UNIVERSITY OF TECHNOLOGY SYDNEY

School of Mathematical and Physical Sciences

37457 Advanced Bayesian Methods

ASSIGNMENT 6

Due time and date: 3:05pm, Friday 7th October, 2022.

Submission method: Hand to Professor Wand at start of Week 9 class.

NOTE: For the benefit of participants requiring assistance with this assignment, a help session will be held at 4pm-5pm on Thursday 6th October 2022 in Room 006, Level 6, Building 7.

This assignment requires that the the R package `HRW` to part of your R environment. If this has not already been achieved, the the command `install.packages("HRW")` is required.

1. This question is concerned with the issue that practical Bayesian analyses with prior specifications such as $\beta_j \overset{\text{ind.}}{\sim} N(0, 10^{10})$ need to be done in such a way that the choice of units (e.g. millimetres versus kilometres for length) does not affect the results.

   Consider a regression-type data set $(x_i^{\text{orig}}, y_i^{\text{orig}})$, $1 \leq i \leq n$, where the superscripts indicate that these are the data in their original form before any transformation takes place. Now define:

   $$x_i \equiv \frac{x_i^{\text{orig}} - \overline{x}^{\,\text{orig}}}{s_x^{\text{orig}}} \quad \text{and} \quad y_i \equiv \frac{y_i^{\text{orig}} - \overline{y}^{\,\text{orig}}}{s_y^{\text{orig}}}, \quad 1 \leq i \leq n, \tag{1}$$

   where $\overline{x}^{\,\text{orig}}$ and s_x^{orig} are the mean and standard deviation of the x_i^{orig} data and $\overline{y}^{\,\text{orig}}$ and s_y^{orig} are the mean and standard deviation of the y_i^{orig} data.

   (a) Suppose that the x variable is temperature. Prove that the x_i data are the same regardless of whether the x_i^{orig} are recorded in degrees Celsius or degrees Fahrenheit.

   Hint: Let $x_i^{\text{orig,C}}$ be the original temperature data measured in degrees Celsius and let $x_i^{\text{orig,F}}$ be the original temperature data measured in degrees Fahrenheit. Note that $x_i^{\text{orig,F}} = 1.8\, x_i^{\text{orig,C}} + 32$. Show that x_i obtained with $x_i^{\text{orig}} = x_i^{\text{orig,F}} = 1.8\, x_i^{\text{orig,C}} + 32$ is identical to that obtained with $x_i^{\text{orig}} = x_i^{\text{orig,C}}$.

   (b) Suppose that we use a Bayesian inference engine to fit the following regression model to the (x_i, y_i) data:

   $$y_i | \beta_0, \beta_1, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1\, x_i, \sigma^2),$$

   $$\beta_0, \beta_1 \overset{\text{ind.}}{\sim} N(0, 10^{10}), \quad \sigma^2 \sim \text{Inverse-Gamma}(0.01, 0.01).$$

   Let $\beta_0^{[g]}$, $\beta_1^{[g]}$ and $(\sigma^2)^{[g]}$, $1 \leq g \leq K$, be the kept samples from the respective posterior distributions (obtained using a Markov chain Monte Carlo scheme). For

interpretation reasons it is common to transform these samples to correspond to the original units of the data as follows:

$$\left(\beta_1^{\text{orig}}\right)^{[g]} = (s_y^{\text{orig}}/s_x^{\text{orig}})\,\beta_1^{[g]}, \quad \left(\beta_0^{\text{orig}}\right)^{[g]} = \overline{y}^{\,\text{orig}} + s_y^{\text{orig}}\left\{\beta_0^{[g]} - \beta_1^{[g]}(\overline{x}^{\,\text{orig}}/s_x^{\text{orig}})\right\} \quad (2)$$

and

$$\left(\sigma^{2,\text{orig}}\right)^{[g]} = (s_y^{\text{orig}})^2(\sigma^2)^{[g]}.$$

i. Write down $y_i \approx \beta_0 + \beta_1\, x_i$ to indicate that the y_i approximately equal $\beta_0 + \beta_1\, x_i$ according to the model.

ii. Replace x_i and y_i by the right-hand sides of the expressions in (1) involving the original data.

iii. Perform algebraic manipulations that justify (informally given the $\approx$ approximate equality) the $\left(\beta_1^{\text{orig}}\right)^{[g]}$ and $\left(\beta_0^{\text{orig}}\right)^{[g]}$ expressions given by (2).

2. This question could either use templating from the file `ratsModel2.R` from Assignment 5 or, if that is not readily available, from the file `ratsModel1.R` on the subject web-site. If your version of `ratsModel2.R` is running correctly then it would be better to template from this file.

   (a) Copy `ratsModel2.R` (or `ratsModel1.R`, see above) to a new file named `ratsModel3.R`.

   (b) Open `ratsModel3.R` in an editor and modify the code (but note hints below) so that the model being fitted is:

$$\begin{aligned} &y_{ij}|\beta_0,\beta_1,\beta_2,u_{0i},u_{1i},\sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} N\Big((\beta_0+u_{0i})+(\beta_1+u_{1i})\,x_{ij}+\beta_2\,x_{ij}^2,\sigma_\varepsilon^2\Big), \\ &\left[\begin{array}{c} u_{0i} \\ u_{1i} \end{array}\right] \Bigg| \sigma_{u0}^2,\sigma_{u1}^2,\rho_u \overset{\text{ind.}}{\sim} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \boldsymbol{\Sigma}_u\right) \text{ where } \boldsymbol{\Sigma}_u \equiv \left[\begin{array}{cc} \sigma_{u0}^2 & \rho_u\sigma_{u0}\sigma_{u1} \\ \rho_u\sigma_{u0}\sigma_{u1} & \sigma_{u1}^2 \end{array}\right] \end{aligned} \quad (3)$$

   and suitable diffuse priors on $\beta_0, \beta_1, \beta_2, \sigma_{u0}^2, \sigma_{u1}^2$ and ρ_u. Note that the 2×2 covariance matrix $\boldsymbol{\Sigma}_u$ in (3) uses the following parameterisation:

$$\sigma_{u0}^2 = \text{Var}(u_{0i}), \quad \sigma_{10}^2 = \text{Var}(u_{1i})$$

   and

$$\rho_u = \text{correlation between } u_{0i} \text{ and } u_{1i}.$$

   Your new script should also update the code for plotting the fitted curves and obtaining a residual plot. Include the two plots in your submission.

   Hints:

   - The script `randIntAndSlpViaStan.R` on the subject web-site contains Stan code for fitting the random intercepts and slopes model, which is model (3) above but without the $\beta_2\, x_{ij}^2$ term. Use the Stan `transformed parameters` and `parameters` code from `randIntAndSlpViaStan.R` to extend from the random intercept structure to the random intercept and slope structure.
   - In `randIntAndSlpViaStan.R` inspect the code for the update of the object `fitMCMC`. Use this to update similar code in `ratsModel3.R`.
   - In `randIntAndSlpViaStan.R` inspect the code for the update of the object `fittedMCMC`. Use this to update similar code in `ratsModel3.R`.

3. Download the file from the `orthodontModel1.R` from the subject web-site. This script fits the Bayesian random intercepts and slopes model

$$\begin{aligned} y_{ij}|\beta_0,\beta_1,\beta_2,u_i,\sigma^2_\varepsilon &\overset{\text{ind.}}{\sim} N\Big((\beta_0+u_{0i})+(\beta_1+u_{1i})\,x_{ij},\sigma^2_\varepsilon\Big), \\ \left[\begin{array}{c} u_{0i} \\ u_{1i} \end{array}\right] \Bigg| \sigma^2_{u0},\sigma^2_{u1},\rho_u &\overset{\text{ind.}}{\sim} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2_{u0} & \rho_u\sigma_{u0}\sigma_{u1} \\ \rho_u\sigma_{u0}\sigma_{u1} & \sigma^2_{u1} \end{array}\right]\right). \end{aligned} \tag{4}$$

to data from a dental study involving 27 children by investigators at the University of North Carolina Dental School, U.S.A. The x_{ij} and y_{ij} data are:

$$\begin{aligned} x_{ij} &= \text{age of the } i\text{th child at the } j\text{th visit}, \\ y_{ij} &= \text{dental measurement on the } i\text{th child at the } j\text{th visit}. \end{aligned}$$

The actual dental measurement is somewhat technical: the distance between the pituitary and the pterygomaxillary fissure. As in Question 2, suitable diffuse prior distributions are placed on the model parameters.

(a) Start an R session and type `source("orthodontModel1.R")` to fit model (4).

(b) Copy `orthodontModel1.R` to a new file named `orthodontModel2.R`. The goal is this part of the question is to extend model (4) to be:

$$\begin{aligned} y_{ij}|\beta_0,\beta_1,\beta_2,u_i,\sigma^2_\varepsilon &\overset{\text{ind.}}{\sim} N\Big((\beta_0+u_{0i})+(\beta_1+u_{1i})\,x_{ij}+\beta_2\,z_i,\sigma^2_\varepsilon\Big), \\ \left[\begin{array}{c} u_{0i} \\ u_{1i} \end{array}\right] \Bigg| \sigma^2_{u0},\sigma^2_{u1},\rho_u &\overset{\text{ind.}}{\sim} N\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2_{u0} & \rho_u\sigma_{u0}\sigma_{u1} \\ \rho_u\sigma_{u0}\sigma_{u1} & \sigma^2_{u1} \end{array}\right]\right) \end{aligned} \tag{5}$$

where

$$z_i = \begin{cases} 1 & \text{if the } i\text{th child is male}, \\ 0 & \text{if the } i\text{th child is female}. \end{cases}$$

Modify `orthodontModel2.R` so that it fits model (5).

(c) Is there a statistically significant difference between the two genders in terms of the mean of the response variable for this model? Use output from the fitted Bayesian model in part (b) to justify your answer.

4. Download the data set file `bacteria.txt` and R script `bacteriaViaStan.R` from the subject web-site. This script analyses data from a drug trial in Northern Territory, Australia, on 50 children with middle ear infection. The children were randomised to the drug or a placebo, and also to receive active encouragement to comply with taking the drug. The presence of the bacteria *H. influenzae* was checked at weeks 0, 2, 4, 6 and 11. Define

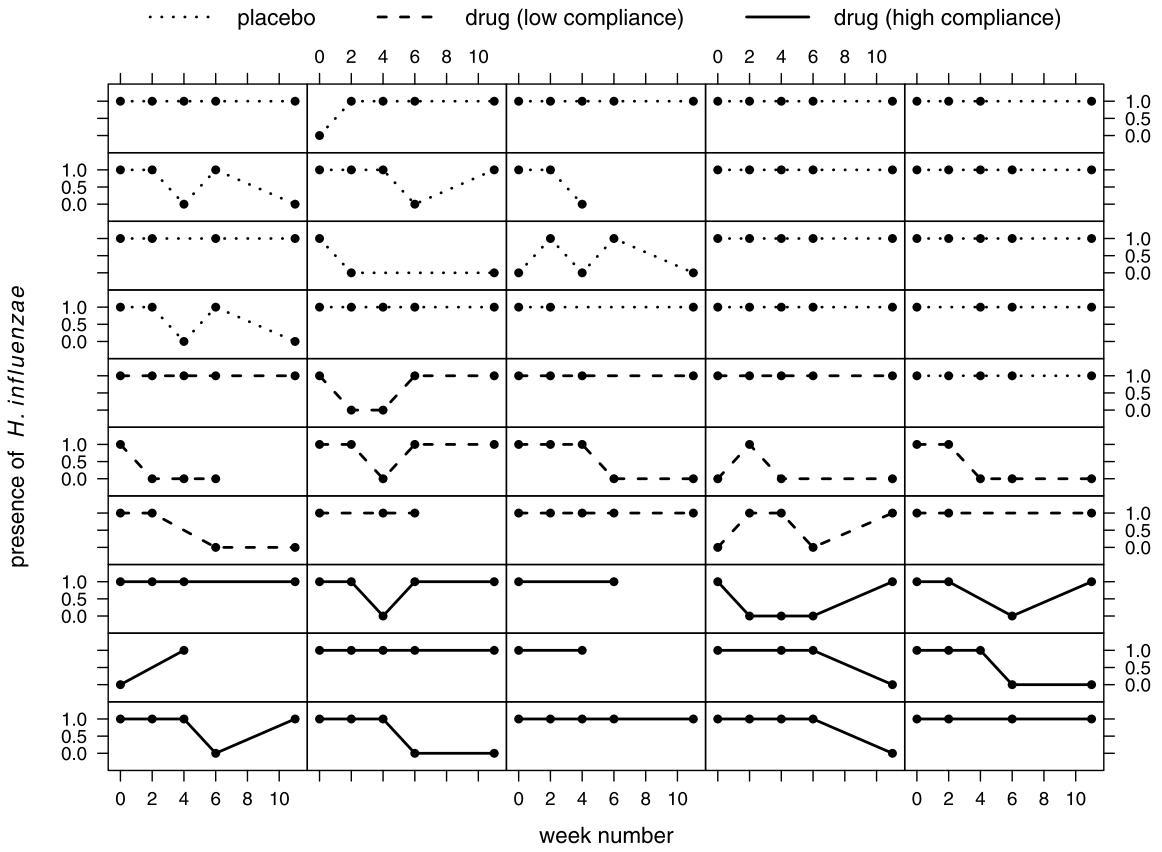
$$y_{ij} = \begin{cases} 1 & \text{if } \textit{H. influenzae} \text{ is present in the } i\text{th child at the } j\text{th check}, \\ 0 & \text{otherwise}. \end{cases}$$

Then define x_{1ij} to be the week number on which y_{ij} was measured,

$$x_{2i} = \begin{cases} 1 & \text{if the } i\text{th child was in the drug group but not actively encouraged to take} \\ & \text{the drug (“low compliance”)}, \\ 0 & \text{otherwise}. \end{cases}$$

$$x_{3i} = \begin{cases} 1 & \text{if the } i\text{th child was in the drug group and was actively encouraged to take the drug (“high compliance”),} \\ 0 & \text{otherwise.} \end{cases}$$

The following figure summarises the data:



The script `bacteriaViaStan.R` fits the following *Bayesian logistic mixed model* for these data:

$$y_{ij}|\beta_0,\beta_1,\beta_2,\beta_3 \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\frac{1}{1+\exp[-\{\beta_0+u_{i0}+(\beta_1+u_{i1})x_{1ij}+\beta_2\,x_{2i}+\beta_3\,x_{3i}\}]}\right),$$

$$\left.\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix}\right| \sigma^2_{u0},\sigma^2_{u1},\rho_u \overset{\text{ind.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \rho_u\sigma_{u0}\sigma_{u1} \\ \rho_u\sigma_{u0}\sigma_{u1} & \sigma^2_{u1} \end{bmatrix}\right)$$

and suitable diffuse priors on $\beta_0,\beta_1,\beta_2,\beta_3,\sigma^2_{u0},\sigma^2_{u1}$ and ρ_u.

(a) Start an R session and make sure that the files `bacteria.txt` and `bacteriaViaStan.R` are in the current working directory.

(b) Enter the command `source("bacteriaViaStan.R")` to run the script.

(c) What conclusion, or conclusions, can be drawn from the output regarding the effect of the the drug on prevalence of *H. influenzae* in the study population?

♦ ♦ ♦