STA 365: Assignment 2

Professor Boris Babic

Due Date: Sunday, April 17, 2022 (11:59pm via Quercus)

Instructions.

This is the second (and final) assignment. It is to be treated like a take-home exam. Accordingly, unlike the homework, this assignment should be completed independently (on your own) without help or consultation from your colleagues, or from anyone else (except the course instructor and TAs). The TAs will be instructed to flag answers that look sufficiently similar.

You are allowed to use the course lecture notes, the recommended textbooks, and external sources. However, you must cite the sources you rely on, and you should be using these sources to inform/help you in developing your own answer, not copying them.

You are encouraged to type out your answers in LaTeX or a word processor. If you need to handwrite your responses, make sure that they are clear and legible, and that you scan a high quality image. What cannot be read will be marked as incomplete.

Where a problem requires the use of R, you must produce your associated R code. While you may use other software, solutions will be provided only in R.

Problem 1. (40 points)

Download the file swim_time.RData from the course page. The data file contains a data matrix Y on the amount of time, in seconds, it takes each of four high school swimmers to swim 50 yards. Each swimmer has six times, taken on a biweekly basis.

(a) For each swimmer $j, (j = 1, 2, 3, 4)$, fit a Bayesian linear regression model which considers the swimming time as the response variable and week as the explanatory variable. To formulate your prior, use the information that competitive times for this age group generally range from 22 to 24 seconds.

(b) For each swimmer $j, (j = 1, 2, 3, 4)$, obtain a posterior predictive distribution for Y_j^*, their time if they were to swim two weeks from the last recorded time.

(c) The coach of the team has to decide which of the four swimmers will compete in a swimming meet in two weeks. Using your predictive distributions, compute $Pr(Y_j^* = \min Y_1^*, ..., Y_4^* | Y)$ for each swimmer j, and based on this make a recommendation to the coach.

Problem 2. (30 points)

In R, load library(MASS) and then consider the dataset **UScrime** which contains crime rates

(y) and data on 15 explanatory variables for 47 U.S. states. A description of the variables can be obtained by typing "?UScrime" in R console.

(a) Fit a Bayesian linear regression model using uninformative priors. Obtain marginal posterior means and 95% credible intervals for coefficients. Describe the relationships between crime and the explanatory variables. Which variables seem strongly predictive of crime rates?

(b) To test how well regression models can predict crime rates based on the explanatory variables, randomly divide the data roughly in half, into training set and a test set. Use the training dataset to fit the model and generate the posterior predictive median of the crime rates given the explanatory variables in the test dataset. Compare the posterior predictive median and the actural crime rate in the test dataset.

(c) Repeat Parts (a) and (b) using spike-and-slab priors for regression coefficients.

Problem 3. (30 points)

In R, load library(geoR) (You need to install the package first if you have not) and then consider the dataset gambia which consists of 2,035 children from 65 villages from The Gambia. It contains eight different variables. A description of the variables can be obtained by typing "?gambia" in R console. Let $Y_i \in \{0, 1\}$ (pos) indicate the presence (1) or absence (0) of malaria in a blood sample taken from child $i, (i = 1, ..., 2035)$. Let $X_i = 1$ (treated) if child i regularly sleeps under a bed-net, and $X_i = 0$ otherwise. Let $v_i \in \{1, ..., 65\}$ denote the village of child i. Note that the dataset only contains the locations of villages instead of the labels. You can use the following R code to obtain v_i.

```
v_loc = unique(gambia[,"x"])
v = match(gambia[,"x"],v_loc)
```

Fit the following logistic regression model:

$$\text{logit}[\Pr(Y_i = 1)] = \alpha_{v_i} + \beta_{v_i} X_i,$$

where α_j and β_j are intercept and slope for village $j, (j = 1, ..., 65)$. The priors are:

$$\alpha_j \sim \text{N}(\mu_\alpha, \sigma_\alpha^2), \beta_j \sim \text{N}(\mu_\beta, \sigma_\beta^2).$$

Choose uninformative priors for the hyperparameters $\mu_\alpha, \mu_\beta, \sigma_\alpha^2, \sigma_\beta^2$. Based on your model fitting, address the following questions:

(a) Scientifically, why might the effect of bed-net vary by village?

(b) Do you see evidence that the slopes and/or intercepts vary by village? You may consider alternative model fitting and perform model comparisons.

(c) Which village has the largest intercept? Slope? Does this agree with the data in these villages?

(d) Are the results sensitive to the priors for the hyperparameters?