

## **Bayesian Statistics: Main Assignment**

*Marks for this assignment: 44 of course total of 88 (50%)*

**See Moodle for submission and deadline**

*This assignment is in three sections. Please include your response to all parts in a single Word document, with the section (and subsection eg (a) where appropriate) clearly labelled. Start each section on a new page.*

### **Section 1**

An entomologist wants to estimate the relative preference that a butterfly species has for laying eggs on four species of plant, and would like some statistical advice from you. The planned experiment involves counting the number of caterpillars on a randomly chosen leaf of a plant that is randomly chosen from all plants in a field. A single researcher will make observations on multiple plants in the same field, for a total of 20 fields. He will also record the species of the plant, which will be one of four common species (plants of other species will be ignored). He will also take a note of the weather (either dry or raining), and the type of field (either grassland or brownfield), as he believes these may have an effect on the number of caterpillars observed. He already knows that eggs are laid in clusters on plant leaves, and that caterpillars of these species are very small and do not move between plants at all before they pupate into butterflies. You may assume that the researcher is able to find and identify caterpillars perfectly, and that there is no other information that is likely to be useful in predicting the number of caterpillars on a plant.

(a) What distribution would the observed data follow, and what information would you include as predictor variables (also indicate what type of effects these would be – ie fixed, linear or random).

*3 marks*

(b) Give any relevant advice on how to tweak the design of the experiment to make statistical analysis easier (without changing the total number of observations or fields).

*2 marks*

(c) Describe, in your own words, the differences between Bayesian and Frequentist philosophy and analysis. Which type do you think is best suited to this problem and

why?

*3 marks*

(d) Where might we be able to find prior information for this problem? The biologist does not believe that there is any useful prior information available, so does not think we can use a prior at all. How would you respond?

*2 marks*

## Section 2

Explain in your own words the advantages and disadvantages of Bayesian Markov chain Monte Carlo (MCMC) relative to a frequentist maximum likelihood (ML) method such as those available in packages such as Stata, SAS or R. Include in your answer the ways that you might minimise the potential for errors in using and interpreting MCMC. Also outline the principle of one algorithm that can be used to generate a Markov chain (you do NOT need to include any code needed to implement this algorithm).

*10 marks*

## Section 3

*Where JAGS code is asked for in this section, you do NOT need to specify the R code needed to run the model, just the JAGS model code itself*

You have been asked by a cardiovascular surgeon to look at some potential risk factors for occurrence of heart disease in 500 men of the same age in the Greater Glasgow area. You have been given the following information taken from patient records:

HeartDisease	- <b>No</b> or <b>Yes</b>
SmokingHistory	- <b>NeverSmoked</b> or <b>PreviousSmoker</b> or <b>CurrentSmoker</b>
DrinkingHistory	- <b>Light</b> or <b>Moderate</b> or <b>Heavy</b>
BMI (or Body Mass Index)	- A continuous variable ranging from approximately 15 to 40
PostcodeArea	- One of 50 postcode areas in the Greater Glasgow area; a clustering variable which is known <i>a priori</i> to have an effect on heart disease through the effect of social status

The data for this problem is contained in the 'AssignmentData.R' file on Moodle – the five variables are named as given above (you can access them by typing `read.table('AssignmentData.R')`). Familiarise yourself with the data, and process them as necessary for the questions below.

You can safely assume that the diagnosis of heart disease is accurate in each case (ie. there are no false positive or false negative diagnoses), and that all of the information given is both complete and correct. Based on previous studies there is a suspicion that there may be an interaction between smoking and drinking history. *Assume that there is no possible interaction between any other effects, and do not attempt to fit a quadratic term to the BMI data.* Also assume that there is no useful prior information available.

(a) Give the JAGS code for all of the candidate models for which you would evaluate the empirical fit to this data (within the constraints given above). For each candidate model, give the initial values that you would use for two chains, and indicate which variables you would monitor.

*10 marks*

(b) Run each of the models you have given in (a) appropriately, and copy/paste the standard summary results (produced by `run.jags`) that you have obtained from each model (this should include the number of iterations run, and a statistic that you can use to compare the empirical fit of the candidate models). Give the details of the steps that you have taken to ensure that the model is efficient and the output is appropriate, including the full details of any data processing or transformation that you may have performed before running the models.

*8 marks*

(c) Write a brief (roughly  $\frac{1}{2}$  page A4) summary report to be read by the cardiovascular surgeon that provided these data. Be sure to provide a complete interpretation of your analysis, including the relative effects of all predictor variables on your outcome. You should include an interpretation of your model fitting results and the appropriate summary statistics for the relevant effects, but do not include any of the JAGS code you have used. You can assume that the surgeon has a reasonable understanding of statistical terminology, but not the expert knowledge that you have!

*6 marks*