Quality Wine is a Quality Time

Landon Anderson, Salem Karren, Emma Larson

April 17, 2020

Introduction

The wine industry has become increasingly popular in recent years. In order to keep up with growing demand, wine companies have looked to invest in new technology for both the making and selling process. To prevent illegal adulteration and disregard of basic safeguards, companies still have to pass quality assessments and obtain certification for all wine sold. The evaluation of wine quality helps identify influential factors that improve wine making and classify wines based on their quality. The certification process involves assessment of alcohol or pH values (measured through physicochemical tests) and taste (through a sensory test). Massive datasets, such as the one assessed in this paper, show trends and patterns that can be used to improve wine companies chances for producing high-quality wine (Cortez et al., 2009).

The purpose of this study is to build a model that can predict the quality of red wine grown in the Vinho Verde region of Portugal using various chemical and physical characteristics of the wine. The data for our model comes from the University of California Irvine (UCI) Machine Learning Repository and is discussed in further detail in section *Data* below. After successfully building our model, we should be able to predict the quality of wine based on certain predictor variables (see Data section) with statistical accuracy. Since higher quality wine has a higher price point, this analysis will be extremely useful to wine producers as well as those performing wine certification and quality evaluations. If the results show that certain factors are more important than others in producing quality wine, companies could focus on these variables and wouldn't waste time and money on things that are statistically insignificant. These results would also be beneficial to consumers as they would know when they pay for premium or luxury wine, it is actually of a higher quality. It could also benefit consumers as companies would likely be able to reduce the price of higher quality wine as they become more effective at producing it. This would give more people the ability to purchase better tasting and higher quality wine. As such, the following regression analysis of the wine dataset is not only interesting, it could potentially be very lucrative for wineries and beneficial for wine consumers.

Data

The dataset we are using contains the following variables:

Variable Name	Variable Type	Variable Description
FixedAcidity	Continuous numerical	Amount of tartaric acid (g/L)
VolatileAcidity	Continuous numerical	Amount of acetic acid (g/L)
Citric Acid	Continuous numerical	Amount of citric acid (g/L)
ResidualSugar	Continuous numerical	Amount of sugar remaining after fermentation (g/L)
Chlorides	Continuous numerical	Amount of sodium chloride (g/L)
FreeSulfurDioxide	Continuous numerical	Amount of free SO_2 at equilibrium in wine (mg/L)
TotalSulfurDioxide	Continuous numerical	Amount of free and bound form SO_2 at equilibrium in wine (mg/L)
Density	Continuous numerical	Density of the wine (g/L)
pH	Continuous numerical	pH of the wine 0-14
Sulphates	Continuous numerical	Amount of potassium sulphate in wine (g/L)
Alcohol	Continuous numerical	Percent alcohol content (vol%)
WineQuality	Discrete numerical	Rating of overall wine quality in whole numbers from 0 (lowest quality) to 10 (highest quality)

Ordinary Least Squares (OLS) regression typically should only be performed on data with a continuous response variable. However, because there is a sufficient number of observations in our dataset, the *WineQuality* (hereafter referred to as *quality*) data can be treated as near-continuous, allowing us to use OLS regression methods on the dataset.

Model Assumptions

Before moving forward with OLS regression, certain model assumptions must be met. Model assumptions are met when (1) the response and predictor variables share a linear relationship, (2) there are no outliers or influential points, and (3) the residuals are independently and identically distributed according to the normal distribution. If any of these assumptions are violated, we must perform remedial measures so the OLS regression is able to provide an accurate, unbiased estimation.

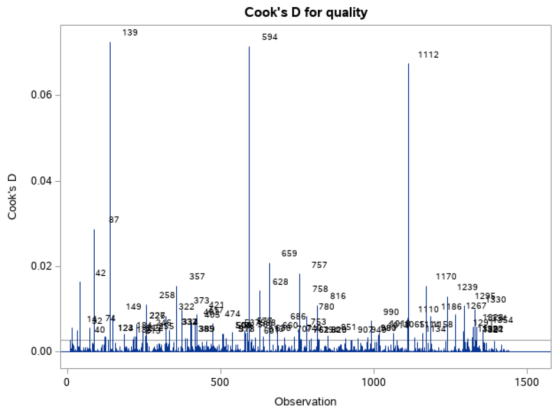
In order to check for a linear relationship between *quality* and the predictor variables, we created a correlation matrix that shows the degree to which each of the variables are correlated with one another (Figure 1). 9 of the 11 variables share a significant linear relationship with *quality*, so the first model assumption is satisfied.

Pearson Correlation Coefficients, N = 1439
Prob > |r| under H0: Rho=0

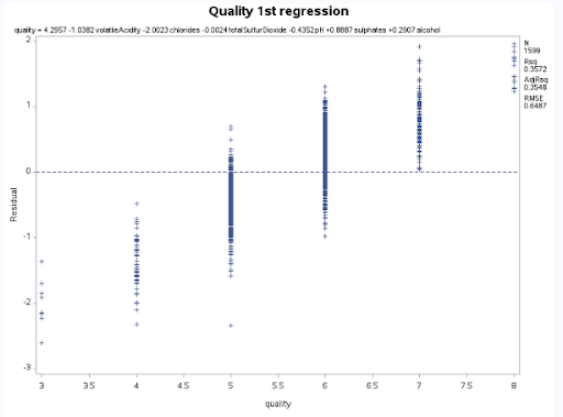
	quality	fixedAcidity	volatileAcidity	citricAcid	residualSugar	chlorides	freeSulfurDioxide	totalSulfurDioxide	density	pH	sulphates	alcohol
quality	1.00000	0.11887 <.0001	-0.40508 <.0001	0.23025 <.0001	0.00949 0.7190	-0.12998 <.0001	-0.05917 0.0248	-0.20973 <.0001	-0.17983 <.0001	-0.04887 0.0638	0.24616 <.0001	0.48043 <.0001

Figure 1: *Correlation between quality and each explanatory variable*

To check for influential points, we constructed a preliminary linear regression model that included each of the 11 explanatory variables. Figure 3a contains graphs from this regression model which indicate 3 highly influential points. A plot of residual vs actual values (Figure 3b) indicates the presence of one outlier. This point can be observed as the lowest value on the chart where $quality = 5$.



(a) *Cook's Distance identifies influential points*

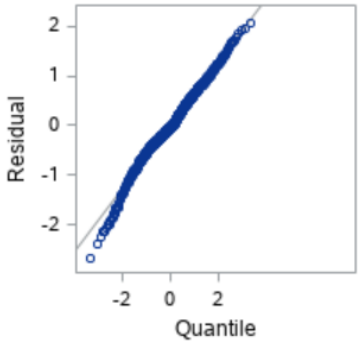


(b) *Histogram of distribution of residuals*

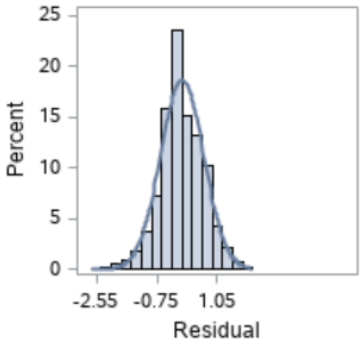
Figure 2: *Residual Vs Actual value plot identifies one outlier*

In order to proceed with OLS regression, these influential points and the outlier must be addressed so that model assumptions are met. See section **Remedial Measures** for a discussion of how these concerns were resolved.

We used output from this same preliminary regression to determine whether the model assumption regarding the distribution of residuals was met. The Normal QQ plot (Figure 3a) and the histogram of the distribution of residuals (Figure 3b) both show that the residuals follow the normal distribution.



(a) *Normal QQ Plot*



(b) *Histogram of residuals*

Figure 3: *Residuals are normally distributed*

Remedial Measures

In order to resolve the issues with influential points and outliers, we first explored the possibility of a transformation of the variable *quality* be conducting a Box-Cox analysis (Figure 4). The lambda value of 1 returned by the test indicates that a transformation of *quality* would not improve our predictions.

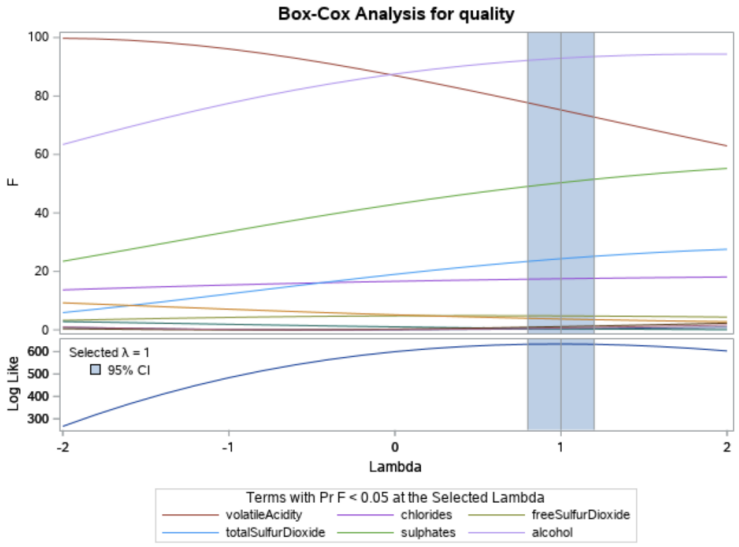
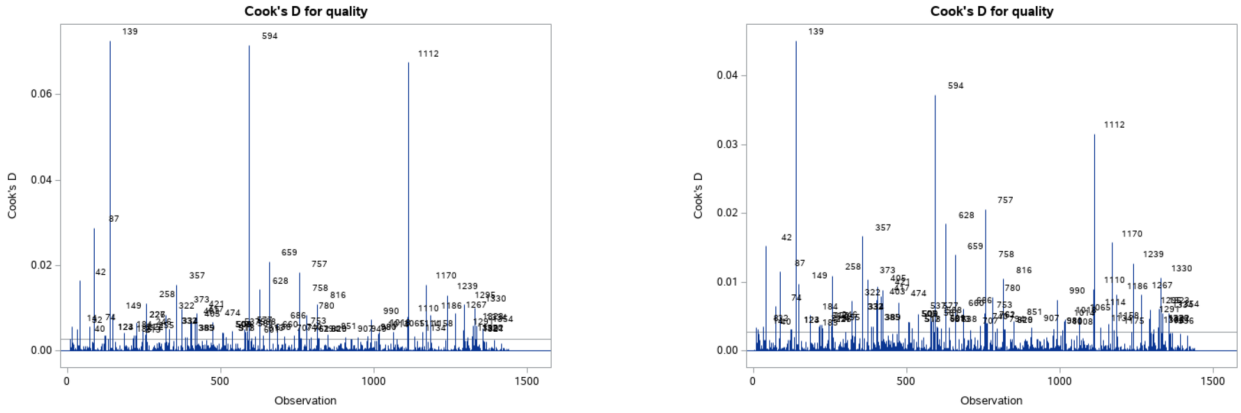


Figure 4: *Results of Box-Cox analysis*

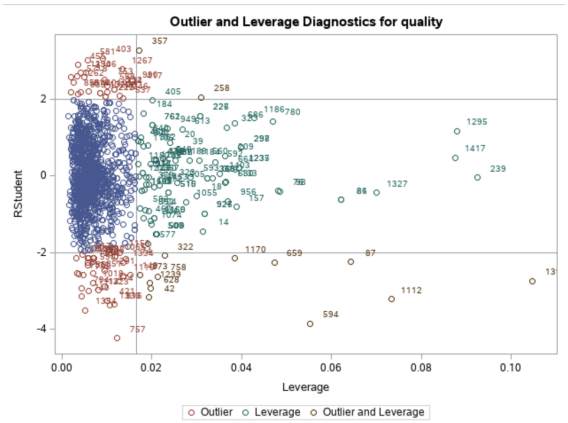
Since a transformation of *quality* would not be useful, we looked at the specific variables responsible for the outlier and three influential points identified above. To reduce the influence of these points on the model, we performed log transformations on the following explanatory variables: *FixedAcidity*, *ResidualSugar*, *Alcohol*, and *Sulphates*. Though they did not completely remove all influential points, these transformations were successful in reducing the influence of these points in the model and reducing the amount of points that could be classified as outliers/influential points (Figures 5, 6).



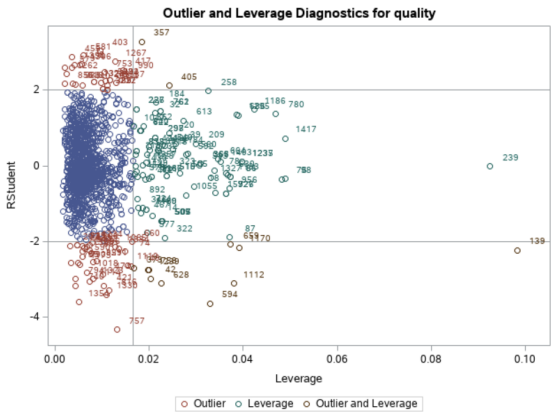
(a) *Cook's Distance before transformation*

(b) *Cook's Distance after transformation*

Figure 5: *Variable transformation reduced model influence*



(a) *Outlier Leverage plot before transformation*



(b) *Outlier Leverage plot before transformation*

Figure 6: *Outliers and Leverage are reduced by variable transformation*

Having resolved the issues with influential points and outliers, our dataset was now ready to continue with OLS regression.

Variable Selection

When a model has many predictor variables, it can become very complex and hard to understand. In order to make the prediction model easier to interpret and explain, we explored two methods of variable selection: Stepwise Selection and LASSO. These variable selection methods provide output that helped us determine which combination of predictor variables was able to most accurately predict wine quality at a reduced level of complexity.

Stepwise selection follows an algorithm for adding and removing variables from the model until an optimal model is obtained. The Stepwise selection results suggest the final model includes 7 variables: $VolatileAcidity, Chlorides, FreeSulfurDioxide, TotalSulfurDioxide, pH, log\_sulphates$, and $log\_alcohol$ (Figure 8).

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1.22508	0.56078	1.94327	4.77	0.0291
volatileAcidity	-0.99220	0.10653	35.32445	86.75	<.0001
chlorides	-1.95622	0.40479	9.50968	23.36	<.0001
freeSulfurDioxide	0.00503	0.00224	2.05147	5.04	0.0249
totalSulfurDioxide	-0.00377	0.00073783	10.61728	26.08	<.0001
pH	-0.45601	0.12192	5.69573	13.99	0.0002
log_sulphates	0.73406	0.08590	29.73713	73.03	<.0001
log_alcohol	3.00860	0.19274	99.20888	243.65	<.0001

Figure 7: *Results from Stepwise selection*

Results from the LASSO selection method are shown in figure 8a. This method suggests we only

keep 6 of the variables, excluding FreeSulfurDioxide and keeping the other 6 suggested by the Stepwise method.

LASSO Selection Summary				
Step	Effect Entered	Effect Removed	Number Effects In	SBC
0	Intercept		1	-615.9523
1	log_alcohol		2	-858.5727
2	volatileAcidity		3	-1036.6945
3	log_sulphates		4	-1199.9184
4	totalSulfurDioxide		5	-1219.6860
5	chlorides		6	-1233.7235
6	log_fixedAcidity		7	-1230.4262
7	pH		8	-1230.1641
8		log_fixedAcidity	7	-1238.5138*
9	density		8	-1233.5022
10	log_fixedAcidity		9	-1229.1249
11	freeSulfurDioxide		10	-1224.1792
12	log_residualSugar		11	-1226.2331
13	citricAcid		12	-1220.6284
* Optimal Value of Criterion				

(a) *LASSO selection process*

Parameter Estimates		
Parameter	DF	Estimate
Intercept	1	0.261981
volatileAcidity	1	-1.068201
chlorides	1	-1.336182
totalSulfurDioxide	1	-0.002262
pH	1	-0.155960
log_sulphates	1	0.683471
log_alcohol	1	2.984293

(b) *Variables selected by LASSO*

Figure 8: *Results from LASSO selection*

Because the suggested models from the selection techniques are very similar in both the variables selected for the model and the effectiveness of the model, measured by Adjusted R-Squared, we see no significant difference between the two sets of variables identified for inclusion. For simplicity, we will proceed with the results from LASSO, as this model achieves similar performance while including one less explanatory variable (Figure 8b).

Interaction Terms

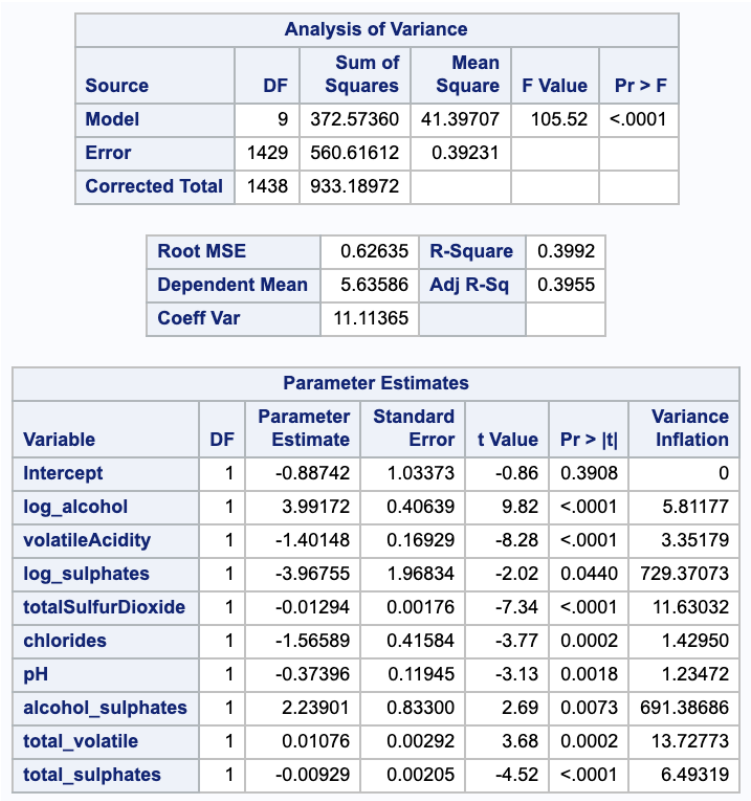
Sometimes the effect of a predictor variable on the response variable will depend on the values of other explanatory variables. This conditional effect on the response is called an interaction effect. To explore the possibility of interaction in our model, we checked for significant interaction between 6 pairs of the explanatory variables in our model. Of the pairs we tested, 3 yielded significant results:

$$alcohol\_sulphates = log\_alcohol * log\_sulphates$$

$$total\_sulphates = totalSulfurDioxide * log\_sulphates$$

$$total\_volatile = totalSulfurDioxide * volatileAcidity$$

After identifying the 3 interaction terms to include in the model, we produced our final model (Figure 9).



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	372.57360	41.39707	105.52	<.0001
Error	1429	560.61612	0.39231		
Corrected Total	1438	933.18972			

Root MSE	0.62635	R-Square	0.3992
Dependent Mean	5.63586	Adj R-Sq	0.3955
Coeff Var	11.11365		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.88742	1.03373	-0.86	0.3908	0
log_alcohol	1	3.99172	0.40639	9.82	<.0001	5.81177
volatileAcidity	1	-1.40148	0.16929	-8.28	<.0001	3.35179
log_sulphates	1	-3.96755	1.96834	-2.02	0.0440	729.37073
totalSulfurDioxide	1	-0.01294	0.00176	-7.34	<.0001	11.63032
chlorides	1	-1.56589	0.41584	-3.77	0.0002	1.42950
pH	1	-0.37396	0.11945	-3.13	0.0018	1.23472
alcohol_sulphates	1	2.23901	0.83300	2.69	0.0073	691.38686
total_volatile	1	0.01076	0.00292	3.68	0.0002	13.72773
total_sulphates	1	-0.00929	0.00205	-4.52	<.0001	6.49319

Figure 9: *Final Model Parameters*

Model Interpretation

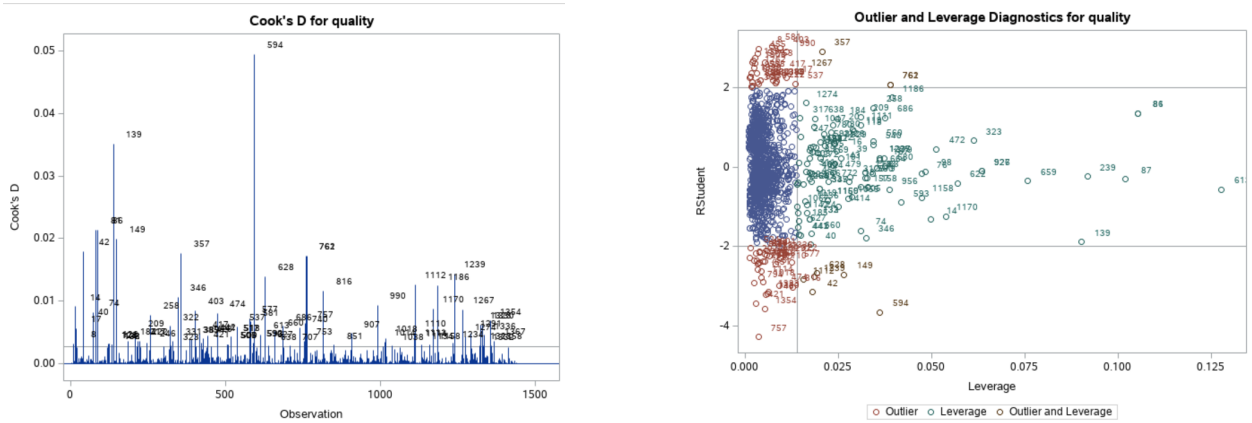
$$\begin{aligned}\hat{Y} = & -0.887 - 3.992 * log\_alcohol - 1.401 * volatileAcidity \\ & - 3.968 * \log_s ulphates - 0.013 * totalSulfurDioxide - 1.566 * chlorides \\ & - 0.374 * pH + 2.239 * alcohol\_sulphates + 0.011 * total\_volatile \\ & - 0.0093 * total\_sulphates\end{aligned}$$

The coefficient for chlorides can be interpreted to mean that holding all other variables constant, a one unit increase in chlorides will lead to an average decrease in wine quality of 1.566. Looking at the pH variable, we can interpret its coefficient to mean that holding all other variables constant, a one unit increase in pH leads to an expected decrease of 0.374 in wine quality.

One way this information could be used is by wine producers attempting to increase the quality of their wine. From these two interpretations we can see that the predictor variable chlorides (Amount of sodium chloride (g/L)) in the wine) seems to have a greater impact in affecting wine quality than pH level does. In other words, when all other variables are held constant, pH seems to have a less negative effect on wine quality than chlorides does. Knowing this, a wine company looking to improve wine quality would increase their wine quality faster by focusing on reducing the amount of sodium chloride in the wine rather than reducing the pH level.

Outliers and Influential Points

An effective model will account for outliers and influential points. In order to check our model's effectiveness, we analyzed the same plots as when we checked original model assumptions. (See section **Remedial Measures**.) The plots are reported below (Figure 10).



(a) *Cook's Distance plot shows reduced influence*

(b) *Outlier Leverage plot shows fewer outliers*

Figure 10

Multicollinearity

Another issue that needs to be checked for is multicollinearity. When two predictor variables share a strong linear relationship, independent of the response variable, multicollinearity is present. Where interaction terms have to do with the relationship between the response and predictor variables, multicollinearity has nothing to do with the response variable. Multicollinearity can create some issues in the model because it makes it hard to interpret model coefficients as it wouldn't make sense to hold all other predictors constant. The variance of model coefficients could also be inflated/ seem contradictory to what you would assume the relationship to be (ie. a negative value when you know the relationship is positive). Multicollinearity affects model inference but it does not affect a model's predictive ability.

Because our model includes several interaction terms, it has multicollinearity. However, standardizing the data reveals that the multicollinearity is not problematic, as the VIF values are all close to 1 (figure 11).

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.00036814	0.00055169	-0.67	0.5047	0
log_alcohol	1	0.36476	0.02301	15.85	<.0001	1.25984
volatileAcidity	1	-0.20038	0.02340	-8.56	<.0001	1.30289
log_sulphates	1	0.23660	0.02450	9.66	<.0001	1.42798
totalSulfurDioxide	1	-0.12359	0.02148	-5.75	<.0001	1.09738
chlorides	1	-0.09231	0.02451	-3.77	0.0002	1.42950
pH	1	-0.07133	0.02278	-3.13	0.0018	1.23472
alcohol_sulphates	1	2.33779	0.86975	2.69	0.0073	1.22564
total_volatile	1	2.88583	0.78408	3.68	0.0002	1.12633
total_sulphates	1	-3.15879	0.69808	-4.52	<.0001	1.19827

Figure 11: *Multicollinearity is not an issue*

Alternative Method

One alternative to OLS is Regression Trees. Regression Trees divide the data into groups at many different levels in such a way that the sum of squared errors is minimized. Regression trees work very well with data that has a discrete response variable, as our dataset does. Our regression tree results indicate that $Alcohol, Sulphates, and volatileAcidity$ are the most important factors in determining wine quality.

Variable Importance			
	Training		
Variable	Relative	Importance	Count
alcohol	1.0000	14.6179	4
sulphates	0.6429	9.3984	4
volatileAcidity	0.6185	9.0407	5
totalSulfurDioxide	0.2453	3.5858	2
fixedAcidity	0.1910	2.7916	1
pH	0.1437	2.0999	1

Figure 12: *Most important predictor variables*

Accuracy

The MSPR (Mean Squared Prediction Error) of the intercept only test set shows the predictive ability of the model without any predictor variables. This will essentially serve as a control to ensure that our model- which includes predictor variables- is better than a model without any variables.

The three other MSPR values shown are for the full model, the final model, and the alternative model. Essentially the purpose in comparing these MSPR values is to see which model is the best at predicting wine quality. We would hope to see the final model have the lowest MSPR value, meaning that the values are dispersed more closely around the mean. Though our final model has a slightly larger MSPR value than the full model, we still prefer the final model because it is more simple/ easy to comprehend and has a similar MSPR value.

Conclusion

To conclude, through remedial measures, log transformations, variables selection techniques, checking for interaction terms, and other methods used in this paper, we have produced a final model that predicts wine quality fairly accurately. This regression model will help those in the wine industry know what variables are most important in producing higher quality wine and be able to cut costs by focusing on the most significant factors. When the wine companies are able to produce wine at a higher average quality and lower price, it would be natural for the price of higher quality wines to be reduced and thus allow more people the opportunity to purchase it. Using this regression model in wine production would be of great benefit to the wine economy and the first step in giving more people an opportunity to have a quality time with some quality wine.

Looking into future directions for the research, it would be interesting to analyze other types of wine with the same model we produced. Since this dataset only included red wine, we could attempt to answer the question, Does this model work with other types of wine besides red? With the hypothesis that this regression function does work for all wine types, further analysis could be run to determine if that hypothesis is valid or if there is evidence to reject it.

Another possible direction to pursue would be looking into the variables most influential in predicting wine quality and studying ways to maximize (if they positively influence quality) or minimize (if they negatively influence quality) their presence. Using one of these final predictor variables as the response variable in a new equation would allow us to look into specific explanatory variables that will contribute to it such as growing techniques, processing, fertilization, etc. If successful, it would be possible to provide wineries with a formula of sorts that is statistically likely to produce higher quality wine more frequently.

References

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.

Appendix - SAS Code

```
Appendix:
%macro resid_num_diag(dataset,datavar,label='requested variable',predvar=' ',predlabel='predicte

/*Upload csv*/
%web_drop_table(WORK.winequality);

/*Change path subjectively*/
FILENAME REFFILE '/home/u45035027/Final Project/winequality-red.csv';

PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=WORK.winequality;
GETNAMES=YES;
RUN;

PROC CONTENTS DATA=WORK.winequality; RUN;

%web_open_table(WORK.winequality);

data wine;
set winequality;
rename
'fixed acidity'n=fixedAcidity
'volatile acidity'n=volatileAcidity
'citric acid'n=citricAcid
'residual sugar'n=residualSugar
'free sulfur dioxide'n=freeSulfurDioxide
'total sulfur dioxide'n=totalSulfurDioxide
;
run;

/* Histogram of quality */
proc univariate data=wine;
 var quality;
 histogram quality / midpoints=(0 to 10 by 1);
run;

/*order plot*/
data wine; set wine;
  order = _n_;
proc sgplot data=wine;
   series x=order y= quality / lineattrs=(pattern=solid) ;
  title1 'quality order plot';
```

```
proc surveyselect data=wine seed=12345 out=wine2
     rate=0.1 outall; /* Withold 20% for validation */
run;

/*separate into training and test sets */
data train; set wine2;
if Selected = 0;
run;

data test; set wine2;
if Selected = 1;
run;


proc reg data=train;
model quality = volatileAcidity;
run;

/*transform x-variables*/
data train; set train;
log_residualSugar = log(residualSugar);
log_sulphates = log(sulphates);
log_fixedAcidity = log(fixedAcidity);
log_alcohol = log(alcohol);
run;

/*check model assumptions and outliers */
proc reg data=train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model quality = fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxid
           totalSulfurDioxide density pH sulphates alcohol;
         output out = train r=resid p=pred;

proc corr data=train;
var quality fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxide
           totalSulfurDioxide density pH sulphates alcohol;
title1 'Correlation matrix';
run;

/*Recheck model assumptions after performing log transformations*/
proc reg data = train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
         model quality = log_fixedAcidity volatileAcidity citricAcid log_residualSugar chlorides
           totalSulfurDioxide density pH log_sulphates log_alcohol;
         output out=train r=resid p=pred;
run;

%resid_num_diag(dataset=train, datavar=resid,
label='residual', predvar=pred, predlabel='predicted');
```

```
/*box cox analysis */
proc transreg data=train;
model boxcox(quality / lambda=-2 to 3 by 0.2)
= identity(fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxide
           totalSulfurDioxide density pH sulphates alcohol );
run;

/*R square, mallows Cp, AIC, SBC */
proc reg data = train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
         model quality = log_fixedAcidity volatileAcidity citricAcid log_residualSugar chlorides
           totalSulfurDioxide density pH log_sulphates log_alcohol / selection=AdjRSq Cp AIC SBC
run;

/*Stepwise variable select */
proc reg data = train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
         model quality = log_fixedAcidity volatileAcidity citricAcid log_residualSugar chlorides
           totalSulfurDioxide density pH log_sulphates log_alcohol / selection=stepwise slentry=
run;

/*Lasso */
proc glmselect data=train plots=(criterion ase);
model quality = log_fixedAcidity volatileAcidity citricAcid log_residualSugar chlorides freeSulf
           totalSulfurDioxide density pH log_sulphates log_alcohol / selection=lasso(adaptive ch
output out=out1 p=predlasso;
run;

/*interaction terms*/
data train; set train;
alcohol_sulphates = log_alcohol*log_sulphates;
alcohol_volatileAcidity = log_alcohol*volatileAcidity;
sulphates_volatileAcidity = log_sulphates*volatileAcidity;
total_alcohol = totalSulfurDioxide*log_alcohol;
total_sulphates = totalSulfurDioxide*log_sulphates;
total_volatile = totalSulfurDioxide*volatileAcidity;


proc reg data=train;
model quality = log_alcohol log_sulphates alcohol_sulphates / vif;
title1 'Check for interaction';
run;

proc reg data=train;
model quality = log_alcohol volatileAcidity alcohol_volatileAcidity / vif;
title1 'Check for interaction';
run;

proc reg data=train;
model quality = log_sulphates volatileAcidity sulphates_volatileAcidity / vif;
```

```
title1 'Check for interaction';
run;

proc reg data=train;
model quality = totalSulfurDioxide volatileAcidity total_volatile / vif;
title1 'Check for interaction';
run;

proc reg data=train;
model quality = totalSulfurDioxide log_sulphates total_sulphates / vif;
title1 'Check for interaction';
run;

proc reg data=train;
model quality = totalSulfurDioxide log_alcohol total_alcohol / vif;
title1 'Check for interaction';
run;

/*final model*/
proc reg data=train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model quality = log_alcohol volatileAcidity log_sulphates totalSulfurDioxide
chlorides pH alcohol_sulphates total_volatile total_sulphates / VIF Collin;
store regModel;
run;

proc stdize data=train out=std_train
method=std mult=.0264;
run;
data std_train; set std_train;
alcohol_sulphates = log_alcohol*log_sulphates;
total_sulphates = totalSulfurDioxide*log_sulphates;
total_volatile = totalSulfurDioxide*volatileAcidity;
run;
proc reg data=std_train;
model quality = log_alcohol volatileAcidity log_sulphates totalSulfurDioxide
chlorides pH alcohol_sulphates total_volatile total_sulphates / VIF Collin;
title1 'Check for interaction (standardized scale)';
run;


%resid_num_diag(dataset=train2, datavar=resid3,
label='residual', predvar=pred3, predlabel='predicted');


/*Full Model*/
proc reg data=train noprint;
model quality = fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxid
           totalSulfurDioxide density pH sulphates alcohol;
```

```
store regModel2;
run;

/*Fit a third model with NO variables*/
proc reg data=train noprint;
model quality = ;
store regModel3;
run;

/*Making Predictions*/
proc plm restore=regModel;
score data=test out=newTest predicted;
run;

proc plm restore=regModel2;
score data=test out=newTest2 predicted;
run;

proc plm restore=regModel3;
score data=test out=newTest3 predicted;
run;

/*Estimating errors*/
data newTest; set newTest;
ASE = (quality - Predicted)**2;
run;

data newTest2; set newTest2;
ASE = (quality - Predicted)**2;
run;

data newTest3; set newTest3;
ASE = (quality - Predicted)**2;
run;

/*Calculating means*/
proc means data = newTest;
var ASE;
run;

proc means data = newTest2;
var ASE;
run;

proc means data = newTest3;
var ASE;
run;
```

```
/* Fit a regression tree */
proc hpsplit data=train seed=123 maxdepth=15 maxbranch=2;
model quality = fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxid
              totalSulfurDioxide density pH sulphates alcohol;
code file='/home/u45035027/Final Project/Tree.sas'; /* This saves the tree to a file (need to ch
output out=out2;
run;

/*Scatter plot for regression tree*/
proc sgplot data=out2;
scatter x=quality y=p_quality / markerattrs=(symbol=circlefilled size=6pt);
run;

/* Call the test data and include the tree, this will make predictions on the tree */
data scored;
set test;
%include '/home/u45035027/Final Project/Tree.sas';
run;

/* Now calculate the MSPR as we did in OLS */
data testTree;
set scored;
ASE = (quality - P_quality)**2;
run;
proc means data = testTree;
var ASE;
run;

/* random forest */
proc hpforest data=train seed=134 scoreprole=oob;
input fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxide
          totalSulfurDioxide density pH sulphates alcohol;
target quality;
ods output FitStatistics=fitstats VariableImportance=varimp;
run;



proc corr data=train;
var quality fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxide
              totalSulfurDioxide density pH sulphates alcohol ;
title1 'Correlation matrix';
run;

proc sgscatter data=train;
matrix quality fixedAcidity volatileAcidity citricAcid residualSugar chlorides freeSulfurDioxide
           totalSulfurDioxide density pH sulphates alcohol/
              markerattrs=(symbol=CIRCLEFILLED size=2pt);
```

```
title1 'wine data';
run;
```