EFFECTS ON WINE QUALITY

By: Laila Khalilieh, Teshinee Kukamjad, Humairah Djafar, Daniella Kalaie, Gavin Cardeno

INTRODUCTION

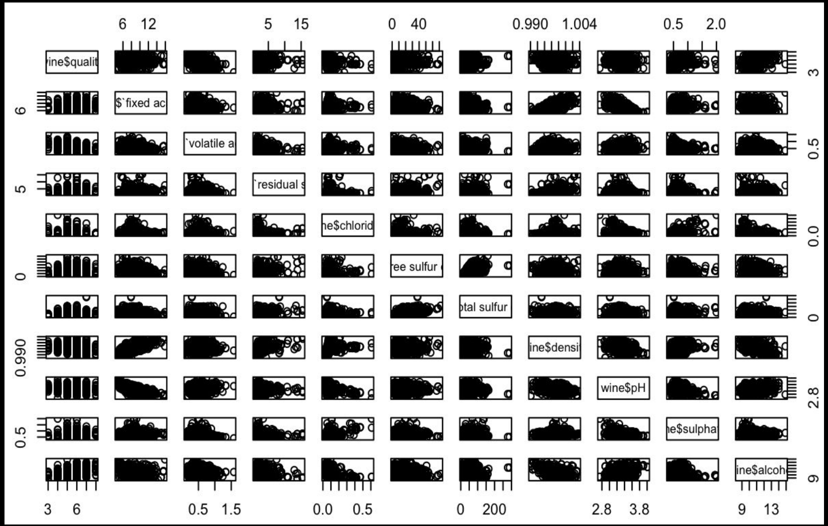
We are interested in determining what factors affect the quality of the red wine variant of the Portuguese “Vinho Verde” based on several different tests.

Approach:

- After modeling the full model, we will transform the data based on concerns regarding the model assumptions validity using variable transformations.
- Next, we will use model subsetting to reduce the instance of overfitting our model considering the large amount of predictions that are in the full model.
- Lastly, we will draw conclusions regarding the relationship between our chosen independent variables and our dependent variable, wine quality.

DATA DESCRIPTION

To assess the linear relationship between wine quality and its predictors, we begin by examining multivariable scatter plots of the variables.



From the scatter plot analysis, it seems that there exists a positive correlation between alcohol content and wine quality. Conversely, we observe negative correlations between quality and variables such as volatile acidity, pH, residual sugar, and total sulfur dioxide.

DATA DESCRIPTION

Summary Statistics:

Employed all ten predictors to construct the full model.

Output:

```
Call:
lm(formula = wine$quality ~ wine$`fixed acidity` + wine$`volatile acidity` +
    wine$`residual sugar` + wine$chlorides + wine$`free sulfur dioxide` +
    wine$`total sulfur dioxide` + wine$density + wine$pH + wine$sulphates +
    wine$alcohol)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6896 -0.3698 -0.0464  0.4563  2.0247

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  2.224e+01  2.120e+01   1.049   0.2943
wine$`fixed acidity`         1.426e-02  2.447e-02   0.583   0.5601
wine$`volatile acidity`     -1.003e+00  1.023e-01  -9.807  < 2e-16 ***
wine$`residual sugar`        1.534e-02  1.498e-02   1.024   0.3062
wine$chlorides              -2.011e+00  4.045e-01  -4.972 7.33e-07 ***
wine$`free sulfur dioxide`   4.799e-03  2.143e-03   2.240   0.0253 *
wine$`total sulfur dioxide` -3.504e-03  7.028e-04  -4.986 6.84e-07 ***
wine$density                -1.811e+01  2.164e+01  -0.837   0.4027
wine$pH                     -4.055e-01  1.915e-01  -2.117   0.0344 *
wine$sulphates               9.126e-01  1.143e-01   7.983 2.72e-15 ***
wine$alcohol                 2.713e-01  2.619e-02  10.358  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6481 on 1588 degrees of freedom
Multiple R-squared:  0.3599,  Adjusted R-squared:  0.3559
F-statistic:  89.3 on 10 and 1588 DF,  p-value: < 2.2e-16
```

Seven variables show statistical significance:

Volatile Acidity, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, pH, Sulphates, and Alcohol.

RESULTS AND INTERPRETATION of Full Model

Full Model Analysis:

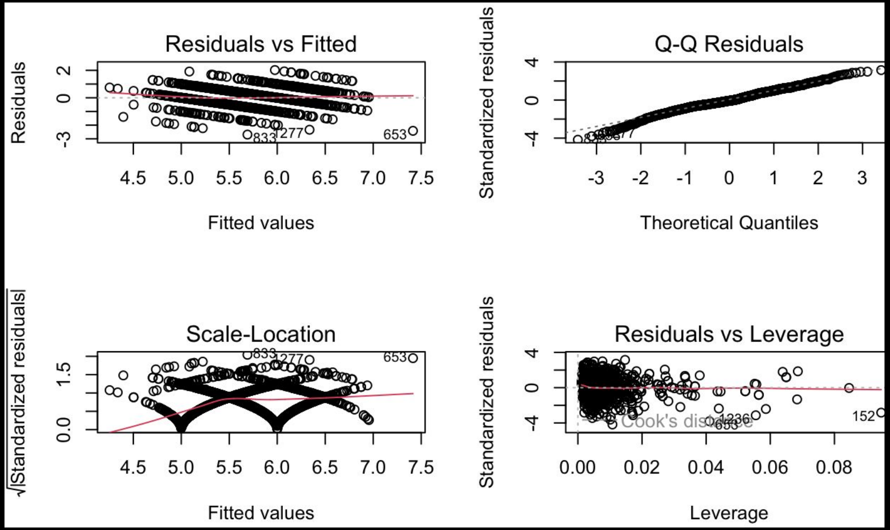
From the diagnostic plots, we evaluate our regression model's adherence to key assumptions: normality, linearity, independence, homoscedasticity.

Residuals vs Fitted: The plot is impacted by the discrete nature of the response variable but the regression line is plotted at the horizontal zero line so this indicates that the residuals equal zero which is a sign of constant variance.

Q-Q Plot: Points follow a 45-degree angle indicating that the normality of errors assumption has held.

Scale-Location Plot: The curved pattern suggests varying residual variance, possibly due to a discrete response variable

Residuals vs Leverage Plot: Notable clusters and high-leverage points are observed, potentially influencing the regression coefficients.



RESULTS AND INTERPRETATION of Full Model (cont)

Full Model Analysis:

To assess multicollinearity, we examine the variance inflation factor (VIF) for each variable::

Output:

```
   wine$`fixed acidity`     wine$`volatile acidity`        wine$`residual sugar`
               6.904684                    1.276333                     1.697730
         wine$chlorides  wine$`free sulfur dioxide` wine$`total sulfur dioxide`
               1.378875                    1.911189                     2.033274
           wine$density                     wine$pH               wine$sulphates
               6.343291                    3.325828                     1.428426
           wine$alcohol
               2.963442
```

From the plot, it’s evident that fixed acidity and density demonstrate high multicollinearity. Thus, we will attempt a regression subset model later in our analysis to decrease this association.

RESULTS AND INTERPRETATION for Transform Model

Transform Model Analysis:

Since we have many leverage points, we transform both the predictor and response variables simultaneously.

Output:

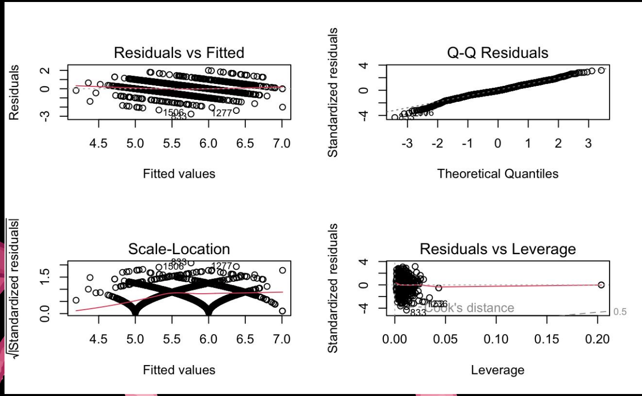
```
bcPower Transformations to Multinormality
    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1      0.9525        1.00       0.7234       1.1816
Y2     -0.2497       -0.33      -0.3906      -0.1088
Y3      0.3533        0.33       0.2466       0.4601
Y4     -1.0599       -1.00      -1.1529      -0.9668
Y5     -0.4627       -0.50      -0.5182      -0.4073
Y6      0.0664        0.07       0.0151       0.1178
Y7     -0.0672       -0.07      -0.1210      -0.0133
Y8    -49.2186      -49.22     -60.5256     -37.9116
Y9      1.0775        1.00       0.4924       1.6627
Y10    -1.1892       -1.19      -1.3511      -1.0274
Y11    -1.4616       -1.46      -1.8731      -1.0502

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
LR test, lambda = (0 0 0 0 0 0 0 0 0 0 0)     1269.917     11 < 2.22e-16

Likelihood ratio test that no transformations are needed
LR test, lambda = (1 1 1 1 1 1 1 1 1 1 1)     7985.31      11 < 2.22e-16
```

The likelihood ratio test indicates the need for transformation, rejecting both no transformation and all log options. We apply the Box Cox multivariable transformation, incorporating the recommended power transformations into a linear model.

RESULTS AND INTERPRETATION of Transform Model

Summary:

- Multiple variables exhibit negative correlation with the response, while others show a positive relationship.
- Significant ANOVA p-value and coefficients, except for one.
- A large F-statistic suggests the model significantly outperforms the null hypothesis.
- After the transformation, the QQ plot follows 45 degrees, indicating the normality of the models errors. Moreover, in the residual vs leverage plot, we mitigated some high-leverage points.
- The scale-location plot shows slight improvements and the residual vs fitted plot once again shows that the sum of the residuals is almost perfectly zero.



Transform Model Analysis:

Summary Output:

```
Call:
lm(formula = wine$quality ~ talc + tchlor + tsulp + tfixacid +
    tressug + tvolacid + tfsulp + ttsulp + tdense + wine$pH)
Residuals:
     Min       1Q   Median       3Q      Max
-2.75966 -0.37368 -0.03611  0.44277  2.00310
Coefficients:
              Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   4.746939   1.266589    3.748   0.000185 ***
talc        -49.611857   6.026909   -8.232   3.81e-16 ***
tchlor        0.032577   0.008819    3.694   0.000228 ***
tsulp        -0.460476   0.043605  -10.560    < 2e-16 ***
tfixacid    -13.894802   4.944817   -2.810   0.005015 **
tressug      -0.389592   0.168991   -2.305   0.021272 *
tvolacid     -2.671300   0.307313   -8.692    < 2e-16 ***
tfsulp        1.055501   0.472556    2.234   0.025648 *
ttsulp        2.673514   0.728602    3.669   0.000251 ***
tdense        1.146044   0.372790    3.074   0.002146 **

wine$pH      -0.100869   0.183918   -0.548 0.583464

---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.6409 on 1588 degrees of freedom

Multiple R-squared:  0.374,        Adjusted R-squared:  0.3701

F-statistic: 94.89 on 10 and 1588 DF,  p-value: < 2.2e-16
```

METHODS AND MOTIVATION

Subsetting the Transformed Model

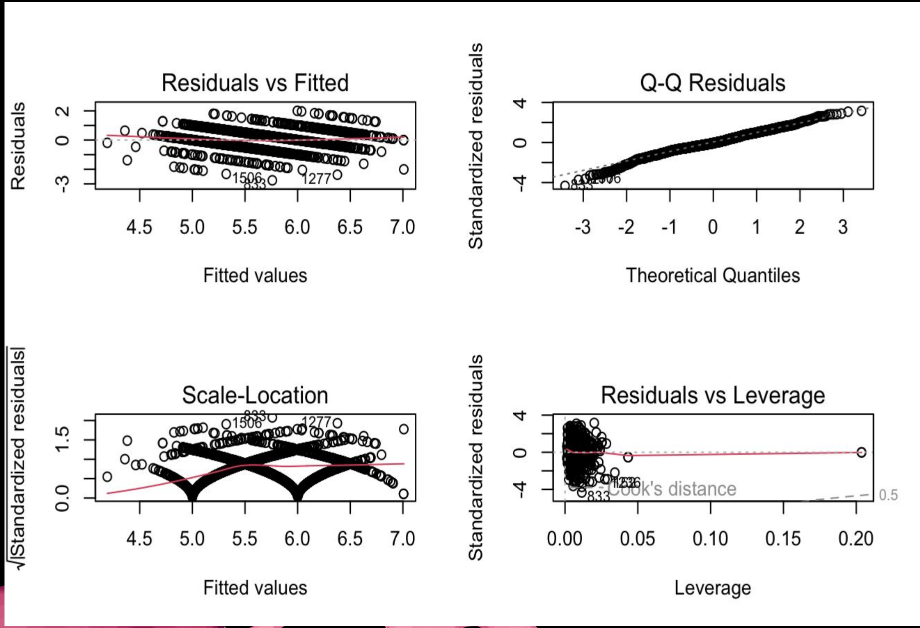
size <int>	Radj2 <dbl>	AIC <dbl>	AICc <dbl>	BIC <dbl>
1	0.2044582	-6527.092	-6527.092	-6516.337
2	0.3016665	-6734.485	-6734.485	-6718.354
3	0.3346264	-6810.796	-6810.796	-6789.288
4	0.3387789	-6819.810	-6819.810	-6792.924
5	0.3438645	-6831.159	-6831.159	-6798.896
6	0.3458817	-6835.086	-6835.086	-6797.446
7	0.3479465	-6839.146	-6839.146	-6796.129
8	0.3483312	-6839.095	-6839.095	-6790.701

According to the analysis of the best subset model method, the 7-variable model has the highest adjusted R square and the lowest AIC and AICc.
Additionally, we used forward stepwise subset analysis which also suggested the use of the 7 variable model.

```
Call:
lm(formula = wine$quality ~ talc + tvolacid + tsulp +
tchlor + wine$pH + ttsulp + tfsulp)
Residuals:
     Min       1Q   Median       3Q      Max
-2.74041 -0.36253 -0.04016  0.44978  1.95303
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.685041   1.114831   5.996 2.49e-09 ***
talc       -63.539105   4.044576 -15.710  < 2e-16 ***
tvolacid    -2.775970   0.304414  -9.119  < 2e-16 ***
tsulp       -0.422602   0.041284 -10.237  < 2e-16 ***
tchlor       0.036766   0.008486   4.333 1.56e-05 ***
wine$pH     -0.477540   0.114738  -4.162 3.32e-05 ***
ttsulp       2.851313   0.713266   3.998 6.69e-05 ***
tfsulp       1.184944   0.470155   2.520   0.0118 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
' 1
Residual standard error: 0.6424 on 1591 degrees of freedom
Multiple R-squared:  0.3699,     Adjusted R-squared:
0.3672
F-statistic: 133.5 on 7 and 1591 DF,  p-value: < 2.2e-16
```

- The summary output for our final transformed 7 variable model indicates that all of our variables have a significant p-value and the overall ANOVA p-value is also significant.
- Four of our 7 variables have a negative association with the response variable. The R^2 of this model is very slightly lower than the non-subsetted model (by about 0.05) which is negligible given the fact that the F-statistic is significantly higher than just the transformed full model.

RESULTS AND INTERPRETATION OF SUBSET MODEL



After subsetting the plots show no significant effect in comparison with the transformed model which indicates that our model assumptions still hold with this new subset model.

VIF values for all predictor variables are less than 5, indicating low correlation among predictor variables and no multicollinearity in our model.

talc	tvolacid	tsulp	
1.327055	1.278241	1.217210	
tchlor	wine$pH	ttsulp	tfsulp
1.230384	1.214964	2.920351	2.789261

CONCLUSION

Summary

In our report, we started out with a full 10 variable model and then assessed the diagnostic criteria for said model. We then determined that our full model would be able to better fit the data if we transformed it as suggested by the box-cox multivariable transformation function and then analyzed this models diagnostics. Finally, to prevent overfitting our data and to eliminate multicollinearity, we proceeded to do regression subset analysis on our transformed model which led us to our final 7-variable transformed model.

Real-World Applicability

Understanding what factors impact the quality of wine most significantly will allow these industries to gain a better understanding of consumer preference and therefore allow them to continue to flourish. Additionally, there is some cause for concern about the real world applicability of our model due to the fact that certain variables like alcohol have a very large negative coefficient in our final model yet it had a near zero coefficient in the full model.

Limitations and Future Directions

The quality of wine is a discrete response variable. To improve the accuracy of our predictions, further research should be conducted using a multiclass classification model, which is better suited for predicting discrete response variables. Additionally, the nature of the transformations that were applied to our variables in our analysis makes us unable to determine a concrete association between our response variable and the predictors or interpret the model coefficients.