An Analysis of Wine Quality

PSTAT 126
Fall 2019

Maia Young (Tuesday 6:00-6:50 pm)
Romtin Toranji (Wednesday 3:00-3:50 pm)

Table of Contents

1. Introduction

Wine tasting has been practiced for hundreds of years and is considered by many to be an important tradition and even an art form. There is a high value placed on what is considered to be high quality wine and people spend their lives learning the craft of tasting and rating wine. The dataset examined for this project contains the overall rankings given to a large set of Portugese wine samples by professional wine tasters. The rankings vary from scores of 0 to 10 where 0 is the worst possible quality score and 10 is the best. The variables fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol are taken into account as possible predictors the overall quality score. A 12th possible predictor is added to the set, id, as a way to differentiate between the red and white wine varieties that are included in the set. Three research questions emerged when considering this dataset, the first regarding the effectiveness of the 12 variables in predicting wine quality, the second in predicting quality based on alcohol content and the third about the difference between the mean quality responses for red wine vs. white wine.

Data:

fixed acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily)
volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste
citric acid	found in small quantities, citric acid can add 'freshness' and flavor to wines
residual sugar	the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
chlorides	the amount of salt in the wine
free sulfur dioxide	the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

total sulfur dioxide	amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
density	the density of water is close to that of water depending on the percent alcohol and sugar content
pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
sulphates	a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant
alcohol	the percent alcohol content of the wine
quality	output variable (based on sensory data, score between 0 and 10)
id	White wine = 0 Red wine = 1

2. Questions of Interest

A. How do fixed acidity, volatile acidity, citric acid, residual sugar, chlorides free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content and wine variety predict the quality of wine when ranked on a scale of 0 to 10?

B. What is the predicted response (quality rating) for wine that has an alcohol content of 14.5%?

C. Is there a difference in mean wine quality for white and red wines?

3. Regression Method

Regression analysis can be used to determine which of the 12 predictors are relevant in predicting wine quality and to eliminate any issues of collinearity which may arise due to the related nature of many of the possible predictors. A general F test that all the coefficients are equal to zero is appropriate in this case. Stepwise regression with AIC can be used to determine the proper order of predictors and their corresponding coefficients and calculating cook's distances will reveal if the dataset contains any influential points. This will ultimately lead to a best fit model for predicting wine quality. To answer the second research question a 95% prediction interval will be used given an alcohol content of 14.5% and the average values for all other predictors in the model. The final question can be answered by conducting a two sample t test for a difference in means for the two types of wine present in the dataset.

4. Regression Analysis, Results, and Interpretations

A. Question 1:

Analysis details:

The appropriate, untransformed model is found to include the predictors: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, pH, sulphates and alcohol. Density and id were removed from the best fit model because of large vif (variance inflation factor) values (greater than 5) and total sulfur dioxide because of a strong correlation with free sulfur dioxide of .72093408. Additionally, total sulfur dioxide had a larger vif value than the rest of the remaining predictors.

	chlorides	free.sulfur.dioxide	t
fixed.acidity	0.29819477	-0.28273543	
volatile.acidity	0.37712428	-0.35255731	
citric.acid	0.03899801	0.13312581	
residual.sugar	-0.12894050	0.40287064	
chlorides	1.00000000	-0.19504479	
free.sulfur.dioxide	-0.19504479	1.00000000	
total.sulfur.dioxide	-0.27963045	0.72093408	

The remaining predictors show vif values small enough to remain in the model:

fixed.acidity	volatile.acidity	citric.acid
1.673152	1.696675	1.561635
residual.sugar	chlorides	free.sulfur.dioxide
1.456538	1.524603	1.413851
pH	sulphates	alcohol
1.391563	1.345714	1.333655

Therefore the untransformed model is:

$$E(Quality) = \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 citric.acid + \beta_4 residual.sugar + \beta_5 chlorides + \beta_6 free.sulfur.dioxide + \beta_7 pH + \beta_8 sulfates + \beta_9 alcohol$$

F-Test:

Using the hypothesis that:

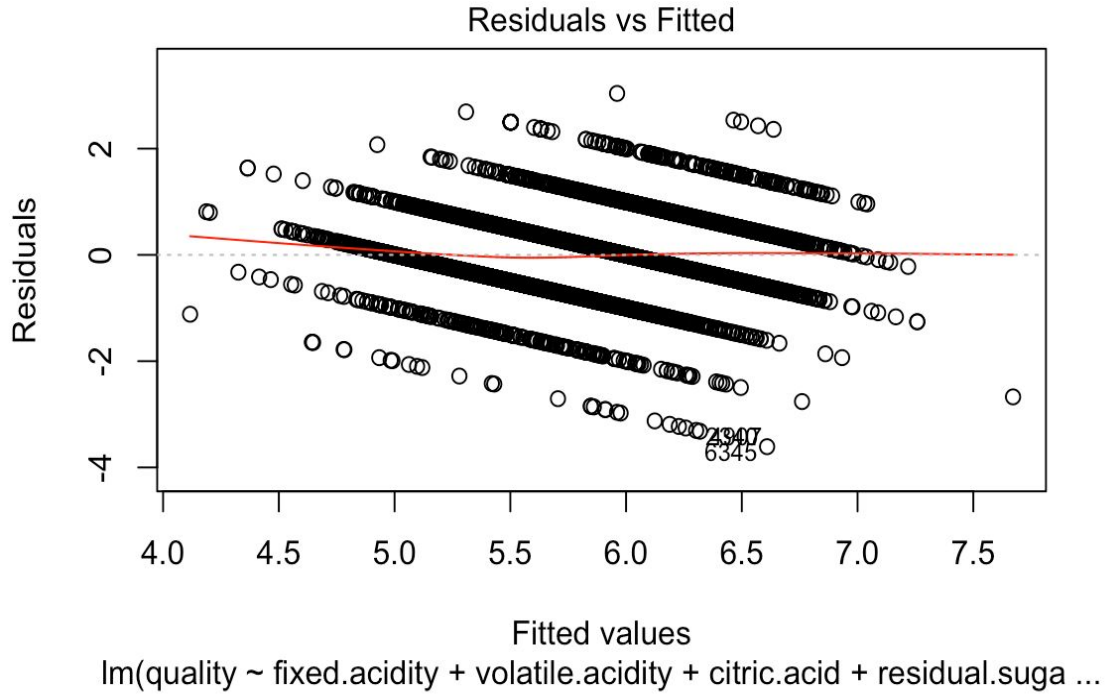
H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$

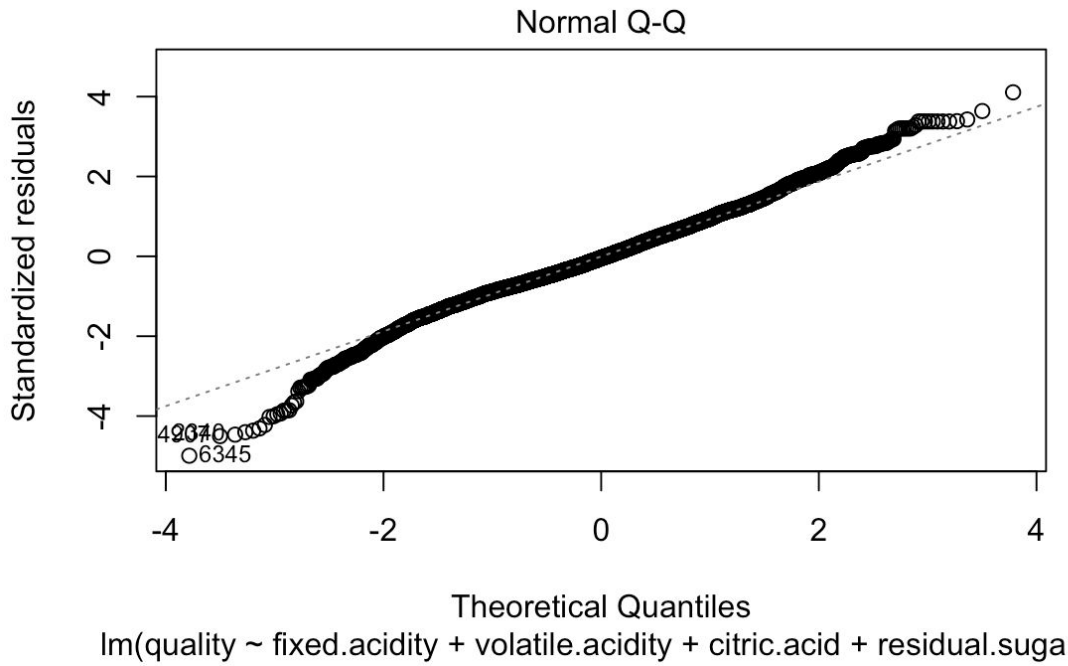
vs. the alternative

H_1: At least one β is not equal to 0

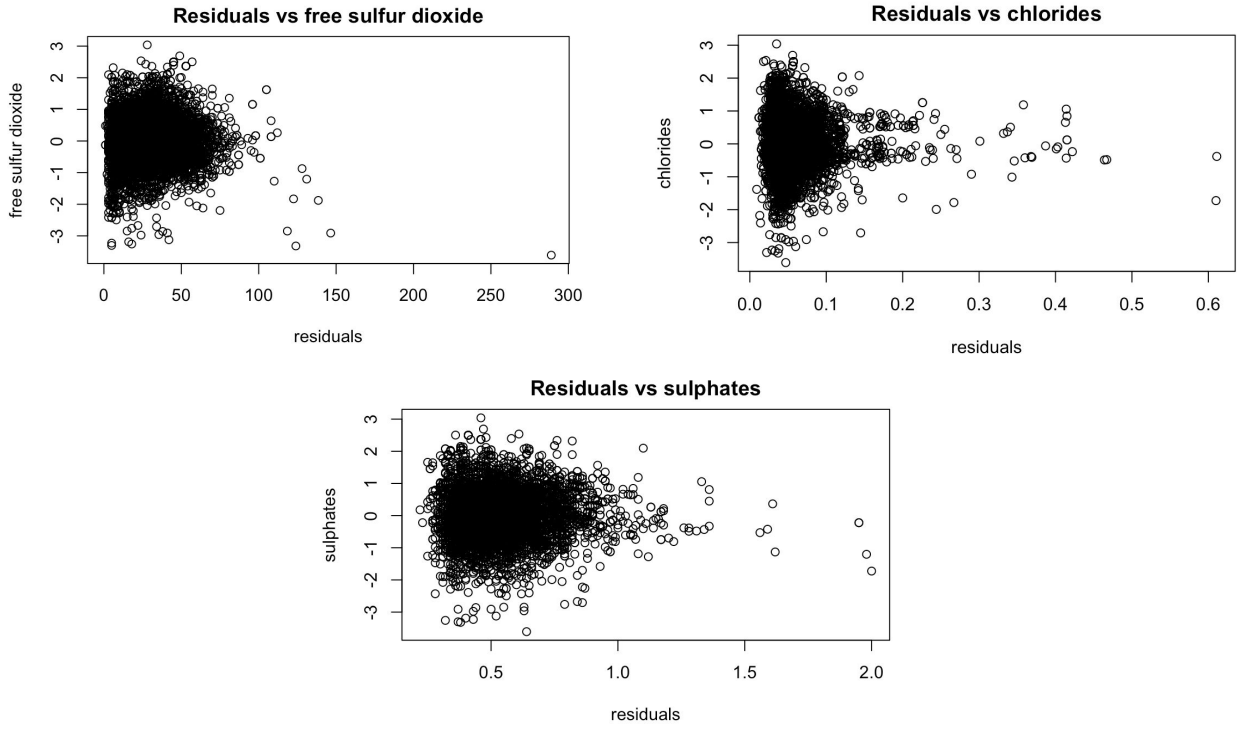
This hypothesis test results in a general F-statistic of 283.6 and a p-value of <2.2e-16. Because this p-statistic is very small, the null hypothesis is rejected and therefore at least one β is not equal to zero. From this point diagnostic checks were done to assess if any transformations needed to be done on the predictors that remain.

Diagnostic checks:

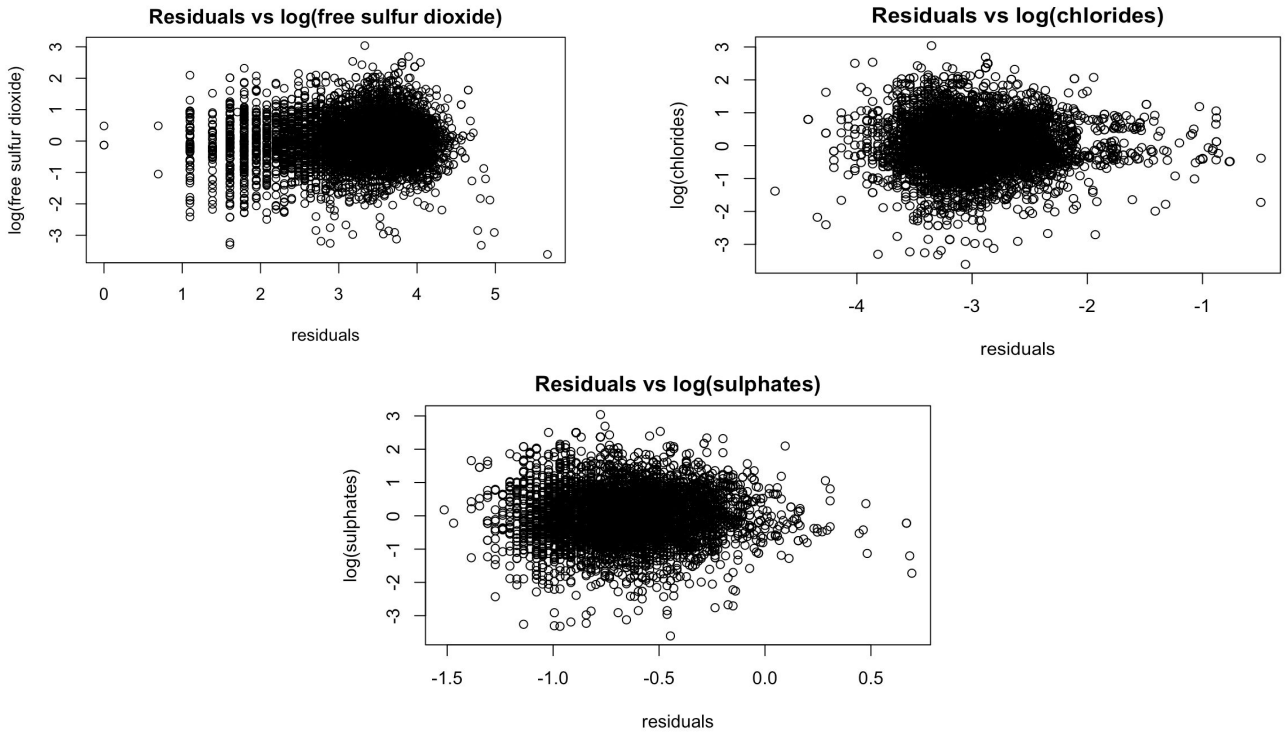




Conducting diagnostics on the untransformed model after density and id were removed show a residual plot which has a horizontal band around the 0 line with no discernable pattern in the plot. Thus we satisfy the Linearity and Equal Variance LINE conditions. Additionally, the QQ plot appears to be fairly linear aside from some slight skew and therefore the Normality condition is satisfied. Taking a closer look at the residuals vs. fit plots of all predictors revealed that log transformations were necessary for three of the remaining predictors: free sulfur dioxide, chlorides and sulphates.

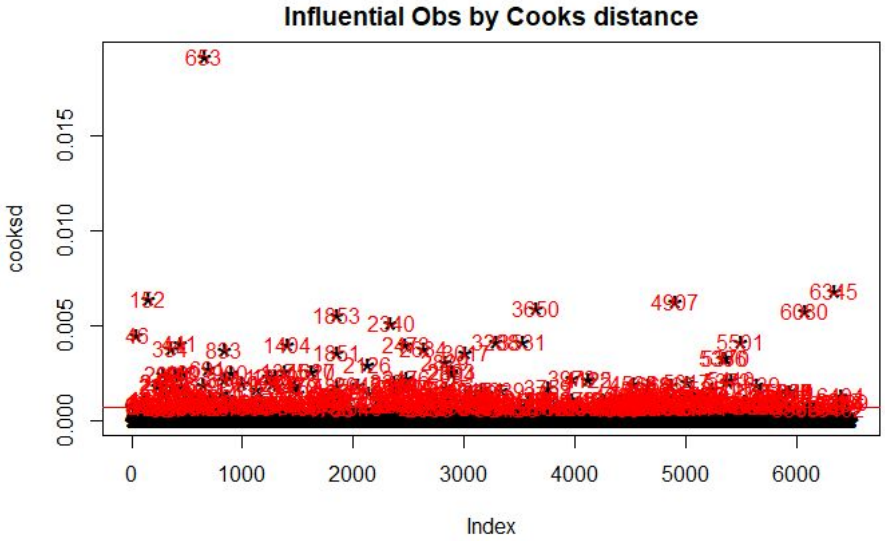


Residuals vs. fitted values for log transformations:

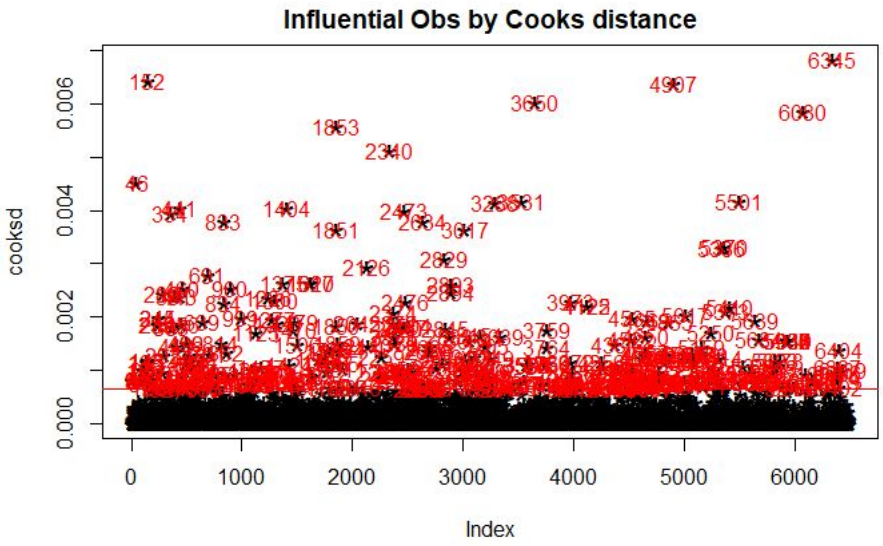


After applying the log transformations the residuals vs. fit plots for the three predictors display residuals that form a horizontal band around the 0 line and are randomly scattered, and therefore satisfy the Linearity and Equal Variances assumptions.

<u>Checking for Influential Points:</u>



Observing the Cook's distance of the model, we can see that there seems to be an influential point. We can remove this point from our dataset and recalculate the model. Below is the new model's Cook's distance:



There does not seem to be any influential points from our new model's Cook's distance. Therefore, we can conclude that our model is not influenced by any highly influential points.

Final Model:

Conducting stepwise regression with AIC on the transformed predictors led to the removal of log(chlorides) and the arrival at the final, transformed model. The ANOVA table shows that the p-values for all remaining predictors are significantly small (<.05) and therefore can remain in the model.

```
Analysis of Variance Table

Response: quality
                          Df Sum Sq Mean Sq   F value    Pr(>F)
alcohol                    1  978.0  977.95 1796.8089 < 2.2e-16 ***
volatile.acidity           1  307.5  307.51  564.9901 < 2.2e-16 ***
log(sulphates)             1   49.4   49.40   90.7566 < 2.2e-16 ***
log(free.sulfur.dioxide)   1   47.5   47.51   87.2847 < 2.2e-16 ***
residual.sugar             1   23.5   23.50   43.1724 5.402e-11 ***
pH                         1    3.8    3.85    7.0655 0.0078776 **
fixed.acidity              1    7.1    7.13   13.0997 0.0002976 ***
citric.acid                1    5.6    5.62   10.3239 0.0013196 **
Residuals               6488 3531.2    0.54
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The summary table for the best fit model gives the estimated coefficients, which results in the model:

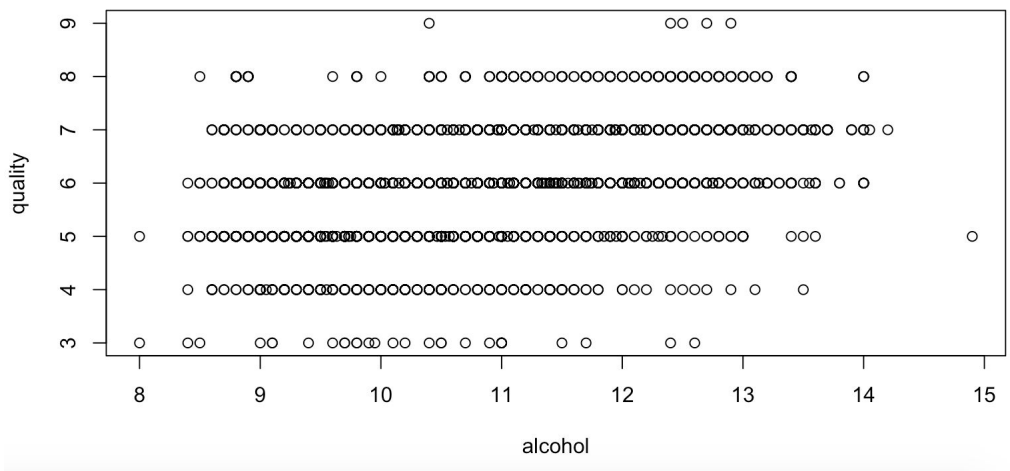
$$Quality = 1.331817 + .354809alcohol - 1.3565576volatile.acidity + .382366log(sulphates) + .124489log(free.sulfur.dioxide) + .017083residual.sugar + .237243pH + .042343fixed.acidity - .250106citric.acid$$

Interpretation:

The first research question considered for this study was: How do fixed acidity, volatile acidity, citric acid, residual sugar, chlorides free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol content and wine variety predict the quality of wine when ranked on a scale of 0 to 10? Through regression analysis it was discovered that alcohol content, volatile acidity, log(sulphates), log(free sulfur dioxide), residual sugar, pH, fixed acidity and citric acid are valid and significant predictors for the white and red Portugese wines sampled in this dataset. The potential predictors of density, total sulfur dioxide id and chlorides were unable to prove significant enough to include in the model as predictors for quality.

B. Question 2:

Analysis Details:



By viewing the relationship between alcohol content and overall wine quality one can see that there is a positive and somewhat linear relationship between the predictor and response. Because of this trend a prediction interval was created for the case where alcohol percentage is 14.5%, which is within the scope of the model because it is smaller than the largest value for alcohol percentage in the dataset(14.9%). In order to hold the rest of the predictors constant their average values were used in the interval.

```
       fit      lwr      upr
1 7.278853 5.830955 8.726751
```

These values gave a predicted quality ranking of 7.278853 and a 95% interval of (5.830955, 8.726751).

Interpretation:

We are 95% confident that when holding all predictors constant at their average values except for alcohol content which is given the value 14.5, that the new response for wine quality is between 5.830955 and 8.726751. The predicted new response(quality) is 7.278853. This result is consistent with the trend displayed in the graph of quality vs. alcohol because the high percentage wine receives a quality rating that is on the upper end of the scale.

C. Question 3

Analysis Details

One can compare the means of the two wines by doing a two sample t-test. From the test the results are:

```
        Welch Two Sample t-test

data:  winequality.red$quality and
winequality.white$quality
t = -10.149, df = 2950.8, p-value < 2.2e-16
alternative hypothesis: true difference in means is not
equal to 0
95 percent confidence interval:
 -0.2886173 -0.1951564
sample estimates:
mean of x mean of y
 5.636023  5.877909
```

Interpretation

If an alpha = .05 is used, one can reject Ho. Therefore, when examining quality ratings for red and white wines it is concluded that the means of the two samples are not equal.

5. Conclusion

The analysis of this dataset led to a final model which includes alcohol content, volatile acidity, log(sulphates), log(free sulfur dioxide), residual sugar, pH, fixed acidity and citric acid as significant predictors for wine quality. Several predictors were removed from the model such as Id, thus showing that the quality ranking given to the wine is not significantly affected by the wine variety(red or white). Despite this, conducting a two sample t test for the mean quality ratings of red and white wines shows that the means are significantly different. The data also revealed that a predicted quality rating for a 14.5% wine while holding other predictors constant is 7.278853 and that we are 95% confident that the true value is contained in the interval (5.830955, 8.726751).

The dataset used for this study contained information on red and white wines of a specific type of Portugese wine and therefore the results can only be used to draw conclusions about that specific variety of wine. The model could be made more applicable by including data from other wine varieties aside from the Portugese wines tasted in the dataset. Utilizing wine outside of this particular dataset would allow for the results to be generalized to more varieties of wine.

6. Appendix

A. R Code

Import datasets into R

```
winequality.red <- read.csv("~/Downloads/winequality-red.csv", sep=";")
View(winequality.red)
winequality.white <- read.csv("~/Downloads/winequality-white.csv", sep=";")
View(winequality.white)
```

Add a new column to the datasets to id red and white: white = 0 red = 1

```
winequality.white$id = 0
winequality.red$id = 1
```

Combine datasets into 1

```
winequality = rbind(winequality.red,winequality.white)
View(winequality)
fixed.acidity=winequality$fixed.acidity
volatile.acidity=winequality$volatile.acidity
citric.acid=winequality$citric.acid
residual.sugar=winequality$residual.sugar
chlorides=winequality$chlorides
free.sulfur.dioxide=winequality$free.sulfur.dioxide
total.sulfur.dioxide=winequality$total.sulfur.dioxide
density=winequality$density
pH=winequality$pH
```

```
sulphates=winequality$sulphates

alcohol=winequality$alcohol

quality=winequality$quality

id=winequality$id

attach(winequality)

plot(quality[id == 1], alcohol[id == 1], col = "black", main = "quality vs alcohol
content",ylab="alcohol",xlab="quality")

points(quality[id==0],alcohol[id==0],col = "yellow")
```

Create a scatterplot matrix and determine correlations between predictors

```
mod=lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar +
chlorides + free.sulfur.dioxide + total.sulfur.dioxide + density + pH + sulphates + alcohol
+ id)

pairs(winequality)

cor(subset(winequality, select = -c(quality)))
```

Create a step function to determine an appropriate linear model

```
library("car", lib.loc="~/R/win-library/3.6")

library(car)

library(faraway)

vif(mod.upper)
```

Drop density from the predictors because it has the largest VIF

```
vif(lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar +
chlorides + free.sulfur.dioxide + total.sulfur.dioxide  + pH + sulphates + alcohol + id))
```

Drop id from predictors because it is over 5

```
vif(lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide  + pH + sulphates + alcohol))
```

Remove total.sulfur.dioxide because of its higher correlation with free.sulfur dioxide and its larger vif value

```
vif(lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide  + pH + sulphates + alcohol ))
```

Since everything is smaller than 2 we keep the rest of the predictors: fit full model

```
full_model = lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar + chlorides + free.sulfur.dioxide  + pH + sulphates + alcohol )

anova(full_model)

summary(full_model)
```

Diagnostic plots for the full model

```
plot(full_model)

plot(winequality$fixed.acidity, resid(full_model), main = "Residuals vs fixed acidity", ylab = "fixed acidity",xlab= "residuals")

plot(winequality$volatile.acidity, resid(full_model), main = "Residuals vs volatile acidity", ylab = "volatile acidity",xlab= "residuals")

plot(winequality$citric.acid, resid(full_model), main = "Residuals vs citric acid", ylab = "citric acid",xlab= "residuals")

plot

plot(winequality$chlorides, resid(full_model), main = "Residuals vs chlorides", ylab = "chlorides",xlab= "residuals")

plot(winequality$free.sulfur.dioxide, resid(full_model), main = "Residuals vs free sulfur dioxide", ylab = "free sulfur dioxide",xlab= "residuals")
```

```
plot(winequality$pH, resid(full_model), main = "Residuals vs pH", ylab = "pH",xlab= "residuals")

plot(winequality$sulphates, resid(full_model), main = "Residuals vs sulphates", ylab = "sulphates",xlab= "residuals")

plot(winequality$alcohol, resid(full_model), main = "Residuals vs alcohol", ylab = "alcohol",xlab= "residuals")
```

Stepwise regression: use log of free sulfur dioxide chlorides and sulphates to normalize the data

```
new.free.sulfur.dioxide=log(free.sulfur.dioxide)

new.chlorides=log(chlorides)

new.sulphates=log(sulphates)

plot(new.free.sulfur.dioxide, resid(full_model), main = "Residuals vs log(free sulfur dioxide)", ylab = "log(free sulfur dioxide)",xlab= "residuals")

plot(new.chlorides, resid(full_model), main = "Residuals vs log(chlorides)", ylab = "log(chlorides)",xlab= "residuals")

plot(new.sulphates, resid(full_model), main = "Residuals vs log(sulphates)", ylab = "log(sulphates)",xlab= "residuals")

mod0 = lm(quality~1)

mod.upper = lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar + log(chlorides) + log(free.sulfur.dioxide)  + pH + log(sulphates) + alcohol )

stepmodel = step(mod0,scope=list(lower = mod0, upper = mod.upper))

anova(stepmodel)

summary(stepmodel)
```

Cook's distance

```
library(olsrr)

cooksd <- cooks.distance(mod.upper)
```

```
plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")  # plot cook's distance

abline(h = 4*mean(cooksd, na.rm=T), col="red")  # add cutoff line

text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),names(cooksd),""), col="red")  # add labels

#testing the model without the influential points

winequality.no.outlier = winequality[-c(653),]

mod0 = lm(quality~1,data = winequality.no.outlier)

mod.upper = lm(formula = quality ~ fixed.acidity +volatile.acidity  + citric.acid + residual.sugar + log(chlorides) + log(free.sulfur.dioxide)  + pH + log(sulphates) + alcohol , data = winequality.no.outlier)

stepmodel = step(mod0,scope=list(lower = mod0, upper = mod.upper))

cooksd <- cooks.distance(mod.upper)

plot(cooksd, pch="*", cex=2, main="Influential Obs by Cooks distance")  # plot cook's distance

abline(h = 4*mean(cooksd, na.rm=T), col="red")  # add cutoff line

text(x=1:length(cooksd)+1, y=cooksd, labels=ifelse(cooksd>4*mean(cooksd, na.rm=T),names(cooksd),""), col="red")  # add labels
```

Testing the regression model

```
set.seed(27)

rows = sample(nrow(winequality))

shuffle_winequality = winequality[rows, ]

testing = shuffle_winequality[1:1000,]

#test the number correct

correct = 0

total = 1000
```

```
for(i in 1:1000){

 if(testing[i,]$quality == round(1.331817 + (0.354809*(testing[i,]$alcohol)) +
(-1.365576*testing[i,]$volatile.acidity) + (log(testing[i,]$sulphates)     *0.382366) +
(log(testing[i,]$free.sulfur.dioxide)*0.124498 )+(testing[i,]$residual.sugar * 0.017083
)+(testing[i,]$pH*0.237243)+(testing[i,]$fixed.acidity*0.042343)+(testing[i,]$citric.acid
* -0.250106) )){

   correct = correct + 1 } }

testing[1,]$alcohol

print(correct)

print(correct/total)
```

Therefore our model has about 54% accuracy which is much more accurate than a random guess (10%)

Prediction interval

```
new=data.frame(alcohol=14.5, volatile.acidity=mean(volatile.acidity),
sulphates=mean(sulphates), free.sulfur.dioxide=mean(free.sulfur.dioxide),
residual.sugar=mean(residual.sugar), pH=mean(pH), fixed.acidity= mean(fixed.acidity),
citric.acid=mean(citric.acid))

pi=predict(stepmodel, new, interval='prediction', level=.95)

pi
```

Two-sample t-test

```
res<-t.test(winequality.red$quality,winequality.white$quality)

Res
```

B. Citation

Citation

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.