Final Project of MA5771 – Applied Generalized Linear Models

Kasia Krueger

Identifying a Good or Bad Wine: Predicting the Effect of Wine Quality from Physiochemical Properties Using Bernoulli GLMs

Section 1 Introduction

Wine is a 364.25-billion-dollar industry globally (Wine, 2020), but what makes a wine "good" or "not good"? Researchers from University of Minho, Guimarães, Portugal, studied eleven physiochemical properties of 6497 wines (red wine - 1599; white wine – 4898) and their resulting quality scores, a numerical score from 1 to 10. The data sets were collected from red and white variants of the Portuguese "Vinho Verde" wine. This report aims to find if physiochemical properties can be used to determine the quality of a wine in the absence of wine price, region, and labels. The objectives are to find which model and explanatory variables have the most predictable power in the response, and how the quality of a good wine is affected by the physiochemical properties.

Each of the quantitative variables in the data are continuous: (1) alcohol content describes the percentage of alcohol; (2) chlorides describe the amount of salt (sodium chloride - g / dm^3); (3) volatile acidity is the amount of acetic acid found in wine – too high of levels can lead to a vinegar taste (acetic acid - g / dm^3); (4) density, which is affected by alcohol, sugar, and salt content (g / cm^3); (5) fixed acidity are the primary acids in wine (do not evaporate readily) (tartaric acid - g / dm^3); (6) citric acid is added to wines to increase acidity but can add microbial instability (g / dm^3); (7) residual sugar is the amount of sugar remaining after fermentation stops (g / dm^3); (8) free sulfur dioxide prevents microbial growth and the oxidation of wine (mg / dm^3); (9) total sulfur dioxide: amount of free and bound forms of SO2 (mg / dm^3); (10) pH describes the acidity, most wines are between 3-4 on the pH scale; (11) sulphates are an additive which can contribute to sulfur dioxide gas (SO2) levels, which acts as an antimicrobial and antioxidant and maintains the flavor and freshness of wine (potassium sulphate - g / dm3). By combining two separate data sets for red and white wines, the (12) "color" (red or white) will be the qualitative variable. The response is a quality score (median of at least 3 evaluations made by wine experts) between 0 (very poor) and 10 (very excellent).

Section 2 Statistical Methods

All analyses have been conducted using RStudio Version 1.3.1093, using an overall significance level of 0.05 (95% confidence). The analysis in this report is conducted using a generalized linear model using *glm* in the stats package, *step* in the stats package, and *qresid* in the statmod package (R Core Team, 2020).

Section 2.1 Exploratory Data Analysis

Summary statistics are used to explore the data set. The summary statistics show that for the qualitative explanatory variables *color*, red and white wines have very similar means and variances, with outliers in both color categories. The summary statistics for the quantitative explanatory variables show that higher quality wines have fewer chlorides, higher alcohol content, less volatile acidity, less density, fixed acidity (tartaric acid) around the mean of 7 g / dm^3, less citric acid, less residual sugar, less free and total sulfur dioxide, and mid-range (7) pH than lower quality wines. The summary statistics show all of the variables have outliers but residual sugar, free sulfur dioxide, density, citric acid, and sulphates have outliers and max values that far exceed the mean values

Section 2.2 Generalized Linear Model

The quality of wine is categorized as a “good” wine if it has a score greater than 5 (1), and a “not good” wine if it has a score equal to or less than 5 (0). Since the outcome is binary, the generalized linear model used will be the Bernoulli (binomial) EDM with logistic regression. The estimated dispersion will be found using Pearson statistic (the sum of squared Pearson residuals) to determine if the Bernoulli generalized linear model is appropriate for this data. The final model will be selected using stepwise AIC, BIC. The quantile residuals will be obtained since they are preferred for discrete EDMs to avoid distracting patterns in the residuals, and assumptions of equal variance, normality, independence, and outliers will be checked using appropriate residual plots. The significance of each exploratory variable of the final model will be tested using the Wald test with normal distribution since phi is not estimated for binomial GLMS.

Section 3 Results and Conclusion

The initial generalized model includes all 12 variables; from these, chlorides and citric acid are not significant at 5%. The lowest AIC model determined that the generalized linear model without chlorides is the best model with all exploratory variables significant at 5%. The generalized linear model with the lowest AIC includes 11 variables, fixed acidity, volatile acidity, citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, color. The AIC step-wise procedure found the best model with AIC 6719.613 and 6800.962. However, the generalized linear model with the lowest BIC is a model with only 6 variables: volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, sulphates, alcohol, and citric acid. The BIC is a more accurate model since it penalizes models for adding too many predictor variables that are not necessary in the model. The BIC step-wise procedure found the best model with AIC 6729.596 and 6783.828. The estimated dispersion found by dividing the deviance over the residual degrees of freedom, found that the estimated dispersion [1.0346] indicates the model is not overdispersed since the estimated dispersion is close to the theoretical dispersion of 1 from the binomial distribution, therefore, the binomial distribution is reasonable for this set of data.

The assumptions are plotted using a plot of the quantile residuals against the fitted values transformed to the constant-information scale (A); a plot of the working responses against the linear predictors (B); a Q-Q plot of the quantile residuals or the standardized deviance residuals (C); and a scatter plot of the Cook’s distance *D* (D) (Figure 1). The plot of quantile residuals against transformed mu (A) is for checking the systematic component including link function, linearity, variance function, and dispersion parameter. There are no patterns apparent, and the residuals seems to be evenly distributed around 0, however, there are residuals > |3|, indicating outliers, but this condition is satisfied due to the larger residuals not being influential or affecting the data. The Working Responses vs Linear Predictors plot (B) is used to evaluate the link function. Due to the Bernoulli model having a binary response, two lines appear. Three extreme outliers have been removed from the model and the assumption is satisfied. The Q-Q plot of the sample quantiles vs. the theoretical quantiles (C) is used for evaluating the random component and homoscedasticity. The QQ-plots of residuals against the covariates do not show any trends or patterns so this assumption is satisfied. The Cook’s distance plot (D) evaluates if there are influential observations using Cook’s distance *D*. An observation is influential if its Cook’s distance *D* is > 1; the maximum Cook’s Distance is 0.0783, therefore, this assumption is satisfied. All assumptions are met and satisfied for this model.

The results of the Wald test indicates all explanatory variables in the final model are significant at 1% and necessary to predict the quality of a wine and indicates very small standard errors for all variables (Table 1).

	Z-test statistic	p-value	s.e.
Volatile acidity	28.1316	0.0000	0.0339
Residual sugar	-19.2246	0.0000	0.2379
Free sulfur dioxide	8.6842	0.0000	0.2306
Total sulfur dioxide	8.6823	0.0000	0.0075
Sulphates	6.6274	0.0000	0.0025
alcohol	-8.8913	0.0000	0.0008
Citric acid	-3.0769	0.0021	0.2215

Table 1: Wald Test Results

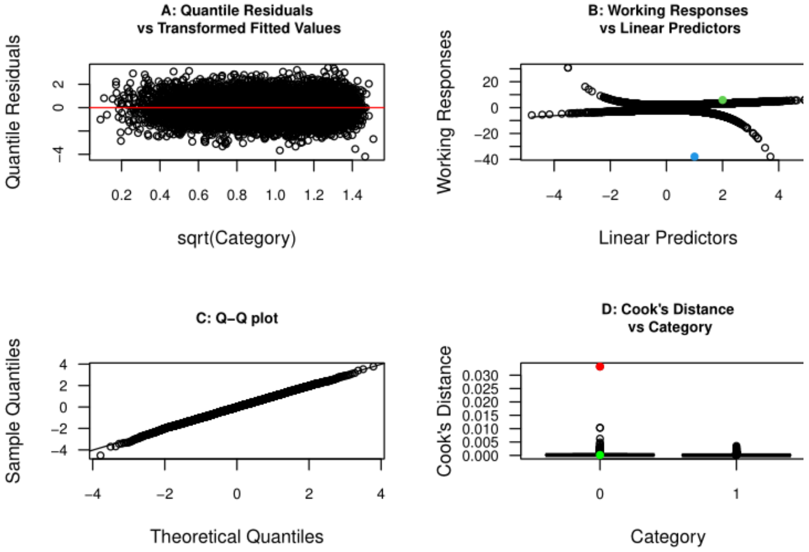

Figure 1: Model Assumption Plots

The response of the final model, quality, is affected by the exploratory variables in the following ways: **Volatile acidity**: For each unit (g / dm^3) increase in acetic acid it is exp (-4.574) = 0.010x less likely to be a good wine. **Residual sugar**: For each unit (g / dm^3) increase in residual sugar it is exp (0.065) = 1.067x more likely to be a good wine. **Free sulfur dioxide**: For each unit (mg / dm^3) increase in free SO2 it is exp (0.017) = 1.017x more likely to be a good wine. **Total sulfur dioxide**: For each unit (mg / dm^3) increase in free and bounded SO2 it is exp (-0.007) = 0.993x less likely to be a good wine. **Sulphates**: For each unit (g / dm^3) increase in potassium sulphates it is exp (2.002) = 7.406x more likely to be a good wine. **Alcohol**: For each percent increase in alcohol it is exp (0.953) = 2.593x more likely to be a good wine. **Citric acid**: For each unit (g / dm^3) increase in citric acid it is exp(-0.681) = 0.509x less likely to be a good wine. The odds of a good wine decrease when more Volatile acidity, Total sulfur dioxide, and Citric acid are present in a wine. The odds of a good wine increase slightly when more Residual sugar, Free sulfur dioxide, and Alcohol are present in a wine – and the odds of a good wine increase much more when more sulphates are present in a wine.

Section 4 Discussion

The analysis found that while all the variables from the data set with the exception of chlorides were significant in predicting wine quality, only 6 variables were included in the final model as the variables that had the most predictable power for predicting wine quality without oversaturating the model. From the highest Wald test score, the explanatory variables with the most predictable power in determine the quality of wine are alcohol content (%) and volatile acidity, distantly followed by sulphates, total sulfur dioxide and residual sugar. Increased units of volatile acidity, which can provide a vinegar-like taste, total sulfur dioxide, and citric acid decrease the quality of a wine, while increased units of residual sugar, free sulfur dioxide, sulphates and alcohol percent increase the quality of wine. This indicates that a good wine typically has less acidity and more sweetness. The most surprising result is that the probability of a good wine increases 7x for each unit increase in sulphates, which act as preservatives in a wine, and 2.5x for each unit increase of alcohol content, and that color did not have a significant effect on predicting wine quality in the Bernoulli generalized linear model. This analysis found that the Bernoulli generalized linear model with logistic link is an appropriate model of the data and found that the quality of wine can be predicted from the significant physiochemical properties in this dataset.

References

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009, https://archive.ics.uci.edu/ml/datasets/wine+quality.

Wine Market Size, Share & Industry Analysis, By Type (Sparkling Wine, Still Wine, and Others), Flavour (Red Wine, White Wine, and Rose Wine), Distribution Channel (On-trade and Off-trade), and Regional Forecast, 2020 – 2027. (2020, May). Fortune Business Insights. Retrieved November 20, 2021, from https://www.fortunebusinessinsights.com/wine-market-102836