Wine taste prediction by GLM

Hyeongcheol Park

2017-12-11

Introduction

Ubiquitos era is blurring out physical barrier. People from different countries consume different goods from all over the world. the Wine market is no exception of this phenomenon. Deviating the Europe continent, the wine products are getting consumed by more and more areas in the world.

Meanwhile, the wine production areas is currently enlarging. Not only old wine regions in Europe, but also new wine regions such as China, America continents, and Africa yield large amount of wine products. As the quality and taste of the wine heavily relies on the climate and environment of the region, the diversity of wines grow rapidly.

In this situation, traditional role of oenologist reach its limit. To match a wine product to the exact world market, we need another way of wine analyzing. GLM can be a way to conduct that task. By analyzing physicochemical ingredients of the wine which people in a region prefer the best, we can distribute wine products in very effecient way. In order to do that, we can conduct regression model analyzing to figure out the portion of the components of wine products.

This project aims at analyzing the relationship between the taste preference of traditional market and physicochemical ingredients of wine produced in traditonal region. This method can be used further to analyze the relationships between new wine market and new wine regions. As the result, we can easily match the demand and consumption in a easier way.

Data Discription

2. Understand data. Describe how the data is collected and what the data is. -> data description

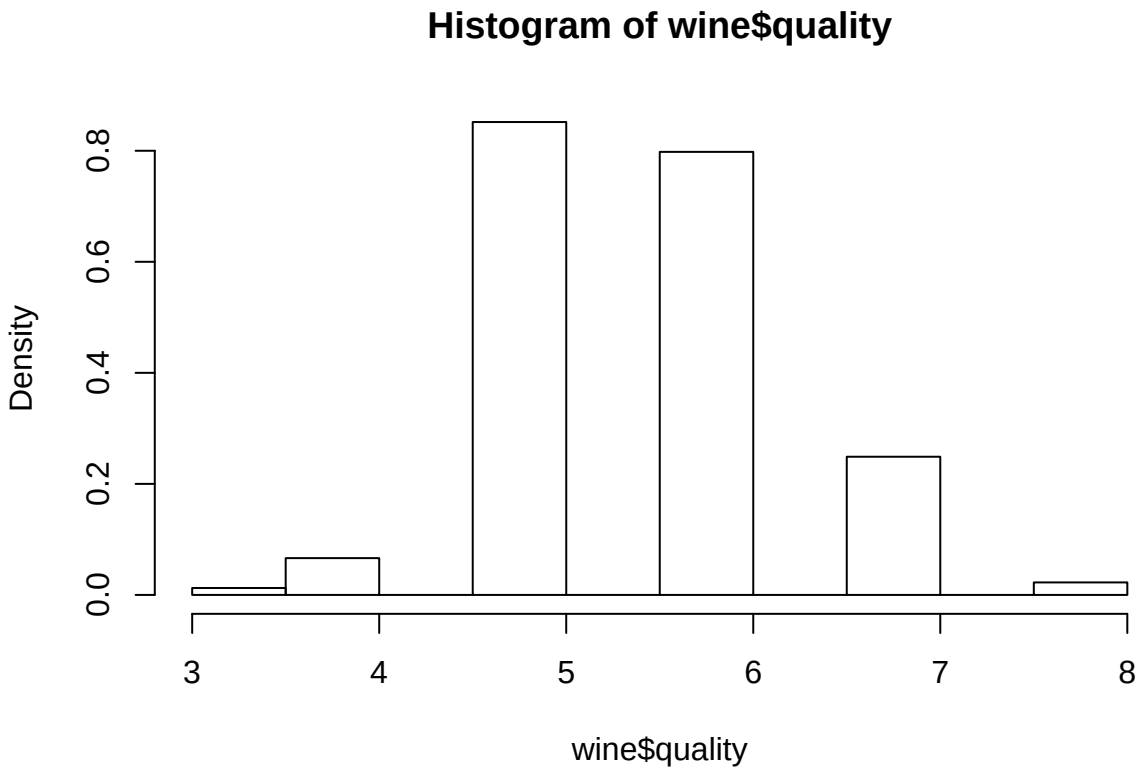
The wine from Portugal, vinho verde is the subject of this project which is a unique product from the Minho (northwest) region of Portugal. The data were collected from May/2004 to February/2007 with random sample method. The data was collected in two ways: physicochemical ingredients data and sensory data. Official certification entity (CVRVV), collected basic physicochemical statistics of red and white wine. For

the sensory data, three sensory assessors evaluted each sample in a scale that ranges from 0 (very bad) to 10 (excellent). The final sensory score is given by the median of these evaluations.

Original dataset contains red wine data and White wine data. In this project, we will consider only red wine. In the dataset, there are 11 variables of physicochemical statistics and 1 variable from sensory score, quality.

There are 1599 samples.

Attribute (units) Red wine	Min	Max	Mean	Median
Fixed acidity (g(tartaric acid)/dm3)	4.6	15.9	8.3	7.9
Volatile acidity (g(acetic acid)/dm3)	0.1	1.6	0.5	0.52
Citric acid (g/dm3)	0.0	1.0	0.3	0.26
Residual sugar (g/dm3)	0.9	15.5	2.5	2.2
Chlorides (g(sodium chloride)/dm3)	0.01	0.61	0.08	0.079
Free sulfur dioxide (mg/dm3)	1	72	14	14
Total sulfur dioxide (mg/dm3)	6	289	46	38
Density (g/cm3)	0.990	1.004	0.996	0.99
pH	2.7	4.0	3.3	3.31
Sulphates (g(potassium sulphate)/dm3)	0.3	2.0	0.7	0.62
Alcohol (vol.%)	8.4	14.9	10.4	10.2
quality	3	8	5.6	6

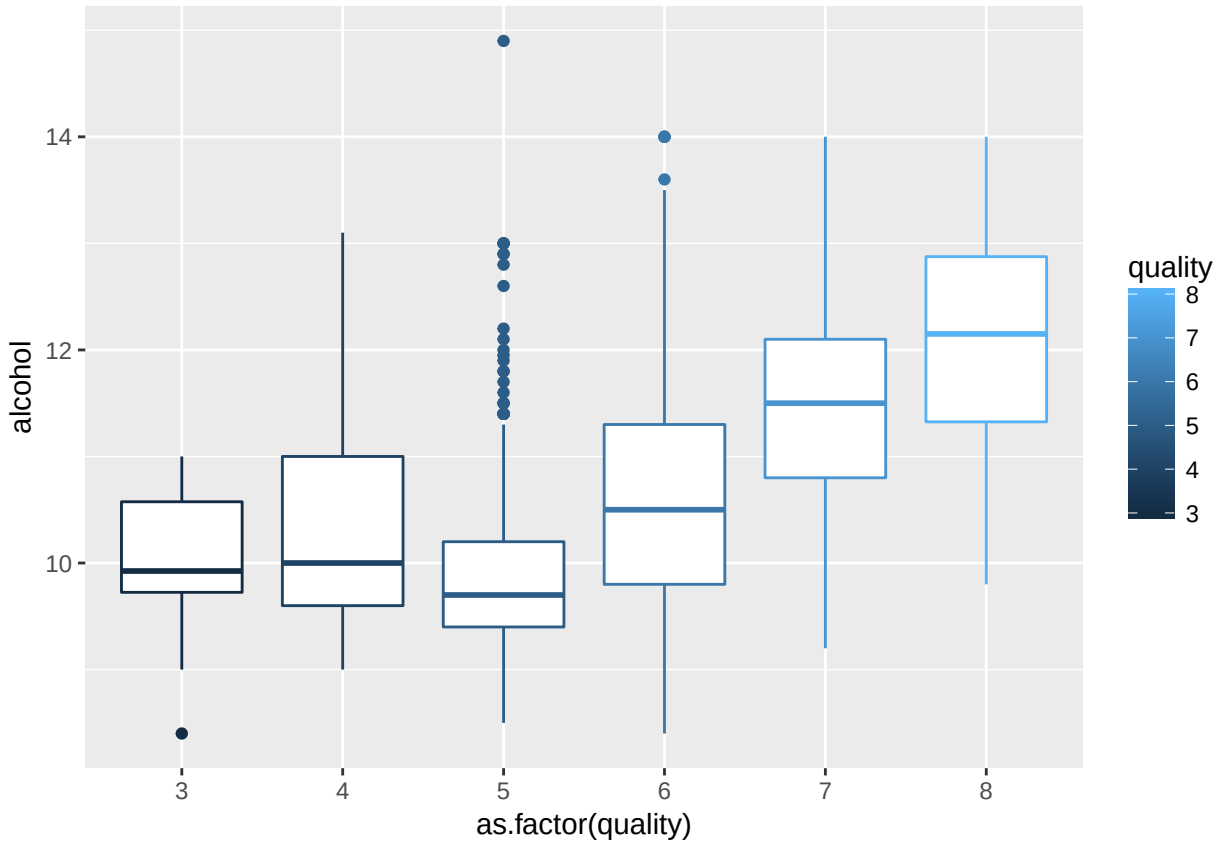


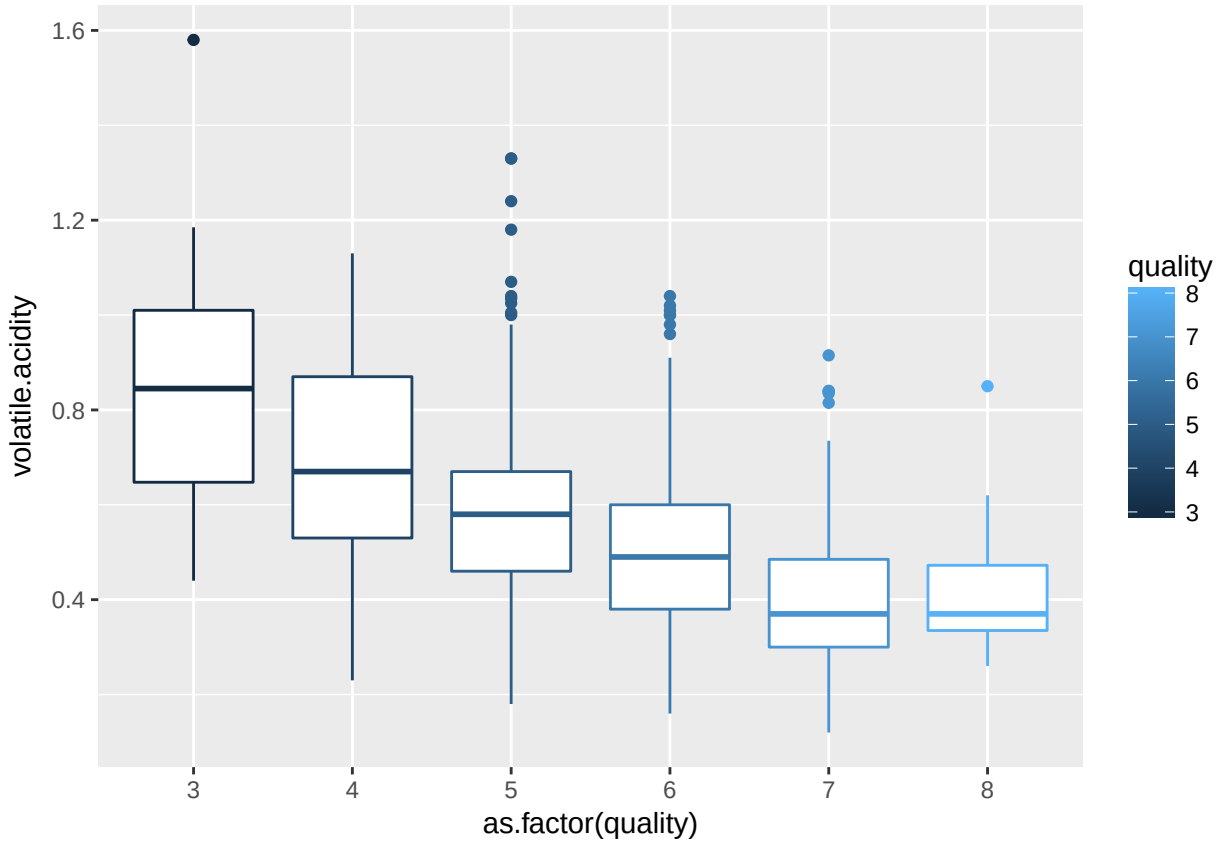
The histogram of quality seems to follow normal distribution.

3. Exploratory Analysis

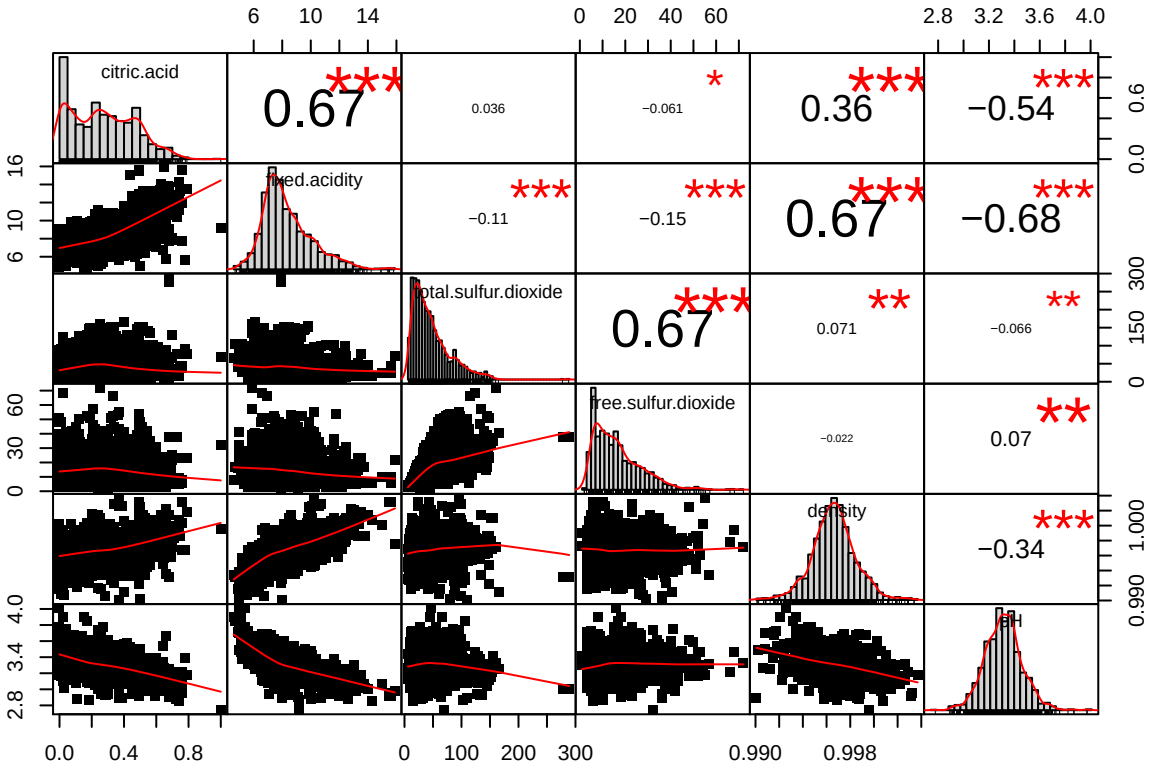
3. Use summary statistics / exploratory data analysis to understand the data better or to identify any obvious patterns. -> exploratory data analysis

Observing correlation matrix, there exists weak relationship between two predictor, alchole 0.47 and volatile.acidity -0.39 for quality. As the correlation was not strong, we can see that the quality of the wine is not dependent on any single covariate.


alcohol
14
12
10
3
4
5
6
7
8
as.factor(quality)
quality
8
7
6
5
4
3


volatile.acidity
1.6
1.2
0.8
0.4
3
4
5
6
7
8
as.factor(quality)
quality
8
7
6
5
4
3

```
##
## The downloaded binary packages are in
##  /var/folders/l7/y2ztr08x2qb06_dtl31rdcz00000gn/T//RtmpKCOoSX/downloaded_packages
## Loading required package: xts
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
##
## Attaching package: 'xts'
## The following objects are masked from 'package:dplyr':
##
##     first, last
##
## Attaching package: 'PerformanceAnalytics'
## The following object is masked from 'package:graphics':
##
##     legend
```



In this stage, there are many correlation between covariates found.

Variables	Correlation coefficient
citric.acid, fixed.acidity	0.67
total.sulfur.dioxide, free.sulfur.dioxide	0.67
density, fixed acidity	0.67
pH, fixed acidity	-0.68

Four pairs of covariates has high correlations. However, since all of the kinds of acidity hugely affects the tastes of the wines, we do not delete any of acid covariates. It is obvious that density affects taste, and pH is affected by the level of acid. so we do not delete density nor pH.

Dioxide is used for presevation of wine. free sulfur dioxide is to prevents microbial growth and the oxidation of wine. Total sulfur dioxide accounts for amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine. As total sulfur dioxide is more affective to the taste, we delete free.sulfur.dioxide from the dataset.

4. Confirmatory data analysis

We conduct Generalized Linear Model and linear regression model. We first conduct Poisson model and see which variables are adequate for the modeling. Then we try other models: logit model and linear regression model.

In the exploratory analysis, we figured out that alchole and volatile.acidity play roles in the taste of wine, but not determinately. Hence we predict that confimatery analysis should contain both, and more variables.

1. Possion Model.

Following the property of response variable, the most adequate model is Poisson model. Considering all variables simultaneously yield insignificant covariates included. Thus stepwise variable selection was done and the smaller model was selected by ANOVA. The smaller model contains only three covariates:volatile.acidity, sulphates, alcohol. Since based on common knowledge, the taste of wine is determined by compositions of many acid and other ingredients. The model is already simple so we stop dropping variables. After that, Dispersion parameter for quasipoisson family taken turns out to be 0.07654819, which is too small. As the assumption for poisson model is severely violated, quasipoisson model is choosed. There happenes to be no outlier after checking cook's distance.

```
##
## Call:
## glm(formula = quality ~ volatile.acidity + sulphates + alcohol,
##     family = quasipoisson, data = wine_subset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.24673  -0.16362  -0.03129   0.19520   0.85793
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.210564   0.034226  35.369  < 2e-16 ***
## volatile.acidity -0.219826   0.017382 -12.647  < 2e-16 ***
## sulphates         0.119394   0.017495   6.824 1.25e-11 ***
## alcohol           0.053026   0.002725  19.462  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 0.07654819)
##
##     Null deviance: 185.16  on 1598  degrees of freedom
## Residual deviance: 123.72  on 1595  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

This result corresponds to the exploratory analysis. The important two variables, alcohol and volatile.acidity are included and there happens to be sulphates added. Sulphates is a wine additive which can contribute to sulfur dioxide gas (S02) levels, wich acts as an antimicrobial and antioxidant. Total sulfur dioxide means the amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine. but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine. It makes sense that sulphates affects taste and quality of wine, as over 50ppm of S02 becomes evident in the nose and taste of wine.

2.Logit model

As seen in the histogram, the quality of the wine takes values between 3 to 8, and seems to follow normal distribution. Logit model was conducted bisecting the qualities of the wines: 0(low quality), 1(high quality). The quality values 3, 4, 5 was regarded as low quality, and the qualities 6, 7, 8 was regard as high quality.

Considering all variables simultaneously yield insignificant covariates included. Thus stepwise variable selection was done and the smaller model was selected by ANOVA. The smaller model contains 7 covariates:fixed.acidity, volatile.acidity, citric.acid, chlorides, total.sulfur.dioxide, sulphates, alcohol. As an attempt to yield simpler model, AIC was compared by drop1 function. Deleting chlorides increase smallest AIC. Hoevever, there exists significant difference between the two models, which was confirmed by ANOVA test. Hence the complex model with chlorides was chosen. After that The dispersion parameter was checked by using quasibinomial and was 1.14. Since there exists no outlier after checking cook's distance. We keep the binomial model with dispersion parameter 1.14.

```
##
## Call:
## glm(formula = quality1 ~ fixed.acidity + volatile.acidity + citric.acid +
```

```
##     chlorides + total.sulfur.dioxide + sulphates + alcohol, family = quasibinomial,
##     data = wine_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3496  -0.8492   0.3191   0.8573   2.3224
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -9.179517   1.015657  -9.038  < 2e-16 ***
## fixed.acidity         0.130762   0.054572   2.396   0.0167 *
## volatile.acidity     -3.588698   0.507502  -7.071 2.29e-12 ***
## citric.acid          -1.516679   0.591458  -2.564   0.0104 *
## chlorides            -3.422795   1.619108  -2.114   0.0347 *
## total.sulfur.dioxide -0.010593   0.002153  -4.921 9.51e-07 ***
## sulphates             2.723256   0.469523   5.800 7.98e-09 ***
## alcohol               0.924676   0.078495  11.780  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 1.145734)
##
##     Null deviance: 2209.0  on 1598  degrees of freedom
## Residual deviance: 1665.2  on 1591  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

3. Another logit model as the reponse variables take integer values from 3 to 8, it is obvious that linear regression model is not appropreate. The previous logit model take the response value 5 to 0, and 6 to 1, which is too extreme bisection, eventhough there is not critical difference. Hence, discarding response values 5 and 6, another logit model was conducted with response values 3,4 to low quality(0), and 7and 8 to high quality(1).

The new subsample has 280 sample size. Considering all variables simultaneously yield insignificant covariates included. Thus stepwise variable selection was done and the smaller model was selected by ANOVA. The smaller model contains 8 covariates: volatile.acidity,citric.acid , residual.sugar, chlorides , density , pH , sulphates , alcohol. As an attempt to yield simpler model, AIC was compared by drop1 function. Deleting citric.acid increase smallest AIC. Hoevever, there exists significant difference between the two models, which was confirmed by ANOVA test. Hence the complex model with citric.acid was chosen. After that The dispersion parameter was checked by using quasibinomial and was 0.59. Since there exists one outlier(14th datum) after checking cook's distance, we deleted that datum and the dispersion parameter was 0.49. Therefore, we use qusibinomial model with 0.49 dispersion parameter.

```
##
## Call:
## glm(formula = quality3 ~ volatile.acidity + citric.acid + residual.sugar +
##     chlorides + density + pH + sulphates + alcohol, family = quasibinomial,
##     data = wine_subset3)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.36721   0.03647   0.11942   0.26731   2.45042
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -442.8152   162.6984  -2.722 0.006917 **
## volatile.acidity   -8.8184     1.6884  -5.223 3.52e-07 ***
## citric.acid        -4.9947     1.7716  -2.819 0.005168 **
## residual.sugar     -0.3557     0.1082  -3.288 0.001142 **
## chlorides          -3.8109     5.8201  -0.655 0.513166
## density           449.3215   163.3534   2.751 0.006351 **
## pH                 -7.2839     1.7577  -4.144 4.57e-05 ***
## sulphates           5.9222     1.5619   3.792 0.000185 ***
## alcohol             2.2262     0.3086   7.213 5.52e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## (Dispersion parameter for quasibinomial family taken to be 0.4923372)
##
##     Null deviance: 295.58  on 278  degrees of freedom
## Residual deviance: 109.44  on 270  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```

5. Interpret the results. -> results/discussions

The three models respresent the same result with the one of exploratory analysis. Everymodel has alcohol, and volatile.acidity for its covariates and they are very significant as they have very small p values. We can say that the conclusions consistent with univariate analysis even there were some correlation problems between some covariates.

6. Conclude. -> conclusion

In this project, we performed generalized linear modeling(glm) to the wine data. The tastes of the wines is well known of their diversity based on the ingredients of the wines. In order to keep pace with the growing of wine produce regions and of the wine market, we concluded that glm can play its role in matching the proper wine products and its adequate market. As the appetite propensity could be analyzed by this method, we can develop the wine market more efficiently less depending on the traditional way.