# Executive Summary

The purpose of our regression analysis is to research and identify the statistically significant independent variables (predictors) that enable the production of quality (“good”) wine. We utilized the following variables: **Type, Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, Alcohol, and Quality Score.** The variable Quality Score was used as the response variable. We conducted multiple forms of analysis and found that **Type, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Sulphates, Alcohol** were the statistically significant factors contributing to wine quality.

Based on influential magnitude, statistical significant level, negative and positive, we would recommend producers focus on **decreasing the level of volatile.acidity and chlorides and increasing the level of sulphates** during the production process to improve their wine quality in both existing and new wines.

# Introduction

## Background

According to ‘State of the Wine Industry 2018’, over 770 million gallons of wine is consumed per year and with more discerning, internet savvy customers the competition amongst wine producers to make quality wines is high. Our goal is to analyze what factors contribute the most to producing a wine that meets the customers standards of being “good”. We utilized the key chemical components of fixed acidity, volatile acidity, critic acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol as predictors in the analysis. We also created a Type predictor variable to distinguish between red and white wines. The data was chosen from www.kaggle.com which contains a variety of datasets that can be used for various analyses.

## Purpose of Report

This project was undertaken to identify those key variables that help formulate a good quality wine. This information would greatly benefit wine producers in fine tuning their winemaking or vinification process. Our research and analysis were conducted to identify those factors that we can depend on to reliably and consistently produce a good quality wine. Our goal is to provide wine producers with the additional insight and knowledge to produce new high-quality wines as well as improve those wines already being produced.

# Research Methodology

## Explanation of Variables

Prediction variables:

1 - **fixed acidity**: most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

2 - **volatile acidity**: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste

3 - **citric acid**: found in small quantities, citric acid can add ‘freshness’ and flavor to wines

4 - **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

5 - **chlorides**: the amount of salt in the wine

6- **free sulfur dioxide**: the free form of SO2SO2; exists in equilibrium between molecular SO2SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine

7 - **total sulfur dioxide**: amount of free and bound forms of SO2SO2; in low concentrations, SO2SO2 is mostly undetectable in wine, but at free SO2SO2 concentrations over 50 ppm, SO2SO2 becomes evident in the nose and taste of wine

8 - **density**: the density of water is close to that of water depending on the percent alcohol and sugar content

9 - **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

10 - **sulphates**: a wine additive which can contribute to sulfur dioxide gas (S02S02) levels, which acts as an antimicrobial and antioxidant

11 - **alcohol**: the percent alcohol content of the wine

Output variable (based on sensory data):

12 - **quality** (score between 0 and 10)

## Transformations

We initially find an ascending trend of residual variance when fitting multiple linear regression with all the variables except quality score. In order to satisfying the assumption of linear regression, a transform of response is needed. By applying the method of Box Cox in R, we transform the response to the power of 0.7475.

A close up of a map

Description automatically generated

We diagnose the residuals, the mean of residuals on each level is very close to zero, which indicates the linear relationship is satisfied. Besides, the constancy of residual variance improves slightly. Therefore, it is reliable to use the transformed response throughout the analysis.

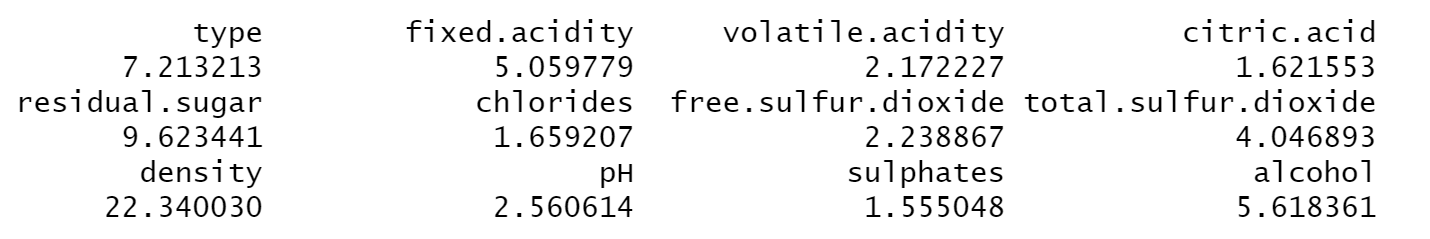
A close up of a map

Description automatically generated

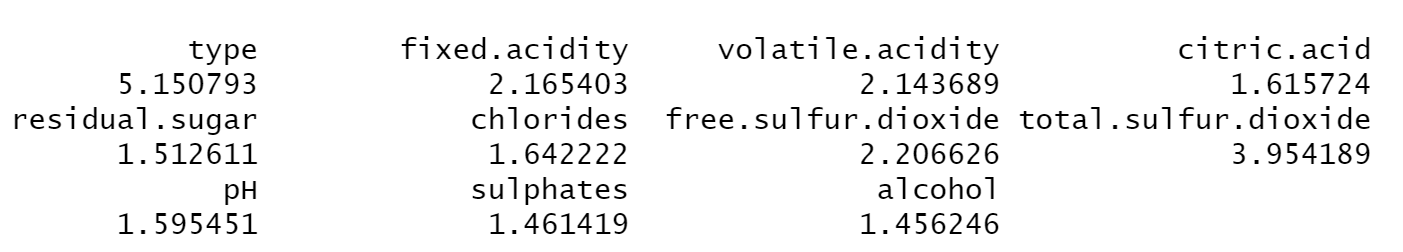
## Challenges

The main challenges we encountered are multicollinearity, variable selection and strong influential points.

First, there are multicollinearity between current predictors. Our model contains a series of indexes, some of them seem to have relationships based on basic logical inference, for example: there are two kinds of sulfur dioxide, free and bound, free sulfur dioxide accounts for the former while both combine to form the total sulfur dioxide. We calculate the VIF of our model to make sure whether there is multicollinearity or not. The VIF of each predictor is as follows,



By looking at the VIF of each predictor, the VIF of density is nearly 22, which indicates a strong multicollinearity. Obviously, the overall density of water is decided by all the density of chemical elements in the wine. There are different ways to solve the multicollinearity problem. Our model has no polynomial term and our goal is not prediction. Here, we remove density from the predictors. New VIF of each predictor is showed as follows. It can be shown that the multicollinearity problem is solved.

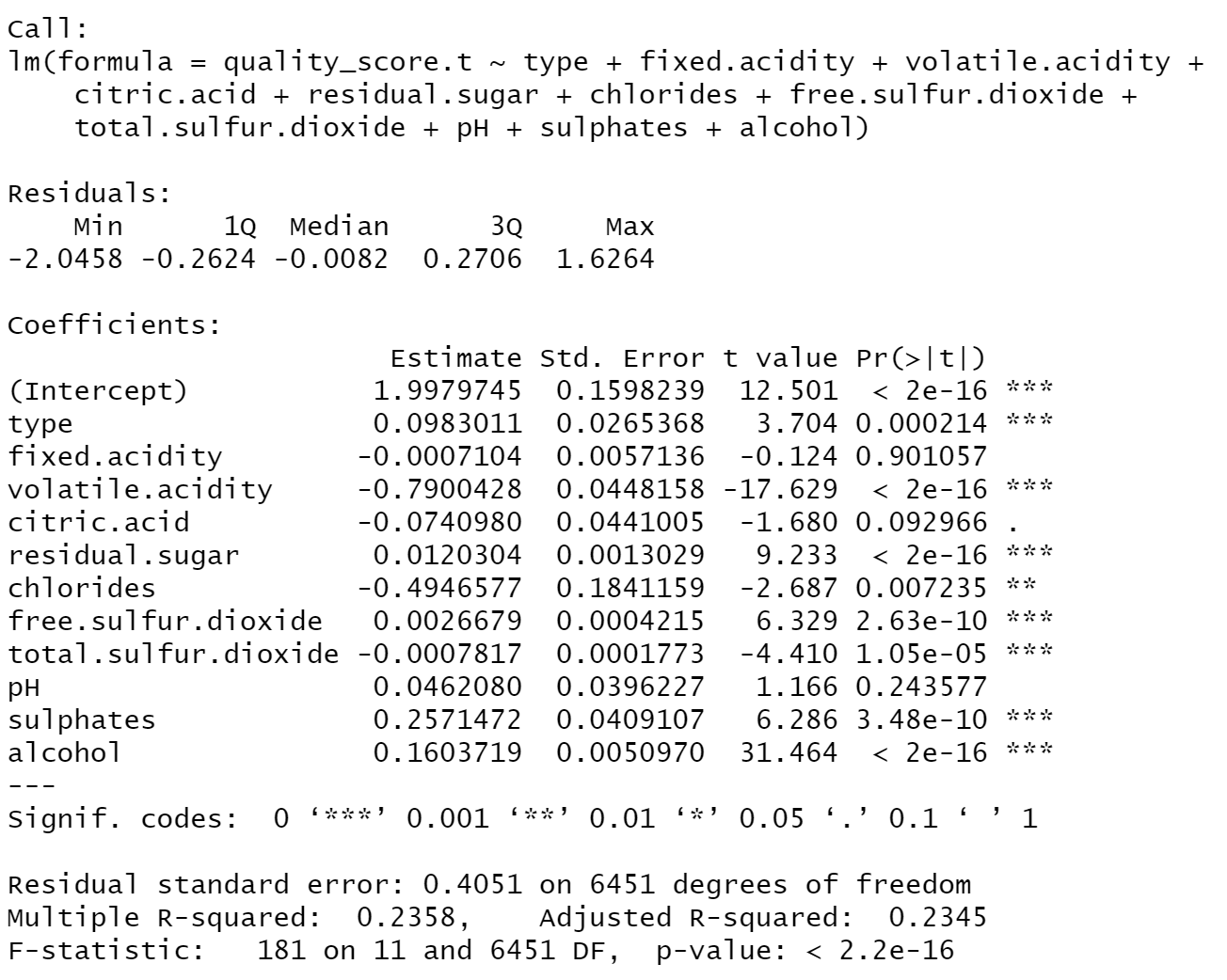


Our hypothesis:

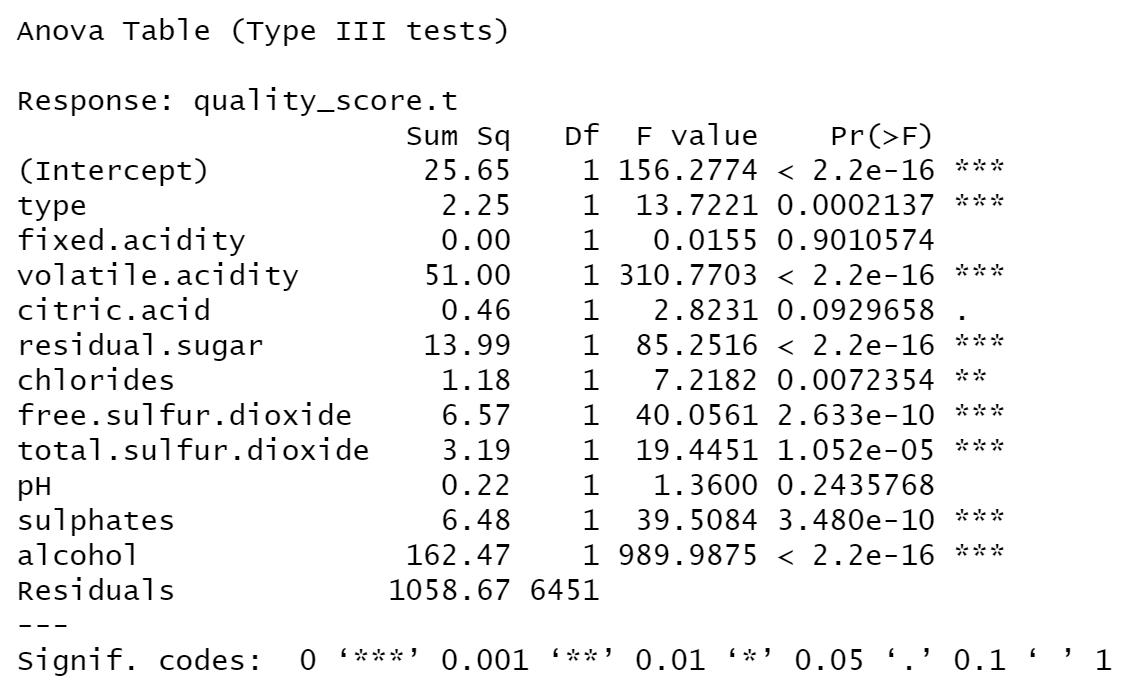
H0: Neither type, fixed.acidity, volatile.acidity, critic.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates nor alcohol predictors has a statistical impact on final quality score of wine

H1: At least one of the type, fixed.acidity, volatile.acidity, critic.acid, residual.sugar, chlorides, free.sulfur.dioxide, total.sulfur.dioxide, pH, sulphates and alcohol predictors has a statistical impact on final quality score of wine

From the following summary table, we can see the p-value of F-statistic is extremely small, which means we can reject the null hypothesis. So, we can say that our linear model is effective.



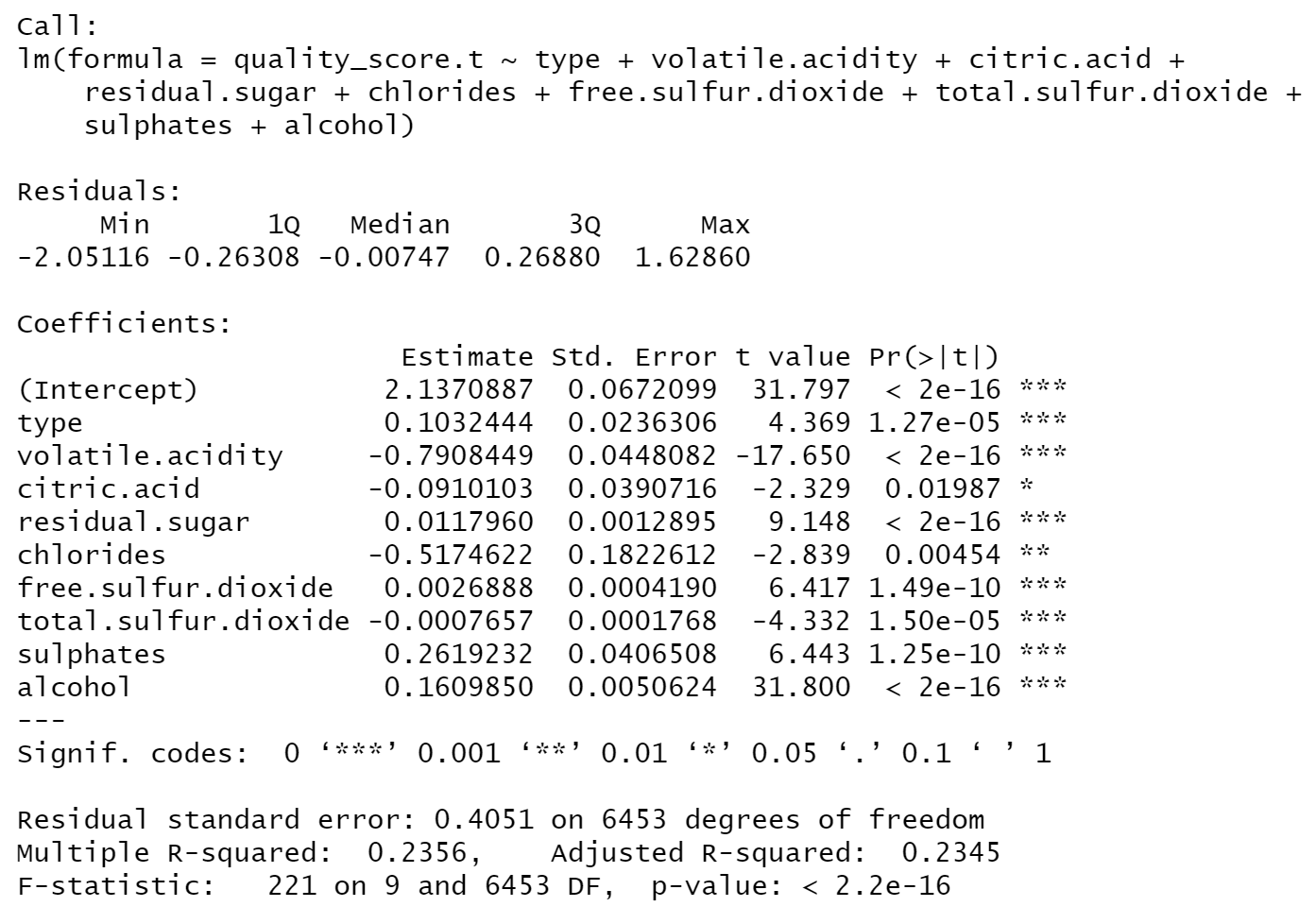
After proper transformation, we check the significance of all the predictors by extra SSE. The result shows that fixed acidity, citric acid and pH do not significantly help to explain the quality score.



However, we cannot simply delete all these variables since some combination of them may be significantly helpful. In order to decide which predictor should be included, we use the method of stepwise regression (back and forward stepwise return the same result) and get the potential best model as follows

**lm(quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar + type + free.sulfur.dioxide + total.sulfur.dioxide + chlorides + citric.acid)**

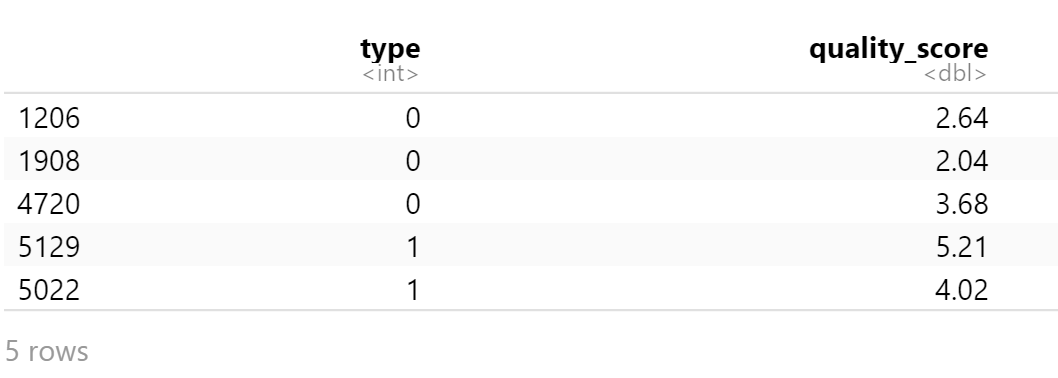
The F-statistic value of this model is 221 and the p-value is less than 2.2\*10^ (-16), which is extremely small. All the predictor variables in the model are significant based on 0.05 significance level. The p-value of variable “citric.acid” is 0.0198, which is relatively greater than p-values of other variables. R-squared value is not very high, which shows that our model can only explain part of change in quality score. Also, our original full model does not have a high R-squared, which means these predictors are just part of the elements that explain the quality of wine. There are some unknown predictors which can explain the remaining variability in the quality of wine. All these numbers tell us that the model is not bad.



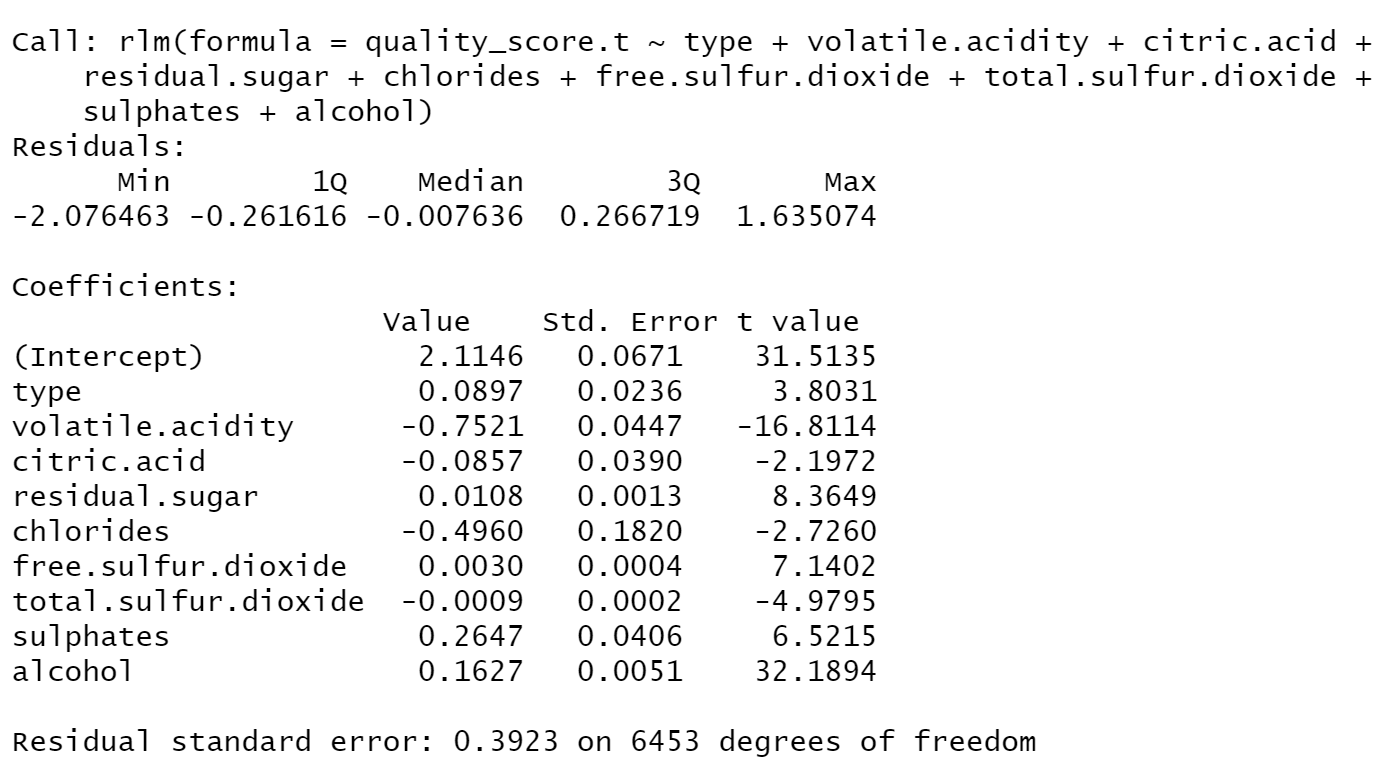
We then perform best subsets regression to get support for the above model. We found that the model contains 9 predictors best fit all of the criteria, which matches the model given by stepwise regression. To get to this conclusion, we compare the adjusted R2, BIC and Mallow’s Cp. We search for the one with lower BIC and higher R2, while having Mallow’s Cp no higher than the number of predictors plus 1. The model contains 9 predictors perfectly satisfies all the requirements, which implies that the model does not have substantial bias and it best explains the data.

Besides, we found several influential points that may strongly influence the regression line. By closely looking at those influential points, we can see that most of the points are from low quality wine. The residuals of them are negative, which indicates the current model overestimates those points.

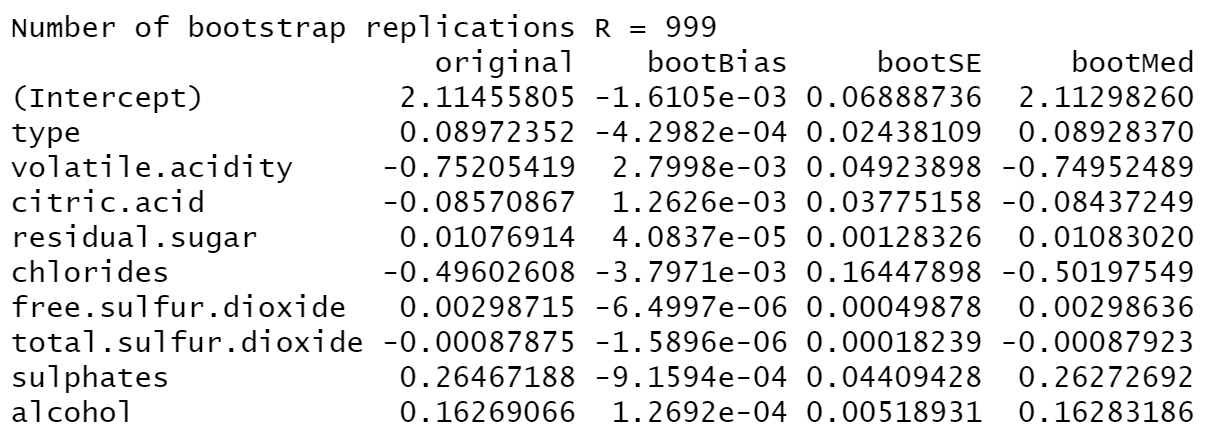
A screenshot of a cell phone

Description automatically generated

Keeping putting the same weights on all the predictors will probably overestimate the quality on each level. Therefore, we use Huber’s weights and apply IRLS, the result shows as follows,



In order to get the confidence interval of each coefficients, we conduct bootstrap for regression analysis. Results show that there are not significant differences between the original estimation and the bootstrap estimation, which means we can still use the original estimation of coefficients.

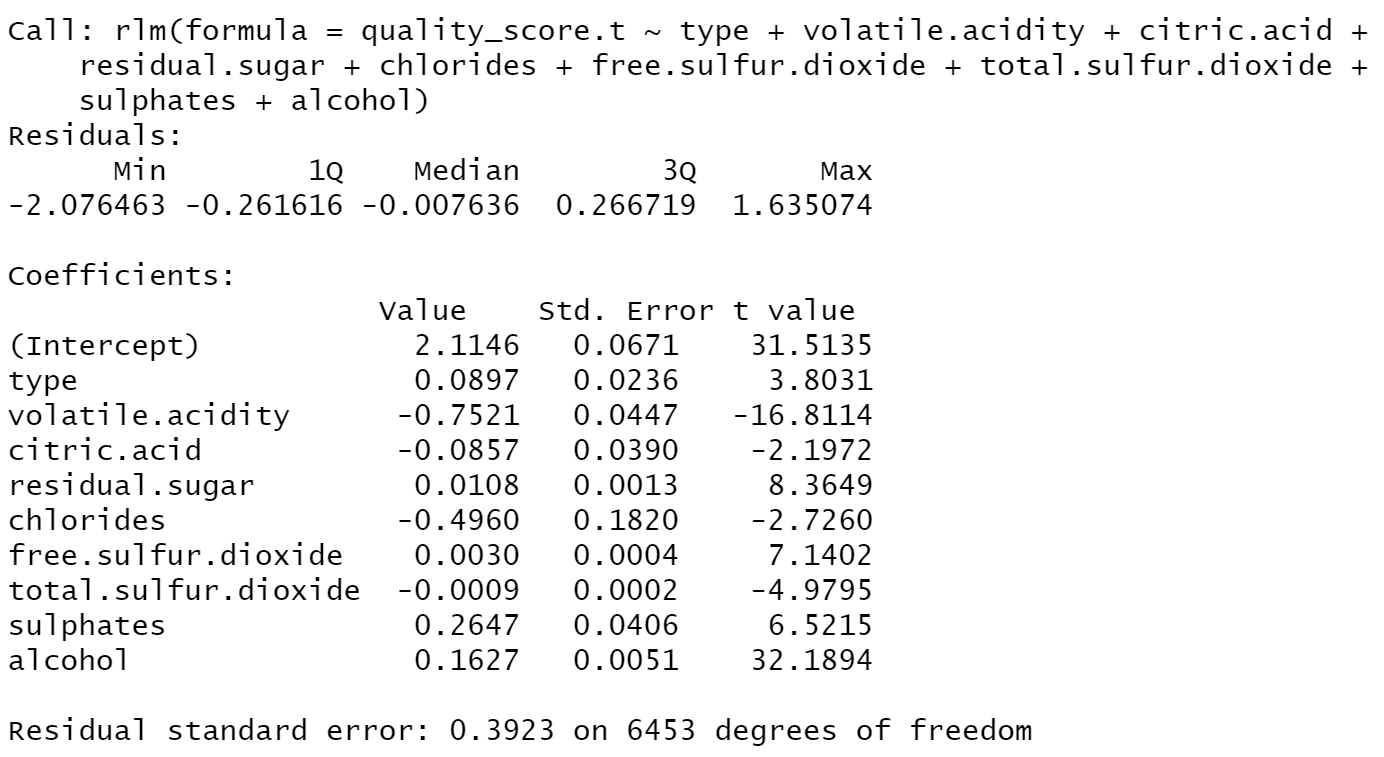


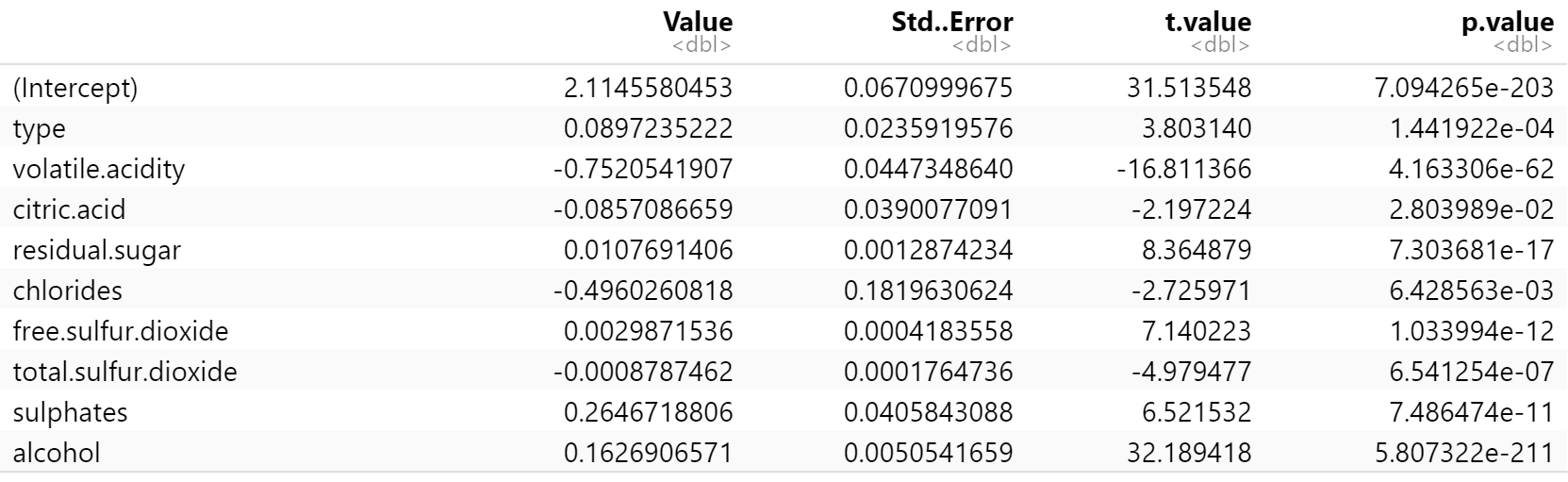
# Results

Let’s review the result of our final model.

The model line is:

The Summary table (we add p values in the table):





From the summary table, we can clearly see the residual standard error getting lower from 0.4051 to 0.3923, which means our model become better explaining the relationship between predictors and the response. We calculate the adjusted-R square of final RLM model, the value is very close to OLS model, which is 0.2345. The number means our model can interpret 23.45% of the whole change in wine quality.

Diagnostic Plots:

A close up of a map

Description automatically generated

From the result of these four plots, we can fairly say that the assumptions of linear regression model are fulfilled well in our model. The first plot shows the linear relationship is perfect. The second plot shows the normal assumption is also perfectly fulfilled. The third plot shows the constant variance can be basically fulfilled and the last plot shows the high residuals and leverage points exist but don’t generate serious influence on fitting the model. So, this model will be our final model.

# Summary and Recommendations

Based on our regression model, we can provide useful insight to wine producers on ways to improve their manufacturing process. We can see that all the predictors have passed the significant test in our final model, but the degree of their statistical impact are different.

For our only binary qualitive predictor---wine type (0 means white wine while 1 means red wine), the coefficient is 0.09, a positive number, which means when others remain the same, if the product is red wine, then it will have a bit higher quality score than white one. There are several reasons to explain this point. First, because our model can only explain nearly 24% change of the total change in quality score, so the source of difference between the red and white wine may be some unknown variables which are the attributes that can be inherently different in red and white wine. Secondly, people may have some preferences toward red wine thus they tend to give red wine a bit higher score. However, based on the current model and information, we cannot give a certain conclusion, so the indication of this point is that we should make a further investigation about difference between red and white wine and try to find some indexes which may influence the quality of wine but can be significantly different in two types.

For other quantitive variables, they are the chemical component indexes of wine. Let’s analyze the model from the perspective of negative or positive coefficients. We can observe coefficients of volatile.acidity, chlorides, total.sulfur.dioxide, critic.acidity are negative, which means they have negative relationship with wine quality score. For this point, if we can lower the level of these three indexes during our wine procedure, then we may gain a better-quality wine. For other variables which have positive coefficients (residual.sugar, free.sulfur.dioxide, sulphates, alcohol) they tell us that producers should increase the level of these indexes during wine production.

Now let’s look at the influencial magnitude of each coefficient. We can clearly see that the absolute number of coefficients of volatile.acidity, chlorides, sulphates have relatively larger numbers than others. So when we change the same unit for all these variables, these three can generate the larger impact on wine quality score. That is to say, we should focus more on these three indexes when we are able to make some changes among all the indexes.

Third point is from the perspective of significance level of the coefficients of all our predictors. We can clearly see that pH and chlorides have larger p-value than others, which means they have less statistical impact than other predictors. Regardless, the magnitude of coefficient of chlorides is relatively large. As always, producers will always have to face some trade-offs and it may be very hard to improve all the aspects at the same time and at same cost.

In conclusion, our recommendation to wine producers is they should focus more on ways to decrease the level of volatile.acidity and chlorides and increase the level of sulphates during vinification to improve their wine quality.

# Appendix

Result of stepwise regression (forward):

Start: AIC=-9951.77

quality\_score.t ~ 1

Df Sum of Sq RSS AIC

+ alcohol 1 216.479 1168.9 -11047.9

+ volatile.acidity 1 80.703 1304.7 -10337.7

+ chlorides 1 45.128 1340.3 -10163.8

+ type 1 14.295 1371.1 -10016.8

+ citric.acid 1 7.126 1378.3 -9983.1

+ fixed.acidity 1 6.494 1378.9 -9980.1

+ free.sulfur.dioxide 1 3.754 1381.6 -9967.3

+ total.sulfur.dioxide 1 1.917 1383.5 -9958.7

+ sulphates 1 1.458 1383.9 -9956.6

+ residual.sugar 1 0.896 1384.5 -9953.9

+ pH 1 0.547 1384.8 -9952.3

<none> 1385.4 -9951.8

Step: AIC=-11047.89

quality\_score.t ~ alcohol

Df Sum of Sq RSS AIC

+ volatile.acidity 1 70.702 1098.2 -11449.1

+ free.sulfur.dioxide 1 21.657 1147.2 -11166.8

+ residual.sugar 1 21.597 1147.3 -11166.4

+ type 1 10.670 1158.2 -11105.2

+ chlorides 1 9.175 1159.7 -11096.8

+ citric.acid 1 7.939 1161.0 -11089.9

+ total.sulfur.dioxide 1 6.748 1162.2 -11083.3

+ sulphates 1 1.608 1167.3 -11054.8

+ fixed.acidity 1 1.296 1167.6 -11053.1

+ pH 1 1.082 1167.8 -11051.9

<none> 1168.9 -11047.9

- alcohol 1 216.479 1385.4 -9951.8

Step: AIC=-11449.13

quality\_score.t ~ alcohol + volatile.acidity

Df Sum of Sq RSS AIC

+ sulphates 1 10.558 1087.7 -11510

+ type 1 8.616 1089.6 -11498

+ residual.sugar 1 7.946 1090.3 -11494

+ free.sulfur.dioxide 1 2.848 1095.4 -11464

+ pH 1 1.564 1096.6 -11456

+ total.sulfur.dioxide 1 1.536 1096.7 -11456

+ fixed.acidity 1 0.512 1097.7 -11450

<none> 1098.2 -11449

+ citric.acid 1 0.154 1098.0 -11448

+ chlorides 1 0.036 1098.2 -11447

- volatile.acidity 1 70.702 1168.9 -11048

- alcohol 1 206.478 1304.7 -10338

Step: AIC=-11509.57

quality\_score.t ~ alcohol + volatile.acidity + sulphates

Df Sum of Sq RSS AIC

+ residual.sugar 1 11.375 1076.3 -11576

+ free.sulfur.dioxide 1 4.400 1083.2 -11534

+ type 1 2.631 1085.0 -11523

+ chlorides 1 1.095 1086.5 -11514

+ citric.acid 1 0.854 1086.8 -11513

+ pH 1 0.640 1087.0 -11511

<none> 1087.7 -11510

+ total.sulfur.dioxide 1 0.308 1087.3 -11509

+ fixed.acidity 1 0.024 1087.6 -11508

- sulphates 1 10.558 1098.2 -11449

- volatile.acidity 1 79.652 1167.3 -11055

- alcohol 1 206.013 1293.7 -10390

Step: AIC=-11575.52

quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar

Df Sum of Sq RSS AIC

+ type 1 7.359 1068.9 -11618

+ total.sulfur.dioxide 1 3.848 1072.4 -11597

+ pH 1 1.983 1074.3 -11585

+ citric.acid 1 1.586 1074.7 -11583

+ free.sulfur.dioxide 1 1.276 1075.0 -11581

+ chlorides 1 0.373 1075.9 -11576

<none> 1076.3 -11576

+ fixed.acidity 1 0.011 1076.3 -11574

- residual.sugar 1 11.375 1087.7 -11510

- sulphates 1 13.987 1090.3 -11494

- volatile.acidity 1 65.997 1142.3 -11193

- alcohol 1 212.088 1288.4 -10415

Step: AIC=-11617.86

quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar +

type

Df Sum of Sq RSS AIC

+ free.sulfur.dioxide 1 3.763 1065.2 -11639

+ citric.acid 1 1.911 1067.0 -11627

+ chlorides 1 1.737 1067.2 -11626

+ fixed.acidity 1 1.126 1067.8 -11623

+ pH 1 1.109 1067.8 -11623

<none> 1068.9 -11618

+ total.sulfur.dioxide 1 0.158 1068.8 -11617

- sulphates 1 4.728 1073.6 -11591

- type 1 7.359 1076.3 -11576

- residual.sugar 1 16.103 1085.0 -11523

- volatile.acidity 1 64.820 1133.7 -11239

- alcohol 1 218.785 1287.7 -10416

Step: AIC=-11638.65

quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar +

type + free.sulfur.dioxide

Df Sum of Sq RSS AIC

+ total.sulfur.dioxide 1 3.485 1061.7 -11658

+ citric.acid 1 1.880 1063.3 -11648

+ chlorides 1 1.772 1063.4 -11647

+ pH 1 0.810 1064.3 -11642

+ fixed.acidity 1 0.694 1064.5 -11641

<none> 1065.2 -11639

- free.sulfur.dioxide 1 3.763 1068.9 -11618

- sulphates 1 4.288 1069.4 -11615

- type 1 9.846 1075.0 -11581

- residual.sugar 1 12.001 1077.2 -11568

- volatile.acidity 1 62.073 1127.2 -11275

- alcohol 1 222.541 1287.7 -10414

Step: AIC=-11657.83

quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar +

type + free.sulfur.dioxide + total.sulfur.dioxide

Df Sum of Sq RSS AIC

+ chlorides 1 1.802 1059.9 -11667

+ citric.acid 1 1.370 1060.3 -11664

+ pH 1 0.942 1060.7 -11662

+ fixed.acidity 1 0.548 1061.1 -11659

<none> 1061.7 -11658

- type 1 2.197 1063.9 -11646

- total.sulfur.dioxide 1 3.485 1065.2 -11639

- sulphates 1 5.309 1067.0 -11628

- free.sulfur.dioxide 1 7.091 1068.8 -11617

- residual.sugar 1 13.764 1075.4 -11577

- volatile.acidity 1 55.838 1117.5 -11329

- alcohol 1 188.843 1250.5 -10602

Step: AIC=-11666.81

quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar +

type + free.sulfur.dioxide + total.sulfur.dioxide + chlorides

Df Sum of Sq RSS AIC

+ citric.acid 1 0.890 1059.0 -11670

+ pH 1 0.610 1059.2 -11668

+ fixed.acidity 1 0.468 1059.4 -11668

<none> 1059.9 -11667

- chlorides 1 1.802 1061.7 -11658

- type 1 2.937 1062.8 -11651

- total.sulfur.dioxide 1 3.515 1063.4 -11647

- sulphates 1 6.456 1066.3 -11630

- free.sulfur.dioxide 1 7.154 1067.0 -11625

- residual.sugar 1 13.179 1073.0 -11589

- volatile.acidity 1 53.404 1113.3 -11351

- alcohol 1 165.146 1225.0 -10733

Step: AIC=-11670.24

quality\_score.t ~ alcohol + volatile.acidity + sulphates + residual.sugar +

type + free.sulfur.dioxide + total.sulfur.dioxide + chlorides +

citric.acid

Df Sum of Sq RSS AIC

<none> 1059.0 -11670

+ pH 1 0.299 1058.7 -11670

+ fixed.acidity 1 0.079 1058.9 -11669

- citric.acid 1 0.890 1059.9 -11667

- chlorides 1 1.323 1060.3 -11664

- total.sulfur.dioxide 1 3.079 1062.0 -11654

- type 1 3.133 1062.1 -11653

- free.sulfur.dioxide 1 6.757 1065.7 -11631

- sulphates 1 6.813 1065.8 -11631

- residual.sugar 1 13.732 1072.7 -11589

- volatile.acidity 1 51.120 1110.1 -11368

- alcohol 1 165.948 1224.9 -10731

Result of stepwise regression (backward):

Start: AIC=-11668.09

quality\_score.t ~ type + fixed.acidity + volatile.acidity + citric.acid +

residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +

pH + sulphates + alcohol

Df Sum of Sq RSS AIC

- fixed.acidity 1 0.003 1058.7 -11670

- pH 1 0.223 1058.9 -11669

<none> 1058.7 -11668

- citric.acid 1 0.463 1059.1 -11667

- chlorides 1 1.185 1059.8 -11663

- type 1 2.252 1060.9 -11656

- total.sulfur.dioxide 1 3.191 1061.9 -11651

- sulphates 1 6.484 1065.2 -11631

- free.sulfur.dioxide 1 6.574 1065.2 -11630

- residual.sugar 1 13.991 1072.7 -11585

- volatile.acidity 1 51.000 1109.7 -11366

- alcohol 1 162.466 1221.1 -10747

Step: AIC=-11670.07

quality\_score.t ~ type + volatile.acidity + citric.acid + residual.sugar +

chlorides + free.sulfur.dioxide + total.sulfur.dioxide +

pH + sulphates + alcohol

Df Sum of Sq RSS AIC

- pH 1 0.299 1059.0 -11670

<none> 1058.7 -11670

- citric.acid 1 0.580 1059.2 -11668

+ fixed.acidity 1 0.003 1058.7 -11668

- chlorides 1 1.185 1059.9 -11665

- type 1 2.656 1061.3 -11656

- total.sulfur.dioxide 1 3.199 1061.9 -11653

- sulphates 1 6.494 1065.2 -11632

- free.sulfur.dioxide 1 6.674 1065.3 -11632

- residual.sugar 1 14.030 1072.7 -11587

- volatile.acidity 1 51.018 1109.7 -11368

- alcohol 1 163.694 1222.4 -10743

Step: AIC=-11670.24

quality\_score.t ~ type + volatile.acidity + citric.acid + residual.sugar +

chlorides + free.sulfur.dioxide + total.sulfur.dioxide +

sulphates + alcohol

Df Sum of Sq RSS AIC

<none> 1059.0 -11670

+ pH 1 0.299 1058.7 -11670

+ fixed.acidity 1 0.079 1058.9 -11669

- citric.acid 1 0.890 1059.9 -11667

- chlorides 1 1.323 1060.3 -11664

- total.sulfur.dioxide 1 3.079 1062.0 -11654

- type 1 3.133 1062.1 -11653

- free.sulfur.dioxide 1 6.757 1065.7 -11631

- sulphates 1 6.813 1065.8 -11631

- residual.sugar 1 13.732 1072.7 -11589

- volatile.acidity 1 51.120 1110.1 -11368

- alcohol 1 165.948 1224.9 -10731

Result of best subsets regression:

