# Understanding Wine Quality: Talking Points

Alison Lanski

Spring 2019

## Introduction

When people think of wine quality, they might immediately turn to the "point" system, which rates wines out of a scale of 100. Factors that go into this rating include color, clarity, nose, flavor, finish and more. All of these elements are determined directly by people. But what if we wanted to understand where these kinds of ratings come from, scientifically? The wine dataset, with information about both red and white wines, provides scientific measures of 11 features related to wine. With this objective and testable information, we can explore what underlying chemistry takes a wine from low to high quality in the mind of the tasters. With a promising model, we could even test new wines before drinking them to provide an preliminary categorization and set expectations.

The wine data has 6497 wines included, with 1599 red wines and 4898 white wines, all of which are variants of the Portuguese "Vinho Verde" wine[[1]](#footnote-0). Twelve variables are available, including the wine's color, ABV, pH, density, and quantities of sulphates, chlorides, total and free sulfur dioxide, volatile and fixed acidity, residual sugars, and citric acid. All wines have measures for all variables. The wines were originally rated using sensory factors on a scale from 0 to 10; for this analysis, I have collapsed ratings of 0-5 as "low quality" and 6-10 as "high quality". To start, I split the data into a training set (60% of wines) and test set (40% of wines) to allow better testing of predictions and avoid a situation where the model fits only this particular dataset. I explore all variables through graphing, with a selection in the slides. Next, I begin to find an appropriate model that explains quality, first by using all variables but quickly eliminating several to simply the analysis. Then, I evaluate the characteristics of the model and assess its performance. Finally, I discuss the predictive power of the model with reference to the test data, make a final recommendation, and conclude.

## Exploring the Data

* Split data into training/test to avoid overfitting the model and to make decisions without knowing all outcomes
* Graph the values of each variable and use color to show which wines are high or low quality: two examples here
* Sulphates scatterplot: does not distinguish very well between high and low quality
  + sulphate quantity (the y axis) for each wine (x axis) in the data. Orange indicates low quality; green for high
  + Points are jumbled together in a single band, mostly
* Interleaved histogram tells the same story:
  + sulphate quantity is on the x axis and the height of each bar shows us how many wines (out of the % of the total) have that particular quantity.
  + Each pair of orange and green bars shows how many wines of high or low quality have that particular amount of sulphates:
  + the bars are mostly of similar height
* ABV distinguishes between low and high quality wines better than sulphates:
  + the scatterplot dots have a wider range, with green (high quality) more common in the higher band and orange (low quality) more common in the lower band.
  + The pairs of histogram bars also show different heights and patterns

## Variable Selection

* Use all the variables as predictors (training data) and run step (AIC) to simplify: tests each variable in turn and see if if the AIC goes down by enough to remove something *(model deviance + 2 \* (number of predictors + 1)*; keeps doing this until the AIC stops improving.
* AIC only removes one item; I also remove chloride b/c the AIC value is the same and the deviance is only slightly worse (model 3)
* Want an even simpler model: look at correlations to reduce multicollinearity. Remove total sulfur dioxide and density to eliminate some strong correlations (model 4), then rerun AIC one more time, remove fixed acidity as a result (model 5)
* Model 5 is the final model: AIC and Deviance are highest out of models shown, but not by a large % difference -- worth it because only 7 predictors is much simpler than having 12 and should be more stable after removing some correlated variables
* AIC results were surprising --- (1) sulphates stays in even though graphs showed poor discrimination and (2) clearly correlated variables like the sulfur dioxides were both kept. I see no clear reasons for what AIC chose to remove. Maybe a wine expert would understand.

## Assess Model Fit

* HL test compares predicted probabilities with the actual classification of low or high quality wines: essentially producing a sum of errors; this is used as a test stat in a chisq distribution to test if the model fit is acceptable
* HL test: p value 0.4131 > alpha = 0.05: keep the null, assume model fit is reasonable
* Test model assumptions via loess plots to see if the continuous predictors are linearly related to the logodds of the model --- all of the plots fail
* In each plot:
  + X axis shows the value the named variable
  + Y axis shows the logodds when we use that variable only to predict wine quality, and when we do so with localized regression, e.g. instead of running our model on all points at the same time, we run it repeatedly on groups of points that are close to each other -- this allows for varying intercepts and beta-hats
* To resolve: break each variable into pieces (splines) that can be graphed with a linear relationship: quadratics into 2 pieces, cubics into 3 pieces, quartics into 4 pieces, then rerun the model with these new values
* Re-test with HL --- model fit still okay

## Model Inferences

* Splines make the model more difficult to interpret. For example, each spline within a model has a different intercept, error, and p-value which increases the complexity and are not necessarily meaningful individually
* The pvalues shown are for the predictors that have the strongest overall pvalues; the ones not shown: 2 of the 4 residual sugar are significant; 2 of the pH are significant, 2 of the sulphates are significant. The intercept has a p-value of 0.15.
* CIs can be calculated but don't mean much
* Some parameters are a lot larger/smaller than others (in magnitude) because the data is on different scales
* Cook's Distance: See how much removing a single point moves the model
  + Add: (model estimates - model estimates without a particular wine)^2
  + Divide by: number of parameters \* mean squared error of the estimates
  + Look for anything > 3 in magnitude (we don't have any)

## Assess Predictive Power

* ROC curve has AUC > .8 which means the model has "excellent" discrimination between high and low quality wines (1 is max, 0.5 is min)
  + computed by comparing how many high quality wines we get right with how many low quality wines get predicted as high quality
  + do this 1000 times, rounding model predictions based on a split point of 0.001 all the way up to 0.999
* Next plot shows predicted probabilities vs outcomes: upper right is actual high quality wines predicted as high quality, upper left is actual high quality wines predicted as low quality; etc.
  + Red color indicates where the model was correct (rounding at 0.5)
  + Point distribution shows that we identify a higher % of High-Quality wines than Low-Quality wines; we also get more right (overall) than wrong
* To try our model with the test data, must re-fit splines because the datasets are different sizes
  + use the same break points (knots) calculated from training data
  + refit the model with the same variables
* Model fit is still acceptable
  + HL test accepts the null
  + ROC curve looks good with AUC of 0.822 -- even better than the training data
  + Accuracy rate is .76 vs .75 for original splined
* Interpretation of final model
  + wines with high odds of being classified as high-quality are red wines with:
    - lower volatile acidity
    - higher residual sugar
    - higher free sulfur dioxide
    - lower PH (more acidic)
    - higher sulphates
    - higher ABV

## Conclusion

In the end, this model does a reasonable job of explaining what makes a wine high-quality. All tests (HL, AUC, Accuracy) are acceptable or good, and the model produced on training data fits the test data equally well, if not better, which suggests that we do not have a situation of overfitting.

I would say yes, this model is satisfactory for this data. However, there are many ways the situation would be improved. The data is imbalanced because there are about 3x as many white wines as red wines. Additionally, there are almost 2x as many high quality wines as low ones. This means that high-quality white wines are overrepresented in the model and may bias the results. It might be more effective to create evenly balanced bootstrapped samples to fit future models. I would also be willing to consider other modeling approaches, like a mixed model to separate wine types (red/white) or other classification methods like nearest neighbor or random forest.

In addition, all data is only using Portuguese "Vinho Verde" wine, which means that the results may not be suitable for application to other types of grapes. That said, it would be interesting to apply this model to a new test set with other grapes to see how consistent the relationships are between chemical composition and quality. If the model fits well, that would indicate that all grapes and wines share fundamental chemical relationships with quality. If the model fits poorly, I would want to re-run with a wider range of grapes in the data. This analysis might help produce a more general sense of what makes a wine high-quality instead of limiting our interpretations to this single varietal.

Finally, I would like to consult with an actual wine expert or a chemistry expert on any results to find out if 1) if there are explanations for why these variables might matter and 2) if the resulting features that are predictive of a high quality wine are related to any particular flavors or nose, e.g. from reading a bottle's description, could you apply the information from this model to select a high quality wine over a low quality wine.

1. Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). "Modeling wine preferences by data mining from physicochemical properties." *Decision Support Systems*, Elsevier, 47(4):547-553. [↑](#footnote-ref-0)