Advanced Bayesian Methods (4H/M)
Level M Assessment

Total Marks: 20

You are allowed to make use of any notes, codes, or materials from this class. All answers must be written in your own words (you can cite any references if necessary). Please submit a report electronically in PDF format (through Moodle) answering all the questions below, and also separately submit your R code as a text file.

You are provided a data set, `foodpoisoning.csv`, containing information from an investigation concerning the occurrence of an outbreak of acute gastrointestinal illness on a national handicapped sports day in Thailand in 1990[1]. The data set has the following variables:

```
sex: a numeric vector (0 = female, 1 = male)
age: a numeric vector- age in years
beefcurry: a numeric vector- whether the subject had eaten beef curry
eclair: a numeric vector- pieces of eclair eaten
       80 = ate but could not remember how much
water: a numeric vector- whether the subject had drunk water
case: 1 if the person was a victim of food poisoning, 0 if not
```

The question of interest is to trace the possible cause of the food poisoning outbreak and compare the severity of the outbreak among various demographical segments of the population, using a Bayesian logistic regression model. Let us denote the response variable (`case`) for individual i as Y_i ($i = 1, \ldots, n$), the set of explanatory variables (including intercept term) as $\boldsymbol{X}_i = (1, X_{i1}, \ldots, X_{ip})$, and regression coefficients $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$. For a single observation y_i, we then have,

$$f(y_i|\boldsymbol{\beta}, \boldsymbol{X}_i) \propto (\text{logit}^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta}))^{y_i}[1 - \text{logit}^{-1}(\boldsymbol{X}_i'\boldsymbol{\beta})]^{1-y_i},$$

where $\text{logit}(\theta) = \log[\theta/(1-\theta)]$.

Based on the description above, answer the questions on the following page.

[1] Thaikruea, L., Pataraarechachai, J., Savanpunyalert, P., Naluponjiragul, U. (1995). An unusual outbreak of food poisoning. *Southeast Asian J Trop Med Public Health* 26(1):78-85.

1. Using the expression above, and a non-informative prior for $\boldsymbol{\beta}$, write down an expression proportional to the posterior density of $\boldsymbol{\beta}|y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$.

[2 marks]

2. Do preliminary exploratory analysis of the data to check for any possible issues, such as outliers or missing values, and to get a sense of how the different variables appear to be related. State your preliminary conclusions.

[1 mark]

3. Do a logistic regression analysis of the data in `R`, using *glm*. Report your conclusions.

[1 mark]

4. Write down the steps of a Metropolis-Hastings (MH) sampler to get draws from the posterior density of $\boldsymbol{\beta}|y_1, \ldots, y_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. Choosing a Normal$(\hat{\boldsymbol{\beta}}, \sigma^2 I)$ proposal (jump) density, write down the form of the acceptance ratio for the sampler ($\hat{\boldsymbol{\beta}}$ denotes the estimates of $\boldsymbol{\beta}$ from the classical logistic regression analysis in question 3, and I is an identity matrix).

[4 marks]

5. Next, implement the MH sampler in `R`, by either writing your own code for the Metropolis-Hastings update function, or adapting similar codes used in the class (provided on Moodle).

[4 marks]

6. Run the sampler for different choices of σ^2, and for each choice, record the sampler's characteristics in terms of convergence (or non-convergence), making use of different convergence criteria (graphical and numeric) discussed in class. Present a table with the different choices of σ^2, numeric values of the calculated convergence criteria, and the runtimes of the sampler (for a pre-set number of iterations). State your conclusions based on the above. [For this part you can make use of functions in `coda` or other packages.]

[4 marks]

7. Once you are satisfied with the level of convergence of your MH sampler, present summaries from the posterior distribution (numerically and/or graphically), and state your final inference and conclusions from the data.

[3 marks]

8. Compare your results in question 7 with a classical logistic regression analysis of the data (in question 3), and comment on what you observe, and whether or not this is what you expect.

[1 mark]