Modeling wine preferences by data mining from physicochemical properties

Bob Ding, Lynn Fan, Alice Jiang

Dec. 2nd, 2018

Introduction

Project Goal: **Explanation**
To identify variables that are important in explaining variation in the response.

We are interested in researching **what factors contribute to the quality of wine for different types of red vinho verde from Portugal**. This data set was used to predict quality of wine for future wine certification, complementary to human wine tasters, in the paper we cited (Modeling wine preferences by data mining from physicochemical properties). We believe that this dataset can also be used to analyze what chemical factors are attributable to the final rating of wine (measured by the variable quality). If we can understand how chemical factors affect the wine quality, it may shed light on future R&D directions for chemical methods that could improve/preserve wine quality.

Regression Methods

We tried several regression methods. First, we started with multiple linear regression with variables and their interactions. In the selected model, we found that assumptions of normality and constant variance not met because quality is categorical. We believe that logistic regressions would be a better fit for this data set, so we worked on logistic regression. We divide the quality as "good" and "not good" with a cutoff of quality at 6 and include both main effects and interactions in the analysis. We found some outliers and patterns on residuals, and we are concerned that for explaining the variables' effects on quality, binary quality variable is not sufficient for our research goal. So we turned to ordinal logistic regression as quality is a naturally ordered response variable. With main effects in the ordinal logistic model, we found it relatively accurate at predicting quality value and the p-values for the explanatory variables are all extremely small. We conclude that the model is a good fit, although more improvements can be made.

Data & Data Exploration

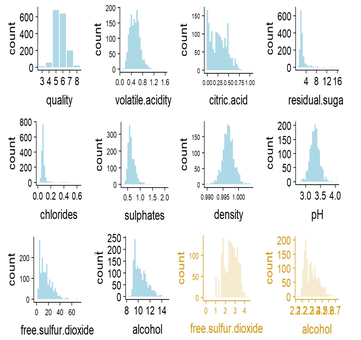
Response Variable:

- **quality:** the quality of the wine (a score between 0 and 10)

Explanatory Variables:

- **fixed.acidity:** (g(tartaric acid)/dm3) the amount of acid in wine that's not volatile (do not evaporate fast)
- **volatile.acidity:** (g(acetic acid)/dm3) the amount of acetic acid in wine; at high levels can lead to an unpleasant and vinegar taste in wines
- **citric.acid:** (g/dm3) citric acid is found in small quantities, and can add freshness and flavor to wines
- **residual.sugar:** (g/dm3) the amount of sugar left after fermatation stops; generally greater than 1 gram/liter in wines and wines with greater than 45 grams/liter are considered sweet
- **chlorides:** (g(sodium chloride)/dm3) the amount of salt in the wine
- **free.sulfure.dioxide:** (mg/dm3) the free form of S02S02 exists in equilibrium between molecular S02S02 (as a dissolved gas) and bisulfite ion
- **total.sulfur.dioxide:** (mg/dm3) amount of free and bound forms of S02S02; at free S02S02 concentration over 50ppm, S02S02 becomes evident in the smell and taste of the wine
- **density:** (g/cm3) the density of the liquid, which is close to the density of water depending on the percent alcohol and sugar content in the wine
- **pH:** the indicator of the acidicity or basic property of the wine on a scale from 0 (very acidic) to 14 (very basic)
- **sulphates:** (g(potassium sulphate)/dm3)a wine additive which can contribute to sulfur dioxide gas S02S02 levels
- **alcohol:** (vol. the percent alcohol content of the wine

```
##  fixed.acidity   volatile.acidity  citric.acid    residual.sugar
##  Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900
##  1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##  Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##  Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##  3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##  Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide
##  Min.   :0.01200   Min.   :0.000       Min.   :  6.00
##  1st Qu.:0.07000   1st Qu.:1.946       1st Qu.: 22.00
##  Median :0.07900   Median :2.639       Median : 38.00
##  Mean   :0.08747   Mean   :2.546       Mean   : 46.47
##  3rd Qu.:0.09000   3rd Qu.:3.045       3rd Qu.: 62.00
##  Max.   :0.61100   Max.   :4.277       Max.   :289.00
##     density             pH          sulphates         alcohol      quality
##  Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   :2.128   3: 10
##  1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.:2.251   4: 53
##  Median :0.9968   Median :3.310   Median :0.6200   Median :2.322   5:681
##  Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :2.339   6:638
##  3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:2.407   7:199
##  Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :2.701   8: 18
```



We can see that free.sulfur.dioxide and alcohol have an obvious rightly skewed distribution. residual.sugar, cholorides, and sulphates also have a slightly rightward skewedness. citric.acid at first appears to have a bimodal distribution, but this is because there are some wines with zero citric acid, so we see a spike at 0 in the histogram. Based on the data definition, we know it is possible for wines to have citric acid of 0. The distribution of citric acid is overall fairly symmetric. All other variables: quality, volatile.acidity, density, and pH are unimodal and fairly symmetric.

Notice that now free.sulfur.dioxide,and alcohol is logged

Conclusions for Linear Regression model

$$\begin{aligned} quality = {} & 4.2245 - 0.0512 * residual.sugarCent - 0.6498 * citric.acid \\ & + 0.0204 * free.sulfur.dioxideCent - 0.1836 * sulphatesCent \\ & + 57.4648 * densityCent - 0.1355 * alcoholCent - 0.0438 * citric.acid * free.sulfur.dioxideCent \\ & - 1.9995 * citric.acid * sulphatesCent + 0.0103 * free.sulfur.dioxideCent * alcoholCent \\ & - 1.3477 * sulphatesCent * alcoholCent \end{aligned}$$

Holding all else constant: : (interpretation for interaction hasn't been added)

- An average wine should have a mean of quality rating as 4.22
- An average increase in residual sugar will lead to a mean of 0.05 decrease in quality
- An average increase in citric acid will lead to a mean of 0.65 decrease in quality
- An average increase in free sulfur dioxide will lead to a mean of 0.02 increase in quality
- An average increase in sulphates will lead to a mean of 0.18 decrease in quality
- An average increase in density will lead to a mean of 57.46 increase in quality
- An average increase in alcohol will lead to a mean of 0.14 decrease in quality

Conclusions for Logistic Regression model

$$\begin{aligned} \frac{\pi_{good}}{\pi_{bad}} = \exp(& 0.539269 + 0.996689 * alcoholCent - 3.519343 * volatile.acidityCent \\ & + 3.367093 * sulphatesCent - 2.259180 * chloridesCent - 0.014257 * free.sulfur.dioxideCent \\ & - 0.976473 * citric.acid + 1.549278 * alcoholCent * sulphatesCent) \end{aligned}$$

Holding all else constant: (interpretation for interaction hasn't been added)

- The log odd of being good wine vs bad wine for an average wine is 0.54
- An increase in alcohol will lead to a 1.00 increase in log-odd of good wine vs bad wine
- An increase in volatile acidity will lead to 3.52 decrease in log-odd of good wine vs bad wine
- An increase in alcohol will lead to a 3.37 increase in log-odd of good wine vs bad wine
- An increase in alcohol will lead to a 2.25 increase in log-odd of good wine vs bad wine
- An increase in alcohol will lead to a 0.01 increase in log-odd of good wine vs bad wine
- An increase in alcohol will lead to a 0.98 increase in log-odd of good wine vs bad wine
- An increase in alcohol will lead to a 1.55 increase in log-odd of good wine vs bad wine

Conclusions for Ordinal Model

$$log(\frac{P(quality \leq 3)}{P(quality > 3)}) = -6.0844790 - [-3.5986801 * volatile.acidityCent - 0.7397748 * citric.acid - 5.2592883 * chloridesCent + 2.7344891 * sulphatesCent - 1.4775465 * pHCent + 0.9493381 * alcoholCent]$$
$$log(\frac{P(quality \leq 4)}{P(quality > 4)}) = -4.1434489 - [-3.5986801 * volatile.acidityCent - 0.7397748 * citric.acid - 5.2592883 * chloridesCent + 2.7344891 * sulphatesCent - 1.4775465 * pHCent + 0.9493381 * alcoholCent]$$
$$log(\frac{P(quality \leq 5)}{P(quality > 5)}) = -0.4952119 - [-3.5986801 * volatile.acidityCent - 0.7397748 * citric.acid - 5.2592883 * chloridesCent + 2.7344891 * sulphatesCent - 1.4775465 * pHCent + 0.9493381 * alcoholCent]$$
$$log(\frac{P(quality \leq 6)}{P(quality > 6)}) = 2.3184245 - [-3.5986801 * volatile.acidityCent - 0.7397748 * citric.acid - 5.2592883 * chloridesCent + 2.7344891 * sulphatesCent - 1.4775465 * pHCent + 0.9493381 * alcoholCent]$$
$$log(\frac{P(quality \leq 7)}{P(quality > 7)}) = 5.3309018 - [-3.5986801 * volatile.acidityCent - 0.7397748 * citric.acid - 5.2592883 * chloridesCent + 2.7344891 * sulphatesCent - 1.4775465 * pHCent + 0.9493381 * alcoholCent]$$

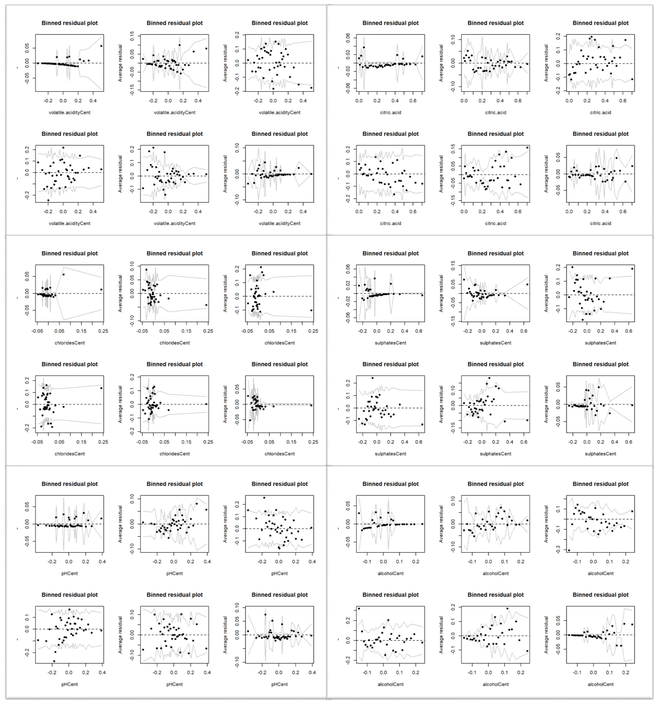
- For one unit increase in volatile.acidity, the odds of the wine falling at or below quality i multiply by a factor of 0.027359812, given that all other variables are held constant.
- For one unit increase in citric.acid, the odds of the wine falling at or below quality i multiply by a factor of 0.477221357, given that all other variables are held constant.
- For one unit increase in chlorides, the odds of the wine falling at or below quality i multiply by a factor of 0.005199004, given that all other variables are held constant.
- For one unit increase in sulphates, the odds of the wine falling at or below quality i multiply by a factor of 15.401872706, given that all other variables are held constant.
- For one unit increase in pH, the odds of the wine falling at or below quality i multiply by a factor of 0.228196887, given that all other variables are held constant.
- For one unit increase in % alcohol content, the odds of the wine falling at or below quality i multiply by a factor of 2.583998874, given that all other variables are held constant.

Ordinal Regression Model Detail

term	estimate	std.error	statistic	coefficient_type
volatile.acidityCent	-3.5986801	0.3777477	-9.526676	coefficient
citric.acid	-0.7397748	0.3751994	-1.971684	coefficient
chloridesCent	-5.2592883	1.3069001	-4.024247	coefficient
sulphatesCent	2.7344891	0.3620702	7.552373	coefficient
pHCent	-1.4775465	0.4145199	-3.564477	coefficient
alcoholCent	0.9493381	0.0579898	16.370780	coefficient
3|4	-6.0844790	0.3430896	-17.734373	zeta
4|5	-4.1434489	0.1783815	-23.228019	zeta
5|6	-0.4952119	0.1160465	-4.267356	zeta
6|7	2.3184245	0.1366369	16.967775	zeta
7|8	5.3309018	0.2700297	19.741910	zeta

This is the final model after selection and dropping out insignificant predictors. We see there are 5 possible intercepts/models. The T-statistics for the model each variables are bigger than 1.96, which means that their p-values are all smaller than 0.05. Therefore, we can choose this as or final ordinary regression model.

Ordinal Regression Assumption Check



The binned residual plot shows normal distribution, independence, constant variances for most of the graphs. There are some minor concerns:

- In chloridesCent, the binned residual plots show a possible chloridesCent outlier.
- In sulphatesCent, the binned residual plot shows some potential linear trend and outliers in three categories of quality. We might need to transform sulphatesCent.
- In pHCent, there is also some linear trends in the binned residual plots for pHCent. We should also transform this variable.

names	x
volatile.acidityCent	6.284542
citric.acid	1.355600
chloridesCent	1.352646
sulphatesCent	1.566299
pHCent	1.486717
alcoholCent	1.345261

The vif function shows that in the ordinal regression model, there isn't serious multi-collinearity going on.

Ordinal Regression Prediction

pred.comp	3	4	5	6	7	8
4	1	NA	NA	NA	NA	NA
5	8	39	510	223	9	NA
6	1	14	167	386	138	10
7	NA	NA	3	29	52	8
8	NA	NA	1	NA	NA	NA

Shown is a confusion matrix from the ordinal regression model's prediction. We can see that the model is not making prediction from quality= 1,2, 9 and 10. This is due to the lack of according quality data in the raw data we've got from the internet, as other categories have completely dominated the prediction in odds probability.

Besides, we can easily see that the model is doing a fairly good job in prediction, as there are 10 categories, the prediction accuracy, which is about 57.3% is way higher than the baseline 10%. But this is still far away from being a perfect model. We can improve this by getting more data, doing transformation, or even by reduce the number of ordinal categories into 5.

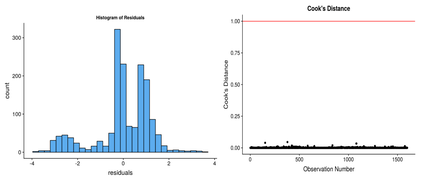
Linear Regression Model Detail

term	estimate	std.error	statistic	p.value
(Intercept)	4.2245320	0.0592509	71.2990038	0.0000000
residual.sugarCent	-0.0512202	0.0256054	-2.0003694	0.0456304
citric.acid	-0.6497755	0.1867955	-3.4785390	0.0005178
free.sulfur.dioxideCent	0.0204311	0.0051881	3.9380530	0.0000857
sulphatesCent	-0.1836422	0.3561499	-0.5156318	0.6061833
densityCent	57.4648324	24.3560389	2.3593669	0.0184266
alcoholCent	-0.1355440	0.0386503	-3.5069347	0.0004659
citric.acid:free.sulfur.dioxideCent	-0.0437867	0.0157370	-2.7824077	0.0054596
citric.acid:sulphatesCent	-1.9995045	0.8489262	-2.3553337	0.0186271
free.sulfur.dioxideCent:alcoholCent	0.0103289	0.0029290	3.5264524	0.0004331
sulphatesCent:alcoholCent	-1.3476692	0.1940602	-6.9445928	0.0000000

We can see that p-value of all the coefficients are less than the threshold of 0.05, except for sulphatesCent. Therefore, these main and interaction effects are significant predictors of wine quality (besides sulphatesCent). So we will make this our final model.

Linear Regression Assumption Check

names	x
residual.sugarCent	1.402748
citric.acid	1.425081
free.sulfur.dioxideCent	3.169724
sulphatesCent	3.922514
densityCent	2.274231
alcoholCent	1.825881
citric.acid:free.sulfur.dioxideCent	3.249458
citric.acid:sulphatesCent	3.832337
free.sulfur.dioxideCent:alcoholCent	1.097922
sulphatesCent:alcoholCent	1.187895



Overall, the final model does not have any major multicollinearity concerns - all VIF values are fairly low and below the threshold of 10. However, multiple regression model does not seem to be the best fitting for our data set. Given we have a categorical response variable, we will try different types of logistic regresison models.

The distribution of residuals appear to be bimodal. The QQ plot also suggests the same conclusion, since we can see a very prominent deviation from the diagonal normal line on the left side. Overall, the Normaltiy Assumption seems to be violated. This could be due to the fact that we are dealing with a categorical response varaible quality.

We can see that there is one point with a significantly high leverage around 0.4, comparing with other observations. This could be an outlier and might be an influence point.

Logistic Regression Model Detail

term	estimate	std.error	statistic	p.value
(Intercept)	0.4951634	0.1207867	4.099486	0.0000414
alcoholCent	0.9510158	0.0709582	13.402476	0.0000000
volatile.acidityCent	-3.5486411	0.4515594	-7.858636	0.0000000
sulphatesCent	2.7030432	0.4463930	6.055299	0.0000000
chloridesCent	-3.7700684	1.4668058	-2.570257	0.0101623
free.sulfur.dioxideCent	-0.0132536	0.0058580	-2.262476	0.0236680
citric.acid	-0.8755285	0.3918921	-2.234106	0.0254761

All the variables have p-values much smaller than 0.05, so there is significant evidence that they are important predictors of the log-odds (and therefore odds) of wine quality (good vs. not good). Based on the model, it appears that the percent alcohol content of the wine is the strongest predictor of wine quality. alcoholCent has the largest test statistic magnitude of 13.402476. The positive test statistic value also shows that as the % alcohol content of the wine increases, the logs-odds of good vs. not good wine quality increases. The amount of acetic acid in wine (volatile.acidityCent) and the amount of wine additive (sulphatesCent) are also strong predictors.

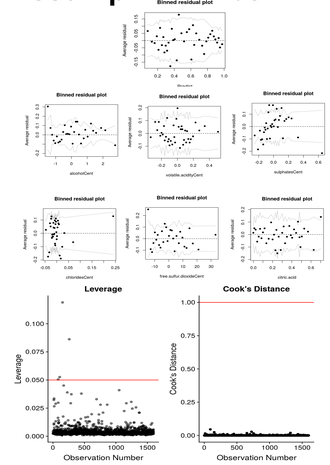
Logistic Regression Assumption Check

Most of the binned residual plots show random pattern and raises no major concerns of violations. The binned residual plot of alcoholCent and free.sulfur.dioxideCent seems to show an outlier on the top left, suggesting high average residuals for that bin.

There is also one outlier on the binned residual plot of chloridesCent on the far right. The binned residual plot also reveals that sulphatesCent might need a transformation, since there appears to be some linear trend.

names	x
alcoholCent	1.050002
volatile.acidityCent	1.542159
sulphatesCent	1.416972
chloridesCent	1.433665
free.sulfur.dioxideCent	1.018900
citric.acid	1.650265

VIF values are very small, so there is no major concerns of multicollinearity.



Overall, there is no significant influence point.

Logistic Model Prediction

From the ROC curve and AUC calculation, we can see the curve is fairly close to the top left corner (area under the curve is close to 1).

This shows that the logistic model is able to distinguish between good and not good quality, so this is a pretty good model.