Analyzing The Significant Variable That Affects White Wine Quality?

Kaying Ng and Hanying Li

191005668 and 197001004

Submitted to: Lynn A. Agre, MPH, PhD

Statistics 467: Section H6

August 18, 2022

Abstract

This study investigates the relationship between sensory tests - subjective white wine quality of Vinho Verde by 3 assessors and physicochemical tests - variation in the number of ingredients, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. Physicochemical tests and sensory tests as measures of white wine quality are defined using the quality of the white wine of the Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal 2009. The dataset contains 4898 wine samples. The level of quality is from 0 - 10, which is a discrete variable. And other variables are all continuous variables. The bivariate relationship between white wine quality and all the factors from physicochemical tests that can affect quality such as alcohol, pH, etc are investigated with chi-square test statistics, logistic regression model, and other graph analyses to examine the effect of the white wine quality. In multivariate analysis, a new variable–quality_2 was created. This new variable is obtained by classifying the original variable quality. According to different quality scores, quality with a score of 7 or above was classified as high quality and represented by 1 in the dataset, while quality with a score of less than 7 was classified as low quality and represented by 0 in the dataset. Therefore, the new variable quality_2 is composed of the numbers 0 and 1; it is a binary form of variable, and its existence is more convenient to analyze intuitively. Furthermore, another new variable quality_3 is also created for more detailed grouping and more earise for discriminant analysis. In variable quality_3, according to the score of original quality, 1 represented the score of quality range between 3 to 4; 2 represented the score of quality range between 5 to 6; 3 represented the score quality range between 7 to 9. In multivariate analysis, all of logistic regression , multiple linear regression, forward stepwise regression , discriminant linear and MANOVA are used to analyze the relationship between white wine quality and all the factors that can affect quality of wine.

Keywords: quality, wine, physicochemical

Introduction

Our thesis is to explore and analyze all the physical factors contained in white wines that may affect the quality of white wines. When comparing the quality of white wines, we consider both the influence of physical factors and the direct sensory (mouthfeel). According to Richard Gawel, Paul A. Smith, Sara Cicerale & Russell Keast(2017), the taste of white wines includes tactile, chemosensory, and mouthfeel attributes. The results of the sensory test in the dataset we used were obtained from 3 professional assessors.

Although the taste test is subjective, experts matched wine descriptions better than non-experts through the research of HARRY T. LAWLESS (1984). Jan Schiefer, Christian Fischer (2008) also mentions that experts are more accurate than consumers when it comes to evaluating white wines. Furthemore, this view is also supported by Helene Hopfer and Hildegarde Heymann(2014)'s research. Also, wine experts usually use descriptive terminology terms when describing wines, which help us more directly and accurately derive the differences in quality exhibited by different white wines. Select expert taste test results allow us to analyze the quality of white wines in a way that reduces the inaccuracies of subjective testing.

In exploring the impact of physical factors on the quality of white wines, we used the study summarized by Scott Schulfer and Sina Tech to gain a deeper understanding of the function and classification of the acid variable, including fixed acidity, volatile acidity, and citric acid in the dataset we used.

Furthermore, according to Paul Kaan's (2018), it is better to group the variable residual sugar because it cannot be turned into a normal distribution by simple transformation. At the same time, we also used the studies of P.R. Jones; R. Gawel, I.L. Francis; and E.J. Waters(2008) to determine the extent to which the variable alcohol in our datasets affects the quality and taste of white wines. In addition, through the studies of Rocío Gutiérrez-Escobar, María José Aliaño-González, and Emma Cantos-Villar(2021), we learned that the polyphenols contained in wine are highly correlated with health-promoting properties (antioxidant and cardioprotective, among others).

According to J. T. Williams, C. S. Ough, and H. W. Berg(1978), white wines do not differ significantly in absolute quality, and there are several variables (both physical and chemical) that the winemaker can control to influence the taste and quality of white wines. Therefore, the purpose of our study is to analyze all physical variables in our selected dataset to determine the magnitude of the correlation between the effects of different variables on the quality of white wines. This can help winemakers analyze how to produce rich tasting and high-quality white wines, and also can help consumers promote health to some extent through high-quality white wines.

Also, through Günter Schamel's (2006) study, the analysis of wine can help the development of the wine market. Higher quality wines can make a brand more influential and reap more benefits.

According to Uli Fischer, Ann C. Noble and Am J Enol Vitic's research (1994), the pH value plays a relatively large role in the quality of white wine, because the pH value can control the acidity and bitterness of white wine taste, and the mouthfeel of white wine is one of the important criteria for quality evaluation. In the multivariate test, we will use different regression models to select the factors that have a greater impact on the quality of white wine and compare them with this study.

According to the IOWA state university extension and outreach's studies (2018) in white wine, the variable total sulfur dioxide includes variable free sulfur dioxide and other chemicals such as aldehydes, pigments, or sugars. In order to make the multivariate variables test more accurate, the variable free sulfur dioxide has been removed in the regression model, only maintaining the variable total sulfur dioxide.

After selecting the factors that greatly affect the quality in the dataset, we still consider that the factors that affect the taste of white wine are not all determined by the factors provided in the dataset. As Puckette, M. (2013) mentioned, the quality of white wine is also affected by different temperatures and different grape varieties. Therefore, we will also consider more factors in the subsequent data analysis and research.

In the multivariate analysis, we refer to the mathematical formula of linear regression and logistic regression as described by Jeff Sauro and James R. Lewis's research (2016). For the discriminant regression, we refer to the mathematical formula explained by Saed Sayad.

In this study, **the first hypothesis is:**

The null hypothesis (H0): Both the full and simple models fit the data equally well. As a result, the simple model should be used.

The alternative hypothesis (H1): The full model significantly outperforms the simple model in terms of data fit. As a result, the full model should be used.

The likelihood ratio test for the regression parameter will be used to examine the results and decide whether to accept the null hypothesis.

The second hypothesis is:

The null hypothesis (H0): there is no significant relationship between variable fixed acidity and variable alcohol, which means the beta coefficient is equal to 0.

The alternative hypothesis (H1): there is a significant relationship between variable fixed acidity and variable alcohol, which means the beta coefficient is not equal to 0.

The multiple logistic regression should be used to examine the model and calculate the p-value and decide whether to accept the null hypothesis.

The third hypothesis is:

The null hypothesis (H0): The full model of logistic regression and the nested model of logistic regression fit the data equally well. Thus, you should use the nested model of logistic regression.

The alternative hypothesis (H1): The full model of logistic regression fits the data significantly better than the nested model of logistic regression. Thus, you should use the full model of logistic regression.

Using likelihood ratio test to examine the hypothesis test.

Methods

The dataset for this study focuses on the wine quality of the Vinho Verde brand of wine from a small region in northern Portugal. It was gathered between May 2004 and February 2007 by the Viticulture Commission of the Vinho Verde Region (CVRVV), as well as P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. The datasets contain 2 main parts, the physicochemical test, and the sensory test. The dataset has 4898 numbers of samples and 12 variables. But we randomly select 100 samples when quality equal to 3 to 4, quality equal to 5 to 6, and quality equal to 7 to 9. Total sample size is 300 observations.

The fixed acidity (grams per liter, or g/L) means low volatility acids, mostly tartaric acids; they do not easily evaporate. This variable is a continuous variable with a normal distribution; the range is between 3.8 to 14.2 g/L. Volatile acidity (grams per liter, or g/L)) is the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant vinegar taste. This variable is a continuous variable with a right-skewed distribution range between 0.08 to 1.10 g/L. The citric acid (grams per liter, or g/L), found in small quantities, can add 'freshness' and flavor to wines. It is a continuous variable with sharp normal distribution from 0 to 1.66 g/L. Residual sugar (grams per liter, or g/L) is the amount of sugar remaining after fermentation stops. In the raw dataset, the variable is continuous with right-skewed distribution varied from 0.6 to 65.8 g/L. It is a wise idea to switch the residual sugar variable from continuous to discrete since the right-skewed distribution is not an acceptable shape for analysis; it has 4 levels of sweetness in white wine; according to the residual sugar chart of wine by Paul Kaan (2018), the residual sugar has 5 levels; 0-3 g/L names Brut Zero, since it is unlikely to have 0 residual sugar, it means nearly zero sugar remaining; 4-6 g/L names Extra Brut, which consider bone dry; 7-12 g/L names Brut, which consider dry; 12-17 g/L names Extra Dry, which consider fruity; larger than 17 m/L names Sec. It is rare to find the amount of residual sugar in white wine larger than 45 g/L. There is only one sample greater than 45, therefore let this sample integrate with Sec. Chlorides (grams per liter, or g/L) are the amount of salt (sodium chloride) in the wine. The variable is continuous with the right-skewed distribution; the range is from 0.012 to 0.099 g/L. According to the covariance table, free sulfur dioxide has strong overlap with the variable total sulfur dioxide, thus free sulfur dioxide is being deleted from the datasets. According to the research, free sulfur dioxide is somehow included in total sulfur dioxide (Admin, 2018). Total sulfur dioxide (milligrams per liter, or g/mL) is the amount of free and bound forms of SO2; 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. This variable is continuous with a normal distribution from 9 mg/L to 440 mg/L. Density (grams per milliliter, or g/mL) refers to the density of water that is close to that of water depending on the percent alcohol and sugar content. This variable is continuous with a normal distribution from 0.9871 to 1.0390 g/ml. pH represents how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); pH is a continuous variable with a normal distribution from 2.720 - 3.820. Sulphates (milligrams per liter, or g/mL), which are mainly potassium sulphate, are a wine additive that can contribute to sulfur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant. This variable is a continuous variable with a normal distribution from 0.22 g/L to 1.08 g/L. Alcohol (% by volume) is the percent alcohol content of the wine. It is a continuous variable with right-skewed distribution from 8% to 14.2%. This study would not delete any variables from datasets and no missing values. Quality is an output discrete variable (rating from 0 to 10), which is based on a sensory dataset; the range of quality in the raw dataset is from 3 to 9. For doing logistic regression and discriminant analysis, the dataset adds 2 more variables which recode from quality variables, named quality_2 and qulaity_3. According to the description of the original dataset, the score of quality equal or higher than 7 means the white wine has a high quality, and the score of quality lower than 7 means low quality wine. Therefore, based on the difference in quality scores, white wines were divided into two groups based on quality, which is named as quality_2. One group represents high quality (denoted by 1) and the other group represents low quality (denoted by 0). Also, quality_3 is divided into three groups for exploring discriminant analysis; coding quality equals 3 to 4 as 0, quality equals 5 to 6 as 1, and quality equals 7 to 9 as 2, respectively; represent low, middle, high quality of white wines.

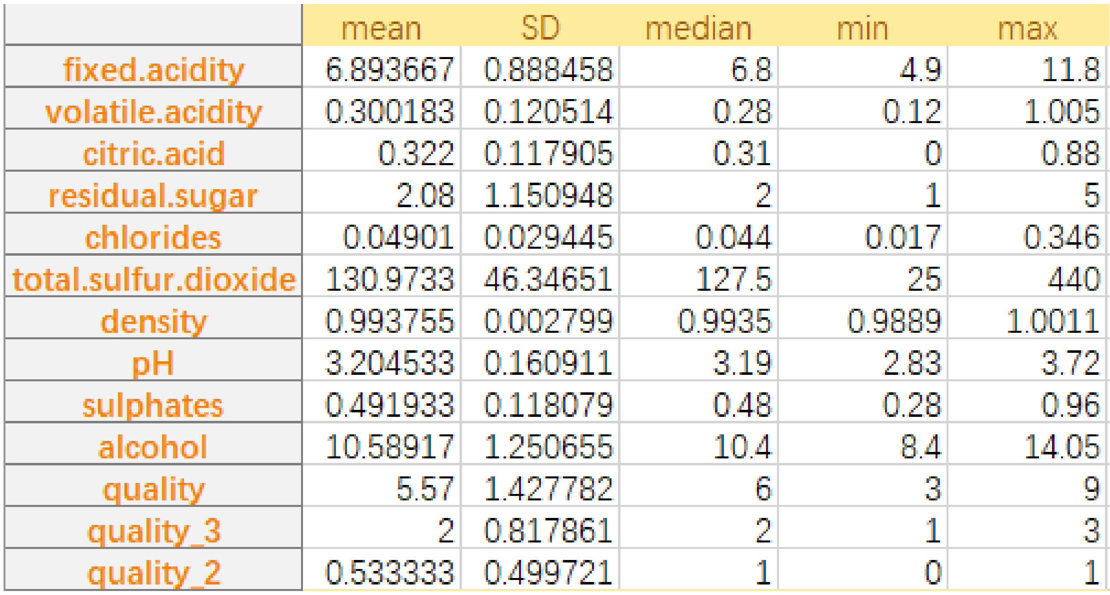
In the multiple variable analysis used in the selected dataset, due to the smaller sample size and distribution of all variable graphs displayed as normal distribution and right skewed, no variable was transformed in the analysis. In addition, since the values of many variables in the dataset are very small, such as pH and chlorides, if such variables with small values are transformed, the dataset will become negative, which is not conducive to the analysis of multivariate variables. Therefore, transform is not a good choice for the selected dataset and all variables are discussed and analyzed in their original state. Since our dependent variable is a discrete variable, this research will only focus on 3 models, linear regression model, logistic regression models, and discriminant analysis.

In the multiple linear regression, one important equation has been used which is $\widehat{Y} = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_p * X_p$, where $\widehat{Y}$ is the predicted value of the dependent variable, in this multiple variable analysis; $\widehat{Y}$ means the original output variable quality in the multiple linear regression model. X_1 through X_p are predictor variables, in this multiple variable analysis, p equal to 10 because only 10 independent variables are selected as predictor variables for analysis. Every variable in the dataset has a corresponding X, for example, X_1 means the first predictor variable fixed acidity in the multiple linear regression model. b0 is the value of Y when all of the independent variables (X_1 through X_p) are equal to zero, which is an intercept. b_1 through b_p are the estimated regression coefficients for each independent variable. Each regression coefficient represents the change in Y (output variable) relative to a one unit change in the respective predictor variable (X_1 through X_p). In the multiple logistic regression, equation $ln[\frac{Y}{1-Y}] = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_p * X_p$ has been used. The Y variable is the probability of obtaining a particular value of the nominal variable, which has the value range from 0 to 1. The meaning of other factors in multiple logistic regression equations is the same with the factors' meaning in multiple linear regression equations. For linear discriminant analysis, equation $Z = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_d x_d$ is applied into the analysis. β stands for linear model coefficient; equation is $\beta = C^{-1}(\mu_1 - \mu_2)$; μ_1, μ_2 are mean vectors. C stands for pooled covariance matrix; equation is $C = \frac{1}{n_1+n_2}(n_1 C_1 + n_2 C_2)$; C_1 and C_2 are covariance matrices.

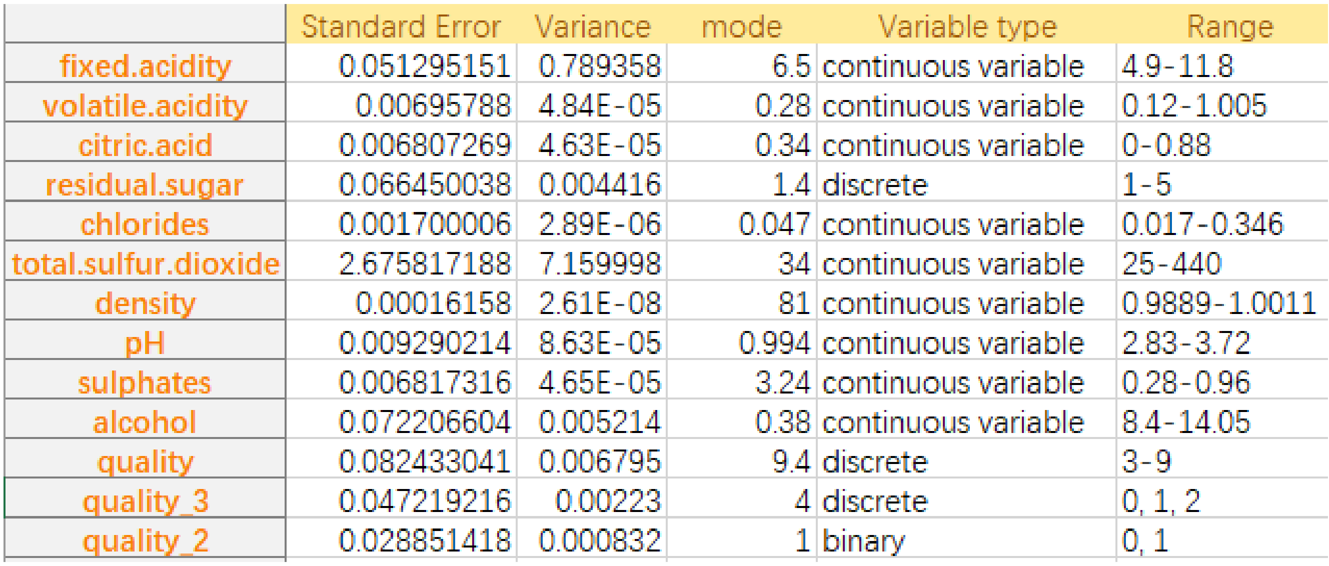
Results

Table I

Summary table of all variables; (n = 300)



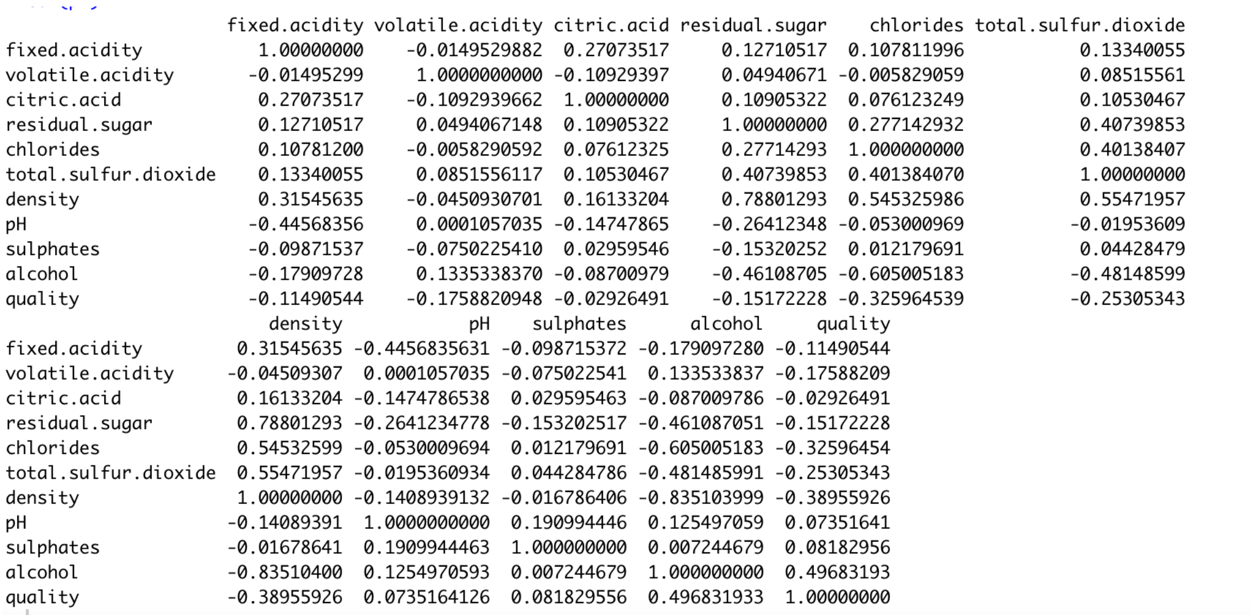
	mean	SD	median	min	max
fixed.acidity	6.893667	0.888458	6.8	4.9	11.8
volatile.acidity	0.300183	0.120514	0.28	0.12	1.005
citric.acid	0.322	0.117905	0.31	0	0.88
residual.sugar	2.08	1.150948	2	1	5
chlorides	0.04901	0.029445	0.044	0.017	0.346
total.sulfur.dioxide	130.9733	46.34651	127.5	25	440
density	0.993755	0.002799	0.9935	0.9889	1.0011
pH	3.204533	0.160911	3.19	2.83	3.72
sulphates	0.491933	0.118079	0.48	0.28	0.96
alcohol	10.58917	1.250655	10.4	8.4	14.05
quality	5.57	1.427782	6	3	9
quality_3	2	0.817861	2	1	3
quality_2	0.533333	0.499721	1	0	1



	Standard Error	Variance	mode	Variable type	Range
fixed.acidity	0.051295151	0.789358	6.5	continuous variable	4.9-11.8
volatile.acidity	0.00695788	4.84E-05	0.28	continuous variable	0.12-1.005
citric.acid	0.006807269	4.63E-05	0.34	continuous variable	0-0.88
residual.sugar	0.066450038	0.004416	1.4	discrete	1-5
chlorides	0.001700006	2.89E-06	0.047	continuous variable	0.017-0.346
total.sulfur.dioxide	2.675817188	7.159998	34	continuous variable	25-440
density	0.00016158	2.61E-08	81	continuous variable	0.9889-1.0011
pH	0.009290214	8.63E-05	0.994	continuous variable	2.83-3.72
sulphates	0.006817316	4.65E-05	3.24	continuous variable	0.28-0.96
alcohol	0.072206604	0.005214	0.38	continuous variable	8.4-14.05
quality	0.082433041	0.006795	9.4	discrete	3-9
quality_3	0.047219216	0.00223	4	discrete	0, 1, 2
quality_2	0.028851418	0.000832	1	binary	0, 1

Table II

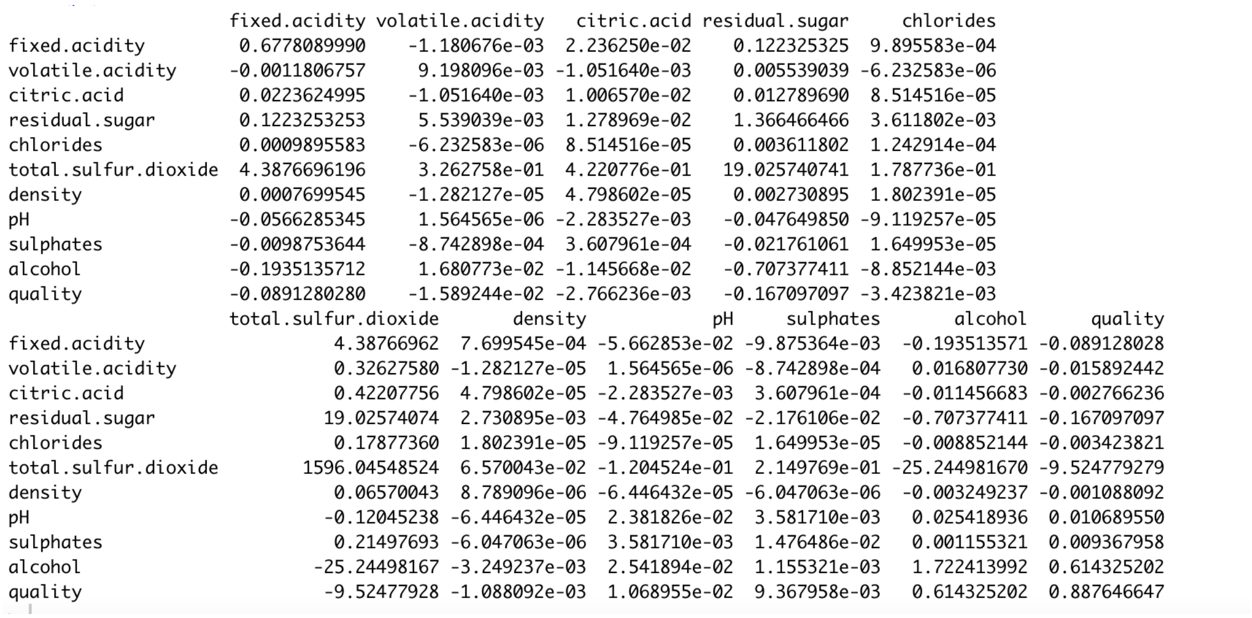
Correlation matrix for all variables; sample size (n =300)



```
                     fixed.acidity volatile.acidity citric.acid residual.sugar    chlorides total.sulfur.dioxide
fixed.acidity           1.00000000    -0.0149529882  0.27073517     0.12710517  0.107811996           0.13340055
volatile.acidity       -0.01495299     1.0000000000 -0.10929397     0.04940671 -0.005829059           0.08515561
citric.acid             0.27073517    -0.1092939662  1.00000000     0.10905322  0.076123249           0.10530467
residual.sugar          0.12710517     0.0494067148  0.10905322     1.00000000  0.277142932           0.40739853
chlorides               0.10781200    -0.0058290592  0.07612325     0.27714293  1.000000000           0.40138407
total.sulfur.dioxide    0.13340055     0.0851556117  0.10530467     0.40739853  0.401384070           1.00000000
density                 0.31545635    -0.0450930701  0.16133204     0.78801293  0.545325986           0.55471957
pH                     -0.44568356     0.0001057035 -0.14747865    -0.26412348 -0.053000969          -0.01953609
sulphates              -0.09871537    -0.0750225410  0.02959546    -0.15320252  0.012179691           0.04428479
alcohol                -0.17909728     0.1335338370 -0.08700979    -0.46108705 -0.605005183          -0.48148599
quality                -0.11490544    -0.1758820948 -0.02926491    -0.15172228 -0.325964539          -0.25305343
                        density            pH    sulphates      alcohol     quality
fixed.acidity         0.31545635 -0.4456835631 -0.098715372 -0.179097280 -0.11490544
volatile.acidity     -0.04509307  0.0001057035 -0.075022541  0.133533837 -0.17588209
citric.acid           0.16133204 -0.1474786538  0.029595463 -0.087009786 -0.02926491
residual.sugar        0.78801293 -0.2641234778 -0.153202517 -0.461087051 -0.15172228
chlorides             0.54532599 -0.0530009694  0.012179691 -0.605005183 -0.32596454
total.sulfur.dioxide  0.55471957 -0.0195360934  0.044284786 -0.481485991 -0.25305343
density               1.00000000 -0.1408939132 -0.016786406 -0.835103999 -0.38955926
pH                   -0.14089391  1.0000000000  0.190994446  0.125497059  0.07351641
sulphates            -0.01678641  0.1909944463  1.000000000  0.007244679  0.08182956
alcohol              -0.83510400  0.1254970593  0.007244679  1.000000000  0.49683193
quality              -0.38955926  0.0735164126  0.081829556  0.496831933  1.00000000
```

Table III:

Variance - Covariance matrix for variables; sample size (n =300)



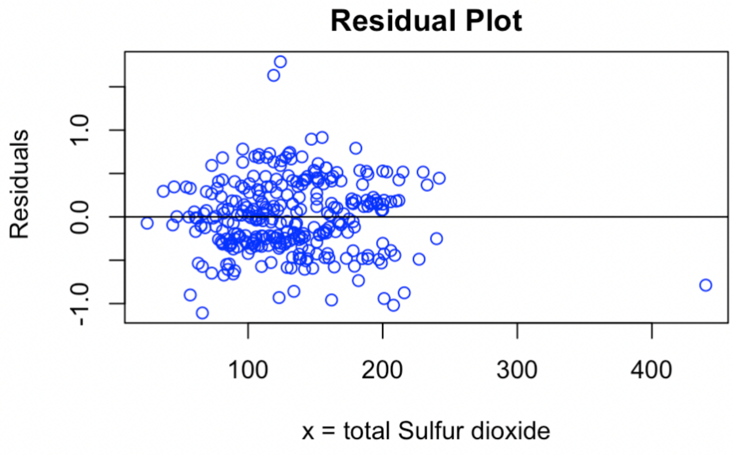
```
                     fixed.acidity volatile.acidity   citric.acid residual.sugar     chlorides
fixed.acidity         0.6778089990    -1.180676e-03  2.236250e-02    0.122325325  9.895583e-04
volatile.acidity     -0.0011806757     9.198096e-03 -1.051640e-03    0.005539039 -6.232583e-06
citric.acid           0.0223624995    -1.051640e-03  1.006570e-02    0.012789690  8.514516e-05
residual.sugar        0.1223253253     5.539039e-03  1.278969e-02    1.366466466  3.611802e-03
chlorides             0.0009895583    -6.232583e-06  8.514516e-05    0.003611802  1.242914e-04
total.sulfur.dioxide  4.3876696196     3.262758e-01  4.220776e-01   19.025740741  1.787736e-01
density               0.0007699545    -1.282127e-05  4.798602e-05    0.002730895  1.802391e-05
pH                   -0.0566285345     1.564565e-06 -2.283527e-03   -0.047649850 -9.119257e-05
sulphates            -0.0098753644    -8.742898e-04  3.607961e-04   -0.021761061  1.649953e-05
alcohol              -0.1935135712     1.680773e-02 -1.145668e-02   -0.707377411 -8.852144e-03
quality              -0.0891280280    -1.589244e-02 -2.766236e-03   -0.167097097 -3.423821e-03
                     total.sulfur.dioxide       density            pH     sulphates       alcohol      quality
fixed.acidity                  4.38766962  7.699545e-04 -5.662853e-02 -9.875364e-03  -0.193513571 -0.089128028
volatile.acidity               0.32627580 -1.282127e-05  1.564565e-06 -8.742898e-04   0.016807730 -0.015892442
citric.acid                    0.42207756  4.798602e-05 -2.283527e-03  3.607961e-04  -0.011456683 -0.002766236
residual.sugar                19.02574074  2.730895e-03 -4.764985e-02 -2.176106e-02  -0.707377411 -0.167097097
chlorides                      0.17877360  1.802391e-05 -9.119257e-05  1.649953e-05  -0.008852144 -0.003423821
total.sulfur.dioxide        1596.04548524  6.570043e-02 -1.204524e-01  2.149769e-01 -25.244981670 -9.524779279
density                        0.06570043  8.789096e-06 -6.446432e-05 -6.047063e-06  -0.003249237 -0.001088092
pH                            -0.12045238 -6.446432e-05  2.381826e-02  3.581710e-03   0.025418936  0.010689550
sulphates                      0.21497693 -6.047063e-06  3.581710e-03  1.476486e-02   0.001155321  0.009367958
alcohol                      -25.24498167 -3.249237e-03  2.541894e-02  1.155321e-03   1.722413992  0.614325202
quality                       -9.52477928 -1.088092e-03  1.068955e-02  9.367958e-03   0.614325202  0.887646647
```

Table II and Table III show the correlation relationship and covariance relationship for all variables in datasets. Based on the original dataset of the univariate variable analysis, only two new variables – quality_2 and quality_3 are added in this multivariate analysis. However, this new variable was created for logistic regression and discriminant analysis, it is binary form. Therefore, it does not make sense to put these new variables into the correlation and covariance matrix. Hence, the correlation relationship and covariance relationship between all variables shown here are the same as those in the univariate analysis report. The only difference is that only 300 samples in the original dataset are randomly selected in this multivariate analysis. The sample in the univariate analysis report is 4898 of the original dataset.

Bivariate Testing

Table IV:

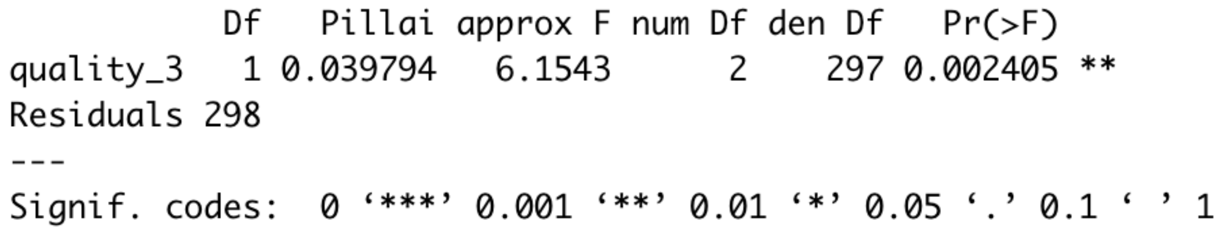
Residual diagnostic plot for the variable total sulfur dioxide; sample size (n=300)



This residual plot shows the points are randomly around the line residuals = 0. Thus, the random patterns demonstrate a good fit for a linear model.

Table V:

MANOVA test of response variables Sulphate and total sulfur dioxide versus group variable quality_3 for testing Hypothesis 1; sample size (n =300)

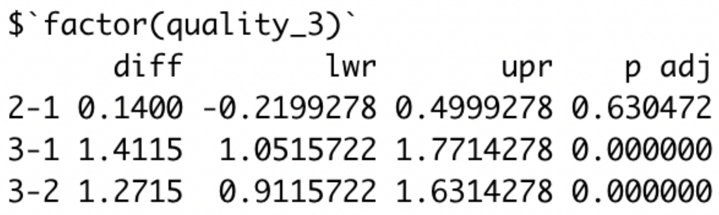


```
           Df   Pillai approx F num Df den Df   Pr(>F)   
quality_3   1 0.039794   6.1543      2    297 0.002405 **
Residuals 298                                            
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The table demonstrates the MANOVA testing for quality_3 versus sulphates and total sulfur dioxide. As a result, total.sulfur.dioxide and sulphates are significantly different among quality with 3 groups.

Table VI

Simultaneous Confidence Intervals of variable alcohol with dependent variable quality_3 (low/ mid/ high quality) by using Tukey HSD test ; sample size (n = 300)

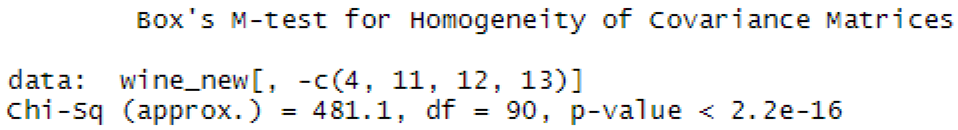


```
$`factor(quality_3)`
      diff        lwr       upr    p adj
2-1 0.1400 -0.2199278 0.4999278 0.630472
3-1 1.4115  1.0515722 1.7714278 0.000000
3-2 1.2715  0.9115722 1.6314278 0.000000
```

Above table shows that groups 3-1 and 3-2 of quality are significantly different at the 95 percent confidence level, according to the Tukey HSD test. This means that based on this Tukey HSD test concludes that group 1-2 is insignificantly different from group.

Table VII:

Testing for Equality of Covariance Matrices by using Box's M-test for Homogeneity of Covariance Matrices; (n=300)



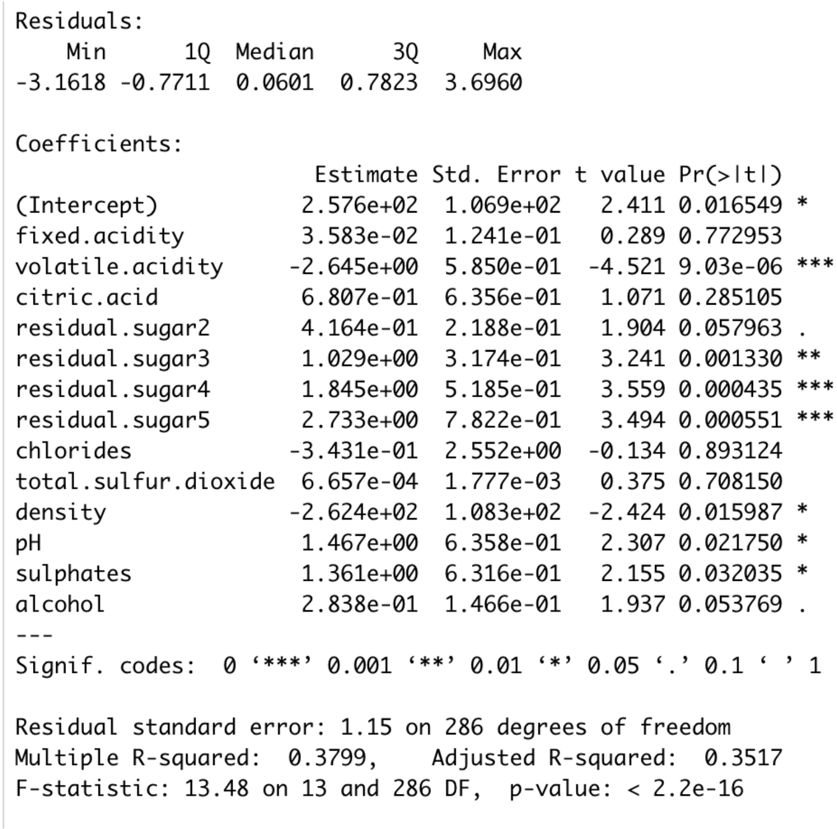
```
        Box's M-test for Homogeneity of Covariance Matrices

data:  wine_new[, -c(4, 11, 12, 13)]
Chi-Sq (approx.) = 481.1, df = 90, p-value < 2.2e-16
```

The datasets are divided into three parts for this test according to quality 3 (1,2,3) and test whether their covariance matrices are significant differences. The result shows that p-value is less than 0.05, which rejects the null hypothesis. The alternative hypothesis is that covariance matrices between 3 different groups according to quality_3 are significant differences.

Linear Regression

Table VIII:

Classical Linear Regression result for variables; sample size (n = 300)



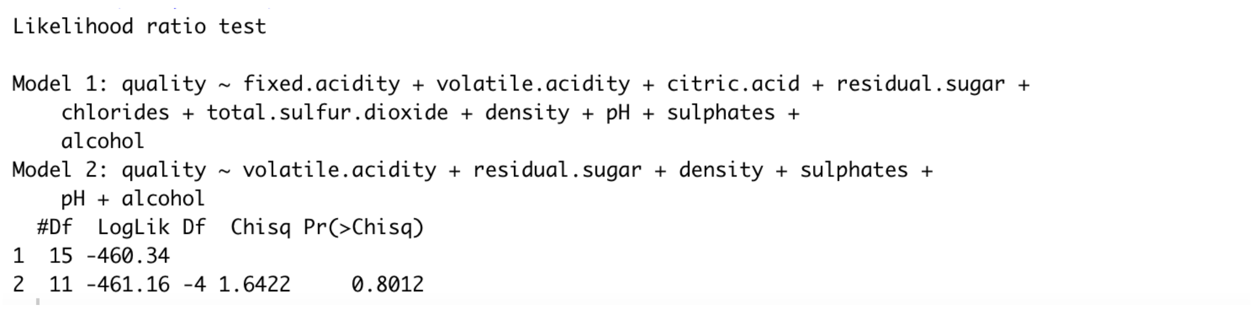
```
Residuals:
    Min      1Q  Median      3Q     Max
-3.1618 -0.7711  0.0601  0.7823  3.6960

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.576e+02  1.069e+02   2.411 0.016549 *
fixed.acidity         3.583e-02  1.241e-01   0.289 0.772953
volatile.acidity     -2.645e+00  5.850e-01  -4.521 9.03e-06 ***
citric.acid           6.807e-01  6.356e-01   1.071 0.285105
residual.sugar2       4.164e-01  2.188e-01   1.904 0.057963 .
residual.sugar3       1.029e+00  3.174e-01   3.241 0.001330 **
residual.sugar4       1.845e+00  5.185e-01   3.559 0.000435 ***
residual.sugar5       2.733e+00  7.822e-01   3.494 0.000551 ***
chlorides            -3.431e-01  2.552e+00  -0.134 0.893124
total.sulfur.dioxide  6.657e-04  1.777e-03   0.375 0.708150
density              -2.624e+02  1.083e+02  -2.424 0.015987 *
pH                    1.467e+00  6.358e-01   2.307 0.021750 *
sulphates             1.361e+00  6.316e-01   2.155 0.032035 *
alcohol               2.838e-01  1.466e-01   1.937 0.053769 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.15 on 286 degrees of freedom
Multiple R-squared:  0.3799,    Adjusted R-squared:  0.3517
F-statistic: 13.48 on 13 and 286 DF,  p-value: < 2.2e-16
```

As the table shows, the variable volatile acidity and the variable density has the negative coefficient, whereas other variables have the positive coefficients. The negative coefficient indicates that as the predictor variable increases, the response variable (quality) tends to decrease. Hence, the relationship between variable volatile acidity and the response variable is inversely proportional; and the variable density also has the inverse relationship with the output variable quality. However, all of the other variables that have positive coefficients have a direct proportion relationship between the response variable (quality), which means as these variables increase, the quality also increases. From the last column in the table, the variable with asterisk

(*) means that they are significant variables. More stars means more importance. Therefore, it is concluded that in multiple linear regression, Predictor variables that have an important impact on response variable (quality) include volatile acidity, residual sugar (set all of the residual sugar group as together), density, pH, and sulphates. And variable alcohol can also be considered as a less important factor because of its low p-value. Compared with the contents of the previous article mentioned in section introduction, the influence of pH value on the quality of white wine is indeed very important in the dataset we selected. The R-squared shown on the table represented the measure of how close the data are to the fitted regression line. R-sqaured is 0.3799 in this linear regression model. This value is lower than 0.5, which means a weak effect size. Adjusted R-squared is a measure of model accuracy for linear models it is always less than or equal to R-square. A value of adjusted R-squared is 1 means this model perfectly predicts values in the target field, whereas a value that is less than or equal to 0 means the model that has no predictive value. In this model shown in table, adjusted R-squared is 0.3517, which means that the prediction accuracy of this model is only a relatively average level. The p-value is less than 2.26e-16, which is less than 0.05. Hence, this model fits the data well.

Table IX:

Likelihood Ratio Tests result for the Regression Parameters; sample size (n = 300),

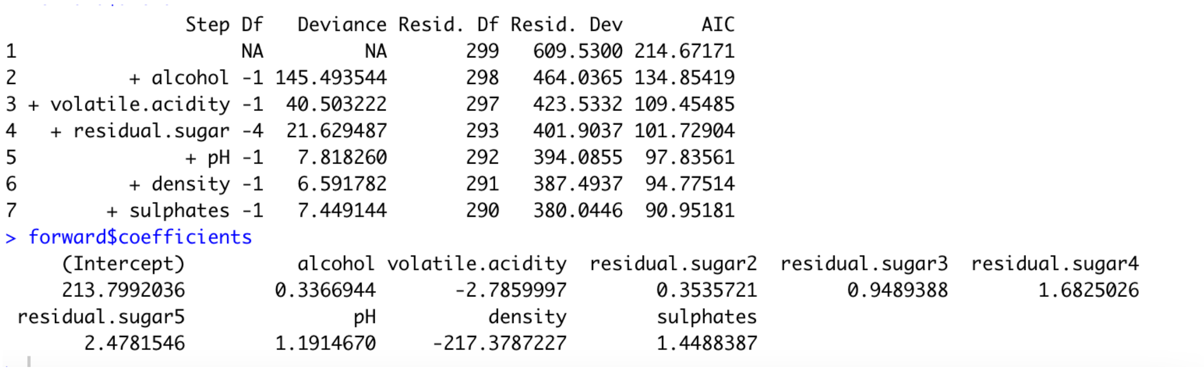


```
Likelihood ratio test

Model 1: quality ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + total.sulfur.dioxide + density + pH + sulphates +
    alcohol
Model 2: quality ~ volatile.acidity + residual.sugar + density + sulphates +
    pH + alcohol
  #Df  LogLik Df  Chisq Pr(>Chisq)
1  15 -460.34
2  11 -461.16 -4 1.6422     0.8012
```

As the table IXshows above, the simple model depends on the significant regression parameters obtained by Table I, which include volatile.acidity, residual sugar, density, sulphates, and alcohol. The full model depends on all of the variables included in the dataset. The Chi-Squared test statistic is 1.6422 and the corresponding p-value is 0.8012, as shown in the output. So the null hypothesis will be accepted because the p-value is more than 0.05. This means that the simple model should be used in linear regression. This result is matched with table I shows that these significant variables have more important effect on the response variable quality. This result will be taken as the answer to the first hypothesis mentioned in the section introduction.

Table X:

Forward Stepwise Regression result; sample size (n = 300)



```
                Step Df    Deviance Resid. Df Resid. Dev       AIC
1                    NA          NA       299   609.5300 214.67171
2          + alcohol -1 145.493544       298   464.0365 134.85419
3 + volatile.acidity -1  40.503222       297   423.5332 109.45485
4   + residual.sugar -4  21.629487       293   401.9037 101.72904
5               + pH -1   7.818260       292   394.0855  97.83561
6          + density -1   6.591782       291   387.4937  94.77514
7        + sulphates -1   7.449144       290   380.0446  90.95181
> forward$coefficients
     (Intercept)          alcohol volatile.acidity  residual.sugar2  residual.sugar3  residual.sugar4 
     213.7992036        0.3366944       -2.7859997        0.3535721        0.9489388        1.6825026 
 residual.sugar5               pH          density        sulphates 
       2.4781546        1.1914670     -217.3787227        1.4488387 
```

Forward stepwise regression starts with no predictors in the model, then adds the most contributive predictors, and stops when the improvement is no longer statistically significant. First, we fit the intercept-only model, this model had an AIC of 214.67171.

Next, we fit every possible one-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the intercept-only model used the predictor alcohol and this model had an AIC of 134.85419.

Then, we fit every possible two-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the one-predictor model added the predictor volatile.avidity. This model had an AIC of 109.45485.

Next, we fit every possible three-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the two-predictor model added the predictor residual sugar. This model had an AIC of 101.72904.

Then, we fit every possible four-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the three-predictor model added the predictor pH. This model had an AIC of 97.83561.

Next, we fit every possible five-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the four-predictor model added the density. This model had an AIC of 94.77514.

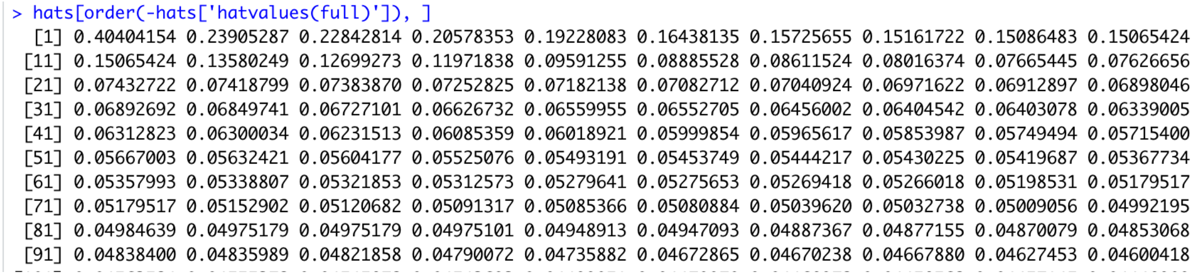
Next, we fit every possible six-predictor model. The model that produced the lowest AIC and also had a statistically significant reduction in AIC compared to the five-predictor model added the sulphates. This model had an AIC of 90.95181..

Next, we fit every possible five-predictor model, it turned out that none of these models produced a significant reduction in AIC, thus we stopped the procedure. And the final model should be quality~alcohol+ volatile.acidity+residual.sugar + pH + density + sulphates. The final model (the better model) obtained from forward stepwise regression is different from the result shown in likelihood ratio test (Table IX) and summary of classical linear regression (Table VIII).

Table XI:

Leverage values of all variables in descending order; sample size (n = 300)

(note:Because the dataset used contains 300 data, showing all the values is too large, so only the first 99 data with large leverage values are truncated)



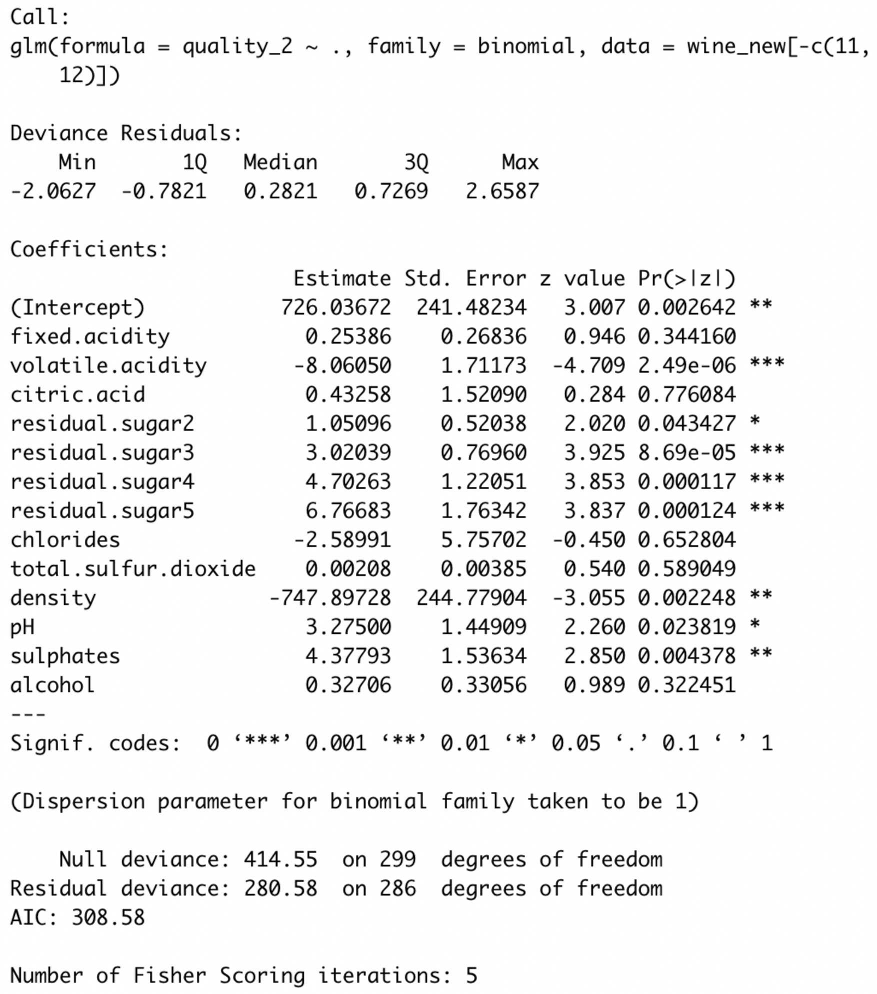
```
> hats[order(-hats['hatvalues(full)']), ]
  [1] 0.40404154 0.23905287 0.22842814 0.20578353 0.19228083 0.16438135 0.15725655 0.15161722 0.15086483 0.15065424
 [11] 0.15065424 0.13580249 0.12699273 0.11971838 0.09591255 0.08885528 0.08611524 0.08016374 0.07665445 0.07626656
 [21] 0.07432722 0.07418799 0.07383870 0.07252825 0.07182138 0.07082712 0.07040924 0.06971622 0.06912897 0.06898046
 [31] 0.06892692 0.06849741 0.06727101 0.06626732 0.06559955 0.06552705 0.06456002 0.06404542 0.06403078 0.06339005
 [41] 0.06312823 0.06300034 0.06231513 0.06085359 0.06018921 0.05999854 0.05965617 0.05853987 0.05749494 0.05715400
 [51] 0.05667003 0.05632421 0.05604177 0.05525076 0.05493191 0.05453749 0.05444217 0.05430225 0.05419687 0.05367734
 [61] 0.05357993 0.05338807 0.05321853 0.05312573 0.05279641 0.05275653 0.05269418 0.05266018 0.05198531 0.05179517
 [71] 0.05179517 0.05152902 0.05120682 0.05091317 0.05085366 0.05080884 0.05039620 0.05032738 0.05009056 0.04992195
 [81] 0.04984639 0.04975179 0.04975179 0.04975101 0.04948913 0.04947093 0.04887367 0.04877155 0.04870079 0.04853068
 [91] 0.04838400 0.04835989 0.04821858 0.04790072 0.04735882 0.04672865 0.04670238 0.04667880 0.04627453 0.04600418
```

From table XI, the largest leverage value for all of the observations is 0.40404154. Since the largest leverage value is still less than 2, it can be concluded that none of the observations in our dataset have high leverage, which means that none of the individuals have extreme predictor x value in our dataset.

Logistic regression

Table XII:

Summary table of logistic regression model by using quality_new as predictors; sample size (n=300)

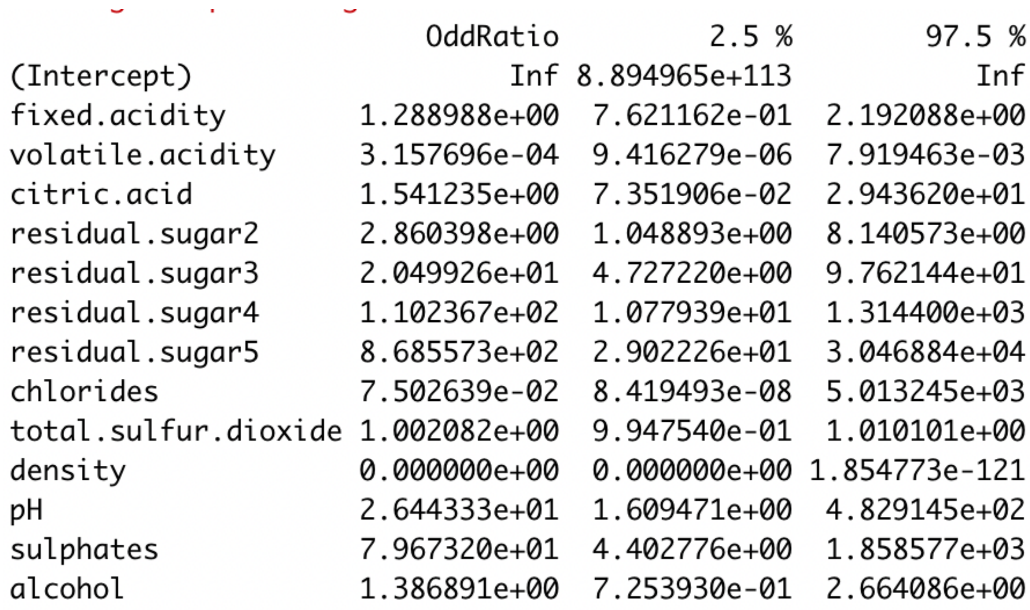


```
Call:
glm(formula = quality_2 ~ ., family = binomial, data = wine_new[-c(11,
    12)])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0627  -0.7821   0.2821   0.7269   2.6587

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           726.03672  241.48234   3.007 0.002642 **
fixed.acidity           0.25386    0.26836   0.946 0.344160
volatile.acidity       -8.06050    1.71173  -4.709 2.49e-06 ***
citric.acid             0.43258    1.52090   0.284 0.776084
residual.sugar2         1.05096    0.52038   2.020 0.043427 *
residual.sugar3         3.02039    0.76960   3.925 8.69e-05 ***
residual.sugar4         4.70263    1.22051   3.853 0.000117 ***
residual.sugar5         6.76683    1.76342   3.837 0.000124 ***
chlorides              -2.58991    5.75702  -0.450 0.652804
total.sulfur.dioxide    0.00208    0.00385   0.540 0.589049
density              -747.89728  244.77904  -3.055 0.002248 **
pH                      3.27500    1.44909   2.260 0.023819 *
sulphates               4.37793    1.53634   2.850 0.004378 **
alcohol                 0.32706    0.33056   0.989 0.322451
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 414.55  on 299  degrees of freedom
Residual deviance: 280.58  on 286  degrees of freedom
AIC: 308.58

Number of Fisher Scoring iterations: 5
```

This is the summary table of the logistic regression full model. Since residual.sugars are discrete variables, it would be factored before using a logistic regression model. The p-value of the intercept shows this model is significant. Volatile.acidity, and residual.sugars from group 3 to 5 are the most significant variables among other variables. And then density, sulphates, residual sugar in group 2, and pH are the second significant variables. And the AIC of this full model is 308.58.

Table XIII:

Table of odds ratios with 95% confidence intervals for the full Logistic Regression model; sample size (n=300)



```
                        OddRatio          2.5 %         97.5 %
(Intercept)                  Inf 8.894965e+113            Inf
fixed.acidity       1.288988e+00  7.621162e-01   2.192088e+00
volatile.acidity    3.157696e-04  9.416279e-06   7.919463e-03
citric.acid         1.541235e+00  7.351906e-02   2.943620e+01
residual.sugar2     2.860398e+00  1.048893e+00   8.140573e+00
residual.sugar3     2.049926e+01  4.727220e+00   9.762144e+01
residual.sugar4     1.102367e+02  1.077939e+01   1.314400e+03
residual.sugar5     8.685573e+02  2.902226e+01   3.046884e+04
chlorides           7.502639e-02  8.419493e-08   5.013245e+03
total.sulfur.dioxide 1.002082e+00  9.947540e-01   1.010101e+00
density             0.000000e+00  0.000000e+00 1.854773e-121
pH                  2.644333e+01  1.609471e+00   4.829145e+02
sulphates           7.967320e+01  4.402776e+00   1.858577e+03
alcohol             1.386891e+00  7.253930e-01   2.664086e+00
```

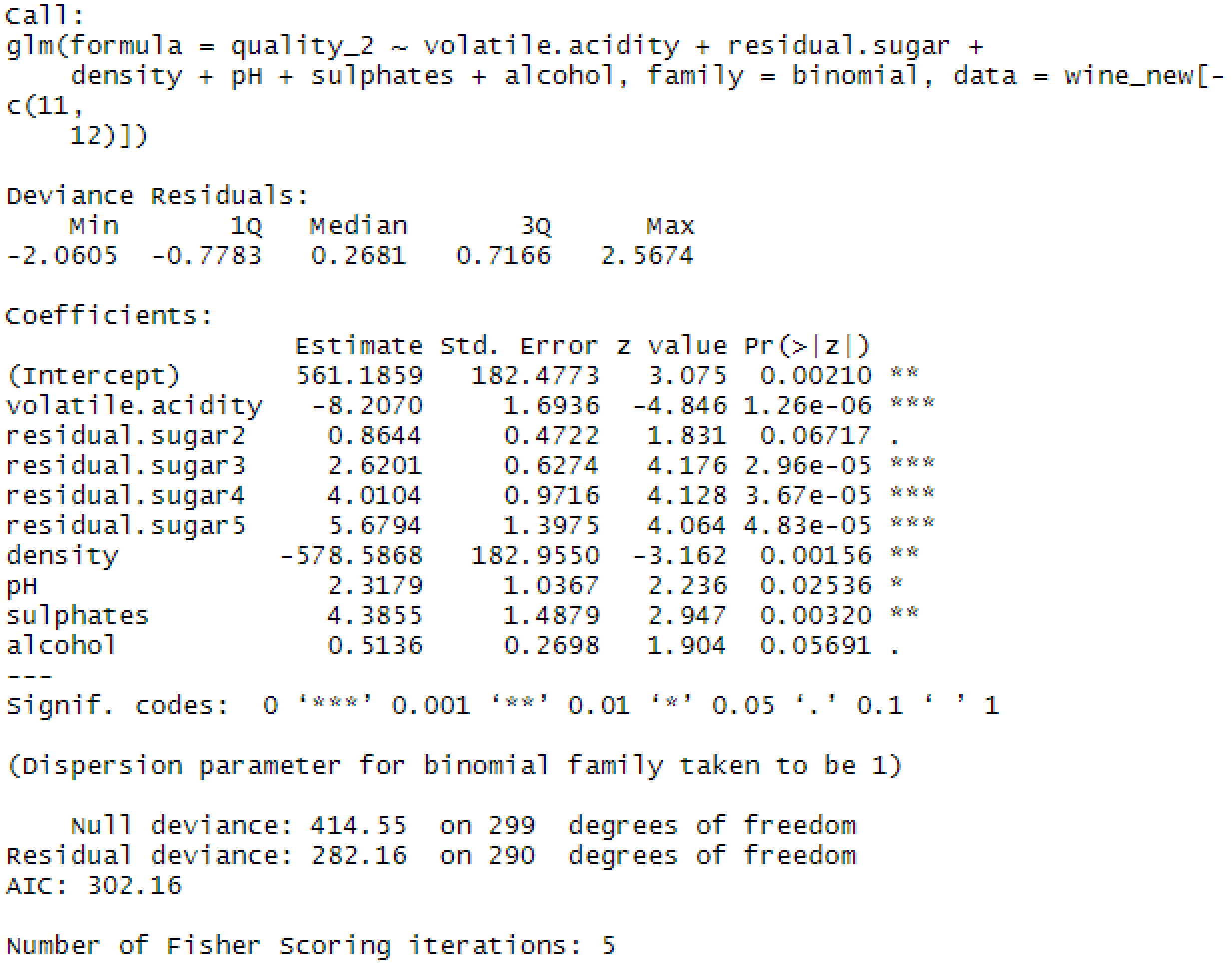
Null hypothesis: odd ratio = 1

Alternative hypothesis : odd ratio ≠ 1

The null hypothesis of 95% confidence interval for odd ratio in logistic regression is odd ratio = 1; alternative hypothesis is odd ratio ≠ 1. But we can also use the range of confidence intervals to find out the significant variables. According to the 95% confidence interval of odd ratio, fixed.acidity, citric acid, chlorides, and total.sulfur.dioxide are included 1 in the interval, which proves that they are statistically insignificant variables.

Table XV:

Table of the best nested model by AIC in a backward stepwise; (n = 300)

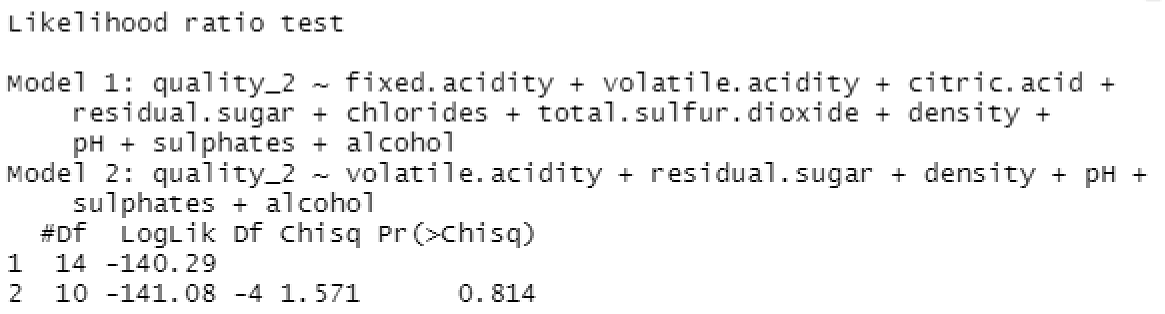


```
Call:
glm(formula = quality_2 ~ volatile.acidity + residual.sugar + 
    density + pH + sulphates + alcohol, family = binomial, data = wine_new[-
c(11, 
    12)])

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.0605  -0.7783   0.2681   0.7166   2.5674  

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)    
(Intercept)       561.1859   182.4773   3.075  0.00210 ** 
volatile.acidity   -8.2070     1.6936  -4.846 1.26e-06 ***
residual.sugar2     0.8644     0.4722   1.831  0.06717 .  
residual.sugar3     2.6201     0.6274   4.176 2.96e-05 ***
residual.sugar4     4.0104     0.9716   4.128 3.67e-05 ***
residual.sugar5     5.6794     1.3975   4.064 4.83e-05 ***
density          -578.5868   182.9550  -3.162  0.00156 ** 
pH                  2.3179     1.0367   2.236  0.02536 *  
sulphates           4.3855     1.4879   2.947  0.00320 ** 
alcohol             0.5136     0.2698   1.904  0.05691 .  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 414.55  on 299  degrees of freedom
Residual deviance: 282.16  on 290  degrees of freedom
AIC: 302.16

Number of Fisher Scoring iterations: 5
```

This table shows the best nested model by using backward stepwise, according to the AIC value. The p-value of the intercept is smaller than the critical value which proves that it is statistically different from zero. Volatile.acidity, residual.sugar from 3 to 5 are the most significant variables. Then density, sulphates, and pH are the second most significant variables. Alcohol has a p-value that is greater than 0.05 but yet very close to 0.05. Thus, that is part of the reason to consider it into the model. The AIC for this best nested model is 302.16 which is slightly lower than the AIC of the full model. Thus, a likelihood ratio test can be used to determine whether the models are different?

Table XVI

likelihood ratio test between full model and nested model of logistic regression; (n=300)



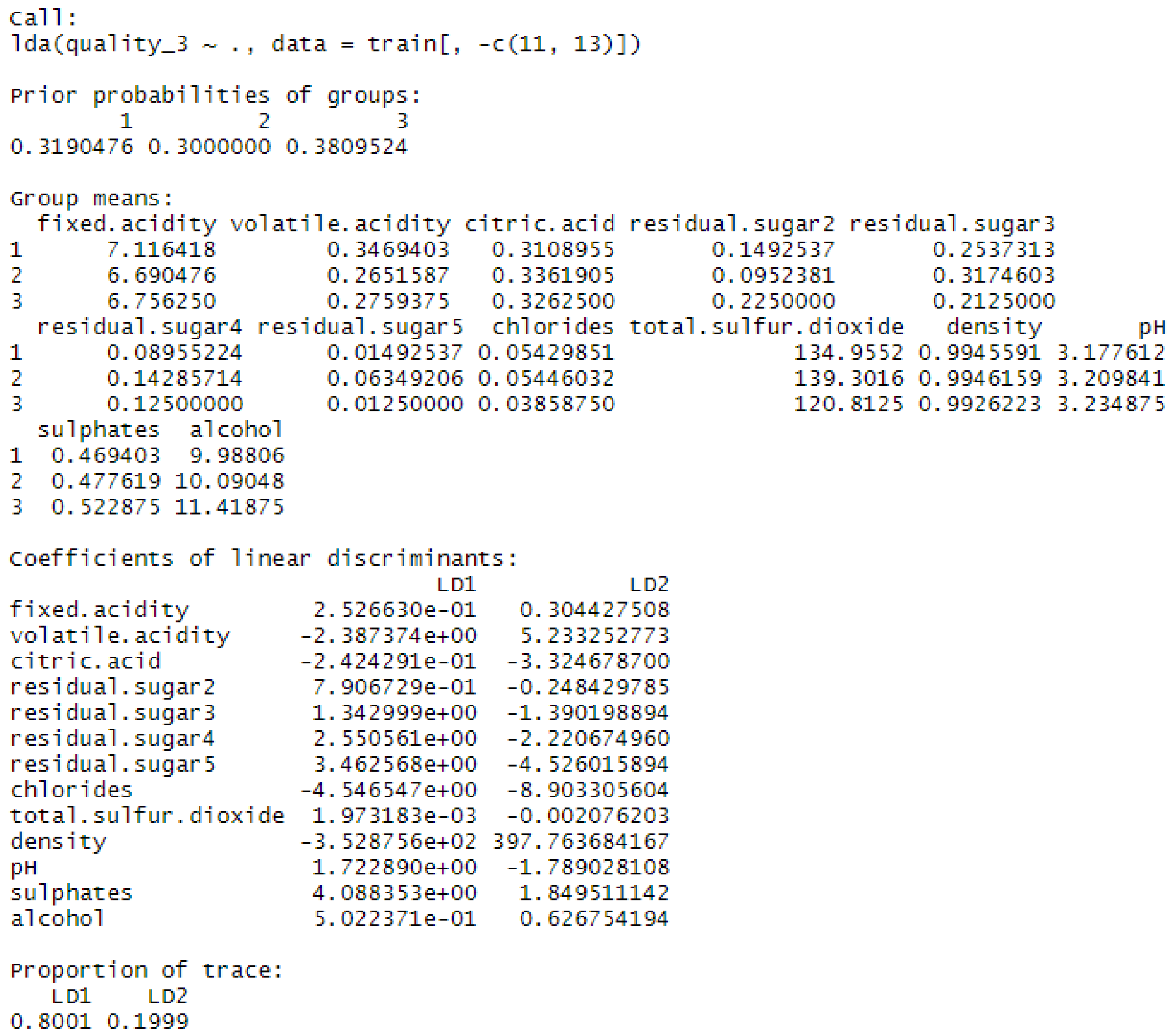
```
Likelihood ratio test

Model 1: quality_2 ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + total.sulfur.dioxide + density +
    pH + sulphates + alcohol
Model 2: quality_2 ~ volatile.acidity + residual.sugar + density + pH +
    sulphates + alcohol
  #Df  LogLik Df Chisq Pr(>Chisq)
1  14 -140.29
2  10 -141.08 -4 1.571      0.814
```

The null hypothesis for this likelihood ratio test is: The full model of logistic regression and the nested model of logistic regression fit the data equally well; Then nested models will be used. The above table shows that the p-value of the likelihood ratio test is 0.814 which is larger than the critical value 0.05. Thus, the null hypothesis can not be rejected. Thus, the nested model is being used rather than the full model.

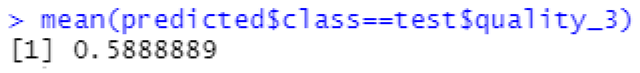
Linear Discriminant Analysis

Table XVII

Linear Discriminant for full model; sample size (n =300)



```
Call:
lda(quality_3 ~ ., data = train[, -c(11, 13)])

Prior probabilities of groups:
        1         2         3 
0.3190476 0.3000000 0.3809524 

Group means:
  fixed.acidity volatile.acidity citric.acid residual.sugar2 residual.sugar3
1      7.116418        0.3469403   0.3108955       0.1492537       0.2537313
2      6.690476        0.2651587   0.3361905       0.0952381       0.3174603
3      6.756250        0.2759375   0.3262500       0.2250000       0.2125000
  residual.sugar4 residual.sugar5  chlorides total.sulfur.dioxide   density       pH
1      0.08955224      0.01492537 0.05429851             134.9552 0.9945591 3.177612
2      0.14285714      0.06349206 0.05446032             139.3016 0.9946159 3.209841
3      0.12500000      0.01250000 0.03858750             120.8125 0.9926223 3.234875
  sulphates  alcohol
1  0.469403  9.98806
2  0.477619 10.09048
3  0.522875 11.41875

Coefficients of linear discriminants:
                               LD1          LD2
fixed.acidity         2.526630e-01   0.304427508
volatile.acidity     -2.387374e+00   5.233252773
citric.acid          -2.424291e-01  -3.324678700
residual.sugar2       7.906729e-01  -0.248429785
residual.sugar3       1.342999e+00  -1.390198894
residual.sugar4       2.550561e+00  -2.220674960
residual.sugar5       3.462568e+00  -4.526015894
chlorides            -4.546547e+00  -8.903305604
total.sulfur.dioxide  1.973183e-03  -0.002076203
density              -3.528756e+02 397.763684167
pH                    1.722890e+00  -1.789028108
sulphates             4.088353e+00   1.849511142
alcohol               5.022371e-01   0.626754194

Proportion of trace:
   LD1    LD2 
0.8001 0.1999 
```



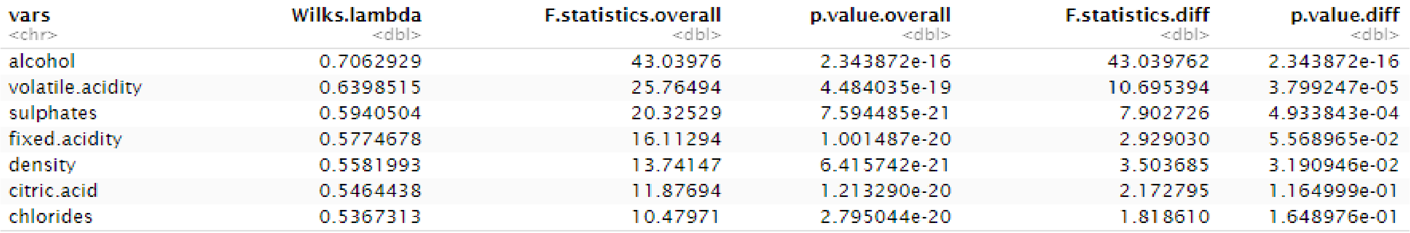
```
> mean(predicted$class==test$quality_3)
[1] 0.5888889
```

The dataset contains 10 variables (seen all of the residual sugar group as one factor) and 300 total observations. Variables fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, total.sulfur.sioxide, density, pH, sulphates and alcohol are used in the linear discriminant model.

The response variable quality_3 is divided into three groups, 1 represents the quality score range from 3 to 4; and 2 represents the score of quality range from 5 to 6; and 3 represents the score of quality range from 7 to 9. 70% dataset used as training set and remaining 30% used as test set. Prior probabilities of groups represent the proportions of every quality in the training set. For example, 31.90476% of all observations in the training set have quality scores from 3 to 4. Group displays the mean values for each predictor variable for low and high quality. Coefficients of linear discriminants means the linear combination of predictor variables that are used to form the decision rule of the LDA model. After prediction, it can be concluded that the model correctly predicted the quality_3 for around 58.89% of the observations in our test dataset.

Table XVIII:

Stepwise forward variable selection for discriminant analysis model; (n=300)

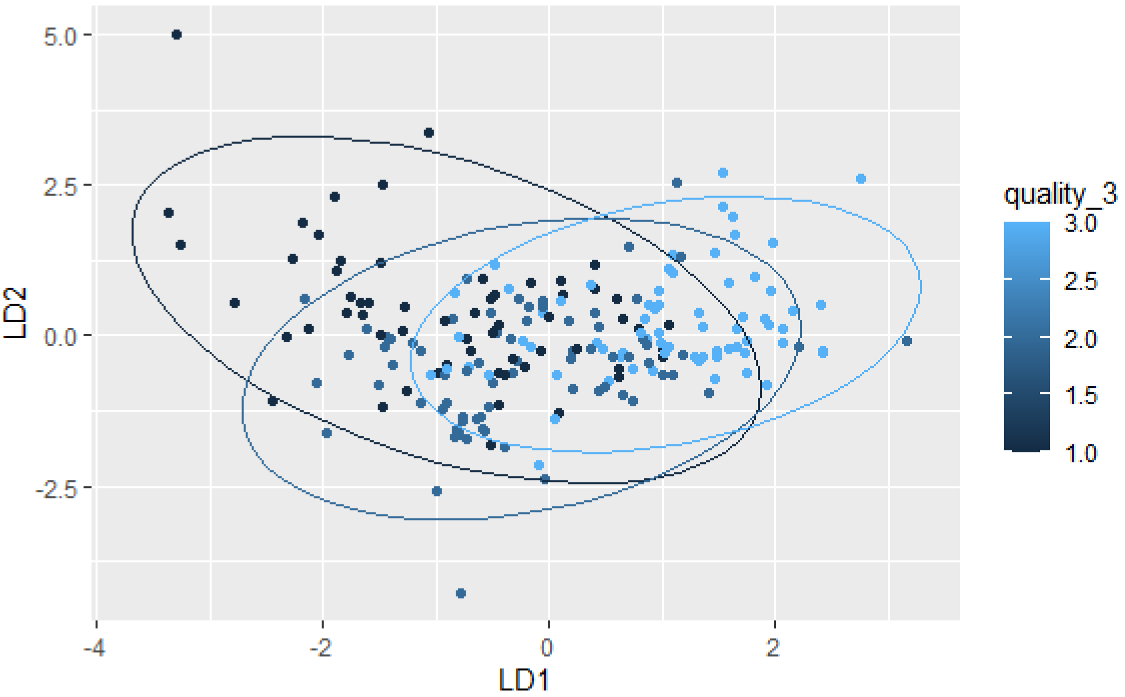


vars <chr>	Wilks.lambda <dbl>	F.statistics.overall <dbl>	p.value.overall <dbl>	F.statistics.diff <dbl>	p.value.diff <dbl>
alcohol	0.7062929	43.03976	2.343872e-16	43.039762	2.343872e-16
volatile.acidity	0.6398515	25.76494	4.484035e-19	10.695394	3.799247e-05
sulphates	0.5940504	20.32529	7.594485e-21	7.902726	4.933843e-04
fixed.acidity	0.5774678	16.11294	1.001487e-20	2.929030	5.568965e-02
density	0.5581993	13.74147	6.415742e-21	3.503685	3.190946e-02
citric.acid	0.5464438	11.87694	1.213290e-20	2.172795	1.164999e-01
chlorides	0.5367313	10.47971	2.795044e-20	1.818610	1.648976e-01

This stepwise forward selection method is based on Wilk's Lambda criterion (scale range from 0 to 1). Wilk's lambda in linear discriminant analysis evaluates the contribution of each level of independent variable to the model. In this table, the leftmost column shows the significant variables for building the linear discriminant analysis model. There are alcohol, volatile.acidity, sulphates, fixed.acidity,density, citric.acid, and chlorides. Alcohol has the stand out Wilks lambda, which means it is the most significant variable among others. Sulphates to chlorides are weight similar to the model.

Figure I:

Plot of linear discriminant Analysis(LDA); LD1 vs LD2; (n=300)



This plot shows the 3 groups of quality_3 (groups = 1,2,3). Three ellipses represent different regions due to groups. The sky blue represents quality_3 = 3 (high quality); blue-gray represents quality_3 = 2 (middle quality); dark-blue represents quality_3 = 1 (low quality). The ellipse of high quality is on the right side of the plot. The ellipse of middle quality is placed on the center of the plot. And the ellipse of low quality is placed on the left side of the plot. However, they have large shared regions between the ellipses.

Discussion

In the analysis process, univariable analysis focuses more on the individual characteristics of a single variable, while multivariate analysis not only focuses on the differences of each variable, but also demonstrates the relationship between different variables. In the univariate analysis, Hotelling's T^2 Confidence Intervals has been used as an important method to test whether a variable is significant. Likelihood ratio test also been used as another method to test the variable selected by method Hotelling's T^2 Confidence Intervals. However, use single method to test every variable in order is time consuming and tedious. Hence, in bivariate and multivariate analysis, a full model has been created to test all variables at the same time. Through the summary of the model, the significant level of every variable is obvious, and we got the same result about variable importance with the univariate analysis. In addition, in univariate analysis, the characteristics of variables are usually shown using figures, while in multivariate analysis, the characteristics of variables and the relationships between variables are usually shown using summary results of different models. Thus, multivariate analysis examines the relationship between variables in greater depth.

In the bivariate results, the result of Simultaneous Confidence Intervals of variables shows that dependent variable quality_3 (groups = 1,2,3) means of group 3 and group 2 are significantly different; means of group1 and group 3 are significantly different. Additionally, the homogeneity of covariance matrices as determined by Box's M-test for equality of covariance matrices reveals that there are notable disparities in quality 3 across each group. However, the LAD plot in the discriminant analysis reveals that group 3 and group 2's ellipses share a significant portion of space, and group 1 and group 3's as well. In general, this could be the cause of the variables' non-normal distribution since some right-skewed distributions cannot be transformed into a normal distribution. In this project, we focused more on finding the best model and the most influential variables, so the results of the important variables selected by regression and the significant variables derived from the analysis of univariable are not different. Finally, we got the variables alcohol, volatile.acidity, residual.sugar, pH, density and sulphates are significant. The reason why variables volatile.acidity, pH, and sulphates are important is that these factors all affect the acidity of white wines. Acidity is a sensitive factor to the human palate, and white wine's quality score is based in part of people's mouthfeel. Therefore, these variables are important for linear regression and logistics regression models. In addition, the dryness of wine is determined by the sugar content, so the variable residual sugar is also important. Besides, alcohol can affect the odor and intensity of white wine, which will also affect the taste of wine tasters. Therefore, variable alcohol is also an important factor. Therefore, the winemaker should pay more attention to these important variables in the white wine making process in order to produce a higher quality white wine.

In this multivariate analysis, only 2 new variables are created recoding output variables for building logistic and discriminant analysis. In the logistic regression model, the new variable quality_2 makes the analysis more convenient. Unlike linear regression, logistic regression usually used to handle classification problems, and it provides discreet output. Therefore, using the grouped quality as a new response can make logistic regression more efficient. In the discriminant analysis, the new variable quality_3 is used as the response. Although in the logistic regression, we can conclude that there is no difference in the group of 2-1 by Tukey, in the discriminant linear mode, if we use quality_2 as the response, the correctness is very low compared to using quality_3 as the response. Therefore, we insist on using quality_3 as the output variable in discrimiant model. Because it will improve the correctness of full model.

Since these new variables are not continuous variables but categorical, they are not included in the correlation matrix and the variance-covariance matrix. So we maintain the result of correlation and variance-covariance matrix in the first paper, which eliminates the variable free.sulfur.dioxide due to high covariance with the variable total.sulfur.dioxide.

Since the analysis has a long way to go. Thus, this research can provide one sight of logistic regression and discriminant analysis. In discriminant analysis, several ellipses overlap one another. It might be the cause of the sample size's possible undersize. Additionally, it might be the cause of the need to restructure qualit_3; perhaps groups for numbers 3-6, 7 and 8-9 should be created. Since groups 6-7 make up a significant percentage of the raw dataset. As a result, the sample size will be too small if we aim to have an equal sample size for each group based on the previous groups. Thus, for future research, we suggest that cluster analysis might be a good method to analyze the dataset. For example, fixed.acidity, volatile.acidity, citric.acid, and pH can be clustered together as an acid group. And sulphates, total.sulfur.dioxide, and free.sulfur.dioxide can be clustered as sulfur components. The sample size may contain more than a thousand observations if the techniques were applied, making the model construction more accurate. The sample size is limited by the equal group size because the sample was drawn from three distinct groups in quality_3. The model might not be as accurate as more sample size.

The dataset includes some suggestions that might be enhanced in a later investigation. We learn that the grape type and fermentation temperature, if they were included in the datasets, may have an effect on the wine's quality in accordance with Puckette's article describing the production of white wine (Puckette, 2013). The tastes, sugar contents, and chemical makeup of different grapes vary. That's another factor in figuring out how good a wine is. The quality may also be impacted by the temperature of the fermentation process. The yeast may react more quickly if the temperature is a little warmer. Grape variety and fermentation temperature might be identified in the dataset for examination overall to investigate the quality of wine.

Reference

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Gawel, R., Smith, P. A., Cicerale, S., & Keast, R. (2017). The mouthfeel of white wine. *Critical Reviews in Food Science and Nutrition*, *58*(17), 2939–2956. https://doi.org/10.1080/10408398.2017.1346584

Gutiérrez-Escobar, R., Aliaño-González, M. J., & Cantos-Villar, E. (2021). Wine Polyphenol Content and Its Influence on Wine Quality and Properties: A Review. *Molecules*, *26*(3), 718. https://doi.org/10.3390/molecules26030718

Jones, P., Gawel, R., Francis, I., & Waters, E. (2008). The influence of interactions between major white wine components on the aroma, flavour and texture of model white wine. *Food Quality and Preference*, *19*(6), 596–607. https://doi.org/10.1016/j.foodqual.2008.03.005

Kaan, P. (2018, September 27). *Residual Sugar*. WINE DECODED. Retrieved August 1, 2022, from https://winedecoded.com.au/wine-words/residual-sugar/

Klatsky, A. L., Armstrong, M. A., & Friedman, G. D. (1997). Red Wine, White Wine, Liquor, Beer, and Risk for Coronary Artery Disease Hospitalization. *The American Journal of Cardiology*, *80*(4), 416–420. https://doi.org/10.1016/s0002-9149(97)00388-3

Lesschaeve, I. (2007, June 1). *Sensory Evaluation of Wine and Commercial Realities: Review of Current Practices and Perspectives*. American Journal of Enology and Viticulture. Retrieved August 1, 2022, from https://www.ajevonline.org/content/58/2/252.article-info

Schamel, G. (2006). Geography versus brands in a global wine market. *Agribusiness*, *22*(3), 363–374. https://doi.org/10.1002/agr.20091

Schiefer, J., & Fischer, C. (2008). The gap between wine expert ratings and consumer preferences. *International Journal of Wine Business Research*, *20*(4), 335–351. https://doi.org/10.1108/17511060810919443

Williams, J. T., Ough, C. S., & Berg, H. W. (1978, January 1). *White Wine Composition and Quality as Influenced by Method of Must clarification*. American Journal of Enology and Viticulture. Retrieved August 1, 2022, from https://www.ajevonline.org/content/29/2/92

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.

Puckette, M. (2013, April 24). *How White Wine is Made From Grapes to Glass*. Wine Folly. Retrieved August 18, 2022, from https://winefolly.com/deep-dive/how-is-white-wine-made/

Fischer, U. (1994, January 1). *The Effect of Ethanol, Catechin Concentration, and pH on Sourness and Bitterness of Wine*. American Journal of Enology and Viticulture. Retrieved August 18, 2022, from https://www.ajevonline.org/content/45/1/6.abstract

Sauro, J., & Lewis, J. R. (2016). An introduction to correlation, regression, and ANOVA. *Quantifying the User Experience*, 277–320. https://doi.org/10.1016/b978-0-12-802308-2.00010-2