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

DS4ling Final Paper

Jukun Zhang

2025-05-09 (updated: 2025-05-09)

# 1 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 R² = 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 SO₂ (FSO2), sulfates (Sulphates) and alcohol (Alcohol), and the negative impact comes from volatile acid (VolAcid), chloride (Chlor), and total SO₂ (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.

**Key words:** OLS; Lasso Regression; Bayes Regression; Logistics Regression

# 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](https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-red.csv). 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:

Table1 : ata\_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.

Table2 : 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 |

The Five-Number Summaries (Table 1) 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 SO₂ and free SO₂ showed the largest absolute ranges (TSO₂ up to 289 mg/L, FSO₂ up to 72 mg/L), indicating different SO₂ 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.