STA442H1 S Assignment 2
Due on March 17, 2023 11:59 PM on CrowdMark
All relevant work must be shown for credit.

Note: In any question, if you are using R, all R codes and R outputs must be included in your answers. You should assume that the reader is not familiar with R outputs and so explain all your findings, quoting necessary values form your outputs. Please note that academic integrity is fundamental to learning and scholarship. You may discuss questions with other students. However, the work you submit should be your own. If I feel suspicious of any assignment (e.g. if your work doesn't appear to be consistent with what we have discussed in class), I will not mark the assignment. Instead, I will ask you to present your work in my office and your grade will be assigned based on your presentation. Assignments can be hand written but the R codes and outputs should be printed.

1. For this question you have to simulate a dataset. Let's assume the outcome Y depends on 50 covariates $X_1, X_2, \ldots, X_{50}$ linearly. That is the relationship is presented with the following equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{50} X_{50} + \epsilon \qquad (1)$$

(a) [**8 Marks**] Perform the following simulations.

- Generating training sets of size 100. That is,
  - Generate 100 random values of all X variables from standard normal distribution. That is $X_1 \sim N(0,1), X_2 \sim N(0,1) \ldots X_{50} \sim N(0,1)$
  - Generate ϵ also from standard normal $\epsilon \sim N(0,1)$
  - Generate βs from some Uniform distribution where β_1 to β_{20} are simulated from Uniform(0.5, 1.5) and β_{21} to β_{50} are simulated from Uniform(0.2, 0.4)
  - Then generate Y using (1)
- Generating test set of size 1000,
  - Generate 1000 random values of all X variables from standard normal distribution. That is $X_1 \sim N(0,1), X_2 \sim N(0,1) \ldots X_{50} \sim N(0,1)$
  - Generate ϵ also from standard normal $\epsilon \sim N(0,1)$
  - Use the same βs generated for the training set.
  - Then generate Y using (1)

(b) [**5 Marks**] Fit a linear regression where Y is the outcome and $X_1, X_2, ..., X_{50}$ are the predictors. Calculate the prediction error from the test set.

(c) [**5 Marks**] Fit a Ridge regression and again calculate the prediction error from the test set.

(d) [**5 Marks**] Fit a LASSO and again calculate the prediction error from the test set.

(e) [**5 Marks**] Which method in (b) - (d) provides the lowest prediction error on the test set? Explain why.

(f) [**15 Marks**] Change the training set size to 10000 from 100. Perform (a) - (d) again. Which method now provides the lowest prediction error? Explain why?

2. For this problem you need to load the NHANES dataset using the following command

   ```
   ## If the package is not installed then use ##
   install.packages('NHANES') ## And install.packages('tidyverse')
   library(tidyverse)
   library(NHANES)
   small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12"
   & NHANES$Age > 17,c(1,3,4,8:11,13,25,61)])
   small.nhanes <- small.nhanes %>%
   group_by(ID) %>% filter(row_number()==1)
   ```

   This is data collected by US National Center for Health Statistics (NCHS). The preceeding codes creates a small dataset of the original NHANES dataset. With this dataset answer the following questions,

   (a) [**5 Marks**] Randomly select 500 observations from the data. For this selection use your student ID as seed. Fit a logistic regression to predict smoking status (variable `SmokeNow`), using all the other variables (excluding `ID`). Explain your results in few sentences.

   (b) [**5 Marks**] Perform a model selection procedure based on step wise methods (both AIC and BIC) and also using elastic-net. Do they select the same model? Why or why not? For the elastic-net selection, consider $\alpha = 0.5$ and 1.

   (c) [**5 Marks**] Perform an internal validation using cross-validation. Explain your results.

   (d) [**5 Marks**] Construct the Receiver operating characteristic (ROC) curve. Calculate the area under the curve (AUC). How would you interpret the AUC.

   (e) [**5 Marks**] Predict the probabilities for the remaining 310 observations. Calculate the deciles for the predicted probabilities. Does the observed and the predicted probabilities differ for the deciles?

   (f) [**10 Marks**] For this problem you need to load the NHANES dataset but keeping all the rows of the data. You can use the following commands

   ```
   small.nhanes <- na.omit(NHANES[NHANES$SurveyYr=="2011_12"
   & NHANES$Age > 17,c(1,3,4,8:11,13,25,61)])
   ```

   Fit a mixed effects logistic regression. Only consider random intercept for subject ID. Use all the available predictors. Interpret the results.