Assessing the Impact of Physicochemical Properties on Red Wine Quality: A Multivariate (OLS and Lasso) Regression and Bayesian Approach

Jukun Zhang[1]

[1] Rutgers University

Abstract

This study mainly uses three models-OLS, Lasso regression and Bayes model. After preliminary exploratory analysis (descriptive statistics, box plots, correlation heat maps, paired scatter plots), OLS can explain 36% of the variance (adjusted R^2 = 0.356), thus determining that alcohol, volatile acid, sulfate, chloride, pH, free sulfur dioxide and total sulfur dioxide are important predictors. Lasso regression reduces 3 variables by setting the L1 penalty coefficient. Although the model is further optimized, the results are basically the same as OLS. However, the Bayes model extracts more accurate prediction variables through 95% confidence interval and KDE test, and believes that the positive impact of wine quality comes from free SO (FSO2), sulfates (Sulphates) and alcohol (Alcohol), and the negative impact comes from volatile acid (VolAcid), chloride (Chlor), and total SO (TSO2). Finally, the data_wine_tidy data is binarized, and Logistic regression is used to verify that the prediction accuracy of the final confirmed 6 variables for the quality of red wine is 0.818, indicating that these six variables can provide a certain quantitative basis for the prediction of red wine quality.

Keywords: OLS; Lasso Regression; Bayes Regression; Logistics Regression
Word count: X

Correspondence concerning this article should be addressed to Jukun Zhang. E-mail: jz1250@scarletmail.rutgers.edu

2 Introduction & Literature Review

The quality of red wine not only affects consumers' tasting experience and health perception, but is also directly related to the improvement direction of grape planting and brewing technology and the optimization of the industrial value chain. With the continuous advancement of analytical instruments and data acquisition technology, researchers can obtain a series of physical and chemical indicators including fixed acidity, volatile acidity, citric acid content, residual sugar, chloride, free sulfur dioxide, total sulfur dioxide, density, pH value, sulfate and alcohol content. These indicators have their own characteristics in different regions, different years and different brewing process conditions, and have a complex and multidimensional impact on the sensory score (quality) of the final wine.

Traditional research mainly uses ordinary least squares (OLS) regression to evaluate the linear relationship between various physical and chemical properties and scores. Some scholars combine ridge regression or principal component analysis to deal with multicollinearity problems. In recent years, Lasso regression has been widely used in high-dimensional data modeling because of its variable selection and regularization capabilities; while Bayesian regression provides a framework for uncertainty quantification for parameter estimation through prior distribution and posterior inference. However, there are few existing literatures that systematically compare the model performance and variable importance differences of OLS, Lasso and Bayesian methods on the same data set.

Based on the red wine quality data set (n = 1,599) provided by UCI, this study uses OLS, Lasso and Bayesian regression as analysis tools. Through descriptive statistics, visual exploration and cross-validation, this study comprehensively evaluates the relative impact of various physical and chemical properties on wine quality scores, and compares the advantages and disadvantages of the three methods in terms of prediction accuracy and variable explanatory power, providing a quantitative basis for red wine quality control and optimization.

3 Data and Exploratory Analysis

3.1 Data Source & Pre-processing

My data source was find from “UCI Red Wine Quality” link. It has 1599 data points and I choose “Quality” as the response and other 11 variables as predictors which include fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chloride, free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulfate and alcohol. After preprocessing the data, I first read the raw data as *data_wine_raw*, and then checked whether there were missing values in the raw data. I used “anyNA(data_wine_raw)” and “colSums(is.na(data_wine_raw))” to test whether there were missing values and which column of variables had missing values. The test results showed that the data was very intact and there were no missing values. The next step was to continue to abbreviate the column names to make it easier to represent in the output image and make the results easier to read. Part of the data_wine_tidy are show here:

Table 1
data_wine_tidy (Part vars)

FixAcid	VolAcid	Sugar	Chlor	Alc	Qual
7.4	0.70	1.9	0.076	9.4	5
7.8	0.88	2.6	0.098	9.8	5
7.8	0.76	2.3	0.092	9.8	5
11.2	0.28	1.9	0.075	9.8	6
7.4	0.70	1.9	0.076	9.4	5
7.4	0.66	1.8	0.075	9.4	5

3.2 Descriptive Statistics

Then I exported Five-Number Summaries to further analyze the data.

The Five-Number Summaries (Table 2) highlight the central tendency and variance of each physicochemical variable in the red wine dataset. For most acidity measurements (fixed acidity, volatile acidity, citric acid), the interquartile ranges (IQRs) were moderate, about 2.1 units for fixed acidity and 0.25 units for volatile acidity, indicating a fairly consistent distribution of acidity across samples, although the maximum values (15.9 for fixed acidity and 1.58 for volatile acidity) indicate a right skew and a high acidity outlier. Sugar content was clearly positively skewed. The median sugar content was only 2.2 g/L, but the maximum reached 15.5 g/L, reflecting the small number of very sweet wines. Chloride and pH show narrow IQRs (0.02 g/L and 0.19 pH units, respectively), indicating strict quality control over salinity and acidity balance. Density was fairly constant (only 0.996-0.998 from Q1-Q3), confirming little variation in ethanol to water ratio. Sulfate (IQR = 0.18 g/L) and alcohol (IQR = 1.6 % v/v) both show moderate distributions, with the upper quartile (11.1 %) and maximum value (14.9 %) of alcohol highlighting that wines with higher alcohol content may have higher quality scores. Finally, total SO and free SO showed the largest absolute ranges (TSO up to 289 mg/L, FSO up to 72 mg/L), indicating different SO conservation strategies, while the mass itself ranged from 3 to 8, but was concentrated at 5 to 6 (IQR = 1). Combining these summaries allows us to predict which predictors contribute to quality changes and which nonlinear effects or outliers require further investigation.

Table 2

Table 2. Five-Number Summaries for All Variables

variable	Min	Q1	Median	Q3	Max
FixAcid	4.60	7.10	7.90	9.20	15.90
VolAcid	0.12	0.39	0.52	0.64	1.58
CitAcid	0.00	0.09	0.26	0.42	1.00
Sugar	0.90	1.90	2.20	2.60	15.50
Chlor	0.01	0.07	0.08	0.09	0.61
FSO2	1.00	7.00	14.00	21.00	72.00
TSO2	6.00	22.00	38.00	62.00	289.00
Dens	0.99	1.00	1.00	1.00	1.00
pH	2.74	3.21	3.31	3.40	4.01
Sulph	0.33	0.55	0.62	0.73	2.00
Alc	8.40	9.50	10.20	11.10	14.90
Qual	3.00	5.00	6.00	6.00	8.00

3.3 Visualization & Correlation

After completing the above mean and median analysis, I output boxplots, heat maps, and Pairwise Scatterplot Matrix of some variables and wine quality, hoping to use these visualizations to analyze the correlation between each variable and response (quality) and check whether there is multicollinearity. The output images are as follows:

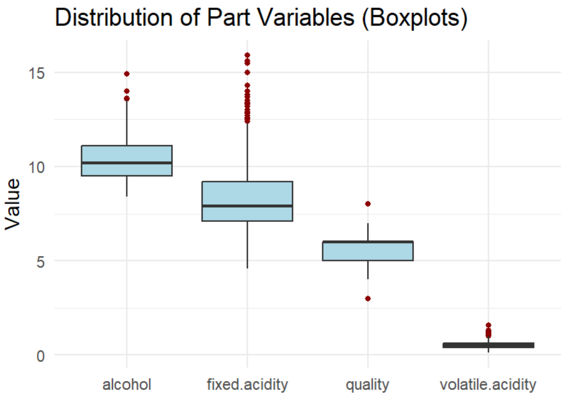

Figure 1. Boxplot of part Variables

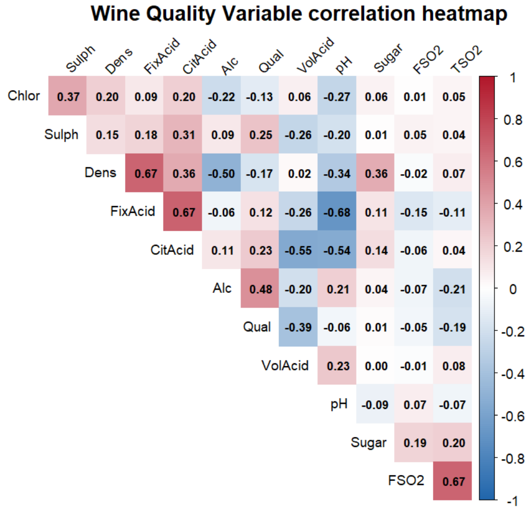

Figure 2. Heatmap of correlations between variables

Figure 2 shows the Pearson correlation coefficient matrix of all eleven variables and the dependent variable wine quality. From the figure, we can see that alcohol and taste show a strong positive correlation (r +0.48), which indicates that wines with higher alcohol content tend to get higher quality scores. On the other hand, volatile acidity is significantly negatively correlated with health (r -0.39), indicating that the higher the volatile acidity, the lower the quality score of the wine. We also found that there may be some multicollinearity between the predictors - free SO2 and total SO2 (r = +0.67) and fixed acidity and density (r = +0.67). In short, based on this figure, we have a deeper understanding of the data, and the conclusions we get will help guide the subsequent OLS and Lasso regression methods.

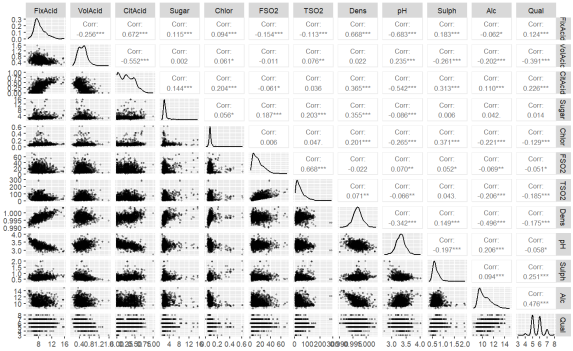

Figure 3. Pairwise Scatterplot Matrix for all variables

Figure 3 uses GGally::ggpairs to output the Pairwise Scatterplot Matrix, where the lower left corner of the diagonal outputs the scatter plots between the variables. From the scatter plots of the variables, we can see that the points of Fixed Acidity and Density are dense and distributed in an upward diagonal, reflecting the conclusion drawn from the heat map that the two variables may be highly collinear. The output in the middle of the diagonal is the kernel density estimates of each variable's marginal distribution, and the output in the upper right corner of the diagonal is the same Pearson correlation coefficient as the heat map. The smooth density curves of each variable reveal the distribution characteristics: for example, Sugar shows an obvious right-skewed long-tail distribution, while Density is highly concentrated in the range of 0.997–0.998. This matrix not only further confirms the known strong associations such as Sulphates and Quality, but also once again shows the multicollinearity problem and the abnormal marginal distribution of some variables, providing a key basis for variable screening and diagnosis of subsequent regression models.

4 Methodology & Results

4.1 Ordinary Least Squares Model (OLS)

I first used the OLS model to linearly model the dependent variable wine quality for all 11 variables. The regression equation is as follows:

$$\mathrm{Qual}_i = \beta_0 + \beta_1 \,\mathrm{FixAcid}_i + \beta_2 \,\mathrm{VolAcid}_i + \cdots + \beta_{11} \,\mathrm{Alc}_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2). \tag{1}$$

Hypothesis tests are performed on the parameters to test the significance of the parameters. The results can be obtained from the P-value column in Table 3. For each regression coefficient β_j, test:

Null Hypothesis(H_0) : $\beta_j = 0$ vs Alternative Hypothesis(H_a) : at least one $\beta_j \neq 0$

Calculate the t statistic $t_j = \hat{\beta}_j/\mathrm{SE}(\hat{\beta}_j)$ and report the two-tailed p value. The results show:

Table 3
OLS regression coefficient estimation and significance test

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	21.97	21.19	1.04	0.30
FixAcid	0.02	0.03	0.96	0.34
VolAcid	-1.08	0.12	-8.95	0.00
CitAcid	-0.18	0.15	-1.24	0.21
Sugar	0.02	0.02	1.09	0.28
Chlor	-1.87	0.42	-4.47	0.00
FSO2	0.00	0.00	2.01	0.04
TSO2	0.00	0.00	-4.48	0.00
Dens	-17.88	21.63	-0.83	0.41
pH	-0.41	0.19	-2.16	0.03
Sulph	0.92	0.11	8.01	0.00
Alc	0.28	0.03	10.43	0.00

Among them, we can find that the parameters of volatile acidity, chloride, total SO_2, free SO_2, pH, sulphates and alcohol are significant in the model, while other variables are not significant at the 5% level. We can preliminarily conclude that these variables shown as significant above have a certain effect on the quality score of red wine. The fitting results of the OLS model show that the coefficient of determination $R^2 = 0.3606$ indicates that the model can explain about 36.1% of the total variance of the quality score. However, after adjustment, $R^2 = 0.3561$, and the F statistic of the model = 81.35 (df = 11, 1587), $p < 2.2 \times 10^{-16}$, rejecting the null hypothesis, indicating that the model is significant.

After that, the OLS model residuals were further tested using the diagnosis() function in the library (ds4ling). The following is the output of the OLS residual model diagnostic graph:

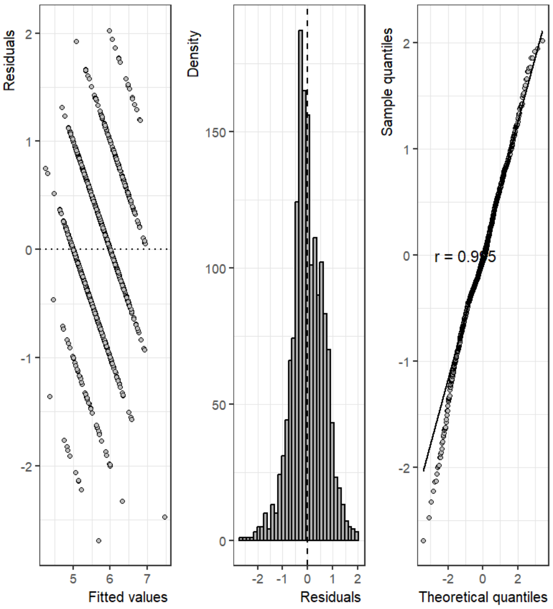

Figure 4. Residual diagnosis for OLS model

The residuals vs. fitted values plot on the left of the test plot shows randomly distributed scatter with no particularly obvious pattern, so the assumption of linearity can be considered supported. However, since the variance seems to increase with the fitted values, there may be slight heteroskedasticity. The Q-Q plot on the right shows that the residuals are roughly distributed along the diagonal (r = 0.95), indicating approximate normality. However, the slight deviation in the tails shows that there may be slight non-normality. The residual density plot in the middle is centered at zero and has a single peak, which is similar to the residual density plot of a good model fit, and is bell-shaped, but a slight skewness is observed.

4.2 Lasso Regression

To mitigate multicollinearity and enforce sparsity in our predictor set, we fit a Lasso model using the **glmnet** package. In matrix notation, letting x_i denote the p-dimensional predictor vector for observation i and y_i its quality score, the Lasso estimate solves

$$\min_{\beta_0,\beta} \frac{1}{N} \sum_{i=1}^{N} \bigl(y_i - \beta_0 - x_i^\top \beta\bigr)^2 \; + \; \lambda \sum_{j=1}^{p} |\beta_j|, \quad \lambda \geq 0. \tag{2}$$

Prior to fitting, all predictors were standardized and an intercept term retained. ### Selection via 10-Fold Cross-Validation Using the glmnet() function, with the penalty parameter α set to 1, we can get a complete Lasso model. We continue to set the response to the quality of the wine, and the dependent variable is the remaining 11 variables. We hope to use the Lasso model to delete as many unimportant variables as possible, making the model parameters simpler - that is, to get a sparse model. This way, we can more clearly understand which physical and chemical properties (variables) have a more significant impact on the quality of the wine.We then output a plot of $Log(\lambda)$ and Mean Square Error to help determine the optimal λ value.

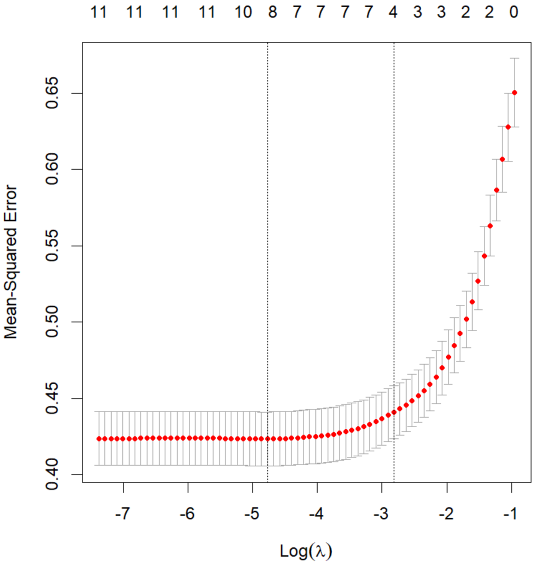

Figure 5. Plot of CV_Lasso

The figure shows that as $Log(\lambda)$ increases from -7 to -1, the MSE of the model also increases. To prevent overfitting, we believe that $Log(\lambda)$ = -3 is the best value, because reducing $Log(\lambda)$ further will penalize the model too much and delete some necessary parameters. The number at the top (11 to 0) indicates the number of factors with non-zero coefficients. As λ increases, fewer predictors are selected, thereby reducing the complexity of the model. At this optimal λ, Lasso retains the following 8 non-zero coefficients.

Compute the cross-validation MSE for a sequence of λ values. The value (denoted as λ_{min}) that minimizes the CV error is found to be. $\hat{\lambda}_{min} = 0.00848$

Table 4

Non-zero Lasso coefficients at $\hat{\lambda}_{\text{min}}$

term	estimate
(Intercept)	4.1628
VolAcid	-1.0243
Sugar	0.0013
Chlor	-1.7093
FSO2	0.0024
TSO2	-0.0027
pH	-0.3825
Sulph	0.8196
Alc	0.2852

Three predictors fixed acidity, citric acid, and density are driven exactly to zero, indicating minimal marginal contribution once penalization is applied. The retained coefficients largely mirror OLS signs and magnitudes, reaffirming volatile acidity and chlorides as the negative drivers of quality and sulphates and alcohol content as positive drivers.

Afterwards, I used 10-fold cross validation to compare the prediction performance of OLS and Lasso models.

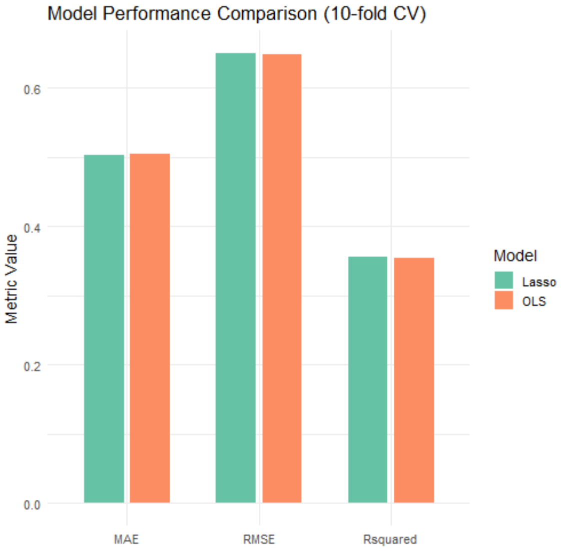

Figure 6. OLS and Lasso Model Performance Comparison

From the comparison of the 10-fold cross validation of the two models, it can be seen that LASSO and OLS perform similarly in predicting wine quality, with little difference in MAE, RMSE, and R^2.