}})$. Platí: $P_{\mathcal{D}}(\hat{\hat{\Gamma}}) = \sum_{i \leq j} s_{ij}(I \otimes V_{ij})$, kde $s = (s_{11}, s_{12}, \ldots, s_{pp})'$ je $p(p+1)/2$-rozmerný vektor, ktorý rieši systém

$$Gs = \lambda,$$

pričom G je Gramova maticas prvkami danými vzťahom $G_{ij,kl} = n\text{tr}\, V_{ij} V_{kl}$, čo v našom prípade po presnejšom odvodení vedie k tvaru

$$G_{ij,kl} = \begin{cases} 0 & i \neq j \neq k \neq l \\ & i \neq k, i = j, k = l \\ n & i = j = k = l \\ 2n & \text{inak} \end{cases}.$$

$p(p+1)/2$-rozmerný vektor pravej strany λ so súradnicami λ_{ij} $i \leq j$ je daný vzťahom $\lambda_{ij} = \operatorname{tr} \hat{\Gamma}(I \otimes V_{ij})$, čo vedie ku vzťahom

$$\lambda_{ij} = \begin{cases} \sum_{k=1}^{n} \left((Y_{ki} - \mu_i^*)^2 - \{\Sigma_k\}_{ii}\right) & i = j \\ 2\sum_{k=1}^{n} \left((Y_{ki} - \mu_i^*)(Y_{kj} - \mu_j^*) - \{\Sigma\}_{ij}\right) & i \neq j \end{cases}.$$

Súradnice vektora s generujú odhad matice Σ, označený $\hat{\Sigma}$, ktorý nazývame *prirodzený odhad* (NE). Platí

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}^{n} \left((Y_k - \mu^*)(Y_k - \mu^*)' - \Sigma_k\right).$$

Jednoduchým postupom sa dajú ukázať nasledujúce vlastnosti odhadu $\hat{\Sigma}$

1. $E(\hat{\Sigma}) = \frac{n-1}{n}\Sigma - \frac{1}{n^2}\sum_{i=1}^{n} \Sigma_i$

2. $\hat{\Sigma}$ je kvadratický odhad, invariantný vzhľadom na posun v strednej hodnote.

Poznamenávame, že prirodzený odhad $\hat{\Sigma}$ nemusí byť vždy nezáporne definitnou maticou. Aby sme sa tomu vyhli, postupujeme nasledovne.

Maticu $\hat{\Sigma}$ je možné rozložiť do tvaru $\hat{\Sigma} = PQP' = \sum_{i=1}^{m} \kappa_i P_i P_i'$, kde koeficienty κ_i sú nenulové vlastné čísla a P_i k nim prislúchajúce vlastné vektory matice $\hat{\Sigma}$. Vytvorme $\hat{\Sigma}_+$, resp. $\hat{\Sigma}_-$ tak, že v rozklade matíc $\hat{\Sigma}_+$, resp. $\hat{\Sigma}_-$ uvažujeme kladné, resp. záporné vlastné čísla κ_i v absolútnej hodnote. Potom platí $\hat{\Sigma} = \hat{\Sigma}_+ - \hat{\Sigma}_-$. Je zrejmé, že pre euklidovskú normu definovanú vzťahom $\|A - B\|^2 = \operatorname{tr}(A-B)^2$ platí, $\|\hat{\Sigma} - \Sigma\| \geq \|\hat{\Sigma}_+ - \Sigma\|$. Za prirodzený nezáporne definitný odhad matice Σ považujeme potom maticu $\hat{\Sigma}_+$. Zároveň však je zrejmé, že budú porušené vlastnosti 1. a 2.. Podrobnejšie pozri [8].

Ako sme uviedli, dvojetapový odhad (4) je za dostatočne všeobecných predpokladov nevychýleným odhadom vektora μ. Môžme však prirodzene postupovať aj ďalej, a to tak, že budeme postupovať iteračne pri odvodení dvojetapového a zároveň prirodzeného odhadu. Postup len naznačíme.

Po odvodení prirodzeného odhadu kovariančnej matice $\hat{\Sigma}$, namiesto OLS odhadu vektora μ budeme uvažovať odhad

$$\mu_{(1)}^* = \left[\sum_{k=1}^{n} (\hat{\Sigma} + \Sigma_k)^{-1}\right]^{-1} \sum_{k=1}^{n} (\hat{\Sigma} + \Sigma_k)^{-1} Y_k.$$

Pomocou $\mu_{(1)}^*$ vyjadríme odhad celej kovariančnej matice Γ, t.j. dostaneme $\hat{\Gamma}_{(1)}$, a postupujeme uvedeným postupom ďalej. Po k krokoch dostávame vzťahy

$$\mu_{(k+1)}^* = \left[\sum_{l=1}^{n} (\hat{\Sigma}_{(k)} + \Sigma_l)^{-1}\right]^{-1} \sum_{l=1}^{n} (\hat{\Sigma}_{(k)} + \Sigma_l)^{-1} Y_l$$

$$\hat{\Sigma}_{(k)} = \frac{1}{n}\sum_{l=1}^{n} \left[(Y_l - \mu_{(k)}^*)(Y_l - \mu_{(k)}^*)' - \Sigma_l\right].$$

Dá sa ukázať, že za tých istých predpokladov, ktoré platia pre dvojetapový odhad, bude aj $\mu^*_{(k)}$ nevychýleným odhadom vektora μ, pričom je zrejmé, že $\hat{\Sigma}_{(k)}$ je párna štatistika v $\underline{Y}$ a invariantná vzhľadom na posun v strednej hodnote.

Treba podotknúť, že podmienky, či už nutné alebo postačujúce, za ktorých popísaný iteračný proces konverguje nie sú zatiaľ známe, v praktických situáciách však sa iterácie stabilizovali po 2–3 krokoch.

3 Najlepší lineárny prediktor (BLUP)

Doteraz sme sa zaoberali odhadmi parametrov μ a Σ. Predmetom záujmu je však veľmi často odhad (prediktor) náhodného efektu μ_i.

Za najlepší lineárny nevychýlený prediktor (BLUP) náhodného efektu μ_i, označme ho napr. $\hat{\mu}_i$, považujeme štatistiku $A\underline{Y} = \hat{\mu}_i$, kde A je hľadaná matica typu $p \times np$, spĺňajúca predpoklady:

$$\begin{aligned} E(A\underline{Y} - \mu_i) &= 0, \quad \text{pre všetky } \mu \in R^p, \Sigma p \times p, \text{ p. s. d.} \\ E(A\underline{Y} - \mu_i)(A\underline{Y} - \mu_i)' &= \min! \text{ v zmysle Löwnerovho usporiadania .} \end{aligned}$$

Po odvodení matice A a dosadení dostávame známy vzťah pre BLUP

$$\hat{\mu}_i = \hat{\mu} + \Sigma(\Sigma + \Sigma_i)^{-1}(Y_i - \hat{\mu}),$$

kde $\hat{\mu} = \left[\sum_{i=1}^{n}(\Sigma + \Sigma_i)^{-1}\right]^{-1} \sum_{i=1}^{n}(\Sigma + \Sigma_i)^{-1}Y_i$ je BLUE v Σ fixného efektu μ.

Pre maticu strednej kvadratickej chyby (MSE) platí:

$$\begin{aligned} \text{MSE}\,(\hat{\mu}_i) &= E(\hat{\mu}_i - \mu_i)(\hat{\mu}_i - \mu_i)' \\ &= \Sigma_i(\Sigma + \Sigma_i)^{-1}\left[\sum_{k=1}^{n}(\Sigma + \Sigma_k)^{-1}\right]^{-1}(\Sigma + \Sigma_i)^{-1}\Sigma_i + \Sigma(\Sigma + \Sigma_i)^{-1}\Sigma_i. \end{aligned} \quad (5)$$

Je zrejmé, že ako BLUP $\hat{\mu}_i$, tak aj MSE $(\hat{\mu}_i)$ závisí od neznámej matice Σ. V praktických situáciách sa často postupuje tak, že sa použije odhad (kvadratický, invariantný) matice Σ a dosadí sa do BLUP. Takto získaný prediktor nazývame *empirický lineárny prediktor* (ELP) a budeme ho označovať $\tilde{\mu}_i(\underline{Y})$. Dostávame teda

$$\tilde{\mu}_i(\underline{Y}) = \tilde{\mu} + \hat{\Sigma}(\hat{\Sigma} + \Sigma_i)^{-1}(Y_i - \tilde{\mu}),$$

kde $\tilde{\mu} = \left[\sum_{i=1}^{n}(\hat{\Sigma} + \Sigma_i)^{-1}\right]^{-1} \sum_{i=1}^{n}(\hat{\Sigma} + \Sigma_i)^{-1}Y_i$. Tak ako v predchádzajúcom paragrafe, dá sa ukázať, že za predpokladu symetrie rozdelenia $\underline{\xi}$ a $\underline{\varepsilon}$, a za predpokladu, že odhad $\hat{\Sigma}$ je párnou a invariantnou funkciou $\underline{Y}$, je $\tilde{\mu}_i(\underline{Y})$ nevychýleným prediktorom náhodného efektu μ_i. (Pozri napr. [2].)

3.1 MSE pre ELP

Je zrejmé, že maticu strednej kvadratickej chyby (MSE) je v prípade empirického lineárneho prediktoru veľmi zložité odvodiť, aj v prípade, že by sme predpokladali normálne rozdelenie vektorov ξ a $\underline{\varepsilon}$ vzhľadom na to, že štatistika $\tilde{\mu}_i(\underline{Y})$ v skutočnosti nie je vôbec lineárna v $\underline{Y}$. V práci [2] Kackar a Harville ukázali, že platí vzťah

$$\begin{aligned} \text{MSE}\ (\tilde{\mu}_i(\underline{Y})) &= E(\tilde{\mu}_i(\underline{Y}) - \mu_i)(\tilde{\mu}_i(\underline{Y}) - \mu_i)' = \\ &= \text{MSE}\ (\hat{\mu}_i) + E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)', \end{aligned}$$

kde $\hat{\mu}_i$ je BLUP efektu μ_i. Platí totiž nasledujúci vzťah:

$$\begin{aligned} E(\tilde{\mu}_i(\underline{Y}) - \mu_i)(\tilde{\mu}_i(\underline{Y}) - \mu_i)' &= E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i + \hat{\mu}_i - \mu_i)(\tilde{\mu}_i(\underline{Y}) - -\hat{\mu}_i + \hat{\mu}_i - \mu_i)' = \\ &= E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i) + E(\hat{\mu}_i - \mu_i)(\hat{\mu}_i - \mu_i)' + 2E\left(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i\right)(\hat{\mu}_i - \mu_i). \end{aligned}$$

Zároveň však za predpokladu symetrie rozdelenia, invariantnosti a párnosti odhadu $\hat{\Sigma}$ platí:

$$\begin{aligned} E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)(\hat{\mu}_i - \mu_i) &= E\left(E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)(\hat{\mu}_i - \mu_i)|\hat{\Sigma}\right) = \\ &= E\left((\hat{\mu}_i - \mu_i)E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)|\hat{\Sigma}\right) = 0. \end{aligned}$$

Člen MSE $(\hat{\mu}_i)$ je daný vzťahom (5). Naším cieľom bude aproximovať druhý člen, t.j. výraz $E(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i)'$.

Podobne ako v práci [4] alebo [1] budeme predpokladať jednoduchšiu situáciu, a to model pre $p = 1$, teda model

$$Y_i = \mu + \xi_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

kde Y_i sú jednorozmerné pozorované náhodné veličiny, $\mu \in R^1$ je neznáma stredná hodnota, t.j. fixný efekt, vektor $\xi = (\xi_1, \ldots, \xi_n)'$ spĺňa predpoklad $\xi \sim N_n(0, \sigma^2 I)$, a podobne pre chybový vektor $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ platí $\varepsilon \sim N_n(0, \text{Diag}\,(\sigma_i^2))$. V tomto prípade Diag (σ_i^2) je diagonálna matica typu $n \times n$ s prvkami σ_i^2, $i = 1, \ldots, n$ na diagonále.

Základnou myšlienkou je využitie aproximácie pomocou Taylorovho rozvoja. Ak zanedbáme zvyšok v Taylorovom rozvoji, platí:

$$\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i \doteq \hat{\mu}_i + \left.\frac{\partial \tilde{\mu}_i(\underline{Y})}{\partial \widehat{\sigma^2}}\right|_{\widehat{\sigma^2} = \sigma^2} (\widehat{\sigma^2} - \sigma^2) - \hat{\mu}_i = \left.\frac{\partial \tilde{\mu}_i(\underline{Y})}{\partial \widehat{\sigma^2}}\right|_{\widehat{\sigma^2} = \sigma^2} (\widehat{\sigma^2} - \sigma^2).$$

Označme $d(\sigma^2) = \left.\dfrac{\partial \tilde{\mu}_i(\underline{Y})}{\partial \widehat{\sigma^2}}\right|_{\widehat{\sigma^2} = \sigma^2}$. Kackar a Harville v [2] uviedli ďalšiu aproximáciu, na základe ktorej platí:

$$E\left(\tilde{\mu}_i(\underline{Y}) - \hat{\mu}_i\right)^2 \doteq E\left(d(\sigma^2)(\widehat{\sigma^2} - \sigma^2)\right)^2 \doteq \text{var}\ (d(\sigma^2))E(\widehat{\sigma^2} - \sigma^2)^2.$$

My využijeme ďalšiu aproximáciu odvodenú v práci [4], v ktorej je ukázané, že zanedbané členy sú rádu menšieho než $o(n^{-1})$.

Označme vektor $b = \sigma^2(\sigma^2 + \sigma_i^2)^{-1} e_i$, kde $e_i = (0, \ldots, 1, \ldots, 0)'$. Platí:

$$\text{var}\ (d(\sigma^2))E(\widehat{\sigma^2} - \sigma^2)^2 \doteq \frac{\partial b'}{\partial \sigma^2} \text{Diag}\,(\sigma^2 + \sigma_i^2) \frac{\partial b}{\partial \sigma^2} E(\widehat{\sigma^2} - \sigma^2)^2.$$

Pre maticu strednej kvadratickej chyby teda dostávame:

$$\begin{aligned}&\text{MSE}\ (\tilde{\mu}_i(\underline{Y}))\\ &\doteq \text{MSE}\ (\hat{\mu}_i) + \sigma_i^2(\sigma^2+\sigma_i^2)^{-3}E(\widehat{\sigma^2}-\sigma^2)^2\\ &= \sigma^2\sigma_i^2(\sigma^2+\sigma_i^2)^{-1} + \sigma_i^2(\sigma^2+\sigma_i^2)^{-2}\left[\sum_{k=1}^{n}(\sigma^2+\sigma_k^2)^{-1}\right]^{-1} + \sigma_i^2(\sigma^2+\sigma_i^2)^{-3}E(\widehat{\sigma^2}-\sigma^2)^2.\end{aligned}$$

Označme postupne

$$t_1(\sigma^2) = \sigma^2\sigma_i^2(\sigma^2+\sigma_i^2)^{-1}, \tag{7}$$

$$t_2(\sigma^2) = \sigma_i^2(\sigma^2+\sigma_i^2)^{-2}\left[\sum_{k=1}^{n}(\sigma^2+\sigma_k^2)^{-1}\right]^{-1} \tag{8}$$

$$t_3(\sigma^2) = \sigma_i^2(\sigma^2+\sigma_i^2)^{-3}E(\widehat{\sigma^2}-\sigma^2)^2. \tag{9}$$

Uvažujeme prirodzený odhad σ^2, čo v predchádzajúcom označení budeme značiť ako $\widehat{\sigma^2}$. Pripomenieme si, že platí:

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{k=1}^{n}\left((Y_k-\bar{Y})^2-\sigma_k^2\right),$$

kde $\bar{Y} = \mu^* = \frac{1}{n}\sum_{k=1}^{n}Y_k$. Po odvodení dostávame nasledujúce vzťahy:

$$\text{var}\ \widehat{\sigma^2} = \frac{2}{n^2}\left((n-1)\sigma^4 + 2\frac{n-1}{n}\sigma^2\sum_{i=1}^{n}\sigma_i^2 + \frac{n-2}{n}\sum_{i=1}^{n}\sigma_i^4 + \frac{1}{n^2}(\sum_{i=1}^{n}\sigma_i^2)^2\right), \tag{10}$$

a ďalej

$$\left|E(\widehat{\sigma^2}-\sigma^2)\right| = \frac{\sigma^2}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2. \tag{11}$$

Po zanedbaní členov rádu vyššieho než $O(n^{-1})$ po dosadení z (10) a (11) dostávame priblíženie:

$$E(\widehat{\sigma^2}-\sigma^2)^2 \doteq \frac{2}{n}\left(\sigma^4 + \frac{2}{n}\sigma^2\sum_{i=1}^{n}\sigma_i^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^4\right). \tag{12}$$

Z predchádzajúceho teda dostávame približný výraz pre t_3:

$$t_3 \doteq \frac{2\sigma_i^2}{n(\sigma^2+\sigma_i^2)^3}\left((\sigma^4 + \frac{2}{n}\sigma^2\sum_{k=1}^{n}\sigma_k^2 + \frac{1}{n^2}\sum_{k=1}^{n}\sigma_k^4\right). \tag{13}$$

A napokon uvedieme bez dôkazu tvrdenie, ktoré je založené na predchádzajúcich úvahách a na Taylorovom rozvoji štatistík $t_1(\widehat{\sigma^2})$, $t_2(\widehat{\sigma^2})$ a $t_3(\widehat{\sigma^2})$, ktoré vzniknú dosadením odhadu $\widehat{\sigma^2}$ za skutočnú hodnotu σ^2. Platí totiž lema (pozri [4])

Lema 1 *V zmysle predchádzajúceho označenia platia nasledujúce vzťahy:*

$$E\left(t_1(\widehat{\sigma^2})\right) = t_1(\sigma^2) - t_3(\sigma^2) + o(n^{-1}), \tag{14}$$

$$E\left(t_3(\widehat{\sigma^2})\right) = t_3(\sigma^2) + o(n^{-1}), \tag{15}$$

$$E\left(t_3(\widehat{\sigma^2})\right) = t_3(\sigma^2) + o(n^{-1}). \tag{16}$$

Tvrdenie 1 *Odhad strednej kvadratickej chyby v modeli (6) empirického lineárneho prediktoru náhodného efektu μ_i je daný vzťahom*

$$\text{MSE}\,(\hat{\mu}_i) = t_1(\widehat{\sigma}^2) + t_2(\widehat{\sigma}^2) + 2t_3(\widehat{\sigma}^2),$$

kde vzťahy $t_i(\widehat{\sigma}^2)$ pre $i = 1, 2, 3$ sú dané vzťahmi (7), (8) a (13).

Pre prípad, keď neznámy parameter σ^2 je odhadnutý Hendersonovým odhadom, odkazujeme čitateľa na prácu [4]. Ukazuje sa, že výsledky sú totožné s výsledkami uvedenými v spomenutej práci, vzhľadom na asymptotické vlastnosti prirodzeného odhadu.

Literatúra

[1] Kleffe J. and Rao J.N.K. Estimation of mean square error of empirical best linear unbiased predictors under a random variance linear model. *Journal of Multivariate Analysis*, 43:1–15, 1992.

[2] R.N. Kackar and D.A. Harville. Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association*, 79:853–862, 1984.

[3] C.G. Khatri and K.R. Shah. On the unbiased estimation of fixed effects in a mixed model for growth curves. *Communication in Statististics — Theory and Methods*, A10(4):401–406, 1981.

[4] N.G.N. Prasad and J.N.K. Rao. The estimation of the mean squared error of small-area estimators. *Journal of the American Statistical Association*, 85(409):163–171, 1990.

[5] C. R. Rao and J. Kleffe. *Estimation of Variance Components and Applications*, volume 3 of *Statistics and probability*. North-Holland, Amsterdam, New York, Oxford, Tokyo, first edition, 1988.

[6] S. R. Searle, G. Casella, and Ch. E. McCulloch. *Variance Components*. Wiley series in probability and mathematical statistics. John Wiley & Sons, Inc., New York, Chichester, Brisbane, Toronto, Singapore, first edition, 1992.

[7] J. Volaufová. On the variance of two-stage estimator in variance-covariance components model. *Applications of Mathematics*, 38(1):1–9, 1993.

[8] J. Volaufová and Komorník, J. Weighted multivariate regression estimates solved by random effects approach. *Biometrical Journal*, 1992. Accepted for publication.