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.
   1. Alcohol. It had the highest correlation of 0.51 and was found to be significant <.0001.
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?
   1. The coefficients produced from the regression show all of the mentioned features absolute values are < 0.1, whereas everyone else is > 0.1 so this is actually a reasonable breakpoint to filter out some noisy variables.
3. Does the full model have a problem with multicollinearity? Explain.
   1. I believe there is a moderate problem with multicollinearity. The maximum correlation between independent and dependent is 0.57, but we see multiple locations where this is close or higher (FixAcid – pH, for example with -0.67).
4. Use the method of forward selection to select the predictor variables for this model. Identify the variables chosen by this method.
   1. Alcohol, VolAcid, FixAcid, Sugar, Chloride, TotalSulfer, Density
5. Use the method of backward elimination to select the predictor variables for this model. Identify the variables chosen by this method.
   1. FixAcid, CitricAcid, Sulphate, FreeSulfer, pH, Chloride
6. Use the stepwise method to select the predictor variables for this model. Identify the variables chosen by this method.
7. Alcohol, VolAcid, FixAcid, Sugar, Chloride, Density, TotalSulfer
8. 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.
   1. The variable removed was FixAcid. During the stepwise regression, after adding chloride just one iteration before, enough information was contained in chloride with the set of predictors than with FixAcid, making it uninformative to the model.
9. Using the model generated by the stepwise method, interpret the slope on the variable Alcohol. (This needs to be one complete sentence.). For every percent increase in the wine, we expect to see an on average increase of 0.3 units of quality.
10. 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?
    1. The standard assumptions of both models appear to be met with residuals centered around 0 and NIID properties met. I prefer backward selection because of the compounding bias that is introduced by entering in the first variable (and preceding). I suppose this could equalize, but it seems if you choose a bad starter variable, then all feature importance’s are contingent on having information gained from that. I like the idea of capturing the entire space in backward selection, then whittling it down to a smaller “cleaner” space with everyone’s (all the variables) opinion in the mix the whole time. I still prefer stepwise over both. 😊
11. 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?
    1. Not really, they both got a 0.43 Adjusted R^2 value, and both beat the original model’s 0.41 Adjusted R^2 value.
12. Evaluate the outliers for each of these two models. Is there anything that would make you prefer one model over the other?
    1. It appears that forward selection indicated more outliers. This could be very beneficial.
13. 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?
    1. Both test statistics I got failed to reject the null, providing no evidence of favor of the full model. This does not give me any preferences.
14. 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?
    1. Forward selection is biased from my point of view, so I feel better with my own work if I use backward selection. It does seem like forward selection is better at detecting outliers. So if the task was to predict new values, the forward selection might be better because of the fit, if the task is just analysis, I would prefer the backward selection.