UNIVERSITY OF TECHNOLOGY SYDNEY

School of Mathematical and Physical Sciences

37457 Advanced Bayesian Methods

ASSIGNMENT 6

Due time and date: 10:05am, Wednesday 9th October, 2024.

Submission methods:

Questions 1,3, and 4: hand to Professor Wand at the start of Class 9.

Question 2: e-mail file to `matt.wand@uts.edu.au`

NOTES:

- For the benefit of participants requiring assistance with this assignment, a help session will be held at 3pm-4pm on Tuesday 8th October 2024 in Room 006, Level 6, Building 7.
- This assignment requires that the R package `HRW` is part of your R environment. If this has not already been achieved, then 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.

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

   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:

$$\sigma_{u0}^2 = \text{Var}(u_{0i}), \quad \sigma_{u1}^2 = \text{Var}(u_{1i})$$

   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`.

What to submit for Question 2:

Please e-mail the updated `ratsModel3.R` to `matt.wand@uts.edu.au`

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.

- Start an R session and type `source("orthodontModel1.R")` to fit model (4).
- A major goal of this 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

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

- Issue the commands:

```
z <- as.numeric(as.character(Orthodont$Sex)=="Male")
print(z)
```

  The new object `z` contains the z_i data.
- Copy `orthodontModel1.R` to a new file named `orthodontModel2.R`.
- Modify `orthodontModel2.R` so that it fits model (5).
- Hints:
  - The `X` matrix should be updated to include `z`
  - In the `Stan` code the specification `matrix[numObs,2] X` should be changed to `matrix[numObs,3] X` to reflect the fact that `X` has an extra column.
  - In the `Stan` code the specification `vector[2] beta` should be changed to `vector[3] beta` since there are now three β_j parameters (i.e. β_0, β_1 and β_2).

(a) Print off the Markov chain Monte Carlo summary plot and include it in your submission.

(b) Using the summary plot from part (a), write down:

i. the Bayes estimate of β_2.

ii. a 95% credible interval for β_2.

(c) Is the gender effect on the mean dental measurement statistically significant?
Hint: The Bayesian inference approach to answering this question involves the 95% credible interval for β_2.

4. Download the data set file `GuatImmun.txt` and R script `GuatImmunViaStan.R` from the subject web-site. The data set is from a study concerned with immunisation statuses of children in Guatemala. The are grouped according to having the same mother. Define

$$y_{ij} = \begin{cases} 1 & \text{the } j\text{th child of the } i\text{th mother received a complete set of immunisations,} \\ 0 & \text{otherwise.} \end{cases}$$

Then define

$x_{1i} =$ the proportion of residents in the ith mother's community that are indigenous,

$$x_{2i} = \begin{cases} 1 & \text{if the } i\text{th mother works outside of the home,} \\ 0 & \text{otherwise,} \end{cases}$$

$$x_{3ij} = \begin{cases} 1 & \text{the } j\text{th child of the } i\text{th mother is two or older,} \\ 0 & \text{otherwise,} \end{cases}$$

$$x_{4i} = \begin{cases} 1 & \text{if the } i\text{th mother's community is in a rural area of Guatemala,} \\ 0 & \text{otherwise,} \end{cases}$$

$$x_{5i} = \begin{cases} 1 & \text{if the } i\text{th mother completed secondary school or higher,} \\ 0 & \text{otherwise,} \end{cases}$$

and

$$x_{6i} = \begin{cases} 1 & \text{if the } i\text{th mother's husband completed secondary school or higher,} \\ 0 & \text{otherwise.} \end{cases}$$

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

$$y_{ij}|\beta_0, \beta_1, \ldots, \beta_6 \overset{\text{ind.}}{\sim} \text{Bernoulli}\left(\frac{1}{1+\exp[-\{\beta_0 + u_{i0} + (\beta_1 + u_{i1})x_{1ij} + \beta_2\, x_{2i} + \ldots + \beta_6\, x_{6i}\}]}\right),$$

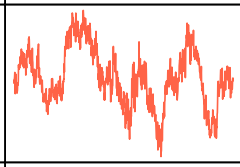
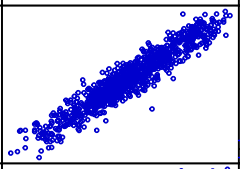
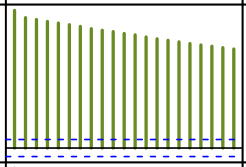
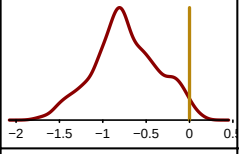
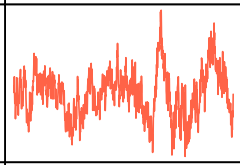
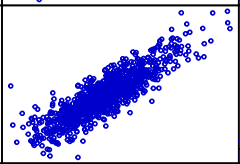
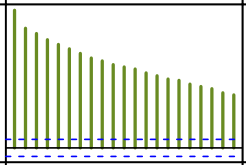
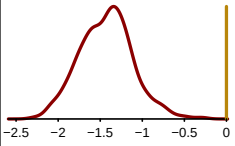
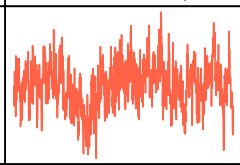
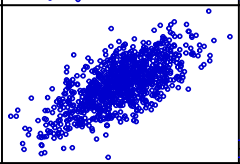
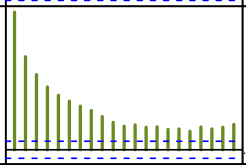
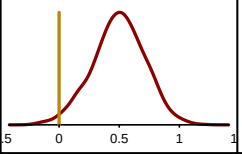
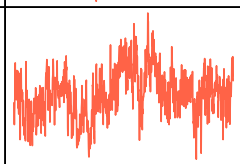
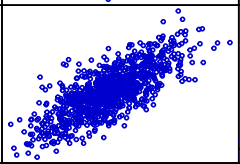
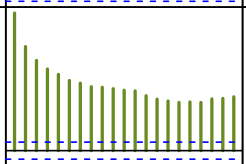
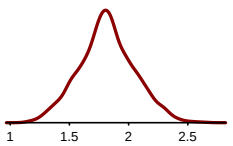
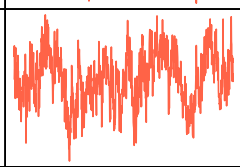
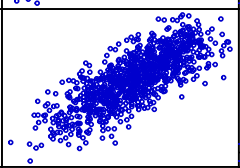
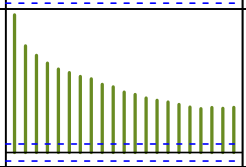
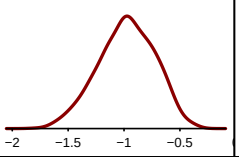
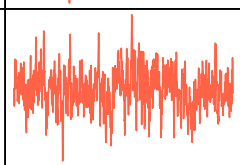
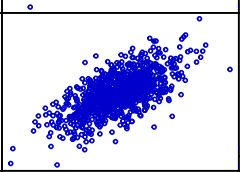
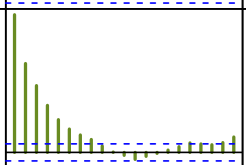
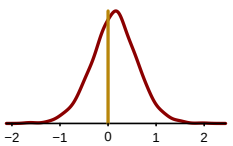
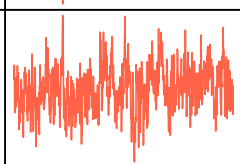
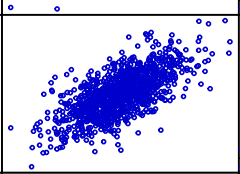
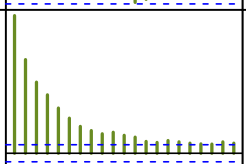
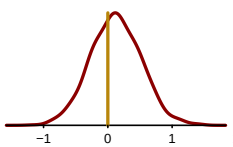
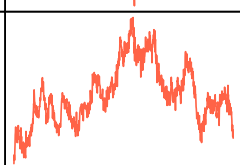
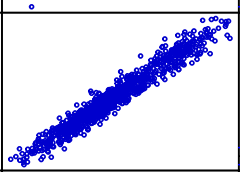
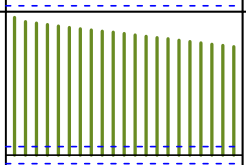
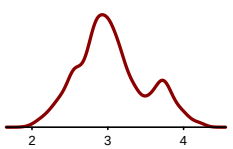
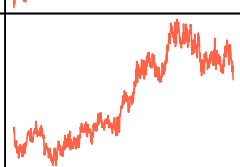
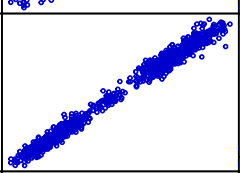
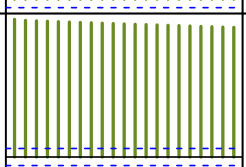
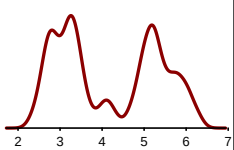
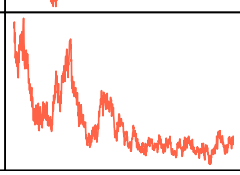
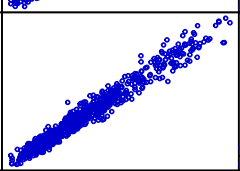
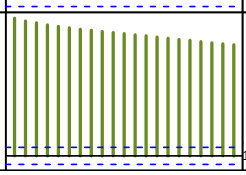
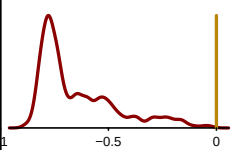
$$\left.\begin{bmatrix} u_{0i} \\ u_{1i} \end{bmatrix}\right| \sigma^2_{u0}, \sigma^2_{u1}, \rho_u \overset{\text{ind.}}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{u0} & \rho_u\sigma_{u0}\sigma_{u1} \\ \rho_u\sigma_{u0}\sigma_{u1} & \sigma^2_{u1} \end{bmatrix}\right)$$

and suitable diffuse priors on $\beta_0, \beta_1, \beta_2, \ldots, \beta_6, \sigma^2_{u0}, \sigma^2_{u1}$ and ρ_u.

Ideally, each class member would run the script `GuatImmunViaStan.R` on their laptop as part of Assignment 6. However, getting the MCMC samples to converge reasonably well requires a very large burnin and the script may take several hours to run. Instead, running `GuatImmunViaStan.R` is left as an optional extra for Assignment 6 and the figure overleaf can be used to complete this assignment question.

In half a page or less, what conclusions can be drawn from the output regarding the effect of the predictors on immunisation probability for the study population?

Hint: Note the hint for Question 3(c).

parameter	trace	lag 1	acf	density	summary
intercept					posterior mean: −0.736 95% credible interval: (−1.48,−0.0345)
proportion indigenous in community					posterior mean: −1.43 95% credible interval: (−2.05,−0.765)
indicator of mother works					posterior mean: 0.496 95% credible interval: (0.0443,0.899)
indicator of child aged over 2 years					posterior mean: 1.82 95% credible interval: (1.36,2.29)
indicator of community in rural area					posterior mean: −0.978 95% credible interval: (−1.49,−0.538)
indicator of mother sec. school					posterior mean: 0.131 95% credible interval: (−0.839,1.06)
indicator of husband sec. school					posterior mean: 0.124 95% credible interval: (−0.63,0.895)
σ_{u0}					posterior mean: 3.05 95% credible interval: (2.26,3.96)
σ_{u1}					posterior mean: 4.17 95% credible interval: (2.51,6.11)
ρ_u					posterior mean: −0.641 95% credible interval: (−0.835,−0.169)

♦ ♦ ♦