Regression Analysis of Vinho Verde Wine Preferences

Roz Huang, Raj Jagannath, Kiara Monahan, Sean Sica

December 11, 2023

Abstract

This paper builds on previous research by employing modern statistical methods to investigate the relationship between the chemical properties of wines and their taste profiles.

1 Overview

The complexity of wine, coupled with its significant cultural and economic impact, presents a unique challenge and opportunity for data-driven analysis. Our research is inspired by a study that utilized machine learning to predict human taste preferences of Vinho Verde wines based on physicochemical properties. The original study, which applies multiple regression, neural networks, and support vector machines, paves the way for employing computational techniques in oenology - the science of wine and winemaking and serves as a basis for our investigation. It demonstrated the feasibility of predicting sensory assessments using data mining techniques, with support vector machines showing superior performance. This approach not only aids certification processes but also enhances the precision of market targeting strategies. Their methodological rigor and application of machine learning algorithms inspire our current work, where we aim to build upon their foundations with our own regression analysis.

In this paper, we build on previous research by using modern statistical methods to explore the relationship between the physicochemical makeup of wines and their taste profiles. We hypothesize that refined regression techniques can yield insights into the subtle interplay of factors that govern consumer preferences, thereby providing actionable guidance for winemakers and marketers alike.

2 Data and Methodology

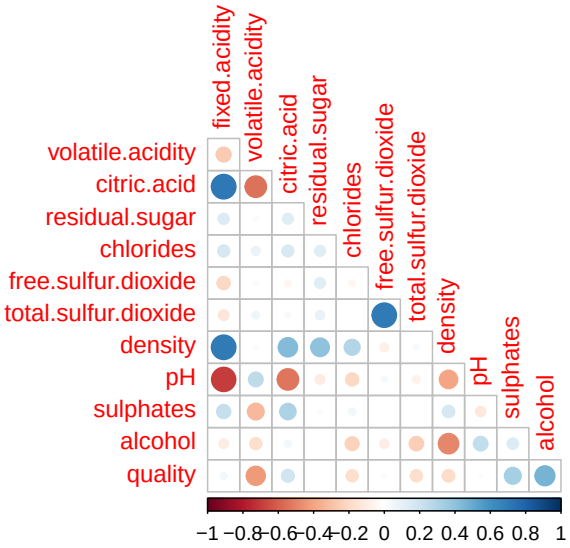
The data in this study were collected between May 2004 and February 2007 by the CVRVV, which is an inter-professional organization with the goal of improving the quality and marketing of Vinho Verde wine produced in Portugal.[1] The dataset offers a comprehensive view of the attributes that influence wine quality and consumer preference, and it is particularly notable for its volume, comprising 1599 red wine samples, each graded by expert tasters, or sommeliers. We performed all exploration and modeling on a randomly selected 30% subsample of the data. The remaining 70%, totaling 1099 rows, were used to generate the statistics in this report.

Each row in the data represents a sample of Vinho Verde wine, and the outcome variable, quality, is a value on a scale of 0-10. Each sample was evaluated by a minimum of three sommeliers and the result was the average of the ratings of the sommeliers. While there may be concerns that the quality variable is a value from the Likert scale, we have confidence in using it as ordinal since it is determined by expert tasters who provide consistency in their ratings beyond the Likert scale. In addition, by taking the average of multiple tasters, we avoid bias by any individual taster for a sample.

[1] Cortez, Paulo, et al. "Modeling Wine Preferences By Data Mining from Physicochemical Properties." University of Minho, Department of Information Systems (2009).

When considering modeling the effects of the various physicochemical properties on the quality of Vinho Verde wine, we took care to address possible collinearity between properties. To do so, we constructed a correlation matrix and removed variables that were highly correlated with others since multicollinearity can inflate variances and reduce the precision of coefficient estimates. Some variables, though available in the dataset, were excluded due to their limited direct relevance to wine quality as per existing oenological research.

Correlation Matrix for Vivho Verde Red Wine



This led us to develop three models outlined below:

- **Model 1**: Focuses on traditional factors known to impact red wine quality, such as volatile acidity and alcohol content.
- **Model 2**: Adds chlorides and total sulfur dioxide to explore the acid balance in red wine.
- **Model 3**: Adds a log-transformation of sulphates, examining the chemical balance in wine.

Our preliminary model was defined as follows: $\widehat{quality} = \beta_0 + \beta_1 \times volatile.acidity + \beta_2 \times alcohol$

where β_1 represents increase in quality per 1 $\frac{g}{dm^3}$ of acetic acid increase in volatile acidity, and β_2 represents increase in quality per 1 percent volume increase in alcohol.

In the second model, we added `chlorides` and `total sulfur dioxide`. In the third model, we added a logarithmic transformation of `sulphates`. This transformation was supported by scatterplots showing a non-linear relationship with the quality and existing oenological theory suggesting diminishing returns of sulphate addition on wine quality. We considered adding other terms in the models but found the resulting coefficients on these terms non-significant and not improving the model, so we removed them for simplicity.

The most performant model is as follows:

$\widehat{quality} = \beta_0 + \beta_1 \times volatile.acidity + \beta_2 \times chlorides + \beta_3 \times total.sulfur.dioxide + \beta_4 \times ln(sulphates) + \beta_5 \times alcohol$

where β_1 represents increase in quality per 1 $\frac{g}{dm^3}$ of acetic acid increase in volatile acidity, β_2 represents increase in quality per 1 $\frac{g}{dm^3}$ of sodium chloride increase in chlorides, β_3 represents increase in quality per 1 $\frac{mg}{dm^3}$ increase in total sulfur dioxide, β_4 represents increase in quality per 1 log-transformed $\frac{g}{dm^3}$ of potassium sulphate increase in sulphates, and β_5 represents increase in quality per 1 percent volume increase in alcohol.

3 Results

Our analysis utilizes three linear regression models to dissect the factors influencing the quality of Vinho Verde red wines. The models are constructed with key physicochemical properties as predictors of wine quality. Below is a summary table of the models:

Table 1: Regression Results

	Dependent variable:		
Log(Sulphates)			0.792^{***} (0.100)
Volatile Acidity	-1.301^{***} (0.119)	-1.283^{***} (0.118)	-1.006^{***} (0.120)
Chlorides		-0.108 (0.406)	-1.658^{***} (0.442)
Total SO2		-0.002^{***} (0.001)	-0.003^{***} (0.001)
Alcohol	0.319^{***} (0.020)	0.305^{***} (0.020)	0.275^{***} (0.020)
Constant	3.002^{***} (0.227)	3.246^{***} (0.248)	3.907^{***} (0.256)
Observations	1,099	1,099	1,099
R^2	0.306	0.314	0.351
Adjusted R^2	0.305	0.312	0.348
F Statistic	242.000^{***} (df = 2; 1096)	125.326^{***} (df = 4; 1094)	118.301^{***} (df = 5; 1093)
Note:			*p<0.1; **p<0.05; ***p<0.01

Our analysis reveals significant insights into the factors affecting Vinho Verde wine quality. The regression models for red wine present a range of statistically significant variables, with each of the variables showing a substantial impact on the quality rating:

- **Volatile Acidity and Chlorides**: Both have negative impacts on quality, with volatile acidity showing a particularly strong effect.
- **Total Sulfur Dioxide and Log of Sulphates**: These variables negatively and positively affect the quality, respectively.
- **Alcohol**: Alcohol content significantly boosts quality.

Additional statistical tests like ANOVA confirm the robustness of our models and their capacity to explain variations in wine quality. However, it is important to interpret these findings through the lens of practical significance.

- **Volatile Acidity**: Given its strong negative impact on quality, winemakers should carefully control fermentation processes to manage volatile acidity levels.
- **Alcohol Content**: The positive correlation with quality suggests a potential area of focus. However, there are legal, health, and sensory considerations that limit the extent to which alcohol content can be increased.

For red wine, the negative impacts of chlorides and volatile acidity provide actionable insights for winemakers. They suggest a nuanced approach and careful management of the additions of chlorides and volatile acids during the winemaking process. The significant impact of alcohol content and sulphates on wine stands out. These are critical factors in determining wine quality and should be a primary focus in winemaking practices. While total sulfur dioxide is significant, its relatively smaller coefficient suggests it are less critical lever for manipulating wine quality.

4 Limitations

In the pursuit of robust regression estimations to unravel the determinants of wine quality, it is important to acknowledge several limitations that may impact the precision and validity of the findings.

The dataset under examination involves red Vinho Verde wine samples originating from the northern part of Portugal. The assumption of independent and identically distributed (IID) observations is challenged by the existence of geographical clusters. These clusters, inherent to the unique terroir of the Vinho Verde region, are influenced by factors such as climate variability, soil composition, and altitude differences. Additionally, the dataset spans a collection period between May 2004 and February 2007. This introduces the possibility of temporal clustering, where wine quality measurements may exhibit similarities or patterns based on the time of collection. Factors such as seasonal variations, annual trends, or evolving winemaking practices over the almost three-year period may influence the outcomes. While acknowledging the presence of geographical clustering and temporal clustering, the current models do not fully account for it. Future investigations may explore advanced techniques like clustered standard errors or mixed-effects models, incorporating geographical fixed effects for a more nuanced understanding.

In considering potential omitted variables in the wine quality models, it is crucial to acknowledge that certain characteristics inherent to Vinho Verde wines, while influential, might not have been explicitly included in the current models. One notable omitted variable is the grape varietal. Omitting grape varietal from the models may introduce bias in the estimated coefficient of volatile.acidity, a key predictor negatively associated with wine quality. In general, it is typically expected that grape varietals with higher acidity contribute positively to the overall quality and sensory profile of the wine. Grape varietals with naturally higher acidity may also exhibit higher levels of volatile acidity. Omitting grape varietal could inflate the estimated coefficient of volatile.acidity. This inflation suggestions a larger negative impact on wine quality than is true.

While the presence of volatile acidity may have causal effect on wine quality, it is also an outcome variable as alcohol can be metabolized to ethyl acetate, which is the source of volatile acidity. Conversely, sulfur is known to decrease bacteria load and increased sulfur may prevent volatile acidity.[2]

5 Conclusion

Our study presents linear regression models that link physicochemical properties to the perceived quality of Vinho Verde wine, developed through correlation matrices and optimizing for high R^2 and significant coefficients. Recommendations for wine quality improvement include increasing alcohol content, reducing volatile acidity, and boosting sulphate content while lowering chlorides. These findings, while subject to limitations like geographical/temporal dataset clustering and exclusion of variables like grape varietal, offer valuable insights and guidelines for stakeholders in the wine industry. Despite its limitations, our research contributes significantly to oenology, offering a deeper understanding of wine quality factors and practical advice for winemakers, marketers, and wine consumers. Future research could overcome current limitations by integrating advanced techniques and additional variables like grape varietal, thereby enriching the scientific study of wine and its quality dynamics. This continued exploration in oenology may benefit many stakeholders in the wine industry.

[2]Horton, Drew and Clark, Matthew. "Volatile Acidity in Winemaking." University of Minnesota (2016).