Project Report: Wine Quality Dataset

Abstract

In this document, we present our investigation of the relationship between a set of physicochemical data and wine quality ratings. Wines are widely consumed products in the United States, and the wine industry contributes about $161 billion to the economy in 2007. An efficient and objective quality control process is very important for the wine industry. Wine quality is mainly measured by a limited number of wine experts, and hence the quality scoring process is both costly and subjective. This motivated us to investigate the relationship between measured physicochemical attributes and wine quality, to see the possibility of designing a wine quality estimation procedure.

We use a dataset based on the physicochemical measurements on the the Portuguese Vinho Verde wines and their quality score. We first analyze the properties of the dataset via several descriptive statistics. We then perform multiple linear regression on the data dataset under the assumption that there is a linear relationship between the wine quality scores and physicochemical factors. From the hypothesis testing for ANOVA, we verify that at least one attribute is relevant. We also assess the adequacy of our linear regression model based on the coefficient of determination (R^2), and it turns out that the R^2 statistic is relative low. In addition, the resulting residuals are highly correlated with the wine quality scores.

In order to increase the quality of our estimation, we tried several other linear regression models. We found a more complicated model involving linear, logarithmic and polynomial terms and their pair-wise interactions, which led to 561 parameters. This complicated model exhibits about twice the R^2 value and its resulting residuals are less correlated with the wine quality scores.

Contents

1 Motivation **2**

2 Dataset **3**

3 Analyses **4**
3.1 Descriptive Statistics 4
3.2 Multiple Linear Regression 5
3.3 A More Complicated MLR Model 8

4 Conclusion **14**
4.1 Discussion 14
4.2 Future Work 15

Chapter 1

Motivation

Wine is a widely consumed product. It can be found in from simple dinner-tables to luxurious settings. In 2010, total wine consumption in the United States was about 784 million gallons [2]. This translates roughly to 2.54 gallons of wine consumed per resident in 2010, which is up from 2.5 gallons the previous year. Wine industry is also a big economical driver, employing 1.1 Million full-time equivalent jobs and contributing $161 Billion to U.S. economy in 2007 [4].

For an industry with such impact, quality control is an important issue. It is difficult to produce wine with consistent quality. This quality is impacted by many factors coming into play during pre-production, production and post-production such as the weather, production methodology, humidity and temperature. The wine quality measurement is mostly done by wine tasting experts which require extensive training. Hence, wine quality assessment is both costly and subjective.

Wine experts have identified some physicochemical factors for wines, which may be able to help determine wine quality. If these features and their effects on the quality can be identified, the quality measurement process would be made more objective and the need for a human expert would be reduced (or eliminated all together). In this project, we are interested investigating the relationship between measured physicochemical factors of wines and their quality ratings as evaluated by the human experts.

Chapter 2

Dataset

To investigate the factors related to wine quality, we chose the 'Wine Quality' dataset [3] which is composed of two datasets related to red and white wines of the Portuguese Vinho Verde [1]. The dataset contains 1599 and 4898 instances of red and white wine measurements, respectively. There are 11 continuous input attributes which were measured via physicochemical tests:

- fixed acidity
- volatile acidity
- citric acid
- residual sugar
- chlorides
- free sulfur dioxide
- total sulfur dioxide
- density
- pH
- sulphates
- alcohol

and one discrete output attribute of quality scored by wine experts, where its value ranges from 0 (poor) to 10 (excellent).

Chapter 3

Analyses

3.1 Descriptive Statistics

Before performing more complex analyses, we first examine and visualize our dataset with several descriptive statistics.

Table 3.1: Sample means, standard deviations, and ranges of the wine dataset

Attributes	**Red wine**				**White wine**			
	Mean	**Std**	**Min**	**Max**	**Mean**	**Std**	**Min**	**Max**
Fixed acidity	8.32	1.74	4.60	15.90	6.85	0.84	3.80	14.20
Volatile acidity	0.53	0.18	0.12	1.58	0.28	0.10	0.08	1.10
Citric acid	0.27	0.19	0.00	1.00	0.33	0.12	0.00	1.66
Residual sugar	2.54	1.41	0.90	15.50	6.39	5.07	0.60	65.80
Chlorides	0.09	0.05	0.01	0.61	0.05	0.02	0.01	0.35
Free sulfur dioxide	15.87	10.46	1.0	72.0	35.31	17.01	2.0	289.0
Total sulfur dioxide	46.47	32.90	6.0	289.0	138.36	42.50	9.0	440.0
Density	1.00	0.00	0.99	1.00	0.99	0.00	0.99	1.04
pH	3.31	0.15	2.74	4.01	3.19	0.15	2.72	3.82
Sulphates	0.66	0.17	0.33	2.00	0.49	0.11	0.22	1.08
Alcohol	10.42	1.07	8.40	14.90	10.51	1.23	8.00	14.20
Quality score	5.64	0.81	3.00	8.00	5.88	0.89	3.00	9.00

Table 3.1 presents sample means, standard deviations, and min-max ranges of the physicochemical attributes as well as quality score. From the sample means and standard deviations, we can easily notice that the attributes are not standardized. Although our main task is not determining whether a given wine is red or white, some of attributes are quite distinguishable between red and white wines. Especially, the levels of free and total sulfur dioxide give good clues for the discrimination.

Our main task is to estimate the quality score of a wine, given its attributes. Thus it is of interest to see the distribution of the quality scores. From the mean values of both white and red wines, the centers of distributions are between 5 and 6. The standard deviations also imply that most of scores are centered around the means. This analysis is often more clear by plotting histograms. The left two plots in Figure 3.1 show the histograms of the red and white dataset. As

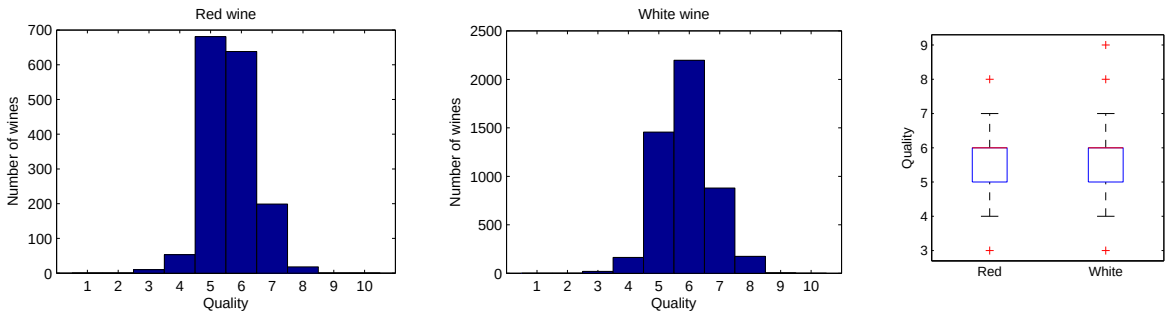

Figure 3.1: **Histograms and boxplot for the red and white wine quality scores.**

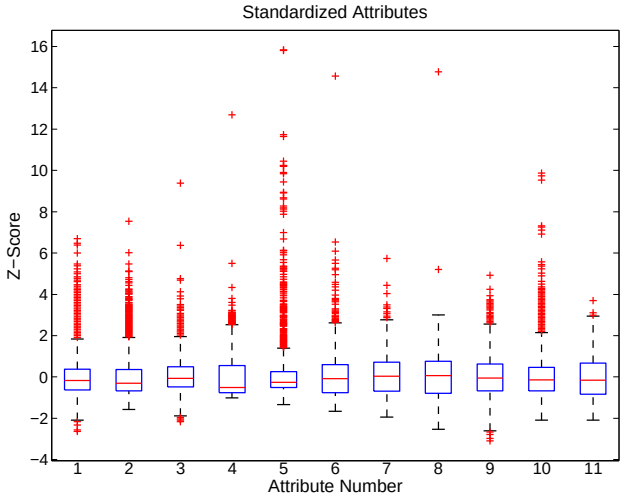

Figure 3.2: **Boxplot of the physicochemical attributes in the dataset.**

we can see, both histograms are symmetric and centered between 5 and 6 scores. It is also worth to note that the number of white wine samples (4898) is significantly more than the number of red wine samples (1599). The right plot in Figure 3.1 presents the boxplot of the wine quality scores. According to the boxplot, the median of both red and white wines are 6. The outliers are also clearly shown; 3 and 8 scores are outliers for both dataset; 9 score is also another outlier for the white wine dataset.

We also look at the boxplots of the attribute data, depicted in Figure 3.2. We pool the data for the red and white wines together for visualization purposes. We standardize the data to be able to compare them at the same scale. From the figure, it is clear that none of the attributes are normally distributed, which individual normal plots also confirm.

3.2 Multiple Linear Regression

Our main task is to find a relationship between the physicochemoical attributes of the wines in our dataset and their quality. As a first step, we run the standard multiple linear regression (MLR) on

our dataset. The matrix formulation for the MLR is as follows

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.1}$$

where $\mathbf{y} \in \mathbb{R}^{\mathsf{N}}$ is the response variable, $\mathbf{X} \in \mathbb{R}^{\mathsf{N}\times 12}$ is the regressor matrix, $\boldsymbol{\beta} \in \mathbb{R}^{12}$ is the intercept and slope, and $\boldsymbol{\epsilon} \in \mathbb{R}^{\mathsf{N}}$ is the random noise, where N is the number of observations. Note that $\mathbf{y}$ correspond to wine quality scores and the matrix $\mathbf{X}$ correspond to the concatenation of the measured physicochemoical attributes.

Recall that the wine quality scores are actually discrete, and can take the values between 0 and 10, although it is between 3 and 9 in our dataset. However, we are going to assume that they are continuous and apply the linear regression framework.

From the Eq. 3.1, the resulting least squares estimate that minimizes the sum of squared residuals is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}. \tag{3.2}$$

By the Eq. 3.2, the regression coefficients both for red and white wine dataset are obtained as follows

$$\hat{\boldsymbol{\beta}}^0_{red} = (21.97, 0.03, -1.08, -0.18, 0.02, -1.87, 0.00, -0.00, -17.88, -0.41, 0.92, 0.28)^{\mathsf{T}} \tag{3.3}$$

$$\hat{\boldsymbol{\beta}}^0_{white} = (150.19, 0.07, -1.86, 0.02, 0.08, -0.25, 0.00, -0.00, -150.28, 0.69, 0.63, 0.19)^{\mathsf{T}}. \tag{3.4}$$

As we mentioned in Section 3.1, attributes of our dataset were not standardized, and hence it is not possible to compare the computed regression coefficients. For the comparison, we first need to standardize our dataset. The regression coefficients computed after standardizing are as follows

$$\hat{\boldsymbol{\beta}}_{red} = (5.64, 0.04, -0.19, -0.04, 0.02, -0.09, 0.05, -0.11, -0.03, -0.06, 0.16, 0.29)^{\mathsf{T}} \tag{3.5}$$

$$\hat{\boldsymbol{\beta}}_{white} = (5.88, 0.06, -0.19, 0.00, 0.41, -0.01, 0.06, -0.01, -0.45, 0.10, 0.07, 0.24)^{\mathsf{T}}. \tag{3.6}$$

From the absolute value of the estimated coefficients, we can see that there is relationship, which can be partially be described as linear, between the attributes and the score. To make sure, we do the test for significance of regression. The hypothesis test for ANOVA is

$$\mathbf{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = \cdots = \boldsymbol{\beta}_k = 0 \tag{3.7}$$

$$\mathbf{H}_1 : \boldsymbol{\beta}_j \neq 0 \text{ for at least one } j \tag{3.8}$$

where $k = 11$ in our case. We know that the test statistic for ANOVA is

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} \tag{3.9}$$

$$= \frac{MS_R}{MS_E}. \tag{3.10}$$

If we estimate the test statistics for both wines, we get

$$F_0^{red} = 81.3479 > F_{0.05,11,1599-12} = 1.7947 \tag{3.11}$$

$$F_0^{white} = 174.3441 > F_{0.05,11,4898-12} = 1.7906. \tag{3.12}$$

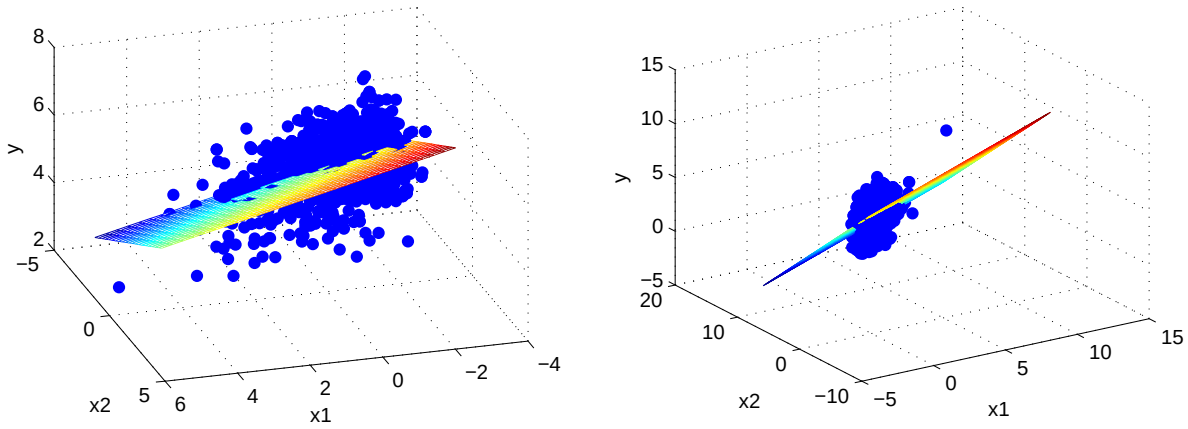

Figure 3.3: **Scatter plots for the red and white wine quality dataset given two dominant coefficients as described in the text.**

We reject the null hypothesis, and thus we can verify that the wine quality score is related to at least one of the attributes.

It is worth mentioning that the bigger the absolute value of the coefficient in Eq. 3.6 (the coefficients after standardization), the more significant contribution it has to the final response value, assuming that the attributes are normally distributed. Thus we can choose dominant attributes that largely determine the wine quality. The second and eleventh attributes are dominant attributes for the red wines, which correspond to *volatile acidity* and *alcohol*. The fourth and eighth attributes are important for the white wines which correspond to *residual sugar* and *density*. Figure 3.3 shows scatter plots with the estimated regression model only for the two most dominant attributes. The resulting regression model is shown as a colored plane. We can see that there is a large variation between the scores and the fitted model.

When we use the multiple linear regression, it is useful to estimate a confidence interval for each coefficient of the regression model $\boldsymbol{\beta}_j$. The 95% ($\alpha = 0.05$) confidence interval on the regression coefficients $\boldsymbol{\beta}_j$ is given by the Table 3.2. We can see that 7 coefficients for the red wine and 8 coefficients for the white wine (in addition to the slope) are significant. The *citric acid* attribute is not significant for both of the wines. We do not get into details of which attributes are important or not since, as will be evident shortly, the regression is not adequate.

When we look at the Table 3.2, we see that coefficients of some of the attributes are significantly different than each other, hence, pooling together the data and doing regression with a dummy variable is not suitable and different estimators are needed for each wine type.

Note that we calculated the confidence intervals of the coefficients using the t-distribution without first establishing that the residuals are normally distributed. However, we have enough number of data points such that the distribution of the coefficients can be approximated as a normal distribution by the central limit theorem and that the t-distribution approaches to the normal distribution as the degrees of freedom gets large.

We are also interested in assessing the adequacy of our linear regression model. The coefficient

Table 3.2: The confidence intervals for the coefficients of the regression and their significance. The first 3 columns are for the red wine and the other 3 columns are for the white wine. The last column represents if the two attributes are significantly different than each other or not.

Coefficients	**Red wine**			**White wine**			**Different**
	Lower	**Upper**	**Significant**	**Lower**	**Upper**	**Significant**	
$\boldsymbol{\beta_1}$	5.60	5.67	Yes	5.86	5.90	Yes	Yes
$\boldsymbol{\beta_2}$	-0.04	0.13	No	0.02	0.09	Yes	No
$\boldsymbol{\beta_3}$	-0.24	-0.15	Yes	-0.21	-0.17	Yes	No
$\boldsymbol{\beta_4}$	-0.09	0.02	No	-0.02	0.02	No	No
$\boldsymbol{\beta_5}$	-0.02	0.06	No	0.34	0.49	Yes	Yes
$\boldsymbol{\beta_6}$	-0.13	-0.05	Yes	-0.03	0.02	No	Yes
$\boldsymbol{\beta_7}$	No	0.09	Yes	0.04	0.09	Yes	No
$\boldsymbol{\beta_8}$	-0.15	-0.06	Yes	-0.04	0.02	No	Yes
$\boldsymbol{\beta_9}$	-0.11	0.05	No	-0.56	-0.34	Yes	Yes
$\boldsymbol{\beta_{10}}$	-0.12	-0.01	Yes	0.07	0.13	Yes	Yes
$\boldsymbol{\beta_{11}}$	0.12	0.19	Yes	0.05	0.09	Yes	Yes
$\boldsymbol{\beta_{12}}$	0.24	0.35	Yes	0.18	0.30	Yes	No

of determination (R^2) is often used as a global statistic to evaluate the fit of the model as

$$R^2 = 1 - \frac{SS_E}{SS_T} \tag{3.13}$$

$$R^2_{red} = 0.3606 \tag{3.14}$$

$$R^2_{white} = 0.2819. \tag{3.15}$$

From the R^2_{red} and R^2_{white} values, we can state that about 36% and 28% of variability in the data can be accounted for by the regression model for red and white wines respectively. Note that sicne the number of data points is large and the number of coefficients are low, the adjusted R^2 values are virtually the same as non-adjusted ones.

We also look at the residuals resulting from this model to test the assumptions of the regression. In Figure 3.4, we plot the fitted wine scores versus the residual values, obtained from the equation $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. In Figure 3.5, we plot the histogram, fitted normal distribution and the normal plots of the residuals. When we look at these plots, the errors do not look like they have constant variance and that they are not normally distributed (especially at the tails). Their correlation with the fitted values can be argued. If we take a deeper look at the data and the residuals, we can see correlation with some of the attributes. We conclude that none of our assumptions for the regression are properly satisfied.

3.3 A More Complicated MLR Model

Given that our R^2 values are low and that the regression assumptions are violated with the simple model, we look for a more complicated model that might be able to perform better. Upon several trial and errors, we selected a model that give us good results. Our selected model is:

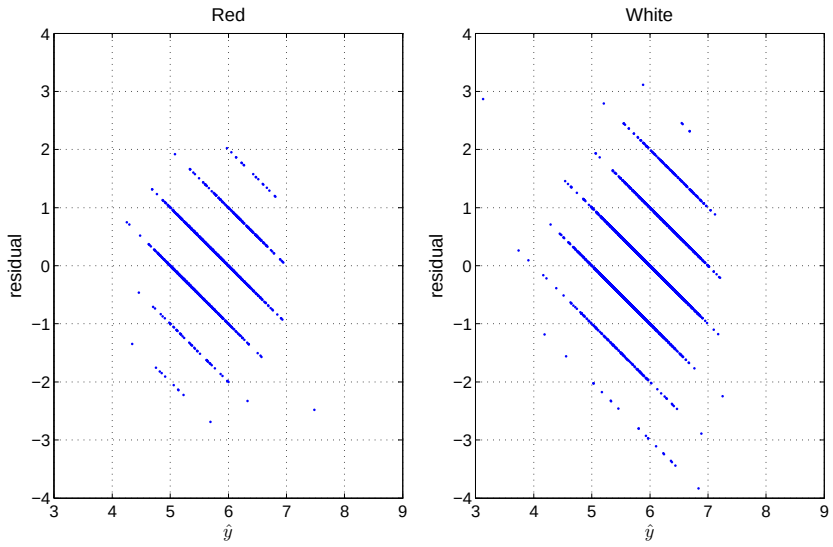

Figure 3.4: **Scatter plots for the fitted values and the residuals.**

$$\mathbf{y} = \boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta}_1 + ln(\mathbf{X})\boldsymbol{\beta}_2 + \mathbf{X}^{\sqrt{2}}\boldsymbol{\beta}_3 + \mathbf{X}_{IM}\boldsymbol{\beta}_4 + \boldsymbol{\epsilon} \tag{3.16}$$

Where, all the function operations are elementwise, $\boldsymbol{\beta}_i$'s are vectors and the matrix $\mathbf{X_{IM}}$ is constructed from the pairwise multiplication of the linear, logarithmic and the polynomial terms. $\mathbf{X}$ has 11 columns and together with the logarithmic and the polynomial terms, they result in 33 coefficients. The pairwise multiplication results in $(33^2 - 33)/2$ additional coefficients. When we add the constant column, we get a model with $33 + (33^2 - 33)/2 + 1 = 562$ coefficients, as compared to the 12 in the previous model.

To increase the performance and numerical stability, we map all the attributes (before forming the regressor matrix) between $[\gamma, 1 + \gamma]$ where γ is a small value (taken as 0.01) by using the following equation:

$$\hat{\mathbf{x}_\mathbf{i}} = \frac{(\mathbf{x_i} - min(\mathbf{x_i})) + \boldsymbol{\gamma}}{max(\mathbf{x_i}) - min(\mathbf{x_i})} \tag{3.17}$$

Furthermore, we reject outliers based on a post-regression analysis. The data points that result in residuals that lie outside the $[Q_{25} - 1.5IQR, Q_{75} + 1.5IQR]$ range, where Q_i is the i^{th} percentile data point and $IQR = Q_{75} - Q_{25}$, are rejected and the regression is run again[1].

The main purpose of the effort is to find a good estimation model, instead of analyzing the relationship between attributes and the wine quality score. Hence, we are only going to report R^2 values and do the residual analysis, skipping the details about the resulting coefficients. Figure 3.6 shows the normal and adjusted R^2 values for both of the mlr models. The complicated model has roughly twice the performance of the simple model. An improvement was expected, since the number of coefficients increased by a factor of about 45. A potential problem with the model is overfitting, given the number of coefficients.

[1] We continue this until there is no change or 3 repetitions are reached.

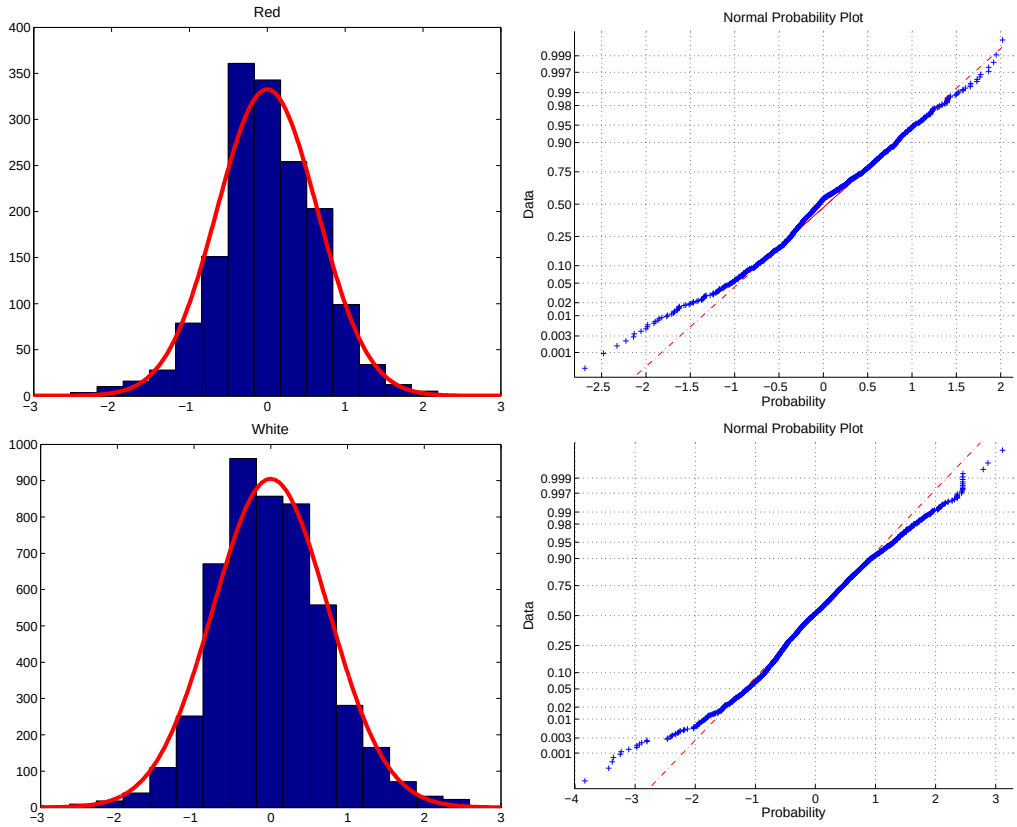

Figure 3.5: **The histogram (left) and the normal plots (right) of the residuals of the regression. The top row corresponds to red wine and the bottom row corresponds to white wine. The red curves in the histogram plots correspond to the fitted and scaled normal distributions.**

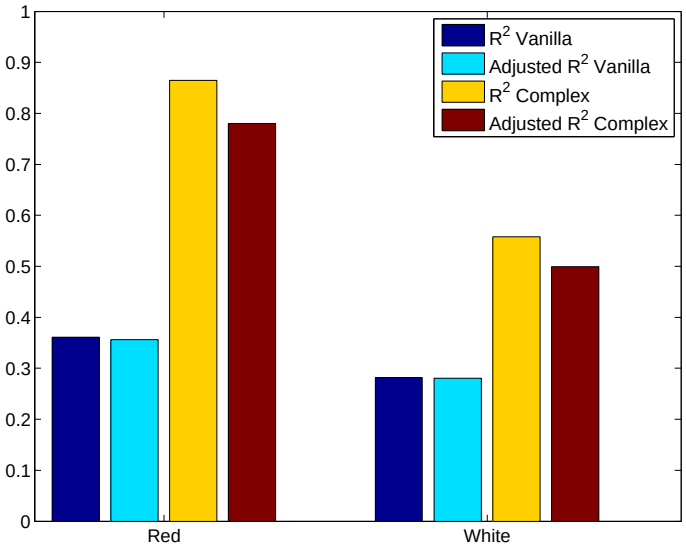

Figure 3.6: **The bar plot of the R^2 values from both of the models. Vanilla refers to the simple model described in the previous section.**

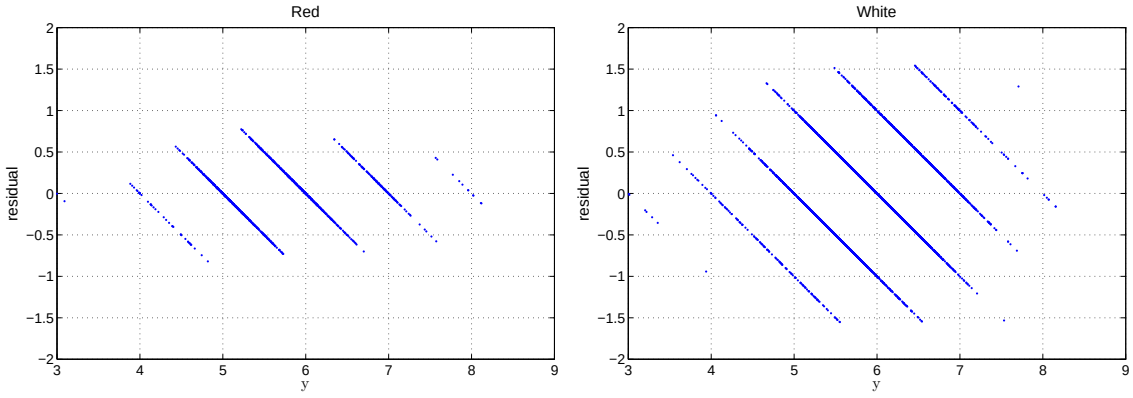

Figure 3.7: **Scatter plots for the fitted values and the residuals.**

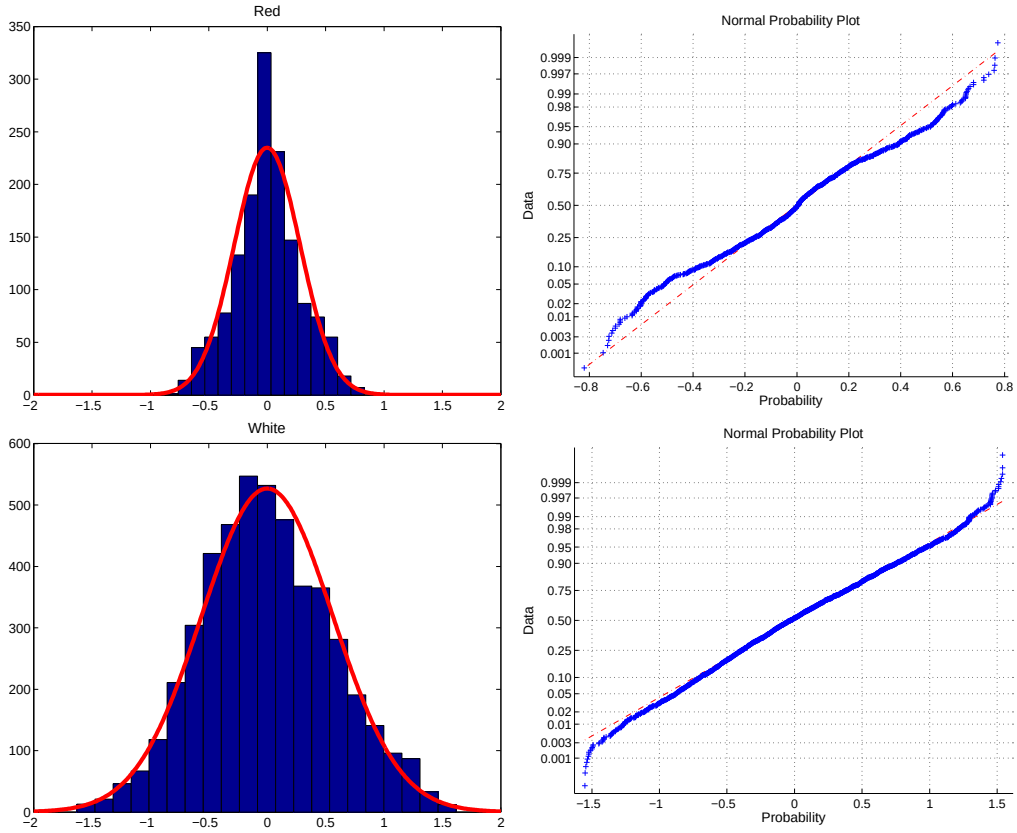

Figure 3.8: **The histogram (left) and the normal plots (right) of the residuals of the regression. The top row corresponds to red wine and the bottom row corresponds to white wine. The red curves in the histogram plots correspond to the fitted and scaled normal distributions.**

The Figure 3.7, depicts the fitted wine scores versus the residual values, obtained from the equation and the Figure 3.8, depicts the histogram and the normal plots of the residuals. These figures look somewhat better. Overall the residuals are decreased. The residuals for the white wine look more normal whereas the residuals for the red wine look more tamed. Even tough we get much better R^2 values, especially for the red wine, it is difficult to say that this estimator is a good one, the assumptions are still not obeyed. However, for practical purposes, this model seems to be usable.

Chapter 4

Conclusion

4.1 Discussion

We started out with a highly noisy and non-normally distributed data. We applied multiple linear regression using a simple model and found out that our regression was not adequate and the assumptions are violated. We then tried several different models and settled on a more complex model involving nonlinear functions and interaction terms. The regression adequacy, especially for the red wine, was much better. However, the assumptions were still flaky and there is a risk of over-fitting due to the large number of regression parameters. Moreover, such a model does not give us any insight into the relationship between the attributes and the quality score.

We think there are several reasons behind the seemingly bad results we got:

- Our data was noisy and unevenly distributed from the start and simple linear regression was not the best approach.
- Needed a highly complex model to get a reasonable result, hence, the relationship is probably highly complex and nonlinear.
- The attributes were not good enough, at least for a linear model.
- We needed more attributes (i.e., more information).

We want to elaborate on some of the items. The relationship between physicochemoical attributes and a single wine score is probably complex and nonlinear and cannot be captured by a simple model. Since we are not domain experts, we do not know how each attribute interact with each other and do not have insight into how they affect the quality score. A domain expert would help in coming up with a more adequate (e.g., through experience or chemistry) model, instead of the brute force approach we employed.

More information about the data collection process would be helpful. It is quiet possible that several wine testers were used to measure the quality score hence they may not be consistent across each other. Moreover, additional information about the wines such as date of production, color, clarity etc. might be helpful. Discrete but important information such as the date of production can also be used in an ANOVA analysis to further give insight.

Our main conclusion from this study is that the real world problems are hard and do not fit the assumptions very well. It is important to understand the problem and limitations of the applied

techniques and look for other methods that might be helpful, accordingly. Moreover, real world problems need domain expertise, multiple iterations and analysis from several perspectives.

4.2 Future Work

We want to comment on a couple of immediate directions for future work. The obvious suggestions are getting in touch with a domain expert (e.g. a wine maker) and/or collecting more data with more information. However, these would be costly approaches.

There are other lower cost options that should be tried first. Since we found the simple linear regression to be inadequate, we can try different methods. We found out that our assumptions are violated. Robust regression techniques that can handle heteroscedastic residuals is the one of the first next steps. Another step is to try ordinal regression, since the quality scores are discrete.

However, both of these cannot fully address the need for complex models. Nonlinear regression methods can be used to find a more complex relationship. Since the response variable is discrete, techniques from the classification literature (e.g. Support Vector Machines) can be tried as well. Another idea is to use a model with many parameters along with a method that allows robust feature selection such as the lasso method.

Bibliography

[1] Vinho Verde Wine. http://www.vinhoverde.pt/en/. [Online; accessed 13-March-2013].

[2] Wine consumption in the U.S. http://www.wineinstitute.org/resources/statistics/article86. [Online; accessed 13-March-2013].

[3] Wine Quality Dataset. http://archive.ics.uci.edu/ml/datasets/Wine+Quality. [Online; accessed 13-March-2013].

[4] MKF RESEARCH. The impact of wine, grapes and grape products on the american economy 2007: family businesses building value. http://http://ngwi.org/files/documents/Economic_Impact_on_National_Economy_2007.pdf.