Remy Lagrois

PCA Project

March 28, 2016

Deconstructing Wine: How Chemical Make-Up Affects Quality

**Introduction**

The top tier judges of wine make are a rarified group. It can years of practice and many attempts at the test to become a Master Sommelier and many never make it. Part of the problem is the distinguishing characteristics of wine are very subtle and taste is a subjective sense. Preconceptions can have a strong effect; knowing the price of the wine, for example, actually changes how your mind perceives and experiences the tasting experience[1]. It is therefore of interest to determine if and how objective properties of wine affect its (perceived) quality. Such an analysis could help determine if one wine really is of higher quality than another and help in formulating better wines. In order to do this eleven different chemical properties of wine were measured and each wine was tested by a panel of experts on a scale of 0 to 10. The median score was then recorded and used for analysis[2]. In our analysis we will use their data for Principal Component Analysis (PCA) and selected principle components will be used to fit a regression model.

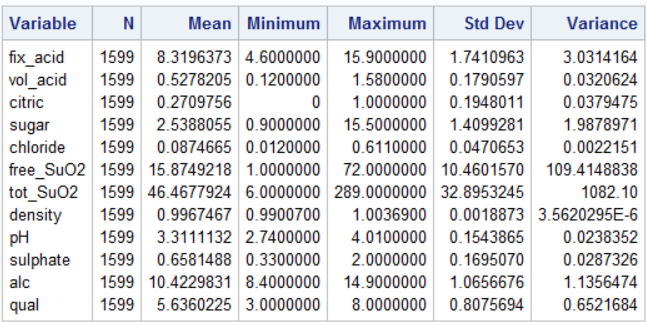
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable | Label | Notes | Variable | Label | Notes |
| fixed acidity | fix\_acid |  | density | density |  |
| volatile acidity | vol\_acid | Carbonic Acid | pH | pH |  |
| citric acid | citric |  | sulphates | sulphate |  |
| residual sugar | sugar |  | alcohol | alc |  |
| free sulfur dioxide | free\_SuO2 | SuO2 not bound | chlorides | chloride |  |
| total sulfur dioxide | tot\_SuO2 | Bound + unbound SuO2 | quality | 10\_qual | Transformed to 100 pt scale |

*Fig 1. The variables used for analysis*

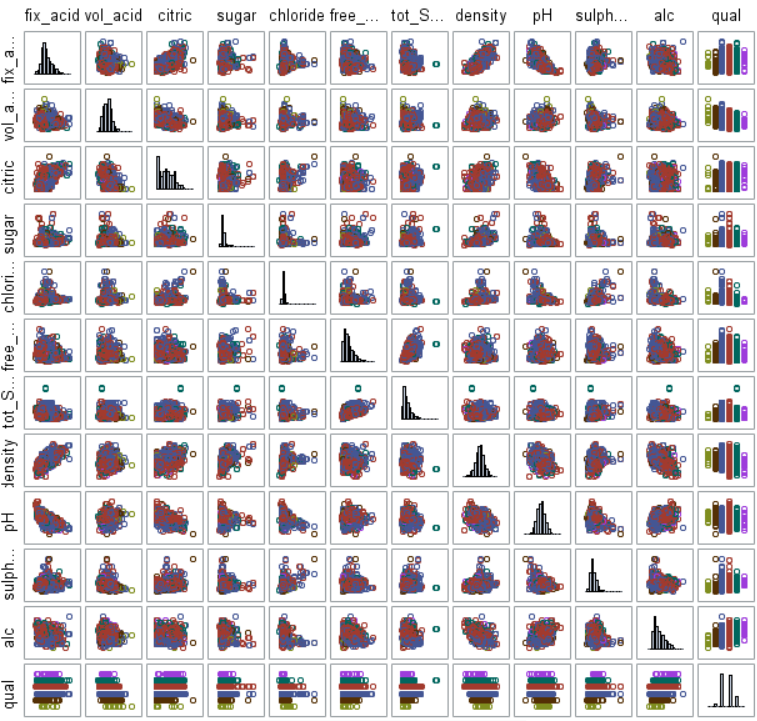
At this time rating quality is still ultimately subjective. Using multiple experts and the median score helps to at least determine a ‘consensus’ opinion that is not heavily influenced by strong bias and outlier ratings. This analysis doesn’t hope to produce some formula which is guaranteed to produce a perfect wine; rather we hope to find which characteristics would lead a plurality of experts (and by extension people) to rate a wine higher or lower than the average.

**Basic Statistics and PCA**

*Descriptive Statistics*

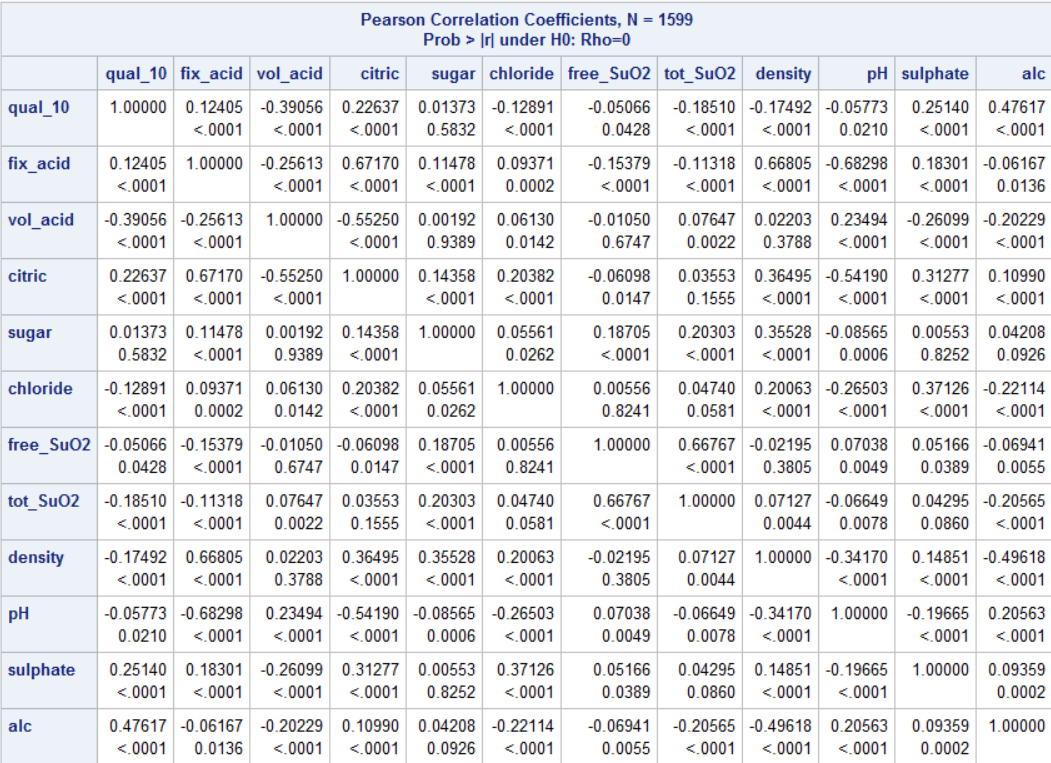


*Fig 2. Descriptive statistics of the variables*



*Fig 3. Scatterplot and histograms of the variables*

There is orders of magnitude differences in the standard deviations and means of the different variables (fig 2). This is not surprising since different variables use different units of measurement which will have varying ranges of possible numbers. The scatter plot shows that for the most part the variables seem to have no real relationship with each other though there are several that do have collinearity (fig 3). Otherwise there is no strong evidence for any non-linear relationships. There is also no evidence against normality in the histograms.



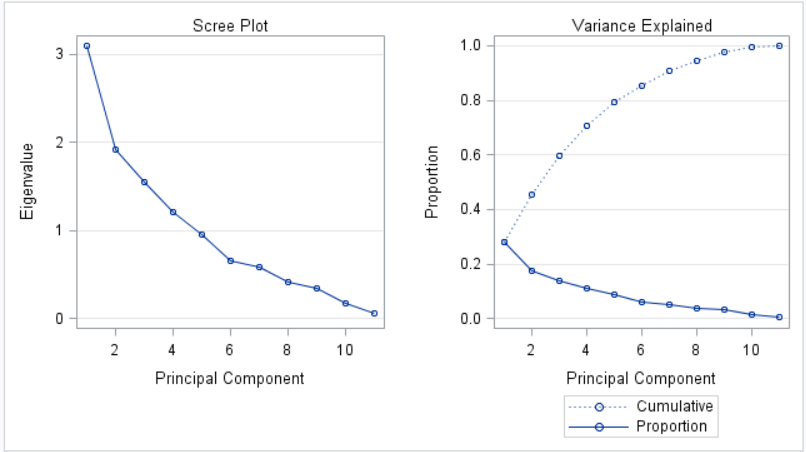
*Fig 4 Correlation of Variables*

The variables that correlate with each other make sense (fig 4). The acids correlate well with pH which is itself a measure of acid while alcohol has a negative correlation with density as alcohol is less dense than water. A few other variables have a positive correlation with density. These correspond to dissolved compounds which have a higher molecular weight than water so wines with more of these compounds will also have a higher density.

Given the number of variables and correlation between them (its existence and that it’s linear) these set of data is a good candidate for PCA using the correlation matrix and multiple linear regression with the selected principle components.

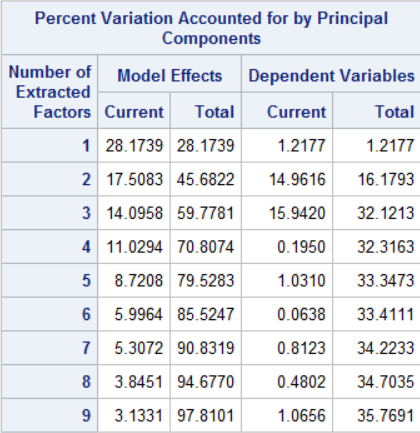
*Principle Component Analysis*

First the full PCA was run using SAS’s PROC PRINCOMP to generate the eigenvectors for the components and a Scree plot to begin to determine which components should be selected for the regression model.

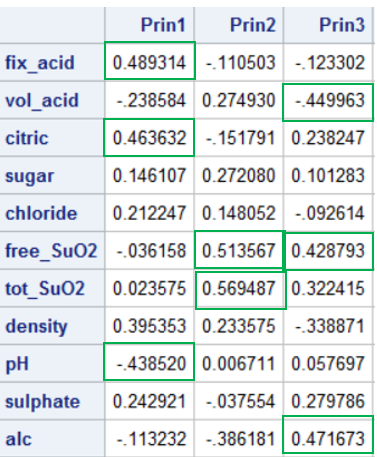


*Fig 5. Scree Plot and Variance Explained*

The Scree plot (fig 5) has no extremely obvious ‘elbow’, there seems to be an inflection point at component 6 and possibly at component 2. The proportion of variance explained also has no obvious leveling off point. The increase slows around points 3-5 and slows even more after component 6. In order to more clearly determine which components to select the PROC PLS with cross validation and the PCR method was run.

*Fig 6. Accounted Variation*

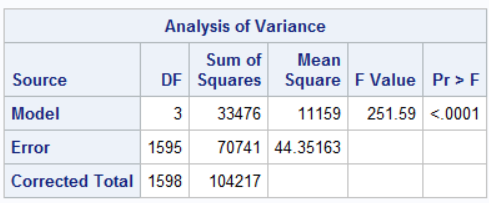
From the table (fig 6) it can be seen that the increases in explained variation starts to decay after point 4 for the model effects and drops precipitously after point 3 for the dependent variables. Given this data the first three principle components were selected to be used as factors in multiple linear regression. As a note: a regression model using the first four principal components was run. While it was significant based on the p-value, zero was barely outside the range of the 95% confidence interval and it had no effect on R2 so the final model used only the first three components.

*Fig 7. Eigenvector for Principal Components*

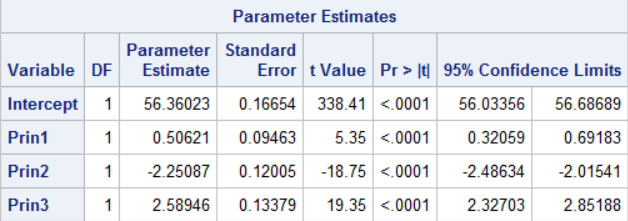
Based on the major factors in each component (fig 7) I would base these as being representative of the following: Prin1 is clearly acidity. The citric acid and the fixed acid are both important while pH has a negative relationship (however as pH increases acidity decreases). Acid would provide more of a sour flavor. Prin2 is the combination of the two measures of sulfur dioxide. Sulfur dioxide is important for keeping wine from going bad due to bacterial growth. However in higher concentrations it can begin to affect the smell and taste of the wine. Prin3 is less clear. It could be either something like overall flavor or age. Higher alcohol content would be due to a longer fermenting process. Higher concentrations of free sulfur dioxide would give the wine a longer shelf life which may explain the positive relationship. Also the volatile acid would increase overtime due to low levels of microbial activity and could give the wine a carbonated feel, even if subtly.

**Multiple Regression**

The model being tested is quality = β0 + β1 Prin1 + β2 Prin2 + β3 Prin3. The coefficient β0 represents the mean wine quality given the data while the other three represent the magnitude and direction (negative or positive) of each principal component used in the model. The F-test resulting from this model is significant (p-value of <0.0001) showing that our model is apporpriate for explaining at least some of the variation (fig 8).



*Fig 8. Overall F-test*

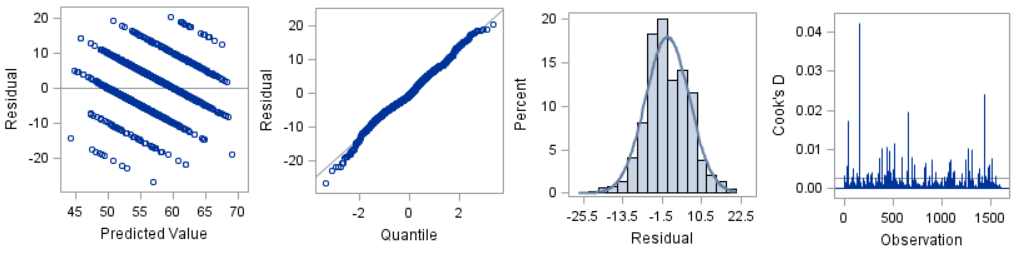


*Fig 9. Parameter Estimates for regression model*

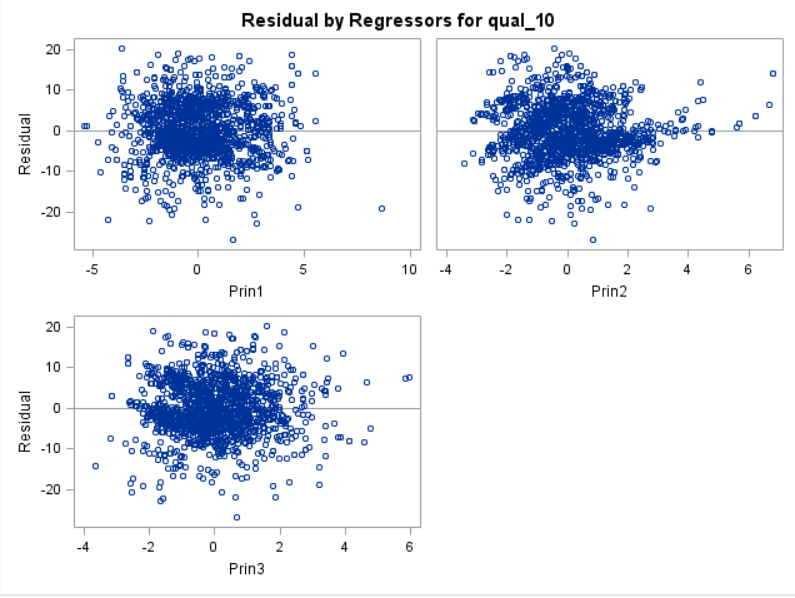
All three of our parameter estimates are significant, each has a p-value of <0.0001 (fig 9). However none of them have a very large absolute value and so can only have modest impacts given real world constraints. For example the slight positive value for Prin1 (0.506) indicates acidity has some small positive effect, presumably increasing acidity would have a positive effect on quality – up to a point. Eventually the sour flavor will start to detract from the quality and going even further you’d end up with a bottle of acid. Prin3 also has a positive value (2.589) which indicates increased age also has a positive effect on quality. It would appear though that increasing sulfur dioxide has a deleterious effect on quality with Prin2’s negative value (-2.251).



*Fig 10. R2 table*

 The R2 value for the model is not very high (0.32), about a third, which leaves room for quite a bit more explanation of the variation. This could be due to subjective effects but it’s difficult to comment since we do not know the details of how the experts tasted the wine beyond what has already been told.

*Fig 11. Model residuals and Cook’s D*



*Fig 12. Residuals for components*

Looking at the residual plot for the model (fig 11) one thing immediately jumps out: the residuals are arranged into streaks. The reason for this is that while the quality variable is continuous it behaves similarly to a categorical variable. This is because it exists as a 10 point scale with a resolution of 1 unit (or 10 for our transformed quality), any predictions that fall between the each possible point will have to be subtracted from an actual value of which there are many of the same. Beyond the streaks though the values symmetrically distributed and have a blob like shape. The histogram and QQ plot both indicate normality. The residual plots for each of the three principal components look healthy and show no strong evidence of problems with non-constant evidence (fig 12). There is one high value of Cook’s D in fig 11 however the actual value of it is about 0.04 which isn’t high enough to cause much concern.

**Conclusions**

Even though only about a third of the variation is explained by our three principal components there are still meaningful conclusions which can be made.

References

# 

1. *Baba Shiv: How a Wine's Price Tag Affect Its Taste.* Lisa Trei, Stanford Business. <https://www.gsb.stanford.edu/insights/baba-shiv-how-wines-price-tag-affect-its-taste>
2. *Modeling wine preferences by data mining from physicochemical properties.* P. Cortez, Decision Support Systems. <http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names>