Homework 04

Math 315, Fall 2018

Due: Oct. 5 by 4 p.m.

Instructions: Complete the following problems and submit them by 4 p.m. on the due date. Please make sure that your solution is neatly written, clearly organized, and stapled (if there are multiple pages).

This is the final version of the homework.

Problem 1: 4H1

- Repeat this problem, but this time assume that you're making predictions about an **invidual** rather than the mean.
- Be sure to include a nicely formatted table as part of your solution. Don't make me read raw code/output exclusively!

Problem 2: An Adaptation of 5H1

(a) Fit a regression model for body weight as a linear function of territory size (`area`) and create a table of the *maximum a posteriori* estimates for the parameters, along with standard deviations and 89% credible intervals.

(b) Fit a regression model for body weight as a linear function of `groupsize` and create a table of the *maximum a posteriori* estimates for the parameters, along with standard deviations and 89% credible intervals.

(c) Plot the results of the regression models from parts (a) and (b), displaying the MAP regression line and a 95% credible interval for the mean response. (Formatting hint: to combine two `ggplot2` objects, use `grid.arrange()` in the `gridExtra` package.)

(d) Based on your 89% credible intervals from parts (a) and (b) and the plots from part (c), is either variable important for predicting fox body weight? Briefly justify your answer.

Problem 3: An Adaptation of 5H2

(a) Fit a multiple linear regression model with `weight` as the outcome and both `area` and `groupsize` as predictor variables. Create a table of the *maximum a posteriori* estimates for the parameters, along with standard deviations and 89% credible intervals.

(b) Create a counterfactual (i.e. effect) plot for each predictor, holding the other predictor constant at its mean.

(c) Based on the credible intervals and/or the counterfactual plots, what does this model say about the importance of each variable?

(d) Why do you get different results than in problem 2 (i.e. 5H1)?

Problem 4: The Behrens-Fisher problem The classic Behrens-Fisher problem is that of estimating normal means from independent populations with unknown (and thus possibly unequal) variances. In this problem, you will consider this problem in the context of an experiment on the effects of magnetic fields on the flow of calcium out of chicken brains. The experiment made measurements on two groups of chickens: an exposed group (A) and a control group (B). Each chicken was measured only once. The data for this problem can be loaded using the command:

```
calcium <- read.csv("https://aloy.rbind.io/data/calcium.csv")
```

(a) Use a two-sample t-test (i.e. frequentist methods) to calculate a 97% confidence interval for $\mu_A - \mu_B$.

(b) Write down the likelihood and link function for a regression model that could be used to estimate the difference in means you considered in part (a).

(c) Fit this regression model using `map()` (that is, using quadratic approximation) and report the *maximum a posteriori* estimates for the parameters, along with standard deviations and 97% credible intervals.

(d) Report and interpret a 97% credible interval for $\mu_A - \mu_B$. How does this interval compare to you confidence interval in part (a)?

Problem 5: Model formulation

Problem 3 For each of the following, write an expression for the mean function of the regression model and carefully define the predictor variables in your model (including dummy variables). Additionally, indicate how the desired quantities could be estimated from the *maximum a posteriori* parameter estimates. You do not need to worry about how to construct the posterior distribution of these quantities.

For example, you answer may be of the form: $\mu_i = \alpha + \beta x_i$, where the quantity of interest is $\alpha + \beta(100)$.

(a) A study is comparing the energy content of constant-sized pieces of firewood from different tree species. If you are burning wood to heat a room or a house, a higher energy content is a good thing.

One complication is that the energy released depends on the moisture content of the firewood, which is hard to standardize. You have studied three species (Red Oak, White Pine and Black Walnut). You believe that the relationship between energy content and moisture is linear with the same slope for each species. You want to estimate the difference in energy content at 10% moisture content between White Pine and Red Oak.

(b) Consider the situation in part (a) again, assuming that the three species have different slopes (for the association of moisture content on energy). You want to estimate the difference in energy content at 10% moisture between White Pine and Red Oak.

(c) Education researchers are studying whether watching television impacts the performance of graduate students. For each student in a class, they have Y, the exam score, and X, the number of hours spent watching television during the week prior to the exam. They assume that the relationship between Y and X is linear up to 20 hours. After X=20 hours, there is no relationship, i.e. the slope is 0 for X>20. They wish to estimate the slope for 0 to 20 hours and the expected difference between light television watching (3 hours) and heavy watching (25 hours).