Modelling Wine Quality with Multiple Linear Regression

Hoang[a], **480380144**[a], **480011455**[a], **and Zain**[1]

[a]The University of Sydney, Camperdown NSW 2006

This version was compiled on November 4, 2019

This report is an attempt to predict wine quality based on its various physiochemical properties. To do this many linear regression models – using physiochemical properties as predictors and the quality as the outcome – were constructed and compared. Comparison involved analysing various metrics: of greatest interest were the root mean square error, the mean absolute error, and the R-squared statistic. Exhaustive search and backward stepwise selection methods were used. In the latter case, the most accurate model – in other words, the one with the smallest calculated metrics – was a regression model with seven predictor variables. These variables were the residual sugars, the density, the alcohol level, the amount of dissolved sulphates, the amount of dissolved sulphur dioxide, the wine's volatile acidity, and finally the pH. This model exhibited the best relative performance, however, in absolute terms, performance left much to be desired.

Regression | Wine quality | Model selection

0.1. nocite: ''.

1. Introduction

Wine quality is a phrase that encompasses a range of sensory features – mouthfeel, taste, aroma, aging potential, visual appeal, and more. The complexity of wine as an alcoholic beverage has brought forth a new paradigm, and this can be summarised with the following question: to what extent can one understand wine quality through its physicochemical components? Literature on this subject is diverse, ranging from studies on phenolic compounds and polysaccharides, to the effects of proteins and ethanol. Each characteristic, it is argued, contributes to wine quality in different ways and with differing intensities. For example, volatile compounds in wine, present at extremely low concentrations, interact in complicated ways to produce certain aromas; fruity smells are a result of acetate esters and ethyl esters, animalic and otherwise unpleasant smells are attributed to the presence of phenols above their threshold level, etc.

In short, the subject is complicated. We contribute to this discussion, however tentatively, by training a regression model to predict wine quality based on certain physiochemical parameters.

2. Data Set

The wine dataset is the making of Paulo Cortez, Associate Professor at the University of Minho, Portugal. It consists of 4898 samples with 12 variables [see figure 1], 11 of which are physiochemical measurements – examples include *fixed acidity*, *citric acid*, and *residual sugar*. We could not find explicit documentation on the units for these measurements. The other variable, *quality*, is a subjective rating based on a sensory test conducted by experts; it is measured on a 0-10 scale, 0 the worst, 10 the best.

3. Analysis

Table 1. Fig. 1. Statistics of Variables

	Min	Max	Mean
fixed acidity	3.80	14.20	6.85
volatile acidity	0.08	1.10	0.28
citric-acid	0.00	1.66	0.33
residual sugar	0.60	65.80	6.39
chlorides	0.01	0.35	0.05
free sulfur-dioxide	2.00	289.00	35.31
total sulfur-dioxide	9.00	440.00	138.36
density	0.99	1.04	0.99
pH	2.72	3.82	3.19
sulphates	0.22	1.08	0.49
alcohol	8.00	14.20	10.51
quality	3.00	9.00	5.88

3.1. *The Model.*

The linear regression model can be succinctly summarised in vector notation as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Where $\boldsymbol{Y}$ and $\boldsymbol{\epsilon}$ are (n x 1), $\boldsymbol{X}$ is (n x p) and $\boldsymbol{\beta}$ is (p x 1) (n being the number of observations and p being the number of predictors).

We will adopt a multiple linear regression evaluation to predict the outcome variable where $\boldsymbol{Y} = (Y_1, Y_2, Y_3..., Y_n)'$ is the response variable *quality*, $\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1, ..., \beta_p)'$ is the vector of least squares of regression coefficients, $\boldsymbol{X} = (\boldsymbol{x_1}, \boldsymbol{x_2}, ..., \boldsymbol{x_n})'$ is the matrix of predictor variable vectors where $\boldsymbol{x}_i = (1, x_1, x_2, ..., x_p)$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$ is the vector of error terms (residuals), $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$.

The model must be able to predict the result of the dependent variable $\boldsymbol{Y}$ based on p independent predictors, $\boldsymbol{X}$. In this project, we are trying to predict what the quality score of white wines, $\boldsymbol{Y}$, based on physicochemical properties, $\boldsymbol{X}$.

3.2. *Assumptions for Regression Model.*

A few assumptions require careful consideration. These are:

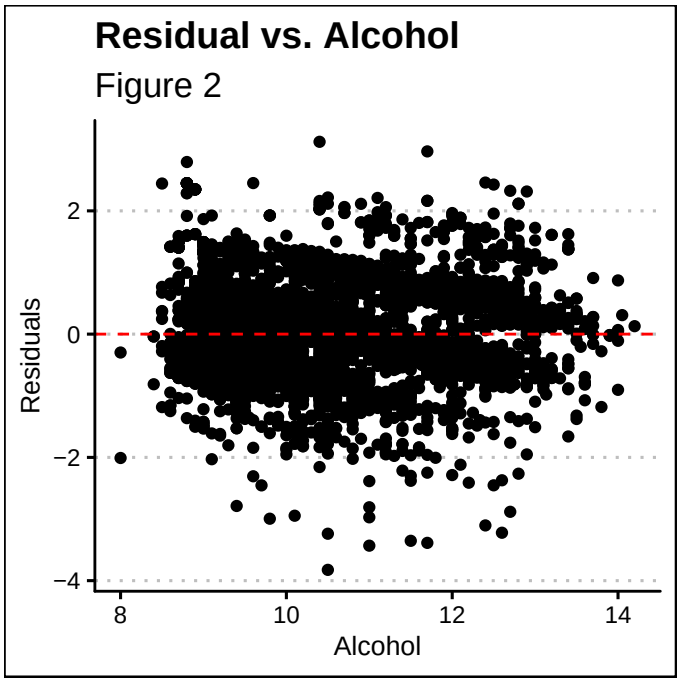
(1) Linearity: our model must be linear, or approximately so, in all the predictor variables.
(2) No Multicolinearity: no two predictor variables are to be highly correlated.
(3) Independence of the residual terms.
(4) Homoscedacisity: the residuals must have constant variance.
(5) Normality: the residuals must be normally distributed.

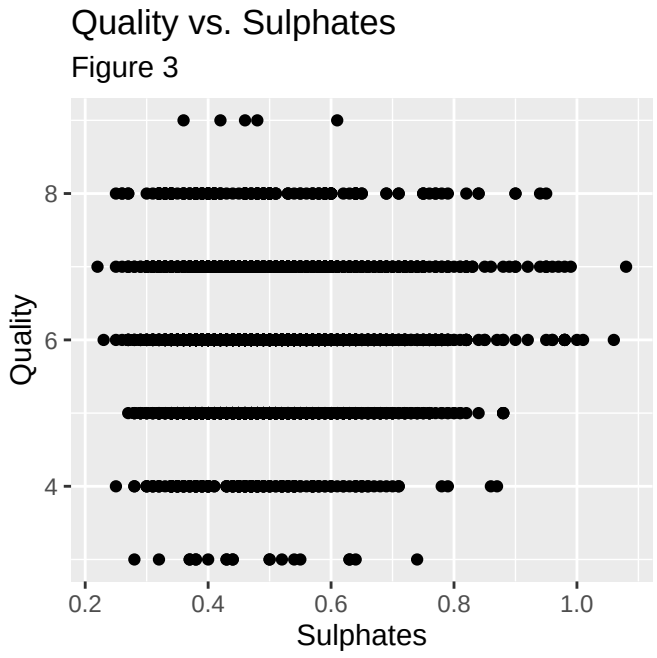
***3.2.1.* Linearity.** To check this we plot each predictor variable against the outcome variable and fit a linear regression line. If the variable appears non-linear, certain transformations (e.g. $log(x)$) may be implemented to enforce linearity. Alternatively, we plot the residuals against each predictor, checking whether they are symmetrically distributed about $y = 0$. If the variable shows little sign of linearity, we delete it.

Figure 2 represents one example of a residual-predictor plot, here done for the alcohol variable. Here the residuals are more or

less scattered evenly above and below the line $y = 0$. All predictor variables exhibited this characteristic – for brevity, we omit them.

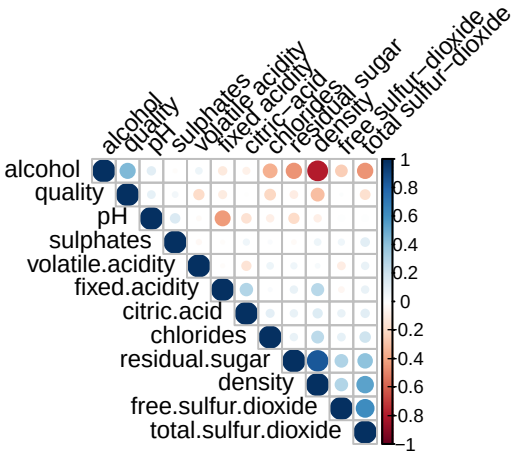
Figure 3 represents the outcome variable, quality, against the sulphates predictor. At first glance this plot appears strange: no real linear relationship is apparent.





3.2.2. **No Multicolinearity.** Based on the correlation plot [Fig 4.], there are some chemical variables that are dependent on each other indicated by high correlation coefficients, namely *density* and *residual sugar*. Even though these two variables are colinear, our aim is to find the model that has the best predictive ability. The "no multicolinearity" assumption is not important regarding the aim of our model.

Figure 4. Correlation Plot

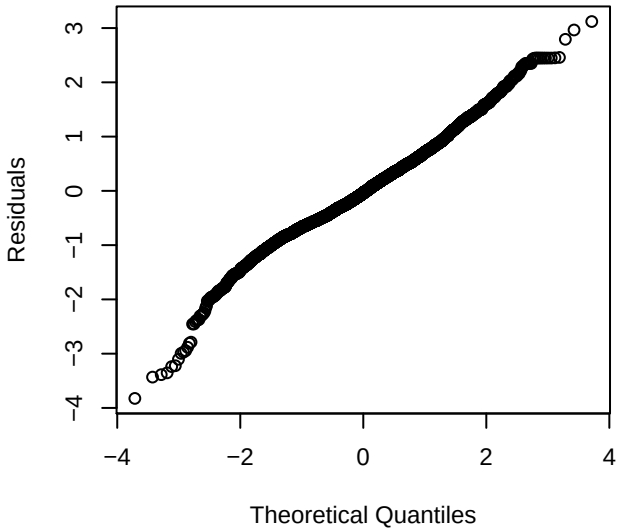


3.2.3. **Independence.** This assumption is mostly a matter of trust. We assume careful data-collection methods were used – if this is true, the residuals must be independent.

3.2.4. **Homoskedasticity and Normality.** To check for homoscedacisity, we analyse the residual plots for each predictor in a similar way to the linearity assumption check. We were interested if the values form a horizontal band that suggest equal variance and thus homoscedacisity. Again, figure 2 is a paradigmatic example of such a plot.

For normality, Q-Q plots were created. Figure 5, below, indicates that the residuals are approximately linear with respect to the normal quantiles.

Figure 5: Normal Q−Q Plot



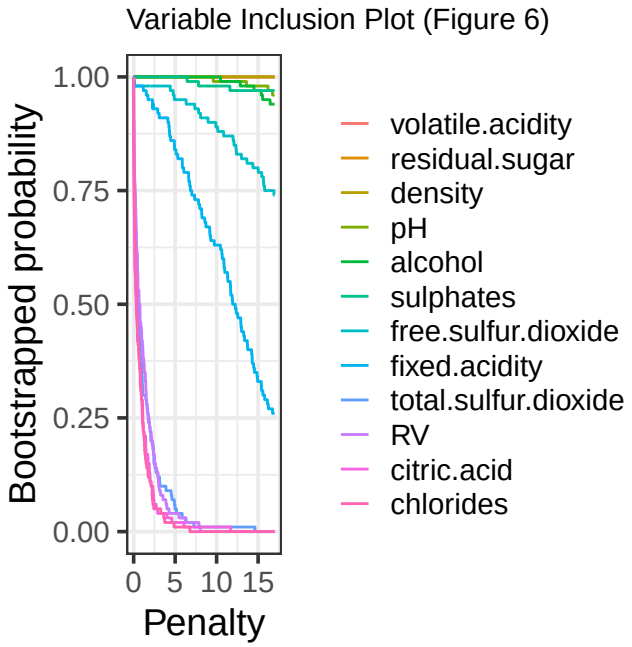
3.3. ***Model Selection.*** Our aim was to find the most accurate multilinear regression model that predicts white wine *quality* with the best in-sample and out-of-sample predictive performance. To do this, we conducted a backward stepwise selection: starting with all eleven physiochemical predictor variables, we trimmed it down to the most essential. For each iteration, we compare the model with

models from previous iterations as well as models suggested by the exhaustive searching method. In detail, our procedure inovlved the following steps:

(1) we have our model M with p number of predictors.
(2) Confirm every necessary assumption is met by the model using assumption checks outlined previously.
(3) Perform a 5-fold cross-validation test to quantify the performance
(4) Compare the results with models from previous iterations and, if it is currently the best model, adopt it as our current best model $\hat{M}$.
(5) Remove the least significant variable and repeat from (1-5).

Overall, the procedure was repeated until the performance became worse than its previous iteration and the final $\hat{M}_7$ became our model.

The model selection process outlined above and found that the p-values of the physiochemical variables *chlorides*, *citric-acid*, *total sulfur-dioxide* and *fixed acidity* were all large, indicating their slopes were not significantly different to the null hypothesis of 0. Thus, the iterative backward selection process omitted these variables. They offered no linear predictive ability in our model. To confirm this, we looked at the plots for each variable and saw that there was no linearity between these variables and *quality*, further validating our omission. Any further omissions were found to make performance worse. This brought our model, $\hat{M}$, down to the most optimal 7 variables. Call this model $\hat{M}_7$.



From the variable inclusion plot (figure 6) we can clearly see our findings from backwards selection visualised. *Chlorides*, *citric acid*, *total sulfur dioxide* and *fixed acidity* all drop off similarly to the random variable *RV*, further validating their omission from the model.

To ensure there are no models with a lesser number of predictor variables that offer better performance better than $\hat{M}_7$, we used exhaustive searching to test all predictor combinations that totalled to less than 7. No model was found to be better than $\hat{M}_7$.

Table 2. Fig.7. Regression Summary Table

	Df	Sum.of.Sq	RSS	AIC	F.value
log(volatile acidity)	1	170.97	2874.84	-2595.81	309.20
residual sugar	1	88.95	2792.81	-2737.59	160.86
log(free sulfur-dioxide)	1	71.79	2775.66	-2767.77	129.84
density	1	50.20	2754.07	-2806.01	90.79
pH	1	14.93	2718.80	-2869.14	27.01
sulphates	1	17.06	2720.93	-2865.31	30.86
log(alcohol)	1	94.45	2798.32	-2727.94	170.82

4. Result

To evaluate the performance of our most optimal model, we used both 10-fold cross validation and a holdout validation. The 10-fold cross validation was conducted in the same way as in the model selection process. The holdout validation involved assigning 80% of the dataset to training and 20% to testing. For training dataset, we used it to fit the model, while the testing group was used for measuring the prediction errors of the model. This latter test was repeated 1000 times, and each time, we resampled the whole dataset, then assigning the data points into training and testing sets. The average RMSE and MAE across these tests were then calculated.

In the end, 7 variables were kept – this represents an optimal balance between simplicity and prediction. The RMSE for this model was 0.74 and the MAE was 0.58. The R-squared value (0.30) is not exactly ideal, but we suspect this is due to the nature of the data: the chemistry of wine is far more complicated, far more nuanced, than anything captured by a regression model.

$$Y_{quality} = 0.54 * sulphates + 0.54 * \log(free\ sulfur\text{-}dioxide) + 0.39 * pH + 0.06 * residual\ s$$

5. Discussion

Overall, our results show that we were able to achieve the aim of finding the multilinear regression model that predicts white wine *quality* with the best in-sample and out-of-sample predictive performance. We found that *volatile acidity* was the best predictor of *quality* (highest F value in figure 7), followed by the remaining six variables in our model. Our final model, $\hat{M}_7$, was arrived at through backward selection from the full model and was shown to be the best performing model by comparison to models found through exhaustive searching.

Despite showing that $\hat{M}_7$ was the best performing linear regression model, it yielded an incredibly low R-squared value of only 0.30. This is similar to the findings of other papers. This is probably due to the incredibly complicated and nuanced nature of the dataset. There are just so many variables become responsible for wine quality and not nearly enough of them were quantified in our report.

Another reason for the poor performance of our model could be due to the subjective nature of the dependent variable. This nature probably resulted in a vastly exaggerated variance that made linear regression very inaccurate.

Ultimately, while we were able to achieve our aim of finding the best linear regression model, it is likely that the nature of the dataset implies that a linear regression model may not be appropriate. Perhaps another model may have performed better and this should be an area for further research.

References

Broman KW (2015). "R/qtlcharts: Interactive graphics for quantitative trait locus mapping." *Genetics*. ISSN 19432631. doi:10.1534/genetics.114.172742.

Dowle M (2016). "Package 'data.table'." *Cran*.

Henry L, Wickham H (2019). "Purrr: Functional programming tools." *R package version 0.3.2*.

Kuhn M (2008). "caret Package." *Journal Of Statistical Software*. ISSN 15487660.

Kutner MH, Nachtsheim CJ, Neter J (2004). *Applied linear models*. ISBN 0072386886. doi:10.1088/1757-899X/149/1/012180.

Lüdecke D (2017). "sjPlot: Data Visualization for Statistics in Social Science, R package version 2.4.0."

Lüdecke D (2018). "sjmisc: Data and Variable Transformation Functions." *Journal of Open Source Software*. ISSN 2475-9066. doi:10.21105/joss.00754.

Lumley T using Fortran code by Miller A (2009). "Leaps: regression subset selection. R package version 2.9." *http:// CRAN.R-project.org/package=leaps*.

R VILLAMOR R (2012). *THE IMPACT OF WINE COMPONENTS ON THE CHEMICAL AND SENSORY PROPERTIES OF WINES*. Ph.D. thesis, Washington State University.

Revelle W (2015). "Package 'psych' - Procedures for Psychological, Psychometric and Personality Research." *R Package*.

Robinson D, Hayes A (2019). "broom: Convert Statistical Analysis Objects into Tidy Tibbles."

Tang Y, Horikoshi M, Li W (2016). "Ggfortify: Unified interface to visualize statistical results of popular r packages." *R Journal*. ISSN 20734859. doi:10.32614/rj-2016-060.

Tarr G, Müller S, Welsh AH (2018). "Mplot: An r package for graphical model stability and variable selection procedures." *Journal of Statistical Software*. ISSN 15487660. doi:10.18637/jss.v083.i09. 1509.07583.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S Fourth edition by*. ISBN 0387954570. doi:10.2307/2685660.

Wei T, Simko V (2016). "The corrplot package." *R Core Team*.

Wickham H (2011). "ggplot2." *Wiley Interdisciplinary Reviews: Computational Statistics*. ISSN 19395108. doi:10.1002/wics.147.

Wickham H (2016a). "Package 'plyr'."

Wickham H (2016b). "readxl: Read Excel files." *R package version 0.1*.

Wickham H (2016c). "tidyverse: Easily Install and Load 'Tidyverse' Packages." *Technical report*.

Wickham H, Francois R, Henry L, Müller K (2019). "Package 'dplyr'. A Grammar of Data Manipulation." *R package version 0.8.0.1*.