



Približné konfidenčné intervaly pre variančný komponent vo všeobecnom prípade modelu s dvomi komponentami

Barbora Arendacká

barendacka@gmail.com

Ústav merania, Slovenská akadémia vied, Dúbravská cesta 9, 841 04 Bratislava

Poster je venovaný porovnaniu vlastností troch približných konfidenčných intervalov pre variančný komponent zodpovedajúci náhodnému faktoru vo všeobecnom prípade zmiešaného lineárneho modelu s dvomi variančnými komponentami s jediným obmedzením, že je k dispozícií viac ako iba minimálny počet pozorovaní. Porovnáme interval navrhnutý Parkom a Burdickom (PB) [1], Hartungom a Knappom (HK) [2] a Thomasom a Hultquistom (TH) [3]. Posledne menovaný bol pôvodne odvodený pre nevyvážený model jednoduchého triedenia, ale, ako naznačíme, takmer bez zmeny je ho možné použiť aj vo všeobecnej situácii.

MODEL

Predpokladáme $y \sim N_n(X\beta, \sigma_1^2 ZZ^T + \sigma^2 I_n)$, kde X, Z sú známe matice, $\beta, (\sigma_1^2, \sigma^2)^T$ sú vektory neznámych parametrov ($\sigma_1^2 \geq 0, \sigma^2 > 0$), $\mathcal{R}(Z) \not\subseteq \mathcal{R}(X)$, kde $\mathcal{R}(A)$ označuje priestor generovaný stĺpcami matice A . Využitím invariance parametra σ_1^2 vzhľadom na posunutie v strednej hodnote, konštrukciu konfidenčných intervalov založíme na minimálnych postačujúcich štatistikách pre maximálny invariant $B^T y$:

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

Predpoklad viac ako minimálneho počtu pozorovaní zodpovedá predpokladu: $\lambda_r = 0$, t.j. $U_r \sim \sigma^2 \chi_{\nu_r}^2$.

$$|BB^T| = M_X = I_n - X(X^T X)^{-1} X^T \quad \text{a} \quad B^T B = I_{n - \text{rank}(X)},$$

$\lambda_1 > \lambda_2 > \dots > \lambda_r \geq 0$, sú vlastné čísla $B^T Z Z^T B$, ν_i sú ich násobnosti a E_i sú k nim prislúchajúce matice zo spektrálneho rozkladu $B^T Z Z^T B$, $E_i E_j = 0, i \neq j, \sum E_i = I$.

$F_{n, m; \alpha}$ a $\chi_{n; \alpha}^2$ označujú α kvantily príslušných rozdelení.

PB INTERVAL

Založený na $S_M^2 = y^T F^T (F Z Z^T F^T)^+ F y / s$ a $S_E^2 = y^T (I - P_{X^*}) y / p$, kde $X^* = [X, Z Z^T]$, $F = M_X - M_{X^*}$, $s = \text{rank}(X^*) - \text{rank}(X)$ a $p = n - \text{rank}(X^*)$, a aproximáciou $s S_M^2 / E(S_M^2) \sim \chi_s^2$. Táto aproximácia je presná, ak $\lambda_i \rightarrow \lambda, i = 1, \dots, r-1$ alebo $\sigma_1^2 / \sigma^2 \rightarrow \infty$ (pozri [1]). Dá sa ukázať, že $s S_M^2 = \sum_{i=1}^{r-1} U_i / \lambda_i$ a $p S_E^2 = U_r$, $s = \sum_{i=1}^{r-1} \nu_i, p = \nu_r$.

Tvar $(1 - 2\alpha)100\%$ -ného PB intervalu, $[L_{PB}, U_{PB}]$:

$$L_{PB} = S_M^2 - S_E^2 / h - (G_1^2 S_M^4 + H_2^2 S_E^4 / h^2 + G_{12} S_M^2 S_E^2 / h)^{1/2}$$

$$U_{PB} = S_M^2 - S_E^2 / h + (H_1^2 S_M^4 + G_2^2 S_E^4 / h^2 + H_{12} S_M^2 S_E^2 / h)^{1/2}$$

kde $h = \sum_{i=1}^{r-1} \nu_i (\sum_{i=1}^{r-1} \nu_i / \lambda_i)^{-1}$, $G_1 = 1 - 1 / F_{s, \infty; 1-\alpha}$, $G_2 = 1 - 1 / F_{\nu_r, \infty; 1-\alpha}$, $H_1 = 1 / F_{s, \infty; \alpha} - 1$, $H_2 = 1 / F_{\nu_r, \infty; \alpha} - 1$, $G_{12} = [(F_{s, \nu_r; 1-\alpha} - 1)^2 - G_1^2 F_{s, \nu_r; 1-\alpha}^2 - H_2^2] / F_{s, \nu_r; 1-\alpha}$, $H_{12} = [(1 - F_{s, \nu_r; \alpha})^2 - H_1^2 F_{s, \nu_r; \alpha}^2 - G_2^2] / F_{s, \nu_r; \alpha}$. Záporné hranice kladieme rovné 0.

+: presný, ak $\sigma_1^2 / \sigma^2 \rightarrow \infty$ (pre bežne používané $\alpha \leq 0.1$)

-: na zlepšenie vlastností nevyplýva zvyšovanie ν_r (ak $r > 2$)

nulu neobsahuje v zhode s testom nulovosti σ_1^2 (ak $r > 2$)

TH INTERVAL

Pôvodne odvodený pre prípad nevyváženého modelu jednoduchého triedenia,

$y_{ij} = \mu + \alpha_i + \epsilon_{ij}, i = 1, \dots, a, j = 1, \dots, n_i$. Je založený na $U_r / \nu_r = S_E^2$ (stredná suma štvorcov rozdielov vo vnútri súborov) a $S_y^2 = \frac{1}{a-1} \sum_{i=1}^a (\bar{y}_i - \frac{1}{a} \sum_{i=1}^a \bar{y}_i)^2$, výberovej variancie priemerov v jednotlivých súboroch a aproximáciou $(a-1) S_y^2 / (\sigma_1^2 + \tilde{n} \sigma^2) \sim \chi_{a-1}^2$, $\tilde{n} = (1/a) \sum_{i=1}^a (1/n_i)$. Aproximácia je presná ak všetky $n_i \rightarrow k$ alebo $\sigma_1^2 / \sigma^2 \rightarrow \infty$. (Ak $n_i \rightarrow k$, tak $\lambda_i \rightarrow \lambda$.) Keďže $(a-1) S_y^2 = y^T C y$, kde tvar C je známy a $s S_M^2 = \sum_{i=1}^{r-1} U_i / \lambda_i = y^T (M Z Z^T M)^+ y$, priamym overením vlastností Moore-Penroseovej inverzie sa dá ukázať, že $(M_X Z Z^T M_X)^+ = C$, a tiež sa dá ukázať, že $1/h = \tilde{n}$. Teda $(a-1) S_y^2 / (\sigma_1^2 + \tilde{n} \sigma^2) = s S_M^2 / E(S_M^2)$ a TH a PB interval sú založené na rovnakých sumách štvorcov a rovnakej aproximácii.

Tvar $(1 - 2\alpha)100\%$ -ného TH intervalu, $[L_{TH}, U_{TH}]$:

$$L_{TH} = s(S_M^2 - \frac{1}{h} S_E^2 F_{s, \nu_r; 1-\alpha}) / \chi_{s; 1-\alpha}^2 \quad U_{TH} = s(S_M^2 - \frac{1}{h} S_E^2 F_{s, \nu_r; \alpha}) / \chi_{s; \alpha}^2$$

Záporné hranice kladieme rovné 0.

+: presný, ak $\sigma_1^2 / \sigma^2 \rightarrow \infty$

-: na zlepšenie vlastností nevyplýva zvyšovanie ν_r (ak $r > 2$)

nulu neobsahuje v zhode s testom nulovosti σ_1^2 (ak $r > 2$)

HK INTERVAL

Ide o presný interval pre σ_1^2 / σ^2 odvodený Waldom [4] vynásobený nevychýleným odhadom pre σ^2 : U_r / ν_r . Tvar $(1 - 2\alpha)100\%$ -ného HK intervalu, $[L_{HK}, U_{HK}]$:

$$L_{HK} = l U_r / \nu_r, \quad U_{HK} = u U_r / \nu_r,$$

kde l, u sú nezáporné riešenia (alebo nuly, ak nezáporné riešenia neexistujú) nasledujúcich rovníc:

$$\sum_{i=1}^{r-1} U_i / (\lambda_i l + 1) = s F_{s, \nu_r; 1-\alpha} U_r / \nu_r \quad \sum_{i=1}^{r-1} U_i / (\lambda_i u + 1) = s F_{s, \nu_r; \alpha} U_r / \nu_r$$

Konkrétne riešenia môžeme získať napr. Newtonovou-Raphsonovou metódou.

+: presný, ak $\nu_r \rightarrow \infty$

nulu obsahuje v zhode s Waldovým testom o nulovosti σ_1^2 (na hl. významnosti α)

-: limitná pravdepodobnosť pokrytia, keď $\sigma_1^2 / \sigma^2 \rightarrow \infty$, $P_{2\alpha, s, \nu_r}$, nie je presne $1 - 2\alpha$. $P_{2\alpha, s, \nu_r} = P(s F_{s, \nu_r; \alpha} \leq \chi_s^2 \leq s F_{s, \nu_r; 1-\alpha})$.

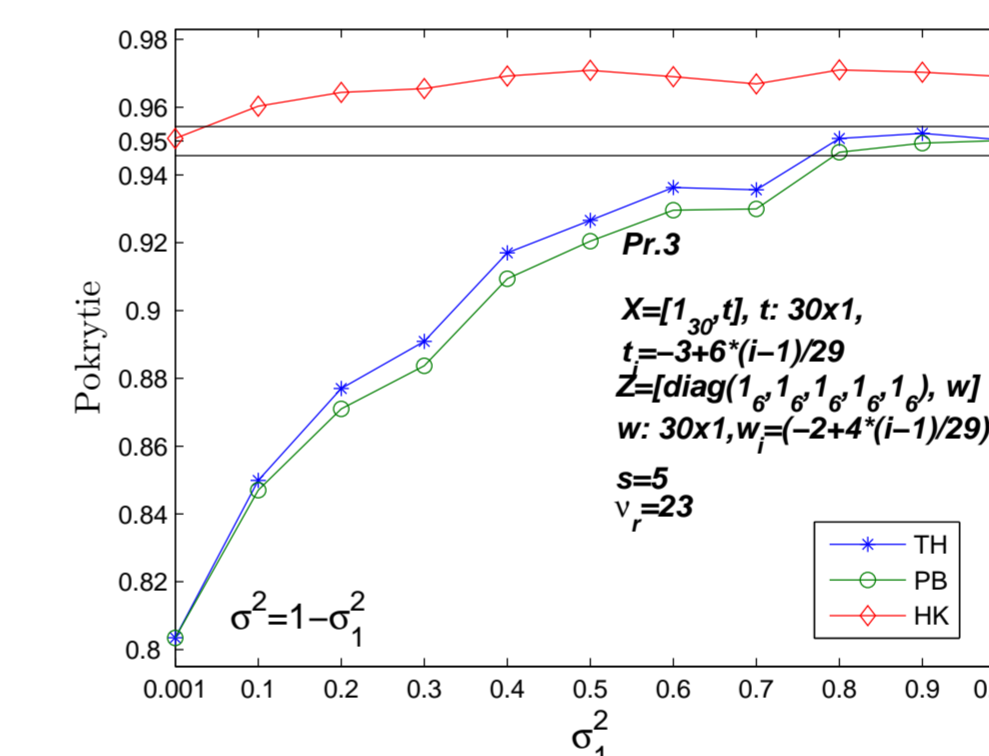
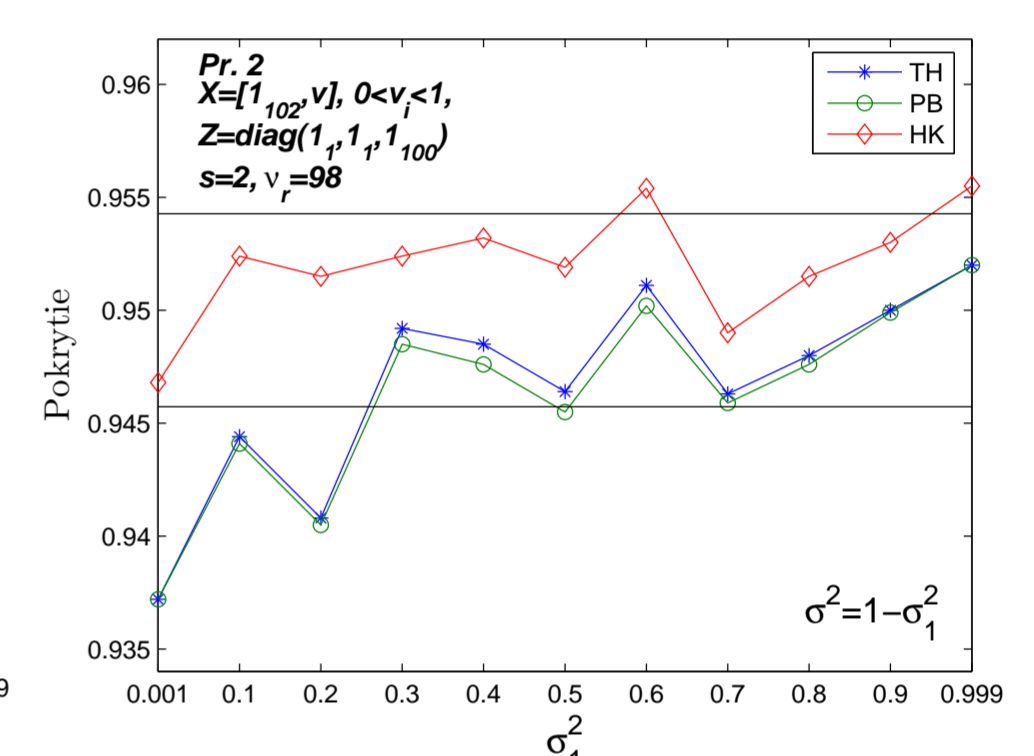
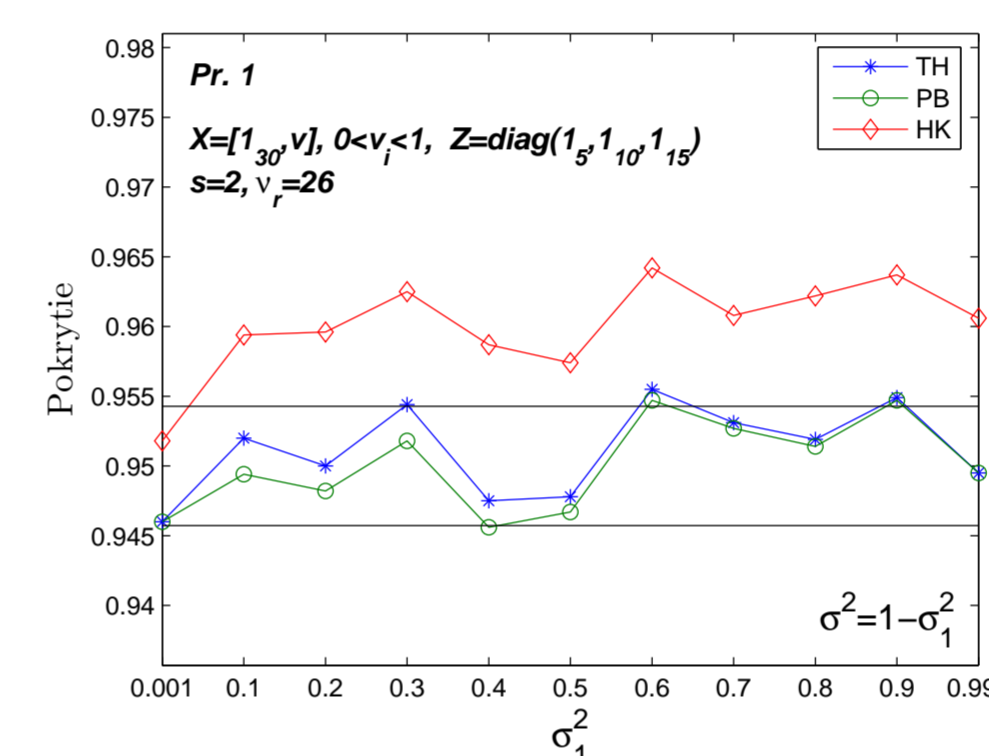
Rozdiely medzi $P_{2\alpha, s, \nu_r}$ a $1 - 2\alpha$ pre niektoré konfigurácie s, ν_r sú v nasledujúcej tabuľke. Prirodzene, tieto rozdiely sú menšie, ak je s podstatne menšie ako ν_r .

To, že je HK interval má tendenciu byť konzervatívny pre veľké σ_1^2 / σ^2 , konštatovali aj Hartung a Knapp [2] na základe výsledkov ich simulačnej štúdie.

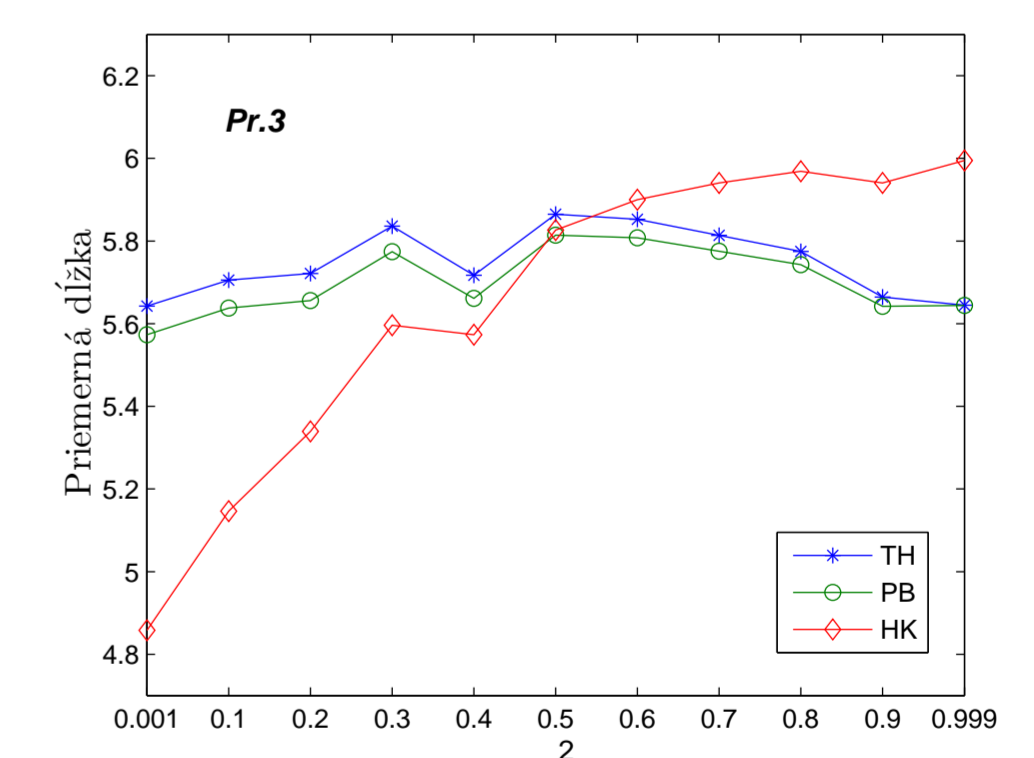
| | $P_{2\alpha, s, \nu_r} - (1 - 2\alpha)$ | | | | |
|------------------|---|--------------|--------------|--------------|--------------|
| | $s = 2$ | $s = 2$ | $s = 10$ | $s = 10$ | $s = 30$ |
| | $\nu_r = 4$ | $\nu_r = 25$ | $\nu_r = 25$ | $\nu_r = 50$ | $\nu_r = 50$ |
| $2\alpha = 0.1$ | 0.0484 | 0.0160 | 0.0479 | 0.0294 | 0.0614 |
| $2\alpha = 0.05$ | 0.0248 | 0.0113 | 0.0283 | 0.0187 | 0.0355 |
| $2\alpha = 0.01$ | 0.0050 | 0.0036 | 0.0067 | 0.0051 | 0.0083 |

ILUSTRÁCIA NA PRÍKLADOCH

Pre konkrétnu voľbu X, Z , pevne zvolený vektor β a uvedené hodnoty σ_1^2, σ^2 bola na základe 10 000 simulácií spočítaná pravdepodobnosť pokrytia a priemerná dĺžka 95%-ných PB, TH a HK intervalov. Výsledky sú v nasledujúcich grafoch. V tých, ktoré zobrazujú pokrytia sú tiež naznačené hranice, v ktorých by sa na 95% mali nachádzať nasimulované hodnoty, ak skutočná pravdepodobnosť pokrytia je 0.95.



Pr.1 zastupuje situáciu s nie príliš odlišnými $1/\lambda_i$, ale nízkym ν_r , v Pr. 2 je $\nu_r \gg s$, ale $1/\lambda_i$ sú pomerne odlišné a v Pr. 3 je skombinované nízke ν_r a navzájom dost odlišné $1/\lambda_i$. (Optimálnosť intervalov TH a PB pre $\sigma_1^2 = 0$ narastá čím menej sa $1/\lambda_i$ líšia, $1/\lambda_i = 1/\lambda \Rightarrow r = 2$.) Priemerné dĺžky intervalov v Pr.1 a Pr.2 (neuvadené) sa navzájom veľmi nelíšia (max. rozdiel v Pr.2 je 0.082, v Pr.1 iba 0.015). Priemerné dĺžky pre Pr.3 uvádza nasledujúci graf.



ZÁVERY

- simulačne sa potvrdil známy fakt, že PB a TH intervaly v niektorých modeloch nedosahujú požadovanú konfidenčnú úroveň pre malé hodnoty σ_1^2 / σ^2

- v modeloch s nízkym ν_r a výrazne rôznymi $1/\lambda_i$, intervaly TH, PB nemusia dosahovať požadovanú konfidenčnú úroveň ani pre pomerne veľké hodnoty σ_1^2 / σ^2

- vo všetkých príkladoch TH interval dosahoval rovnakú alebo vyššiu konfidenčnú úroveň ako PB interval

- HK interval dosahoval konfidenčnú úroveň aspoň takú, aká bola požadovaná, jeho konzervatívnosť sa pre $\nu_r \gg s$ stráca

- v modeloch s ν_r nie oveľa väčším ako s a výrazne odlišnými $1/\lambda_i$, bola priemerná dĺžka HK intervalu kratšia pre nižšie hodnoty σ_1^2 / σ^2 , pre vyššie hodnoty tohto podielu však boli v priemere kratšie TH a PB intervaly

Z uvedeného vyplýva, že TH a PB intervaly je vhodné použiť len ak hodnoty $1/\lambda_i$ v modeli nie sú príliš odlišné a vzhľadom na vyššiu dosahovanú pravdepodobnosť pokrytia je potom vhodné uprednostniť TH interval pred PB intervalom. Ak sú však hodnoty $1/\lambda_i$ výrazne rôzne, je aj napriek konzervatívnosti (pre malé ν_r) vhodnejší HK interval.

PodĎakovanie. Práca bola podporená grantom VEGA 1/3016/06 a VEGA 2/4026/04.

Referencie.

- [1] Park, D.J., Burdick, R.K. (2003) *Performance of confidence intervals in regression models with unbalanced one-fold nested error structures*. COMMUNICATIONS IN STATISTICS Simulation and Computation, 32, 717-732
- [2] 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
- [3] 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
- [4] Wald, A. (1947) *A note on regression analysis*. The Annals of Mathematical Statistics, 18, 586-589