STAT 4640/7640 **Homework 8**

Due: April 21, 2022

- **Instructions**: Make sure your name is on your paper and your answers are clearly written.

1. The dataset `SKorea_Covid19.Rdata` contains birth year and sex of a sample of coronavirus patients in South Korea. Each of these observations corresponds to an individual that tested positive for the virus. Assume that birth year is normally distributed.

   (a) The median age of South Korea citizens is 40.8, which corresponds to a birth year of approximately 1979.5. In the Bayesian framework, conduct a hypothesis test to identify if the average birth year of coronavirus patients in South Korea is equal to the median age of the citizens. Make sure to include the null and alternative hypothesis, the code used to conduct the test, and an interpretation of the results.

   (b) In the Bayesian framework, conduct a hypothesis test to identify whether the average age of infected individuals is the same for men and women. Make sure to include the null and alternative hypothesis, the code used to conduct the test, and an interpretation of the results.

2. Using the Homes example as your guide, fit the Bayesian multiple linear regression model using the three different prior specifications for the `lakesN.Rdata` data. This dataset comes from a research project I am involved with (see `https://lagoslakes.org/`), and contains observations of total nitrogen in 745 lakes across the northeast US, as well as possible important covariates for explaining total nitrogen. The following variables are included in the data.

   - $logTN$ - log of total nitrogen (response variable)
   - $hu8\_baseflow$ - watershed measure of flow
   - $hu8\_no3depo$ - watershed measure of nitrate deposition
   - $hu8\_totalndepo$ - watershed measure of total nitrogen deposition
   - $hu8\_runoff$ - watershed measure of runoff
   - $urban$ - percent of urban area in watershed
   - $rowcrop$ - percent of agriculture in watershed
   - $pasture$ - percent of pasture area in watershed
   - $forest$ - percent of forest area in watershed
   - $wetland$ - percent of wetland area in watershed
   - $lake\_area$ - lake area
   - $maxdepth$ - maximum depth
   - la_wa_ratio - ratio of lake size to watershed size
   - nhd_lat - latitude

- *nhd_long* - longitude
- *july* - indicator variable if water sample taken in July (September is base case)
- *august* - indicator variable if water sample taken in August (September is base case)
- *DRstream* - indicator variable if lake is located below a stream (below a lake and stream is base case)
- *headwater* - indicator variable if lake is a headwater, meaning no stream or lake feeding into it (below a lake and stream is base case)

Your task is to conduct multiple regression with $logTN$ as the response variable and the remaining variables as possible predictor variables. Recall that the Bayesian multiple linear regression model is

$$Y_i \sim Normal(\beta_0 + \sum_{j=1}^{p} X_{ij}\beta_j, \sigma^2).$$

Compare the models for the three different priors for the coefficients $\beta_1, \ldots, \beta_p$.

(a) Uninformative Gaussian: $\beta_j \sim Normal(0, 1000)$

(b) Gaussian shrinkage: $\beta_j \sim Normal(0, \sigma_b^2\sigma^2)$ with $\sigma_b^2 \sim InvGamma(0.1, 0.1)$

(c) Bayesian Lasso: $\beta_j \sim DE(0, \sigma_b^2\sigma^2)$ with $\sigma_b^2 \sim InvGamma(0.1, 0.1)$