Executive Summary - DATA2002 Assignment 2

500520320, 480025267, 500505430, 500586901, 500555424

University of Sydney

This version was compiled on November 12, 2021

The following paper analyses the quality of Vinho Verde red wine from Northern Portugal, measured from a scale of one to ten. Using multiple regression, this paper aims to model the quality of various red wines on other factors such as the alcohol concentration or total sulfur dioxide concentration. Various potential predictor variables were removed from the model for not meeting key assumption criteria such as co-linearity with red wine quality, normality, or residual homoscedasticity. Transformations were attempted on variables that didn't meet the co-linearity assumption, with the only successful transformation being the log of total sulfur dioxide. Multiple regression was then computed using both a step-back and step-forward AIC-based model. These both netted the same model; a regression of red wine quality against alcohol concentration, volatile acidity, the log of total sulfur dioxide and density. In-sample performance of this model showcased an R^2 value of 0.32, indicating that the model may not be particularly strong, and an RMSE of 0.669. Performing 10-fold cross validation suggested that our selected model wasn't subject to obvious overfitting issues. The model generated was as follows:

$$\mathbf{Quality = -22.22 + 0.32AC - 0.67VA - 0.06log(TSD) + 23.31D}$$

1. Introduction

Backgroud. This dataset highlights different factors which contribute to an overall quality rating of the Portuguese red wine "Vinho Verde". Red wine quality may be of interest to producers of wine who are trying to better understand the factors that contribute to overall quality, as well as consumers who are interested in specific characteristics of red wines from speciality regions such as Portugal.

Dataset Overview. The dataset had 1599 rows corresponding to different wine observations, and was collected in 2009 from "Vinho Verde" wine in the north of Portugal. The input variables presented to us were collected from chemical tests, and consisted of fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output variable is the wine quality, graded on a scale of 0 to 10. This was determined by the median score of at least three evaluations made by wine experts.

The first of the 12 variables is the fixed acidity. The fixed acidity is a measure of the tartaric, malic, citric and succinic, found within grapes, which are measured by a steam distillation of a sample of the wine. The volatile acidity (being the second variable) however measures the acetic acid levels in the wine which impacts on the flavour of the wine negatively if the acetic acid is too high. The third variable is citric acid which is used in wine in order to assist in increasing flavour by in turn increasing acidity. Moving away from acidity now, the data set analyses the residual sugars in "Vinho Verde". The residual sugar in wines is not the artificially added sugars, but the raw sugar from grapes found post fermentation, with higher levels of residual sugar unsurprisingly resulting in sweeter wine. Chlorides are used in this data set as a factor contributing towards an overall rating for the wine due to its impact on the saltiness of the final product. The density of wine is deemed to be a major contributor to the overall quality of wine, with higher density resulting in higher quality. pH is simply used in wine making in order to "determine ripeness in relation to acidity". Lower pH is far more desirable as it reduces susceptibility to bacteria growth, while also increasing the quality of taste by making the final product "tart and crisp". Sulphate levels are the next criteria the researchers used to determine overall wine quality rankings, with the sulphate being used in wine in order to protect the final product from bacteria. Evidently, alcohol concentration is a major contributing factor towards the final quality of "Vinho Verde", with it having a large impact on flavour and desirability for the intended market. Finally, the final quality ranking summarises all of the previous factors into a final value.

2. Analysis

Assumption Validation. Multiple regression relies upon four key assumptions: 1. Linearity. 2. Independence. 3. Normality. 4. Homoscedasticity.

For each of these assumptions, the individual relationships between the input variables and quality of wine must meet the assumptions themselves, as well as the final model, and thus were checked individually for each input variable against the quality of wine. For independence, collection methodologies were not explicitly outlined in the dataset, however we can assume that each wine was independent of each other for the purposes of this report.

Linearity refers to how well the input variables can be linearly mapped to the output variable. Violations included free sulphur dioxide, fixed acidity and pH, the latter of which is shown as an example here. [] is also shown as an example of meeting the linearity assumption. Additionally, for any variable which violated the linearity assumptions, data transformations were attempted to improve linearity. The only successful case of this was the log transformation of Total Sulphur Dioxide, as it allowed for the multiple orders of magnitude of total sulphur dioxide to be better expressed linearly against wine quality. (FIgure xxx)

Normality refers to the fact that residuals should be normally distributed. Violations included residual sugar which can be seen below, as compared to a normal example of [].

Finally, homoscedasticity refers to the even distribution of the residuals around the mean. Any fanning of the residuals suggests heteroscedasticity, violating this assumption. Violations include sulphates and chlorides as they displayed distinct residual distribution patterns. Variables which violated assumptions were excluded from the dataset used to construct the model. As such, we were left with alcohol, volatile acidity, the log of total sulphur dioxide, density and citric acid as the remaining input variables which met all assumptions.

Model Selection. Both step-forward and step-back AIC variable selection methods for building a multiple regression model of the quality of red wine came to the same conclusion as can be shown in table 1 and table 2 below. Of the predictor variables that were used to build the model, i.e. those that met the fundamental multiple regression assumptions outlined previously, only the predictors in the two tables were significant. Therefore, the combination of predictors we are using to predict the quality of red wine are the alcohol concentration, volatile acidity, total sulphur dioxide, and the density.

Hence, our model is as follows:

$$\mathbf{Quality = -22.22 + 0.32AC - 0.67VA - 0.06log(TSD) + 23.31D}$$

This model had an AIC of 3259.415, an R^2 of 0.32 and an RMSE of 0.669.

Table 1. Step-Forward Method

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-21.22	10.31	-2.06	0.04
Alcohol Concentration	0.32	0.02	16.99	0.00
Volatile Acidity	-0.67	0.05	-13.64	0.00
Total Sulfur Dioxide	-0.06	0.02	-2.29	0.02
Density	23.31	10.25	2.27	0.02

Table 2. Step-Back Method

	Estimate	Std. Error	t value	Pr(>|t|)
Intercept	-21.22	10.31	-2.06	0.04
Alcohol Concentration	0.32	0.02	16.99	0.00
Volatile Acidity	-0.67	0.05	-13.64	0.00
Total Sulfur Dioxide	-0.06	0.02	-2.29	0.02
Density	23.31	10.25	2.27	0.02

3. Results & Discussion

In Sample Performance. As stated, the selected model had an R^2 of 0.32 and an RMSE of 0.669. This means that approximately 32% of the total variance in red wine quality can be explained by our regression model. This isn't particularly high, indicating this linear model may not be performing strongly. This could be due to the somewhat poor co-linearity observed with most predictor variables versus wine quality in this dataset. However, to ensure that our model isn't subject to overfitting, we want to test it using out of sample performance.

Out of Sample Performance. Using the caret package, we performed a 10-fold cross validation. The purpose of this was to test how well our model with predictor variables of alcohol concentration, volatile acidity, the log of total sulphur dioxide and density performed out of sample. This method allows for error to only be judged based on performance on the testing subset of the original sample, such that our judgement on the performance of the model isn't impacted by the data that was used to produce the model in the first place.

Using 10-fold cross validation, we divided our sample randomly into 10 folds, of which one was allocated to be the testing set, while the rest were training folds used to build the model using the significant predictor variables. This process was repeated 10 times, where a different fold would be the testing set each time. The R^2, root mean squared error and mean average error of each prediction, compared to the real value in the testing set was then computed.

The following graph shows the distribution of R^2, MAE and RMSE for each of the 10 total cross validations performed. We have compared the performance of three separate models. The 'full' model represents all of the potential predictors that were initially present in the red wine quality data frame – even those that didn't meet assumptions such as linearity. The 'selected' model represents the predictor variables we chose that were outlined previously. Finally, the simple model was the regression of red wine quality using only alcohol concentration which was the most significant predictor.

The full model had a median R^2 of 0.35, which was higher than the median R^2 of both the selected and simple models. However, as the full model had predictors that violated assumptions, it is an inappropriate multiple regression model so we must take this result with a grain of salt. The higher R^2 is likely to do with the fact that the larger number of predictors used would result in overfitting of the data, and artificially increasing the R^2 even though the relationship between the wine quality and some of the predictors may not have been linear in the first place. In comparison, the median R^2 of our model was 0.32, which was higher than that of the simple model which was 0.23. Interestingly, as the R^2 of our selected model on the training data was very similar to the R^2 during in-sample performance, it appears that our selected model didn't suffer from overfitting.

Again, discounting the full model, the selected model had the lowest median MAE 0.527 and RMSE (0.671) of the models being compared, indicating that it performed better on the testing set of red wine quality compared to the simple model. (figure name)

Assumption Re-Validation. Additionally, our multiple regression model was also checked against the key assumptions outlined previously – normality and homoscedasticity. The following figures show that there is no obvious fanning pattern in the residuals, indicating homoscedasticity, and that the residuals of the full model are also normally distributed.

Limitations. This model has a few limitations. The linearity of most predictor variables against red wine quality, even after attempting transformations, wasn't very clear and so we had to be generous when determining what met this assumption for the purpose of analysis. An R^2 of 0.32 for the selected model is fairly low, which indicates that the in-sample performance of the model isn't particularly strong, even though it appears the model didn't run into overfitting issues when performing 10-fold cross-validation

There is also not much information available on the data collection process, so we are unsure about whether the different observations are independent or not. For example, red wines from neighbouring regions may not be independent and the quality of wine could be influenced as such. This is an assumption we have had to make for the purpose of the analysis.

4. Conclusion

Overall, we were able to model the quality of red wine by linear regression, using alcohol concentration, volatile acidity, log of total sulfur dioxide, and density. We determined from this that there was a positive relationship between wine quality and alcohol quantity as well as density, whereas a negative relationship was determined between wine quality and volatile acidity as well as total sulphur dioxide.

This could be of interest to both consumers of wine who are interested in characteristics of wine that may determine quality, as well as producers who are trying to make wines of certain qualities at certain price ranges.

5. References

Acknowledgments. We gratefully acknowledge all the help from the rticles package (Allaire *et al*, 2017) which not only introduced us to the powerful PNAS LaTeX style class, but also provided useful code templates to study in the other mode as the fine macros. The pandoc document converter (MacFarlane J , 2017) is the much-admired driving force behind the document manipulation. Additionally, we acknowledge the fantastic support and teaching provided by G. Tarr and the DATA2002 team throughout the semester.

References

Allaire J, R Foundation, Wickham H, Journal of Statistical Software, Xie Y, Vaidyanathan R, Association for Computing Machinery, Boettiger C, Elsevier, Broman K, Mueller K, Quast B, Pruim R, Marwick B, Wickham C, Keyes O, Yu M (2017). *rticles: Article Formats for R Markdown.* R package version 0.4.1, URL https://CRAN.R-project.org/package=rticles.

MacFarlane J (2017). *Pandoc: A Universal Document Converter.* Version 1.19.2.1, URL http://pandoc.org.

Xie Y (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R.* R package version 1.17, URL https://yihui.name/knitr/.

Wickham et al., (2019). *Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686.* URL https://doi.org/10.21105/joss.01686.

McLean MW (2017). *RefManageR: Import and Manage BibTeX and BibLaTeX References in R. The Journal of Open Source Software.* URL https://doi.org/10.21105/joss.00338.

McLean MW (2014). *Straightforward Bibliography Management in R Using the RefManager Package. arXiv: 1403.2036 [cs.DL].* URL https://arxiv.org/abs/1403.2036.

Schloerke B, Cook D, Larmarange J, Briatte F, Marbach M, Thoen E, Elberg T and Crowley J (2021). *GGally: Extension to ggplot2.* R package version 2.1.2. URL https://CRAN.R-project.org/package=GGally.

Fox J and Weisberg S (2019). *R Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage.* URL https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Tang Y, Horijoshi M and Li W (2016) *ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages.* The R Journal 8.2 (2016): 478-489.

Lüdecke D (2021) *sjPlot: Data Visualization for Statistics in Social Science.* R package version 2.8.9 URL https://CRAN.R-project.org/package=sjPlot.

Anderson D, Heiss A, Sumners J (2021). *equatiomatic: Transform Models into LaTeX Equations.* R package version 0.3.0 URL https://CRAN.R-project.org/package=equatiomatic.

Faraway J (2016). *faraway: Functions and Datasets for Books.* R package version 1.0.7 URL https://CRAN.R-project.org/package=faraway.

Broman K (2015). *R/qtlcharts: interactive graphics for quantitative trait locus mapping*. URL doi:10.1534/genetics.114.172742.

Kuhn M (2021). *caret: Classification and Regression Training*. R package version 6.0-90. URL https://CRAN.R-project.org/package=caret.

Zhu H (2021). *ableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.3.4. URL https://CRAN.R-project.org/package=kableExtra..

Firke S (2021). *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 2.1.0. URL https://CRAN.R-project.org/package=janitor.

Wickham H, François R, Henry L and Müller K (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.7. URL https://CRAN.R-project.org/package=dplyr.

Lumley T(2020). *leaps: Regression Subset Selection*. R package version 3.1.0. URL https://CRAN.R-project.org/package=leaps.

Tarr G (2021). *Week three Lecture L08: Testing for independence – who was more likely to die on the Titanic?*. [Powerpoint Slides]. DATA2002, University of Sydney, Sydney, Australia.

Tarr G (2021). *Week eleven Lecture L26: Simple linear regression*. [Powerpoint Slides]. DATA2002, University of Sydney, Sydney, Australia.

Tarr G (2021). *Week eleven Lecture L27: Multiple regression and model selection*. [Powerpoint Slides]. DATA2002, University of Sydney, Sydney, Australia.

Tarr G (2021). *Week eleven Lecture L28: Prediction internals and performance assessment*. [Powerpoint Slides]. DATA2002, University of Sydney, Sydney, Australia.

Tarr G (2021). *Week ten Live Lecture*. [Video file]. DATA2002, University of Sydney, Sydney, Australia.

Tarr G (2021). *Week eleven Live Lecture*. [Video file]. DATA2002, University of Sydney, Sydney, Australia.