PSTAT 126 Project: Understanding Red Wine Quality

Hanya Ansari (Section: Thursday, 12 PM)
Carina Yuen (Section: Wednesday, 3 PM)
Daren Aguilera (Section: Wednesday, 3 PM)

Introduction

This project examines a set of observations on a number of red wine varieties from Portugal and their respective physicochemical properties based on the attributes of the dataset "Wine Quality Based on Physicochemical Tests" from the UCI Machine Learning Repository. The dataset contains 1599 number of instances and includes 12 predictors. 11 continuous variables were examined with the goal to create a model that accurately predicted the wine quality. The variables used in our linear model are Fixed Acidity (Acids involved with wine – Fixed or Nonvolatile), Volatile Acidity (Amount acetic acid in wine), Citric Acid (Adds 'freshness' and flavor to wine), Residual Sugar (Amount of sugar remaining after fermentation is completed), Chlorides (Amount of salt in the wine), Free Sulfur Dioxide (Free form of SO2 that exists in equilibrium between molecular SO2 and Bisulfite ion), Total Sulfur Dioxide (Amount of free and bound forms of S02), Density (Measures how close the wine is to the density of water), pH (Measures acidity on a scale from 0 to 14), Sulphates (Wine additive that contributes to SO2 levels), Alcohol (Percent of alcohol content in the wine), and Quality (This discrete variable evaluates each wine sample with a minimum of three sensory assessors (using blind tastes) on a scale from 1 to 10).

The Quality Score is the response variable in our ultimate model, fit_normal 3, which comprises 11 quantitative predictors, as shown below:

Model in R:

```
fit_normal3 <- lm(volatile.acidity~ x_f+x_v+x_c+x_r+x_ch+x_fs+x_ts+x_d+x_p+x_s+x_a,
data)
```

$$Y = -9.17*10^{-15} + 6.70*10^{-17}\, x\_f - 3.82*10^{-18}\, x\_r - 6.97*10^{-17}\, x\_ch - 2.56*10^{-19}\, x\_fs + 9.83*10^{-20}$$
$$x\_ts + 7.59*10^{-15}\, x\_d - 1.48*10^{-16}\, x\_p + 2.53*10^{-17}\, x\_s + 6.76*10^{-18} x\_a$$

For our chosen model, fit_normal3, the model with coefficients is as shown above.

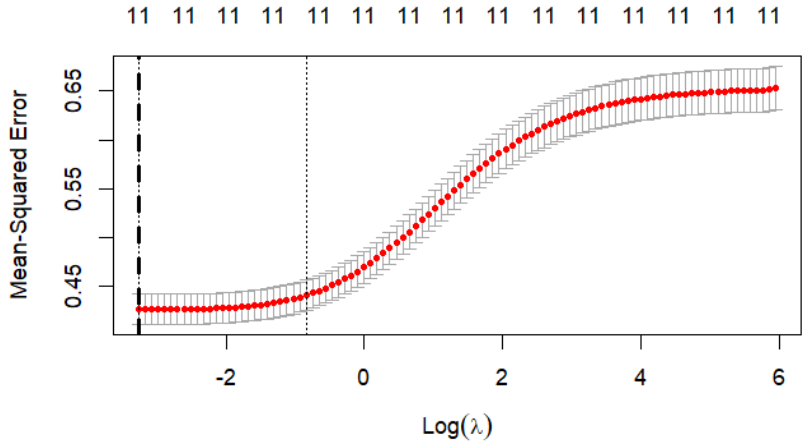
Questions of Interest

- Which predictors can be used to create the best model to predict the quality score and volatile acidity value of wine?
- What is the predicted volatile acidity when selection consistency is prioritized based on the model created?
- How can we adjust winemaking practices to ensure an optimal amount of volatile acidity?

Shrinkage Methods: Ridge and Lasso Regression

Ridge regression shrinks the coefficients towards zero. Upon using coefficients (best_model), we got the following estimates: 10.42 for the intercept, 0.987 for alcohol, -0.0025 for free sulfur dioxide, and -0.0043 for total sulfur dioxide. Using cv.glmnet, we found that the optimal lambda is 0.03105, as a dotted vertical line marks the value of Log(λ) that minimizes the MSE value.

Analysis of the model using the optimal lambda:
Below is a plot of the test MSE by the lambda value for Ridge Regression.

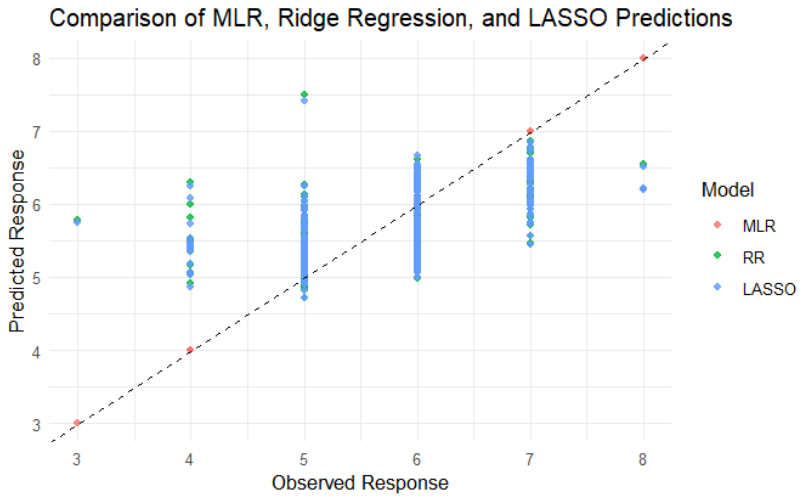


Lasso Regression (Least Absolute Shrinkage and Selection Operator): Upon using Lasso Regression, we noted that all the predictors except alcohol are dropped from the model. In the output matrix, the coefficients are "empty" values. Since there is only one "non-zero" coefficient of alcohol (besides the intercept), it indicates the regularization parameter lambda might be causing too much regularization. On one hand, it was helpful in narrowing down our many

predictors to only the impactful ones, but perhaps we lost certain insights by completely eliminating the other terms. Here, we can see the limitations of using certain criteria for determining the calculated best_lambda is 0.03105506. The MSE plot for the Lasso Regression looks similar to that of Ridge Regression, with the vertical dotted line being plotted slightly more to the left at -3.75 instead of -3.5.

Analyze and differentiate the new ridge and lasso models with the final MLR model from previous task

The plot below measures the prediction using three regression models. Comparing the points to the dotted line, we see that the LASSO regression method has clustered values of the predicted response at observed response 4-7. The Ridge Regression Model also has clustering, but it seems there is greater variance in the response than that of the LASSO model. This makes intuitive sense as LASSO regression tends to choose more sparse models, so there is less variation amongst the chosen model. Lastly, the red points indicating the MLR fall right on the line as expected.



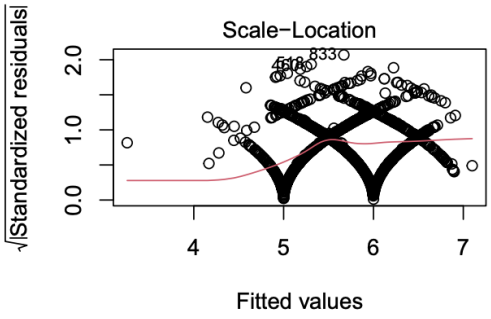
Regression Method

To determine our variables of interest, we looked at the summary statistics output from skimr, parameters such as fixed acidity (Min: 4.6, Max: 15.9) and total sulfur dioxide (Min: 6, Max:

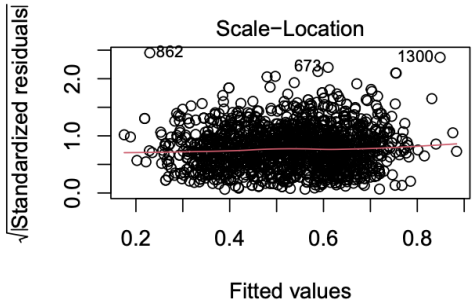
289.00) had a large range compared to other variables. On the other hand, comparing 6 fixed variables and volatile acidity, we noticed that relative variation in acidity levels was lower.

The two main regression methods we used were Stepwise Regression with Backward Elimination and the Bayesian Information Criterion (BIC) method. Our first fitted model of interest is fit3_quality, which is a full polynomial model of the quality data with the highest order 3. To perform Backward Elimination, we dropped the Density term. Then we conducted an analysis comparing the models, and noticed that dropping the term did not change the RSS much, in fact, it increased it from 624.49 to 625.42. We experimented with dropping terms to minimize the RSS, and did not see major changes, so we concluded that our model with the Density term was the best. For our continuous variable analysis, we decided to use the BIC method to predict Volatile Acidity scores, as shown in fit_normal3. This had a BIC value of -107962.4.BIC prioritizes selection consistency, and a lower BIC value implies a better fitted model. Comparing the full model and the BIC, we can see the Scale-Location plot improves drastically in regards to violating the homoscedasticity assumption. The graph below shows the improvement of the plot, as the points are relatively evenly and randomly bouncing around the horizontal line.

Full Model



BIC Model

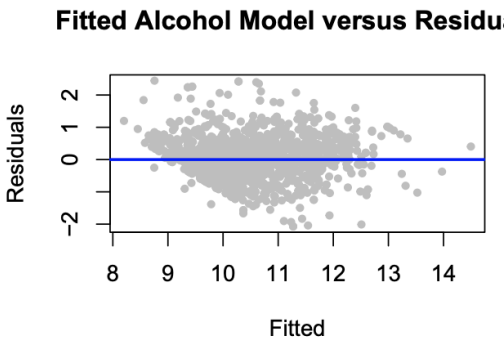


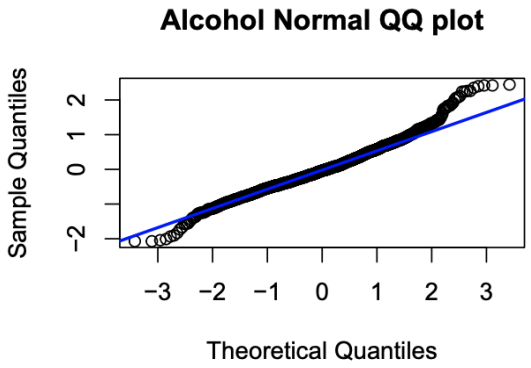
Regression Analysis, Results and Interpretation

Interpretation of regression for predicting discrete variable: Quality Score

We looked into factors that can affect Quality Score and were able to deduce that alcohol has the highest collinearity with wine quality, implying that there was a preference for wine with higher alcohol content. In our initial project step, we did ANOVA testing of alcohol content and final quality score, the F value was 468.3, which indicates that alcohol content likely causes changes in Final Quality Score for each wine. Looking at our ggpairs plot, the correlation value is 0.476, with three asterisks besides it. This indicates a p-value less than 0.001, which suggests very significant correlation.

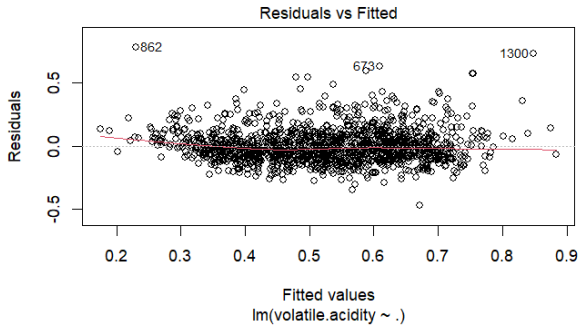
Our chosen model for quality score is fit3_quality, which is a full polynomial model of the quality data with highest order 3. For fitalc (Alcohol Content and Density), the F value is 521.6, indicating that it is more likely alcohol content is causing the variation in density. This makes sense as the density parameter is in g/cm^3, with pure water having density of about 1 g/cm^3 and the density of pure alcohol (ethanol) being 0.789 g/cm^3.

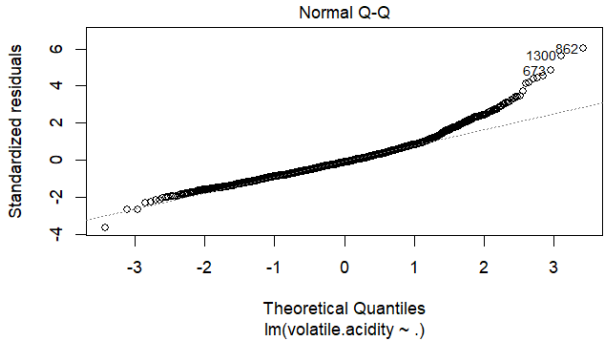




Interpretation of regression for predicting continuous variable: Volatile Acidity

We looked into Volatile Acidity because while Quality Score and enjoyment of the wine is important, further background research revealed that Volatile Acidity mainly measures the acetic acid content of the wine, which contributes to the smell and taste of vinegar in the wine. This further shows how the presence of a high acetic acid content can negatively affect the Quality Score as it is reflective of wine that has either gone bad or developed bacterial growth.





Innovation: New Regression Method with Weighted Least Squares Regression (WLS)

In addition to the Ordinary Least Squares Regression, we looked into the new method, Weighted Least Squares Regression. WLS Regression adjusts for the violation in the assumption because each weight is made inversely proportional to error variance, so our results weigh observations with lower variance heavier. The intuition behind this weight The WLS Regression coefficients for the variables volatile acidity, free sulfur dioxide, total sulfur dioxide, and alcohol are 1, $-8.507*10^{-20}$, and $2.815*10^{-18}$, respectively.

We chose this as our new method because it helps when the data violates the homoscedasticity assumption. From our Scale Location plot of the full linear model with the response being volatile acidity, we can see there is a defined pattern in the graph (see page 5). While we were able to mostly address this by using BIC to pick our best model, using the WLS method.

Conclusion

After examination of the data, our business insights indicated that monitoring volatile acidity content and alcohol content is crucial, as there was a high correlation with the final quality score. In terms of data quality, this dataset has the advantage of being larger than most wine-related data, with a sample size of 1599 for red wine. For variable selection, the data might be examined to exclude variables with high collinearity such as total sulfur dioxide and free sulfur dioxide. The variable, however, is still important, as sulfur dioxide is used to reduce the bacterial load from native yeast fermentation and contamination from transportation. Further investigation on

how volatile acidity affects the taste, and therefore the distribution of the quality scores, could be made with a smaller model, or a focused study on volatile acidity.

Regarding implementation for further investigation, it is valuable to look into the vinegar content because while most consumers would think they can easily detect and discern such a taste, commonly used vinegar containing products are 3-9% acetic acid by volume, which is 30-90 times the amount permitted in wine. As a result, volatile acidity is not easily detected and should be measured at different stages of the wine production process to provide an accurate estimate. Practices such as cold-soaking the grapes prior to fermentation, and barrel aging increase the risk of acetic acid bacterial contamination and growth. The issue is easier to deal with through prevention. Reducing and monitoring such growth can help improve overall wine quality in terms of consumption safety, and controlling desired flavor.

Works Cited

Lesson 13: Analysis of Variance for Simple Linear Regression. Stat 501 - Regression Methods | Online Statistics Course. The Pennsylvania State University, n.d. Web. 13 Dec. 2023.

Volatile Acidity in Wine. Penn State Extension. The Pennsylvania State University, n.d. Web. 13 Dec. 2023.

Volatile Acidity in Wine Making. Enology | University of Minnesota, n.d. Web. 13 Dec. 2023.