Regression on Vinho Verde

Raymond Chang

1. Background

For this multivariate regression project, I used a wine quality score dataset from the Machine Learning Repository at University of California, Irvine's Center for Machine Learning and Intelligent Systems. Vihno Verde is a Portuguese wine that originated in the Minho province in the north of Portugal. Though it means "green wine", it translates to "young wine" since this type of wine is harvested and consumed within three to six months. Vinho Verde can be white, red, or rosé.

2. Research Question

What is an optimal regression model that can be used to predict wine quality score given its alcohol percentage, sugar levels, pH, density, and other chemical compositions? The **dependent variable** is the wine quality score that is to be predicted. The **independent variables** are the features that act as predictors to the wine score such as alcohol percentage and pH.

3. Exploratory Data Analysis

3.1 Data Cleaning

After importing the white wine dataset, I reordered the columns so the dependent variable, wine quality score, is in the first column. I also renamed the columns for readability and checked for missing values (NA), not a number values (NaN), and infinite values which there were none.

3.2 Data Dictionary

Attribute	Units
Alcohol Percentage	% volume
Sulphates	g(potassium sulphate)/dm^3
Density	g/cm^3
Total Sulfur Dioxide	mg/dm^3
Free Sulfur Dioxide	mg/dm^3
Residual Sugar	g/dm^3
Citric Acid	g/dm^3
Volatile Acidity	g(acetic acid)/dm^3
Fixed Acidity	g(tartaric acid)/dm^3

3.3 Splitting the Dataset

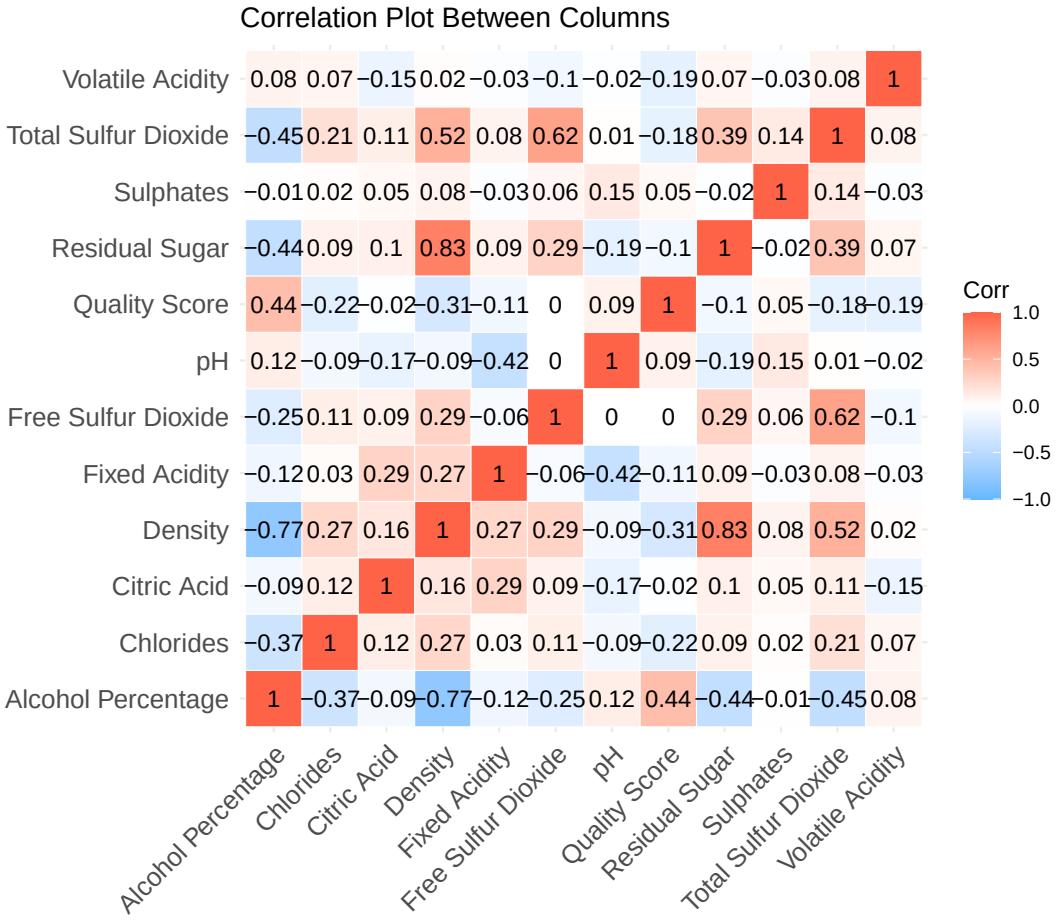
I split the dataset into testing and training sets using a 80/20 ratio. Although this ratio is arbitrary, I chose 80/20, instead of another common ratio like 70/30, because of the Pareto Principle which I personally find fascinating.

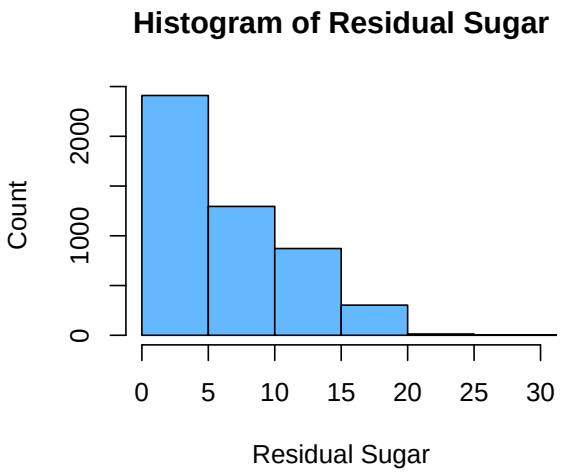
I used sample with default replace = FALSE because the testing set should be tested using a model that didn't already "learn" the quality score of wines with given parameters.

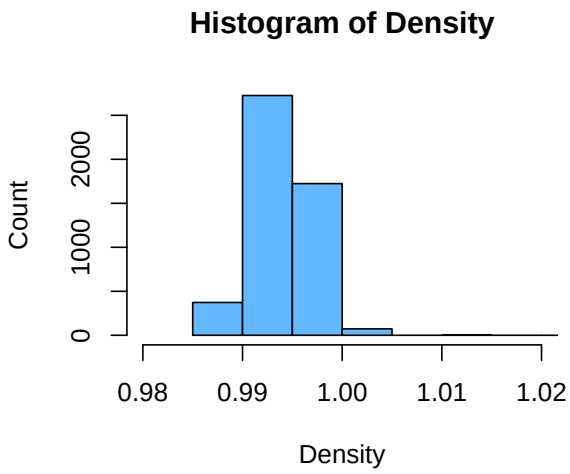
4. Assumptions

Because multivariate regression is a specific type of parametric test, there are stricter requirements than nonparametric tests because stronger inferences are being made from the dataset and model. I used a multivariate regression to test for a *cause-and-effect* relationship, observing the effect of the 11 predictor variables on the 1 outcome variable (wine quality score). The four assummptions are: independence, normality, linearity, and homoscedasticity. I will check for homoscedasticity after modeling.

4.1 Independence of Predictors

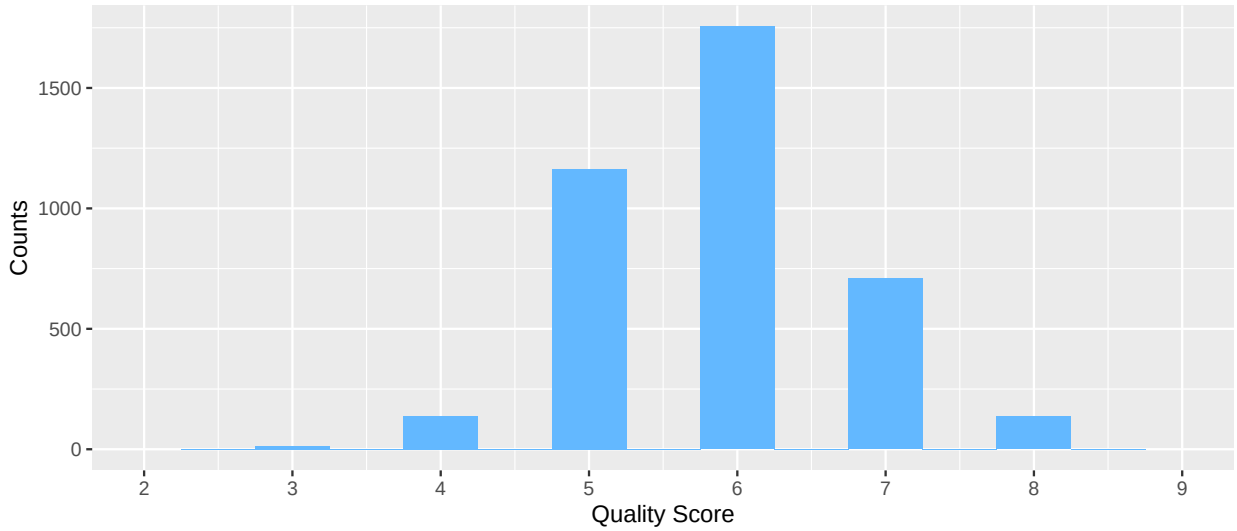






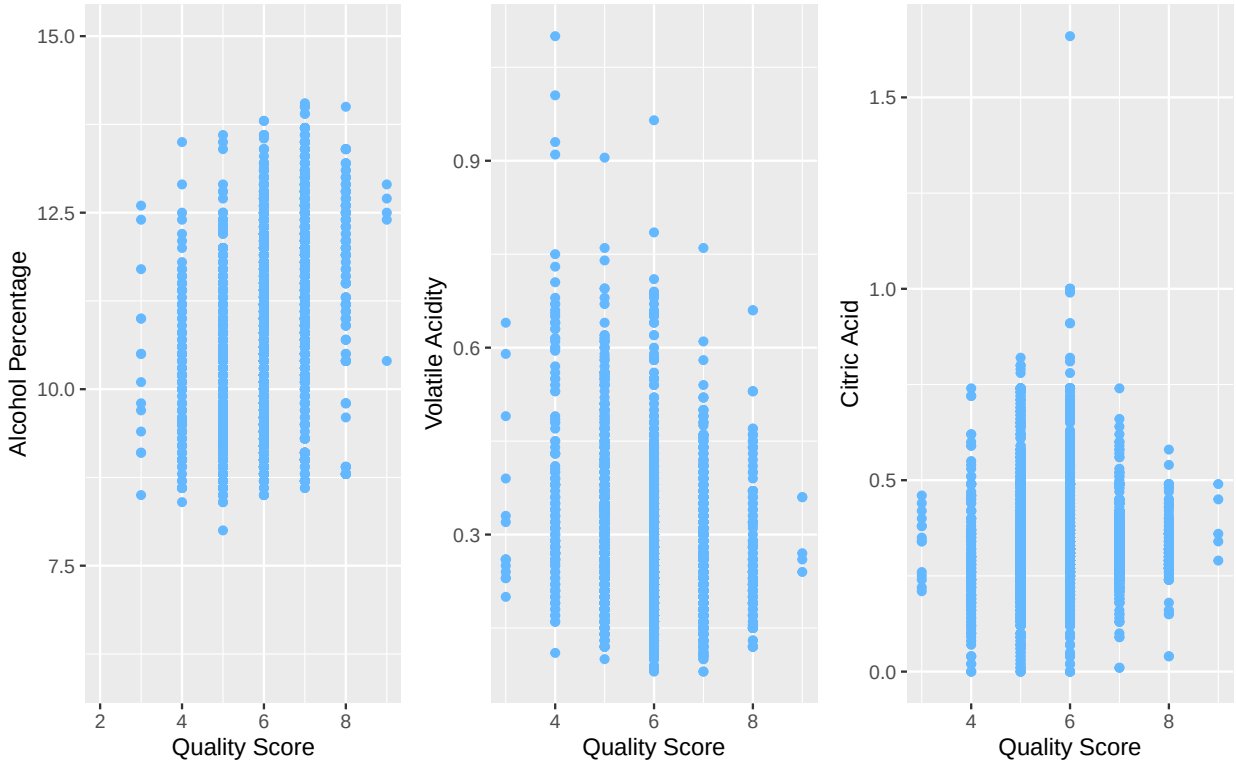
To observe the correlations between the predictor variables, I plotted a heatmap of all columns from the dataset using `ggcorrplot`. The arbitrary cut-off value of r was 0.8, and Density and Residual Sugar had a correlation value of 0.83 which exceeded the cut-off value. To determine whether to remove Residual Sugar or Density, I plotted the histograms of counts of Residual Sugar and Density to observe the distributions of both features. I removed the feature with lower variance, which is Density, because I wanted to retain as much of the dissimilarity in the dataset as possible.

4.2 Normality



The distribution of the dependent variable is roughly normal or bell-shaped.

4.3 Linearity



To check for linearity, I plotted the scatterplots of 3 independent variables (Alcohol Percentage, Volatile Acidity, Citric Acid) against Quality Score and assumed linearity for the other independent variables.

4 Modeling

4.1 Mathematical Model

My goal is to come with a linear model to estimate the quality score of each wine as a linear combination of 10 independent variables. The equation to model a wine's quality score the following equation:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_{10} x_{10} + \epsilon$$

Variable	Interpretation
y	response output, which is the predicted wine score
β_0	regression intercept or baseline score of a wine with 0 in all independent variables
β_1	increase in value for each increase of 1 in first dependent variable (alcohol percentage)
x_1	value of the first dependent variable (alcohol percentage)
β_{10}	increase in value for each increase of 1 in tenth dependent variable (chlorides)
x_{10}	value of the tenth dependent variable (chlorides)
ϵ	random noise or error term which we assume follows a normal distribution with mean 0 and constant variance σ^2

4.2 OLS Model

To select the optimal model, I used a stepwise forward selection method. I started with an empty model and added variables until a parsimonious model was reached. Because a good model should fit data well and be parisomonious, it should only be as complex as necessary to describe a dataset. The *bias-variance tradeoff* is the conflict in balancing the bias error and variance in supervised learning. A model with high bias but low variance runs the risk of underfitting and missing relationships between the dependent and independent variables. A model with high variance but low bias runs the risk of overfitting the training set and the random noise.

I prioritized variables based on how low their p-values were, which indicates their significance to the model. For Model #1, I selected 3 variables (alcohol, free_sulfur, sugar) because they all have extremely low p-values of below $< 10^{-12}$ which meant they were highly significant factors. For Models #2-4, I added the variable with the next lowest p-value to each model.

Model	Variables Included
1	Alcohol Percentage, Free Sulfur Dioxide, Residual Sugar
2	Alcohol Percentage, Free Sulfur Dioxide, Residual Sugar, Total Sulfur Dioxide
3	Alcohol Percentage, Free Sulfur Dioxide, Residual Sugar. Total Sulfur Dioxide, Sulphates
4	Alcohol Percentage, Free Sulfur Dioxide, Residual Sugar. Total Sulfur Dioxide, Sulphates, pH
5	Alcohol Percentage, Free Sulfur Dioxide, Residual Sugar. Total Sulfur Dioxide, Sulphates, pH, Chlorides

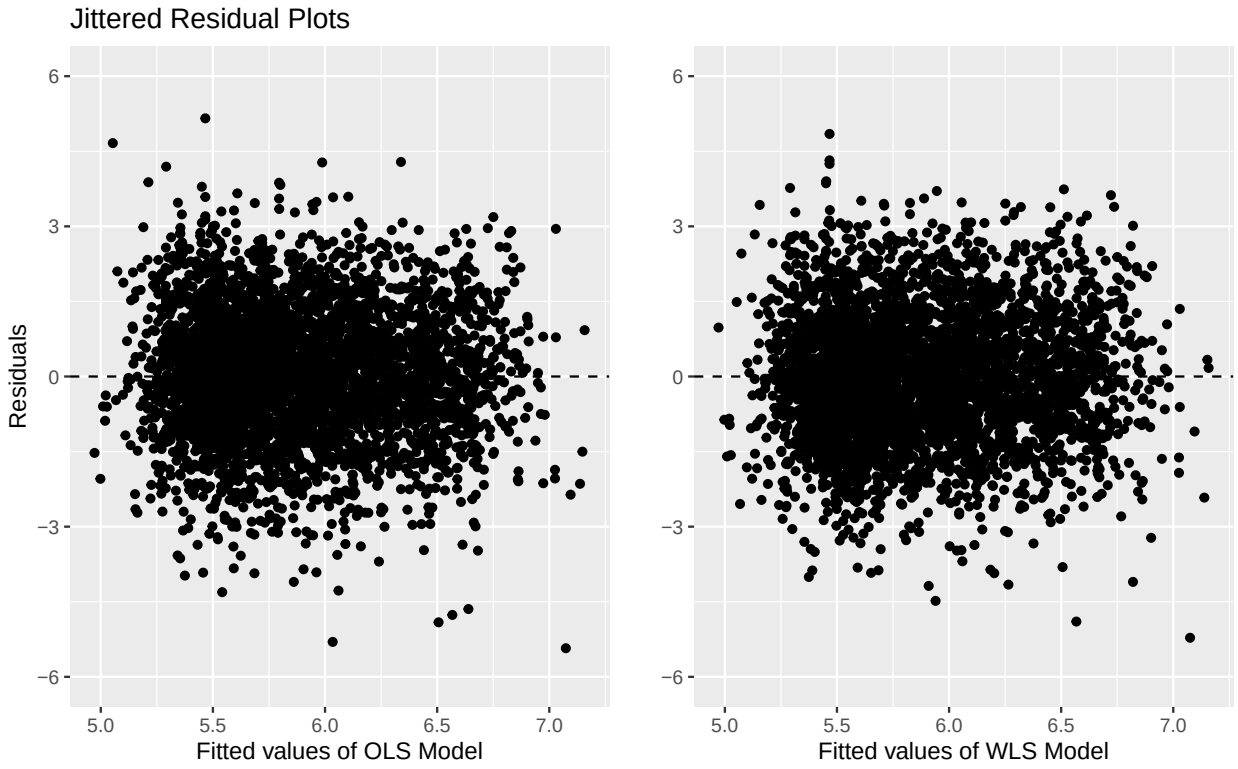
To compare between the models, I used `anova` to interpret whether or not the addition of another feature significantly improved the model. This ANOVA tested whether or not the addition of total_sulfur led to a significant improvement over using just alcohol, free_sulfur, and sugar.

```
## Analysis of Variance Table
##
## Model 1: score ~ alcohol + free_sulfur + sugar
## Model 2: score ~ alcohol + free_sulfur + sugar + total_sulfur
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   3914 2427.5
## 2   3913 2413.2  1    14.287 23.166 1.542e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA results show a Df (degree of freedom) of 1, which is expected since it's indicating that the more complex model has one additional parameter. The p-value is very close to 0, which means that adding total_sulfur to the model did lead to a significantly improved fit over Model 1.

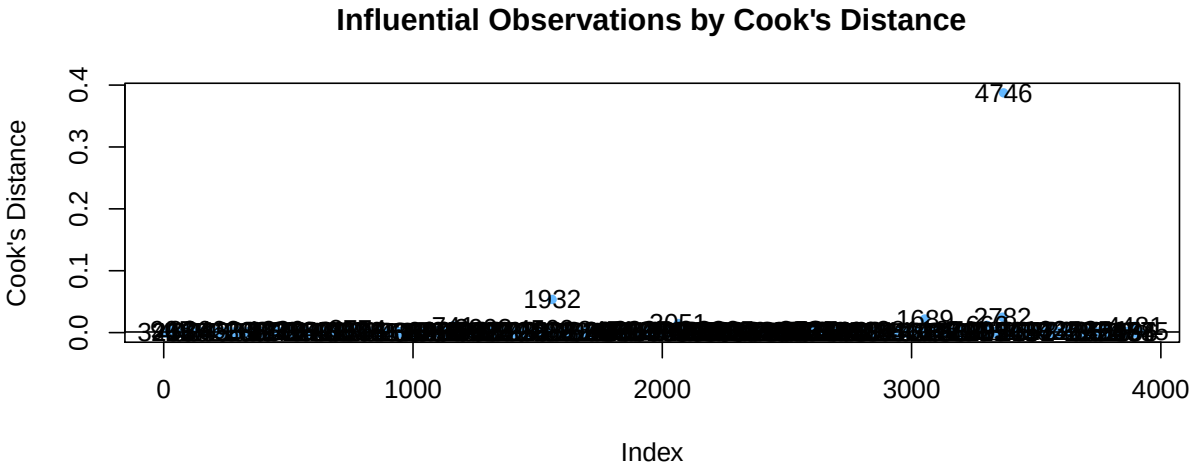
4.3 WLS Model

Because the OLS assumption of constant variance in the errors was violated, as seen in the heteroscedasticity of residuals below, I tried a WLS (weighted least squares) model which corrects the nonconstant variance by weighting each observation by the reciprocal of its estimated variance. In my specific model, the WLS fit was more heavily weighted on the end points where the variance is lower. I graphed the adjusted residuals from the OLS model on the left and the new adjusted residuals from the WLS model on the right. The new WLS model showed a slight improvement as the residuals are now less clustered towards the lower fitted values.



4.4 Outliers

A major drawback of linear modeling is that they are sensitive to outliers. An outlier given an inappropriate weight could dramatically skew the regression results. I identified outliers by using Cook's distance, a commonly used estimate of the influence of a data point in a least-squares model. To determine how "big" is "big", I used the standard cutoff of 4/n where n is the number of observations. I plotted the observations considered influential by Cook's distance below, and since the cutoff line of 4/n is quite close to 0, there are many influential observations that are unidentifiable from the plot below. In total, I removed 206 influential points from the original dataset with 3918 observations.



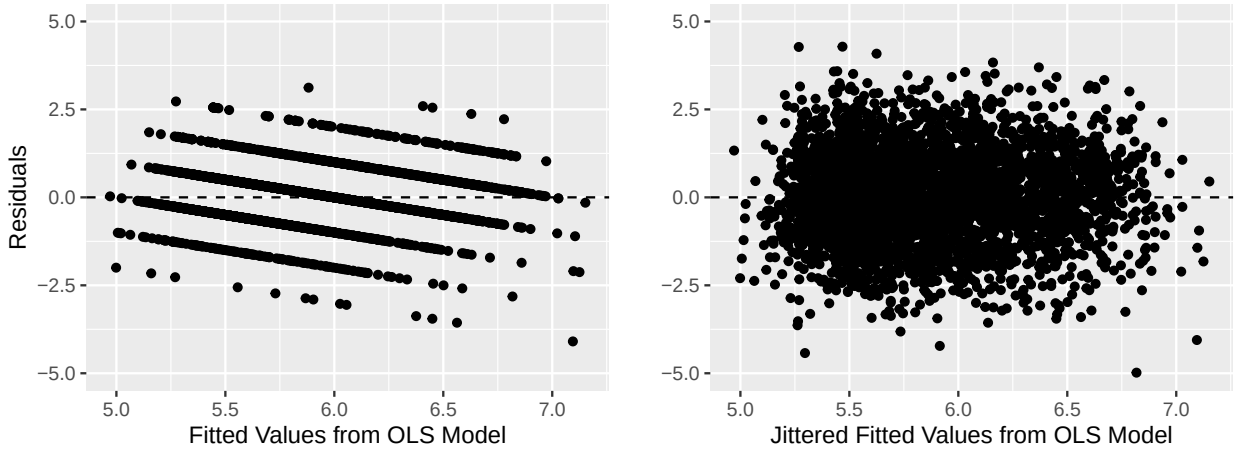
4.5 Model Interpretation

Coefficient Estimates of Model 2	Interpretation
Intercept: 2.171992	The estimated quality score for a wine with alcohol percentage, free sulfur dioxide, residual sugar, and total sulfur dioxide at zero. However, this does not provide for any meaningful interpretation.
Alcohol: 0.3420803	For every one unit increase in alcohol, the model predicts a 0.3420803 increase in quality score.
Free Sulfur Dioxide: 0.0073861	For every one unit increase in free sulfur dioxide, the model predicts a 0.0073861 increase in quality score.
Residual Sugar: 0.0188162	For every one unit increase in residual sugar, the model predicts a 0.0188162 increase in quality score.
Total Sulfur Dioxide: -0.0019954	For every one unit increase in total sulfur dioxide, the model predicts a 0.0019954 decrease in quality score.

Some metrics worth pointing out from the model summaries include the standard error, t-value, p-values of the coefficients, and adjusted r^2. The standard error measures the average amount the coefficient estimates vary from the actual average value of the response variable. It can also be used to compute confidence intervals and statistically test the hypothesis of the existence of a relationship between the dependent and independent variables. The t-values are coefficient estimates divided by their standard errors. The p-value for each term tests the null hypothesis that the coefficient is equal to zero. Because the p-values are so low, I can confidently reject the null hypothesis of the coefficient being zero. R^2 aims to estimate the proportion of variance explained by the regression equation. Adjusted r^2 is a modication of r^2 that adjusts for the number of explanatory terms in the model. R^2 is always a value between 0 and 1, and because my OLS and WLS models had adjusted r^2 values of 0.2155 and 0.2142 respectively, it can be interpreted as the model explaining around 21% of the variance. It is important to keep in mind that studies attempting to predict human behavior generally have r^2 values less than 0.5, because people are hard to predict. There is an inherent amount of unexplainable variability in all studies. Even with a low r^2, statistically significant p-values can still identify valuable relationships, and the models' coefficients still have the same interpretation.

4.6 Residuals

To check for homoscedasticity, I plotted the residual plots from the OLS model against the fitted values to see whether or not the observed errors (residuals) are consistent with stochastic error. Ideally, a residual plot should be centered around 0 and be homoscedastic or equally distributed throughout its fitted values. The residual plot from the model is plotted below on the left. Because the response variable (wine score) only takes integer values, it created a residual plot that was difficult to check for homoscedasticity, I mitigated this issue by adding jitter to the y-values by randomly generating a normal distribution with mean = 0 and standard deviation = 1 for interpretability purposes. From the jittered residual plot, I observed that the residuals were slightly heteroscedastic since they were more clustered towards the lower fitted values. I suspected that the heteroscedasticity came from my third assumption, because I assumed linearity for all variables even though I only checked 3 independent variables. This could also be because I used `lm` which provides a simple linear model based on OLS (ordinary least squares).



5. K-Fold Cross Validation

In addition to using a stepwise forward model selection method, I also used k-fold cross validation to create another potential model. K-fold cross validation splits a dataset into k-subsets, and each subset is held out while the model is trained on all other subsets. This process is repeated k times, and an overall estimate of coefficients is determined from the average of the k estimates. Though arbitrary, I selected k=10 because a larger k means the model will have less bias towards overestimating the true expected error since there are more training folds the model is being ran on.

Results from 10-fold Cross Validation Model:

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0739 -0.5217 -0.0211  0.4688  3.1119
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1719483  0.1539318  14.110  < 2e-16 ***
## alcohol       0.3421053  0.0120026  28.503  < 2e-16 ***
## free_sulfur   0.0073745  0.0009274   7.952 2.39e-15 ***
## sugar         0.0187336  0.0028648   6.539 6.99e-11 ***
## total_sulfur -0.0019902  0.0004135  -4.813 1.54e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7853 on 3913 degrees of freedom
## Multiple R-squared:  0.2148, Adjusted R-squared:  0.214
## F-statistic: 267.5 on 4 and 3913 DF,  p-value: < 2.2e-16
```

The intercept and all coefficients in the k-fold cross validation model were highly significant due to their extremely low p-values. The adjusted r^2 from the k-fold cross validated model is 0.214, which is lower than both adjusted r^2 values of the OLS and WLS models. All coefficients in the k-fold model were also highly significant.

6. Success Metrics

To test how well the models predicted wine scores, I examined metrics based on their in-sample and out-of-sample performance. In-sample metrics show how well the model models the underlying trends and noise of the given wine quality scores. Out-of-sample metrics show how well the models would perform on new information that it has not previously seen. I used adjusted r^2 values as the in-sample metric and compared Residual Sum of Squares (RSS) and Akaike's Information Criterion (AIC) as out-of-sample metrics. RSS measures the discrepancy between the dataset and the three models; a small RSS would indicate a tight fit of the model to dataset. AIC estimates the quality of each model, taking into account the trade-off between goodness of fit and model simplicity. This mitigates the risk of overfitting/underfitting. A higher r^2 value means the model explains a greater proportion of variance of the dataset, and a lower value in both RSS and AIC indicate a better model.

I have gathered these metrics in the table below and bolded the best performancce within each category.

Model	Adjusted r^2	Residual Sum of Squares (RSS)	AIC
OLS	**0.2155**	594.993	**8834.579**
WLS	0.2142	**594.8207**	9230.603
K-Fold CV	0.214	594.8532	9232.069

7. Conclusion

To examine the relationship between wine quality score and its alcohol percentage, pH, density, and chemical compositions, I first conducted exploratory data analysis on the dataset. I cleaned the dataset, created a data dictionary for the units of measurements, and held out 20% of the dataset by sampling without replacement. This 20% served as the test set while the remaining 80% was used as the training set for my models.

I then stated the four main assumptions I made to proceed with linear regression, which were independence of predictors, normality, linearity, and homoscedasticity. I checked for independence between the independent variables by plotting their correlations using a heatmap and found Residual Sugar and Density to both be highly correlated. I wanted to retain the feature with greater variance because it allows the model to be trained on more diverse data.

I generated my first linear model by using a stepwise forward selection method with the standard ordinary least squares (OLS) method. For the base model, I chose the 3 variables with extremely low p-values, which indicate their high significance. I then added variables based on the next lowest p-value to create 5 total OLS models. I used `anova` to compare the p-values of each additional variable, which tests the null hypothesis that the coefficient is equal to zero (having no effect) and selected Model 2 as the best OLS model out of the 5. For my second linear model, I used a weighted least squares (WLS) approach because I noticed from the residuals plot that the OLS assumption of constant variance in errors was violated. The weights in WLS are inversely proportional to the variances of each observation, and it can be shown by the Gauss-Markov Theorem that picking weights to minimize the variance in the WLS estimate has such a unique solution. In my third linear model, I took a k-fold cross validation approach which is a robust method for estimating the accuracy of a model. The algorithm splits the data into k-subsets or folds. One subset is reserved, and the model is trained on all other subsets. The model is tested on the reserved subset, and this process is repeated until all k subsets have been served as the test set. Finally, the average is computed between the k models in a process called cross-validation.

I then compared the 3 metrics of success for my models: adjusted r^2, RSS, and AIC. Though the WLS model edged out in RSS values, the difference in RSS between the models is minimal, and the OLS model performed the best as evidenced by its highest adjusted r^2 and lowest AIC.

8. Sources

1. Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez A. Cerdeira, F. Almeida, T. Matos and J. Reis. Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal @2009

2. https://archive.ics.uci.edu/ml/datasets/wine+quality

3. http://www3.dsi.uminho.pt/pcortez/wine5.pdf