

Cvičení č. 3, od 27.3.2023:

Datový soubor `toenail.txt` (hodnoty oddělené mezerami) pochází z longitudinální dermatologické klinické studie, jejímž hlavním cílem bylo porovnat účinnost dvou ošetření na potlačení infekce nehtů na nohou. Proměnné mají následující význam:

idnr identifikační číslo pacienta;

infect indikátor síly infekce (0 = bez infekce nebo slabá infekce, 1 = střední nebo vážná infekce);

trt indikátor ošetření (0 = ošetření A, 1 = ošetření B);

time čas návštěvy (měsíce);

visit číslo návštěvy.

Jako $Y_{i,j}$ označme náhodnou veličinu reprezentující indikátor síly infekce u *itého* pacienta při *jté* návštěvě ($i = 1, \dots, n$, $j = 1, \dots, n_i$), která proběhla v čase $t_{i,j}$ měsíců. Hodnota $x_i \in \{0, 1\}$ nechť odpovídá indikátoru ošetření, které bylo použito u *itého* pacienta.

Uvažujte následující (hierarchický) model („čisté“ parametry ani regresory nejsou uváděny v podmínkách při specifikaci jednotlivých rozdělání):

$$\begin{aligned} B_i &\sim \mathcal{N}(\beta_0, \tau_0^{-1}), & i = 1, \dots, n, \\ Y_{i,j} | B_i &\sim \mathcal{A}\left(\pi(B_i)\right), & i = 1, \dots, n, j = 1, \dots, n_i, \\ \log\left\{\frac{\pi(B_i)}{1 - \pi(B_i)}\right\} &= B_i + \beta_1 x_i + \beta_2 t_{i,j} + \beta_3 x_i t_{i,j}, & i = 1, \dots, n, j = 1, \dots, n_i. \end{aligned}$$

V nebayesovské terminologii se jedná o model logistické regrese s náhodným absolutním členem. Jako primární „čisté“ parametry uvažujte:

$$\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \tau_0)^\top, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\top.$$

Předpokládejte následující apriorní rozdělání pro primární parametry:

$$\begin{aligned} p(\boldsymbol{\beta}, \tau_0) &= p(\boldsymbol{\beta}) p(\tau_0), \\ \boldsymbol{\beta} &\sim \mathcal{N}_4(\mathbf{0}, \text{diag}(10^2, \dots, 10^2)), \quad \tau_0 \sim \text{Ga}(1, 0.005). \end{aligned}$$

Pro bayesovskou specifikaci modelu uvažujte jako parametry též hodnoty náhodných efektů $\mathbf{B} = (B_1, \dots, B_n)^\top$.

1. Odvoďte (stačí rukou na papír) plně podmíněné hustoty (stačí tvar hustoty známý až na multiplikativní konstantu) pro implementaci Gibbsova algoritmu, který by v jednotlivých krocích generoval (i) $\boldsymbol{\beta}$ (sdruženě), (ii) τ_0 , (iii) \mathbf{B} (sdruženě).

Dále odpovězte na následující otázky:

- Odpovídá některá z odvozených hustot některému z „pojmenovaných“ rozdělání? To jest, lze snadno určit normující konstantu?

- Liší se Gibbsův algoritmus, který v části (iii) generuje po jednom hodnoty B_1, \dots, B_n od výše uvedeného algoritmu, který generuje sdružené hodnoty \mathbf{B} ?
2. Implementujte výše uvedený model v JAGSu a vygenerujte dva markovské řetězce, jejichž limitním rozdělením bude aposteriorní rozdělení pro uvažovaný model.
 3. Nakreslete trajektorie (**traceplots**) pro primární parametry modelu a také pro devianci¹ modelu (kreslete oba řetězce do jednoho obrázku dvěma různými barvami). Nakreslete odhady autokorelačních funkcí (pro alespoň jeden z vygenerovaných řetězců).
Posuďte, zda lze předpokládat konvergenci markovského řetězce k limitnímu rozdělení a zda řetězec vykazuje přijatelnou autokorelovanost.
 4. Posuďte, zda s ohledem na variabilitu aposteriorního rozdělení pro β lze považovat použité apriorní rozdělení pro β za slabě informativní.
 5. Spočtete základní charakteristiky aposteriorního rozdělení pro následující parametry:
 - (a) $d_0 = \tau_0^{-1/2}$ (směrodatná odchylka náhodných efektů).
 - (b) γ_1 = střední směrnice logitu pravděpodobnosti střední nebo silné infekce ve skupině s ošetřením A. O jakou funkci primárních parametrů se jedná?
 - (c) γ_2 = střední směrnice logitu pravděpodobnosti střední nebo silné infekce ve skupině s ošetřením B. O jakou funkci primárních parametrů se jedná?
 - (d) γ_3 = parametr hodnotící odlišnost v účinnosti obou ošetření. O jakou funkci primárních parametrů se jedná?
 6. Pro výše definované parametry $d_0, \gamma_1, \gamma_2, \gamma_3$ spočtete 95% věrohodnostní intervaly (ET i HDP) a nakreslete odhady aposteriorních hustot.
 7. Pro parametr γ_3 spočtete (pomocí vygenerovaného markovského řetězce) hodnotu p splňující

$$p = \inf \{ \alpha : 0 \notin C(\alpha) \},$$

kde $C(\alpha)$ je $(1 - \alpha)100\%$ ET věrohodnostní interval pro γ_3 .

Uvědomte si, že spočtené p lze interpretovat jako P-hodnotu testu nulové hypotézy $\gamma_3 = 0$.

Deadline pro odevzdání vypracovaného úkolu (e-mailem na komarek [AT] karlin.mff . . .) je pondělí 10.4. ve 22:04 CEST.

¹Mezi monitorované parametry je potřeba přidat též "deviance". Dále monitorujte tyto veličiny: "pd", "popt", "dic", "ped". Jejich význam bude později vysvětlen.