STSCI 4740 - Data Mining and Machine Learning
Yang Ning
Final Paper

Erika Hauschild - esh 79
Cesar Cisneros - cec343
Jahin Aishee - sta48
Albina Chowdhury - ac2523
Rahma Tasnim - rt429

INTRODUCTION

Wine connoisseurs and amateurs alike try to determine the quality of the wine they drink. However, tasting all the wines in the world would take too much time and too many resources. Any rating that is then made about the wine quality would then need to be confirmed and validated by other connoisseurs. Instead, we can predict the quality of wine without even needing to taste it. We can do this by looking at a data set that looks at different wines and notes their different attributes, including the wine quality. These attributes impact the quality of the wine and we need to see which attributes are most important. The most significant attributes will then help us determine which wines have the best quality even when we have not tasted them.

DATA DESCRIPTION

The *Wine Quality Data Set* consists of two data sets with a set of attributes that describes red wine and white wine. We had decided to use one complete data set that combined the two data sets which added a new column called “type” which would either be red or white. The different attributes of the wine include “fixed acidity”, “volatile acidity”, “citric acid”, “residual sugar”, “chlorides”, “free sulfur dioxide”, “total sulfur dioxide”, “density”, “pH”, “sulphates”, “alcohol”, and “quality”. We use these attributes to predict the quality of the wine.

DATA EXPLORATION

Before making any models or predictions, we explore the data to understand any relationships between the different predictors. Initially, we look at the summary of the data. Most notable points are that the wine quality values range from 3 to 9. After looking at the count of each of the quality values, we see that most wines are assigned a score of 5 or 6 indicating that most wines are mediocre. Very few wines are assigned a score of 3 or 9, indicating that the best wines are those with a score of 3 whereas the worst wines are the ones with a score of 9.

Next, we want to find any relationships between the predictors so that we can verify their relationship when we run regression models. We mutate the data using the dplyr library to make the categorical feature: type, a binary feature to run the pairs function from the tidyverse library. After running the pairs function (Figure 1), there are no evidently clear relationships between the predictors.

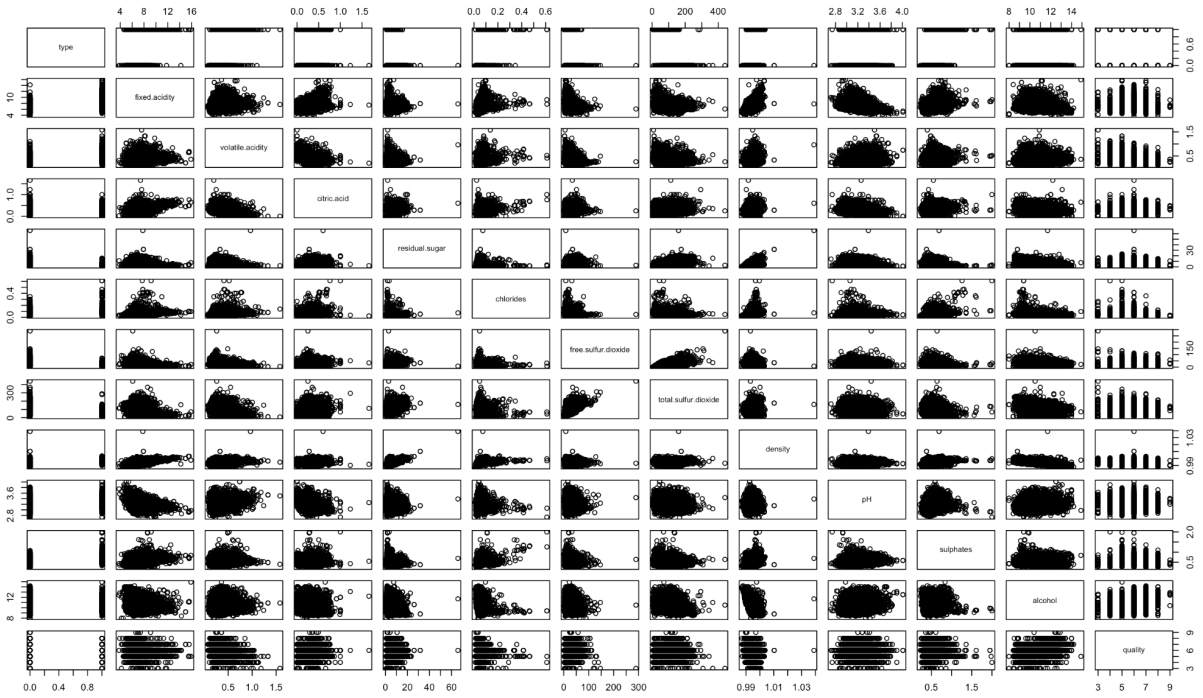

Figure 1: Pair plot of all the predictors. From top to bottom and left to right: type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, quality.

To look for more relationships, we look for any correlations between the predictors. We make and plot a correlation matrix (Figure 2) using the corrplot library. With higher correlations, the darker the blue is whereas with smaller correlations, the darker the red is. We can see that the most correlated predictors are free sulfur dioxide and total sulfur dioxide. This makes sense because total sulfur dioxide is a combination of the bounded sulfur dioxide and unbounded sulfur dioxide. However, we are most concerned with any correlations between quality and the predictors. The only majorly correlated predictor with quality is alcohol. This also makes sense because the main component in wine is alcohol, so the quality will be affected.

We also looked at the correlation matrix (Figure 3) with specific values as well to verify our correlation plot. It shows that free sulfur dioxide and total sulfur dioxide has a correlation of 0.79. The correlation between alcohol and quality is 0.44. It is not the largest correlation value between two predictors but it is the largest correlation that quality has with another predictor. Any other correlation values that quality has are close to 0. This means that we must fit some regression models to determine which predictors play a role in predicting the quality of wine. We are unable to make conclusions about quality using these graphs.

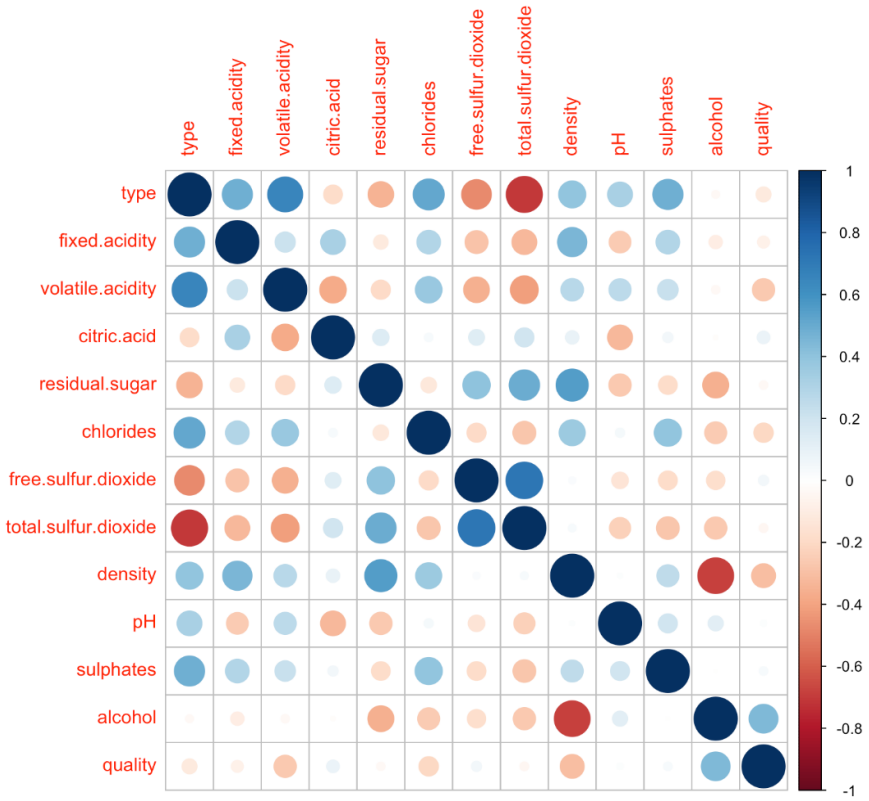

Figure 2: Correlation plot of all the predictors.

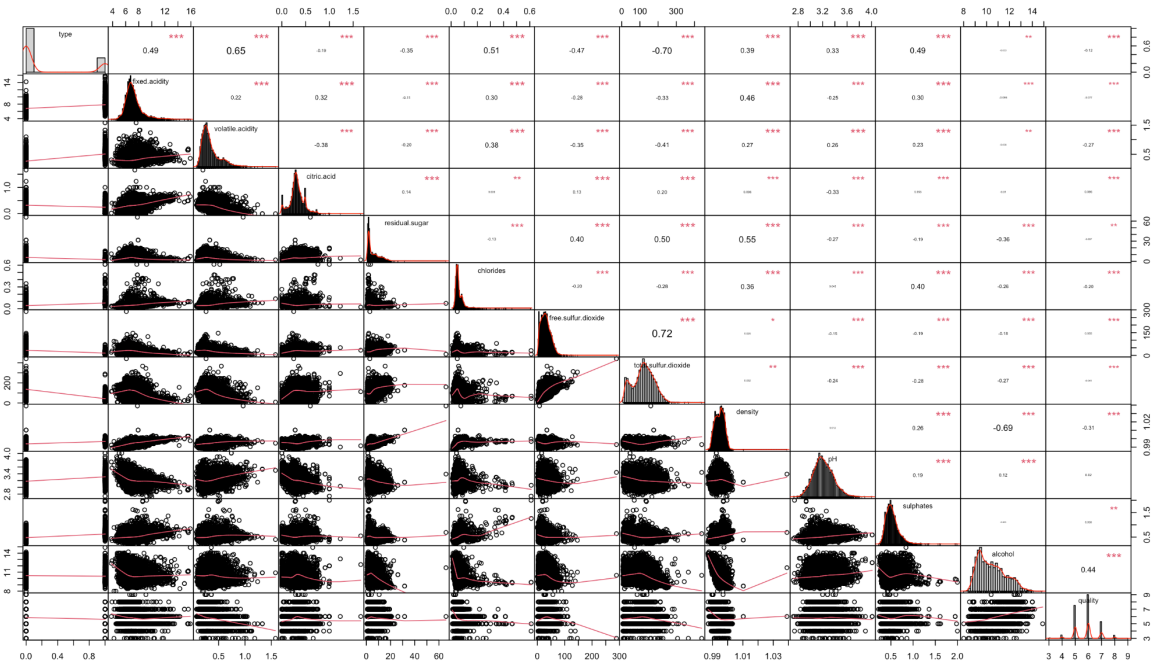

Figure 3: Correlation Matrix of all the predictors.

FORWARD SELECTION

In order to evaluate the importance of the given predictors in this dataset, we conducted three stepwise regressions: forward selection, backwards selection, and best subset selection.

Forward selection is a process in which there is a null model and through each step of the regression, variables are added in one by one. Each one of the variables that are added in from each forward step is supposed to give us a single best improvement to our model. Essentially, this process helps us determine which predictors are most likely to give us a more fitted model. Using the package olsrr, we were able to complete forward selection using a linear model on the predictors of the dataset which contained combined information of both types of wine (red and white). From the results of the regression model, when looking at the Adjusted R-squared values, we found that the predictors with the four highest values are: fixed_acidity (0.2965), pH (0.2934), chlorides (0.2927), total.sulfur.dioxide (0.2914). Citric acid is the only predictor that was not included in the list of strong predictors.

BACKWARD SELECTION

Backward selection is a process which is similar to forward selection, however, instead of starting with a null model, this process begins with all of the predictors. With each step, a predictor is removed because it is statistically insignificant. This process continues until all of the insignificant predictors are completely removed. From the results of the regression model generated by our code, we found that citric acid should be omitted because it is not a significant predictor in fitting this model.

BEST SUBSET SELECTION

Best subset selection is a process in which the goal is to find the subset of predictors that best fit the linear prediction model. This process considers all possible combinations of the predictors and finds the one with the highest R^2 value. From our results, we can see that the model with all of the predictors excluding the predictor citric acid has the highest Adjusted R squared value (0.2953).

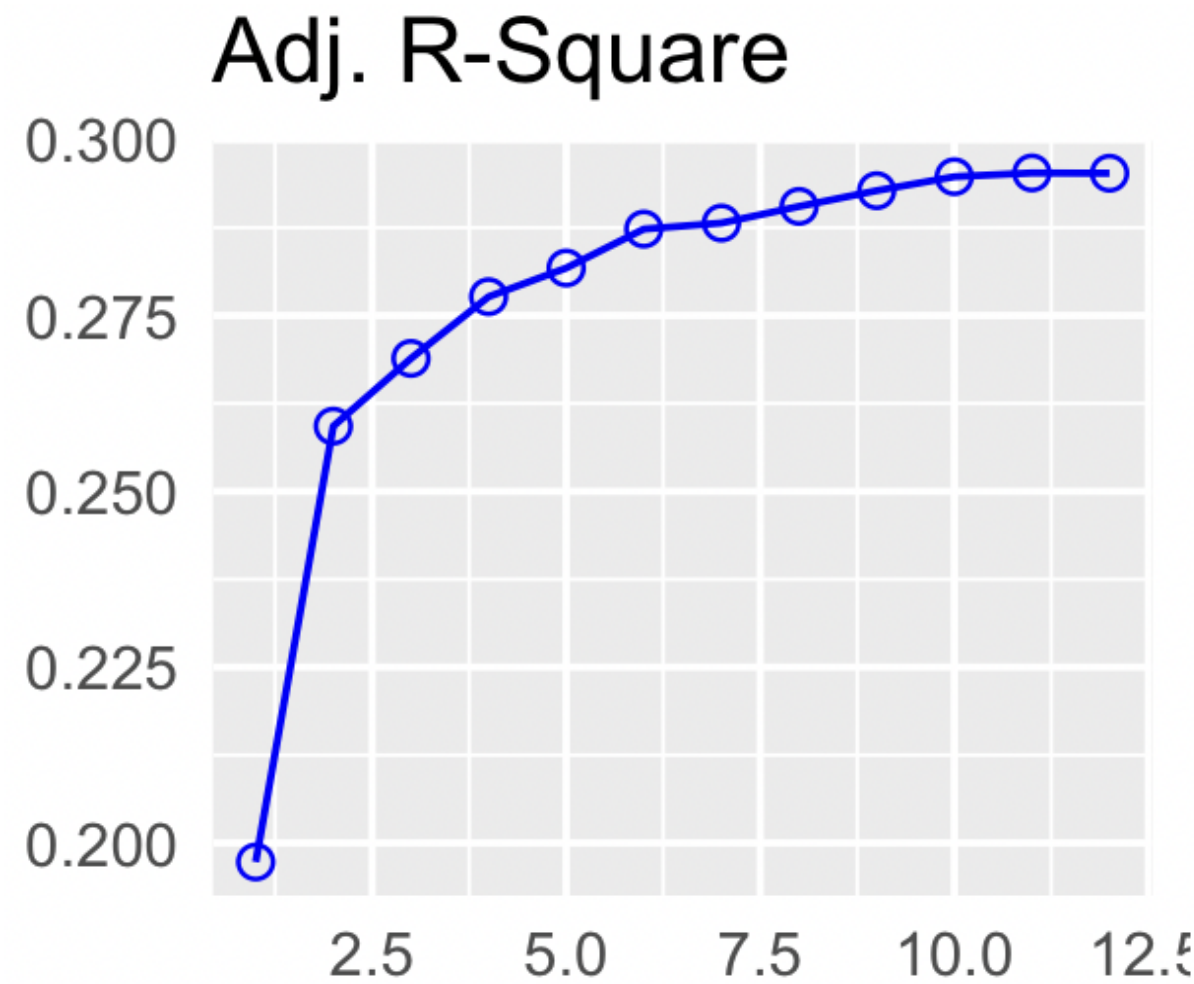

Figure 4: Adjusted R-Squared graphed using best subset selection.

From all of the stepwise functions, we can see that the predictor citric acid is the most statistically insignificant predictor from the entire dataset for fitting a linear model.

RESAMPLING DATA

To determine which regression model is the best at predicting the quality of wine, we resample our data using K Fold Cross Validation and then calculate the errors. The regression model with the lowest errors is the best model. For our three models: linear regression, polynomial regression, and ordinal logistic regression (OLR), we want to determine the error rate using k fold cross validation. Because ordinal logistic regression is used as a classification method, we look at the misclassification errors instead. We see that there is a lower misclassification error (0.1891642) with the model fit with the OLR on the best variables, when comparing it to the misclassification error (0.1897799) from the OLR model with all the variables. For linear regression, after running the k fold cross validation on the model with all the predictors and on the model with only the best predictors we find that the test errors for the model with the best variables is lower (0.5699147 0.5698385) than the model with all the variables (0.569924

0.569844.) For the polynomial regression, the error rate (0.70610552) with the model fit on the best variables is lower than the error rate (0.70998596) with the model fit on all the variables.

LINEAR REGRESSION

After selecting the most impactful variables, to predict the data we chose to start with linear regression because quality, the response variable, is continuous. Also, linear regression is the simplest but also one of the most effective approaches when it comes to supervised learning and predicting quantitative responses.

Linear regression tends to be divided into two - simple linear regression which is useful for predicting the response based on a single predictor and multiple linear regression which extends the simple linear regression model to directly accommodate for multiple predictors. In our case we want to build a model for a set of p predictors X_p, which predict the level of quality Y expressed by the following equation:

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

Here Y is the outcome and p is the number of predictors. Since we selected the best subset of predictors during variable exploration, our model becomes:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} + \beta_{11} x_{11}$$

where x_1 is type, x_2 is fixed acidity, x_3 is volatile acidity, x_4 is residual sugar, x_5 is chlorides, x_6 is free sulphur dioxide, x_7 is total sulphur dioxide, x_8 is density, x_9 is pH, x_{10} is sulphates, and x_{11} is alcohol. With the help of linear regression, we built a function that best fits the data.

In order to test that our crafted model provides a more accurate prediction of quality than the generic model which includes all predictors we browsed through available scientific literature. In the Decision Support Systems paper, by Cortez et al. Mean Absolute Deviation(MAD) is mentioned as a commonly used error metric the shown in the equation below:

$$MAD = \sum_{i=1}^{N} |y_i - y_i| / N$$

Here $\hat{y}_k$ is the predicted value for the k input pattern. The Mean Absolute Deviation measures variability by looking at the average distance between observations and their mean. It uses the original units of data which makes interpretation much simpler. Here larger values indicate that the data points are far from the average and lower values lower values indicate that the data

points are quite close to the mean. It is quite similar to standard deviation and they both measure variability but using very different calculations.

Using MAD as a cost function for the K-fold Cross Validation, we then calculated the prediction or test error rate. The test error rate here represents the percentage of errors made when predicting the quality of the wine. We found that our linear regression model with the best subset had an adjusted prediction or test error rate of 0.5698385 which was .00001 less than the existing generic model. The reason behind using the adjusted prediction error rate is to account for bias amongst the model.

ORDINAL LOGISTIC REGRESSION

Following our results from performing variable selection on our dataset, certain predictors appear to be more fitting for predicting wine quality. We understood that nearly all predictors but citric acid are important in predicting wine quality. This prompted us to perform ordinal logistic regression (olr): an extension of simple logistic regression, which is a classification method. The difference in ordinal logistic regression lies precisely in the word “ordinal”, or the dependent variable, which is categorical in simple logistic regression and ordinal in olr, meaning there is an explicit ordering in the categories. The dependent variable in our case is wine quality. The variable is measured in an ordinal scale and can be equal to one of the categories or levels: low-quality wine, medium-quality wine, and high-quality wine. These categories were determined by a cutoff of 3 so that it is equally distributed among 3 categories since there our maximum quality is 9. Thus, category 1 equals high quality, 2 medium quality, and 3 low quality.

As expounded, we ran olr to predict quality on (1) all predictors in the dataset and (2) all predictors except citric acid, which we deemed unimportant based on our best subset selection results. Overall, our model excluding citric acid gave us a better count of both good and high quality wine. For our first model with all predictors, 0 category 1 observations are identified correctly (high quality), 4944 category 2 (medium quality), and 320 category 3 (low quality). Conversely, we obtained 0 category 1 observations identified correctly, 4945 observations category 2, and 323 category 3. We can notice a slight increase in observations identified with good quality after removing citric acid and leaving only the significant predictors. Similarly, the observations in category 3 increase after removing citric acid, the category accounting for lowest wine quality. No changes occur for category 1, or high quality. Overall, the trend seems to suggest the exclusion of citric acid yields better results for wine quality.

POLYNOMIAL REGRESSION

We used polynomial regression in case the relationship between the predictors and the wine quality is non-linear. This way, we could improve upon our results from the linear regression model. Polynomial regression models the relationship between the response variable and predictors using a polynomial function of any degree, as shown in the figure below. It is often used to extend a linear model, by adding polynomial terms to the linear regression model. Also, the coefficients in a polynomial regression model can be easily estimated using least squares linear regression because each polynomial term can be substituted with a linear term. Larger degrees of d (like greater than 3 or 4) can allow us to produce a very non-linear curve that is highly flexible. This could allow our model to fit well to our data.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i,$$

I created polynomial regression models using the predictors selected using best subset selection. I tested polynomial models with degrees of 0 through 20 and compared their prediction accuracies using k-fold cross validation. As shown in the graph below, the prediction accuracy on the y-axis decreased as the degree of the polynomial increased. Additionally, the highest prediction accuracy was 0.29389448 when the degree was 1 and the model was linear. This implies that the relationship between the wine quality and predictors is linear.

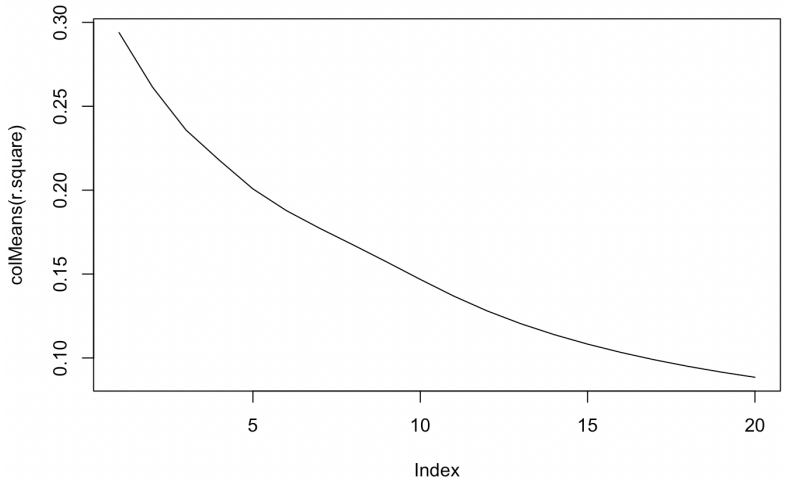

Figure 5: Plot that shows the relationship between the degree of the regression model (on the x-axis) and the prediction accuracy (on the y-axis) of the polynomial regression models

RIDGE REGRESSION

Lastly, we used lasso and ridge regression in order to shrink the coefficient estimates towards zero. This reduces the flexibility of the model, leading to a decreased variance but increased bias. Ridge regression has a shrinkage penalty that is small when the coefficients are close to zero, therefore favoring lower coefficients. The tuning parameter λ controls how much impact this shrinkage penalty has. When $\lambda = 0$, the penalty term has no effect and the ridge regression produces least square estimates. As $\lambda \rightarrow \infty$, the impact of this term grows and the coefficients approach zero. Therefore, selecting the right value for λ is critical. I trained ridge models for different values of lambda and compared them to identify the best value of lambda. I calculated the training and validation r-squared scores for each of the ridge models. Then, I plotted how the training and validation r-squared values changed with respect to ln(lambda).

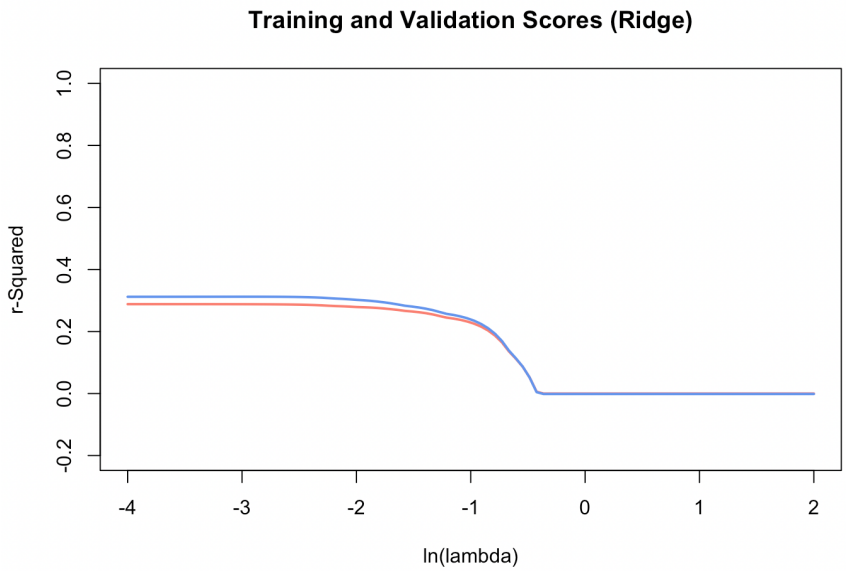

Figure 6: Plot that shows the relationship between the r-squared value of a ridge regression model (on the y-axis) and the value of ln(λ) (on the x-axis)

I found that the optimal r-squared value was 0.312575 when λ=100. This corresponds to the point where ln(λ)= -2 on the graph above. This shows that the ridge regression model that fits the data best has a low variance and high bias.

LASSO REGRESSION

A disadvantage of ridge regression is that it includes all p predictors in the final model instead of helping to select predictors like best subset, forward stepwise, and backward stepwise selection. This is because the penalty term will shrink the coefficients towards zero, but never actually set any of them to zero. This can make it difficult to interpret a model if it has many predictors. Lasso regression solves this problem. Like ridge regression, lasso regression shrinks the coefficient estimates towards zero. However, lasso forces some of the coefficient estimates to be exactly zero if the tuning parameter λ is large enough. Therefore, lasso regression can help us with variable selection, unlike ridge regression. It also provides us with models with sparse models that are easier to interpret. I started off by training 100 LASSO models with different values of λ and including all predictors in the model. Then, I generated predictions based one each LASSO model that I trained. Finally, I calculated the training and validation r-squared scores for each LASSO model. The optimal r-squared value was 0.3122914 when λ=83. This corresponds to ln(λ) = -2.969697 in the plot below. Therefore, the best lasso model had a high value for λ and therefore had a lower variance and higher bias. However, it had a higher variance and lower value of λ than the optimal ridge regression model. Additionally, the lasso regression model removed the predictors: the dummy variable representing if the wine is red, free sulfur dioxide, total sulfur dioxide, and citric acid.

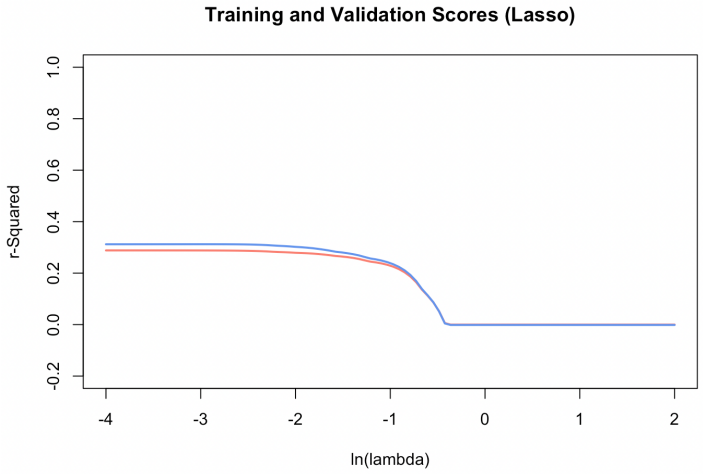

Figure 7: Plot that shows the relationship between the r-squared value of a lasso regression model (on the y-axis) and the value of ln(λ) (on the x-axis)

CONCLUSION

Our first major finding was that citric acid was the weakest predictor of wine quality. Forward, backward and best subset selection agreed that citric acid was the weakest predictor of quality. The model excluding citric acid had an R^2 value of 0.2953, higher than the model including citric acid. Our results from ordinal logistic regression also showed that by removing citric acid as a predictor, we predicted more data points correctly. Specifically, our model with all predictors, classified 5264 wines correctly into low, medium, and high quality. Conversely, we correctly classified 5268 wines correctly with our model excluding citric acid. Lastly, using lasso regression, our most accurate model excluded citric acid, and 3 other predictors: the dummy variable representing if the wine is red, free sulfur dioxide, and total sulfur dioxide.

Another major finding was that a linear regression model performed best on this data. Our linear regression model excluding citric acid had an adjusted prediction rate of 0.5698385. Additionally, we found that our model's prediction accuracy decreased as the degree of our polynomial regression model increased. This shows the data is likely linear. We also found using the ridge and lasso regression models that models with higher values of λ, and therefore lower variance and higher bias, performed better on our data.