Stat306 Group Project

Group: D7

Group member:

Harbor Zhang #54007349

Xiaotong Tao #24111791

Introduction

Background

Nowadays, the wine industry shows a recent growth spurt as social drinking is on the rise. Among all kinds of wines, red wine is always the most popular choice for people to bring on parties or drink along at night. The production of red wine usually takes many steps, including grape processing, destemming and crushing, cooling, inoculation and fermentation and so on. After a batch of red wine was successfully produced, a series of tests would be implemented on the red wine in terms of its alcoholicity, acidity, sugar, PH, density…etc. Recording and analyzing these data would not only help the producers know whether the quality of the red wine is good or not, but it will also reveals which elements in the red wine production significantly contribute to the red wine quality and therefore can helps improve manufacturing technique in the future.

Objectives

In this report, we are interested in how others factors during the production affect the alcohol in the red wine. We want to know whether there is a linear relationship between alcohol and other variables. If so, we try to figure out a best model for this relationship.

Methodology

Data Source and Acknowledgement

The dataset that we will use in our project contains the data of some variables in the red-wine (specifically, Portuguese "Vinho Verde" wine) making process. The dataset is named “Red Wine Quality” and we downloaded it directly from Kaggle(https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009). This dataset is also available from the UCI Machine Learning repository. (Cortez, Cerdeira, Almeida, Matos and J. Reis, 2009)

Data Overview

The dataset consists of 1599 observations, with one response variable alcohol, and 10 explanatory variables, which are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, and sulphates respectively.

Method

We will start with visualizations of the dataset by using suitable plots, a summary of key features will be provided. Then we will fit a linear regression model to the data with all factors included. An analysis of this first model will then be given and followingly, we will try to improve our model by deleting some variables if they are not significant or adding interaction terms to find out whether interactions between variables exist. Residual plots and QQ plots will be used to see how the models fitted perform. We keep the significance level at 5% in the whole report.

Analysis

We first read the data into R by using the read.csv() function, then we use the summary() function to provide an overall look of the dataset.

```
> summary(wine)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037
       pH          sulphates         alcohol
 Min.   :2.740   Min.   :0.3300   Min.   : 8.40
 1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
 Median :3.310   Median :0.6200   Median :10.20
 Mean   :3.311   Mean   :0.6581   Mean   :10.42
 3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
 Max.   :4.010   Max.   :2.0000   Max.   :14.90
```

We see that in the dataset, we have two variables about acidity of the red wine, namely fixed acidity and volatile acidity. We are curious about if correlation exists between these two types of acidity, so we use cor() function in R to make the calculation. The value of the correlation between these two variables is -0.256, which indicates that there exists a relatively weak negative correlation between them.

Next, we would like to fit a linear regression model to the data using all ten explanatory variables in order to find out the relationship between these variables and our response variable, alcohol. We use the lm() function in R to generate the linear regression model and we name this first model reg1.

```
> summary(reg1)

Call:
lm(formula = alcohol ~ ., data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.07175 -0.39267 -0.04056  0.35396  2.44365

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.072e+02  1.308e+01  46.419  < 2e-16 ***
fixed.acidity         5.324e-01  2.064e-02  25.796  < 2e-16 ***
volatile.acidity      3.608e-01  1.144e-01   3.154 0.001638 **
citric.acid           8.306e-01  1.379e-01   6.024 2.11e-09 ***
residual.sugar        2.844e-01  1.229e-02  23.135  < 2e-16 ***
chlorides            -1.462e+00  3.956e-01  -3.696 0.000227 ***
free.sulfur.dioxide  -2.143e-03  2.057e-03  -1.042 0.297517
total.sulfur.dioxide -2.296e-03  6.881e-04  -3.336 0.000868 ***
density              -6.174e+02  1.342e+01 -45.998  < 2e-16 ***
pH                    3.762e+00  1.551e-01  24.263  < 2e-16 ***
sulphates             1.247e+00  1.037e-01  12.020  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.614 on 1588 degrees of freedom
Multiple R-squared:  0.6701,    Adjusted R-squared:  0.668
F-statistic: 322.5 on 10 and 1588 DF,  p-value: < 2.2e-16
```

We then provide the summary of the model reg1. From the above picture of R output, we observe that the p-value of the explanatory variable named free sulfur dioxide (0.298) is larger than 0.05, which indicates that it is not significant in the model. As for other variables, we see that their p-values are all smaller than 0.05, thus, we decide to omit the variable free sulfur dioxide and fit a new model without it. We name our second model reg2 and we provide the summary of reg2 below.

```
> summary(reg2)

Call:
lm(formula = alcohol ~ . - free.sulfur.dioxide, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.06145 -0.39706 -0.03917  0.34928  2.44848

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.059e+02  1.302e+01  46.535  < 2e-16 ***
fixed.acidity         5.300e-01  2.051e-02  25.846  < 2e-16 ***
volatile.acidity      3.809e-01  1.128e-01   3.377 0.000749 ***
citric.acid           8.548e-01  1.359e-01   6.289 4.12e-10 ***
residual.sugar        2.827e-01  1.219e-02  23.198  < 2e-16 ***
chlorides            -1.487e+00  3.949e-01  -3.766 0.000172 ***
total.sulfur.dioxide -2.775e-03  5.123e-04  -5.416 7.02e-08 ***
density              -6.160e+02  1.335e+01 -46.125  < 2e-16 ***
pH                    3.739e+00  1.534e-01  24.369  < 2e-16 ***
sulphates             1.242e+00  1.036e-01  11.984  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.614 on 1589 degrees of freedom
Multiple R-squared:  0.6699,    Adjusted R-squared:  0.668
F-statistic: 358.2 on 9 and 1589 DF,  p-value: < 2.2e-16
```

We observe from above that the p-values of all the explanatory variables are less than 0.05, this means that all the variables are significant in this new model fitted. However surprisingly, we see that the value of adjusted R-squared does not increase compared to the first model after we delete the free sulfur dioxide. This means that the effect of free sulfur dioxide in the model is near zero and ignoring it would not improve our model. However, since deleting this variable makes our model simpler, we can still conclude that reg2 is better than reg1. Thus, in the following context, we would use this model reg2 to replace model reg1.

Then, we would like to explore if interaction exists between fixed acidity and volatile acidity. We

think that a change in one type of acidity can affect the coefficient of the other type of acidity towards to alcohol. Thus, we add an interaction term between these two variables based on model reg2, we call this new model reg3. Here is the summary of model reg3.

```
> summary(reg3)

Call:
lm(formula = alcohol ~ . - free.sulfur.dioxide + fixed.acidity * 
    volatile.acidity, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.16529 -0.39147 -0.03431  0.36055  2.58475 

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 6.251e+02  1.294e+01  48.314  < 2e-16 ***
fixed.acidity               3.310e-01  3.086e-02  10.724  < 2e-16 ***
volatile.acidity           -3.565e+00  4.780e-01  -7.458 1.44e-13 ***
citric.acid                 7.708e-01  1.334e-01   5.780 8.99e-09 ***
residual.sugar              2.874e-01  1.194e-02  24.076  < 2e-16 ***
chlorides                  -1.308e+00  3.869e-01  -3.379 0.000744 ***
total.sulfur.dioxide       -3.117e-03  5.028e-04  -6.199 7.21e-10 ***
density                    -6.346e+02  1.325e+01 -47.898  < 2e-16 ***
pH                          4.017e+00  1.536e-01  26.144  < 2e-16 ***
sulphates                   1.289e+00  1.015e-01  12.694  < 2e-16 ***
fixed.acidity:volatile.acidity  4.813e-01  5.673e-02   8.484  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6008 on 1588 degrees of freedom
Multiple R-squared:  0.6842,    Adjusted R-squared:  0.6822 
F-statistic:   344 on 10 and 1588 DF,  p-value: < 2.2e-16
```

We observe from the above figure that the p-value of the interaction term is less than 0.05, this means that it is significant in the model. Also, we see that the adjusted R-squared (0.682) is larger than 0.668, which is the adjusted R-squared of model reg2. Thus, we can conclude than after adding the interaction term between these two types of acidity, the model has been improved.

We will then apply the best subset selection algorithms to the full model with all ten variables included to see what is the best model selected by R. The R command we use here is regsubsets() and the method being used is exhaustive.

```
> s <- regsubsets(alcohol~., data = wine, method="exhaustive")
> ss <- summary(s)
> ss$which
  (Intercept) fixed.acidity volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide total.sulfur.dioxide density    pH sulphates
1        TRUE          FALSE            FALSE       FALSE          FALSE     FALSE               FALSE                FALSE    TRUE FALSE      FALSE
2        TRUE           TRUE            FALSE       FALSE          FALSE     FALSE               FALSE                FALSE    TRUE FALSE      FALSE
3        TRUE           TRUE            FALSE       FALSE          FALSE     FALSE               FALSE                FALSE    TRUE  TRUE      FALSE
4        TRUE           TRUE            FALSE       FALSE           TRUE     FALSE               FALSE                FALSE    TRUE  TRUE      FALSE
5        TRUE           TRUE            FALSE       FALSE           TRUE     FALSE               FALSE                FALSE    TRUE  TRUE       TRUE
6        TRUE           TRUE            FALSE        TRUE           TRUE     FALSE               FALSE                FALSE    TRUE  TRUE       TRUE
7        TRUE           TRUE            FALSE        TRUE           TRUE     FALSE               FALSE                 TRUE    TRUE  TRUE       TRUE
8        TRUE           TRUE            FALSE        TRUE           TRUE      TRUE               FALSE                 TRUE    TRUE  TRUE       TRUE
> ss$adjr2
[1] 0.2457224 0.3768794 0.5069384 0.6221989 0.6559881 0.6596492 0.6641424 0.6658248
> ss$cp
[1] 2033.42970 1402.62571  777.88738  224.98675   63.71401   47.12108   26.56273   19.49344
> ss$rsq
[1] 0.2461944 0.3776593 0.5078641 0.6231446 0.6570645 0.6609271 0.6656136 0.6674978
```

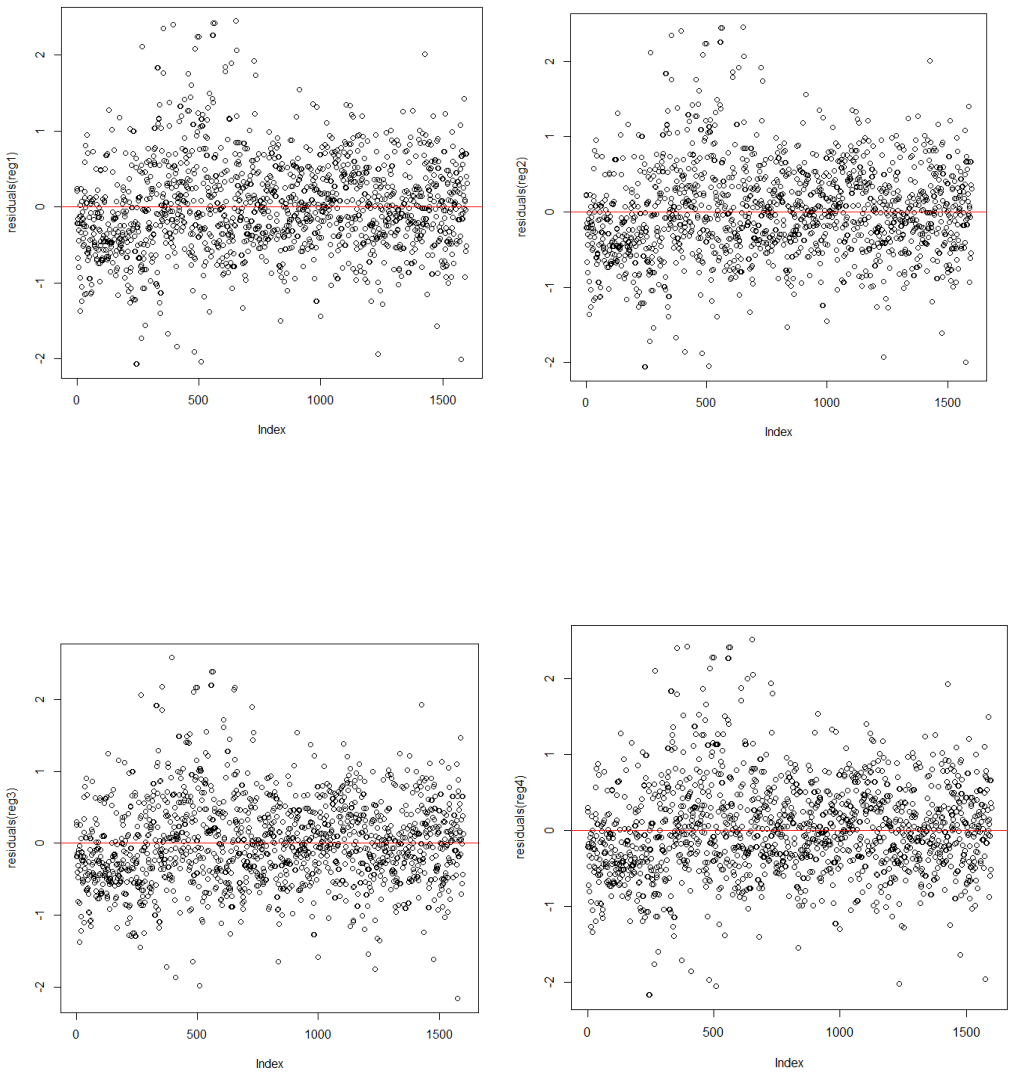
We see that the best model includes eight variables, free sulfur dioxide and volatile acidity are not being selected. The adjusted R-squared is 0.666, which is the largest among all. The values of Cp are all much greater than k+1, this means that the best model is still biased. We omit these two variables and fit this best model selected by R, call it reg4.

```
> summary(reg4)

Call:
lm(formula = alcohol ~ . - free.sulfur.dioxide - volatile.acidity,
    data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.17462 -0.39643 -0.04228  0.34482  2.51179

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.004e+02  1.296e+01  46.325  < 2e-16 ***
fixed.acidity         5.364e-01  2.048e-02  26.190  < 2e-16 ***
citric.acid           6.172e-01  1.167e-01   5.290 1.40e-07 ***
residual.sugar        2.826e-01  1.223e-02  23.109  < 2e-16 ***
chlorides            -1.151e+00  3.833e-01  -3.002  0.00273 **
total.sulfur.dioxide -2.547e-03  5.095e-04  -4.998 6.42e-07 ***
density              -6.104e+02  1.329e+01 -45.913  < 2e-16 ***
pH                    3.768e+00  1.537e-01  24.519  < 2e-16 ***
sulphates             1.169e+00  1.017e-01  11.497  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.616 on 1590 degrees of freedom
Multiple R-squared:  0.6675,    Adjusted R-squared:  0.6658
F-statistic:   399 on 8 and 1590 DF,  p-value: < 2.2e-16
```

Compared the adjusted R-squared of all four models fitted, we conclude that model reg3 is the best among all. Besides, by using the AIC () function in R, we also see that model reg3 perform better than others.

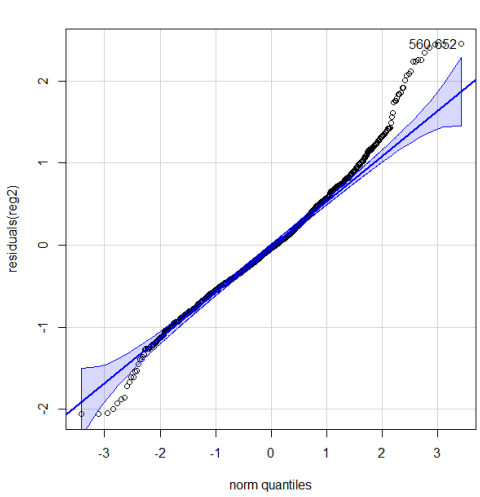
```
> AIC(reg1)
[1] 2990.959
> AIC(reg2)
[1] 2990.052
> AIC(reg3)
[1] 2921.167
> AIC(reg4)
[1] 2999.49
```

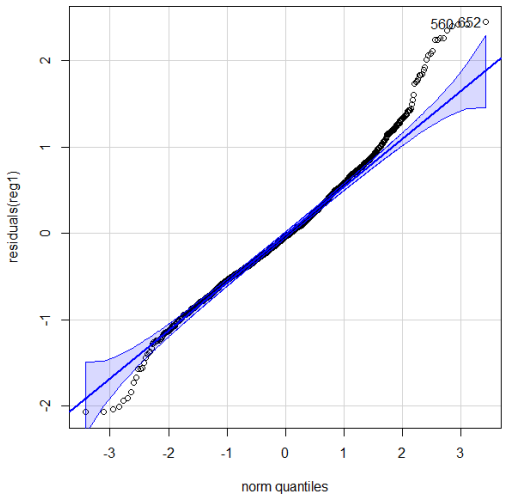
We now use the residual plots and QQ plots of residuals to analyze the residuals of the four models fitted.

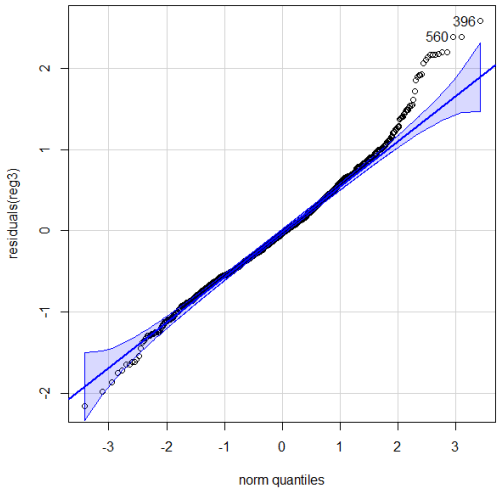


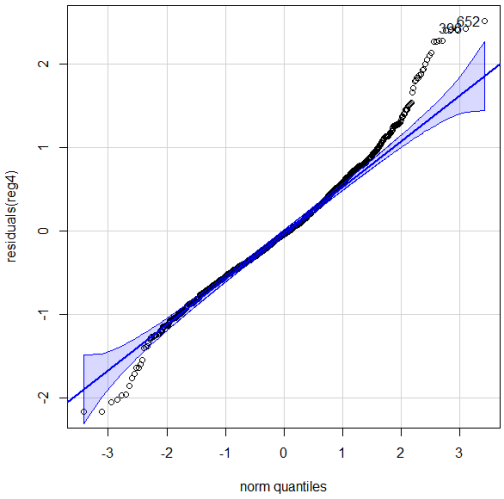
From the residual plots, we observe that the residuals distribute randomly and evenly around zero and there is no clear pattern from all four plots. This indicates that all four models perform well, however, the residuals of reg3 seems to be more dense near zero, which means it is slightly better than others.

We use R command qqPlot() to generate the QQplots of residuals for these four models.









From the four QQ plots above, we observe that most of the data are normally distributed, but a little bit heavy-tailed. Also, we see that model reg3 capture the lower quantile better than others, which indicates it performs better.

Conclusion

Being curious about the relationship of alcohol in red wine and ten different factors which might shade an influence on it, our core objective in this research is to figure out a best-fit model for the alcohol and different known influencing factors.

We firstly did a simple linear regression of the ten different influencing factors and alcohol data, called reg1. By reading the summary of reg1, we deleted one of the ten factors, which is free sulfur dioxide and has little significance. Thus, a new model reg2 was built, and is proved to be better than reg1 by comparing their adjusted R squared values.

Followingly, we guess that there might exist interaction between the fixed acidity and the volatile acidity. From a common sense of point of view, the existence of a correlation between the two factors is something that can be easily associated. In order to verify this conjecture, we add the interaction when doing linear regression, which made a new model reg3. As the adjusted R squared value is obviously larger, reg3 is a better fitted model.

Based on reg3, we apply the best subset selection algorithms, and delete volatile acidity and free sulfur dioxide from the model. Doing linear regression for this new reg4, we noticed that the adjusted R squared value turned smaller, which indicated that the model did not get improved. Therefore, reg3 is still the best model till now.

Now for further validation that reg3 is a best-fitted model, we successively apply methods like Akaike's Information Criteria (AIC), drawing residual plots and QQ plots, all of which present the same conclusion as before, that reg3 is the best fitted model.

To conclude, the model of alcohol and ten influencing factors except free sulfur dioxide with consideration of the interaction of fixed acidity and volatile acidity turned out to be the best-fitted one.

This model could well interpret the relationship of alcohol and ten influencing factors in the basic dataset.

However, our research still has some shortcomings and areas for improvement. In the process of doing AIC, we noticed that the cp values showed much larger than expected, which indicated that there might exist more factors which affect the alcohol data of red wine. In the future, we would focus on this, and try expanding our dataset for a more rigorous result.

Reference

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.