Applying Linear Regression Methods to Model the Quality of Red Wines

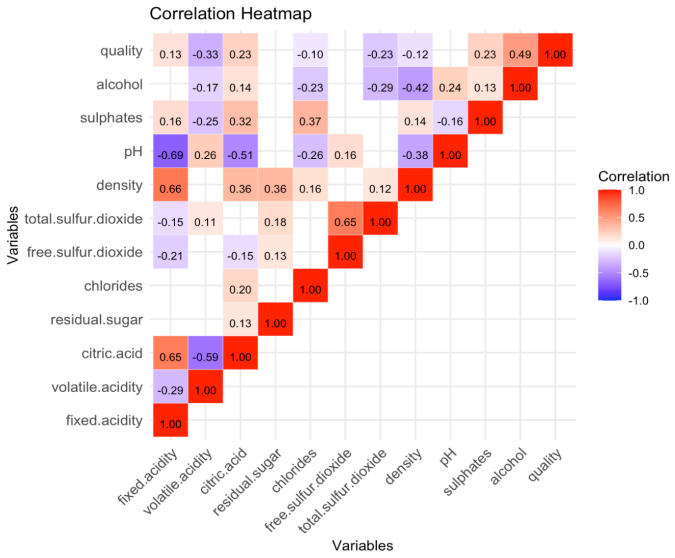
Abel Tekle

According to “5 Most Consumed Alcohols in the World,” by Sultan Khalid, wine is the second most consumed alcoholic beverage in the world. In the study, Khalid analyzed the 2022 global market sizes of different types of alcoholic beverages, as they relate directly to the consumption levels, to make an accurate list of the top five beverages. It is no wonder that with a 2022 global market size of $441.6 billion, winemakers worldwide would aim to continue to produce quality wines every year.

The Red Wine Quality dataset, provided by the UC Irvine Machine Learning Repository, contains 1500 samples of red wines from Portugal. For each of the 1500 samples, the dataset includes a quality score, as rated by experts, and 11 chemical properties that make up the wines, such as the alcohol content, the pH level, and the density of the wine. The goal of this study was to provide an elementary analysis of the Red Wine Dataset by fitting regression models to the dataset in an attempt to predict the quality of red wines accurately. In order to do this, five steps were taken over the course of 10 weeks.

Getting to Know the Data

In order to begin an analysis of the dataset, in Step 1, we first developed an understanding of the 12 variables in the dataset, and how they relate to one another. More importantly, we sought to determine the relationships of the chemical properties of the wine with the target variable, quality, which is a rating of the wine from 1 to 10, as rated by experts. This included plotting a correlation heat map, a graph of the correlations, or how each variable was affected by every other variable. The heat map allowed for a visual of the variables with the highest correlation with the quality of the wine, by color coordinating each correlation to be darker red if the variable had a high positive relationship with the quality, and dark purple if the variable had a high negative relationship with the quality.



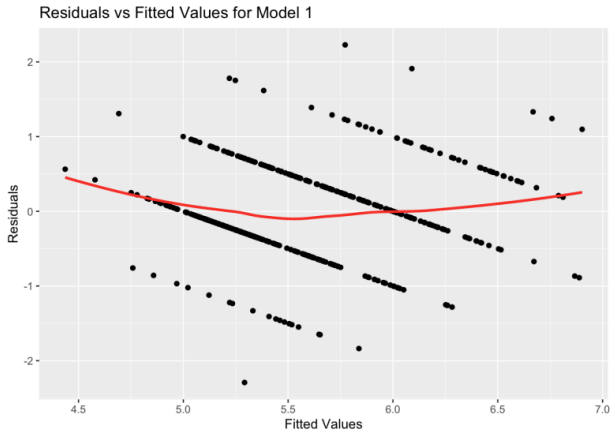
The correlation heatmap presented allowed for us to grasp an idea of the most influential factors of the wine on the quality of the wine. For example, the percentage of the alcohol in the wine, or alcohol, had the highest correlation with the quality of the wine. The correlation, 0.49, was positive, suggesting that the higher the percentage of alcohol in the wine, the higher the rating would be. Similarly, volatile acidity, or the amount of acetic acid in the wine, had a relatively high correlation, however, it was negative, suggesting that the more acetic acid in the wine, the lower the wine would be rated. This was reasonable, as too much acetic acid in a wine can lead to an unpleasant, vinegar-like taste.

Modeling

In Steps 2 and 3 of the study, we aimed to introduce linear modeling to the data, approximating equations that would allow us to enter values of the different chemical properties in order to approximate the quality of wine. In Step 2, we focused on simple linear regression, in order to gain an understanding of how modeling works with respect to the Red Wine Dataset. As a simple example, rather than approximating the quality of the wine, the density of the wine, measured in grams/decimeters^3, was approximated with alcohol as the predictor. Density was decided as the target or response variable, as it has a high correlation with alcohol and it is a qualitative variable. While quality is a score from 1 to 10, it is qualitative, indicating that it can't be measured in units, but represents the satisfaction of the experts that rated the wine. This was purely a learning experience aimed to prepare us for multiple linear regression.

For Step 3, we started to apply multiple linear regression to the dataset. While we considered models based solely on interest, we primarily focused on models chosen through statistical methods. In order to select the variables for the model, a stepwise approach was used. This automated approach took a model that predicted quality using all of the chemical properties in the dataset, and step-by-step, consider adding or removing variables to pick the variables best suited for prediction. This produced the stepwise model, or Model 1. Model 1 consisted of the volatile acidity level, amount of residual sugars, sulfur dioxide level, amounts of sulphates, and alcohol content as the predictors.

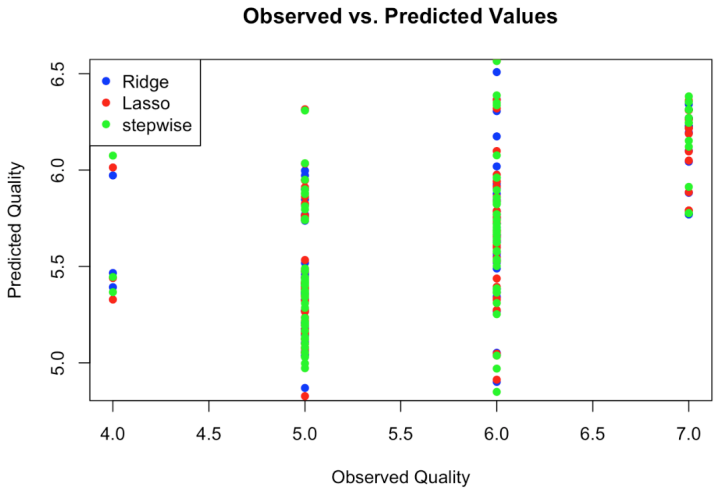
In order to determine that the model was reliable in its ability to predict the quality of the wines, the assumptions of the errors or residuals of the model were checked. This was done by examining a residuals vs. predicted plot of the model.



The residuals plot for Model 1 shows the differences between the observed quality scores and those predicted by the model. The red line represents a "loess" fit to the residuals, which should ideally be flat if the model's assumptions are fully met. However, the curve in the red line suggests that there are systematic deviations from the model's predictions, particularly evident at the extremes of fitted values. This could mean that the relationship between the predictors and response is not perfectly linear or that other variables not included in the model could better explain the variability. Nevertheless, since the model was more reliable than any of our earlier experimental models, we chose to move forward with the stepwise model. Below is the fitted linear model rounded to three decimal places(see Step 3 for a summary of the model)

$$quality_i = 2.672 - 0.96*acidity_i + 0.066*residual.sugar_i - 0.003*total.sulfur.dioxide_i + 0.607*sulphates_i + 0.29*alcohol_i$$

In step 4, shrinkage methods were introduced to the dataset in the hopes of improving the predictive power of the model. To do this, both Ridge Regression and LASSO Regression were performed on all of the variables of the dataset. The Ridge Regression model, or Model 2 showed slightly higher coefficients for most variables, indicating less penalization, compared to the LASSO Regression model, Model 3, which set several coefficients to zero, creating a much smaller model. The stepwise model, introduced in the previous step, selected a middle ground, with a subset of predictors based on AIC. In terms of MSE, our stepwise model showed an MSE of 0.42, while the Ridge and LASSO models resulted in MSEs of 0.388 and 0.389, respectively. The Ridge model has the lowest MSE and is considered to have the best predictive performance on our test set. Considering the complexity, the LASSO model provided a more simple solution with fewer predictors. Below is a plot of the predicted values vs. the observed values of quality for Models 1 through 3.



As indicated by the plot, the predictions resulting from the Ridge Regression, LASSO Regression, and stepwise models are quite close to each other, as indicated by the overlapping blue, red, and green dots. This indicates that Models 1 through 3 performed similarly when applied to the dataset. Overall, the graph indicates that applying the shrinkage methods(Ridge and LASSO) and stepwise selection have provided regression models with comparable predictions of wine quality.

The final goal of our study was to incorporate a machine learning method with the intention of creating a more complex model that provided more accurate predictions of the quality of wine.

For this, we decided to build a basic neural network. Due to the complex nature of wine quality determination, the flexibility and adaptability of neural networks make them ideally suited for this task. As expected, Model 4 produced the superior R^2 and MSE values, suggesting that to an audience looking to learn more about red wine, Model 4 would be the best model for predicting the quality.

Conclusion

Arguably the most intriguing discovery in the analysis was that while the models selected provided some insight into the quality of wine, at most 36% of the variability was accounted for, as indicated by the R^2 values. Considering additional factors of wine such as the age of the wine or the location it was made(as all of the samples from the database are from Portugal) and exploring more advanced modeling techniques may provide more information regarding the subtle changes in the characteristics that compose the wines. Additionally, the majority of the wines fell in the 5 to 6 range of scores, with the median being a 6 and the mean being a 5.61, and none falling outside of the 3 to 8 range. This eliminates any insight as to what makes a very highly-rated wine(rating of 9 or 10), or, subsequently, a poorly rated wine(rating of 1 or 2). Ultimately, the models created in the study would be helpful in understanding the basics of rating red wine, and could help winemakers in the more general selection of the characteristics of their wine to ensure a high quality.

References:

"5 Most Consumed Alcohols in the World" by Sultan Khalid:
https://www.insidermonkey.com/blog/5-most-consumed-alcohols-in-the-world-1204057/4/

https://finance.yahoo.com/news/20-most-consumed-alcohols-world-055902397.html

UC Irvine Machine Learning Repository: Red Wine Quality
https://archive.ics.uci.edu/dataset/186/wine+quality