DATA2X02 Group Assignment

Group: W13A Early 8

This version was compiled on November 20, 2020

1. Abstract

We aim to test the importance of alcohol content as a significant predictor of red and white wine quality in a dataset containing the physicochemical tests of red and white wine. This was measured by comparing the R^2 in a simple linear regression (SLR), containing only alcohol content and quality, to a multi-linear regression (MLR). The MLR was built through stepwise model selection, using inputs with a correlation of greater than 0.2 with wine quality. The results show that alcohol content is both significant and the most important predictor of red and white wine quality(p-value less than 0.001) as the there was only a 7-11% increase in the MLR R^2 compared to the SLR R^2 (19-23%).

2. Introduction and data exploration

We observed 12 variables, 11 independent continuous variables and 1 dependent discrete variable, which were produced as a result of physicochemical tests performed on red and white wine from Portugal. The data also did not include useful variables such as brand, grape type or price. The data set was acquired from the University of California, Irvine. The data had a far larger number of observations for white wine compared to red. Additionally, most of the quality results were medium quality with few being low or high. This was measured between 1 – 10, with 10 being of higher quality.

On observation, alcohol(vol%) has the strongest correlation of 0.44 and 0.48 for white and red wine respectively. This prompted us to wonder **the extent to which alcohol content is the most important predictor for wine quality?**

3. Analysis

Linear Regression Model. Investigating this by modelling a simple linear regression of alcohol against quality gave an R^2 value of 0.19 for white wine and 0.23 for red wine. This shows that 19% of the variability in white wine and 23% of the variability in red wine can be explained by the change in alcohol. Additionally, both white and red wine show p-values that are low enough to reject the null hypothesis (less than 0.001), suggesting the slope parameter is statistically significant for both wines.

	R^2	Adjusted R^2
White	0.1897	0.1896
Red	0.2267	0.2263

Table 1: *Summarize goodness of fit of SLR models.*

In order to measure the extent that alcohol influences wine quality we build a multiple regression model for each wine to measure the change in R^2 as we add new predictors. Care was taken to ensure all predictors are linearly correlated with the response variable, quality. Violations of linearity can result in a systematic error and may seriously hinder the quality of our models' predictions. Taking the most correlated variables for white gives chlorides, volatile acidity, density and fixed acidity while for red we have citric acid, volatile acidity and sulphates. This can be seen in the shiny app whose link is provided in the appendix.

Model interpretation. The multi-linear regression models were built using forward and reverse stepwise selection for both datasets. In both cases, the red and white wine models were found to be identical. This evidence suggests that these models are robust and are suitable for use in our analysis.

The red wine model incorporates sulphates, alcohol and volatile acidity, with a equation of

$$quality = 2.61 + 0.31(alcohol) - 1.22(volatile\,acidity) + 0.68(sulphates) + \epsilon.$$

All variables included in the model have highly significant p-values less than 0.001. Showing in table 2, the R^2 and R^2 adjusted of 0.336 and 0.335 were observed respectively.

More variables were found useful in constructing the white wine, with a equation of

$$quality = -47.65 + 0.4(alcohol) - 2.09(volatile\_acidity) + 50.91(density) - 0.1(fixed\_acidity) - 1.32(chlorides) + \epsilon.$$

This is potentially a result of the increased sample size. Within the white wine dataset, alcohol, volatile acidity, density, fixed acidity and chlorides were found to be significant, where chlorides had a p-value of 0.014

with all other variables measuring below 0.001. As shown in table 2, the model produced an R^2 and R^2 adjusted values of 0.256 and 0.255 respectively.

The estimate for density was very large compared to the other variables to account for the low mean and variation in the values. Density has a positive estimate in the model despite its correlation with quality being negative. This is likely due to interlinearity between the independent variables, which can lead to an estimate of opposite sign to the correlation coefficient.

	R^2	Adjusted R^2
White	0.2561	0.2554
Red	0.3359	0.3346

Table 2: *Summarize goodness of fit of MLR models.*

Assumption checking. Despite moderately weak linear relationships between predictors and wine quality, there is a lack of patterns in the residual plots, excluding striations which occur from measuring the discrete 'quality' variable, reinforcing an approximately linear relationship. Furthermore, the assumption for independence is satisfied as physicochemical tests are measured separately for each bottle of wine. The residual plot figure 2 shows equal variance among red wine residuals but slight patterns of heteroskedasticity in white wine residuals showing by figure 3. Hence, heteroskedasticity corrected standard errors were used to compute p-values. This resulted in all converted p-values measuring below 0.001,including 'Chlorides'. Despite slight deviations in the lower tails, residuals largely followed the QQ-line. Additionally with the large number of observations we can rely on the central limit theorem to satisfy the normality assumption.

Performance evaluation. A 10 fold cross validation test was conducted to measure out-of sample model performance. The R^2 values indicate 26% and 34% of the observations can be explained from white and red wine models, indicating poor insample performance. Further, the RMSE values show on average the predicted quality of white and red wine differed from actual quality by 0.76 and 0.66 respectively. Given rounding, this results in a constant one level difference in our predicted values, indicating fairly poor model performance. Thus, predicted values should be interpreted with care.

4. Results

A 10 fold cross validation test was conducted to measure out-of sample model performance, giving the results from table 3. The R^2 values indicate 26% and 34% of the observations can be explained from white and red wine models, indicating poor out of sample performance. Further, the RMSE values show on average the predicted quality of white and red wine differed from actual quality by 0.76 and 0.66 respectively. Given rounding, this results in a constant one level difference in our predicted values, indicating fairly poor model performance. Suggesting predicted values should be interpreted with care.

	RMSE	R^2	MAE	RMSESD	R^2SD	MAESD
SLR White	0.80	0.19	0.63	0.02	0.04	0.02
SLR Red	0.71	0.23	0.56	0.05	0.07	0.04
MLR White	0.76	0.26	0.60	0.03	0.03	0.02
MLR Red	0.66	0.34	0.52	0.03	0.06	0.02

Table 3: *Summarize performance of models.*

Alcohol appears to be a significant predictor in both red and white wine models with a p-value less than 0.001. For a 1 unit increase in alcohol, there is an increase of 0.31 in white and 0.40 in red quality. As alcohol content explains 19% (red) and 23% (white) of the simple linear model, with only a 7-11% increase in explanatory power for adding additional variables in the multilinear model, this implies alcohol content is the most influential predictor in determining the quality of red and white wine.

5. Discussion and conclusion

Although models developed are proven to be significant, there have been several limitations. The differing number of observations between red and white wine resulted in inconsistent variances of data, which can impact the quality of inferences. However, as we are focused on comparing the R^2 values instead of estimating predictions, this is less of a concern. If grapes are from the same batch, the assumption of independence could be violated, making the inference of results unreliable. In addition, variables are not necessarily linear as the highest correlation coefficient with wine quality measured 0.48. Hence, as supported by Cortez (2019), MLR is the worst model for predicting wine quality for this dataset. Future studies may focus on wine quality predictions by evaluating alternative machine learning techniques such as K-nearest neighbor and random forest which are able to capture non-linearity.

6. References

- 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.
- Alboukadel Kassambara (2020). ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0. https://CRAN.R-project.org/package=ggpubr
- Hao Zhu (2020). kableExtra: Construct Complex Table with 'kable'and Pipe Syntax. R package version 1.3.1. https://CRAN.R-project.org/package=kableExtra
- Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.
- Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

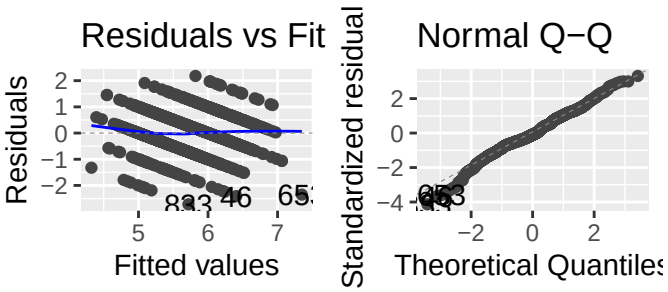

Fig. 2: *Residual plot and normal Q-Q plot of the stepwise regression model for red wine data.*

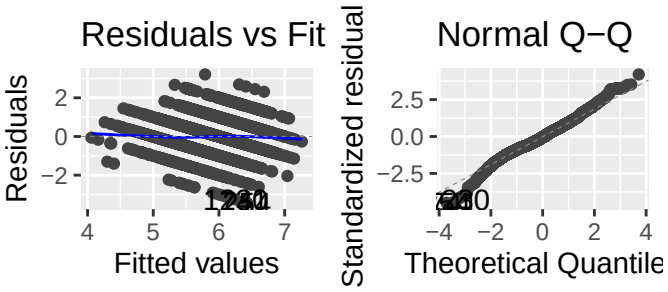

Fig. 3: *Residual plot and normal Q-Q plot of the stepwise regression model for white wine data.*

Shiny app: https://stephan-iie.shinyapps.io/Linearity/?_ga=2.34896929.1856433187.1605152041-1797514058.1605152041

7. Appendix

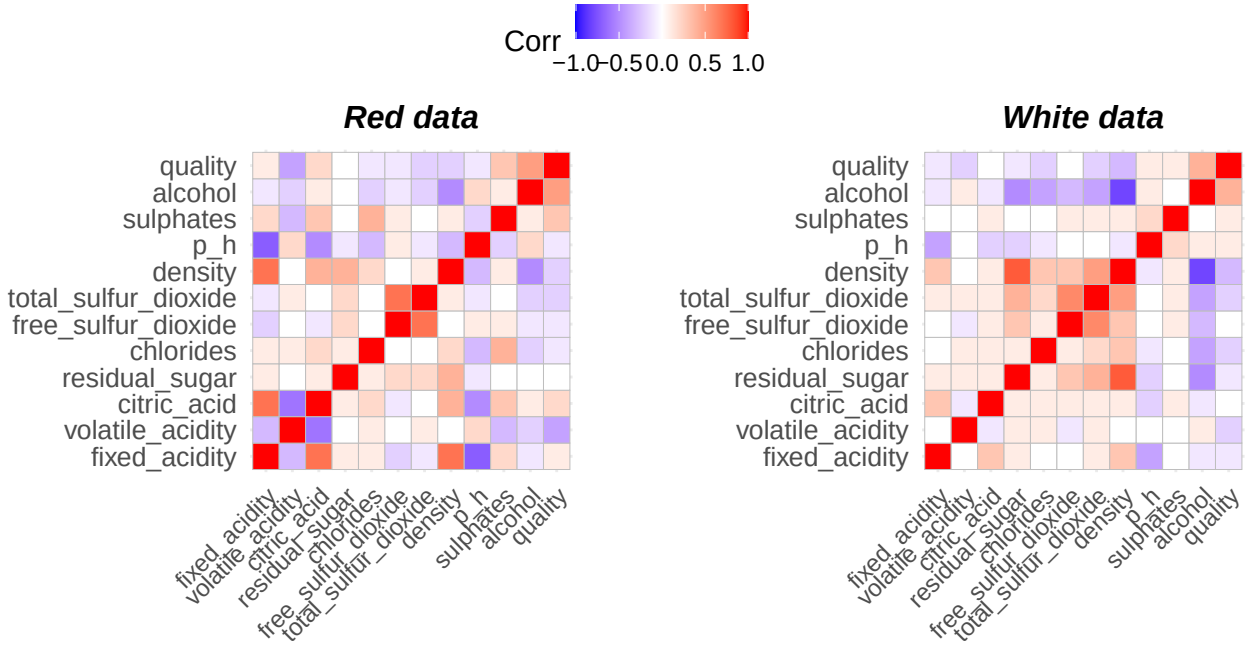

Fig. 1: *Correlation matrices for red and white wine data. The correlation metric is the Pearson's correlatin coefficients.*