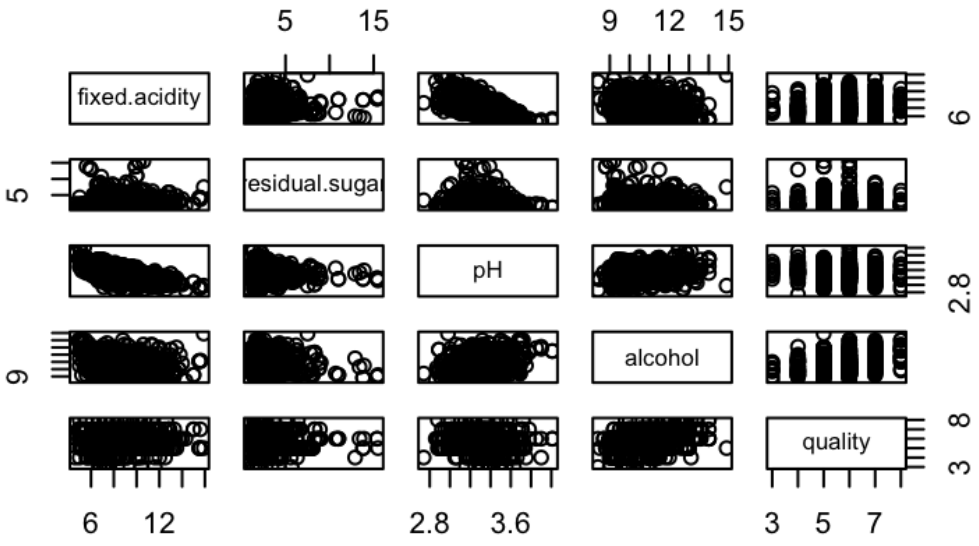
Reuben Lopez
10/12/20

Multiple Regression Analysis

In determining the quality of a red wine, the variables of interest will be fixed acidity, residual sugar, PH level, alcohol level. This part of the work will consist of the findings of a multiple linear regression used to determine the quality of wine using the variables described earlier.

Before any regression model is ran, it is important to check the variables for any correlation as this can impact the accuracy of the model significantly.It is important to note that no interaction variables were used in this model. Upon running a pairs plot on the variables of interest, I expected to see some high correlation between some variables.I picked these variables for this reason. The results were surprising, there did not seem to be much collinearity besides PH level and fixed acidity which yielded a negative correlation of -0.6829. This makes sense as our researched earlier suggested that these two variables could have some correlation.



Upon fitting the model, we see that of the 4 explanatory variables, 3 are significant. The fixed acidity is significant at the 1% significance level, ph and alcohol are also significant at .001. Having less than one in a thousand chance of being wrong.This model also had an R-

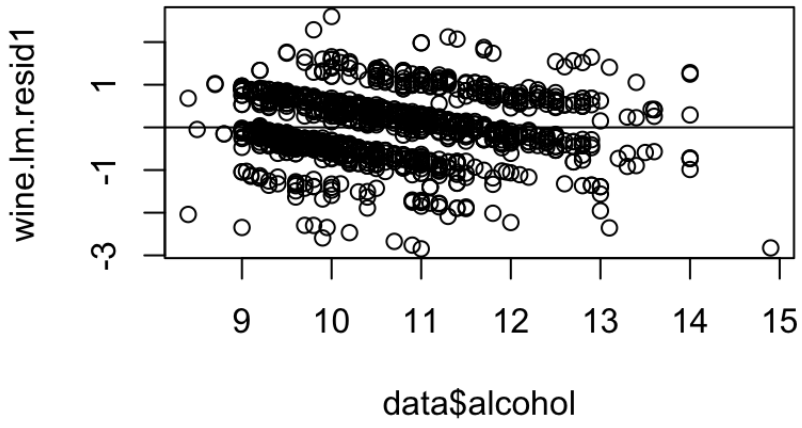
squared of .2565 which is not good at all as well as an adjusted R-squared of 0.2546. This means that our model only explains only 25% of the variability of the data.

For our simpler model, the variables that were eliminated were the ph, and residual sugars variables . From our research earlier, I was inclined to believe that by dropping these two variables the model would improve or at minimum, remain relatively the same. The new models resulted in a R-squared value of 0.2504 and an adjusted R-squared value of 0.2494. This does not guarantee that the model would describe the population well. In fact this model could be deemed useless

If I had to report 1 of the models to my boss I would report the smaller model. The reason for this is because it maintains similar results as the larger model while being more of a simpler model. This means that some of the variables in the larger model were not necessary.

Once the PH and residual sugar variables were dropped, a change that did occur was the significant of the fixed acidity. It went from being significant at a 1% level to being significant at a .01% level meaning there is a less than one in a thousand chance of being wrong.

When plotting the residuals to view which variables contributed to the fit, the residual plot with the alcohol variable exhibited a pattern. Thus proving not to be random.



In summary, This model proved to be insignificant.The originals model produced a very low R-squared, and the improved model also reported a low R-squared.I hoed by including the variables that I did for the original model that there would be an improvement in the model once some variables were dropped and model became simpler. This was not the case. For this specific model we were not able to accurately predict the quality rating (1-10) of a wine based on the chosen variables.