Homework 9

For this homework you will create an R Markdown document that outputs to a PDF and you will upload both the .Rmd and .pdf files to wolfware. Be sure to put your name in the title of the document.

The purpose of this homework is to get practice with fitting and predicting with multiple linear regression models and, similarly, with logistic regression models.

Application: We will use a dataset from the UCI Machine Learning Repository. This data set is about wine quality. You can learn more about the data here.

The data description describes the following variables:

Input variables (based on physicochemical tests)

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10) This will be the response for the multiple linear regression modeling portion. Although it is not a continuous response, our usual multiple linear regression methods can still be used for prediction purposes.

13 - **Create a new output variable that is binary.** If the quality is 6 or higher, assign the variable 1. If the quality is less than 6, assign 0. This will be your response for the logistic regression modeling portion.

I've combined the red and white data and split that dataset into a training and test set.
- You should do all of your model fitting on the training set only! You'll be using cross-validation (details below).
- Only once you've selected your model should you then see how it performs on the test set.

To Do:

Create a document that goes through your process of reading the data, manipulating/creating any variables, and fitting and choosing a final model for both the mulitple linear regression modeling (with quality as your response) and the logistic regression modeling (with your binary variable as the response).

Normally, you'd use some scientific reasoning to know what candidate models to consider. As we don't necessarily have that, you can figure out candidate models however you choose (include interactions, quadratic terms, etc.). When including interactions and polynomial terms, we often standardize the variables (center and scale each observation), but that is up to you. When just exploring models you might want to standardize anyway (or maybe do more - here is an interesting article - Gelmen is a very well known statistician).

- You should compare at least five candidate models for both the MLR and logistic regression modeling.

- Use 5 or 10 fold cross-validation to select your model (again, using the training set only!)
- For the final model you come up with (for both MLR and logistic), see how well it performs on the test set using squared error loss (RMSE) for MLR and accuracy for the logistic modeling.
- Report these values at the end of your report (along with a description of your best model in each case).