Effect of Physicochemical Property on Wine Quality

Weiqi Weng
College of Computer and Information Science, Northeastern University

Introduction

Viticulture Commission of the Vinho Verde Region collected physicochemical data of wine, including Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates and Alcohol, to study what kind of wine is more attractive. Quality is described as sensory data. It may be designated by controlled group of people based on taste, fragrance and so on. This project aims to build a multiple regression model, using Quality of wine as the response variable, predict Quality and discuss the effect of each explanatory variable. $\alpha = 0.05$ is set as the significance level.

Methods

Data Preprocessing

To begin with, we detect potential outliers with box plots and deal with them according to preceding research result and legal restraint. Logarithmic transformation on Residual Sugar fixes significant skewness. In this report, we mean log_e(Residual Sugar) by Residual Sugar.

Waterhouse Lab of University of California Davis presents the major wine components with reference data on their website. For example, levels of tartaric, malic, citric and succinic acid in Fixed Acidity are expected to be 1000 to 4000 mg/L, 0 to 8000 mg/L, 0 to 500 mg/L and 500 to 2000 mg/L. (Doug Nierman, 2004) Therefore, observations with citric acid larger than 700 mg/L are questionable and we have reason to believe that the fixed acid of wine rarely exceeds 14000 mg/L. Similar procedures have been done on other variables. Please refer to Appendix for details.

Fixed Acidity contains Citric Acidity. The similar situation, but more serious with $r = 0.614$, is within Free Sulfur Dioxide and Total Sulfur Dioxide. Therefore, we try to combine their information by adding two more variables, citric acid proportion = (citric acid)/(fixed acidity) and free sulfur dioxide proportion = (free sulfur dioxide)/(total sulfur dioxide).

Model Selection: Data Set Splitting and Exhaustive Subset Selection

After preprocessing, we split the data set equally into building and validation data sets while keeping similar statistical property of Quality. Acids impart the sourness or tartness that is a fundamental feature in wine taste, (Doug Nierman, 2004) which consequently affects people's reflection on Quality and makes it reasonable to assume that Fixed Acidity will be presented in the model. Exhaustive subset selection is performed under Schwartz' Bayesian Information Criterion. However, multicollinearity comes with Density, Alcohol and Free Sulfur Dioxide, which is also shown by the large variance inflation factors in table 1 and correlation matrix in Appendix. We consider two ways to solve it. One is ridge regression and the other is dropping variable while we have to compromise between the pros and cons of both methods.

Multicollinearity Solution: Ridge Regression and Variable Dropping

First, we try ridge regression. We plot ridge trace with ridge constant ranging from 0 to 100 and output the HKB estimator, L-W estimator and the smallest value of GCV. However, the effect is too weak to offset the cost of inference procedures. Therefore, we choose to drop variables.

Exposing Density and Free Sulfur Dioxide to scrutiny, we compute the correlation between Density - Residual Sugar, Density - Alcohol and Residual Sugar - Alcohol. We see the coefficient estimation in standardized model is unstable. As to Free Sulfur Dioxide and Free Sulfur Dioxide Proportion, we compare the standardized model coefficients within their standard error and compare t test statistic to conclude that Proportion is more informative.

Parameter Estimation and Inference

After dropping Density and Free Sulfur Dioxide, we estimate coefficients of Fixed Acidity, Volatile Acidity, Residual Sugar, pH, Alcohol and Free Sulfur Dioxide Proportion by Least Square Method. The ANOVA table shows a fair proportion of variance in Quality has been reduced by the selected variables. The model passes tests for coefficients, F test for regression relation but fails the F test for lack of fit. We move a step further to curvature and interaction investigation.

Curvature and Interaction Investigation

Partial residual plots of selected variables detect obvious nonlinearity within Residual Sugar and pH and slight nonlinearity within Fixed Acidity. We introduce square of each term and interaction between each two of them. Combining T test statistics, standardized model coefficients and standard error with some points in wine production process helps us to keep square of Fixed Acidity, square of Residual Sugar and interaction between Fixed Acidity and Residual Sugar in our model. The refined model passes F test for lack of fit.

Diagnostics: Influential Point Detection, Diagnostic Plots and Tests

Residual vs. fitted value plot and Q-Q plot justify normal assumption. We look close into studentized residual plot, employ Bonferroni outlier test and output the distribution of observations with large hat statistic and Cook's distance in influence plot. Then we delete suspicious observations. Constance of error is tested by Scale-location graph and double-checked by non-constant variance score test. Also we employ Durbin-Watson test to ensure independence of response variable. Finally, a refined model is fitted.

Validation and Prediction: Mean Squared Prediction Error

First, we present the refined tentative model. Secondly, we pick out five observations from the validation data set as new observations to do interval estimation. Thirdly, we compute mean squared prediction error to measure how the model performs. Also refer to Model Discussion for model interpretation.

Relative Weight

Now we want to learn which factor affects more on wine Quality. First we investigate the standardized model coefficients. Then we compare the result with relative weights of each variable.

Results

Data Preprocessing

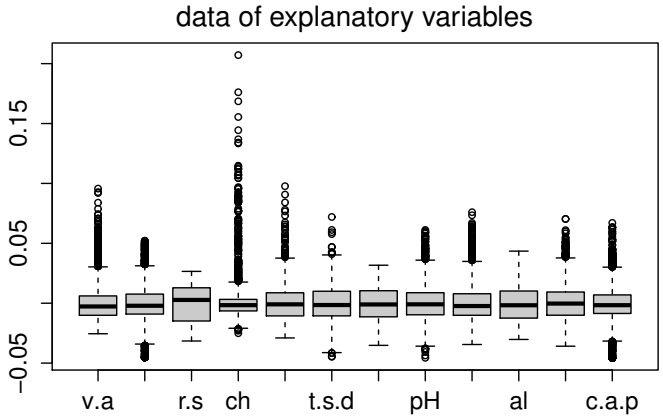

figure 1. box plot of explanatory variables after correlation transformation

Here we present the box plot of explanatory variables after correlation transformation in figure 1.

Model Selection: Data Set Splitting and Exhaustive Subset Selection

For exhaustive subset selection, 8 out of 13 variables are selected. Refer to Appendix for model selection plot in figure 7. The estimates are presented in table 1. The intercept is 49.032 for unstandardized model.

indice	variable name	VIF	coefficient	std. model coefficient
X_1	Fixed Acidity	2.238	0.885	0.846
X_2	Volatile Acidity	1.093	-1.800	-1.798
X_3	Residual Sugar	6.501	-0.111	-0.193
X_4	Free Sulfur Dioxide	3.514	-0.002	0.001
X_5	Density	17.697	-53.221	-3.705
X_6	pH	1.824	0.479	0.322
X_7	Alcohol	7.389	0.310	0.376
X_8	Free Sulfur Dioxide Proportion	3.302	0.813	0.905

table 1. model created with the variables by exhaustive subset selection

Note that VIFs of Density and Residual Sugar reveal multicollinearity problem.

Multicollinearity Solution: Ridge Regression and Variable Dropping

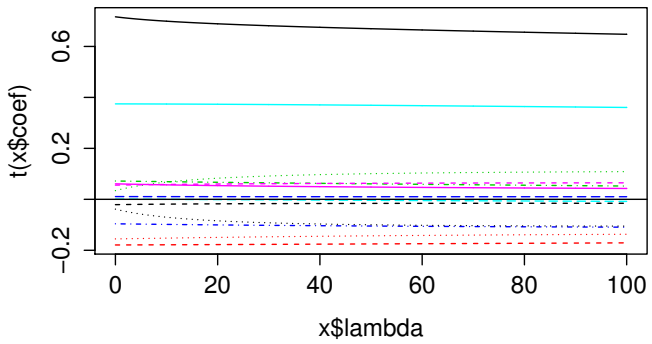

figure 2. ridge trace

The ridge constant estimators, HKB estimator = 9.942, L-W estimator = 14.432, and the smallest value of GCV = 8.6, are all large estimators for all variables. The larger the ridge constant is, the less chance the sampling distribution of a single coefficient will cover its expected value. Ideally we want to keep a good balance between variance and being biased. In addition, from figure 2, we can see that the ridge traces of variables selected by exhaustive subset selection are stable even for very small ridge constant, which indicates that ridge regression does not help a lot here. Besides, statistical inference are not applicable since the exact distribution of coefficients are unknown in ridge regression. Therefore, we choose option 2, dropping Density.

From the correlation matrix in Appendix, we see that information of Density overlaps with both Residual Sugar and Alcohol. However, Residual Sugar and Alcohol do not convey much common information. When we fit the model with selected variables directly, we see that the coefficient of Density, -53.221, is so much larger than other coefficients that it needs a large intercept, 49.032, to offset its effect, while intercept is not our interest. Besides, the estimated standardized model coefficient of Density is -3.705 with standard error 0.193, which makes it an unstable estimate compared to other coefficients. Therefore, it's abandoned.

The standardized model coefficient of Free Sulfur Dioxide Proportion, 0.905, is much larger than that of Free Sulfur Dioxide, 0.001. Besides, Free Sulfur Dioxide Proportion is a relative value which implies more about wine attribute so that we regard it as more descriptive and consequently rule out Free Sulfur Dioxide.

Parameter Estimation and Inference

The following table 2 is a summary of the model with 6 predictor variables plus intercept = -3.66. The ANOVA table shows that $SSE = 203.200, MSE = 0.900, R^2 = 0.893$ and $R_a^2 = 0.893$. 89.3% of the variation in Quality is reduced by the use of selected variables.

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Fixed Acidity	0.844	0.008	102.156	$< 2^{-16}$
Volatile Acidity	-1.786	0.063	-28.198	$< 2^{-16}$
Residual Sugar	-0.194	0.007	-26.494	$< 2^{-16}$
pH	0.318	0.045	7.024	$2.81 \cdot 10^{-12}$
Alcohol	0.376	0.005	69.518	$< 2^{-16}$
Free Sulfur Dioxide Proportion	1.097	0.066	16.612	$< 2^{-16}$

table 2. model summary without multicollinearity

The model passes t tests for coefficients, F test for regression relation but fails the F test for lack of fit. Due to limited space, please refer to Appendix for complete tests.

Curvature and Interaction Investigation

The partial residual plots in figure 3 reveal possible nonlinearity within Residual Sugar, pH and Fixed Acidity.

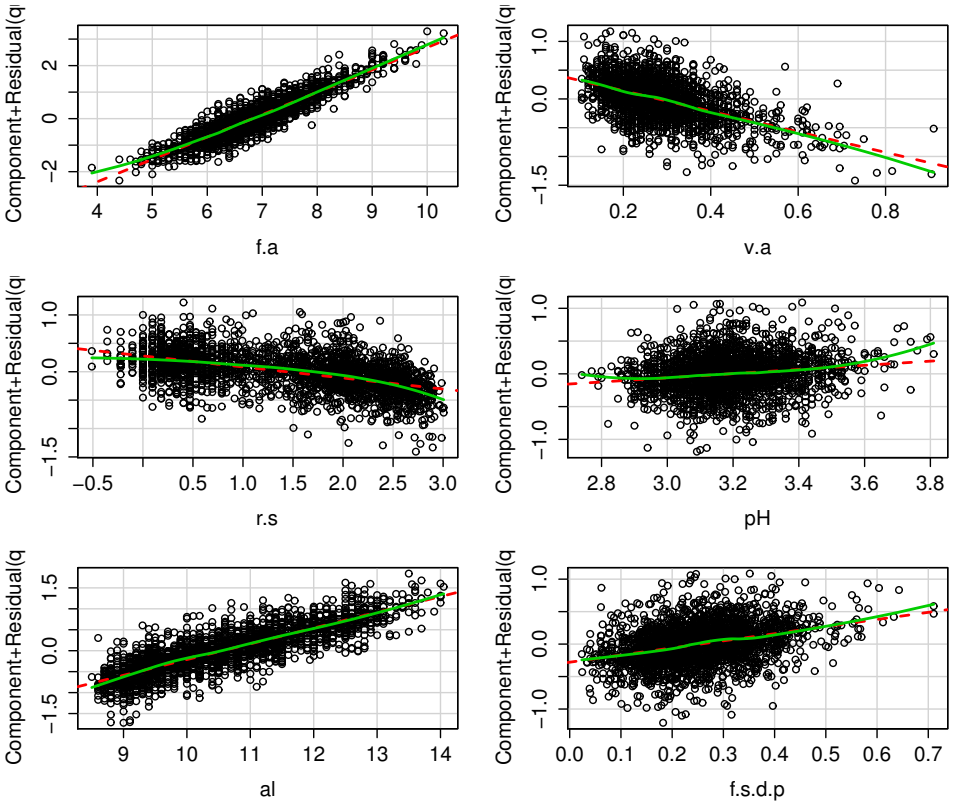

figure 3. partial residual plot

We present the model with all second-order terms involving these factors in Appendix. It makes sense to have a quadratic term of Fixed Acidity since it plays such an important role in wine. From a subjective perspective, what average people care about a wine is its taste, smoothness, fragrance and even color which are controlled by the listed physicochemical property in most cases. For example, Residual Sugar affects taste. Fixed Acidity and Volatile Acidity contributes to smoothness and fragrance. In addition, they make long-term preservation of wine reasonable since it takes time for full fermentation whose by-product is acidity, especially Fixed Acidity. Compared to other property, the effect of pH should be relatively stable for the purpose of long-term preservation. From this point of view, square of pH is inappropriate since it not only negates coefficient of first-order term pH and increases its standard error from 0.045 to 1.638. pH provides far less information than other factors in the second-order interaction term involving pH. Consequently we think interaction between Fixed Acidity and pH, Residual sugar and pH are not necessary, let along their failure to pass the T test for coefficient.

To sum up, we tend to keep terms involving Residual Sugar and Fixed Acidity in a simple model and desire for small standard error of pH coefficient. This motivates us to introduce square of Fixed Acidity, square of Residual Sugar and interaction between Fixed Acidity and Residual Sugar. It turns out the refined model explains 89.7% of the variance in Quality and passes F test for Lack of Fit. Please refer to Appendix for complete test.

Diagnostics: Influential Point Detection, Diagnostic Plots and Tests

The $p-value$ of non-constant variance score test is 0.883 which means lack of evidence for non-constant error issue. From the Q-Q plot in figure 4, we see that the majority of Quality data distribution is normal but data points at bottom and top depart from the straight line, which implies some observations are disproportionally influential to the model and we need to take care of them.

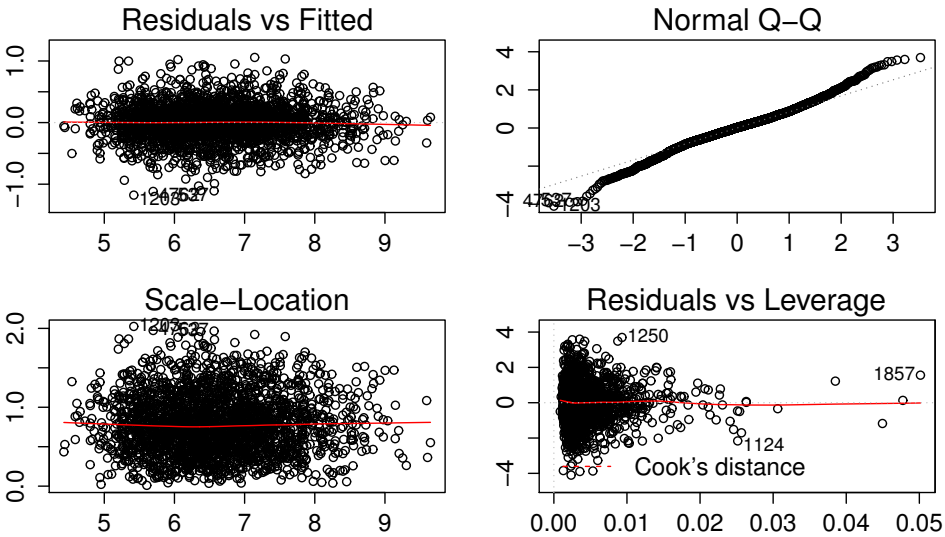

figure 4. diagnostic plots

Bonferroni outlier test labels the 1203th observation in building data set as an outlier. Based on influence plot in figure 5, we eliminate data points with hat value larger than 0.020 and Cook's distance larger than 0.004 as suspicious data points. The following studentized residual plot of the refined model indicates that the situation has been improved and we can rely on the refined model.

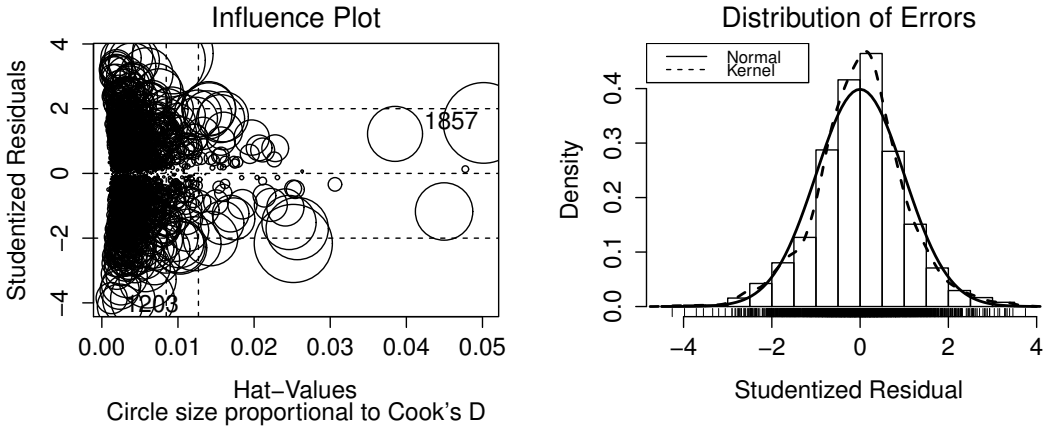

figure 5. influence plot and studentized residual plot

Validation and Prediction: Mean Squared Prediction Error

Table 3 is the summary of refined model.

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$	Estimate (std.)
(Intercept)	-1.697	0.388	-4.378	$1.25 \cdot 10^{-5}$	0
Fixed Acidity	0.378	0.093	4.063	$5.00 \cdot 10^{-5}$	0.757
Volatile Acidity	-1.880	0.063	-29.697	$< 2^{-16}$	-0.205
Residual Sugar	-0.033	0.061	-0.550	0.583	-0.201
pH	0.249	0.044	5.686	$1.46 \cdot 10^{-8}$	0.042
Alcohol	0.360	0.005	66.174	$< 2^{-16}$	0.492
Free Sulfur Dioxide Proportion	1.146	0.063	18.071	$< 2^{-16}$	0.117
Fixed Acidity^2	0.032	0.007	4.930	0.000	2.179
Fixed Acidity $\times$ Residual Sugar	0.011	0.008	1.363	0.173	0.615
Residual Sugar^2	-0.087	0.010	-8.890	$< 2^{-16}$	-3.454

table 3. summary of refined model

We calculate Mean Squared Prediction Error with

$$MSPE = \frac{\sum_{i=1}^{n^*} (Y_i - \hat{Y}_i)^2}{n^*} = 0.85.$$

Also we calculate the prediction intervals for the 5 observations picked out in advance. (They are not included when calculating MSPE.) Because of limited space, please refer to the Appendix for prediction intervals. MSPE and prediction intervals show better performance and validation of the refined model.

Relative Weight

From table 3 above, we see that Fixed Acidity and Residual Sugar have stronger effect on Quality than other factors. Volatile Acidity and Free Sulfur Dioxide Proportion are moderate while pH is relatively weak.

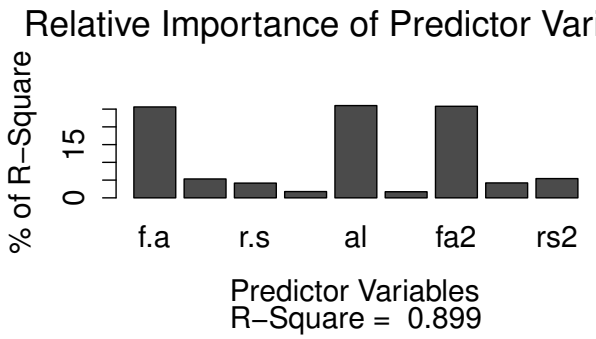

figure 6. bar plot for relative weight of each explanatory variable

From figure 6, Fixed Acidity and Alcohol are the most influential. Volatile Acidity and Residual Sugar are almost at the same level. pH and Free Sulfur Dioxide Proportion have less weight.

From the two methods above, the pivotal role of Fixed Acidity in wine Quality can be sure, just as our priori goes. The effect of pH is limited but stable, which also matches our expectation. Effect of Residual Sugar is more significant than Volatile Acidity, Alcohol and Free Sulfur Dioxide Proportion.

Discussion

Model Discussion

The explanatory variables of final model are presented in table 4.

X_1	Fixed Acidity	X_2	Volatile Acidity	X_3	Residual Sugar
X_4	pH	X_5	Alcohol	X_6	Free Sulfur Dioxide Proportion

table 4. explanatory variables in the final model

Define Y_i as Quality of the ith observation and β as the coefficient vector.

$$X = [1, X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5}, X_{i6}, X_{i1}^2, X_{i1}X_{i3}, X_{i3}^2]^T$$

$$Y_i = \beta X + \varepsilon_i,\ \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

The estimate of β is

$$b = [-1.697, 0.378, -1.880, -0.033, 0.249, 0.360, 1.146, 0.032, 0.011, -0.087].$$

The estimate of coefficients in corresponding standardized model is

$$b' = [0.757, -0.205, -0.201, 0.042, 0.492, 0.117, 2.179, 0.615, -3.454].$$

According to b, we can conclude that

- When other variables are held constant, Quality decreases by 1.880 if Volatile Acidity increases by 1 unit.
- When other variables are held constant, Quality increases by 0.249, 0.360 or 1.457 if pH, Alcohol or Free Sulfur Dioxide Proportion increases by 1 unit.

Fixed Acidity and Residual Sugar are special. We have

$$\frac{\partial Y}{\partial X_1} = 0.378 + 0.064X_1 + 0.011X_3.$$

This implies that how Quality fluctuates given change of Fixed Acidity depends on the Fixed Acidity level itself plus Residual Sugar. The higher level of Fixed Acidity and Residual Sugar makes Quality increase more if Fixed Acidity increases by 1 unit when other variables are held constant. However, is Quality increment guaranteed given increase of Fixed Acidity? Residual Sugar alone is positive but logarithmic transformation makes negative value possible. So, again when the other variables are held constant, Quality will increase if

$$X_1 > \frac{-0.011X_3 - 0.378}{0.064}$$

is satisfied. We do hypothesis testing in Appendix to show that we have 95% confidence to claim this is true for the mean of X_1 and mean of $(-0.011X_3 - 0.378)/0.064$. Therefore, in an average sense, Quality will increase if Fixed Acidity increases by 1 regardless of Residual Sugar when other factors are held constant. This supports the point that Fixed Acidity tends to improve wine Quality, which is not sensitive to minor negative effect from Residual Sugar through interaction. The situation for Residual Sugar is similar. It tends to decrease Quality when other factors are held constant. Please refer to the hypothesis testing and effect plots of figure 8 to 11 in Appendix to see the linear and nonlinear effects.

Reflection

From all the analysis and procedures above, we have learned that 1) Transformation helps to unfold the information in significantly skewed data. This is certainly advantageous to model establishment. However, it also brings difficulty in model interpretation. We hope the model is sticked to what is happening in the real world while transformation is more of a theoretical sense under most circumstances. 2) Outliers have negative impact on model through violation of normal assumption and coefficient estimation. Detecting outliers is not a easy job. A better method combines the subjective elements and objective measurement. 3) Priori, preceding research and even common sense help to build and explain our model. Here, preceding research helps to locate outliers in data preprocessing. Priori from other people helps to select variables and explain coefficients. 4) Almost every method has its own pros and cons. When we have several methods on hand, we have to trade off. In this project, we lay more priority on the exact distribution of coefficient so we choose dropping Density instead of ridge regression. 5) Sometimes a combination of presented variables is more explanatory. For instance, relative index may reveal more information about attribute in some cases than absolute value and vice versa. It is suggested that both of them should be presented to model selection procedure if possible.

There are several ways to improve our analysis. 1) Find a larger data set for both data of less skewed distribution and more predictor variable candidates. Theoretically, data in a valid larger data set will be less skewed so that transformation may not be necessary. More predictor variable candidates will also help to explain the response variable. 2) Investigation in more preceding researches featuring in quantitive method can be done. Although work from others has already helped to finish some job, most of them offer only descriptions and statements and lack mathematical and statistical methodology. Also chemical background in wine production is useful and contributes to better model establishment. 3) If the two steps above can be done, the ridge regression result should be improved. Then we can rely on bootstrap to find the distribution of coefficients in ridge regression. Then compare the two models in terms of explanation and predictive ability and finally choose a better one. 4) We can do more serious coding to visualize dynamic model establishment process. Imagine you can add or delete predictor at will by clicking and dragging. Then the summary and diagnostic plots of the model will be presented automatically. It enables us to build our model in a more convenient way.

References

[1] Doug Nierman. (2004). *Fixed Acidity.* Retrieved From Waterhouse Lab website: http://waterhouse.ucdavis.edu/whats-in-wine/fixed-acidity
[2] Marina Sonegheti COLI, Angelo Gil Pezzini RANGEL, Elizangela Silva SOUZA, Margareth Ferraro OLIVEIRA, Ana Cristina Nascimento CHIARADIA. (2015). Chloride concentration in red wines: influence of terroir and grape type. *Food Science and Technology (Campinas)*, 35(1), 95-99. https://dx.doi.org/10.1590/1678-457X.6493
[3] Vinny. (2009). *What do "pH" and "TA" numbers mean to a wine?.* Retrieved From WineSpectator website: http://www.winespectator.com/drvinny/show/id/5035

Source of Data

Paulo Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009

Associated Publications

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. (2009) Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, Elsevier, 47(4):547-553.

Appendix

Data Preprocessing Complement

- The U.S. legal limits for volatile acidity of white table wine is 1.1 g/L. Threshold for acetic acid which largely composes volatile acidity in wine varies from 600 and 900 mg/L, depending on the variety and style. (Doug Nierman, 2004) We think that volatile acidity ranging from 100 to 1000 mg/L is appropriate and exclude a total of 14 observations.
- No direct reference of chloride concentration is found. limits on maximum chloride concentration vary from country to country. For example, the Brazilian Law establishes the maximum concentration of chlorides in wine at 200 mg/L while it's 607 mg/L in Australia. (Marina Sonegheti COLI, Angelo Gil Pezzini RANGEL, Elizangela Silva SOUZA, Margareth Ferraro OLIVEIRA, Ana Cristina Nascimento CHIARADIA, 2015)
- The observation with residual sugar 65.8 is deleted since it's too far away from the majority.
- The U.S. legal limits of sulfur dioxide is 350 mg/L. (Doug Nierman, 2004) 2 observations are ruled out.
- The observation with density > 1 g/L is deleted since they are too far away from the majority.
- A webpage of WineSpectator tells that most wine pH's fall around 3 or 4 and about 3.0 to 3.4 is desirable for white wines. (Vinny, 2009) The pH level in this data set is appropriate.
- No direct reference of sulphates concentration is found. The data points outside the upper fence are concentrated.
- Quality ranges from 0 to 10. The data points outside this range should be removed.

correlation matrix

	fa	va	ca	rs	ch	fsd	tsd	de	pH	su	al	qu	fsdp	cap
fa	1.000	-0.032	0.289	0.048	0.023	-0.059	0.067	0.247	-0.432	-0.043	-0.109	-0.676	-0.136	-0.077
va	-0.032	1.000	-0.182	0.089	0.075	-0.099	0.084	0.005	-0.025	-0.039	0.064	-0.235	-0.193	-0.188
ca	0.289	-0.182	1.000	0.051	0.069	0.089	0.097	0.123	-0.161	0.055	-0.058	0.018	0.019	0.924
rs	0.049	0.089	0.051	1.000	0.096	0.315	0.408	0.758	-0.172	-0.047	-0.382	-0.055	0.056	0.027
ch	0.023	0.075	0.069	0.096	1.000	0.107	0.223	0.281	-0.084	0.022	-0.364	-0.213	-0.047	0.062
fsd	-0.059	0.099	0.089	0.315	0.107	1.000	0.614	0.307	-0.000	0.055	-0.246	0.025	0.746	0.114
tsd	0.067	0.084	0.097	0.408	0.223	0.614	1.000	0.539	0.016	0.123	-0.452	-0.163	-0.014	0.076
de	0.247	0.005	0.123	0.758	0.282	0.307	0.539	1.000	-0.086	0.055	-0.808	-0.317	-0.063	0.024
pH	-0.432	-0.025	-0.161	-0.172	-0.084	-0.0000	0.016	-0.085	1.000	0.161	0.108	0.092	-0.002	-0.006
su	-0.042	-0.039	0.054	-0.047	0.022	0.055	0.123	0.055	0.162	1.000	-0.008	0.056	-0.018	0.069
al	-0.109	0.064	-0.058	-0.382	-0.364	-0.246	-0.452	-0.808	0.108	-0.008	1.000	0.438	0.074	-0.012
qu	0.676	-0.235	0.020	-0.367	-0.208	-0.126	-0.272	-0.387	0.188	0.056	0.438	1.000	0.071	0.067
fsdp	-0.136	-0.193	0.019	0.056	-0.047	0.746	-0.014	-0.063	-0.002	-0.019	0.074	0.204	1.000	0.072
cap	-0.077	-0.188	0.924	0.027	0.062	0.114	0.076	0.024	-0.007	0.069	-0.012	0.067	0.072	1.000

Model Selection: Data Set Splitting and Exhaustive Subset Selection

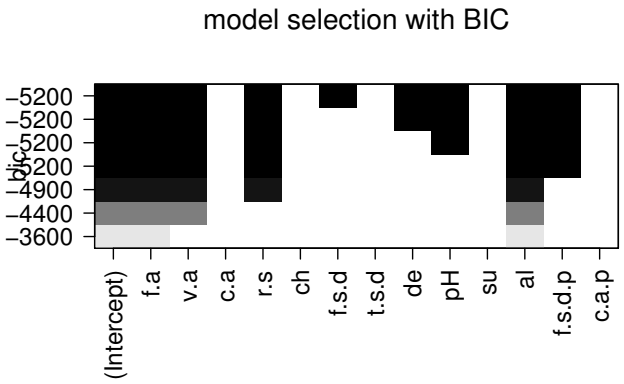

figure 7. exhaustive subset selection under SBC

Parameter Estimation and Inference

Tests for β_k

null hypothesis H_0:

$$\beta_k = 0,$$

the predictor corresponding to β_k is not associated with Quality.
alternative hypothesis H_a:

$$\beta_k \neq 0,$$

the predictor corresponding to β_k is associated with Quality.
test statistic:

$$t^* = \frac{b_k}{s(b_k)}$$

Require $t_{0.975}(2364 - 7) = 1.961$. In table 2, we see $|t^*|$ of each predictor is larger than 1.961, along with an extremely small $p - value$, so we conclude H_a, $\beta_k \neq 0$ for each predictor. This means the selected predictors are associated with Quality.

F Test for Regression Relation

null hypothesis H_0:

$$\beta_1 = \beta_2 = \cdots = \beta_6 = 0,$$

there is no general regression relation between the selected predictors and Quality.
alternative hypothesis H_a:

not all β_k equal to 0,

there is general regression relation between the selected predictors and Quality.
test statistic:

$$F^* = \frac{MSR}{MSE}$$

Require $F_{0.95}(6, 2357) = 2.102$. We use the test statistic $F^* = 3271 > 2.102$ to conclude H_a, that is the existence of a regression relation.

F Test for Lack of Fit

null hypothesis H_0:

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6,$$

the current model is complex enough to explain Quality.
alternative hypothesis H_a:

$$E(Y) \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6,$$

the current model is not complex enough to explain Quality.
test statistic:

$$F^* = \frac{SSLF}{c-p} / \frac{SSPE}{n-c}$$

Here $n = 2364, p = 7, c = 1972$. The test statistic is $F^* = (203.20 - 29.80)/(1972 - 7)/0.076 = 1.161$. Requiring $F_{0.95}(1965, 392) = 1.141$, we have $F^* > F$ and claim that the current model is not complex enough to explain Quality.

Curvature and Interaction Investigation

Coefficient	Estimate	Std. Error	t value	$Pr(>\|t\|)$
Fixed Acidity	0.371	0.2833	1.310	0.190
Volatile Acidity	-1.837	0.063	-29.249	$< 2^{-16}$
Residual Sugar	-0.398	0.201	-1.993	0.046
pH	-1.501	1.638	-0.917	0.359
alcohol	0.358	0.006	63.671	$< 2^{-16}$
free sulfur dioxide proportion	1.145	0.065	17.646	$< 2^{-16}$
Fixed Acidity^2	0.028	0.007	3.782	0.000
pH^2	0.236	0.210	1.121	0.263
Residual Sugar^2	-0.088	0.010	-8.903	$< 2^{-16}$
Fixed Acidity × pH	0.017	0.064	0.263	0.793
Fixed Acidity × Residual Sugar	0.020	0.009	2.177	0.030
Residual Sugar × pH	0.097	0.051	1.905	0.057

table 5. model with all second-order terms involving Fixed Acidity, Residual Sugar and pH

F Test for Lack of Fit

null hypothesis H_0 :

$$E(Y) = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \beta_{11} X_1^2 + \beta_{13} X_3^2 + \beta_{33} X_3^2,$$

the current model is complex enough to explain Quality.
alternative hypothesis H_a:

$$E(Y) \neq \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \beta_{11} X_1^2 + \beta_{13} X_3^2 + \beta_{33} X_3^2,$$

the current model is not complex enough to explain Quality.
test statistic:

$$F^* = \frac{SSLF}{c-p} / \frac{SSPE}{n-c}$$

Here $n = 2364, p = 9, c = 2117$. The test statistic is $F^* = (201.39 - 19.30)/(2117 - 9)/0.078 = 1.107$. Requiring $F_{0.95}(2108, 247) = 1.176$, we have $F^* \leq F$ and claim that the current model is complex enough to explain Quality. The square of Fixed Acidity, square of Residual Sugar and interaction between Fixed Acidity and Residual Sugar help to explain Quality.

Validation and Prediction: Mean Squared Prediction Error

Obs. ID	Quality	Fitted Value	Prediction Interval
4674	6.336	6.456	[5.912, 7.000]
1896	6.967	7.020	[6.476, 7.565]
2229	6.721	6.397	[5.852, 6.941]
1074	6.806	7.041	[6.496, 7.585]
3229	6.559	6.098	[5.554, 6.642]

table 6. prediction intervals

Hypothesis Testing 1

Define μ_1 as the mean of X_1, μ_3 as the mean of $(-0.011X_3 - 0.378)/0.064$. Now we care about the mean of population so Central Limit Theorem comes to help.
null hypothesis H_0:

$$\mu_1 - \mu_3 = 0$$

alternative hypothesis H_a:

$$\mu_1 - \mu_3 > 0$$

test statistic:

$$T = \frac{\bar{X_1} - \bar{X_3}}{\sqrt{s_1^2/n_1 + s_3^2/n_3}} = 1063.038$$

degree of freedom:

$$df = \frac{(s_1^2/n_1 + s_3^2/n_3)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_3^2/n_3)^2/(n_3 - 1)} \approx 5063$$

Require $t_{0.95}(5063) = 1.645$. Since $T > t_{0.95}(5063)$, we conclude H_a.

Hypothesis Testing 2

As to

$$\frac{\partial Y}{\partial X_3} = -0.033 - 0.174X_3 + 0.011X_1,$$

we want to show that we have 95% confidence to claim the mean of $(0.174X_3 + 0.033)/0.011$ is larger than the mean of X_1. Define μ_1 as the mean of X_1, μ_3 as the mean of $(0.174X_3 + 0.033)/0.011$.
null hypothesis H_0 :

$$\mu_3 - \mu_1 = 0$$

alternative hypothesis H_a :

$$\mu_3 - \mu_1 > 0$$

test statistic:

$$T = \frac{\bar{X_3} - \bar{X_1}}{\sqrt{s_1^2/n_1 + s_3^2/n_3}} = 91.926$$

degree of freedom:

$$df = \frac{(s_1^2/n_1 + s_3^2/n_3)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_3^2/n_3)^2/(n_3 - 1)} \approx 4756$$

Require $t_{0.95}(4756) = 1.645$. Since $T > t_{0.95}(4756)$, we conclude H_a.

Model Discussion

Here are the effect plots of Volatile Acidity, Fixed Acidity, interaction of Fixed Acidity and Residual Sugar and Residual Sugar. Residual Sugar increases from 1.2 to 1.6 in the first three plots. Fixed Acidity increases from 6.5 to 6.9 in the last plot. We can see that effect of Volatile Acidity is linear. Effect of Fixed Acidity, Residual Sugar and their interaction is curvilinear.

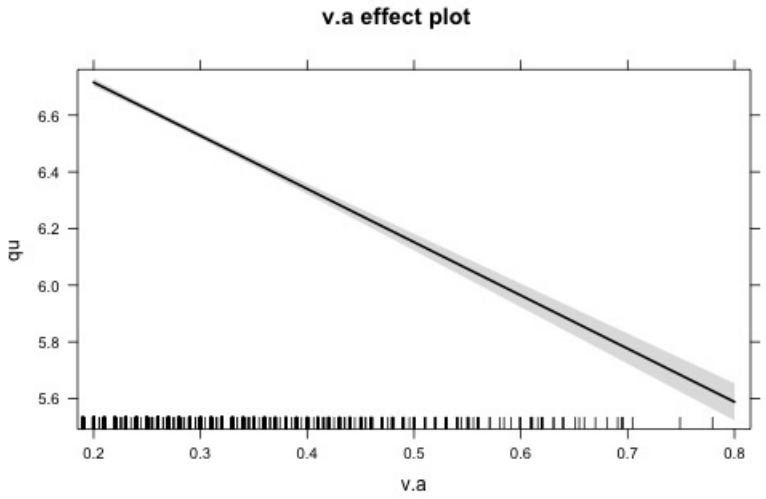

figure 8. effect plot of Volatile Acidity

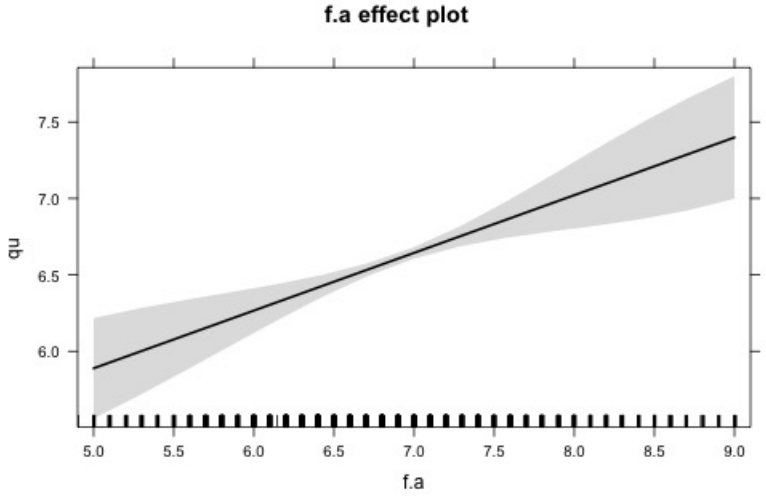

figure 9. effect plot of Fixed Acidity

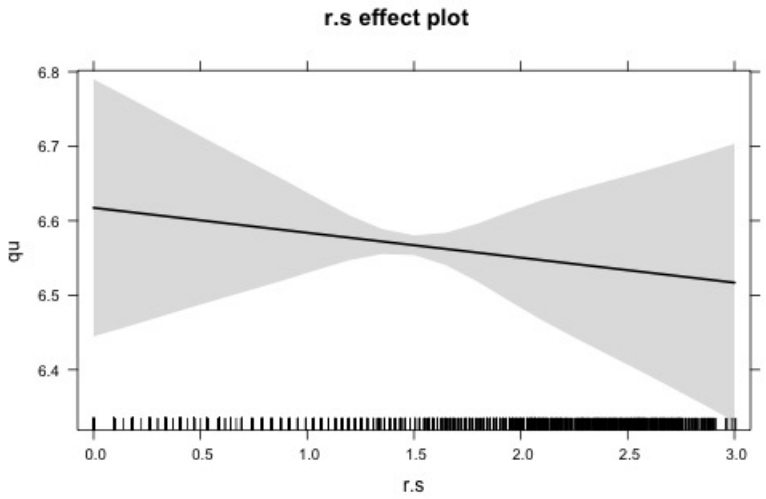

figure 10. effect plot of Residual Sugar

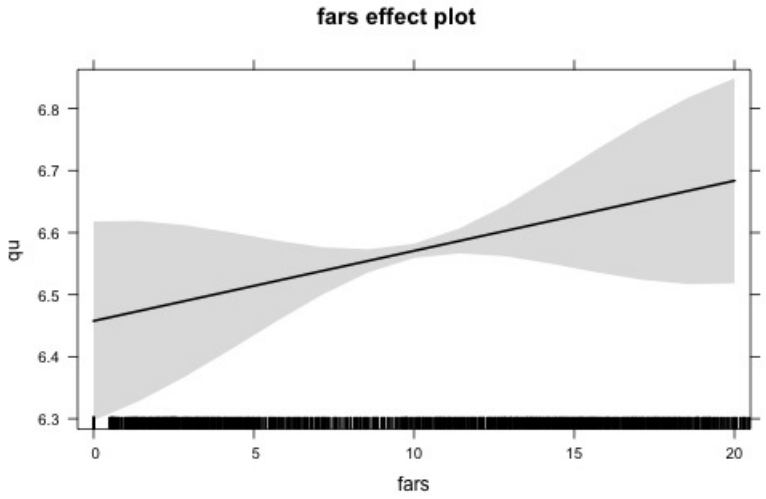

figure 11. effect plot of interaction between Fixed Acidity and Residual Sugar

Statement of contributions

Weiqi Weng is the only author and did all the job.