# AUSTRALIAN NATIONAL UNIVERSITY
# RESEARCH SCHOOL OF FINANCE ACTUARIAL STUDIES, AND APPLIED STATISTICS

INTRODUCTION TO BAYESIAN DATA ANALYSIS (STAT3016/4116/7016)
SEMESTER 2 2016

ASSIGNMENT 4

**DUE DATE: Thursday 6 October 2016, by 4pm**
(12.5% of total course grade)

**INSTRUCTIONS:**

1. All students must hand in an assignment of their own writing.
2. The assignment should be handed in to the assignment box for STAT3016/4116/7016 available on level 4 of the ANUCBE Building 26C. There will be no online submission facility.
3. Ensure you also complete and attach a cover sheet to your assignment (available on the course website)
4. Begin each question on a new page.
5. Where required, provide sufficient computer output to support your answers. Provide enough intermediate numerical calculations to justify working for your final answer.
6. Computer output must be interpreted in written format. A solution solely highlighting the computer output is not acceptable.
7. No late assignments will be accepted.

**COLLABORATION POLICY** (as stated in the course outline)

University policies on plagiarism will be **strictly** enforced. You are encouraged to (orally) discuss your assignments with your classmates, but each student must write up solutions separately. Be sure that you have worked through each problem yourself and that all answers you submit are the results of your own efforts. This includes all computer code and output.

**Problem 1**

The files `school1.dat` through `school8.dat` give weekly hours spent on homework for students sampled from eight diffferent schools. We want to obtain posterior distributions for the true means for the eight different schools using a hierarchical normal model with the following prior parameters:

$$\mu_0 = 7, \gamma_0^2 = 5, \tau_0^2 = 10, \eta_0 = 2, \sigma_0^2 = 15, \nu_0 = 2$$

(a) Run a Gibbs sampling algorithm to approximate the posterior distribution of $\left\{\boldsymbol{\theta}, \sigma^2, \mu, \tau^2\right\}$. Assess the convergence of the Markov chain, and find the effective sample size for $\left\{\sigma^2, \mu, \tau^2\right\}$. Run the chain long enough so that the effective sample sizes are all above 1,000.

(b) Compute posterior means and 95% confidence regions for $\left\{\sigma^2, \mu, \tau^2\right\}$. Also, compare the posterior densities to the prior densities, and discuss what was learned from the data.

(c) Plot the posterior density for $R = \frac{\tau^2}{\sigma^2+\tau^2}$ and compare it to a plot of the prior density on $R$. Describe the evidence for between-school variation.

(d) Obtain the posterior probability that $\theta_7$ is smaller than $\theta_6$, as well as the posterior probability that $\theta_7$ is the smallest of all the $\theta$'s.

(e) Plot the sample averages $\bar{y}_1, ..., \bar{y}_8$ against the posterior expectations of $\theta_1, ...., \theta_8$, and describe the relationship. Also compute the sample mean of all observations and compare it to the posterior mean of $\mu$.

**Problem 2**

Younger male sparrows may or may not nest during a mating season, perhaps depending on their physical characteristics. Researchers have recorded the nesting success of 43 young male sparrows of the same age, as well as their wingspan, and the data appear in the file `msparrownest.dat.` Let $Y_i$ be the binary indicator that sparrow $i$ successfully nests, and let $x_i$ denote their wingspan. Our model for $Y_i$ is logit[ Pr $(Y_i = 1|\alpha, \beta, x_i)] = \alpha + \beta x_i$, where the logit function is given by $\text{logit}[\theta] = \log[\theta/(1-\theta)]$.

(a) Write out the joint sampling distribution $\prod_{i=1}^{n} p(y_i|\alpha, \beta, x_i)$ and simplify as much as possible.

(b) Formulate a prior probability distribution over $\alpha$ and $\beta$ by considering the range of $Pr(Y = 1|\alpha, \beta, x)$ as $x$ ranges over 10 to 15, the approximate range of the observed wingspans.

(c) Implement a Metropolis algorithm that approximates $p(\alpha, \beta|\mathbf{y}, \mathbf{x})$. Adjust the proposal distribution to achieve a reasonable acceptance rate, and run the algorithm long enough so that the effective sample size is at least 1,000 for each parameter.

(d) Compare the posterior densities for $\alpha$ and $\beta$ to their prior densities.

(e) Using output from the Metropolis algorithm, come up with a way to make a confidence band for the following function $f_{\alpha,\beta}(x)$ of wingspan:

$$f_{\alpha,\beta}(x) = \frac{\exp^{\alpha+\beta x}}{1 + \exp^{\alpha+\beta x}}$$

where $\alpha$ and $\beta$ are the parameters in your sampling model. Make a plot of such a band.

**Problem 3**

The file `tplant.dat` contains data on the heights of ten tomato plants, grown under a variety of soil pH conditions. Each plant was measured twice. During the first measurement, each plant's height was recorded and a reading of pH soil was taken. During the second measurement only plant height was measured, although it is assumed that pH levels did not vary much from measurement to measurement.

(a) Using ordinary least squares, fit a linear regression to the data, modelling plant height as a function of time (measurement period) and pH level. Interpret your model parameters.

(b) Perform model diagnostics. In particular, carefully analyse the residuals and comment on possible violations of assumptions. In particular, assess (graphically or otherwise) whether or not the residuals within a plant are independent. What parts of your ordinary linear regression model do you think are sensitive to any violations of assumptions you may have detected?

(c) Hypothesise a new model for your data which allows for observations within a plant to be correlated. Fit the model using a MCMC approximation to the posterior distribution, and present diagnostics for your approximation.

(d) Discuss the results of your data analysis. In particular, discuss similarities and differences between the ordinary linear regression and the model fit with correlated responses. Are the conclusions different?

**Problem 4 [STAT4116/STAT7016 ONLY]**

Non-conjugate hierarchical models: An experiment is conducted to estimate $\theta$, the probability of developing a tumor in a population of female rats that receive a dose of drug X. (Such studies are routinely done in the evaluation of drugs for possible clinical application). Suppose $J$ such experiments have been conducted historically. In the $j^{th}$ historical experiment, let the number of rats with tumors be $y_j$ and let $n_j$ be the total number of rats tested in the $j^{th}$ experiment. We model the $y_j$'s as independent binomial data, given the sample sizes $n_j$, and study-specific means $\theta_j$.

Suppose we assume that the tumor probabilities $\theta$ follow a normal distribution on the log-odds scale, that is, $\text{logit}(\theta_j) \sim N(\mu, \tau^2)$

(a) Write the joint posterior density $p(\boldsymbol{\theta}, \mu, \tau^2|\mathbf{y})$ (where $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)$ and $\mathbf{y} = (y_1, ...., y_J)$).

(b) To obtain the marginal posterior distribution $p(\mu, \tau^2|\mathbf{y})$, we can integrate the joint distribution in (a) over $\boldsymbol{\theta}$. Show that this integral has no closed-form expression.

(c) We can also compute the marginal posterior distribution of $(\mu, \tau^2)$ using the conditional probability formula,

$$p(\mu, \tau^2|\mathbf{y}) = \frac{p(\boldsymbol{\theta}, \mu, \tau^2|\mathbf{y})}{p(\boldsymbol{\theta}|\mu, \tau^2, \mathbf{y})}$$

Why is the above expression not helpful to evaluate $p(\mu, \tau^2|\mathbf{y})$?

In practice, we can solve this problem by normal approximation, importance sampling, and MCMC simulation.