Wine Quality Regression Analysis

Diogo Pimenta - nº 20230498, João Maia - nº 20230746, Pedro Catarro - nº 20230463, Pedro Luis - nº 20230797

[1] Statistics for Data Science, NOVA IMS

Introduction

The wine industry is experiencing significant growth due to the increasing popularity of social drinking. However, the current certification process, relying on human experts and physicochemical tests, is time-consuming and costly. The subjective nature of wine appreciation and variability in opinions among tasters further complicate the process. This project seeks to identify key indicators of wine quality by examining the relationship between human tasting preferences and physicochemical properties. The goal is to enhance and streamline the certification and quality assessment processes in the red wine market.

Research Question

"How do different physicochemical properties, such as acidity, sulfur content, and alcohol, relate to the quality of wine?"

Methods

In our group project, we undertook a comprehensive regression analysis on a cross-sectional dataset to explore the relationships between various properties and wine quality. We began with constructing a Linear Regression Model, which served as the foundation for predicting wine quality based on a range of predictor variables. To diagnose potential issues related to heteroscedasticity, we applied both the Breusch-Pagan Test and the White Test, with the former assessing constant variance in residuals and the latter investigating patterns involving squared and interaction terms. The RESET Test was utilized to evaluate the potential enhancement of our model by incorporating non-linear terms. In response to concerns related to varying residuals' variance, Robust Estimation techniques were collaboratively integrated.

Results

Table 1: Tidy Coefficients Table

term	estimate	std.error	statistic	p.value
(Intercept)	21.9652084	21.1945750	1.0363599	0.3001921
fixed.acidity	0.0249906	0.0259485	0.9630827	0.3356528
volatile.acidity	-1.0835903	0.1211013	-8.9478019	0.0000000
citric.acid	-0.1825639	0.1471762	-1.2404449	0.2149942
residual.sugar	0.0163313	0.0150021	1.0885992	0.2764960
chlorides	-1.8742252	0.4192832	-4.4700697	0.0000084
free.sulfur.dioxide	0.0043613	0.0021713	2.0086353	0.0447450
total.sulfur.dioxide	-0.0032646	0.0007287	-4.4798298	0.0000080
density	-17.8811638	21.6330999	-0.8265650	0.4086079
pH	-0.4136531	0.1915974	-2.1589710	0.0310019
sulphates	0.9163344	0.1143375	8.0142971	0.0000000
alcohol	0.2761977	0.0264836	10.4290143	0.0000000
R^2	0.3605517	NA	NA	NA

The model suggests that volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, and alcohol significantly influence the response variable.

Volatile acidity, chlorides, total sulfur dioxide, and pH have negative associations with the response variable, meaning an increase in these factors tends to decrease the response.

Free sulfur dioxide, sulphates, and alcohol have positive associations with the response variable, indicating that an increase in these factors is associated with an increase in the response.

The model explains about 36% of the variance in the response variable (moderate fit).

The density is not statistically significant, suggesting it may not be essential in the model.

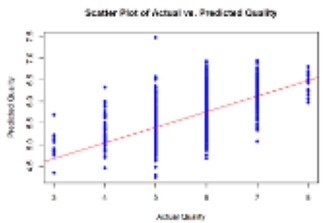

Figure 1: Figure 2

Here we can also visualize how well our data fits the regression line. The closer the points are to the line, the better the fit. This supports our previous conclusion that the model has a moderate fit, based on the R^2 value.

Further issues in our model

We will focus on investigating heteroskedasticity – a condition where the variability of errors is not constant across independent variable levels. This assumption violation can affect the reliability of our model. We will employ employ the Breusch-Pagan Test and the White Test to diagnose potential issues related to heteroscedasticity.

Table 2: Heteroskedasticity Test Results

test	statistic	p_value
Breusch-Pagan Test	84.98902	0
White Test (Full)	115.08705	0
White Test (Special)	115.08705	0

From these results we can conclude that we reject the null hypothesis of homoskedasticity for a 5% significance level. This means that our model is not reliable and we need to use robust estimation techniques to correct for heteroskedasticity.

Robust Estimation

This is a technique used to provide reliable parameter estimates when the assumption of homoskedasticity is violated. We will use the robust standard errors to provide reliable standard errors for coefficients.

Table 3: Robust Standard Errors

	coef.modelo.	robust_se...2.	p_value
(Intercept)	21.9652084	24.7168339	0.3741783
fixed.acidity	0.0249906	0.0327063	0.4448135
volatile.acidity	-1.0835903	0.1380015	0.0000000
citric.acid	-0.1825639	0.1540859	0.2360890
residual.sugar	0.0163313	0.0195255	0.4029262
chlorides	-1.8742252	0.4897887	0.0001299
free.sulfur.dioxide	0.0043613	0.0022685	0.0545361
total.sulfur.dioxide	-0.0032646	0.0007685	0.0000216
density	-17.8811638	25.2915763	0.4795660
pH	-0.4136531	0.2153875	0.0547939
sulphates	0.9163344	0.1367898	0.0000000
alcohol	0.2761977	0.0292505	0.0000000

The robust standard errors provide an adjusted measure of the standard error, taking into account the heteroskedasticity in the model. Some coefficients like volatile acidity, chlorides, total sulfur dioxide, sulfates and alcohol, remain highly significant in the model, even with robust standard errors. This might suggest that these variables are important predictors of wine quality. The standard errors coefficients have increased with the robust estimation, which indicates that the standard errors might have been underestimated duo to heteroskedasticity, now providing a more reliable measure of the standard error.

Table 4: RESET Test Results

p_value	r_squared
0.0001189	0.3671343

The p-value from the RESET test assesses whether the model is adequately specified. In the case for the second degree term, the p-value is less than 0.05 indicating that we reject the null hypothesis, suggesting that the model is not correctly specified for. Concluding that including these non-linear terms improves the model.

Furthermore, the improvement in the R^2 value reinforces this conclusion. The increase in R2R2 indicates a better explanation of the variability in the response variable, supporting the idea that including non-linear terms contributes positively to the model's overall specification. Therefore, it is advisable to consider and include non-linear terms for a more accurate representation of the underlying relationships in the data.

Conclusion

In conclusion, our findings highlight the importance of considering both linear and non-linear terms in modeling wine quality. The robust estimation techniques addressed heteroskedasticity issues, enhancing the reliability of our results. Future research may explore additional factors or data sources to further refine our understanding of the complex interplay between physicochemical properties and wine quality in the evolving landscape of the wine industry.

References

https://www.kaggle.com/datasets/yasserh/wine-quality-dataset
https://towardsdatascience.com/red-wine-quality-prediction-using-regression-modeling-and-machine-learning-7a3e2c3e1f46
https://rpubs.com/Sowjanya_G/1137511