Wine taste prediction by GLM

Hyeongcheol Park

2017-12-11

Abstract

GLM model can be used to obtain some insights of the taste preference on the traditional wine market. This model can benefit the wine production line and be helpful to match the need and supply the wine goods. Logit models and Poission model are effectve to analyze taste preference of traditional wine market

1. Introduction

The ubiquitos era is blurring out physical barriers. 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 are 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 the environment of a region, the diversity of wines grow rapidly.

In this situation, traditional roles 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 the task. By analyzing physicochemical ingredients of the wine that people in a region prefer the best, we can distribute wine products in very effecient way. Moreover, the analysis can be used to enhance the general quality of wine. In order to do that, we conduct generalized linear modeling(GLM) to investigate the best 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 the wine produced in a 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 the consumption in a easier way.

2. Data Discription

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 of wine 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.

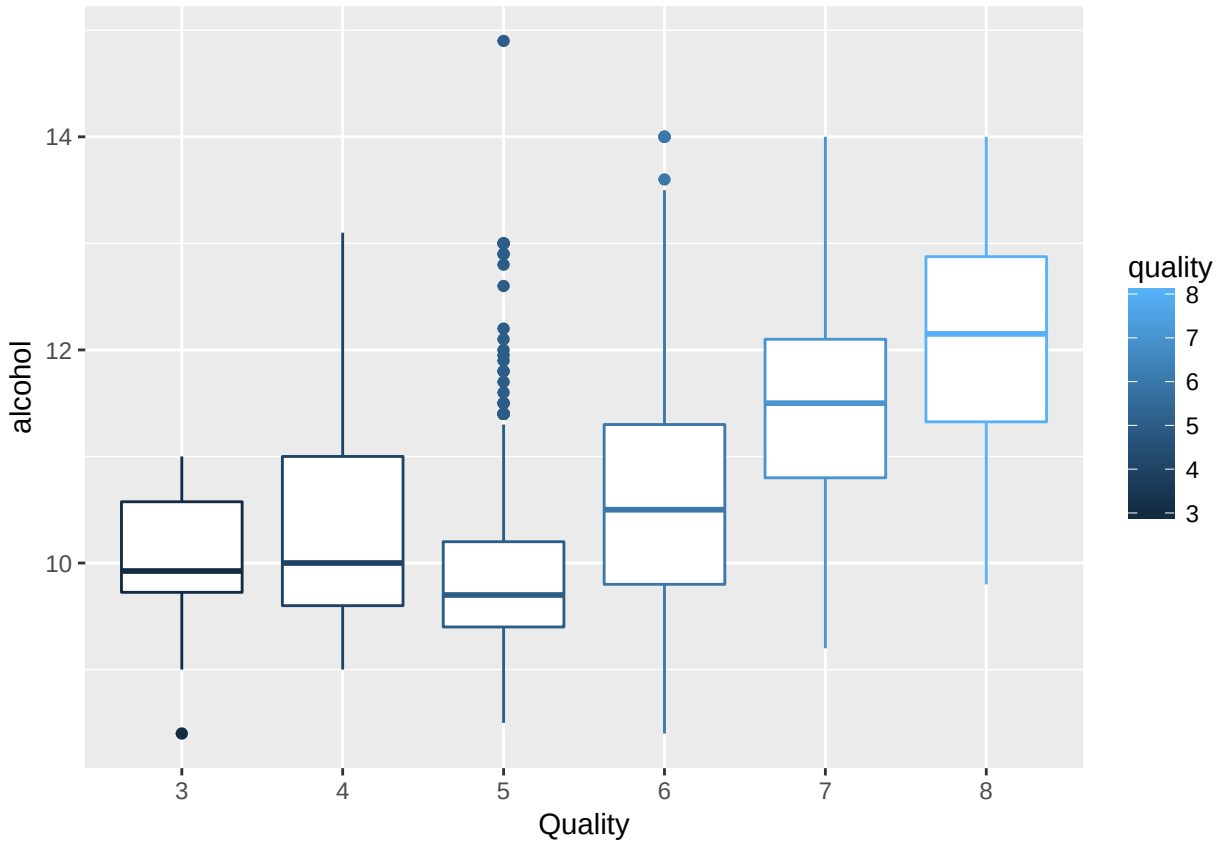
There are 1599 samples.

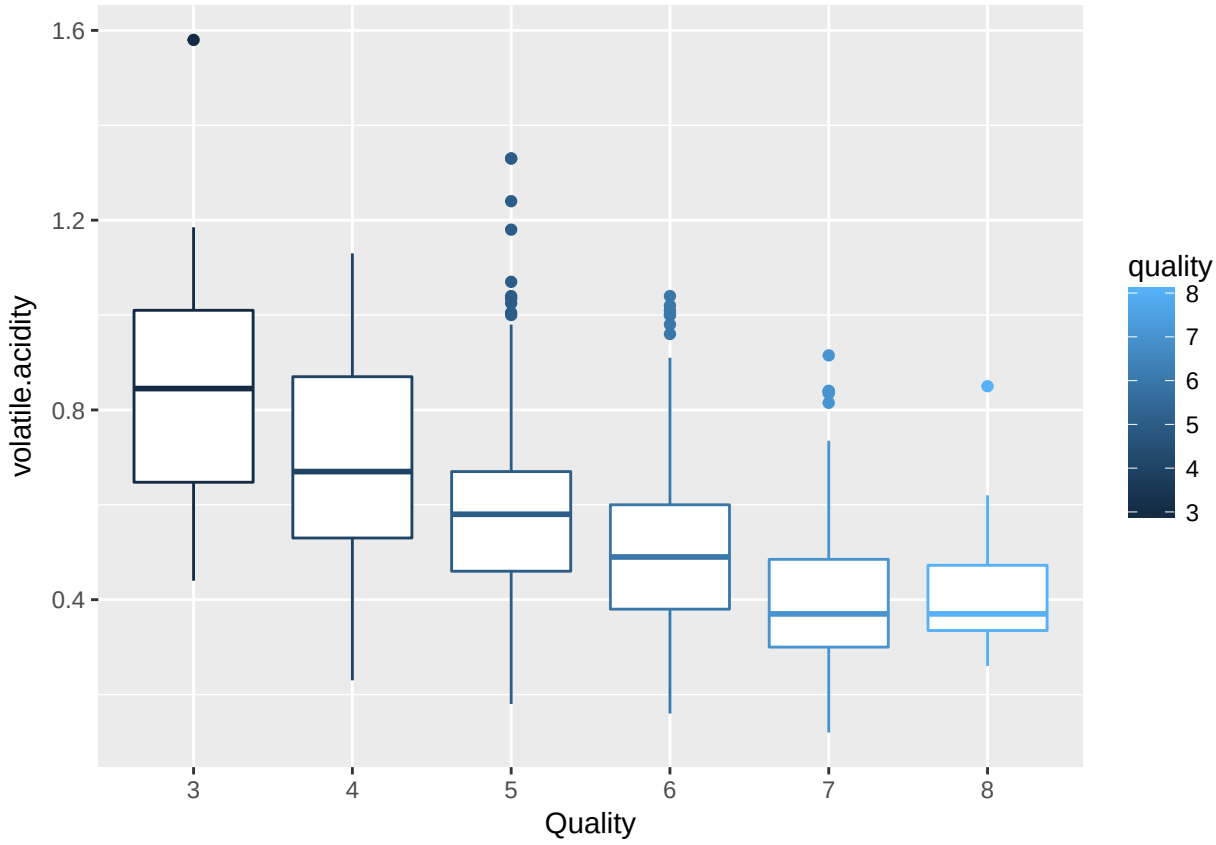
Attribute (units) Red wine	Min	Max	Mean	Median
Fixed acidity (g(tartaric acid)/dm3)	4.6	15.9	8.3	7.9
Volatile acidity (g(acetic acid)/dm3)	0.1	1.6	0.5	0.52
Citric acid (g/dm3)	0.0	1.0	0.3	0.26
Residual sugar (g/dm3)	0.9	15.5	2.5	2.2
Chlorides (g(sodium chloride)/dm3)	0.01	0.61	0.08	0.079
Free sulfur 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

Observing correlation matrix, there seems to exist weak relationships between predictors and response variable; The quality has relatively high correlation coefficient with alchole (0.47) and volatile.acidity (-0.39). Since the correlations are not strong, we can see that the quality of the wine is not heavily dependent on any single covariate.


alcohol
14
12
10
3
4
5
6
7
8
Quality
quality
8
7
6
5
4
3


volatile.acidity
1.6
1.2
0.8
0.4
3
4
5
6
7
8
Quality
quality
8
7
6
5
4
3

<figure1: boxplots between response and two covariates >

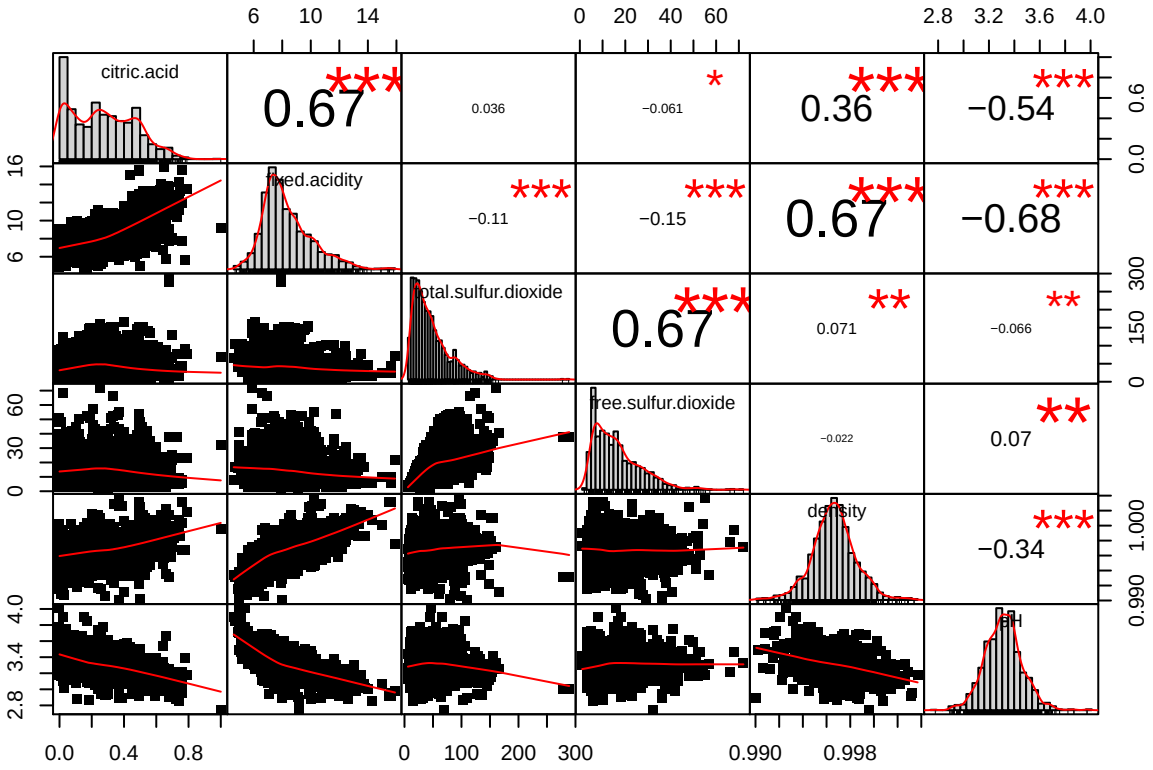
It seems that more alcohol yields better quality, and less volatile.acidity results in better quality of wine.

In an attempt to find any correlation between covariates which might cause multicollinearity, I checked correlation matrix and scatter plot. There are likely to be some correlations between a few random variables: citric.acid, fixed.acidity,total.sulfur.dioxide, free.sulfur.dioxide,density,and pH.

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 neither.

```
##
## The downloaded binary packages are in
##  /var/folders/l7/y2ztr08x2qb06_dtl31rdcz00000gn/T//Rtmpp0PNbp/downloaded_packages
```



<figure2: abbreviated correlation matrix and scatter plot.>

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, but keep total sulfur dioxide.

4. Confirmatory data analysis

We will conduct Generalized Linear Models: poisson model and binomial models. We first conduct Poisson model and see which variables are adequate for the modeling. In the previous 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 yielded 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 acids and other ingredients, the model is already too simple so we stop dropping variables. After that, the dispersion parameter for quasipoisson model turns out to be 0.07654819, which is too small. As the assumption for poisson model is severely violated, we rather choose quasipoisson model. There happenes to be no outlier after checking cook's distance.

Coefficients	Estimate	Std. Error	t value	P value
(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

This result corresponds to the exploratory analysis. The important two variables, alcohol and volatile.acidity are included and the p values are quite small. Meanwhile, there happens to be sulphates added. Sulphates is a wine additive that can contribute to sulfur dioxide gas (S02) levels, which 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. Hence, 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 is likely 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 the rsult of 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.

Coefficients	Estimate	Std. Error	t value	P value
(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)

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 takes the response value 5 to 0(low quality), and 6 to 1(high quality), which is too extreme bisection. This is because there is no critical difference between 5 and 6. 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 sub-sample 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.

Coefficients	Estimate	Std. Error	t value	P value
(Intercept)	-442.8152	162.6984	-2.722	0.006917 **

Coefficients	Estimate	Std. Error	t value	P value
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)

After correction influential point problem, the covariate chlorides seems to be insignificant. However we keep the predictor since it is logically resonable to keep it as we confirmed in the poisson model.

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.

Comparing two logit models, it seems that the smaller model with sample size 279, which discarded some samples could be better model. This is because quality 5 and 6 has no huge difference, and the sample size 279 seems to be still big enough.

6. 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.

There are some limitations about this analysis. The origial paper, by Cortez et al.(2009) mentioned in his paper that support vector machine was a better model than GLM model. The GLM model can conduct its role quite well, but it might not be the best model to do our task. Secondly, the model's has some hidden problem. When we look at the box plot between alcohol and quality, there might be an quadratic relation, not a linear one. That may cause some lack of inference. Moreover, in the other markets, people may like different tastes, which depend on different ingredients of the wine. That means that the model is not valid for predicting other market's preferences.

7.Appendix

```
# #read dataset.
# wine <- read.csv(file="/Users/hyeongcheolpark/Library/Mobile Documents/com~apple~CloudDocs/UBC/2017-f
#
# # Understand data. Describe how the data is collected and what the data is.
# head(wine)
# str(wine)
# summary(wine)
# hist(wine$quality)
# wine <- na.omit(wine)
#
# # 3. Use summary statistics / exploratory data analysis to understand the data better or to identify
# #any obvious patterns. -> exploratory data analysis
# cor(wine)
# pairs(wine$quality)
# wine_subset<-subset(wine, select = c(-free.sulfur.dioxide))
# install.packages("PerformanceAnalytics")
# library("PerformanceAnalytics")
# my_data <- subset(wine, select=c(citric.acid, fixed.acidity,
# total.sulfur.dioxide, free.sulfur.dioxide,density, pH))
# chart.Correlation(my_data, histogram=TRUE, pch=15)
# #alchole 0.47 volatile.acidity -0.39 for quality
# #citric.acid, fixed.acidity0.67,
```

```
# # total.sulfur.dioxide, free.sulfur.dioxide  0.67,
# # density, fixed acidity 0.67,
# # pH fixed acidity -0.68
#
#
# #  Choose an appropriate model and fit the model...  -> confirmatory data analysis
#
# #Poission model
# #Full model-> many insignificant covariates exist.
# fit_poi<-glm(quality~.,family=poisson,data=wine_subset)
# summary(fit_poi)
#
# # stepwise method for variable selection.
# fit_poi1 <- step(fit_poi, direction="both")
# summary(fit_poi1)
#
# anova(fit_poi1,fit_poi)
# qchisq(0.95,7) # There is no difference between two model. So we choose smaller model.
#
# drop1(fit_poi1)
# # Based on common knowledge, the taste of wine is determined by compositions of many acid and other
# #ingredients. The model is too small already so we stop dropping variables.
#
#
# #sulphates: a wine additive which can contribute to sulfur dioxide gas (SO2) levels, wich acts as an
# #it makes sense that sulphates affects taste and quality of wine, as over 50ppm of SO2 becomes eviden
# #total sulfur dioxide: amount of free and bound forms of SO2; in low concentrations, SO2 is mostly un
# #but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
#
# fit_poi_quasi <- glm(quality ~ volatile.acidity + sulphates + alcohol,
#     family = quasipoisson, data = wine_subset)
# summary(fit_poi_quasi)
```

```
# cooks.distance(fit_poi_quasi)
# which(cooks.distance(fit_poi_quasi)>1)
#
# # Dispersion parameter for quasipoisson family taken to be 0.07654819
# # It means that the model seriously violate the assumptions. Hence we use quasi poisson model with di
#
#
#
#
# #binomial model
#
# attach(wine_subset)
# hist(wine_subset$quality)
# summary(wine$quality)
# quality1<-as.numeric(quality>5.5)
# str(quality1)
# summary(quality1)
# hist(quality1)
# count(quality1)
# fit_bin<-glm(quality1~.-quality,family=binomial,data=wine_subset)
# summary(fit_bin)
#
# # stepwise. But still
# fit2_bin <- step(fit_bin, direction="both")
# summary(fit2_bin)
#
# anova(fit2_bin,fit_bin)
# qchisq(0.95,3)# we choose smaller model.
#
# #Comparing AIC by drop1
# drop1(fit2_bin, test = "Chisq")
#
```

```
# #deleting chlorides increase smallest AIC.
# fit3_bin<-glm(quality1 ~ fixed.acidity + volatile.acidity + citric.acid +
#                 total.sulfur.dioxide + sulphates + alcohol,family=binomial,data=wine_subset)
# summary(fit3_bin)
#
# anova(fit3_bin,fit2_bin)
# qchisq(0.95,1)# 5.32>3.841 It means there is difference between simpler model and complex model.
# #so we choose complex model.
#
# #we check the dispersion parameter, By using quasibinomial.
# fit4_bin_quasi<-glm(quality1 ~ fixed.acidity + volatile.acidity + citric.acid +
#                       chlorides + total.sulfur.dioxide + sulphates + alcohol,
#                     family=quasibinomial,data=wine_subset)
# summary(fit4_bin_quasi)#As quasibinimial's dispersion parameter is 1.14,
# #the assumptions for logit model hold well. We use logit model. But we need to check outlier, and Inf
# cooks.distance(fit4_bin_quasi)
# which(cooks.distance(fit4_bin_quasi)>1)
# # There is no outlier in this data. So we keep the binomial model, with dispersion parameter 1.14
#
#
# # Another binomial model
# length(quality)
# quality2 <- c()
# for (i in 1:length(quality))
# {
#   if (quality[i]>6.5)
#   {quality2 <- c(quality2,1)}
#   if (quality[i]<4.5)
#   {quality2 <- c(quality2,0)}
#   if (quality[i]<=6.5&quality[i]>=5)
#   {quality2 <- c(quality2,NA)}
# }
```

```
# wine_subset2 <- wine_subset[-which(is.na(quality2)),]
# str(wine_subset2)
# quality2<- na.omit(quality2)
# str(quality2)
# summary(quality2)
# hist(quality2)
# fit_bin_sub<-glm(quality2~.-quality,family=binomial,data=wine_subset2)
# summary(fit_bin_sub)
#
# fit2_bin_sub <- step(fit_bin_sub, direction="both")
# summary(fit2_bin_sub)
#
# drop1(fit2_bin_sub)
#
# fit3_bin_sub <- glm(quality2 ~ volatile.acidity  + residual.sugar +
#                       chlorides + density + pH + sulphates + alcohol,family=binomial,
#                     data=wine_subset2)
# anova(fit3_bin_sub,fit2_bin_sub)
# qchisq(0.95,1)# There is difference, so we choose complex model.
#
#
# #Quasi!
# fit2_bin_sub_quasi <- glm(quality2 ~ volatile.acidity + citric.acid + residual.sugar +
#                             chlorides + density + pH + sulphates + alcohol, family = quasibinomial,
#                           data = wine_subset2)
# summary(fit2_bin_sub_quasi)
#
#
#
# cooks.distance(fit2_bin_sub_quasi)
# which(cooks.distance(fit2_bin_sub_quasi)>1)
#
```

```
# quality3 <- quality2[-14]
# str(quality3)
# wine_subset3 <- wine_subset2[-14,]
# str(wine_subset3)
#
# fit3_bin_sub_quasi <- glm(quality3 ~ volatile.acidity + citric.acid + residual.sugar +
#                             chlorides + density + pH + sulphates + alcohol, family = quasibinomial,
#                           data = wine_subset3)
# summary(fit3_bin_sub_quasi)
#
#
#
# cooks.distance(fit3_bin_sub_quasi)
# which(cooks.distance(fit3_bin_sub_quasi)>1)
#
```