Cvičení P03, od 22.11.2021:

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 $i$tého pacienta při $j$té návštěvě ($i = 1, \ldots, n$, $j = 1, \ldots, n_i$), která proběhla v čase $t_{i,j}$ měsíců. Hodnota $x_i \in \{0, 1\}$ nechť odpovídá indikátoru ošetření, které bylo použito u $i$tého pacienta.

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

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

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

$$\boldsymbol{\theta} = \big(\boldsymbol{\beta}^\top,\, \tau_0\big)^\top, \qquad \boldsymbol{\beta} = \big(\beta_0,\, \beta_1,\, \beta_2,\, \beta_3\big)^\top.$$

Předpokládejte následující apriorní rozdělení pro primární parametry:

$$p(\boldsymbol{\beta},\, \tau_0) = p(\boldsymbol{\beta})\, p(\tau_0),$$

$$\boldsymbol{\beta} \sim \mathcal{N}_4\big(\mathbf{0},\, \mathrm{diag}(10^2, \ldots, 10^2)\big), \quad \tau_0 \sim \mathrm{Ga}(1,\, 0.005).$$

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

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“ rozdělení? To jest, lze snadno určit normující konstantu?

- Liší se Gibbsův algoritmus, který v části (iii) generuje po jednom hodnoty $B_1, \ldots, B_n$ od výše uvedeného algoritmu, který generuje sdruženě hodnotu $\mathbf{B}$?

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 devianci[1] 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 $\boldsymbol{\beta}$ lze považovat použité apriorní rozdělení pro $\boldsymbol{\beta}$ za slabě informativní.

5. Spočtěte 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 $d_0, \gamma_1, \gamma_2, \gamma_3$ spočtěte 95% věrohodnostní intervaly (ET i HDP) a nakreslete odhady aposteriorních hustot.

7. Pro parametr γ_3 spočtěte (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 čtvrtek 9.12. ve 21:12 CET.

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