Statistics 3080
Homework 6
Due: Monday, December 4

Complete the following problems in a commented R file. Include any output requested as a comment following your code. Include any plots as a single PDF file.

Problem 1 (60 points): This problem will use the data set `geyser` in the `MASS` package. This dataset contains eruption information from the Old Faithful geyser in Yellowstone National Park, Wyoming, from August 1st to August 15th of 1985. Each row records the waiting time until the next eruption and the length of that next eruption. A group is interested in understanding the relationship between length of the previous eruption and its effect on how long a person then has to wait for the next eruption.

(a) Rearrange the data so that each row contains the eruption length and the waiting time following it. (Hint: In the current format, the first waiting time is not useful nor is the last eruption time.)

(b) Create a scatterplot of the data with the explanatory variable on the x-axis and the response variable on the y-axis. (Note: Old Faithful has two chambers that hold water. Shorter eruptions empty only the top chamber while longer eruptions empty both chambers.)

(c) Run the regression model and report the estimated coefficients.

(d) Test whether the model is significant. Explain the test you use and its outcome.

(e) Determine the proportion of variation in the waiting time explained by the model.

(f) Generate the plots needed to evaluate the regression assumptions. Indicate whether each assumption holds or not and explain how you came to each determination.

(g) Estimate the longest times that visitors wait on average after eruptions of 2.5 and 3.5 minutes.

(h) Predict the longest times that visitors wait after eruptions of 2.5 and 3.5 minutes.

Problem 2 (40 points): Recent research has shown that professors are among the most stressed workers. A researcher wants to know exactly what it was about being a lecturer that created this stress and subsequent burnout. She administered several questionnaires to 467 randomly selected professors and measured:

- Burnout (burnt out or not)
- Perceived control (high score = low perceived control)
- Coping style (high score = low ability to cope with stress)
- Stress from teaching (high score = teaching creates a lot of stress for the person)
- Stress from research (high score = research creates a lot of stress for the person)
- Stress from advising (high score = advising creates a lot of stress for the person)

The collected data are in the file *burnout.txt*.

(a) Using the data the researcher has collected and backward model selection, determine an appropriate model to predict burnout. What variables are significant to predicting burnout?

(b) Are all the variables in the model significant? Explain the test you use and how the outcome is determined.

(c) Determine the odds ratio for each variable in the model and give an interpretation for each one in non-statistical terms.

(d) Is there any multicollinearity in the model? Explain how you made your determination.

(e) Are there any outliers or influential points in the model? Explain how you made your determination.