**Rules**

1. Please include the questions in your submission. Keep the numbering and lettering the same or some other format that is very similar. The easiest way to do this is to type your answers into this document. I would greatly appreciate if your comments were in a different color such as blue.
2. You will need to determine the code to solve some of these problems.
3. Include any plots that were helpful to you in reaching your conclusion.
4. Include any R output that was helpful in reaching your conclusions.
5. Regarding points 5 and 6, your final exam should be clear, clean and presented in a professional manner. Do not provide R output or plots without explanation.
6. Related, and possibly repeating myself, do not provide conclusions that were based on R output or plots without also providing the R output and plots.

**Question 1** For this problem use the data set “sat data.csv”.

Load the data into RStudio:

sat\_data <-read.csv("https://sites.google.com/site/bsherwood/sat\_data.csv",header=T)

This data set contains several variables.

state: The state for the data.

expenditure: The average amount the state spends on each student, in thousands of dollars.

percent taking: The percentage of students in the state that take the SAT exam. total sat score: The average total SAT score.

1. Use a model to test the relationship between state expenditure and SAT scores. Clearly state your null and alternative hypothesis and conclusions.
2. Use a model to test the relationship between state expenditure and SAT scores while controlling for the percent of students taking the exam. Clearly state your null and alternative hypothesis and conclusions.
3. Explain the differences between the two models. If you got different results, which model do you trust and why?

**Question 2**  You have been transported back in time to 1985 and have been asked to create a model for wages for your company. The company’s policy is to provide equal pay regardless of race, gender, marital status or union membership. Use the wage data used in class to construct a model of wages for this company. Code to read in data set below.

wage\_data <- read.csv("https://sites.google.com/site/bsherwood/wage\_data.csv")

1. Explain the model you have fit and the steps you have taken to get this final model.
2. What would the model estimate the hourly wages for someone with the following characteristics: (occupation: Sales, sector: Manufacturing, union: no, education: 16, experience: 10, age: 45, gender: male, marital status: yes, race: Hispanic, south: yes).
3. Say you have an exceptional employee, for instance one you would rank at the .975 percentile. Would you recommend using the model’s wage for this employee? Why or why not?

**Question 3** You have data on annual income and years of education and fit a linear mean regression model to the data and got the following equation for the estimated average annual income given the years of education.

Annual Income = 20,000 + 2,000\*(Years of Education) + Error

1. In the data you had a person with 12 years of education and an annual income of $50,000. For the below two questions show work to demonstrate that you know how these values are calculated.
   1. Provide the fitted value for this observation.
   2. Provide the residual for this observation.
2. Correct the following statement. “I estimate that a person with 10 years of education will have an annual income of $40,000.''

**Question 4:**The pima data set contains data from 768 adult female Pima Indians leaving near Phoenix. The “faraway” library will need to be installed to access the data set. After the library has been installed use the following commands to access the data set and look at the first 6 entries.

library(faraway)

head(pima)

Use the following command to get a description of the data set and the variables.

help(pima)

1. Fit a model for the probability of getting a positive test given the other variables in the data set. Write down the estimated model.
2. Perform a hypothesis test of whether the probability of having diabetes changes with the number of pregnancies a women has after controlling for the other variables available in the model.
3. Provide and interpret the confidence interval for the pregnant coefficient from the model fit for part (a). Hint use the confint() function.
4. What does your model predict the probability of getting a positive test for someone who has been pregnant 6 times, has a glucose value of 148, a diastolic value of 72, a triceps value of 35, an insulin value of 0, a bmi of 33.6, a diabetes pedigree function of .627 and is 50 years old. (These are the attributes for the first person in the study, which can be accessed by pima[1,]. That information may or may not be helpful for you depending on how you approach this problem.)

**Question 5:** Say you have data on 100 houses with information on price, number of rooms and the size of the house. You want to compare two linear mean regression models for estimating the price

of a house.

Model 1: Only uses size of a house as predictor.

Model 2: Uses size of the house and number of the rooms as predictors.

Explain how you could use 10-folds cross validation to select the best model. Your explanation should show that you understand all steps of the cross-validation process.

**Question 6:** The supplementary material includes summary output for a linear mean regression model of food expenditure based on income. It also includes a residuals vs fitted plot from that model. The data set includes 234 families. You do **not** have to reproduce the regression output or residuals vs. fitted plot providedto you for this problem.

1. Explain how to use the bootstrap method to get an estimate of the standard error for the income coefficient. Your explanation should make it clear that you understand all the steps of the bootstrap procedure.
2. Why would you want to use the bootstrap method to calculate standard errors for this data?
3. Say the bootstrap method got an estimated standard error of “.1” for the income coefficient. What would the bootstrap t-value be?

**Question 7:** A non-profit has data on years of education and donations from 101 contributors. Do you believe that this data shows higher education is **associated** with larger donations? Explain. Below is the code for reading the data into RStudio.

np\_data <-read.csv("https://sites.google.com/site/bsherwood/nonProfitData.csv",header=T)

**Question 8:** The data set “wages” in the package “TSA” contains monthly data on average hourly wages in the apparel industry. The data can be accessed by doing the following

library(TSA)

data(wages)

A description of the data can be found by doing

help(wages)

Some (potentially) helpful functions for this problem are

plot() - will plot a time series

diff() – will create a differenced time series

acf() – will create an autocorrelation plot

pacf() – will create a partial autocorrelation plot

Note this is monthly data.

1. Create a time series plot of the raw wages data and the monthly change in wages.
   1. Do you believe average wages is a stationary time series? Explain.
   2. Do you believe that monthly difference in average wages is a stationary time series? Explain.
2. What type of AR(p) model would you recommend fitting to this data? Specifically
   1. Should the model be fit to the raw or differenced data? Explain.
   2. Should it be a seasonal or non-seasonal model? Explain.
   3. What choice of p do you recommend? Explain.