

Modeling Wine Quality

BANA 7042



June 22, 2021

Ankush Morey

University Of Cincinnati

Contents

[Part A: Background and Data Exploration 2](#_Toc75295146)

[Introduction 2](#_Toc75295147)

[Summary Statistics 2](#_Toc75295148)

[Part B. Visualization and initial models for a binary response 4](#_Toc75295149)

[Distribution of Variables 4](#_Toc75295150)

[Box plots of variables 5](#_Toc75295151)

[Proportion of excellent Red wine: 5](#_Toc75295152)

[Correlation 6](#_Toc75295153)

[Linear Model 6](#_Toc75295154)

[Logistic Regression 7](#_Toc75295155)

[Part C. Variable selection, Interpretation, and Prediction for a Logistic model 8](#_Toc75295156)

[Chi Square Test 9](#_Toc75295157)

[Drop1 Chi Square Test: 10](#_Toc75295158)

[Collinearity 10](#_Toc75295159)

[Influence of Alcohol, Residual.Sugar and pH 11](#_Toc75295160)

[Confusion Matrix, Specificity and Sensitivity 12](#_Toc75295161)

[Receiver Operating Characteristics (ROC) 12](#_Toc75295162)

[Prediction 13](#_Toc75295163)

[Part D. Link functions and Dispersion parameter 13](#_Toc75295164)

[Probit Link Function 13](#_Toc75295165)

[Complimentary Log Log link function 14](#_Toc75295166)

[Comparison of Logit, Probit and Complimentary Log Log link functions (1 page report) 15](#_Toc75295167)

[Dispersion 16](#_Toc75295168)

[Part E. Modeling the wine quality as a multinomial variable with order 17](#_Toc75295169)

[Kendall Tau Correlation 18](#_Toc75295170)

[Multinomial Model 18](#_Toc75295171)

[Interpretation from Variables 19](#_Toc75295172)

[Logistic and Multinomial Model Comparison 20](#_Toc75295173)

[Prediction using Multinomial model 20](#_Toc75295174)

# Part A: Background and Data Exploration

## Introduction

**Description**: Wine is a traditional alcoholic drink liked and consumed all over the globe. Typically, there are 2 types of wine, red wine and white wine. For our case study we will focus on red wine. There is huge variation in the taste of red wine based on how it is produced, its geographic location, the raw material etc. Our red wine dataset contains 1599 samples of wine with its 11 attributes and 1 quality variable.

**Objective**: We try to establish a relationship between the attributes of wine and the quality. Our aim is to analyze how each variable impacts the quality. This model could be used improving wine, targeting customers depending on their taste buds, rating wines etc.

**Data Dictionary**

|  |  |
| --- | --- |
| Variable | Description |
| Fixed acidity | Fixed acidity in wine (numeric) |
| Volatile acidity | The amount of volatile acid in wine (numeric) |
| Citric acid | Citric acid in wine (numeric) |
| Residual sugar | Sugar after fermentation (numeric) |
| Chlorides | Salts in wine (numeric) |
| Free sulfur dioxide | Amount Free sulfur dioxide in the wine (numeric) |
| Total sulfur dioxide | Free plus the bound sulfur dioxide in wine (numeric) |
| Density | Density of wine |
| pH | Describes how acidic or basic the wine is (numeric) |
| Sulphates | Quantity of Sulphates in wine (numeric) |
| Alcohol | Percent alcohol in wine (numeric) |
| Quality | This is target variable with a score from 1 to 10 with 10 being the best. |

We observe that all our variables are numeric. And hence we will study the distributions of each in the next section.

## Summary Statistics

Dimension on Data:

We have 1599 rows and 12 columns. Quality column is our target variable with a rating of 1 to 10 for every wine based on the taste, with 10 being the highest

Missing Values

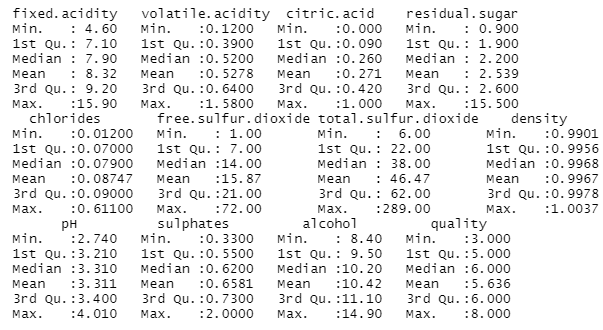
We do not have any missing values in our dataset.

Structure of Dataset:

Text

Description automatically generated

Summary Statistics:



# Part B. Visualization and initial models for a binary response

Binary Variable Excellent:

  
We have 217 samples of excellent wines (quality>=7)

## Distribution of Variables

Diagram, engineering drawing

Description automatically generated

Inferences:

* Apart from density and pH we do not see normal or approx. normal distribution.
* Most of the variables are right skewed, long tail towards right.
* Quality variable also shows that most of the wines are of quality 5 and 6.

## Box plots of variables

Diagram

Description automatically generated

* We observe that some of the variables have some outliers.
* Also, the direction of outlies is consistent with the histogram.

## Chart, pie chart Description automatically generatedProportion of excellent Red wine:

  
We have 217 samples of excellent wines (quality>=7)

That is, 13.57% percent of excellent wine in our dataset.

## Chart, timeline Description automatically generatedCorrelation

|  |  |
| --- | --- |
| Variable | Pearson Correlation with Excellent |
| fixed.acidity | 0.12006104 |
| volatile.acidity | -0.27071153 |
| citric.acid | 0.21471559 |
| residual.sugar | 0.04777895 |
| chlorides | -0.09730764 |
| free.sulfur.dioxide | -0.07174730 |
| total.sulfur.dioxide | -0.13951655 |
| density | -0.15045968 |
| pH | -0.05728334 |
| sulphates | 0.19948521 |
| alcohol | 0.40731485 |

* Excellent shows good correlation with alcohol. Increasing alcohol increases the probability of being labelled excellent.
* Excellent is negatively correlated with volatile acidity
* Whereas, excellent has slightly positive correlation with citric acid

## Linear Model

Text, table

Description automatically generated

**Significant Variables:**Fixed.acidity, volatile.acidity, residual.sugar, chlorides, total.sulfur.dioxide, density, sulphates, alcohol

**Positively correlated significant variables:**  
Fixed.acidity, residual.sugar, sulphates and alcohol

This implies that increasing the above parameters will increase the quality and in turn the chances of being classified as excellent.

Other significant variables are responsible for lowering the rating of wine.

## Logistic Regression

Text, table

Description automatically generated with medium confidence

**Significant Variables:**Fixed.acidity, volatile.acidity, residual.sugar, chlorides, total.sulfur.dioxide, density, sulphates, alcohol

**Positively correlated significant variables:**  
Fixed.acidity, residual.sugar, sulphates and alcohol

This implies that increasing the above parameters will increase the chances of being classified as excellent.

Other significant variables are responsible for lowering the rating of wine.

Comparison Linear and Logistic

* We get exactly same variables as significant in both cases.
* Even the direction is same, i.e. the ones that show positive correlation with linear, same variables show positive correlation with logistic as well.

# Part C. Variable selection, Interpretation, and Prediction for a Logistic model

1. Forward selection wit AIC

Text

Description automatically generated

1. Backward selection wit AIC

Text

Description automatically generated

1. Forward selection with BIC

Text

Description automatically generated

1. Backward selection wit BIC

Text

Description automatically generated

1. Stepwise

Text

Description automatically generated

Since BIC penalizes the model with increase in number of predictors, it gives the simplest model. I will choose BIC as selection criteria and select the model with least BIC. Both Forward and Backward selection process give the same least BIC model:

**excellent ~ fixed.acidity + volatile.acidity + chlorides + total.sulfur.dioxide + sulphates + alcohol**

## Chi Square Test

Chi square test helps us to confirm if the model from variable selection is significantly different from full model or not.

H0: Models are same.  
H1: Models are different.

p value = 0.031  
p value is less than significance level (0.05). Hence, we reject the null hypothesis and confirm the models are statistically different. We need to select another model.

## Drop1 Chi Square Test:

Text

Description automatically generated

drop1 chi square test exactly gives the same model as AIC. I will choose the model with lowest AIC:

**excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density**

Chart, bubble chart

Description automatically generated

## Collinearity

|  |  |
| --- | --- |
| Variable | Pearson Correlation with Excellent |
| fixed.acidity | 0.12006104 |
| volatile.acidity | -0.27071153 |
| citric.acid | 0.21471559 |
| residual.sugar | 0.04777895 |
| chlorides | -0.09730764 |
| free.sulfur.dioxide | -0.07174730 |
| total.sulfur.dioxide | -0.13951655 |
| density | -0.15045968 |
| pH | -0.05728334 |
| sulphates | 0.19948521 |
| alcohol | 0.40731485 |

We see multicollinearity in data. pH and fixed.acidity have high negative correlation. Volatile.acidity and citric.acid are correlated. Citric.acid and fixed.acidity are also correlated. Density and fixed.acidity are also correlated. Citric.acid and pH also display negative correlation. Total.sulfur.dioxide and free.sulfur.dioxide are also correlated.

From these variables, we have the following pairs in our model:  
excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar + density

Density and fixed acidity.  
Removing density, as fixed acidity is correlated with citric.acid, density, and pH, thus it captures information of all these variables.

Final model:  
**excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar**

Text

Description automatically generated

## Influence of Alcohol, Residual.Sugar and pH

* We do not have pH in our final model. We only have Alcohol and Residual sugar
* Alcohol is very significant with p-value close to 0. Its coefficient value of 0.97 implies 1 unit increase in alcohol will increase the odds ratio increases by e0.97
* Residual sugar is also significant at p-value 0.03, it is also positively correlated implying increase in residual sugar content would improve the odds ration by e0.13­­
* We do not have pH in our model, so the pH value does not directly impact our odds ratio. Please note other variables like fixed acidity and citric acid etc. are correlated with pH and might include the effect. But pH indepently and directly will not impact our log odds ratio.

## Confusion Matrix, Specificity and Sensitivity

Cutoff probability = 0.5

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predicted | | |
| Actual |  | 0 | 1 |
| 0 | 1337 | 45 |
| 1 | 146 | 71 |

Specificity (True Negative Rate) = 0.967

Sensitivity (True Positive rate) = 0.327

**We need to lower the threshold to decrease the number of False Negatives.**

Here we see that our model is able to correctly identify the non-excellent wines, but it fails to identify the excellent wines. This shows we have most of the data points including the excellent ones being predicted below 0.5 and hence so many false negatives. Thus, lowering the threshold will help us label these points as 1 or excellent wines.

This is basically the tradeoff between False positives and False negatives.

## Receiver Operating Characteristics (ROC)

Chart, scatter chart

Description automatically generated

**Area Under Curve (AUC) = 0.88**

## Prediction

|  |  |  |  |
| --- | --- | --- | --- |
| Data | Probability | Lower CI | Upper CI |
| 1st Bottle | 0.0096 | 0.0048 | 0.014 |
| 268th Bottle | 0.753 | 0.67 | 0.82 |

So, our model classifies 1st bottle as NOT EXCELLENT  
And 268th bottle as EXCELLENT

This matches the true values(Quality = 5 for Bottle 1 and Quality = 8 for Bottle 2). The classificaton by our model is accurate.

# Part D. Link functions and Dispersion parameter

## Probit Link Function

Text

Description automatically generated with medium confidence

Confusion Matrix:

Text

Description automatically generated

Specificity = 0.97

Sensitivity = 0.31

ROC Curve:

Chart, scatter chart

Description automatically generated

AUC = 0.88

## Complimentary Log Log link function

Text

Description automatically generated

Confusion Matrix:

Text

Description automatically generated with medium confidence

Specificity = 0.972

Sensitivity = 0.304

ROC Curve:

Chart, line chart, scatter chart

Description automatically generated

AUC = 0.877

## Comparison of Logit, Probit and Complimentary Log Log link functions (1 page report)

Comparison of the coefficients:

* The ratio of coefficients of logit to coefficients of probit (including the intercept) is around 1.8 for all values.
* The coefficients of Complimentary log log and logit are much closer than probit and logit.
* Logit and Probit are symmetric link functions. Probit is thinner at both the tails when compared to logit.
* Complimentary log log is not symmetric, it is similar to logit in lower tail but gets thinner in the upper tail.

Model Metrics Comparison

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logit | Probit | Complimentary Log Log |
| AIC | 895 | 891 | 916 |
| BIC | 938 | 954 | 959 |
| Sensitivity | 0.327 | 0.31 | 0.304 |
| Specificity | 0.967 | 0.97 | 0.97 |
| AUC | 0.88 | 0.88 | 0.88 |

* In all the 3 cases we observe that Sensitivity is low and Specificity is high, meaning we need to lower the threshold probability (currently it is 0.5 for all)
* We get same AUC values for all the three cases. This is because, we saw that the ROC curves for all the cures were similar.
* We see that there is not much of a difference in the models when we compare sensitivity, specificity and AUC.
* Logit is able to identify the maximum of excellent wines from confusion matrix(Logit = 71, Probit = 68, Cloglog=66). It is log of odds ratio and easy to interpret. Also, Logit is widely used for logistic regression.

Hence, I would recommend Logit Model.

## Dispersion

Introducing a new parameter gives model the freedom to choose a dispersion parameter(sigma squared) that can describe the dispersion of the distribution.

We get dispersion parameter = 0.828

Final model with new dispersion parameter:

Table

Description automatically generated

Inference

* The dispersion parameter is less than 1 i.e. we have a narrower distribution now.
* The location (center = mean) of the distribution remains same, hence we get same coefficients and intercepts. (No change here)
* The standard error of all the estimates has decreased. This is because as the dispersion parameter has decreased, we have a narrow distribution, tighter confidence intervals.
* And since standard error has decreased, it implies less uncertainty and thus we have smaller p-values.

# Part E. Modeling the wine quality as a multinomial variable with order

Quality is the rating of wine from 1 to 10, with 10 being the best. To analyze the distribution of the quality, we plot frequency of quality.

Chart

Description automatically generated

We observe that the data is spread in 6 categories (3,4,5,6,7,8). Thus, quality is **multinomial ordered variable**. We use **Multinomial Distribution**.

There are 6 categories, wherein quality 3 and quality 8 are sparse. We can merge quality 3 and 4. Also we can merge quality 7 and 8.

Final number of categories = 4. i.e. Category 4, 5, 6 and 7.

Final Count of categories:



## Kendall Tau Correlation

We calculate the Kendal Tau correlation of all the input variables with quality, rank them with 1st being the one with highest correlation

|  |  |  |
| --- | --- | --- |
| Rank | Variable | Correlation |
| 1 | alcohol | 0.38009700 |
| 2 | volatile.acidity | -0.30171854 |
| 3 | sulphates | 0.29969762 |
| 4 | citric.acid | 0.16748100 |
| 5 | total.sulfur.dioxide | -0.15696617 |
| 6 | Chlorides | -0.14847201 |
| 7 | density | -0.13610251 |
| 8 | fixed.acidity | 0.08847661 |
| 9 | free.sulfur.dioxide | -0.04544953 |
| 10 | pH | -0.03382797 |
| 11 | residual.sugar | 0.02619190 |

## Multinomial Model

From the above table, variable selection and the models we built earlier, I have chosen the following variables:  
quality ~ volatile.acidity + alcohol+sulphates + total.sulfur.dioxide + chlorides+ citric.acid

Our output variable quality is ordered multinomial variable and hence we use vglm to model our response.

Text

Description automatically generated

Confusion Matrix

Text

Description automatically generated

In our model we have parallel=TRUE implying that we are building proportional odds model

We have 3 intercepts but same beta coefficients for all the 3 models. The logit functions are parallel. Model 1 calculates the probability of quality <=4, model 2 gives probability of quality<=5, model 3 gives probability of quality <=6. The rest i.e. probability 0f 7, we can calculate once we have these three probabilities.

The predict function with type = ‘link’ gives us the logit value for that model:

Model 1:   
Pr(quality<=4) = ilogit(4.0254 + 3.394545\*volatile.acidity - 0.840836\*alcohol - 2.815713 sulphates + 0.008332\*total.sulfur.dioxide + 4.887978\*chlorides - 0.033615\*citric.acid)

Model 2:   
Pr(quality<=5) = ilogit(4.0254 + 3.394545\*volatile.acidity - 0.840836\*alcohol - 2.815713 sulphates + 0.008332\*total.sulfur.dioxide + 4.887978\*chlorides - 0.033615\*citric.acid)

Model 3:   
Pr(quality<=6) = ilogit(4.0254 + 3.394545\*volatile.acidity - 0.840836\*alcohol - 2.815713 sulphates + 0.008332\*total.sulfur.dioxide + 4.887978\*chlorides - 0.033615\*citric.acid)

Pr(quality =7) = 1 – Pr(quality<=6)

Interpretation from Variables  
Increasing alcohol by 1 unit then odds of quality being lower(4,5,6) decreases by 56.9%  
Increasing total.sulfur.dioxide by 1 unit causes odds of quality (4,5,6) increase by 100.8%  
Increasing volatile.acidity by 1 unit causes odds of quality (4,5,6) increase by 2980%  
Increasing sulphates by 1 unit causes odds of quality being lower(4,5,6) to decrease by 94.1%  
Increasing chlorides by 1 unit causes odds of quality (4,5,6) increase by 13268%

Please note we are getting such high percentages because we are calculating changes per unit of the input parameter, most of the input parameters reflect very small real value changes.

We can conclude from the model increasing alcohol and sulphates improves the quality of wine.

## Logistic and Multinomial Model Comparison

Logistic Model:  
excellent ~ alcohol + volatile.acidity + sulphates + total.sulfur.dioxide + chlorides + fixed.acidity + residual.sugar

Multinomial Model:  
quality ~ volatile.acidity + alcohol + sulphates + total.sulfur.dioxide + chlorides+ citric.acid

* Both models have volatile.acidity, alcohol, sulphates, total.sulfur.dioxide and chlorides common predictor variables
* For both the model alcohol and sulphates show positive direction for increase in quality.
* Consequently, both models show that increase in volatile.acidity, total.sulfur.dioxide and chlorides degrades quality of wine.
* Logistic model has fixed acidity and residual sugar as additional predictors, both are positively correlated with quality.
* Multinomial model has no extra variable (Citric Acid is not statistically significant)

## Prediction using Multinomial model

We get the probability of wine belonging to a particular class by using predict function with type = ‘response’

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Pr(Quality = 4) | Pr(Quality = 5) | Pr(Quality= 6) | Pr(Quality=) 7 |
| Bottle 1 | 0.0813388 | 0.7008397 | 0.2014508 | 0.01637067 |
| Bottle 268 | 0.000677359 | 0.02607724 | 0.288303 | 0.6849424 |

From above table we have, Bottle 1 has quality = 5

Bottle 2 has quality = 7

When we check the actual data quality of bottle 1 is 5 and quality of bottle 268 is 8, our model has correctly identified the quality of bottle 1. Since we have reduced the categories from 6 to 4, i.e. we merged quality 3 and 4 as quality 4, and quality 7 and 8 as quality 7, our model will classify anything below 4 as quality 4 and anything above 7 as quality 7. Hence our model has identified bottle 268 as quality 7.

We can use this model to predict if the bottle of wine is excellent by checking the quality, if our model predicts that the quality is 7 we can label the wine as excellent. Prediction results can be evaluated using confusion matrix and misclassification rate.  
Text

Description automatically generated

To convert this into excellent and not excellent we can merge rows and columns 4,5 and 6. And get a confusion matrix with quality(4,5,6) vs quality(7)

|  |  |  |  |
| --- | --- | --- | --- |
| **Multinomial** | Predicted | | |
| Actual |  | Quality(4,5,6) | Quality(7) |
| Quality(4,5,6) | 1341 | 41 |
| Quality(7) | 150 | 67 |

Accuracy = 88.05%

|  |  |  |  |
| --- | --- | --- | --- |
| **Logistic** | Predicted | | |
| Actual |  | Not Excellent | Excellent |
| Not Excellent | 1337 | 45 |
| Excellent | 146 | 71 |

Accuracy = 88.05%

Here we can see that we have very close results to predict if the wine is excellent using Multinomial and Logistic regression.

Rcode:

