Cvičení č. 3, od 22.4.2024:

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:

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

Předpokládejte následující apriorní rozdělení pro primární parametry:

$$
\begin{gathered}
p(\boldsymbol{\beta},\, \tau_0) = p(\boldsymbol{\beta})\, p(\tau_0), \\
\boldsymbol{\beta} \sim \mathcal{N}_4\big(\mathbf{0},\, \mathrm{diag}(10^2, \ldots, 10^2)\big), \;\; \tau_0 \sim \mathrm{Ga}(1,\, 0.005).
\end{gathered}
$$

Jako skrytá data uvažujte 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 HPD) 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 pátek 3.5. v 15:05 CEST.

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

Exercise #3, since 22/04/2024:

The data file `toenail.txt` (values separated by spaces) comes from a longitudinal dermatological clinical study, whose main objective was to compare the efficacy of two treatments of toenail infection. It contains the following variables:

idnr identification number of a patient;

infect indicator of severity of infection (0 = no or weak infection, 1 = medium or severe infection);

trt treatment indicator (0 = treatment A, 1 = treatment B);

time time of a visit (months);

visit number of a visit.

Let $Y_{i,j}$ denote a random variable representing the indicator of the infection severity for ith patient at the jth visit ($i = 1, \ldots, n$, $j = 1, \ldots, n_i$) which was conducted at time $t_{i,j}$ of months. Let $x_{i,j} \in \{0, 1\}$ denote the treatment indicator of the ith patient.

Assume the following (hierarchical) model („genuine" parameters and regressors are not indicated in the conditions when specifying the distributions):

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

In a non-bayesian terminology this would be the logistic regression with a random intercept. As primary („genuine") parameters, consider the following:

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

Assume the following prior distribution for the primary parameters:

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

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

As latent data consider the random effects values $\mathbf{B} = \big(B_1, \ldots, B_n\big)^\top$.

1. Derive (just by hand on a paper) full conditional densities (just the core of the density known up to a multiplicative constant) to implement a Gibbs algorithm that would generate in individual steps (i) $\boldsymbol{\beta}$ (jointly), (ii) τ_0, (iii) $\mathbf{B}$ (jointly).

   Next, answer the following questions:

   - Does any of the derived densities correspond to any of „named" distributions? That is, is it easy to determine the normalizing constant?
   - Does the Gibbs algorithm differ which in part (iii) generate values of $B_1, \ldots, B_n$ one by one from the algorithm which generates jointly the complete vector $\mathbf{B}$?

2. Implement the above model in JAGS and generate two Markov chains whose limit distribution will be the posterior distribution for the model under consideration.

3. Draw the trajectories (`traceplots`) for the primary parameters of the model and also for the model deviance[2] (draw both chains in one plot with two different colors). Draw the estimates of the autocorrelation functions (for at least one of the generated chains).

   Assess whether the convergence of the Markov chain to the limit distribution can be assumed and whether the chain exhibits acceptable autocorrelation.

4. Assess whether, given the variability of the posterior distribution for $\boldsymbol{\beta}$, the prior distribution used for $\boldsymbol{\beta}$ can be considered as weakly informative.

5. Calculate the basic characteristics of the posterior distribution for the following parameters:

   (a) $d_0 = \tau_0^{-1/2}$ (standard deviation of random effects).

   (b) γ_1 = mean slope of the logit of the probability of a medium or strong infection in a group with treatment A. Which function of the primary parameters is it?

   (c) γ_2 = mean slope of the logit of the probability of a medium or strong infection in a group with treatment B. Which function of the primary parameters is it?

   (d) γ_3 = parameter which quantifies a difference between the two treatments. Which function of the primary parameters is it?

6. For above defined parameters d_0, γ_1, γ_2, γ_3, calculate 95% credible intervals (ET as well as HPD) and plot estimates of posterior densities.

7. For parameter γ_3 calculate (using the generated Markov chain) a value p which satisfies

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

   where $C(\alpha)$ is the $(1-\alpha)100\%$ ET credible interval for γ_3.

   Remember that the calculated p can be interpreted as a P-value of a test of the null hypothesis $\gamma_3 = 0$.

Deadline to deliver the report (e-mail to `komarek[AT]karlin.mff...`): Friday 3 May at 15:05 CEST.

[2] It is necessary to add `"deviance"` among the monitored parameters. Next, monitor the following variables: `"pd"`, `"popt"`, `"dic"`, `"ped"`. Their meaning will be explained later.