Wine Alcohol Content Analysis

An Analysis on the Factors that Affect Alcohol Content of White and Red Wine

Written By

Ishana Narayanan (Section: T 3 PM)
Madeline Li (Section: T 3 PM)

1 Introduction

In this project, we explored the differences in red and white wines, namely which factors impact the alcohol content of each respective wine. To do this, we studied two datasets with the same attributes, the only difference being the type of wine. The red wine dataset has about 1600 entries whereas the white wine dataset has 4898 observations. Both datasets contain the following attributes:

Continuous Numeric Attributes:

- *Fixed acidity* (g/dm^3): mainly comprised of tartaric acid and encompasses most of the acidity in wine
- *Volatile acidity* (g/dm^3): the amount of acetic acid in wine which gives wine a vinegar taste
- *Citric acid* (g/dm^3): weak organic acid responsible for preservation
- *Residual sugar* (g/dm^3): the amount of sugar remaining after alcoholic fermentation finishes
- *Chlorides* (g/dm^3): the amount of salinity in wine
- *Free sulfur dioxide* (g/dm^3): the portion of sulfur dioxide that is not bound to other chemicals
- *Total sulfur dioxide* (g/dm^3): the portion of sulfur dioxide that is free in the wine plus the portion that is bound to other chemicals
- *Density* (g/cm^3): density of the wine
- *pH* (*scale of 0 to 14*): describes the acidity of the wine using the pH scale where 0 is most acidic and 14 is most basic
- *Sulphates* (g/dm^3): range of sulfur compounds that are a natural byproduct of the alcoholic fermentation process that work to preserve against yeast and bacteria
- *Alcohol* (*% by volume*): percent of alcohol content

Ordinal Categorical Attributes:

- *Quality* (*score from 0 to 10*): based on sensory data

We used a first-order linear model with ten quantitative predictors and one qualitative predictor to determine the alcohol percentage in both white and red wine. To complete our research, we performed both stepwise and best subsets regression, analyzed residuals, verified LINE conditions, and examined outliers to determine the best possible model for both white and red wine data. Lastly, we compared our findings and drew conclusions about the relationship between both these wines.

2 Questions of Research

1. Which factors affect alcoholic content for each wine type?

   - We built separate regression models to determine the most relevant predictors
   - We then performed a general linear $F-test$ to test the significance of each predictor

2. Is the quality and density of wine significant in predicting the amount of alcohol in a particular wine type?

   - We performed a general linear $F-test$ with quality and density equal to 0 and test for their significance

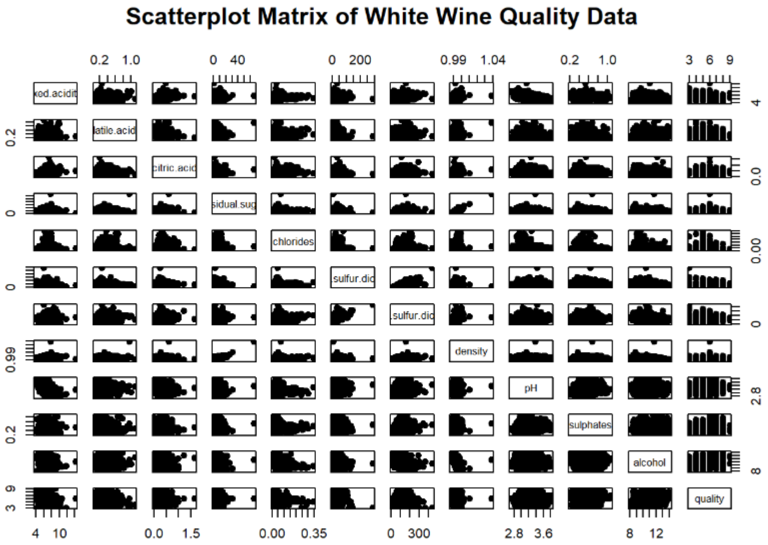
3. What is the expected amount of alcohol in wine when all other factors are average?

   - We calculated a 95% confidence interval with the above specified parameters

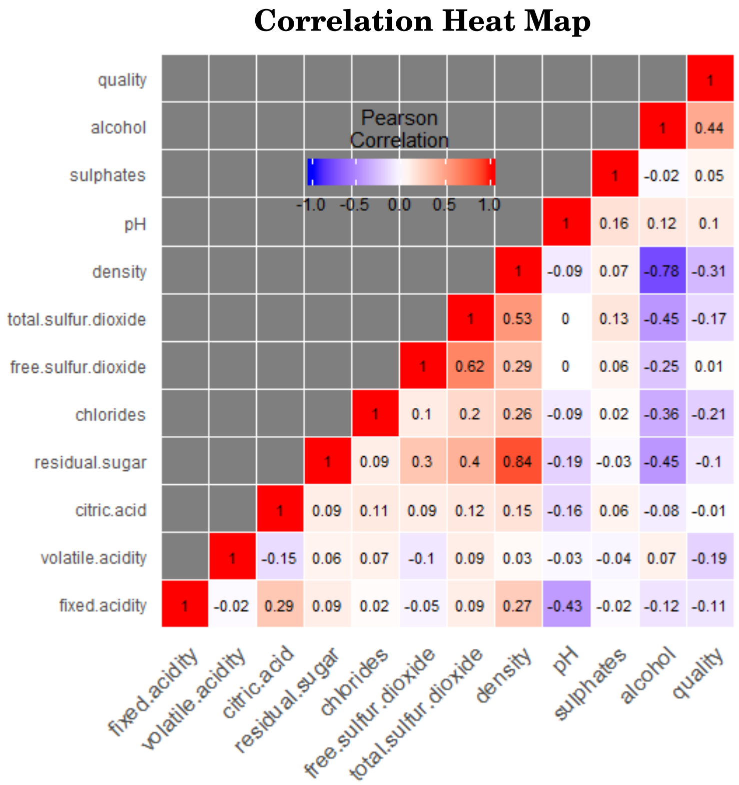
3 White Wine Data Analysis

3.1 Exploratory Data Analysis

To better understand the data, first we graphed a scatterplot matrix to determine any obvious relationships between the variables.



From the scatterplot matrix above, we saw that there is a strong positive correlation between residual sugar and density, and a strong negative correlation between alcohol and density. To further examine variable correlation, we calculated the Pearson correlation coefficients for each parameter.

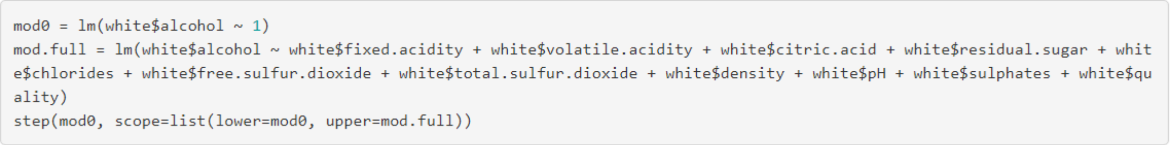


Looking at the following correlation heat map, we saw the strongest positive correlation between density and residual sugar with a value 0.84 and the strongest negative correlation between alcohol and density with a value of -0.78. Moreover, there are moderate positive correlations between density and total sulfur dioxide, free sulfur dioxide and total sulfur dioxide, and residual sugar and total sulfur dioxide with correlation coefficients 0.53, 0.62, and 0.40 respectively. On the other hand, alcohol and residual sugar, alcohol and total sulfur dioxide, and pH and fixed acidity all have moderately negative relationships with correlation values -0.45, -0.45, and -0.43 respectively. From this information, we determined that there might be possible collinearity between residual sugar and density due to the high correlation between these variables.

3.2 Regression Model

3.2.1 Stepwise Regression

To determine which predictors are relevant in predicting alcohol content, we fit a multiple linear regression model to our data. We used the stepwise regression with AIC to identify the order of significant predictors to include in our model.



```
mod0 = lm(white$alcohol ~ 1)
mod.full = lm(white$alcohol ~ white$fixed.acidity + white$volatile.acidity + white$citric.acid + white$residual.sugar + whit
e$chlorides + white$free.sulfur.dioxide + white$total.sulfur.dioxide + white$density + white$pH + white$sulphates + white$qu
ality)
step(mod0, scope=list(lower=mod0, upper=mod.full))
```

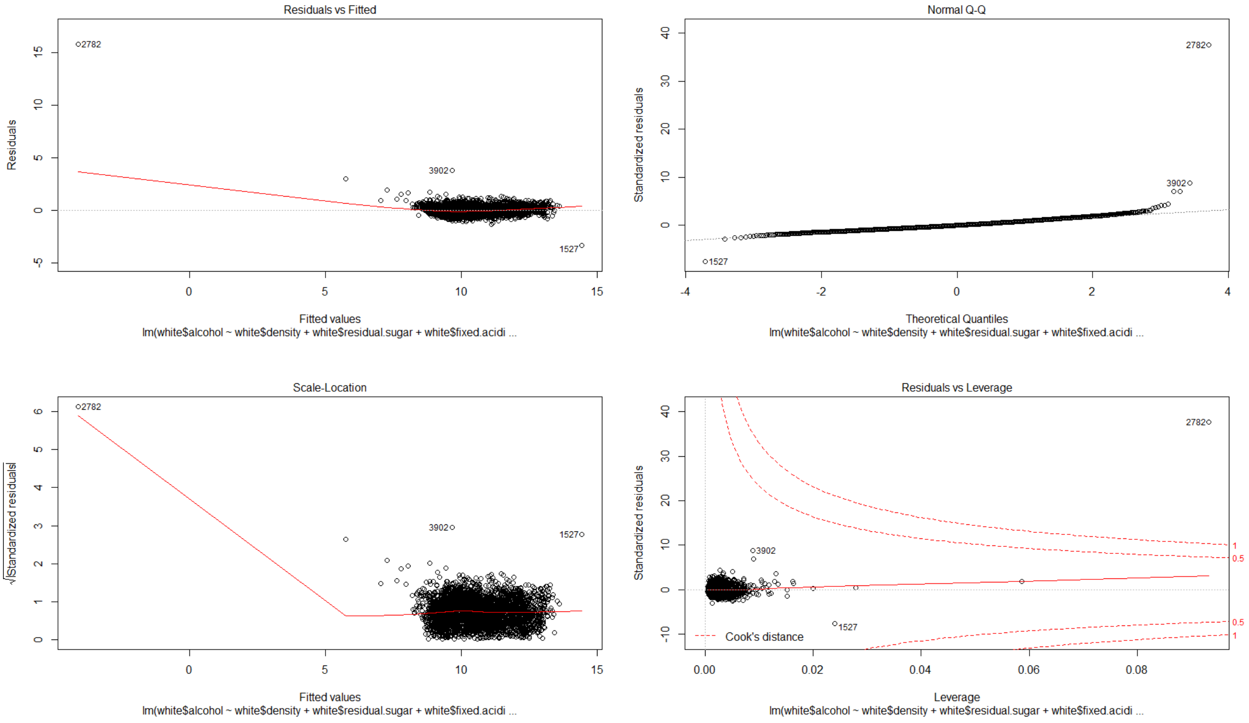
After performing stepwise regression, we found that the best model is one with ten of the eleven predictors in the following order:

$$\begin{gathered} alcohol \sim density + residual\ sugar + fixed\ acidity + pH + sulphates + volatile\ acidity + \\ quality + free\ sulfur\ dioxide + citric\ acid + total\ sulfur\ dioxide \end{gathered}$$

After fitting our mostly full model, we saw that 91.86% (adjusted R^2 value) of the variance in the alcohol content of white wine is explained by the predictors, indicating a significant positive relationship between the predictors and the response.

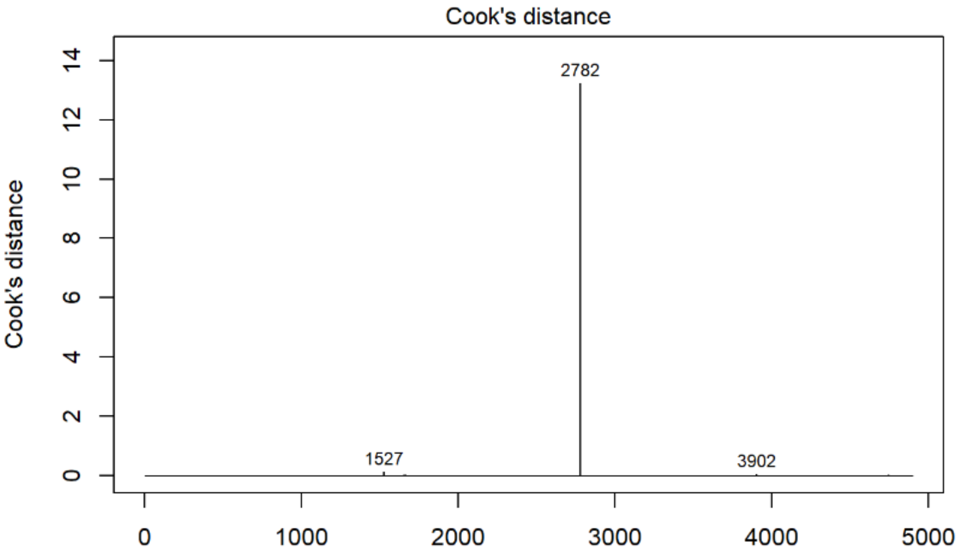
3.2.2 Residual Analysis

To validate this model, we plotted and analyzed the error residuals to verify LINE conditions. According to the Residuals vs. Fit plot, we saw that there are extreme outliers such as 2782 that impact the spread of the data points. The residuals closely follow a linear pattern in the Normal Q-Q plot, indicating that the errors are normally distributed. However, to better understand the distribution of error residuals, we further examined potential outliers and influential data points and their effect on the data.



3.2.3 Outliers and Influential Points

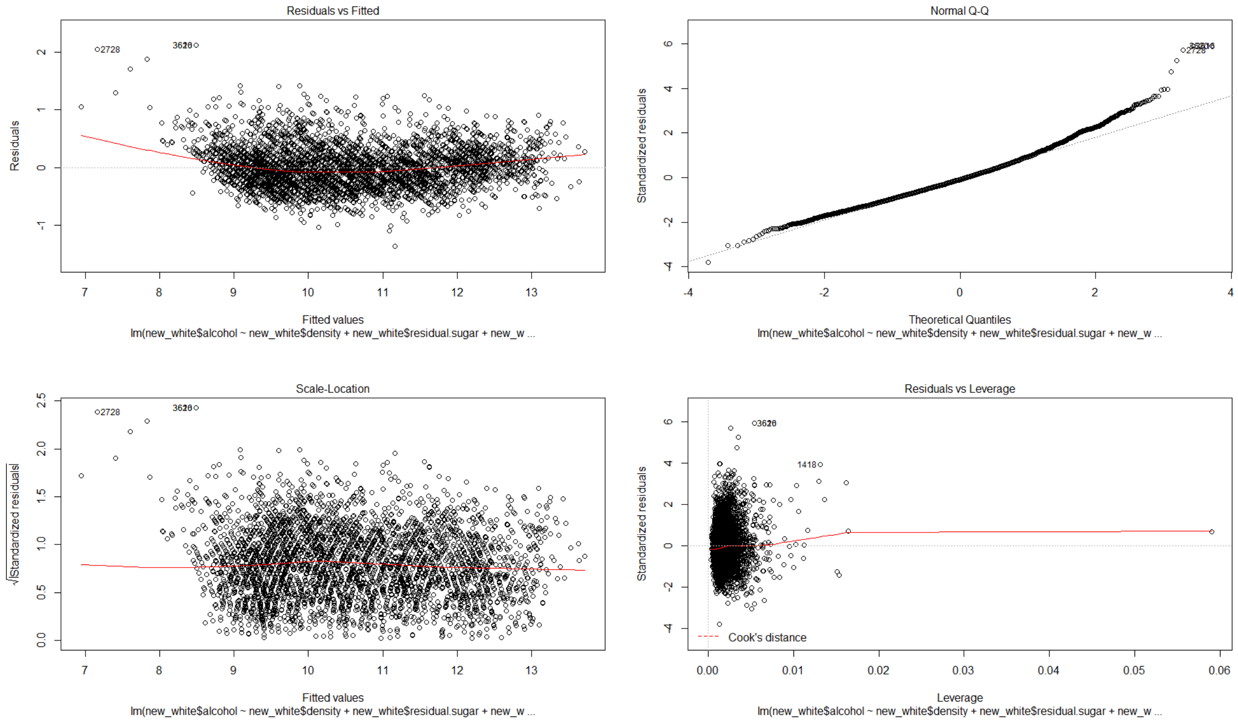
We used the internally studentized residuals and externally studentized residuals to methodologically determine outliers and influential points. The results show that there are five overlapping data points which are considered both an outlier and influential. Using Cook's distance, we determined the extend of the influence and decided to remove the five data points which are classified as both outliers and influential so that their values do not skew our results.



3.2.4 Residual Analysis Pt. 2

With extreme outliers removed, we reanalyzed the residual plots and checked LINE conditions.

Since the residuals in the Residuals vs. Fit plot are evenly spread about x=0 without fanning, we concluded that the error variances are equal. Moreover, the even spread of residuals indicates that a linear model is a good fit. In the Normal Q-Q plot, the residuals are relatively following a linear pattern, so the errors are normally distributed. However, the curve at the top was concerning, so we investigated further to ensure that the model is not violating the normality LINE condition. Lastly, because we did not have a Residuals vs. Order plot, we assumed that the errors are independent.



3.2.5 Best Subsets Regression

We also used the best subsets regression technique to determine the best model for our data.

```
fit_subset = lm(white$alcohol ~ white$fixed.acidity + white$volatile.acidity + white$residual.sugar + white$free.sulfur.diox
ide + white$density + white$pH + white$sulphates + white$quality)

plot(fit_subset)
```

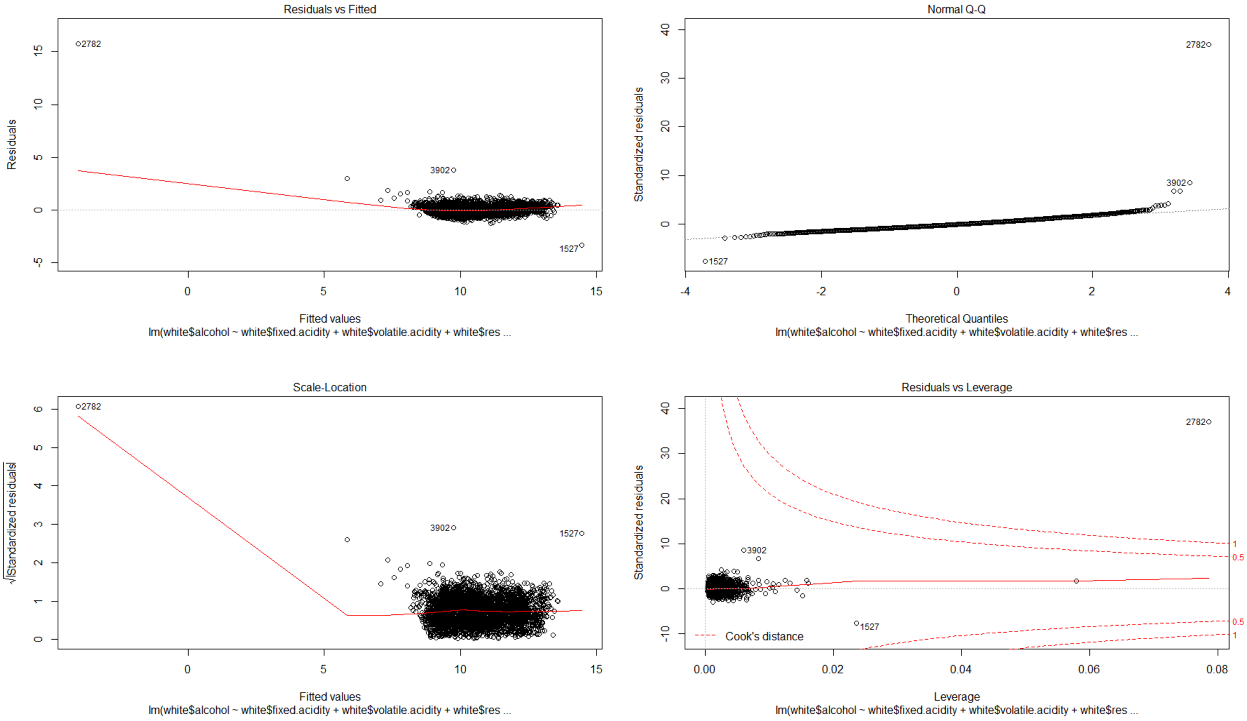
To determine the appropriate number of predictors, we analyzed the corresponding adjusted R^2 values. Since the largest adjusted R^2 value of 0.8705191 corresponds with the eighth model, the optimal model is one which contains eight predictors. We then examined the $summary.mod\$which$ graph to determine the most effective predictors for a model with eight parameters. From this information, we determined that the best model is one with the following eight predictors:

$$alcohol \sim fixed\ acidity + volatile\ acidity + residual\ sugar + free\ sulfur\ dioxide + density + pH + sulphates + quality$$

After fitting the best subsets model, we saw that 91.67% (adjusted R^2 value) of the variance in the alcohol content of white wine is explained by the predictors, indicating a significant positive relationship between the predictors and the response.

3.2.6 Residual Analysis

Like before, to validate this model, we plotted and analyzed the error residuals and verified LINE conditions.



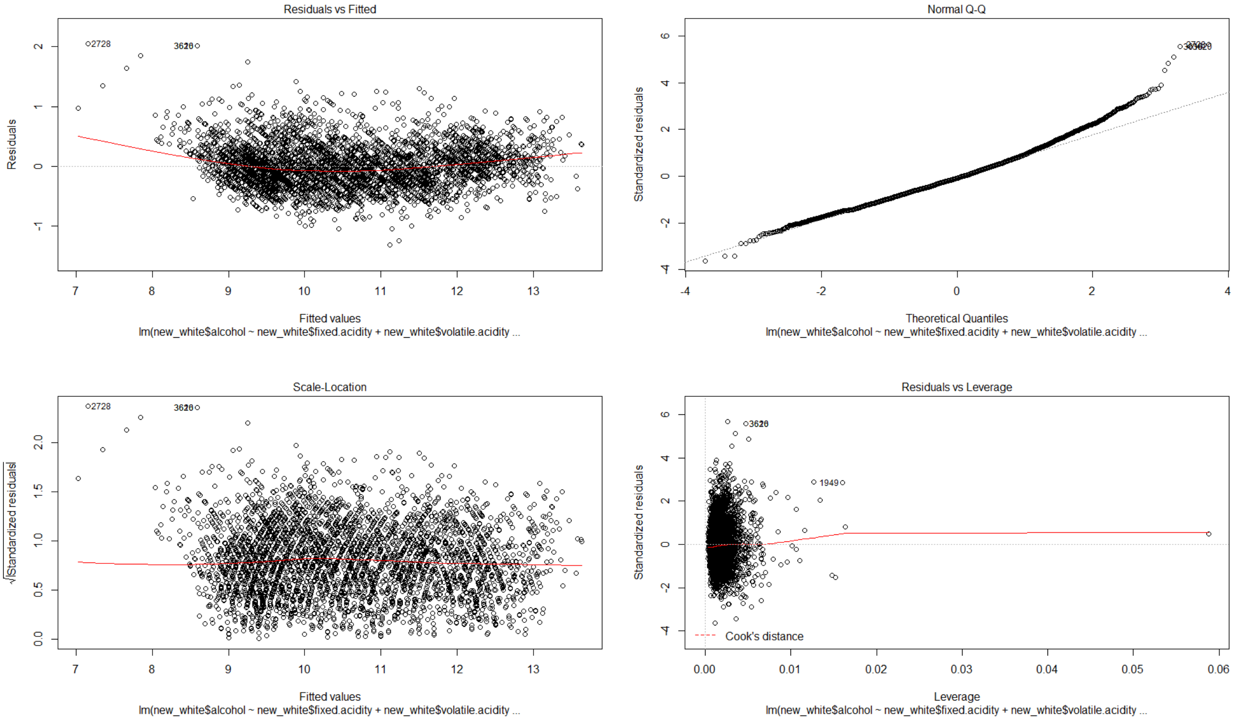
We noticed that the residual plots for the best subsets model closely resemble that of the original stepwise regression. Like with stepwise regression, we concluded that it was necessary to further investigate potential outliers and influential data points.

3.2.7 Outliers and Influential Points

Again, we used internally and externally studentized residuals to determine outliers and influential data points. We found that both the stepwise and best subsets regression have the same outliers and influential data points, so we removed these five data points from our model.

3.2.8 Residual Analysis Pt. 2

Even with the outliers removed, the residual plots for best subsets regression closely resemble that of stepwise regression. The residuals are evenly spread about x=0 without fanning in the Residuals vs. Fit plot, so error variances are equal and a linear model is appropriate. Similar to the stepwise Normal Q-Q plot, the line curves at the top, so we investigated further to ensure that this model does not violate the normality assumption. Lastly, because we did not have a Residuals vs. Order plot, we can assume that the errors are independent.



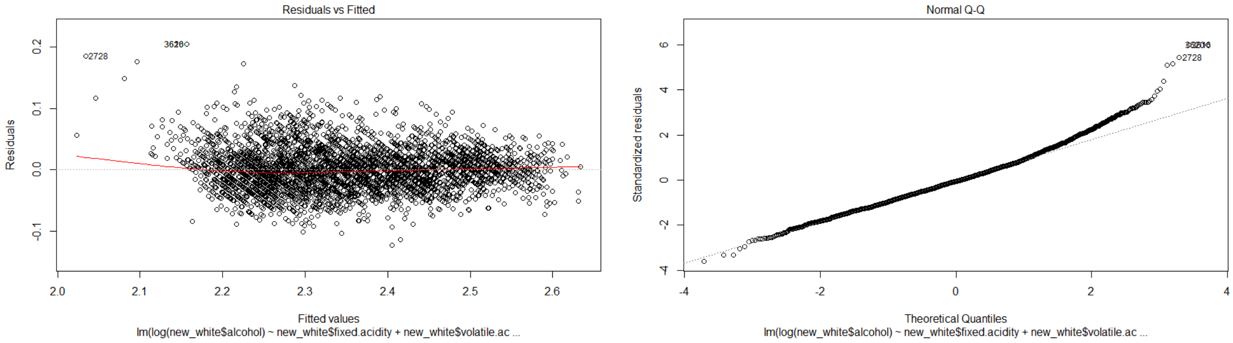
After performing residual analysis on both these models, we saw that there are no striking differences that set these models apart. Both models have a high adjusted R^2 value, indicating that most of the variation in the amount of alcohol content is explained by the predictors. Furthermore, both models do not overtly violate any LINE conditions. The only concern is the curvature at the top of the Normal Q-Q plot. The only significant difference between these two models is the number of parameters: the stepwise regression model has 10 parameters whereas the best subsets model has only 8 parameters. Based on this information, we should choose the model with greater simplicity. Therefore, the best model is the best subsets model with the 8 predictors being fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, density, pH, sulphates, and quality.

3.3 Transformations

Since the residuals possibly violated the normality assumption, we performed transformations on the response variable and determine if a transformed model is a better ft for this data.

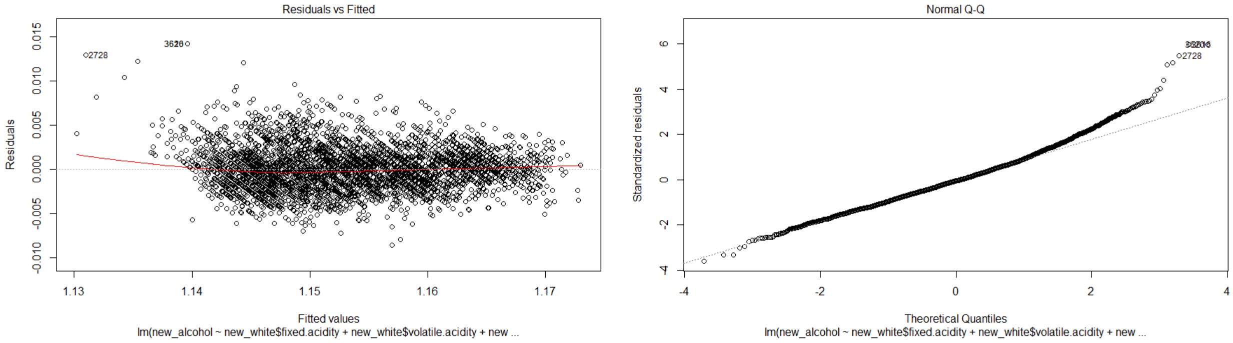
3.3.1 Logarithmic Transformation

First, we applied a logarithmic transformation on the percentage of alcohol.



3.3.2 Boxcox Transformation

Second, we applied a Boxcox transformation on the percentage of alcohol and found λ to be 0.061.



According to the Residuals vs. Fit plot, the residuals are evenly spread with no fanning, so we concluded that the error terms have constant variance and that a linear fit is an appropriate model. Similarly, the residuals follow a linear pattern, so we assumed that the error terms follow a normal distribution. However, note that these residual plots look almost identical to the residual plots for the original best subsets model. This is a clear indication that a transformation is not necessary and that our initial model already meets the LINE conditions criteria.

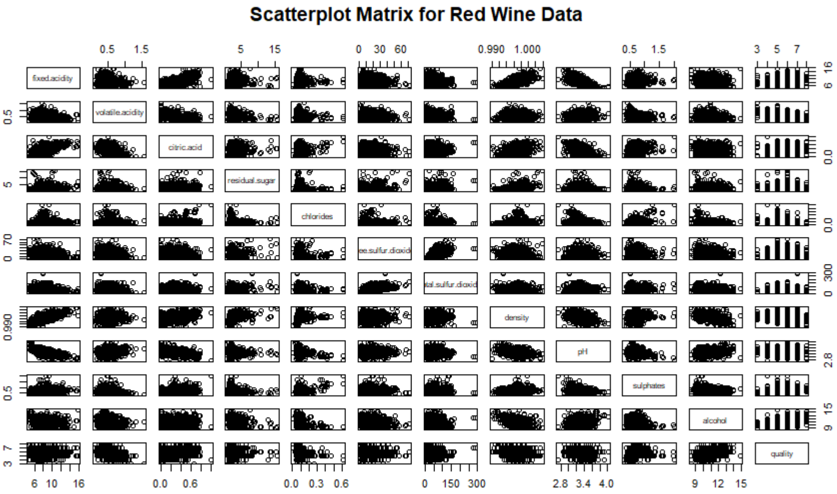
3.4 Final Model

$$alcohol_i = \beta_0 + \beta_1 fixed\ acidity_i + \beta_2 volatile\ acidity_i + \beta_3 residual\ sugar_i + \\ \beta_4 free\ sulfur\ dioxide_i + \beta_5 density_i + \beta_6 pH_i + \beta_7 sulphates_i + \beta_8 quality_i$$

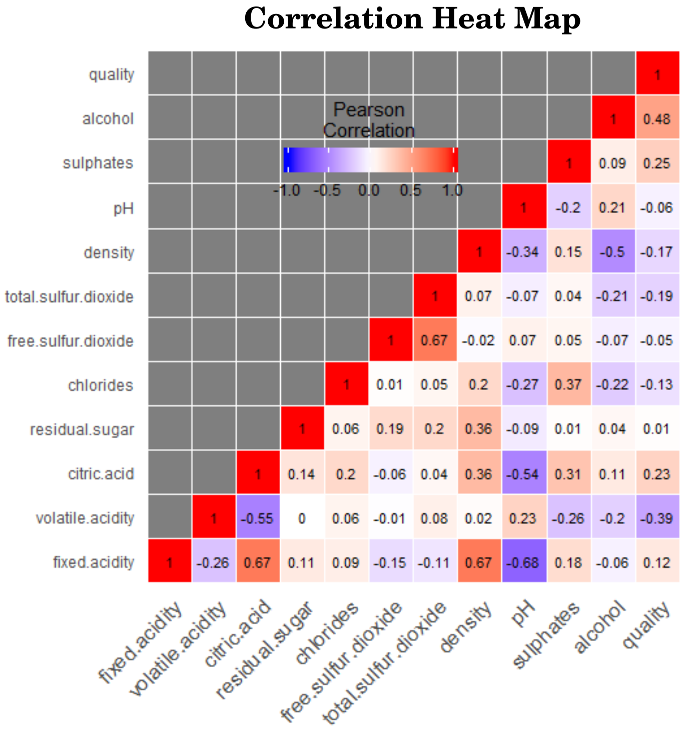
4 Red Wine Data Analysis

4.1 Exploratory Data Analysis

To get a general idea of what predictors would have the greatest effect on the response, we created a scatter plot matrix. However, we found it hard to draw meaningful conclusions due to the sheer volume of data.



Next, we examined the correlation matrix. None of the correlations with alcohol appeared significant aside from density which has a correlation of -0.496 and quality which has a correlation of 0.476. According to the correlation heat map below, there is a moderately positive correlation between fixed acidity and citric acid as well as fixed acidity and density. The strongest negative correlation is between variables fixed acidity and pH with correlation value -0.68.



	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
quality												1
alcohol											1	0.48
sulphates										1	0.09	0.25
pH									1	-0.2	0.21	-0.06
density								1	-0.34	0.15	-0.5	-0.17
total.sulfur.dioxide							1	0.07	-0.07	0.04	-0.21	-0.19
free.sulfur.dioxide						1	0.67	-0.02	0.07	0.05	-0.07	-0.05
chlorides					1	0.01	0.05	0.2	-0.27	0.37	-0.22	-0.13
residual.sugar				1	0.06	0.19	0.2	0.36	-0.09	0.01	0.04	0.01
citric.acid			1	0.14	0.2	-0.06	0.04	0.36	-0.54	0.31	0.11	0.23
volatile.acidity		1	-0.55	0	0.06	-0.01	0.08	0.02	0.23	-0.26	-0.2	-0.39
fixed.acidity	1	-0.26	0.67	0.11	0.09	-0.15	-0.11	0.67	-0.68	0.18	-0.06	0.12

4.2 Regression Model

4.2.1 Stepwise Regression

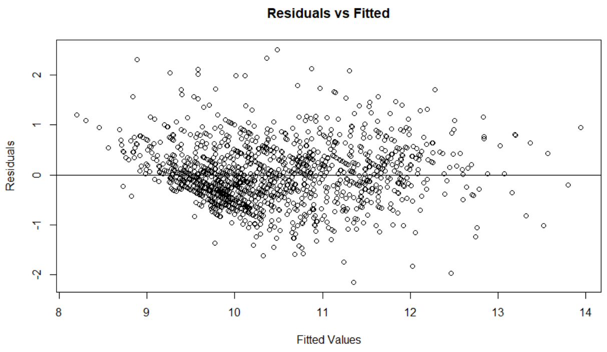
Like with the white wine data, we used stepwise regression with AIC to find our first model. Using this method, we found that the best model contained all eleven predictors in the following order:

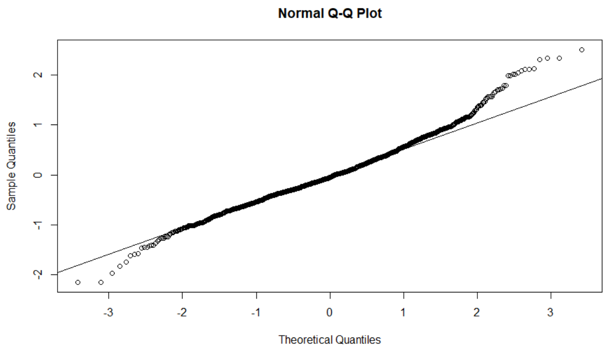
$$\begin{gathered} alcohol \sim density + quality + fixed\ acidity + pH + residual\ sugar + sulphates + \\ free\ sulfur\ dioxide + citric\ acid + volatile\ acidity + chlorides + total\ sulfur\ dioxide \end{gathered}$$

After fitting the best subsets model, we saw that 69.13% (adjusted R^2 value) of the variance in the alcohol content of white wine is explained by the predictors, which is significantly lower than what was found for white wine.

4.2.2 Residual Analysis

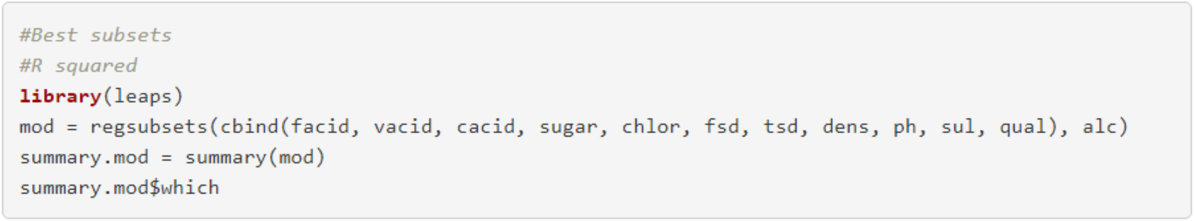
According to the Residuals vs. Fit plot, the residuals appear well-behaved. The points form a horizontal band which fulfill the assumption of equality of variance of error terms. The points also bounce around the zero line randomly satisfying the linearity assumption. However, there is one point in the lower right hand corner that might be an outlier. The Normal Q-Q plot plot looks decent, but there is a little deviation away from the line at the lower left hand corner and the upper right hand corner. Since we did not have a Residuals vs. Order plot, we assumed that the errors are independent. Thus, we concluded that this model meets all LINE conditions criteria.





4.2.3 Best Subsets Regression

Since the stepwise regression model contains all predictors, we also performed best subsets regression in hopes of finding a simplified model.



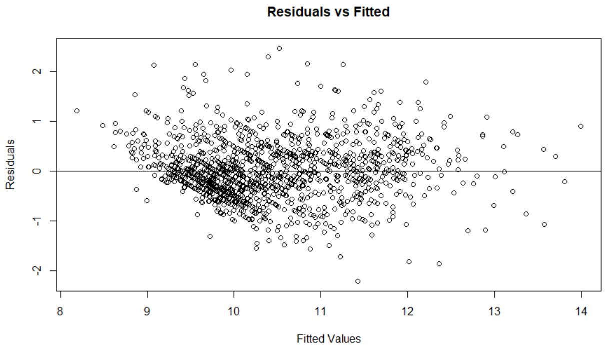
```
#Best subsets
#R squared
library(leaps)
mod = regsubsets(cbind(facid, vacid, cacid, sugar, chlor, fsd, tsd, dens, ph, sul, qual), alc)
summary.mod = summary(mod)
summary.mod$which
```

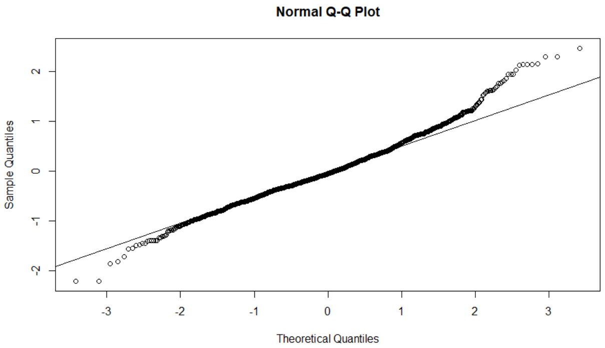
Using adjusted R^2 as well as Mallow's Cp statistic, we found that the best model contains the following eight predictors:

$$alcohol \sim \mathit{fixed\ acidity} + \mathit{volatile\ acidity} + \mathit{citric\ acid} + \mathit{residual\ sugar} + \mathit{density} + pH + \mathit{sulphates} + \mathit{quality}$$

After fitting the best subsets model, we saw that 68.8% (adjusted R^2 value) of the variance in the alcohol content of white wine is explained by the predictors.

4.2.4 Residual Analysis



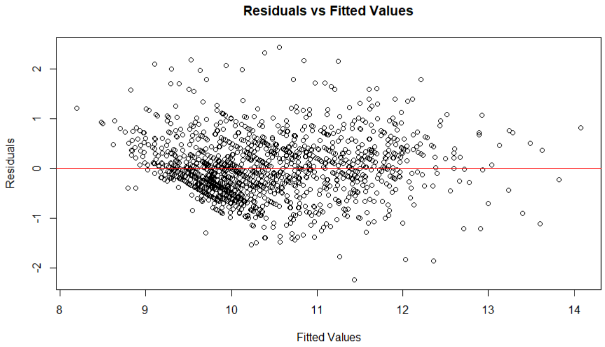


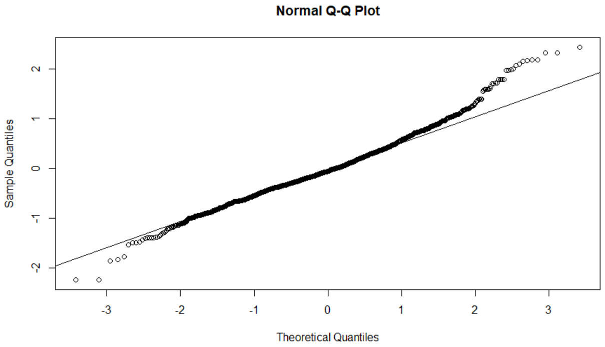
Looking at the residual plots for the best subset regression, we noticed that these plots closely resemble that of stepwise regression. The residuals are evenly spread in the Residuals vs. Fit plot, indicating that the error variances are equal and that a linear model is a good fit. The curvature at the top and bottom of the Normal Q-Q plot was concerning, so we performed a log and Boxcox transformation on Y. However, the residual plots are nearly identical, indicating that the transformations were unnecessary and that this current model does in fact meet all LINE conditions. So, the final model is:

$$alcohol_i = \beta_0\beta_1\mathit{fixed\ acidity}_i + \beta_2\mathit{volatile\ acidity}_i + \beta_3\mathit{citric\ acid}_i + \beta_4\mathit{residual\ sugar}_i + \beta_5\mathit{density}_i + \beta_6 pH_i + \beta_7\mathit{sulphates}_i + \beta_8\mathit{quality}_i$$

4.3 Outliers and Influential Points

First, we identified a list of outliers and high leverage points by calculating internally studentized residuals and hat values. Then, we identified a list of influential points by calculating externally studentized residuals. We compared the two lists and found the common points that were both outliers or high leverage and influential. We then refit our best subsets model and reanalyzed the residual values.





The above plots closely resemble the residual plots found from our best subsets regression. Since there is no significant change, we determined that it was unnecessary to remove influential data points as it had no impact on the model accuracy.

5 Answering Research Questions

5.1 Which factors affect alcoholic content for each wine type?

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$$
$$H_a : \text{at least one } \beta_k \neq 0 \text{ for } k = 1,2,3,4,5,6,7,8$$

To test this hypothesis, we performed a general linear F-test on both the white and red wine final models.

5.1.1 White Wine

With an F-value of 6441 and a corresponding p-value of 2.2e-16 $< \alpha = 0.05$, we must reject the null hypothesis. Therefore, not all $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. There is sufficient evidence to conclude that some of the variation in alcohol content is explained by one or more of these predictors.

5.1.2 Red Wine

With an F-value of 435.69 and a p-value of 2.2e-16 $< \alpha = 0.05$, we must reject the null hypothesis. Therefore, not all $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$. There is sufficient evidence to conclude that some of the variation in alcohol content is explained by one or more of these predictors.

5.2 Is the quality and density of wine significant in predicting the amount of alcohol in a particular wine type?

$$H_0 : \beta_5 = \beta_8 = 0$$
$$H_a : \text{at least one } \beta_k \neq 0 \text{ for } k = 5,8$$

To test this hypothesis, we performed a general linear F-test on both the white and red wine final models.

5.2.1 White Wine

With an F-value of 19070 and a corresponding p-value of 2.2e-16 $< \alpha = 0.05$, we must reject the null hypothesis. Therefore, both $\beta_5 = \beta_8 \neq 0$. There is sufficient evidence to conclude that quality and density of white wine is significant in predicting the amount of alcohol.

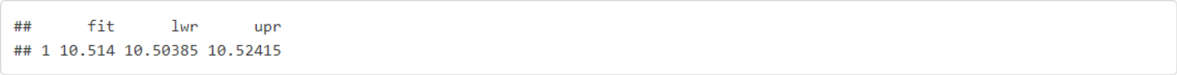
5.2.2 Red Wine

With an F-value of 1395.1 and a p-value of 2.2e-16 $< \alpha = 0.05$, we must reject the null hypothesis. Therefore, both $\beta_5 = \beta_8 \neq 0$. There is sufficient evidence to conclude that quality and density of red wine is significant in predicting the amount of alcohol.

5.3 What is the expected amount of alcohol in wine when all other factors are average?

We used the predict() function in R to determine a 95% confidence interval for alcohol content when all other predictors are averaged.

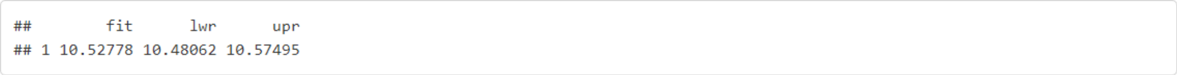
5.3.1 White Wine



```
##      fit      lwr      upr
## 1 10.514 10.50385 10.52415
```

We are 95% confident that the expected alcohol content is between 10.50 and 10.52 percent given that all other variables are averaged.

5.3.2 Red Wine



```
##        fit      lwr      upr
## 1 10.52778 10.48062 10.57495
```

We are 95% confident that the expected alcohol content is between 10.48 and 10.57 percent given that all other variables are averaged.

6 Comparing White and Red Wine

Overall, we found the best model for both the white wine and red wine data using best subsets regression. In fact, both models have exactly eight predictors with seven being the same parameter: fixed acidity, volatile acidity, residual sugar, density, pH, sulphates, and quality. In the white wine model, free sulfur dioxide is considered an essential predictor, whereas in the red wine model, citric acid is included. This is an interesting distinction, given that white wine tends to be more acidic than red wine and also contains more free sulphur dioxide. Furthermore, stepwise regression with AIC produces larger models, containing most if not all predictors for both datasets. Through residual analysis, we found that the addition of the extra variables was insignificant as removing these parameters has no effect on the accuracy of the model. We also noticed that the white wine model has a significantly larger adjusted R^2 value than red wine. Thereby, less of the variation in alcohol content in red wine is explained by the predictors than that of white wine. Perhaps this difference can be attributed to the fact that the white wine data has nearly 3000 more data entries, allowing the model to have greater accuracy.

With regard to individual parameters, both datasets show a negative correlation between density and alcohol with the white wine data having a stronger negative correlation. In general, white wine has more residual sugars, causing the density to be greater than that of red wine. Perhaps this fact is related to the differences in correlation between alcohol and density. Moreover, we found that the average alcohol content in red and white wine is about the same, both having about 10.5%.

7 Conclusion

In conclusion, the purpose of this project was to explore which factors are significant in predicting alcohol percentage. From analysis, we found that eight of the eleven factors are needed to predict alcohol content for both red and white wine. Fixed acidity, volatile acidity, residual sugar, chlorides, total sulfur dioxide, density, pH, sulphates, and quality are included in both models whereas free sulfur dioxide and citric acid are necessary for white and red wine models respectively. We also found that density and quality are necessary in predicting alcohol content and that the average wine contains roughly 10.5% alcohol.

However, there are some ways in which this model can be improved. It must be noted that all the samples were taken from one manufacturer, so our model fits this specific manufacturer's wines best and may not be suitable for other brands. Additionally, all the wine described by the data is made from grapes grown in a specific region of northern Portugal. Making wine in different regions may affect the alcohol content of wines. If we would like to extend our findings to other wines, we would need to collect wine data from different locations and brands. Furthermore, the given data does not distinguish between the different types of red and white wines available, adding a layer of complexity that is not addressed with our current models. We would have to perform some type of classification modeling to understand how this affects our final models.

8 Data Source Citation

Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez
A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009