STAT 462 Final Project

Regression Analaysis for Wine Data

Samuel Fox, Jiaying Liang, Luxin Wang

Statistics Department
Penn State University
December 8, 2018

Abstract

The goal of this research project was to develop the best model to predict the quality of wine. We gathered a large dataset of wine data and took 1000 samples. We first checked assumptions of the dataset to make sure what we had was usable and would result in a working model based off of strong data. After checking for outliers, we did not detect any outlier in the dataset. We deleted density as it is collinear according to VIF output. From there we developed three models. We used backward elimination, RSS, Mallow's Cp, AIC and BIC methods to select models. For our final linear regression model and the final logistic model, we ended up with the similar predictors including volatile acidity, residual sugar, chlorides, free sulfur dioxide, sulphates, alcohol and type. The final ordinal regression model has 4 predictors including volatile acidity, residual sugar, sulphates and alcohol. In addition, we perform an ANOVA test Chi-square test to test whether different types of wine have different quality. The result shows us red wine has significantly higher quality than white wine.

Introduction

The dataset that we used for analysis contained 1000 data points of wine sample from a larger dataset gathered from UCI's wine+quality dataset. The variables considered were: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, ph, sulphate, alcohol, and type with the predicted variable being quality. Questions we asked when considering this dataset included are there any issues such as collinearity or outliers, does the data need to be scaled, what would be the best type of model for this scenario, and which interactions between variables, if any, are important in predicting the wine's quality.
The research objective is to find out the relationship between wine quality and the wine chemical contents. We are interested in the wine quality because it can be used for marketing purposes as well as helping the wine producers to decide how to improve wine quality in the future by determining which aspects to focus on.

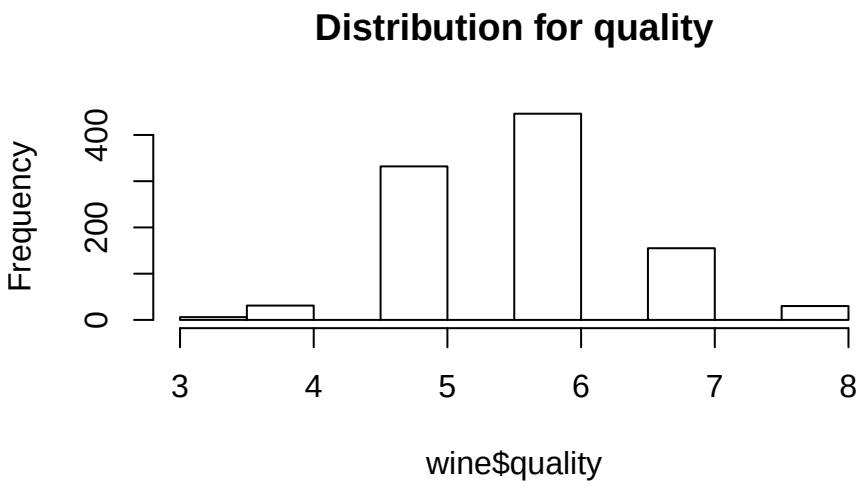
Exploratory Data Analysis

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. We obtained the dataset from: http://archive.ics.uci.edu/ml/datasets/Wine+Quality . The website also provided the following references.
1. Paulo Cortez, University of Minho, Guimar?es, Portugal, http://www3.dsi.uminho.pt/pcortez 2. A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal,2009. We randomly select 1000 samples of the dataset. It contains the following variables (units are not given in the dataset description):

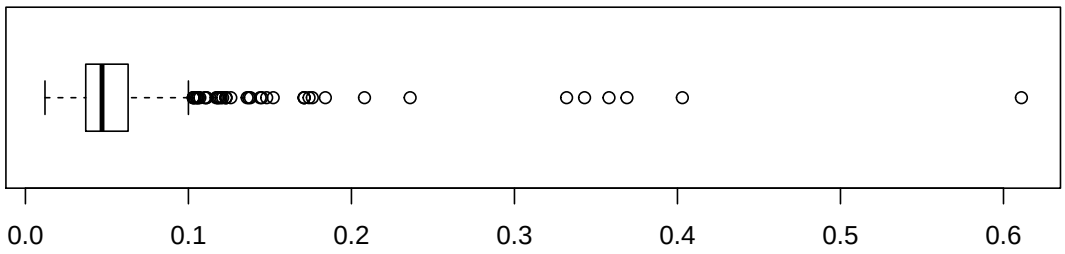
- fixed acidity (grams/liter)
- volatile acidity (grams/liter)
- citric acid (grams/liter)
- residual sugar (grams/liter)
- chlorides (grams/liter)
- free sulfur dioxide (milligrams/liter)
- total sulfur dioxide (milligrams/liter)
- density (grams/cubic centimeter)
- pH: acidity (below 7) or alkalinity (over 7)
- sulphates:potassium sulfate (grams/liter)
- alcohol:percentage alcohol (% volume)
- type: type of wine (red/white)
- Output variable (based on sensory data): quality (score between 0 and 10)

In this report, we are interested in what impact the different qulity of red wine and white wine. We build two different models, treating quality as continuous variable and categorical variable regardingly. Here are some simple summary output of this dataset:

Consider quality as a response, these is the distribution for quality.



After ploting boxplots for all continuous variables. Density is the only variable does not have outliers and roughly symmetric (due to the random sample; few outliers exist when considering the whole dataset). Extreme outliers can be notieced within the chlorides variables (plot shown below). Most variables are skewed to the right with with the outliers on the larger side.



Since the dataset is really big, looking at the pairwise sacatterplots will be relatively hard to identify outliers, we calculate the leverage of the potential predictors. Still considering the quality as response, we calculate the leverange of the potential model.

There are the points with leverage greater then the threshhold 3*p/n, notice observation #745 is the one point with extreme hight chloride content:

```
## 145 153 204 220 224 269 279 298 321 388 402 597 608 655 658 669 696 699
## 145 153 204 220 224 269 279 298 321 388 402 597 608 655 658 669 696 699
## 745 800 856 881 902 932 989
## 745 800 856 881 902 932 989
```

Method

1. Check Collinearity and Re-Scaling X's

From the VIF table we can see density has VIF of 25.061323 which mean severe collinearity. So we have to remove this predictor, and this is the new VIF result.

	vif1
fixed.acidity	5.04
volatile.acidity	2.12
citric.acid	1.59
residual.sugar	10.70
chlorides	1.63
free.sulfur.dioxide	2.38
total.sulfur.dioxide	3.97
density	25.06
pH	2.73
sulphates	1.61
alcohol	6.61
type	7.63

(a) VIF Full Predictor

	vif2
fixed.acidity	2.23
volatile.acidity	2.09
citric.acid	1.59
residual.sugar	1.52
chlorides	1.63
free.sulfur.dioxide	2.37
total.sulfur.dioxide	3.89
pH	1.71
sulphates	1.50
alcohol	1.41
type	4.89

(b) VIF Remove Density

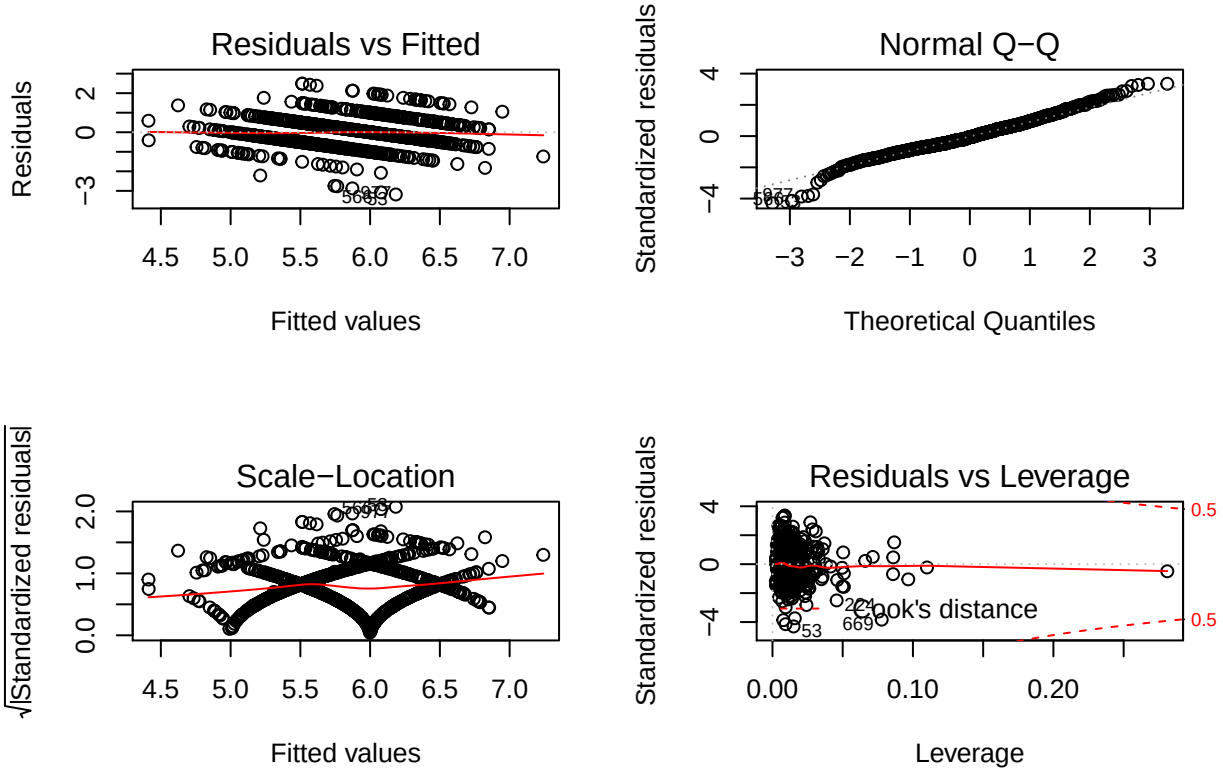
Table 1: VIF table

Looking at the new VIF result, no predictor variabel has VIF largely greater than 4. Thus the collinearity problem is eliminated. We also look in to the scale of X's. Since the orginal full linear model shows there is about 100 times difference between two β, we decided to scale the X's using the method "scale". In the following analysis, all continuous predictors are scaled.

2. Regression Model treating Quality as continuous Variable and Model Selection

2.1 Full model

First we create dummy variable foe type variable, for which type red is 1, and type white is 0. Then we build a full model using quality as response variable. Here shows the full model summary and diagnosit plot



The residuals vs fitted looks like this because quality is a cateforical variable with 6 levels. However, since it has so many level, we treat it as a continous variable. Hence, the residuals vs fitted plot breaks into 6 lines. Accorrding to the Q-Q plot, the residuals follows a normal distribution. According to the residuals vs leverage plot, there is no influential point because there is no Cook's distance exceed 0.5. Independent assumption is also met according to residual vs. observation plot which is not shown above (further discussion of check influencial points is in the last section).

2.2 Select model

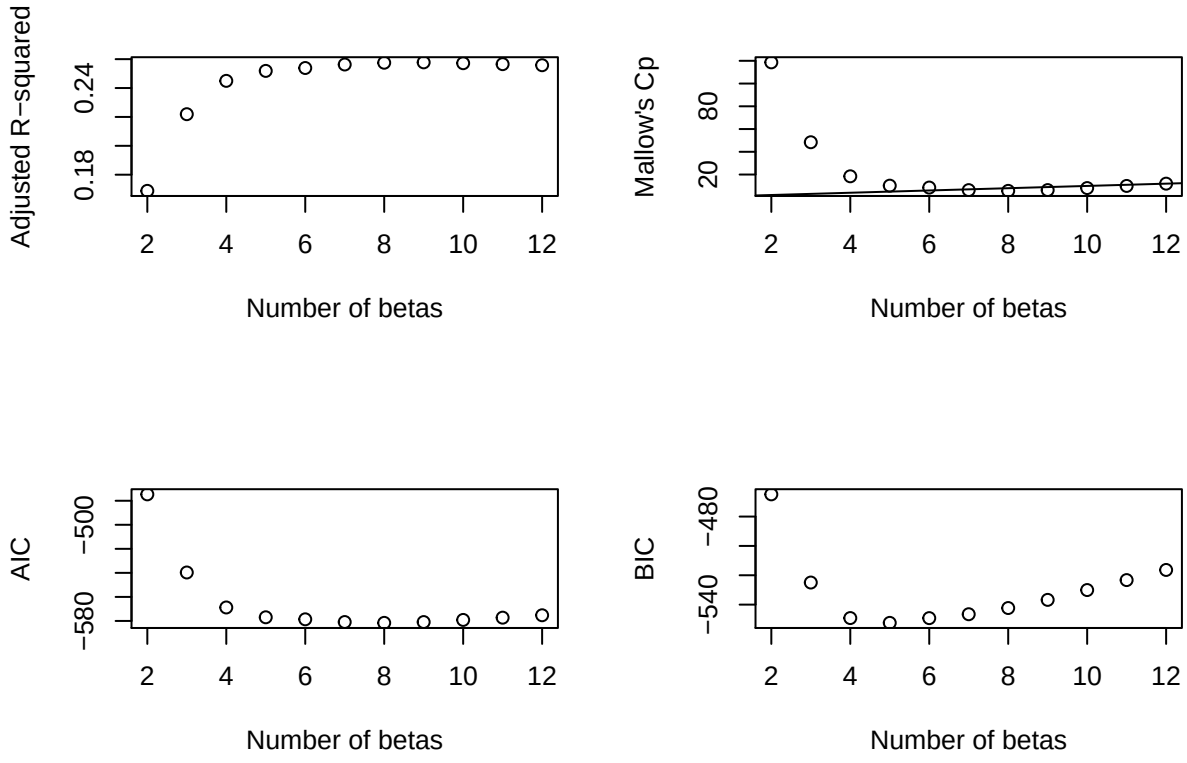
First, we use backward selection with α=0.5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.7505	0.0321	179.21	0.0000
scaled.wine$volatile.acidity	-0.2506	0.0313	-8.01	0.0000
scaled.wine$residual.sugar	0.0972	0.0276	3.53	0.0004
scaled.wine$chlorides	-0.0598	0.0291	-2.06	0.0399
scaled.wine$sulphates	0.1388	0.0283	4.91	0.0000
scaled.wine$alcohol	0.3829	0.0265	14.42	0.0000
dummy	0.2159	0.0896	2.41	0.0162

Table 2: Backward Selected Linear Model

We end up with 6 predictors. The reduced model includes volatile.acidity, residual.sugar, chlorides, sulphates , alcohol and dummy variables.

Then, we try to use R^2_{adj} to select model:



Max adjusted R^2 with 8 predictors. The model include fixed.acidity, volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxide, sulphates, alcohol and dummy variables.

Using C_p: when p=7 , with 6 predictors. The model include volatile.acidity, residual.sugar, chlorides, sulphates, alcohol and dummy variables.

Using AIC: when p=8,with 7 predictors. The model includes volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxide, sulphates, alcohol and dummy variables.

Using BIC: when p= 5, 4 predictors. The model includes volatile.acidity, residual.sugar, sulphates, alcohol and dummy variables.

In conclusion, we should use the model with 7 predictors bacause it has relatively big R^2_{adj}, low C_p, low aic and low bic. The final models includes volatile.acidity, residual.sugar, chlorides, free.sulfur.dioxides, sulphates, alcohol and dummy variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.7409	0.0326	176.07	0.0000
scaled.wine$volatile.acidity	-0.2460	0.0314	-7.83	0.0000
scaled.wine$residual.sugar	0.0871	0.0282	3.09	0.0021
scaled.wine$chlorides	-0.0586	0.0290	-2.02	0.0439
scaled.wine$free.sulfur.dioxide	0.0466	0.0285	1.63	0.1030
scaled.wine$sulphates	0.1364	0.0283	4.82	0.0000
scaled.wine$alcohol	0.3882	0.0267	14.53	0.0000
dummy	0.2557	0.0928	2.76	0.0060

Table 3: Final Reduced Linear Model

According to the summary output, the final model is
y = 5.741 -0.246$X_{volatile.acidity}$ + 0.087$X_{residual.sugar}$ -0.059$X_{chlorides}$ +0.047$X_{free.sulfur.dioxide}$ + 0.136$X_{sulphates}$ + 0.388$X_{alcohol}$ + 0.256d_{type}
However, R^2 for this model is 0.2627 which means this model only represent 26.27% of the quality response. Hence, we need to use other method to find a better model.

3. Logistic Model treating Quality as Categorical Variable and Model Selection

3.1 Prepare Response Variable for Logistic Regression

To successfully perform the logistic regression analysis, we divide the quality variable into two groups, quality 3-5 is marked as low quality, which is assigned as value 0; quality 6-8 is marked as high quality, which is assigned as value 1. Using the function glm with family parameter of binomial. We get a full model of logistic regression.

3.2 Model Selection

Observing the full logsitc regression summary, there are some variable not significantly contribute to the model. Using the function bestglm from the package "bestglm", the best logistic model is selected according to AIC. Then a presudo R^2 is calculated using deviance and null deviance from the model summary. The result is 0.1948436. This is a relatively low R^2, regardless this is not a "true" R^2

However, this logistic regression ignore the fact that quality has 6 different levels. So a better model need to abtained to fully explain the different levels of quality

4. Ordinal Logistic Regression

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.1477	0.2835	4.05	0.0001
volatile.acidity	-0.6499	0.1081	-6.01	0.0000
residual.sugar	0.2046	0.0892	2.29	0.0218
chlorides	-0.1630	0.0870	-1.87	0.0610
free.sulfur.dioxide	0.2226	0.1159	1.92	0.0547
total.sulfur.dioxide	-0.1168	0.1469	-0.80	0.4266
sulphates	0.3920	0.0987	3.97	0.0001
alcohol	1.1093	0.1062	10.45	0.0000
typewhite	-0.5616	0.3596	-1.56	0.1184

Table 4: Best Logistic Model Output

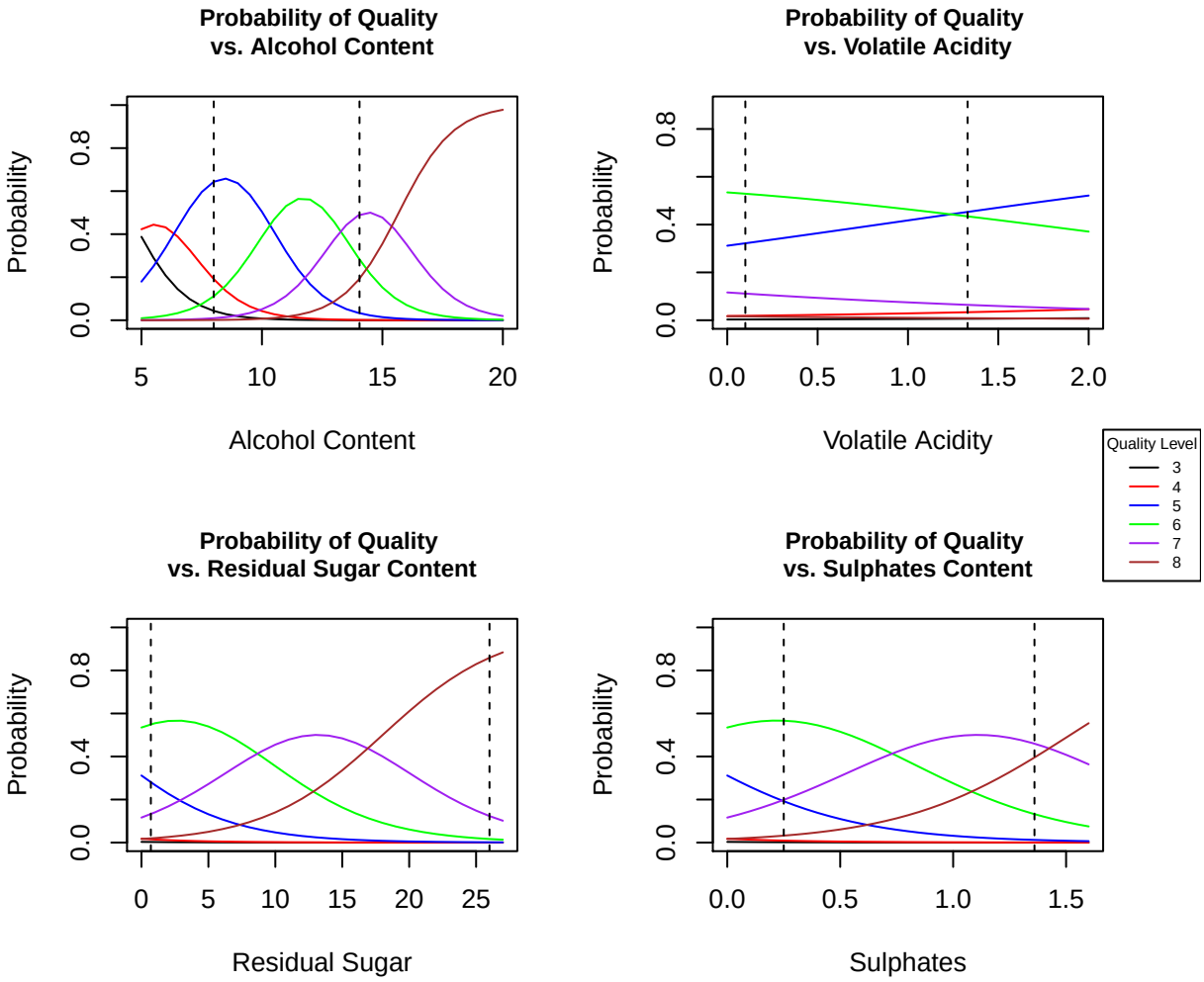
4.1 Build the Ordinal Logistic Regression Model

Even though logistic regression explain the relationship between significant predictors and quality as two level categorical variable, it does not fully explain the true nature of quality as a 6 level categorical variable. Thus, an ordinal logistic regression model is build to fully take consideration of 6 levels of quality.

Using the function polr from package MASS, treating quality as an ordered factor (with ordered level: 3<4<5<6<7<8), the full ordinal logistic model is constructed. Later, the p-value for each individual t-test is performed. Using backward selection, only variable volatile.acidity, residual.sugar, sulphates, and alcohol are left at α=0.05. The best model summary is shown below (a new page called "ordinal" is used, since the clm function has a better ouput than polr function) as well as the plot showing change in probability for each level according to the change in different predictors:

Table 5: Ordinal Logistic Regression Output

volatile.acidity	-3.419^{***}
	(0.419)
residual.sugar	0.047^{***}
	(0.014)
sulphates	2.684^{***}
	(0.461)
alcohol	0.879^{***}
	(0.062)
Log Likelihood	$-1{,}100.949$
Note:	*p<0.1; **p<0.05; ***p<0.01



The graphs above shows the change in probability for each quality level according to change of each significant predictor. And the range between dash lines are the range included by the sample, while range outside the dash lines are prediction. For example, according to the model, when the alcohol content rises up to 15%, the probability of that wine having the quality level of 8 is higher than any other quality.

4.2 ANOVA Analysis and Chi-Square test for Variable Type

During the model selection for ordinal logsitic selection, the dummy variable type caught our attention. It is the last variable to be removed from the backward selection with p-value of 0.0579 which is slightly above 0.05. Will this categorical variable of wine type actually impact the wine quality. A seperated anova analysis is performed along with a histogram overlaying the distribution of quality for red wine and white wine. The ANOVA table shows that the mean quality is significantly different from red wine and white wine.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	1	7.10	7.10	9.58	0.0020
Residuals	998	739.10	0.74		

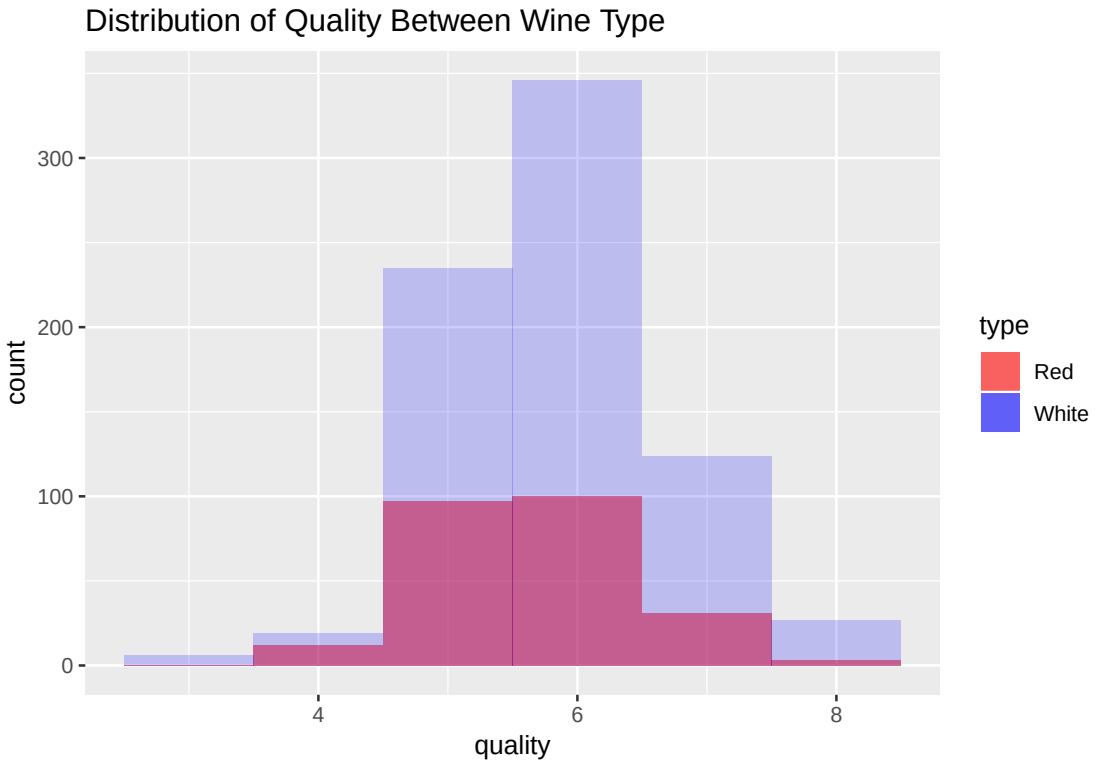
Table 6: ANOVA table

Otherwise, if we treate quality as categorical variable, given the frequency table of different quality levels for red wine and white wine,then perform Chi-square test for homogenity, the p-value is 0.008298, which is smaller than 0.05. Thus we reject the H_0 and conclude that red wine and white wine have different quality distribution.

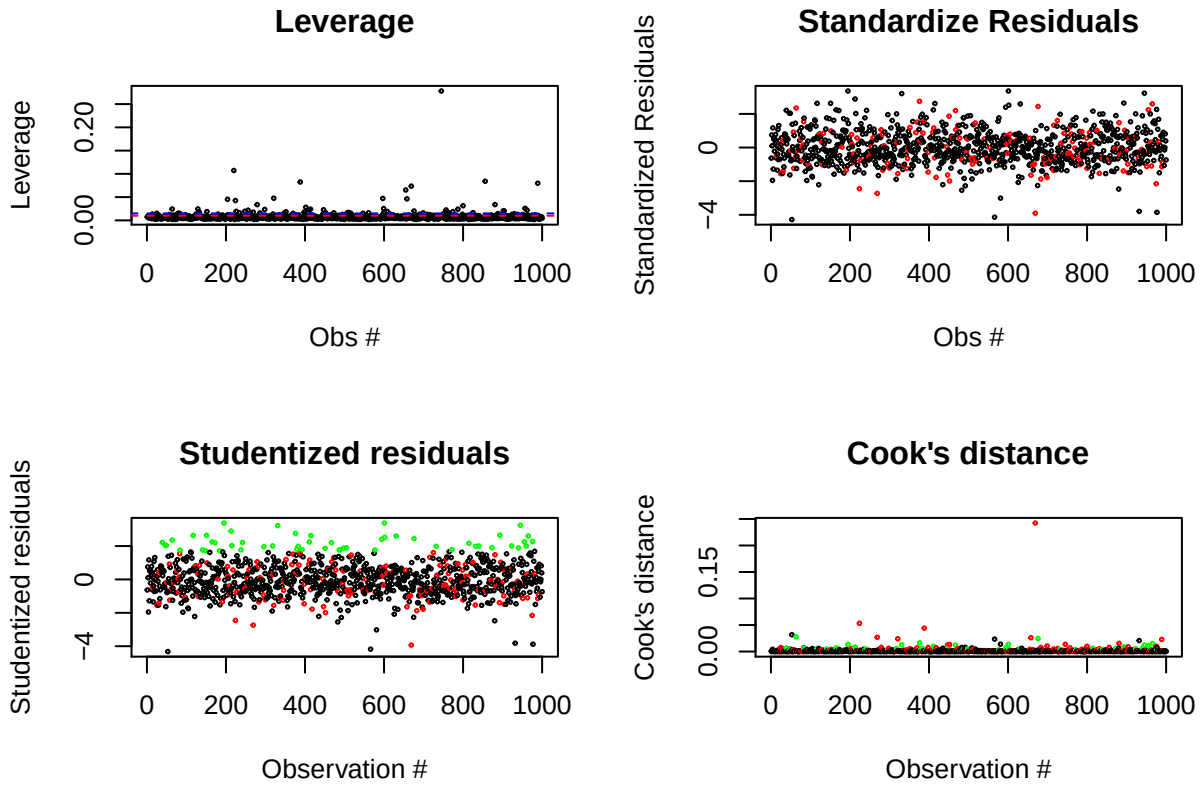
	3	4	5	6	7	8
red	0	12	97	100	31	3
white	6	19	235	346	124	27

Table 7: Frequncy Table

For the ggplot downbelow, the different distribution of quality between red wine and white wine is shown. White wine has a more left skewed trend, showing white wine has more sample in the higher quality range.



5. Check Potential Outlier



Accoding to these plot, there are a lot of observations that have very high leverage (red points in leverage plot are observations with leverage greater than threshold 2p/n) which is most likely due to the fact that a logistic model would better represent them. But no points with Cook's distance greater than 0.5, so no influencial point is detected.

Result

We decided to use an ordinal regression model as our final model. The final fitted model includes volatile.acidity, residual.sugar, sulphates and alcohol.

From this model we can see that volatile.acidity has a negative relationship with quality, and residual.sugar, sulphates, and alcohol have a positive relationship when the other predictors are considered as 0.

Looking at the ANOVA test and Chi-square test, it is apparent that the quality between red and white wine differs significantly.

Conclusion

We decided to use this model because an ordinal regression model has its responses as different ranks which can better represent the quality since the quality contains different levels. The chemical contents that determine the quality of wine are volatile acidity, residual sugar, sulphates and alcohol. We also acknowledged that better quality wine always contains higher alcohol content. In addition, we found out from the ANOVA test that white wine has better quality than red wine does on average.

In this analysis, we need further investment in producing a better linear regression model considering interaction between the predictors. Overall, we think we need more information of wine in order to determine its quality, such as the quality of grapes, the year of its production and the method of wine production, etc.

Team Member Contribution

Sam worked on checking for outliers, leverages, and influential points as well as scaling the data.

Jianying worked on the linear model and checking the linear model assumptions

Luxin worked on the logistic models and their assumptions as well as setting up the final report.

All members contributed to the presentations slides and the final report edits.

Reference

A.I. McLeod and Changjiang Xu (2018). bestglm: Best Subset GLM and Regression Utilities. R package version 0.37. https://CRAN.R-project.org/package=bestglm

Christensen, R. H. B. (2018). ordinal - Regression Models for Ordinal Data. R package version 2018.8-25. http://www.cran.r-project.org/package=ordinal/.

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

John Fox and Sanford Weisberg (2011). An {R} Companion to Applied Regression, Second Edition. Thousand Oaks CA: Sage. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Malshe, Ashwin. "Ordinal Regression in R." Step by Step Solutions: T-Tests: Paired/Dependent and Independent, 20 Oct. 2016,
rstudio-pubs-static.s3.amazonaws.com/220675_90da5cd7a01c4a57b9f22ff2b89bc915.html.

"R Help 15: Logistic, Poisson & Nonlinear Regression." 1.5 - The Coefficient of Determination, r-Squared | STAT 501, Penn State University , onlinecourses.science.psu.edu/stat501/node/433/.

Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0