A report on what makes wine good presentation. Using regression analysis to determine which physicochemical properties make a wine ‘good’… By Anastasia Franio STAT 410

Background

In 2006 wine collector, Rudy Kurniawan, was promoted as possessing “the greatest cellar on Earth.” In fact, he organized the biggest Ponzi scheme in the wine industry existed. Kurniawan was arrested on 8 March, 2012 for and indicted for selling fake high-end wine at auction.

How the market could have prevented it?

Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters. Another key factor is physicochemical tests which take into account different chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification process is more controlled.

Data Description

Two datasets: red wine has 1599 different varieties and white wine and has 4898 varieties.

All wines are produced in a particular area of Portugal. This dataset is related to red variants of the Portuguese "Vinho Verde" wine. Data are collected on 12 different properties of the wines one of which is Quality, based on sensory data, and the rest are on chemical properties.

All chemical properties of wines are continuous variables. Quality is an ordinal variable with possible ranking from 1 (worst) to 10 (best).

The following variables were considered in the dataset:

1) fixed acidity (most acids involved with wine or fixed or nonvolatile (do not evaporate readily);

2) volatile acidity (the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste);

3) citric acid (found in small quantities, citric acid can add 'freshness' and flavor to wines);

4) residual sugar (the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet);

5) chlorides (the amount of salt in the wine);

6) free sulfur dioxide (the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine);

7) total sulfur dioxide (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);

8) density (the density of water is close to that of water depending on the percent alcohol and sugar content);

9) pH (describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale);

10) Sulphates (a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant);

11) Alcohol (the percent alcohol content of the wine).

Methodology

Beta regression was used to identify the most crucial physicochemical properties of wine that influence on its quality. The quality response was factored by 100 in order to be able to use beta regression. After regressing the quality on combined data sets from red and white wine I didn’t receive any significant

predictors and model didn't fit good enough. Therefore, I decided to regress the quality on physicochemical properties using just the red wine data set.
From the regressing the red wine data set, we received several significant predictors like volatile acidity, chlorides, total sulfur dioxide, pH, sulphates, and alcohol. The interpretation of these beta coefficients is the following.

As the volatile acidity increases by one point the estimated mean quality point decreases by (exp {-8.95}-1)*100%=99%, which we can relate to real life because at too high level it can lead to a vinegar, unpleasant taste of wine.

As the chlorides level increases by one point, the estimated mean quality decreases by (exp{-4.47}-1)*100%= 98%, which can be explained by the fact that chlorides level corresponds to the amount of salt in wine which tasters don't really want to feel while tasting a wine.

As the total sulfur dioxide level increases by one point, the estimated mean quality score decreases by 98%, which is to be expected as too high concentrations of sulfur dioxide it becomes evident at the nose of wine.

As the pH level increases by one point, the estimated mean quality score decreases by (exp{-2.16}-1)*100%=88%, which importance can be explained by the fact that 'good' wine shouldn't be too acidic.

As the sulphates level increases by one point, the estimated mean quality score increases by (exp{8}-1)*100%=3000000%, which can be explained by the fact that sulphates level corresponds to antioxidant level, which is good for drinkers.

As the alcohol level increases by one point, the estimated mean score increases by (exp{10.43}-1)*100%=5000000%, which is self-explanatory.

After making sure that model has a good fit, I predicted wine score with the following physicochemical qualities:

Fixed acidity=7.4, volatile acidity=0.7, citric acid=0.05, residual sugar=1.9, chlorides=0.075, free sulfur dioxide=20, total sulfur dioxide=50, density=0.988, pH=3.16, sulphates=0.56, alcohol=10, the predicted wine score was around 0.53.

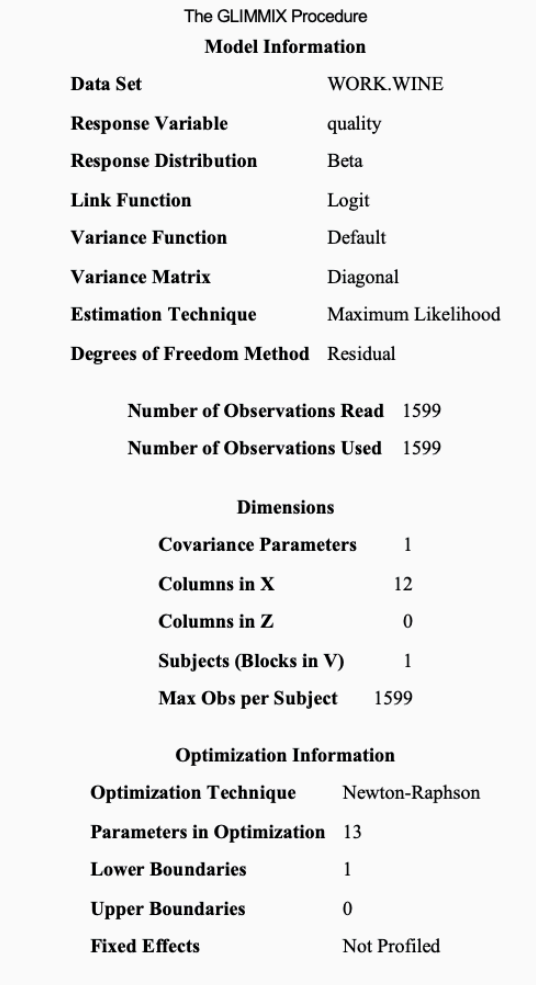
Conclusion

Physicochemical properties can be successfully used for prediction of a wine quality, as the regression analysis above proves. After regressing the wine quality on chemical properties, we identify several significant predictors, like sulphates level, pH, etc. Those properties can be also identified by tasters without using special technologies for measurement of the specific level and can be helpful to know to protect buyers from fake wine. R and SAS analysis had very similar outputs with the same list of the significant predictors and the prediction itself.

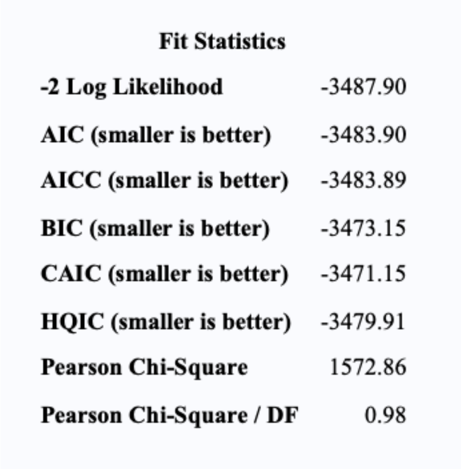
Appendix A (SAS code and output)

```
*import wine data;
proc import out=Wine
  datafile="C:\Users\afranio\Desktop\NewWine.xls"
   dbms=xls replace;
run;

data Wine;
set Wine;
quality = quality/10;
run;

proc print data = Wine (obs = 10);
run;


proc glimmix;
model quality = fixed_acidity volatile_acidity citric_acid residual_sugar
chlorides free_sulfur_dioxide total_sulfur_dioxide density pH sulphates alcohol/dist = beta
link = logit solution;
run;

proc glimmix;
model quality =/dist = beta link = logit;
run;

data deviance_test;
*deviance = -4207-(-3487);
deviance = -3487-(-4207);
pvalue = 1 - probchi(deviance,11);
run;

proc print;
run;

data prediction;
input fixed_acidity volatile_acidity citric_acid residual_sugar chlorides free_sulfur_dioxide
total_sulfur_dioxide density pH sulphates alcohol;
cards;
7.4 0.7 0.05 1.9 0.075 20 50 0.988 3.16 0.56 10
;
```

```
data Wine;
	set Wine prediction;
run;

proc glimmix;
model quality = fixed_acidity volatile_acidity citric_acid residual_sugar
chlorides free_sulfur_dioxide total_sulfur_dioxide density pH sulphates alcohol/dist = beta
link = logit solution;
output out = outdata pred(ilink) = p_quality;
run;

proc print data = outdata (firstobs = 1600 obs = 1600);
var p_quality;
run;
```



```
                 The GLIMMIX Procedure
                  Model Information

    Data Set                      WORK.WINE
    Response Variable             quality
    Response Distribution         Beta
    Link Function                 Logit
    Variance Function             Default
    Variance Matrix               Diagonal
    Estimation Technique          Maximum Likelihood
    Degrees of Freedom Method     Residual

        Number of Observations Read    1599
        Number of Observations Used    1599

                     Dimensions
          Covariance Parameters        1
          Columns in X                12
          Columns in Z                 0
          Subjects (Blocks in V)       1
          Max Obs per Subject       1599

              Optimization Information
      Optimization Technique      Newton-Raphson
      Parameters in Optimization  13
      Lower Boundaries            1
      Upper Boundaries            0
      Fixed Effects               Not Profiled
```



```
             Fit Statistics
 -2 Log Likelihood              -3487.90
 AIC (smaller is better)        -3483.90
 AICC (smaller is better)       -3483.89
 BIC (smaller is better)        -3473.15
 CAIC (smaller is better)       -3471.15
 HQIC (smaller is better)       -3479.91
 Pearson Chi-Square              1572.86
 Pearson Chi-Square / DF            0.98
```

Pearson Chi-Square / DF	1.00

Parameter Estimates

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	8.1733	8.7612	1587	0.93	0.3510
fixed_acidity	0.01124	0.01078	1587	1.04	0.2974
volatile_acidity	-0.4402	0.04973	1587	-8.85	<.0001
citric_acid	-0.07387	0.06063	1587	-1.22	0.2233
residual_sugar	0.007636	0.006203	1587	1.23	0.2185
chlorides	-0.7825	0.1728	1587	-4.53	<.0001
free_sulfur_dioxide	0.001673	0.000894	1587	1.87	0.0613
total_sulfur_dioxide	-0.00129	0.000299	1587	-4.32	<.0001
density	-8.5765	8.9434	1587	-0.96	0.3377
pH	-0.1738	0.07925	1587	-2.19	0.0284
sulphates	0.3842	0.04814	1587	7.98	<.0001
alcohol	0.1150	0.01097	1587	10.48	<.0001
Scale	56.6475	1.9863	.	.	.

Fit Statistics

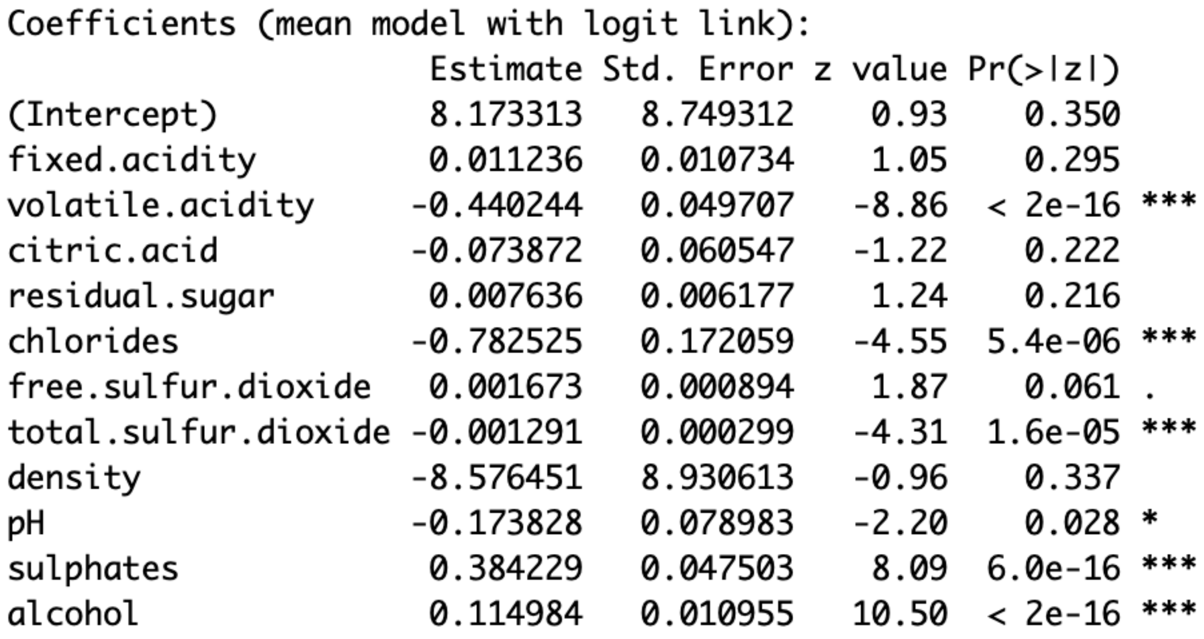
-2 Log Likelihood	-4207.34
AIC (smaller is better)	-4181.34
AICC (smaller is better)	-4181.11
BIC (smaller is better)	-4111.44
CAIC (smaller is better)	-4098.44
HQIC (smaller is better)	-4155.38
Pearson Chi-Square	1588.88
Pearson Chi-Square / DF	1.00

Obs	deviance	pvalue
1	720	0

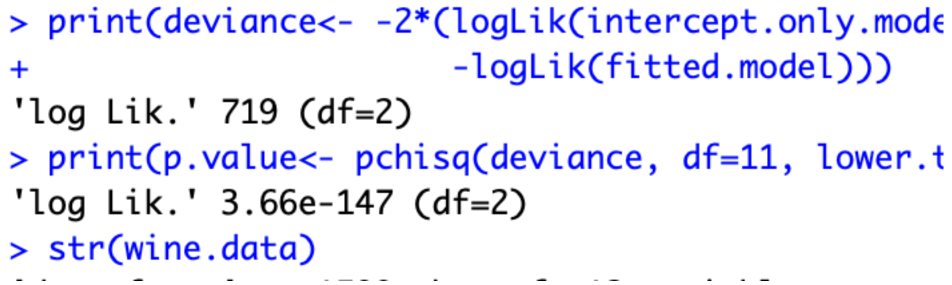
Obs	p_quality
1600	0.55268

Appendix B (R code and output)

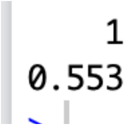
```
wine.data <- read.csv(file = "~/Desktop/winequality-red.csv",
               sep = ";", header = TRUE)

library(dplyr)

wine.data <- wine.data %>%
  mutate(quality = quality/10)

install.packages("betareg")
library(betareg)

#fitting beta regression model
summary(fitted.model<- betareg(quality~., data=wine.data, link="logit"))
intercept.only.model<- betareg(quality ~ 1,
                     data=wine.data, link="logit")
print(deviance<- -2*(logLik(intercept.only.model)
              -logLik(fitted.model)))
print(p.value<- pchisq(deviance, df=11, lower.tail=FALSE))

#using fitted model for prediction
print(predict(fitted.model, data.frame(fixed.acidity=7.4, volatile.acidity=0.7, citric.acid=0.05,
residual.sugar=1.9, chlorides=0.075,
          free.sulfur.dioxide=20, total.sulfur.dioxide=50, density=0.988,
          pH=3.16, sulphates=0.56, alcohol=10)))
```



```
Coefficients (mean model with logit link):
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)           8.173313   8.749312    0.93    0.350    
fixed.acidity         0.011236   0.010734    1.05    0.295    
volatile.acidity     -0.440244   0.049707   -8.86  < 2e-16 ***
citric.acid          -0.073872   0.060547   -1.22    0.222    
residual.sugar        0.007636   0.006177    1.24    0.216    
chlorides            -0.782525   0.172059   -4.55  5.4e-06 ***
free.sulfur.dioxide   0.001673   0.000894    1.87    0.061 .  
total.sulfur.dioxide -0.001291   0.000299   -4.31  1.6e-05 ***
density              -8.576451   8.930613   -0.96    0.337    
pH                   -0.173828   0.078983   -2.20    0.028 *  
sulphates             0.384229   0.047503    8.09  6.0e-16 ***
alcohol               0.114984   0.010955   10.50  < 2e-16 ***
```



```
> print(deviance<- -2*(logLik(intercept.only.mode
+                      -logLik(fitted.model)))
'log Lik.' 719 (df=2)
> print(p.value<- pchisq(deviance, df=11, lower.t
'log Lik.' 3.66e-147 (df=2)
> str(wine.data)
```



```
    1 
0.553 
```