Assignment 2

STAT314/STAT461

Set: Tue, Aug-10. Due: Fri Aug-20

Please type everything in either Word or LaTex, and submit it as a PDF file via Learn. No handwritten submissions! (And no scans of handwritten submissions, please).

Show your workings: equations for theoretical problems; code and (relevant) output for the computational problems. It is not sufficient to report an answer. Don't forget to include intermediate steps and explain your way of thinking. You may get points for thinking in the right direction even if you don't get the answer exactly right.

If you need an extention, please ask for it in advance. Late submissions will not be accepted.

In this Assignment, we are going to take a look at the Penguin data collected and made available by Dr. Kristen Gorman and the Palmer Station Long Term Ecological Research (LTER) Program.

You may need to install the package `palmerpenguins` either via the Packages menu or directly via the Console before you start this exercise.

```
install.packages("palmerpenguins")
```

Then use

```
library(palmerpenguins)
data(penguins)
```

to "activate" the data.

Feel free to do some exploratory analysis first to get familiar with it.

Problem 1. Gaussian model.

For now, we are going to focus on **male Adelie** penguins only.

(a) Assuming that the body mass follows an island-specific normal distribution with the unknown mean μ and (common) known standard deviation $\sigma = 350g$, derive posterior island-specific distributions for the mean population weight of male Adelie penguins. Carefully explain your prior assumptions. (**1pt**)

(b) Which island has the heaviest male Adelie penguins, on average, and how certain are you about it? (**1pt**)

Problem 2. Population means vs. random individual penguins.

The Gentoo penguins were only found on Biscoe island. Let's look at them now.

(a) Again, assuming Gaussian distribution for the body mass, derive posterior island-specific distributions for the mean population weight of male and female Gentoo penguins respectively, assuming that the common $\tau = 500^{-2}$ is known. Explain your prior assumptions. (**1 pt**)

(b) What is the posterior probability that the males are **on average** heavier than females? (**1 pt**)

(c) What is the posterior probability that a random individual male is heavier than a random individual female? (Hint: use posterior predictive distribution.) (**1pt**)

Problem 3. Simple Linear Regression.

Now let's look at all the penguins together, to see whether there is correlation between bill length (x_i) and bill depth (y_i). **NB. R does not deal well with observations containing missing values, so it is a good idea to remove them before modeling:**

```
penguins.clean <- penguins[(!is.na(penguins$bill_length_mm))&
                           (!is.na(penguins$bill_depth_mm))&
                           (!is.na(penguins$species)),]
```

(a) Fit a simple linear model:

$$\text{Length}_i = a + b\text{Depth}_i + \epsilon_i$$

using the `MCMCglmm(bill_length_mm ~ bill_depth_mm,...)` and interpret the coefficients. Do you find the results surprising, and, if so, why. (**2pt**)

(b) Now, let's fit a model which takes species into account:

$$\text{Length}_i = a_{\text{Species}_i} + b_{\text{Species}_i}\text{Depth}_i + \epsilon_i$$

using the `MCMCglmm(bill_length_mm ~ species*bill_depth_mm,...)`

Take care when interpreting the coefficients. The model converts factors into dummy variables and compares everything to the baseline (first species alphabetically). Thus, the intercept for the Chinstrap species will be `(Intercept) + speciesChinstrap` and the slope for the same will be `bill_depth_mm + bill_depth_mm:speciesChinstrap`.

Note, that here the coefficients are species-specific, but the residual inverse variance *tau* is still assumed to be common.

Interpret the coefficients. (**2pt**)

(c) Produce a plot of your data. Use color or point type to differentiate between the species. Plot the fitted regression lines (one for part a and three for part b) on top of the data points. (**1pt**)