Physicochemical Effects on White Wine Quality

STAT632 Section1
Final Project

Hui Huang(hy9489)
Yun Jing(qi2679)
Xuan Zhou(up5758)

1.Introduction

White wine is a wine that is fermented without skin contact. It is produced by the alcoholic fermentation of the non-colored pulp of grapes. It has been existing for at least 2500 years in the world. With the development of technology and the high demand of the white wine all around the world, the white wine industry needs a standard quality control system to test the white wine quality.

In this report, we use the Portuguese "Vinho verde" white wine data set to do the multiple linear regression. Our goal is to find the suitable MLR model to test "Vinho verde" white wine quality using physicochemical factors. The analysis will reveal the important relationship between "Vinho verde" white wine with different physicochemical factors. In order to make sure the model is suitable; we also do the Cross Validation to validate our regression model. Physicochemical factors are very professional laboratory indicator, such as different kinds of acidity, PH value and density. The wine producers and suppliers can use this final model to test the wine quality much easier with standard lab test results. The standardized wine quality control model can also benefit customers when they choose the white wine.

2.Data Description

2.1 Data source

The data sets are public available in https://www.kaggle.com/danielpanizzo/wine-quality. We also collect useful information from http://www.vinhoverde.pt/en/about-vinho-verde official website. There are two data sets, using red and white wine samples in 2009. We choose the white wine data sets to analyze.

2.2 Data dimension

There are 4898 observations on 12 variables in the white wine data set. The data has no missing attribute value.

2.3 Response and predictor variables

The response variable is Quality which is based on sensory data (median of at least 3 evaluations made by wine experts). The score of each wine quality is between 0(very bad) to 10(very excellent). The quality of the white wine data is approximately normally distributed (see Figure1).

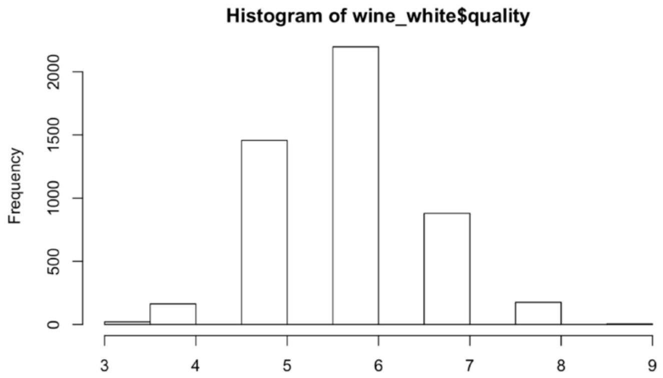
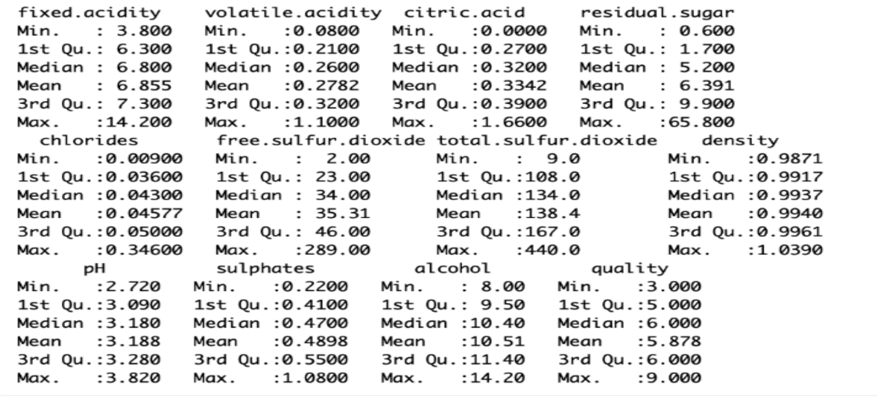


Figure1

The predictors are 11 physicochemical factors. The descriptions are: 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 SO2 exists in equilibrium between molecular SO2 (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 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. 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 (S02) levels, which acts as an antimicrobial and antioxidant. 11.alcohol: the percent alcohol content of the wine.

2.4 Summary Statistics and graphical displays of data set



```
 fixed.acidity    volatile.acidity  citric.acid     residual.sugar
 Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
 1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
 Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
 Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
 3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
 Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
   chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
 1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
 Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
 Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
 3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
 Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
       pH          sulphates         alcohol          quality
 Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
 1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
 Median :3.180   Median :0.4700   Median :10.40   Median :6.000
 Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
 3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
 Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

Figure 2

From the Figure 2, the median and mean values are pretty similar for all variables. The first and third quantile are relatively close to the center. We surmise the distributions might symmetrical and concentrated.

From the data matrix, we can tell there are some relationship between response and predictors. However, due to the data set is large and we have more than 10 predictor it is very hard to see the clear relationships (We do not include the matrix figure here). We need to do further investigation.

3. Methods

For this project, we focus on multiple linear regression model. The methods used in this project includes: variable selection, assumptions check, leverage and outliers check, Box-Cox transformation, and cross-validation.

3.1 Original Full Model

We fit a multiple linear regression model with quality as the response and all the other attributes as the predictors, which is the original full model. Then we use variable selection methods to select the best model for our research questions.

3.2 Model Selection

The purpose of the model selection is to avoid that the model with too many predictors. “Over-fit” model performs poorly when making future predictions. In this step, we use four approaches: the Akaike information criterion (AIC), Bayesian information criterion (BIC), adjusted R^2, and backwards stepwise selection. Since different methods might select different predictors, we used all four approaches to check and guarantee that we can select a best model. The result of the four criterions is shown on the results section. After model selection, we got a reduced model with eight predictors: Fixed.acidity, Volatile.acidity, Residual.sugar, Free.sulfur. Dioxide, Density, pH, Sulphates, and Alcohol.

3.3 Assumptions Check

After we select the most suitable model, we need to use some diagnostic techniques to check assumptions: constant variance, linearity, and normality. Regression diagnostics can also suggest improvements of model, which means model building is an iterative and interactive process. The diagnostic results for our reduced model indicate that we need to do transformation to get a better model to satisfy the assumptions. Diagnostic techniques also help us to check leverage points and outliers.

3.4 Box-Cox Transformation

The assumptions are not satisfied, we use Box-Cox method to estimate transformations. For this data, the quality is an integral variable from 0 to 10, in this case, we do not need to transform response variable. We only consider to transform predictors based on the scatter matrices plot and the summary(powerTransfom()) results. After transformation, we also need to check assumptions (residuals vs. fitted value plot and QQ-plot) using our final model. We also compare the adjusted R^2, and AIC of the reduced model and the final model.

3.5 Cross-validation and Accuracy

Finally, we use cross-validation to validate our model. We split the data into 70% training set and 30% testing set. Cross-validation is a more direct approach is to estimate the test error by holding out a subset of observations from the model fitting process, and then applying the statistical model to make predictions on those withheld observations. Based on the validation set approach, we also fitted ordinary least squares model, backwards stepwise model, ridge model, and Lasso model and compared their RMSE to get the model with best predictive performance.

4. Results

4.1 Models

Original full model: $quality = \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 citric.acid + \beta_4 residual.sugar + \beta_5 chlorides + \beta_6 free.sulfur.dioxide + \beta_7 total.sulfur.dioxide + \beta_8 density + \beta_9 pH + \beta_{10} sulphates + \beta_{11} alcohol + e$

Reduced model: $quality = \beta_0 + \beta_1 fixed.acidity + \beta_2 volatile.acidity + \beta_3 residual.sugar + \beta_4 free.sulfur.dioxide + \beta_5 density + \beta_6 pH + \beta_7 sulphates + \beta_8 alcohol + e$

Final model: $quality = \beta_0 + \beta_1 log(fixed.acidity) + \beta_2 log(volatile.acidity) + \beta_3 sqrt(residual.sugar) + \beta_4 sqrt(free.sulfur.dioxide) + \beta_5 density + \beta_6 1/sqrt(pH) + \beta_7 log(sulphates) + \beta_8 sqrt(alcohol) + e$

4.2 Model Selection

The Figure 3 shows the adjusted R^2, AIC, and BIC versus the number of predictors for the best subset procedure. The adjusted R^2 ,AIC and BIC all select a model with 8 predictors.

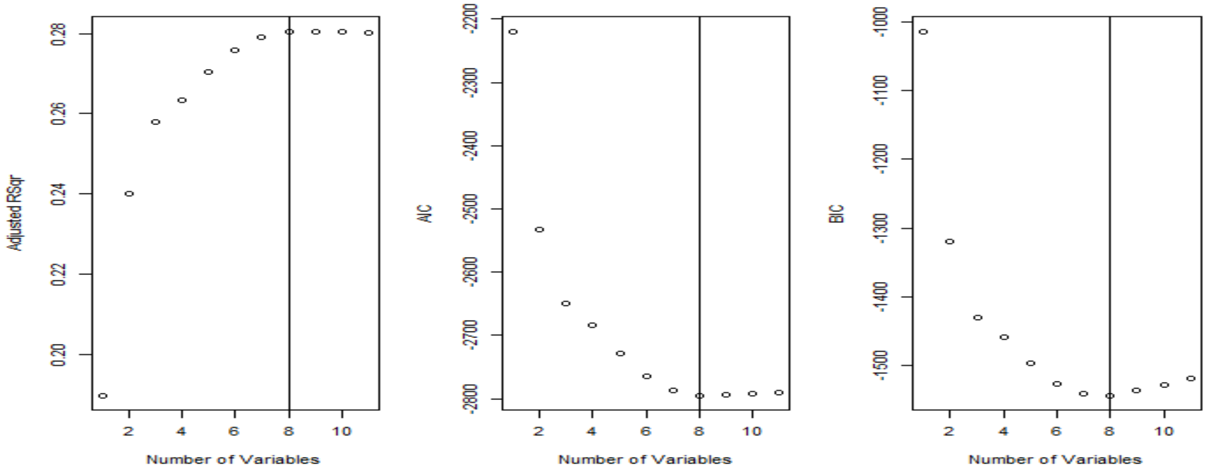
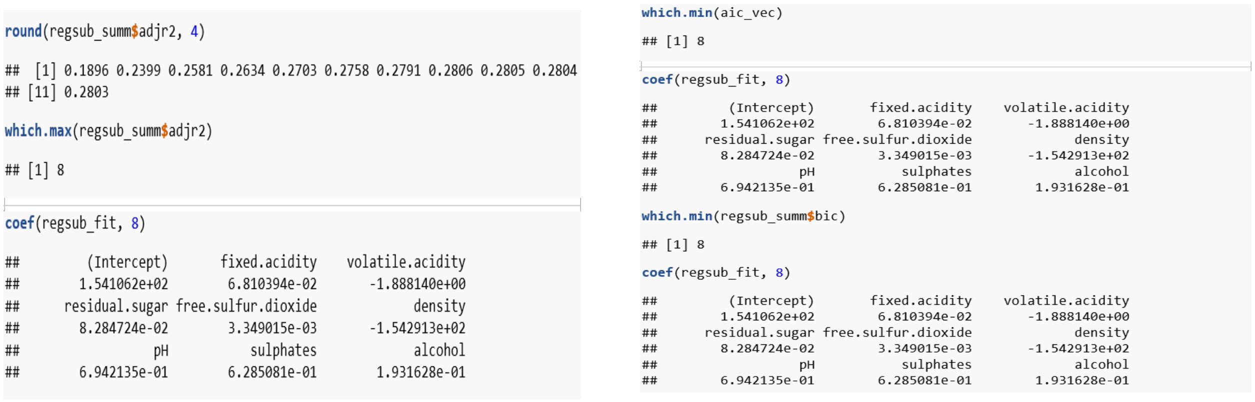


Figure 3

The coefficients result below (Table 1) are the predictors for the best model by using the adjusted R^2, AIC, and BIC, which shows the same 8 predictors.

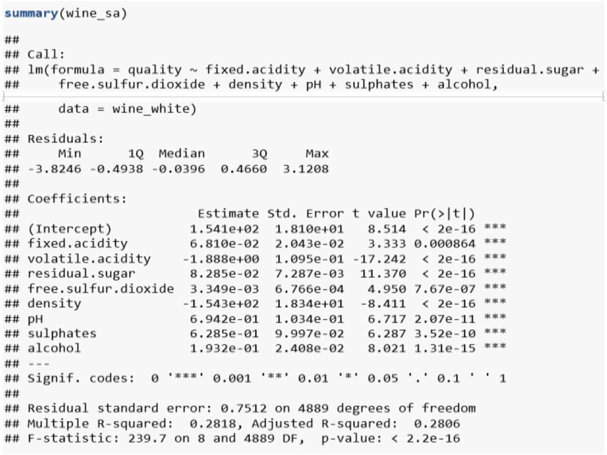


```
round(regsub_summ$adjr2, 4)

##  [1] 0.1896 0.2399 0.2581 0.2634 0.2703 0.2758 0.2791 0.2806 0.2805 0.2804
## [11] 0.2803

which.max(regsub_summ$adjr2)

## [1] 8

coef(regsub_fit, 8)

##        (Intercept)       fixed.acidity    volatile.acidity
##       1.541062e+02        6.810394e-02       -1.888140e+00
##     residual.sugar free.sulfur.dioxide             density
##       8.284724e-02        3.349015e-03       -1.542913e+02
##                 pH           sulphates             alcohol
##       6.942135e-01        6.285081e-01        1.931628e-01
```

```
which.min(aic_vec)

## [1] 8

coef(regsub_fit, 8)

##        (Intercept)       fixed.acidity    volatile.acidity
##       1.541062e+02        6.810394e-02       -1.888140e+00
##     residual.sugar free.sulfur.dioxide             density
##       8.284724e-02        3.349015e-03       -1.542913e+02
##                 pH           sulphates             alcohol
##       6.942135e-01        6.285081e-01        1.931628e-01

which.min(regsub_summ$bic)

## [1] 8

coef(regsub_fit, 8)

##        (Intercept)       fixed.acidity    volatile.acidity
##       1.541062e+02        6.810394e-02       -1.888140e+00
##     residual.sugar free.sulfur.dioxide             density
##       8.284724e-02        3.349015e-03       -1.542913e+02
##                 pH           sulphates             alcohol
##       6.942135e-01        6.285081e-01        1.931628e-01
```

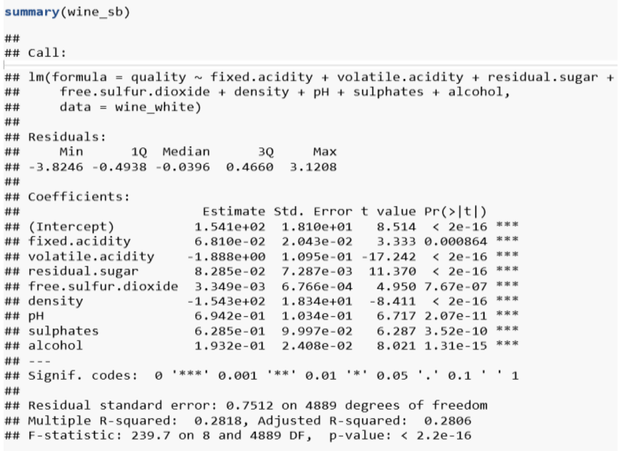
Table 1

The summary() results below(Table 2 is AIC, Table 3 is BIC) are the predictors for the best model by using backwards stepwise selection of AIC and BIC respectively, both which show the same 8 predictors.



```
summary(wine_sa)

##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = wine_white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8246 -0.4938 -0.0396  0.4660  3.1208
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
## fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
## volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
## residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
## free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
## density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
## pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
## sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
## alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7512 on 4889 degrees of freedom
## Multiple R-squared:  0.2818, Adjusted R-squared:  0.2806
## F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

Table2



```
summary(wine_sb)

##
## Call:
## lm(formula = quality ~ fixed.acidity + volatile.acidity + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + alcohol,
##     data = wine_white)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8246 -0.4938 -0.0396  0.4660  3.1208
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.541e+02  1.810e+01   8.514  < 2e-16 ***
## fixed.acidity        6.810e-02  2.043e-02   3.333 0.000864 ***
## volatile.acidity    -1.888e+00  1.095e-01 -17.242  < 2e-16 ***
## residual.sugar       8.285e-02  7.287e-03  11.370  < 2e-16 ***
## free.sulfur.dioxide  3.349e-03  6.766e-04   4.950 7.67e-07 ***
## density             -1.543e+02  1.834e+01  -8.411  < 2e-16 ***
## pH                   6.942e-01  1.034e-01   6.717 2.07e-11 ***
## sulphates            6.285e-01  9.997e-02   6.287 3.52e-10 ***
## alcohol              1.932e-01  2.408e-02   8.021 1.31e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7512 on 4889 degrees of freedom
## Multiple R-squared:  0.2818, Adjusted R-squared:  0.2806
## F-statistic: 239.7 on 8 and 4889 DF,  p-value: < 2.2e-16
```

Table3

4.3 Assumption Check

The Figure 4 below shows the residuals vs. fitted value plot and QQ-plot of the reduced model.

The Figure 5 below shows the residuals vs. fitted value plot and QQ-plot of the final model (after transformation).

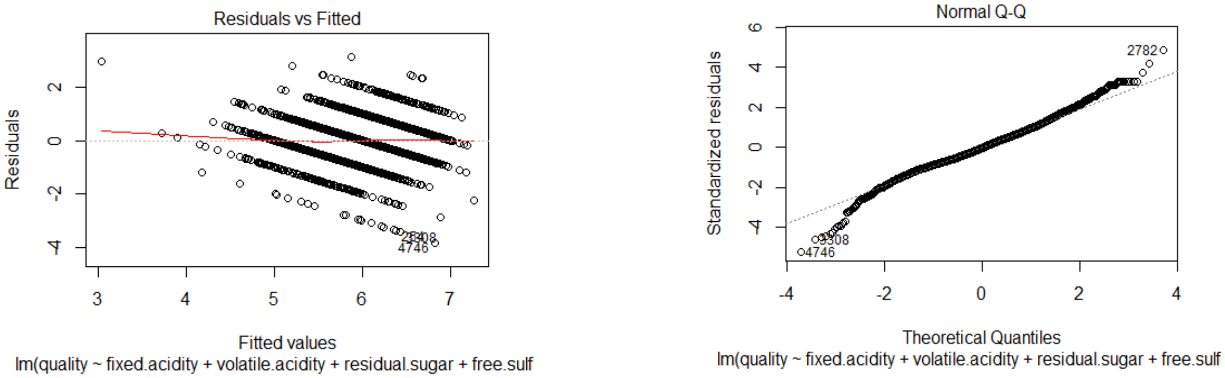


Figure 4

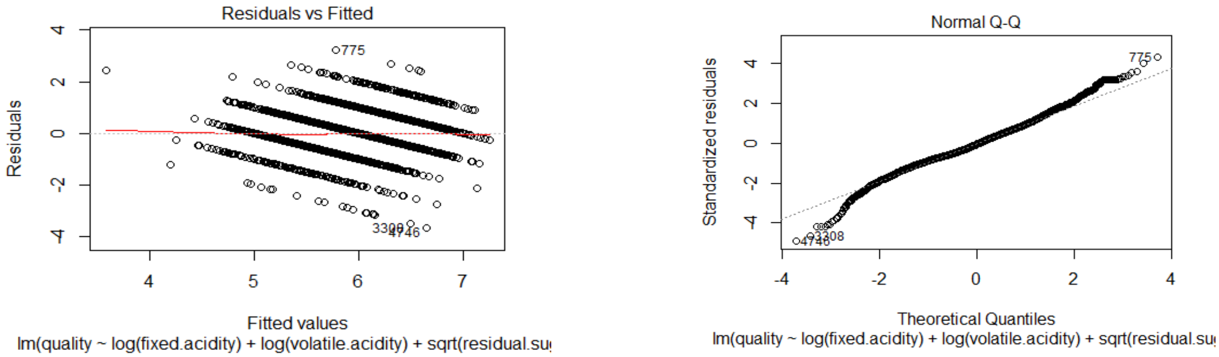


Figure 5

As shown on the Figure 4, the residual plot shows that the red line deviates significantly from the 0 line and the points scatter in a certain pattern, which indicate non-constant variance and non-linearity. The QQ-plot shows that a heavy tails distribution, but most points fall on a straight line. In this case, it might be reasonable to say the residuals follow a normal distribution.

As shown on the Figure 5, the residual plot shows the points scatter more randomly around 0 and the red line is close to 0 line, which indicates that the assumptions of constant variance and linearity are satisfied. The QQ-plot also shows that a heavy tails distribution, but most points still fall on a straight line. After the transformation, the assumptions are satisfied.

4.4 Leverage and Outliers Check

The Figure 6 are the standardized residuals vs. leverage plots. As the plots shown, there are many high leverage points and 4 outliers, 3 of outliers are bad leverage points.

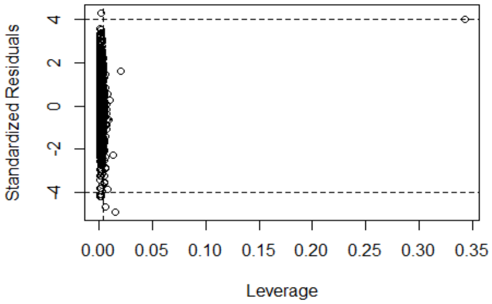
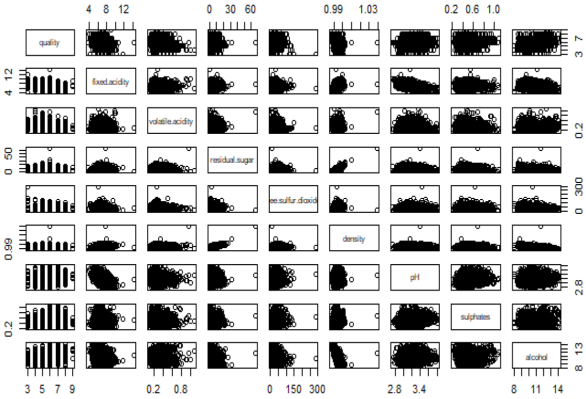


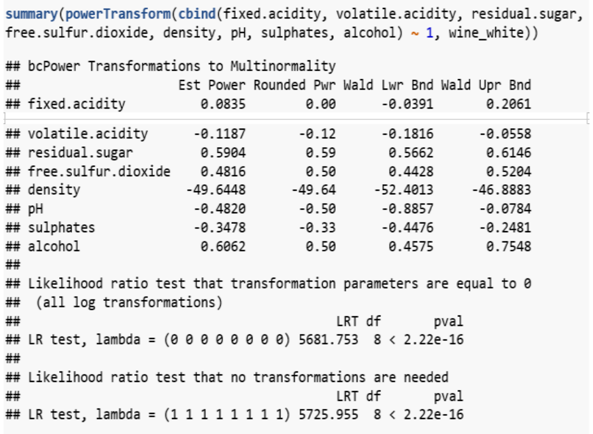
Figure 6

4.5 Box-Cox Transformation

The Figure 7 is the scatter matrices plot of reduced model.

The Table 4 is the result of Box-Cox method to estimate the rounded λ for transformation of predictors.





```
summary(powerTransform(cbind(fixed.acidity, volatile.acidity, residual.sugar,
free.sulfur.dioxide, density, pH, sulphates, alcohol) ~ 1, wine_white))

## bcPower Transformations to Multinormality
##                     Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## fixed.acidity          0.0835        0.00      -0.0391       0.2061
## volatile.acidity      -0.1187       -0.12      -0.1816      -0.0558
## residual.sugar         0.5904        0.59       0.5662       0.6146
## free.sulfur.dioxide    0.4816        0.50       0.4428       0.5204
## density              -49.6448      -49.64     -52.4013     -46.8883
## pH                    -0.4820       -0.50      -0.8857      -0.0784
## sulphates             -0.3478       -0.33      -0.4476      -0.2481
## alcohol                0.6062        0.50       0.4575       0.7548
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                       LRT df       pval
## LR test, lambda = (0 0 0 0 0 0 0 0) 5681.753  8 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                       LRT df       pval
## LR test, lambda = (1 1 1 1 1 1 1 1) 5725.955  8 < 2.22e-16
```

Figure 7

Table 4

As shown on results above, the rounded λ of predictor “density” is -49.64. Since there is usually little justification for making extreme transformations, we do not consider transforming “density”. Thus, the final model can be get through transformation: $quality = \beta_0 + \beta_1 log(fixed.acidity) +$

$\beta_2 log(volatile.acidity) + \beta_3 sqrt(residual.sugar) + \beta_4 sqrt(free.sulfur.dioxide) + \beta_5 density + \beta_6 1/sqrt(pH) + \beta_7 log(sulphates) + \beta_8 sqrt(alcohol) + e.$

The Table 5 is the comparison of the adjusted R^2, and AIC of the reduced model and the final model.

	Adjusted R^2	AIC
Reduced model	0.2805767	11108.29
Final model	0.283624	11086.5

Table 5

As shown on the Table 5, the Final model has a smaller AIC (11086.5) than the reduced model and has a larger adjusted R^2 (0.283624). These means that the transformation is helpful to improve the performance of the model.

4.6 Cross-validation

The Table 6 is the comparison of the RMSE of ordinary least squares model of reduced model (OLS_ori) and final model (OLS_final), backwards stepwise model (OLS_step), ridge model (Ridge), and Lasso model (Lasso).

Model	RMSE
OLS_ori	0.7243308
OLS_final	0.7226316
OLS_step	0.7243308
Ridge	0.7410161
Lasso	0.7298092

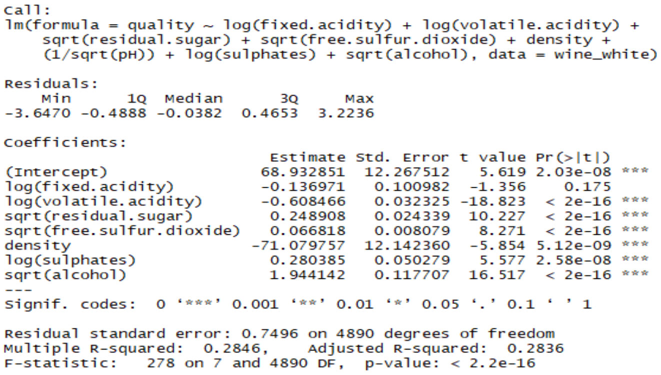
Table 6

As shown on the Table 6, the RMSE of OLS_final is smallest, which indicates that the accuracy of the OLS_final model. that is the best final model.

6. Discussion

6.1 Discussion summary

From the analysis, we figure out not all 11 predictors in the original data set are useful to the model. After log and sqrt transformation for some predictors, the model satisfies assumptions of MLR. From summary (lm_wine2) (Table 7), 7 predictors in the final model are significant. The log (fixed. Acidity) is not significant. We can see that coefficient has relationship with physicochemical factors either negative or positive. From the analysis, we also know not all 11 predictors in the original data set are useful to the



```
Call:
lm(formula = quality ~ log(fixed.acidity) + log(volatile.acidity) +
    sqrt(residual.sugar) + sqrt(free.sulfur.dioxide) + density +
    (1/sqrt(pH)) + log(sulphates) + sqrt(alcohol), data = wine_white)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6470 -0.4888 -0.0382  0.4653  3.2236

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               68.932851  12.267512   5.619 2.03e-08 ***
log(fixed.acidity)        -0.136971   0.100982  -1.356    0.175
log(volatile.acidity)     -0.608466   0.032325 -18.823  < 2e-16 ***
sqrt(residual.sugar)       0.248908   0.024339  10.227  < 2e-16 ***
sqrt(free.sulfur.dioxide)  0.066818   0.008079   8.271  < 2e-16 ***
density                  -71.079757  12.142360  -5.854 5.12e-09 ***
log(sulphates)             0.280385   0.050279   5.577 2.58e-08 ***
sqrt(alcohol)              1.944142   0.117707  16.517  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7496 on 4890 degrees of freedom
Multiple R-squared:  0.2846,    Adjusted R-squared:  0.2836
F-statistic:   278 on 7 and 4890 DF,  p-value: < 2.2e-16
```

Table7

model. In the future, we can use the model to predict white wine quality using these 9 physicochemical factors.

6.2 Discussion weakness

we have three weakness points. Firstly, we do not check multicollinearity. The predictors may correlate with each other. Secondly, since we do box-cox transformation, it become hard to interpret the result. Thirdly, we believe there are more factors have effects on white wine quality like climate, temperature and location, etc. We might include some other crucial factors in our model in the future.

7. Reference

Data resource come from website below:

Daniel S. Panizzo. (2017). Wine Quality: Modeling wine preferences by data mining from physicochemical properties, https://www.kaggle.com/danielpanizzo/red-and-whitewine-quality.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. (2009). Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 01679236.

Method applied from:

Eric Fox. (2019). Linear and Logistic Regression lecture materials and notes, California State University East Bay.

Code Appendix

```
library(tidyverse)
library(ggplot2)
library(glmnet)

##Input data
wine_white <- read.csv(file = "./Data/wineQualityWhites.csv", header =
TRUE)
head(wine_white)
str(wine_white)


##Checking missing values
library(Amelia)
missmap(wine_white, main = "Missing values vs observed")

sapply(wine_white, function(x) sum(is.na(x)))


##Processing data
wine_white <- wine_white[,-1]
wine_white$quality <- as.numeric(wine_white$quality)
glimpse(wine_white)
```

```
##Data Description
###Dimension
dim(wine_white)

###Attributes
names(wine_white)

###Data displays
hist(wine_white$quality)

summary(wine_white)

###Scatter matrices plot of original full model
pairs(quality ~ ., data = wine_white)


##Model selection

###R squared
library(leaps)
regsub_fit <- regsubsets(quality ~ ., data = wine_white, nvmax=11)
regsub_summ <- summary(regsub_fit)

attributes(regsub_summ)

round(regsub_summ$rsq, 4)

round(regsub_summ$adjr2, 4)

which.max(regsub_summ$adjr2)

coef(regsub_fit, 8)


###Plot of R squared, AIC, BIC
n <- nrow(wine_white)
aic_vec <- n*log(regsub_summ$rss/n) + 2*c(1:11)

par(mfrow=c(1,3), mar=c(4.5, 4.5, 1, 1))
####R squared
plot(c(1:11), regsub_summ$adjr2, xlab="Number of Variables",
ylab="Adjusted RSqr")
abline(v=which.max(regsub_summ$adjr2))

####AIC
plot(c(1:11), aic_vec, xlab="Number of Variables", ylab="AIC")
abline(v=which.min(aic_vec))

####BIC
plot(c(1:11), regsub_summ$bic, xlab="Number of Variables", ylab="BIC")
abline(v=which.min(regsub_summ$bic))


###Predictors selected
```

```
which.min(aic_vec)
coef(regsub_fit, 8)

which.min(regsub_summ$bic)
coef(regsub_fit, 8)

###Backwards stepwise selection of AIC
wine_full <- lm(quality ~ ., data = wine_white)
wine_sa <- step(wine_full)
summary(wine_sa)

###Backwards stepwise selection of BIC
wine_sb <- step(wine_full, k=log(n))
summary(wine_sb)

###Reduced model
lm_wine1 <- lm(quality ~ fixed.acidity + volatile.acidity + residual.sugar
+ free.sulfur.dioxide + density + pH + sulphates + alcohol, data =
wine_white)
summary(lm_wine1)

###Scatter matrices plot of reduced model
pairs(quality ~  fixed.acidity + volatile.acidity + residual.sugar +
free.sulfur.dioxide + density + pH + sulphates + alcohol, data =
wine_white)

###Test relationship
wine_null <- lm(quality ~ 1, data = wine_white)
anova(wine_null ,lm_wine1)

###Check assumptions for reduced model
####QQ-plot
plot(lm_wine1, which = 2)

####Residual plot
plot(lm_wine1, which = 1)


##Box-cox Transformation

###Estimate transformations

library(car)
summary(powerTransform(cbind(fixed.acidity, volatile.acidity,
residual.sugar, free.sulfur.dioxide, density, pH, sulphates, alcohol) ~ 1,
wine_white))

###Final model after transformations
lm_wine2 <- lm(quality ~ log(fixed.acidity) + log(volatile.acidity) +
sqrt(residual.sugar) + sqrt(free.sulfur.dioxide) + density + 1/sqrt(pH) +
log(sulphates) + sqrt(alcohol), data = wine_white)
summary(lm_wine2)

###Check and compare AIC and adjusted R squared
```

```
AIC(lm_wine1)

AIC(lm_wine2)

summary(lm_wine1)$adj.r.squared

summary(lm_wine2)$adj.r.squared

###Scatter matrices plot of final model
pairs(quality ~ log(fixed.acidity) + log(volatile.acidity) +
sqrt(residual.sugar) + sqrt(free.sulfur.dioxide) + density + 1/sqrt(pH) +
log(sulphates) + sqrt(alcohol), data = wine_white)

###Check assumptions final model
####QQ-plot
plot(lm_wine2, which = 1)

####Residual plot
plot(lm_wine2, which = 2)

###Compare the plots of the observed versus predicted values
par(mfrow=c(1,2), mar=c(2.5, 2.5, 2, 2))

plot(predict(lm_wine1), wine_white$quality, xlab="Fitted Values",
ylab="quality")
lines(lowess(predict(lm_wine1), wine_white$quality), col='red')
abline(0,1)

plot(predict(lm_wine2), wine_white$quality, xlab="Fitted Values",
ylab="quality")
lines(lowess(predict(lm_wine2), wine_white$quality), col='red')
abline(0,1)

###Identify leverage points and outliers
plot(lm_wine2,which = 5)

p <- 8
n <- nrow(wine_white)
plot(hatvalues(lm_wine2), rstandard(lm_wine2), xlab= 'Leverage' , ylab=
'Standardized Residuals')
abline(h = c(-4,4),v = 2*(p+1)/n, lty=2)


##Cross-validation
###Split data into 70% for train and 30% for test
set.seed(99)

wine <- model.matrix(quality ~ ., data=wine_white)[, -12]
q <- wine_white$quality

train_idx <- sample(n, size = floor(0.7 * n))

wine_train <- wine[train_idx, ]
nrow(wine_train)
```

```
wine_test <- wine[-train_idx, ]
nrow(wine_test)

q_train <- q[train_idx]

q_test <- q[-train_idx]

###Fit model based on cross validation
####Fit reduced model on training set
lm_wine3 <- lm(quality ~ fixed.acidity + volatile.acidity + residual.sugar
+ free.sulfur.dioxide + density + pH + sulphates + alcohol, data =
wine_white, subset = train_idx)

####Fit final model on training data
lm_wine4 <- lm(quality ~ log(fixed.acidity) + log(volatile.acidity) +
sqrt(residual.sugar) + sqrt(free.sulfur.dioxide) + density + 1/sqrt(pH) +
log(sulphates) + sqrt(alcohol), data = wine_white, subset = train_idx)

####Fit ordinary least squares model w/ stepwise selection on training set
lm_step_wine <- step(lm_wine4, trace=F)

####Fit ridge model on training set
ridge_wine <- cv.glmnet(wine_train, q_train, alpha=0)

####Fit lasso model on training set
lasso_wine <- cv.glmnet(wine_train, q_train, alpha=1)

###Compute RMSE
####Create function
compute_rmse <- function(y, y_pred) {
  n <- length(y)
  sqrt((1 / n) * sum((y - y_pred)^2))
}

####Reduced model
wine_pred1 <- predict(lm_wine3, newdata = wine_white[-train_idx, ])
rmse_ori <- compute_rmse(q_test, wine_pred1)

####Final model
wine_pred2 <- predict(lm_wine4, newdata = wine_white[-train_idx, ])
rmse_fin <- compute_rmse(q_test, wine_pred2)

####Step
wine_step_pred <- predict(lm_step_wine, newdata = wine_white[-
train_idx, ])
rmse_step <- compute_rmse(q_test, wine_step_pred)

####Ridge
wine_ridge_pred <- predict(ridge_wine, newx = wine_test, s = "lambda.min")
wine_ridge_pred <- as.numeric(wine_ridge_pred)
rmse_ridge <- compute_rmse(q_test, wine_ridge_pred)

####Lasso
```

```
wine_lasso_pred <- predict(lasso_wine, newx = wine_test, s = "lambda.min")
wine_lasso_pred <- as.numeric(wine_lasso_pred)
rmse_lasso <- compute_rmse(q_test, wine_lasso_pred)

###Compare RMSE
data.frame(Model = c('OLS_ori', 'OLS_final', 'OLS_step', 'Ridge',
'Lasso' ), RMSE = c(rmse_ori, rmse_fin, rmse_step, rmse_ridge,
rmse_lasso))
```