The Influence of Various Factors in Red Wine on Alcohol Concentration

Project group: 32

Members :

Muhua Li B00753847

Yuhao Chen B0084929

Wenke Wang B00849234

Abstract

Red wine is an indispensable item in daily life. Wine has a long history, Studying the ingredients in red wine is what researchers have been doing. The aim of this work was to investigate the factors that influence the alcohol content of red wine. Alcohol content is an important biochemical indicator in red wine. The work could help wine producers control the alcohol content of their wines precisely so that they can produce wines suitable for consumers. This article uses data from Cortez's experiment in 2009. The original data has 11 independent variables and one variable. We take the alcohol content as the dependent variable and the other variables as the independent variables. In this article, we use R to analyze this dataset. In the end, we find the closest model for influencing the alcohol content of red wine.

Introduction

Red wine is an essential drink in People's Daily life. With the development of modern technology, wine production technology is becoming more and more perfect. The most important factor in red wine as an alcoholic drink is its alcohol content. The difference in alcohol content can affect consumers' purchase intention. For example, some people prefer red wine with low alcohol accuracy, while others prefer red wine with high alcohol accuracy. Therefore, for red wine manufacturers, accurate regulation of alcohol degree is one of the ways to expand sales and cater to consumers. In this article, our theme is to explore the factors that influence the alcohol content of red wine. We expected to find a perfect model to describe the relationship between alcohol content and other factors. This research has practical implications. This model could provide guidance for wine manufacturers to more accurately regulate the alcohol content of their wines. As Masterclass said (2020), The alcohol content can affect the flavor of red wine. For wine producers, this model is of great interest. Many researchers have done similar studies. For example, Rodrigues and other researchers investigate environmental factors influencing the efficacy of different yeast strains for alcohol level reduction in wine by respiration (2015). This paper uses data from experiments conducted by Cortez and other researchers (2009). The difference between our study and previous studies is that our study is more comprehensive. Our experiment explored the relationship between many factors and the alcohol content of red wine, not just one factor. This paper uses data from experiments conducted by Cortez and other researchers (2009). Next we will describe the data and build the model.

Data Visualization

Before we do the analysis, we first introduce a new set of data. This set of data accords with the law of original data set, which means this data set is normal. The following of details of this new data set: fixed acidity is 7.1; volatile eacidity is 0.54; citric acid is 0.23; residual sugar is 2.1; chlorides is 0.089; free sulfur dioxide is 17; total sulfur dioxide is 53; density is 0.9977; Ph is 3.44; sulphates is 0.61; alcohol is 9.4; quality is 5. Now let us do the data description.

Due to the relationship between the quality and the rest variables, we transfer the quality to become the vector form and analyze the relationship the quality and alcohol.

Through the data, we visualize the data and analyze the influence of the alcohol and various variables through the visual graph. According to the Figure.1, Figure.3 and Figure.4, we can see the approximate relationship between variables and variables. In this project, we focus on the relationship among the alcohol and the other variables. We can see that the absolute values of correlation of alcohol are not so big. Meanwhile, the citric. acid, residual. sugar, Ph, and sulphates have the positive relationship with the alcohol. Furthermore, the correlation value of density and alcohol is the maximation that is -0.49 so the density is negatively correlated with the alcohol. The Figure.2 shows the comparation of the values between alcohol and other variables. The high correlation means the two variables have a linear relationship and the sign of the correlation indicates the direction of the line. For the rest variables, the fixed. acidity has a negative linear relationship with the pH value. The volatile. acidity and citric.acid have the negative relationship. The citric. acid is highly correlated with the fixed.acidity in the positive direction. The relationship between residuals. sugar and density are a positive correlation. For the chlorides, it is highly correlated with sulphates. The free. sulfur.dioxide is the most relevant with the total.sulfur.dioxide. The density is also high correlated with thefixed. acidity.

For the quality, in this dataset, the quality can be divided into 6 levels: 3,4,5,6,7,8. In the Figure. 5, the red wine with quality is 5 has the lowest alcohol and the highest alcohol is the quality 8 in average. We can see from the boxplot, the alcohol and the quality almost have a positive relationship it means that the red wine with the high quality is more possible to have the high alcohol concentration.

Method

First, the first step after completing the Data Description is to create a new model. We build an initial model. The dependent variable is still using alcohol and combined with the data to build an initial model. We can get BP = 177.96, df = 15, p-value <2.2e-16. When the number of independent variables is larger, then check this model, and we find that this model cannot meet the constant variance for its p-value is less than 0.05. Through the box-cox transformation method, we can get a number -1.434343, which represents the power exponent, and we can regard it as power as -1.434343 Use the Stepwise selection method to delete variables from the model, build the first model and find the variables that do not meet the criteria and delete them. According to the calculated data, we can find that the P-value of free Sulphur dioxide is greater than 0.05. According to the results, we need to delete the set of variables free. Sulphur.dioxide. Then, we use the best subset regression to confirm the best variable. The number of variables obtained by combining the data is up to 11. We need to view the data we have filtered through some visual data graphs.

The first picture of Figure.6 is the relationship between the number of variables and R-square. We can see that there is a downward trend between variables. When the variable gets bigger and bigger, R-square gets smaller and smaller. The more features, the less R-square.

The second picture of Figure.6 shows the number of variables and Adjusted R-square. Through the figure, we can see that when the variable larger than 10, it gets better and better. When the number of features is 11, Adjusted R2 is maximum.

The third picture of Figure.6 shows the relationship between the number of variables and CP, and the data shows a downward trend. When the number of variables is larger, the CP value is smaller, the better. When the number of features is 11, Cp is the minimum.

The fourth picture of Figure.6 shows the relationship between the number of variables and AIC. When the number of variables is larger, the AIC is smaller. When the number of features is 11, AIC is the minimum.

Using regfit.full1 combined with the results of the data, According to the result, chlorides and free.sulfur.dioxide is deleted.

Finally, establish the second model and verify it. According to the results, we can find that all the results meet the requirements. So, delete chlorides and free. Sulphur. dioxide as the second model. When we get the model, we need to verify the model. Use the Anova test to verify which model one or model two is right. According to the result, fit2(Best subset regression) is better. So, the final model is the second model.

For the Figure.8, firstly, observe the chart in Residuals VS Fitted, we can find that each error is distributed on the red line, so it meets the assumption. In Normal Q-Q, the data conforms to a normal distribution. In Scale-Location, it is an equal variance. In Residuals VS Leverage, the error is distributed on the red line. Finally, the VIF is less than 10, so Variable independent. Pass the above verification. The second model is the final model.

Result

The final model is

$$Alcohol = \ -2.382\ - 0.002234 fixed.acidity\ - 0.002401 volatile.acidity$$

(6.181e-02) (9.033e-05) (5.083e-04)

$$- 0.003214 citric.acid - 0.001117 residual.sugar + 9.896e - 06 total.sulfur.dioxide$$

(5.917e-04) (5.610e-05) (2.329e-06)

$$+ 2.508 density\ - 0.01644 Ph - 0.00371 sulphates - 0.001181 quality4$$

(6.329e-02) (6.649e-04) (4.519e-04) (9.525e-04)

$$- 0.001605 quality5 - 0.002941 quality6 - 0.003908 quality7 - 0.004795 quality8$$

(8.904e-04) (8.941e-04) (9.185e-04) (1.109e-03)

Adjusted R-squared = 0.6694

It totally contains 13 variables that includes the different level of the quality. In this model, the R-squared is 0.6721 and adjusted R-squared is 0.6694 that means this model can explain the 66.94% of the changes of alcohol after adjustment. Meanwhile, most variables are strongly significant in this model that their p-values are so small. For those quality variables, only the quality4 and quality5 are insignificant but the higher quality has the stronger significance, so it also means the alcohol and quality are influenced by each of themselves.

What’s more, the F-statistic of the model is 11.554 and the p-value is 0.000693 that is a very small p-value, so this model is a significant model.

For the coefficients of the model, those values are not very large so if we change the values of variables, they can influence the alcohol, but the change of the alcohol is not so large. In the variables, only the total.sulfur.dioxide and density have the positive sign so it means that if we increase the values of these two variables, the value of alcohol will also increase. Furthermore, the maximum coefficient is density (2.508) that means the change of density can influence the most change of the alcohol. However, other variables have the opposite direction that means the alcohol will decrease with the increase of the other variables. The Ph have the most negative coefficient (-0.01644) so the increase of the Ph will lead to the largest decrease of alcohol. In addition, the four quality variables also have the negative coefficients so that means the alcohol and quality are negative relationship.

Conclusion

In this project, we discuss the which factors can influence the alcohol in the red wine. We can see from the final model, there are nine variables can influence the alcohol and most of them are negatively correlated with the alcohol. Meanwhile, for the red wine, the alcohol is not so high, so the coefficients of the explanatory variables are not so big. According to Varela, C et al. (2015), in the winemaking process, the acidity, Ph, density and other factors will influence the alcohol concentration in the wine. Meanwhile, King, E., Dunn, L., & Heymann (2013), the alcohol will change the palate of a wine so the quality of a wine needs to be measured carefully and the outcome can be influenced easily. So, we are difficult to judge the relationship between alcohol and quality but in this project, they have the negative relationship.

In conclusion, the alcohol of a wine is influenced by many variables, so we need to maintain the inputs to control the alcohol. Only in this way can we let the alcohol have a proper concentration.

References

Rodrigues, A. J., Raimbourg, T., Gonzalez, R., & Morales, P. (2016). Environmental factors influencing the efficacy of different yeast strains for alcohol level reduction in wine by respiration. *LWT-Food Science and Technology*, *65*, 1038-1043.

MasterClass. (2020, November 08). Learn About Alcohol Content in Wine: Highest Lowest ABV Wines-2020. Retrieved December 11, 2020, from https://www.masterclass.com/articles/learn-about-alcohol-content-in-wine-highest-to-lowest-abv-wines

King, E. S., Dunn, R. L., & Heymann, H. (2013). The influence of alcohol on the sensory perception of red wines. Food Quality and Preference, 28(1), 235-243

Varela, C., Dry, P. R., Kutyna, D. R., Francis, I. L., Henschke, P. A., Curtin, C. D., & Chambers, P. J. (2015). Strategies for reducing alcohol concentration in wine. Australian Journal of Grape and Wine Research, 21, 670-679.

Appendix

Figure.1 Scatter Plot

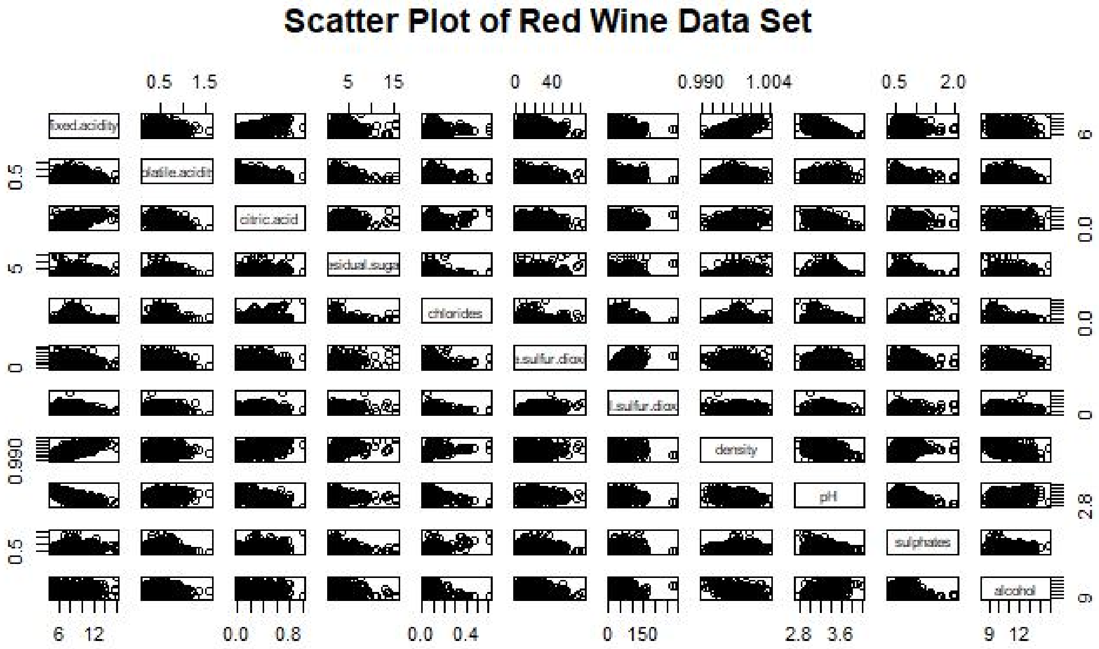
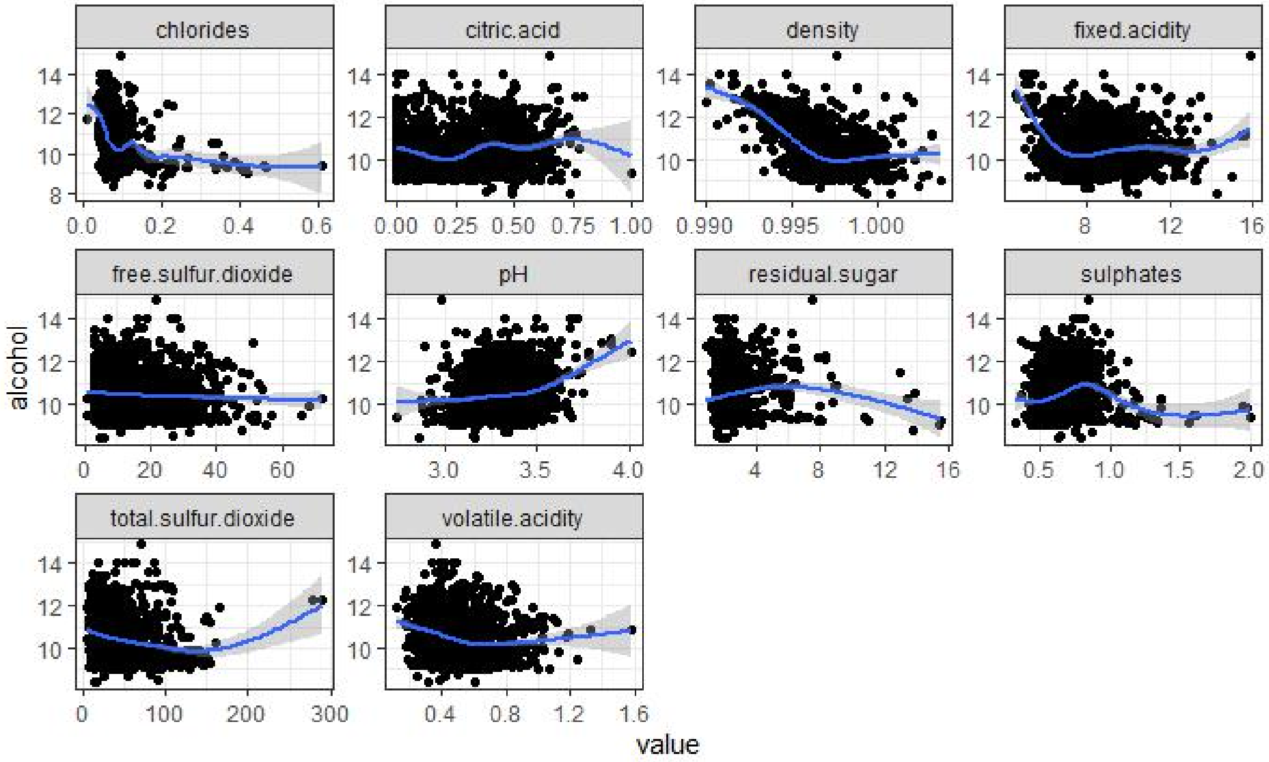


Figure.2 The relationship of alcohol and other variables



Fogure.3 The correlation table

```
##                         fixed. acidity volatile. acidity citric. acid residual. sugar
## fixed. acidity          1.00000000    -0.256121026 0.67168324    0.114891960
## volatile. acidity      -0.25612103     1.000000000 -0.55249619    0.001904582
## citric. acid            0.67168324     -0.552496189 1.00000000   0.143611766
## residual. sugar         0.11489196      0.001904582 0.14361177   1.000000000
## chlorides              0.09367651      0.061299049 0.20381574   0.055601491
## free. sulfur. dioxide  -0.15381716    -0.010499198 -0.06099121    0.187021719
## total. sulfur. dioxide -0.11324966      0.076477398 0.03550597   0.202980581
## density                0.66767030      0.022045928 0.36484657   0.355145966
## pH                    -0.68309024     0.234921276 -0.54188839   -0.085793618
## sulphates              0.18309739     -0.260991806 0.31279519   0.005582099
## alcohol               -0.06122070     -0.202270294 0.10999631   0.042248827
##                         chlorides free. sulfur. dioxidetotal. sulfur. dioxide
## fixed. acidity         0.093676509        -0.153817156          -0.11324966
## volatile. acidity      0.061299049        -0.010499198           0.07647740
## citric. acid           0.203815740        -0.060991214           0.03550597
## residual. sugar        0.055601491         0.187021719           0.20298058
## chlorides             1.000000000         0.005564317          0.04740391
## free. sulfur. dioxide   0.005564317         1.000000000           0.66766916
## total. sulfur. dioxide 0.047403914         0.667669160           1.00000000
## density               0.200626547        -0.021910027          0.07132564
## pH                  -0.264951298        0.070418055          -0.06637560
## sulphates             0.371245204         0.051636975          0.04290995
## alcohol             -0.221096332       -0.069452664          -0.20571135
##                         density          pH   sulphates     alcohol
## fixed. acidity         0.66767030 -0.68309024 0.183097391 -0.06122070
## volatile. acidity      0.02204593 0.23492128 -0.260991806 -0.20227029
## citric. acid           0.36484657 -0.54188839 0.312795191 0.10999631
## residual. sugar        0.35514597 -0.08579362 0.005582099 0.04224883
## chlorides             0.20062655 -0.26495130 0.371245204 -0.22109633
## free. sulfur. dioxide -0.02191003 0.07041805 0.051636975 -0.06945266
## total. sulfur. dioxide 0.07132564 -0.06637560 0.042909949 -0.20571135
## density               1.00000000 -0.34133400 0.148401098 -0.49630042
## pH                   -0.34133400 1.00000000-0.196748067 0.20502756
## sulphates             0.14840110 -0.19674807 1.000000000 0.09373591
## alcohol              -0.49630042 0.20502756 0.093735907 1.00000000
```

Figure.4 The ggplot of the correlation

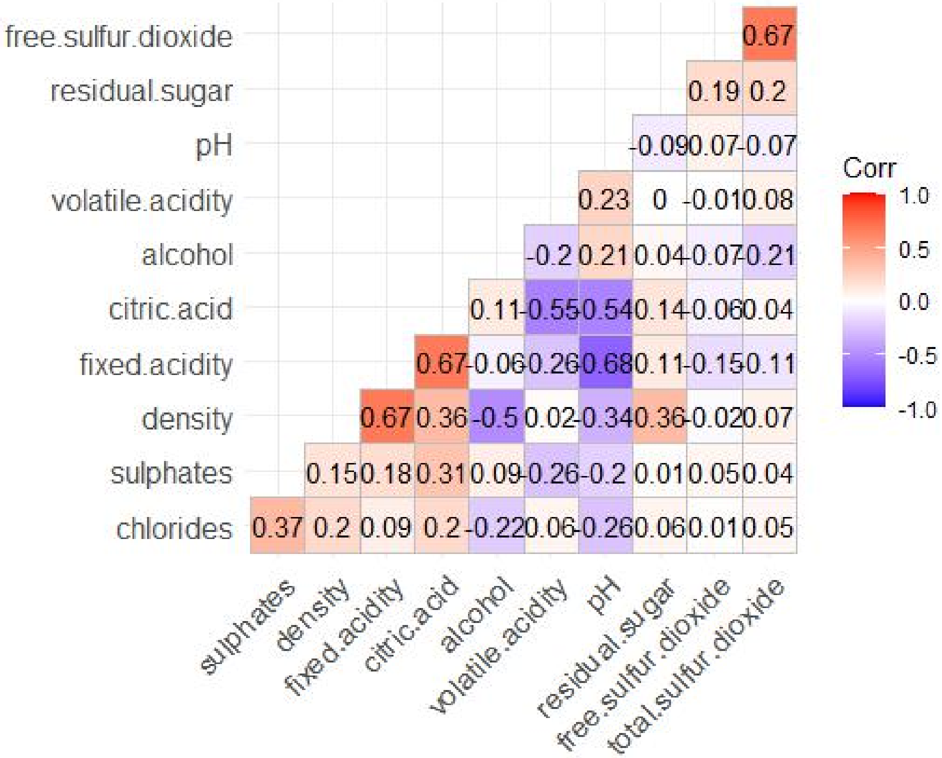


Figure. 5 The boxplot of the alcohol and quality

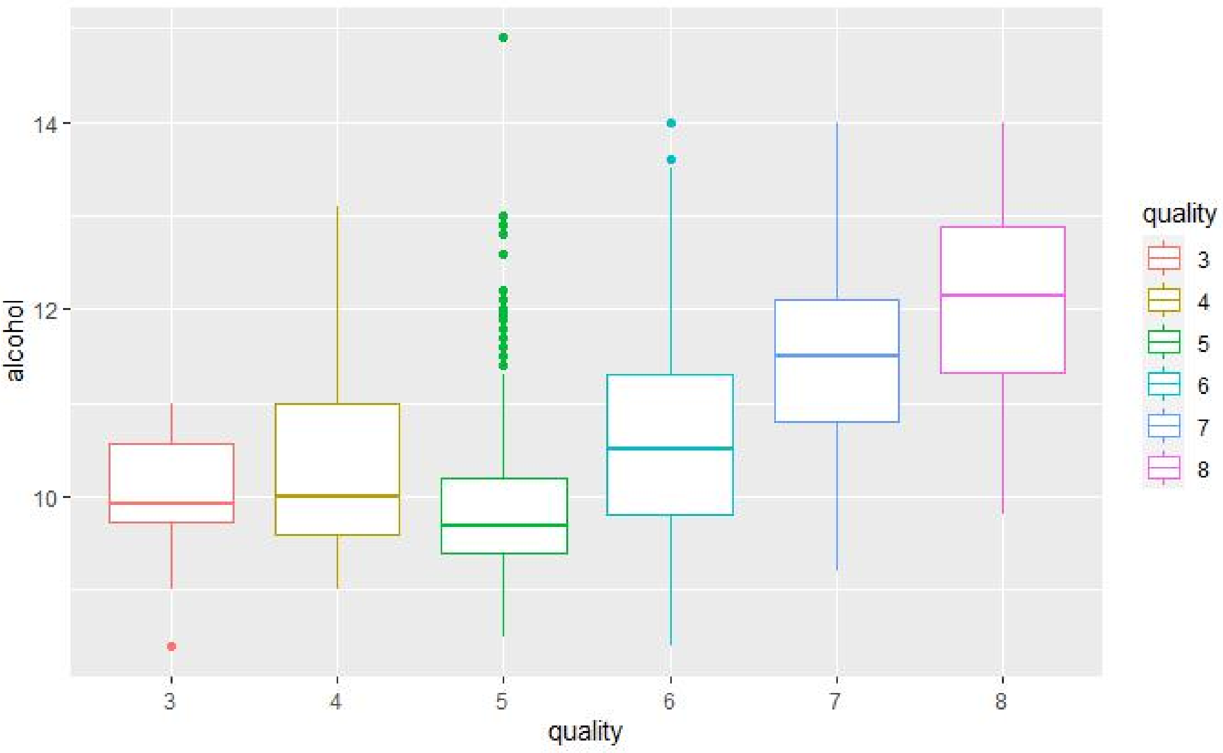
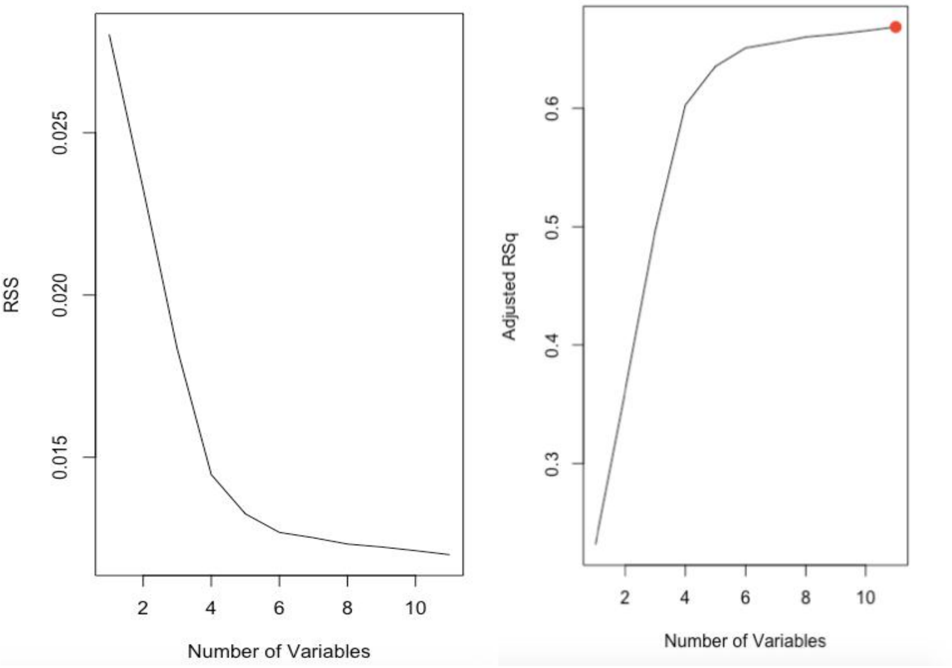


Figure.6 the citations of the model



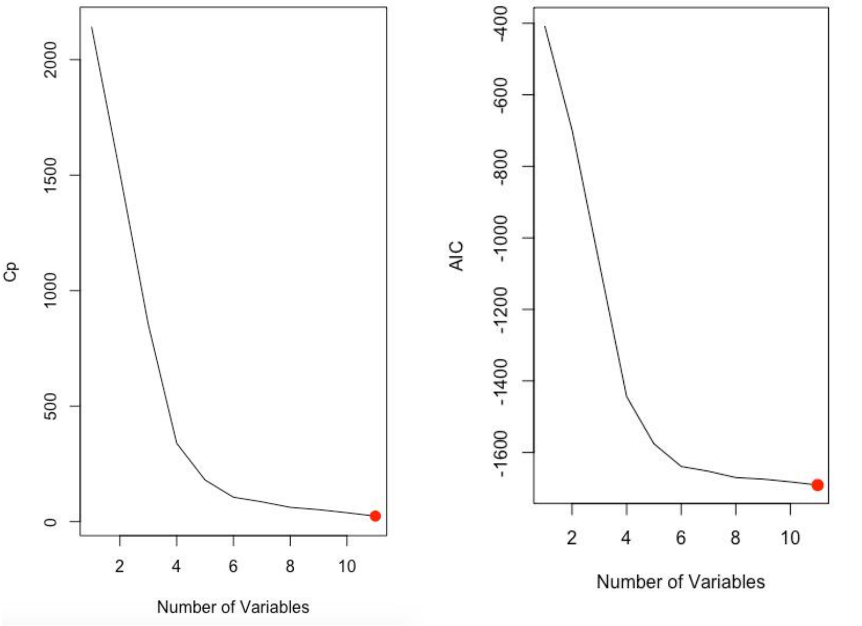
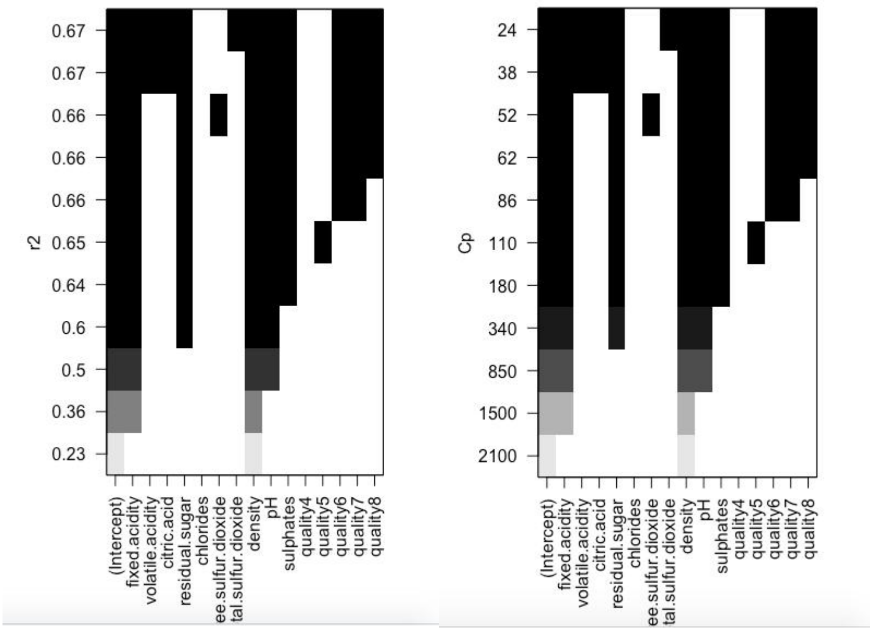

Cp
2000
1500
1000
500
0
2
4
6
8
10
Number of Variables
AIC
-400
-600
-800
-1000
-1200
-1400
-1600
2
4
6
8
10
Number of Variables

Figure.7


r2
0.67
0.67
0.66
0.66
0.66
0.65
0.64
0.6
0.5
0.36
0.23
(Intercept)
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
ee.sulfur.dioxide
tal.sulfur.dioxide
density
pH
sulphates
quality4
quality5
quality6
quality7
quality8
Cp
24
38
52
62
86
110
180
340
850
1500
2100
(Intercept)
fixed.acidity
volatile.acidity
citric.acid
residual.sugar
chlorides
ee.sulfur.dioxide
tal.sulfur.dioxide
density
pH
sulphates
quality4
quality5
quality6
quality7
quality8

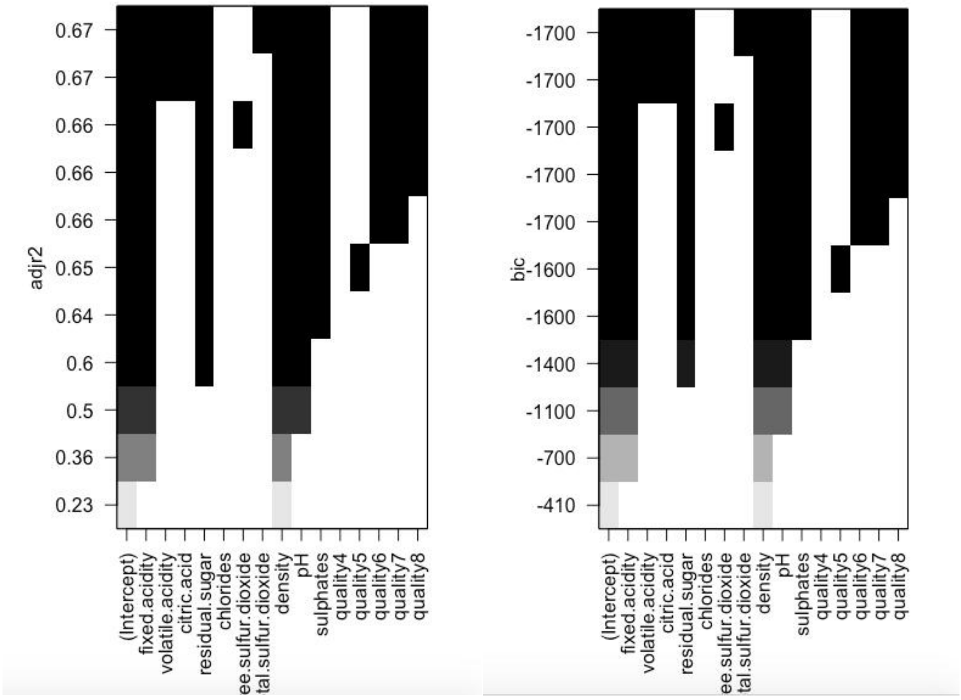
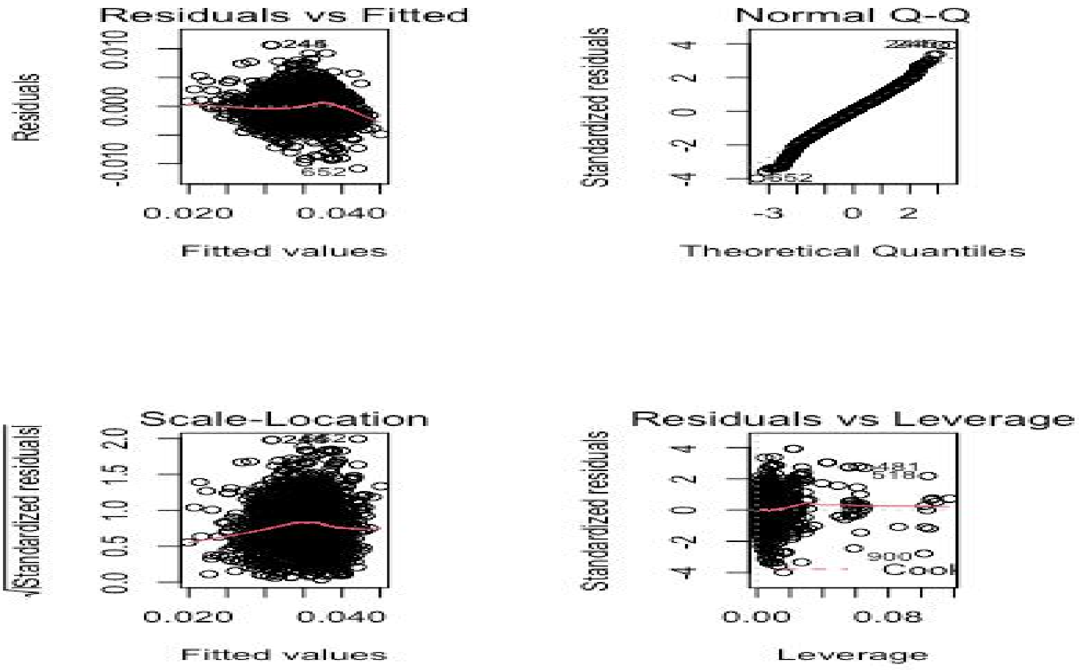


Figure.8 the plots of fit2


Residuals vs Fitted
Residuals
0.010
0.000
-0.010
0245
652
0.020
0.040
Fitted values
Normal Q-Q
Standardized residuals
4
2
0
-2
-4
652
-3
0
2
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
2.0
1.5
1.0
0.5
0.0
0.020
0.040
Fitted values
Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
481
518
900
Cook
0.00
0.08
Leverage