Analysis of Different Properties in the Assessment of Red Wine Quality

Zhehan (Andrew) Shi
Mathematical Statistics
New York University
May 13, 2021

1 Introduction

What makes red wine good? Wine consumption is on the rise due to quarantine measures related to the COVID-19 pandemic. There are many red wine connoisseurs globally; however, it is extremely difficult for ordinary people to discern if a glass of red wine is good or bad without reference to either price or brand. It would be a subject of interest to know how the different chemical properties of red wine influence its quality.

The analysis involved the wine data[1] provided by UCI Machine Learning Repository, which sourced the data from a study done by Dr. Cortez[2]. The red wine was of a variant of Portuguese wine, *Vinho Verde*. Due to the reason of confidentiality, only physicochemical properties were provided. The dataset had 1599 rows, and it contained 12 different fields, 11 of them were different chemical properties, including alcohol, residual sugar, acidity, and others, and the remaining one was the quality score. Quality score was an ordinal variable with a possible value ranging from 1 (worst) to 10 (best). This analysis focused on the red wine dataset, which would be simply referred to as the dataset later in the article.

This project involved many ideas, including but not limited to ideas from the class, such as ordinary least squares (OLS) regression and Pearson correlation, and new techniques, such as principal component analysis (PCA).

2 Summary Statistics

The quality score came from the previously mentioned Portuguese wine. It should be emphasized that taste was the least understood of the human senses; therefore,

there would involve some arbitrariness of wine experts' judgment of the wine quality.

The following were the summary statistics of different fields for the red wine dataset.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	8.319637	0.527821	0.270976	2.538806	0.087467	15.874922
std	1.741096	0.179060	0.194801	1.409928	0.047065	10.460157
min	4.600000	0.120000	0.000000	0.900000	0.012000	1.000000
25%	7.100000	0.390000	0.090000	1.900000	0.070000	7.000000
50%	7.900000	0.520000	0.260000	2.200000	0.079000	14.000000
75%	9.200000	0.640000	0.420000	2.600000	0.090000	21.000000
max	15.900000	1.580000	1.000000	15.500000	0.611000	72.000000

Figure 1: Statistics Summary

	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000	1599.000000
mean	46.467792	0.996747	3.311113	0.658149	10.422983	5.636023
std	32.895324	0.001887	0.154386	0.169507	1.065668	0.807569
min	6.000000	0.990070	2.740000	0.330000	8.400000	3.000000
25%	22.000000	0.995600	3.210000	0.550000	9.500000	5.000000
50%	38.000000	0.996750	3.310000	0.620000	10.200000	6.000000
75%	62.000000	0.997835	3.400000	0.730000	11.100000	6.000000
max	289.000000	1.003690	4.010000	2.000000	14.900000	8.000000

Figure 2: Statistics Summary

The following was the visual representation of the histogram for different fields.

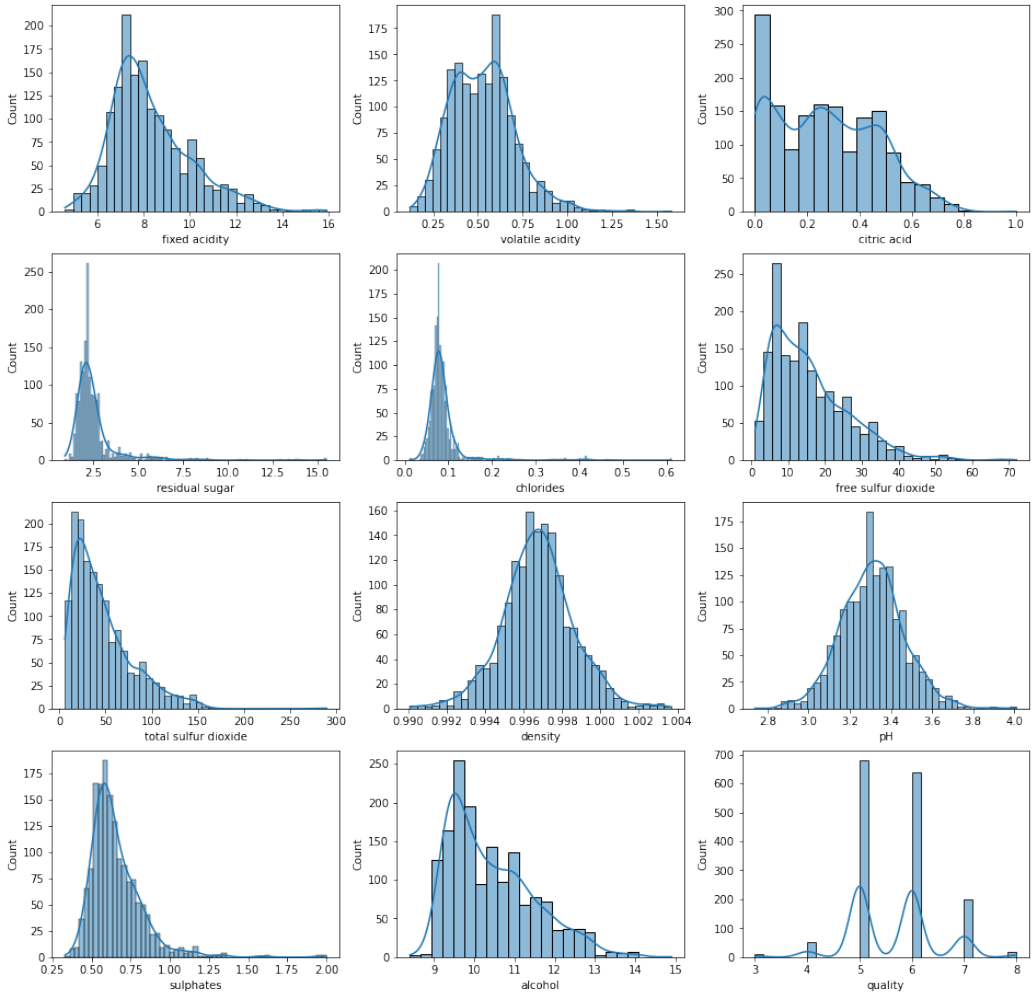

Figure 3: Histogram

3 Multiple Linear Regression

3.1 Model Selection

It was reasonable to assume that dissimilar fields contribute to the overall quality of the wine. I attempted to use a static regression model to model the relationship

between the individual fields and the quality of the wine. For simplicity, I made an assumption that there existed a linear relationship between the predictors and the response variables. I utilized use ordinary least squares regression (OLS) for such a model.

Similar to other models, a multiple linear model was only appropriate if its underlying assumptions were valid; therefore, all the regression conditions would be verified to ensure the validity of the regression analysis.

3.2 Formula Review

For linear model, the regression equation was $Y_i = \sum_{j=1}^{k} \beta_j X_{ij} + \epsilon$. The outcome vector is $Y = X\beta + \epsilon$. β is the coefficient vector, ϵ is the residual vector and X is the covariate matrix. The outcome vector could also be written as the expression of residual vector, $\epsilon = X\beta - Y$.

$$\begin{aligned} ||\epsilon||^2 &= ||X\beta - Y||^2 \\ &= (X\beta - Y)^T (X\beta - Y) \\ &= Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta \end{aligned}$$

$$\begin{aligned} \frac{\partial L}{\partial \beta} &= \frac{\partial (Y^T Y - Y^T X\beta - \beta^T X^T Y + \beta^T X^T X\beta)}{\partial \beta} \\ &= -2X^T Y + 2X^T X\beta \end{aligned}$$

$$\begin{aligned} -2X^T Y + 2X^T X\beta &= 0 \\ X^T Y &= X^T X\beta \end{aligned}$$

The least squares estimate is the following

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

3.3 Applied Regression

The multiple linear regression was applied using the formula above and verified by a built-in regression function from scikit-learn, a Python library.

Fields	Coefficients
intercept	21.965208
fixed acidity	0.024991
volatile acidity	-1.083590
citric acid	-0.182564
residual sugar	0.016331
chlorides	-1.874225
free sulfur dioxide	0.004361
total sulfur dioxide	-0.003265
density	-17.881164
pH	-0.413653
sulphates	0.916334
alcohol	0.276198

Table 1: **coefficients**

Furthermore, the r-squared for the regression was 0.3605517030386881. The response, quality, could be calculated as the following equation using other predictor variables.

$$\text{Quality} = \text{intercept} + \beta_1 \times \text{fixed acidity} + \beta_2 \times \text{volatile acidity} + ... + \beta_{11} \times \text{alcohol}$$

The following was a partial visual representation between the predicted value and actual value. There were 1599 predictions, and it was not feasible to show all of them.

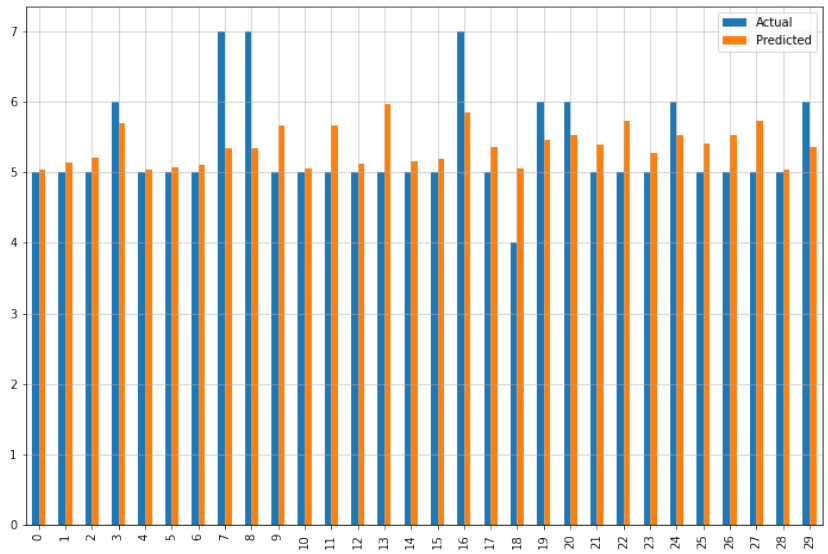

Figure 4: Bar Graph

3.4 Regression Conditions Verification

For a multiple linear regression to be valid, the model needed to meet four conditions, **linearity**, **normality**, **homoscedasticity** and **no multicollinearity**.

3.4.1 Linearity

A linear model worked as long as the underlying assumptions were met. It was therefore imperative to verify if the model was linear.

A simple linear model was not difficult to check; however, this was not the case for multiple linear regression. In order to solve this, I calculated the residual value, which was the difference between the predicted responses and actual responses. Then, I plotted the residual value against the actual responses to verify if the multiple linear regression had linearity. The following graph was the result, the x-axis was for residual value, and the y axis for predicated responses. The actual responses were discrete, whereas the predicted responses were continuous. It was not hard to see from the scatter plot that the relationship was not linear.

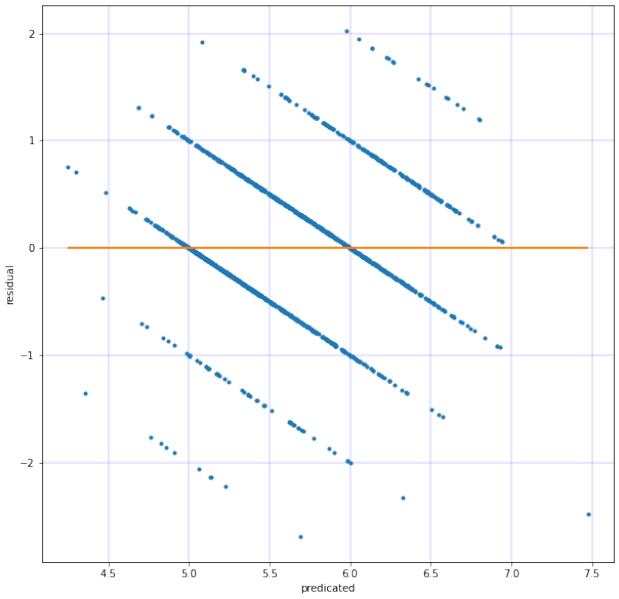

Figure 5: Between Residual Value and Predicted Responses

3.4.2 Homoscedasticity

Homoscedasticity referred to a condition in which the variance of the error term in a regression model was constant. It was also known as equal variances.

Figure 5 could also be used to check for the existence of homoscedasticity. Given different levels of variance at residual value, there did not exist homoscedasticity but heteroscedasticity.

3.4.3 Normality

Shapiro-Wilk Test

$$W = \frac{(\sum_{i=1}^{n} a_i e_{(i)})^2}{\sum_{i=1}^{n}(e_i - \bar{e})^2}$$

The e_i pertains to the i^{th} largest value of the error terms and the a_i values are calculated using the means, variances, and covariances of the e_i. W was compared against tabulated values of this statistic's distribution. Small values of W would lead us to reject the null hypothesis.

The p-value was $1.9549368346361007e^{-08}$, which was significantly less than 0.05. The null hypothesis was rejected. The errors did not follow a normal distribution.

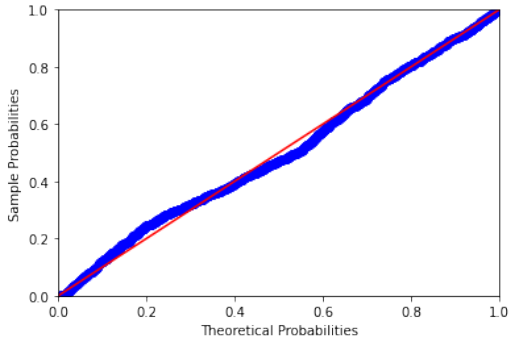

Figure 6: Normal Probability Plot

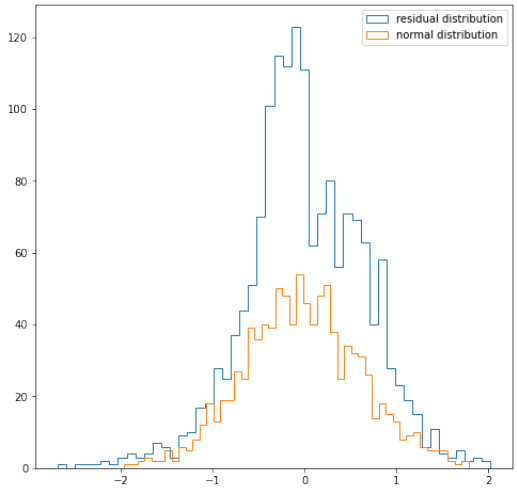

Figure 7: Comparison

The orange line of figure 7 was generated using the same mean and standard deviation as the residuals. The side-by-side graph comparison confirmed that the residual distribution was not normal. Despite the fact that Normal Probability Plot might trick the untrained eyes into thinking that the distribution of residuals was normal, it was not. Moreover, both the normality test and figure 7 could show that residuals were not normally distributed.

3.4.4 Multicollinearity

In order to detect multicollinearity, variance inflation factor (VIF)[4] was used.

$$VIF_i = \frac{1}{1-R_i^2}$$

If the predictors were correlated with each other, the standard errors of the coefficient estimates would be larger than if the predictors were uncorrelated. R_i^2 was a statistical measure that represented the proportion of the variance for a dependent variable that was explained by an independent variable or variables in a regression model.

Feature	VIF Factor
fixed acidity	74.452265
volatile acidity	17.060026
citric acid	9.183495
residual sugar	4.662992
chlorides	6.554877
free sulfur dioxide	6.442682
total sulfur dioxide	6.519699
density	1479.287209
pH	1070.967685
sulphates	21.590621
alcohol	124.394866

Table 2: **Variance Inflation Factors**

According to statistical tradition, VIF factor should be less than 5 to prove there was no multicollinearity. The fields in the dataset clearly had multicollinearity;

Pearson Correlation

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

The r means correlation coefficient. x_i and y_i mean the value of x-variable and y-variable = values of the x-variable in a sample. $\bar{x}$ and $\bar{y}$ means the average respectively.

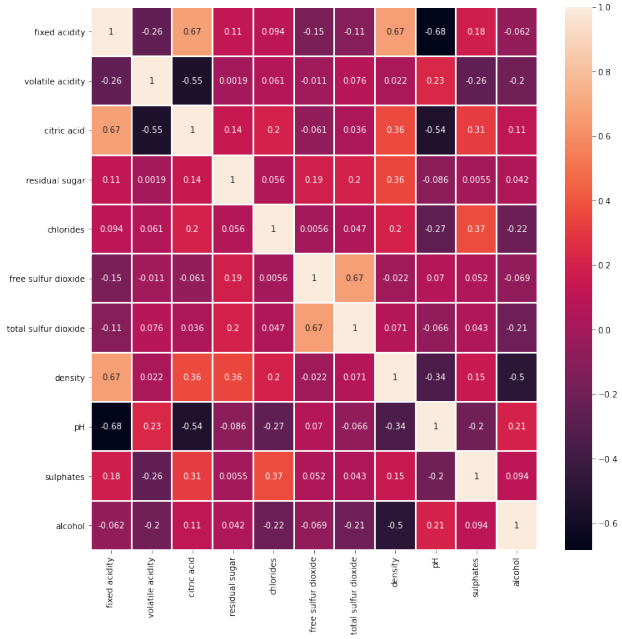

Figure 8: Correlation Matrix

In conclusion, the multiple linear regression relied on four assumptions, **linearity**, **normality**, **homoscedasticity** and **no multicollinearity**. All assumptions were not valid. The verification of each assumption proved that this dataset was not suitable for multiple linear regression. Therefore, the previously constructed multilinear regression model was **invalid**.

4 Principal Component Analysis

The main idea of principal component analysis (PCA)[3] was to reduce the dimensionality of a dataset made of many interrelated variables while keeping as much as possible of the variation present in the dataset.

The following graph showed the influence of the different number of components in determining red wine quality.

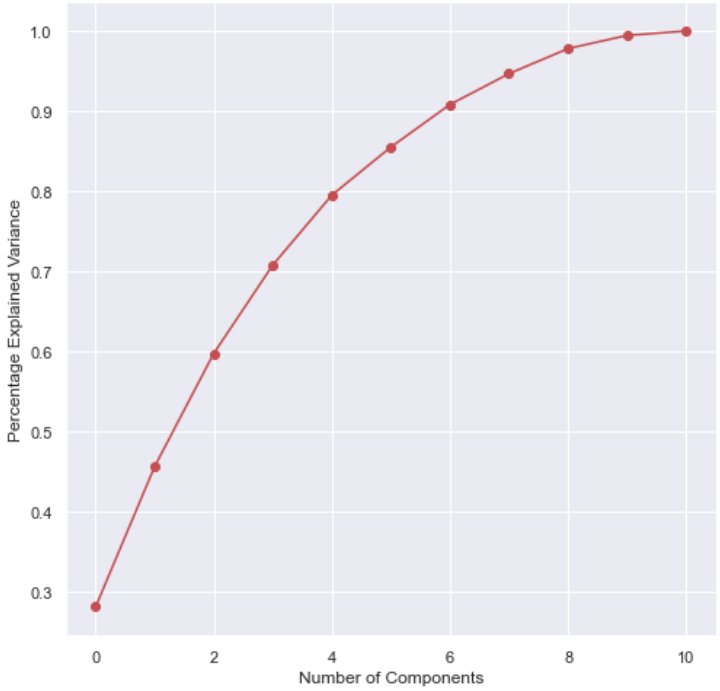

Figure 9: Principal Component Analysis

It could be seen from graph above that 6 components explained 90% of the variance.

	Acidity	Sulfides	More alcohol	Chlorides	More residual sugar	Less pH
0	-1.619530	0.450950	-1.774454	0.043740	0.067014	-0.913921
1	-0.799170	1.856553	-0.911690	0.548066	-0.018392	0.929714
2	-0.748479	0.882039	-1.171394	0.411021	-0.043531	0.401473
3	2.357673	-0.269976	0.243489	-0.928450	-1.499149	-0.131017
4	-1.619530	0.450950	-1.774454	0.043740	0.067014	-0.913921
5	-1.583707	0.569195	-1.538286	0.023750	-0.110076	-0.993626
6	-1.101464	0.608015	-1.075915	-0.343959	-1.133382	0.175000
7	-2.248708	-0.416835	-0.986837	-0.001203	-0.780435	0.286057
8	-1.086887	-0.308569	-1.518150	0.003315	-0.226727	-0.512634
9	0.654790	1.665207	1.209476	-0.824635	1.718501	-0.476497

They were weighted linear combinations of the original variables, for example: $\text{PC1} = (\text{fixed acidity} \times 0.489314) + (\text{volatile acidity} \times -0.238584) + \ldots + (\text{alcohol} \times -0.113232)$

5 Metrics and Further Exploration

5.1 R-squared for Multiple Regression

Even if the multiple regression was valid, the r-squared value was 0.360551703 0386881, which was not significant enough. It basically meant that this multiple linear regression model could only explain about 36.1% of the variation in the response variable around its mean. Not to mention this regression model was faulty due to the reason that none of the four assumptions was sufficiently met.

5.2 K-Nearest Neighbor Attempt

Both Lasso and Ridge regressions had the same assumptions as those of the multiple linear regression. As stated earlier, all these assumptions were not met; therefore, it was not advisable to employ either Lasso or Ridge regression.

Traditional regression techniques failed disastrously; however, other statistical learning methods were promising, and one of them was K-Nearest Neighbor(KNN). In order to build a functioning model, I attempted the technique by following the tutorial[5]. In the KNN method, I used 6 weighted components from PCA to achieve 78.2% accuracy. This result was better than using the same KNN model with 3 neighbors with 11 original predictor variables, which ultimately resulted in 75.5% accuracy. For more details, please check out the code[6].

6 Conclusion

It was indispensable to verify the assumptions of multiple linear regression models. Multiple linear regression might ostensibly achieve satisfactory results, which were shown in figure 4, Bar Graph; however, this was built on shaky foundations. In order to evaluate the red wine dataset more thoroughly, it was paramount to explore other statistical methods to take advantage of these characteristics inherent in the dataset. Furthermore, the red wine quality score was discrete, whereas the predicated value calculated from the faulty regression model was continuous. This difference once again elucidated the limitation of multiple linear regression. More potential studies could be done to make the most use of the dataset. It was hard to pin down precisely the feature that contributed the most to quality; however, PCA did shed light on the number of features needed to capture the variation of wine quality score.

References

[1] Wine Dataset from UCI
https://archive.ics.uci.edu/ml/datasets/wine+quality

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009. https://www.sciencedirect.com/science/article/abs/pii/S0167923609001377

[3] 11.1 - Principal Component Analysis (PCA) Procedure — STAT 505. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat505/lesson/11/11.1

[4] 12.4 - Detecting Multicollinearity Using Variance Inflation Factors — STAT 501. (n.d.). PennState: Statistics Online Courses. https://online.stat.psu.edu/stat501/lesson/12/12.4

[5] KNN Algorithm - Finding Nearest Neighbors - Tutorialspoint. (n.d.). Tutorialspoint. https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm

[6] My Code in Github
https://github.com/cyberzzhhss/statistical_analysis_on_wine/blob/master/red_wine_statistical_analysis.ipynb