Assignment 2
STAT 705, Spring 2020

This is a graded assignment, worth 25 points.

The provided dataset was collected from red wine manufacturers. It provides a variety of measurements taken on each of several wine specimens and an average quality score assigned to each specimen by experienced wine tasters. The goal is estimate the quality of the wine based on the other measurements. The variables are

ID	a numeric identifier for the wine specimen
FixAcid	fixed acidity (acid that does not readily evaporate)
VolAcid	volatile acidity (high levels produce a vinegar taste)
CitricAcid	citric acid (can produce 'freshness' flavor)
Sugar	residual sugar (produces 'sweet' flavor)
Chloride	sodium chloride (salt)
FreeSulfur	free sulfur dioxide (prevents oxidation)
TotalSulfur	total sulfur dioxide (high levels can produce undesirable smell)
Density	density of the wine (as compared to the density of water)
pH	on a scale from 0 (very acidic) to 14 (very basic)
Sulphate	sulphates added as antimicrobial and antioxidant agent
Alcohol	percent of alcohol in the wine
Quality	average quality of the wine assigned by experienced wine tasters

(Note: The units for these measurements are not particularly important for this assignment. In case you are interested, the three acids, sugar, chlorides and sulphates are measured in grams per cubic decimeter. Free sulfur and total sulfur are measured in milligrams per cubic decimeter, and the density is measured in grams per cubic meter.)

For all of this assignment, you will exclude the variable ID. It is just an identifier, not a potential predictor.

As with the first assignment, you will write the SAS code to answer the following questions and then enter your answers directly into Canvas. In addition to answering these questions, you will need to to upload your SAS program file. This is the file with the "sas" extension, and it contains the statements that appear on the Code tab in SAS Studio. Please do <u>not</u> upload your SAS output file, the SAS log file or the HTML file. I need your original program file so that I can execute it and generate your output for myself.

The questions are on the next page.

We want to develop a multiple regression model that will estimate the quality of the wine (i.e., the variable Quality) using the other variables in the dataset (excluding ID).

1. Generate the correlation matrix for all the variables (except ID). Based solely on the correlation matrix, identify the apparent "best" predictor for Quality. Briefly explain why you chose this variable.
2. Fit a model using all the predictors in the dataset. (This is called the 'full model'.) Based on the information in the Parameter Estimates table, would it be reasonable to remove FixAcid, CitricAcid, FreeSulfur and Sulphate from this model? Why or why not?
3. Does the full model have a problem with multicollinearity? Explain.
4. Use the method of forward selection to select the predictor variables for this model. Identify the variables chosen by this method.
5. Use the method of backward elimination to select the predictor variables for this model. Identify the variables chosen by this method.
6. Use the stepwise method to select the predictor variables for this model. Identify the variables chosen by this method.
7. During the stepwise method, one variable entered the model and was later removed from the model. Identify this variable, and briefly explain why it was removed.
8. Using the model generated by the stepwise method, interpret the slope on the variable Alcohol. (This needs to be one complete sentence.)

For the remaining questions, focus on the model generated by forward selection (in Question 4) and the model generated by backward elimination (in Question 5). For these questions, do not consider any other model besides these two.

9. Evaluate the assumptions for these two models. Does either model appear to violate the assumptions? Is there anything about the assumptions that would make you prefer one model over the other?
10. Evaluate the goodness of fit the these two models. Is there anything about the goodness of fit that would make you prefer one model over the other?
11. Evaluate the outliers for each of these two models. Is there anything that would make you prefer one model over the other?
12. Perform a nested model F test to compare these two models. Provide the test statistic and the p-value of this test. Does this test make you prefer one model over the other?
13. If you were going to do additional analysis on this dataset, which model would you use: the model generated by forward selection or the model generated by backward elimination?