April 30, 2021

Alcohol Content in Portuguese Red Wine

Taylor Hartman, Isabel Heard, Corinne Steuk

Abstract:

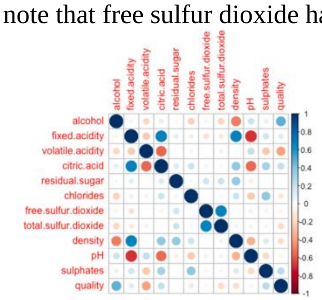
There are several attributes that can be analyzed in wine, such as acidity and concentration of sulphates. Our raw data set included 12 attributes, including the response variable (k=11) and 1,599 observations (n=1,599). The main research question we looked to answer was of the 11 predictors which were significant predictors for determining alcohol content in Portuguese Red Wine. We hypothesized that all 11 of the given predictors would be best suited for predicting the attribute and response variable, alcohol content. To test this hypothesis, we analyzed Anova type III tables, stepwise AIC functions, partial F-tests, residual graphs, and Box-Cox methods. From our findings, the model with predictors fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulphates, and quality is the best model to determine alcohol content in Portuguese red wine.

Introduction:

For our project, we are analyzing the red wine variant of the Portuguese "Vinho Verde" red wine and its attributes. Out of the 12 attributes, we chose alcohol as our response variable because it is a continuous numerical variable. The other 11 attributes act as predictors to best predict because alcohol content is a continuous numerical value. Our research question is to find out which predictors provide the most information about the wine's alcoholic content. Our predictors include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulphates, and quality. We made quality a dummy variable since it is a categorical variable.

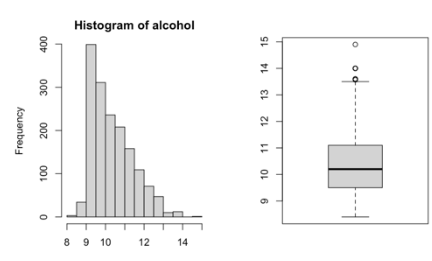
```
[1] 1599   12
    alcohol      fixed.acidity   volatile.acidity  citric.acid     residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates         quality     
 Min.   : 8.40   Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300   Min.   :3.000  
 1st Qu.: 9.50   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.:5.000  
 Median :10.20   Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200   Median :6.000  
 Mean   :10.42   Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :5.636  
 3rd Qu.:11.10   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:6.000  
 Max.   :14.90   Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :8.000  
```

Above is a table of all our attributes and their minimums, 1st quartiles, medians, means, 3rd quartiles, and maximums. We note that free sulfur dioxide and total sulfur dioxide have very high maximums. In turn, free sulfur dioxide and total sulfur dioxide also had very high standard deviations as well. When we look at the summary output for our model f, we note that free sulfur dioxide has the highest p-value. We will use this information to help us create our new model. To the left we have our correlation plot. The predictors that have the best linear relationship out of all the other ones with alcohol are quality and density. Just about all of the other predictors have a light coloration, meaning they do not have much of a linear relationship with alcohol. To the right we have our variance inflation factor output. We use this to determine if there is severe multicollinearity within our data. Since none of these values are greater than 10, we conclude that there is no severe multicollinearity within our model.

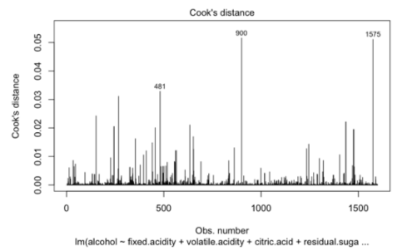



```
                         GVIF Df GVIF^(1/(2*Df))
fixed.acidity        5.749245  1        2.397758
volatile.acidity     1.842350  1        1.357332
citric.acid          2.996653  1        1.731084
residual.sugar       1.326675  1        1.151814
chlorides            1.492344  1        1.221615
total.sulfur.dioxide 1.267179  1        1.125691
density              3.130913  1        1.769439
pH                   2.457958  1        1.567788
sulphates            1.429858  1        1.195767
factor(quality)      1.687309  5        1.053706
```

On the right we have a histogram and box plot of alcohol. As we can see the histogram is skewed to the left, meaning that we do not have normal distribution. Furthermore, with the boxplot we can see that we have a few distinct outliers that we will have to be aware of.

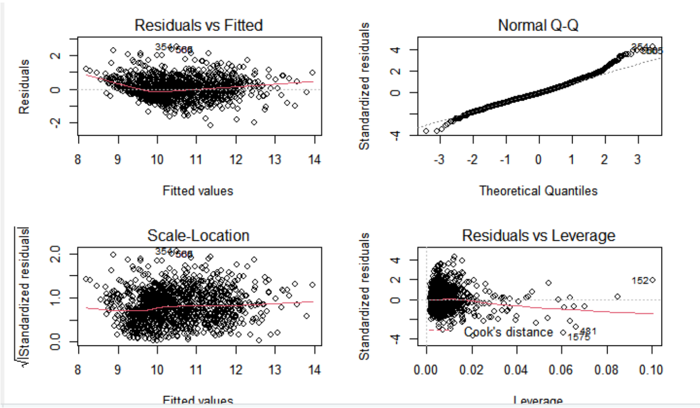


Next, we tested for potential outliers and influential points. First, we checked studentized residuals for outliers, there were a few points that had a value greater than 3 but were not deemed influential. When testing Cook's distance, data point 900 had a value greater than 0.5, but because the value did not exceed 1, it was deemed that the point was not influential.



Methods & Analysis:

For our project to use multiple linear regression, we are naming attribute alcohol content as the response variable because this is a continuous, numerical value. We will also only consider a subset of the other attributes to use as predictor variables.



We can note from the plots that the Normal Q-Q plot has a significant S-shaped curve. Since they do not follow the Y = X line, the normality assumption is not met. Likewise, from the bottom left graph, the red line is slightly curved, especially as the fitted values increase. Although the line does not have a large curvature, we should be cautious of the linearity assumption. From Residuals vs. Fitted, the left side of the graph closely resembles a megaphone shaped, leading to the conclusion that the constant variance assumption is not met either. From the Residuals vs. Leverage, it looks like some points are past or very close to the +/- 3.5 range. We should be cautious of and further investigate all possible outliers and influential points.

```
Anova Table (Type III tests)

Response: alcohol
                      Sum Sq   Df   F value    Pr(>F)
(Intercept)           596.87    1 1690.1752 < 2.2e-16 ***
fixed.acidity         199.09    1  563.7680 < 2.2e-16 ***
volatile.acidity        8.73    1   24.7215 7.344e-07 ***
citric.acid            12.97    1   36.7282 1.694e-09 ***
residual.sugar        159.49    1  451.6378 < 2.2e-16 ***
chlorides               2.03    1    5.7614   0.01650 *
free.sulfur.dioxide     0.86    1    2.4490   0.11780
total.sulfur.dioxide    1.11    1    3.1410   0.07654 .
density               585.89    1 1659.0796 < 2.2e-16 ***
pH                    195.62    1  553.9375 < 2.2e-16 ***
sulphates              28.10    1   79.5714 < 2.2e-16 ***
factor(quality)        39.68    5   22.4726 < 2.2e-16 ***
Residuals             559.02 1583
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the Anova Type III function and testing predictor significance, we found that predictor free sulfur dioxide has a p-value of .1178 that we can use in a hypothesis test where H0: Beta (free.sulfur.dioxide) = 0 Ha: Beta (free.sulfur.dioxide) != 0. The p-value is greater than alpha of .05, so we fail to reject null and conclude that there is not enough evidence that this predictor is significant to Y if all other predictors are included. We have concluded to eliminate this predictor from the model because of the hypothesis testing and the presence of predictor total.sulfur.dioxide that we chose to keep instead.

```
lm(formula = alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + total.sulfur.dioxide + density +
    pH + sulphates + factor(quality), data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.13482 -0.36894 -0.04298  0.33494  2.53012

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.585e+02  1.359e+01  41.100  < 2e-16 ***
fixed.acidity        4.854e-01  2.048e-02  23.698  < 2e-16 ***
volatile.acidity     5.944e-01  1.127e-01   5.272 1.53e-07 ***
citric.acid          8.462e-01  1.322e-01   6.402 2.01e-10 ***
residual.sugar       2.577e-01  1.215e-02  21.210  < 2e-16 ***
chlorides           -9.664e-01  3.860e-01  -2.504 0.012395 *
total.sulfur.dioxide -1.912e-03 5.089e-04  -3.756 0.000179 ***
density             -5.678e+02  1.394e+01 -40.725  < 2e-16 ***
pH                   3.554e+00  1.510e-01  23.532  < 2e-16 ***
sulphates            9.302e-01  1.049e-01   8.866  < 2e-16 ***
factor(quality)4     1.889e-01  2.064e-01   0.915 0.360093
factor(quality)5     2.972e-01  1.930e-01   1.540 0.123822
factor(quality)6     5.627e-01  1.939e-01   2.902 0.003765 **
factor(quality)7     7.778e-01  1.995e-01   3.899 0.000101 ***
factor(quality)8     1.108e+00  2.411e-01   4.596 4.66e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5945 on 1584 degrees of freedom
Multiple R-squared:  0.6915,    Adjusted R-squared:  0.6888
F-statistic: 253.6 on 14 and 1584 DF,  p-value: < 2.2e-16
```

The model f2, introduced to the left, is the original f model with the removal of the free sulfur dioxide predictor. Looking at the coefficients in the model summary, we can conclude that chlorides, total sulfur dioxide, and density negatively affect alcohol content, whereas the rest of the predictors are positively correlated. Density has the largest estimated coefficient meaning this predictor changes alcohol content the most (a decrease of 571.9) when it is increased one by 1 unit. This seems like a surprisingly steep slope, but density varies only slightly when since all examples are liquids of relative thickness. pH is the largest positive coefficient and will change the response variable by 5.59 for every increase in one unit. Both of these increments given by the coefficients table imply that the rest of the predictors are held constant.

We think that model f2 seems like a possible valid model to use for our project. The p-values are all less than alpha now and the r squared is .6915 which is the proportion of the variance of Y that can be explained through the predictors. However, we are going to use the partial F-test to see if a reduced model could be a better option.

```
Analysis of Variance Table

Model 1: alcohol ~ fixed.acidity + volatile.acidity + residual.sugar +
    total.sulfur.dioxide + density + pH + factor(quality)
Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + total.sulfur.dioxide + density + pH + sulphates +
    factor(quality)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1587 607.52
2   1584 559.89  3    47.629 44.916 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We think that a reduced model could simplify the project and still be a significant model to predict Y. To test this hypothesis, we performed a partial F-test with f2 and a reduced model of our choosing. Looking at the anova table image shown above, the p-value is less than alpha of .05 so we can reject null and conclude there is enough evidence that the full model (f2) is significantly better.

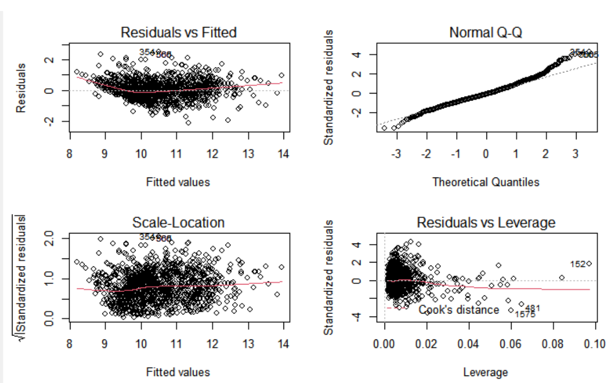
```
Initial Model:
alcohol ~ 1

Final Model:
alcohol ~ factor(quality) + density + fixed.acidity + pH + residual.sugar +
    sulphates + citric.acid + volatile.acidity + total.sulfur.dioxide +
    chlorides


                   Step Df   Deviance Resid. Df Resid. Dev        AIC
1                                          1598  1814.7645   204.3971
2       + factor(quality)  5 483.937982      1593  1330.8266  -281.5416
3               + density  1 294.826383      1592  1036.0002  -679.9838
4         + fixed.acidity  1  95.247430      1591   940.7527  -832.1949
5                    + pH  1 175.047758      1590   765.7050 -1159.4025
6        + residual.sugar  1 157.192137      1589   608.5128 -1524.8190
7             + sulphates  1  29.904908      1588   578.6079 -1603.3972
8           + citric.acid  1   4.961212      1587   573.6467 -1615.1668
9      + volatile.acidity  1   7.350183      1586   566.2965 -1633.7873
10 + total.sulfur.dioxide  1   4.193174      1585   562.1034 -1643.6713
11            + chlorides  1   2.215462      1584   559.8879 -1647.9860
```

We performed a step AIC function (stepwise selection) to compare all alternative models. Note that the final model, which includes all predictors has the lowest AIC value. Using (step3$anova) allows us to consolidate all the AIC values and compare them side-by-side.

When our group considered this data set more, we realized that we should possibly exclude 'quality' from the predictors because this may be an opinion predictor and not have much relation to alcohol content. We are using another partial F test to test this hypothesis. f2 is the model with quality and f3 is excluding this variable. With the small p-value, we conclude that there is evidence that the full model with quality is significant for predicting alcohol content.

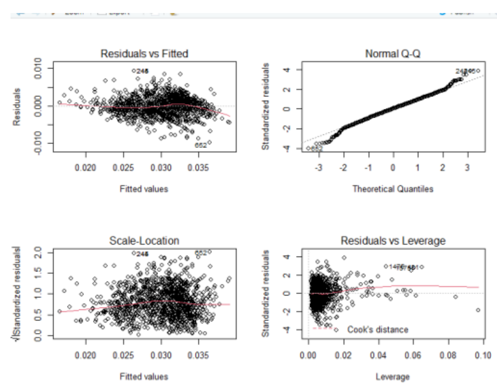
```
Analysis of Variance Table

Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + total.sulfur.dioxide + density + pH + sulphates +
    factor(quality)
Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
    chlorides + total.sulfur.dioxide + density + pH + sulphates
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1584 559.89
2   1589 599.11 -5   -39.225 22.194 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

These residuals graphs show that even with our best model (f2) the variance assumption is not met because it has a megaphone shape. Although the Normal Q-Q plot follows an S-shape (the normality assumption is not met), we can still pass this assumption since the data set is very large.



The assumptions are not met using the various hypothesis tests because the p-value are less than the alpha of .05 (original test shown on the next page). We will check to see if the Box-Cox method will provide any solution. Using the Box-Cox method, we found lambda should equal -1.5 based on the log-likelihood graph.

The residual plots below represent the new transformed model. The normality s-shaped and the megaphone-shape of the top two plots have significantly decreased.



The two images on the next page show a comparison of the numeric diagnostic tests for the normality, constant variance, and independence assumptions using the Breusch-Pagan and Anderson-Darling, Shapiro-Wilk, and Durbin-Watson tests, respectively. The original model f2 is on the left, while the transformed model is on the right. While the p-values are still less than alpha of .05, they have significantly increased from the transformation.

```
        studentized Breusch-Pagan test

data:  f2
BP = 179.96, df = 14, p-value < 2.2e-16

        Anderson-Darling normality test

data:  f2$residuals
A = 5.8559, p-value = 2.057e-14

        Shapiro-Wilk normality test

data:  f2$residuals
W = 0.98276, p-value = 6.306e-13
```

```
        studentized Breusch-Pagan test

data:  f2_bc
BP = 198.13, df = 14, p-value < 2.2e-16

        Anderson-Darling normality test

data:  f2_bc$residuals
A = 1.7547, p-value = 0.0001714

        Shapiro-Wilk normality test

data:  f2_bc$residuals
W = 0.99246, p-value = 2.704e-07
```

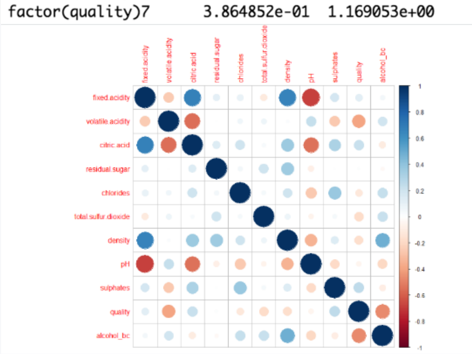
From the results of the Box-Cox method, we predict that the Weighted Least Squares method of transformation may be a better option to relieve the constant variance and normality assumption violations. In future research, we recommend trying this method.

Results:

```
> # Confidence Interval
> confint(f2, level=.95)
                            2.5 %        97.5 %
(Intercept)           5.318561e+02  5.851645e+02
fixed.acidity         4.451947e-01  5.255431e-01
volatile.acidity      3.732494e-01  8.155147e-01
citric.acid           5.869299e-01  1.105397e+00
residual.sugar        2.338712e-01  2.815340e-01
chlorides            -1.723624e+00 -2.092670e-01
total.sulfur.dioxide -2.909732e-03 -9.131852e-04
density              -5.951932e+02 -5.404940e+02
pH                    3.257867e+00  3.850345e+00
sulphates             7.244250e-01  1.136004e+00
factor(quality)4     -2.158904e-01  5.937793e-01
factor(quality)5     -8.139546e-02  6.757697e-01
factor(quality)6      1.823095e-01  9.431101e-01
factor(quality)7      3.864852e-01  1.169053e+00
```

This first confidence interval tests the coefficients of each predictor Beta(i) = 0 with i = [1, 10].

We are 95% confident that the slope for the regression line of alcohol content on fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulphates, factor(quality)6, factor(quality)7, and factor(quality)8 do not include 0, so we can reject the null hypothesis that H_0: Beta(i)=0 and conclude that the coefficients are non-zero.



Using our transformed model, the predictors that have the best linear relationship out of all the other ones with alcohol content is density and quality because they are closest to +/-1. Quality has a negative correlation, while density is positively correlated. Since the alcohol_bc response is the original response variable raised to the -1.5 power, we noted how the correlations are opposite from the original. For example, quality was originally positively correlated with the response alcohol content.

```
> #Prediction for means, Quality = 4
> pred4 = predict(f2, data.frame(fixed.
ides), total.sulfur.dioxide = mean(tota
> pred4
$fit
       fit      lwr      upr
1 10.14531 8.967737 11.32288

$se.fit
[1] 0.08341888

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

```
> #Prediction for means, Quality = 5
> pred5 = predict(f2, data.frame(fixed.acid
ides), total.sulfur.dioxide = mean(total.su
> pred5
$fit
       fit      lwr      upr
1 10.25355 9.086373 11.42072

$se.fit
[1] 0.0249869

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

```
> #Prediction for means, Quality = 6
> pred3 = predict(f2, data.frame(fixed.a
ides), total.sulfur.dioxide = mean(total
> pred3
$fit
       fit      lwr      upr
1 10.51907 9.351974 11.68617

$se.fit
[1] 0.02402432

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

```
> #Prediction for means, Quality = 7
> pred7 = predict(f2, data.frame(fixed.a
ides), total.sulfur.dioxide = mean(total
> pred7
$fit
       fit      lwr      upr
1 10.73413 9.564312 11.90395

$se.fit
[1] 0.04722329

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

Above are the prediction intervals for quality fixed at 4, 5, 6, and 7, while all other variables are held at their mean values. We are 95% confident that, the alcohol content for red wine with a quality of 4 is somewhere between 8.97 and 11.32. We are 95% confident that, the alcohol content for red wine with a quality of 5 is somewhere between 9.07 and 11.42. We are 95% confident that, the alcohol content for red wine with a quality of 6 is somewhere between 9.35 and 11.69. We are 95% confident that, the alcohol content for red wine with a quality of 7 is somewhere between 9.56 and 11.90.

```
> #Confidence Intervals
> #Confidence Intervals for means, Quality = 4
> con4 = predict(f2, data.frame(fixed.acidity = m
des), total.sulfur.dioxide = mean(total.sulfur.di
> con4
$fit
       fit      lwr      upr
1 10.14531 9.981683 10.30893

$se.fit
[1] 0.08341888

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

```
> #Confidence Intervals for means, Quality = 5
> con5 = predict(f2, data.frame(fixed.acidity = me
des), total.sulfur.dioxide = mean(total.sulfur.dic
> con5
$fit
       fit      lwr      upr
1 10.25355 10.20454 10.30256

$se.fit
[1] 0.0249869

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

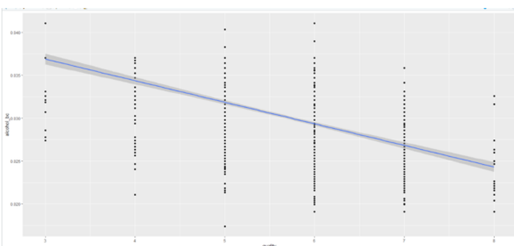
```
> #Confidence Intervals for means, Quality = 6
> con6 = predict(f2, data.frame(fixed.acidity = mear
des), total.sulfur.dioxide = mean(total.sulfur.dioxi
> con6
$fit
       fit      lwr      upr
1 10.51907 10.47195 10.56619

$se.fit
[1] 0.02402432

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

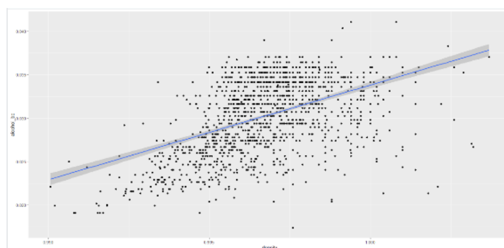
```
> #Confidence Intervals for means, Quality = 7
> con7 = predict(f2, data.frame(fixed.acidity = mean(
des), total.sulfur.dioxide = mean(total.sulfur.dioxic
> con7
$fit
       fit     lwr      upr
1 10.73413 10.6415 10.82676

$se.fit
[1] 0.04722329

$df
[1] 1584

$residual.scale
[1] 0.5945289
```

Above are the confidence intervals for quality fixed at 4, 5, 6, and 7, while all other variables are held at their mean values. We are 95% confident that, for all wine with a quality of 4, the alcohol content on average is somewhere between 9.98 and 10.31. We are 95% confident that, for all wine with a quality of 5, the alcohol content on average is somewhere between 10.20 and 10.30. We are 95% confident that, for all wine with a quality of 6, the alcohol content on average is somewhere between 10.47 and 10.57. We are 95% confident that, for all wine with a quality of 7, the alcohol content on average is somewhere between 10.64 and 10.83. We found it interesting that as quality increases, both the prediction and confidence intervals slightly increase their lower and upper bounds.

All these values VIF<10 indicate that there is no multicollinearity issue in our best fit model.

```
> vif(f2_bc)
                         GVIF Df GVIF^(1/(2*Df))
fixed.acidity        5.749245  1        2.397758
volatile.acidity     1.842350  1        1.357332
citric.acid          2.996653  1        1.731084
residual.sugar       1.326675  1        1.151814
chlorides            1.492344  1        1.221615
total.sulfur.dioxide 1.267179  1        1.125691
density              3.130913  1        1.769439
pH                   2.457958  1        1.567788
sulphates            1.429858  1        1.195767
factor(quality)      1.687309  5        1.053706
```





After determining the predictors that were most correlated to alcohol content (quality and density), ggplots were constructed to create a prediction bond, about the most correlation predictor by holding other predictions constant. The plot on the left is for quality, and it shows the negative correlation because we are comparing quality to the transformed response, alcohol_bc. The plot on the right shows density's positive correlation. As density increases, the predicted value for alcohol content also increases. This is also using the transformed response variable, alcohol_bc, showing the opposite correlation as the matrices above.

Conclusion and Discussion:

Based on all the findings above, the model with predictors fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, total sulfur dioxide, density, pH, sulphates, and quality is the best model to determine alcohol content in Portuguese red wine. Our hypothesis that all predictors should be included in the model were wrong when we determined that free sulfur dioxide did not need to be included in the model. Our other hypothesis, that a smaller subset of the model, would increase the R^2 value and be more significant when predicting also proved to be wrong. While we used the original f2 for the confidence and prediction intervals, we used the transformed model for the plots and visualizations in order to investigate and explore the effects of the transformed model. Since we did not perform many visuals for the transformation models in class, we aimed to incorporate this piece in our project. We advise further researchers to explore other transformation methods to remediate the transformation limitations we faced as a group.

A potential follow-up question from this work would be to determine if this model can be applied to other types of wine, like white wines or other brands of red wine. This could be accomplished by collecting data for the same predictors as this study and interpreting the results through a regression analysis.

References:

Learning, UCI Machine. “Red Wine Quality.” Kaggle, 27 Nov. 2017, www.kaggle.com/uciml/red-wine-quality- cortez-et-al-2009.

Mir, Ferran Rodriguez. “2019, Aug 07.” Red Wine Quality - Data UAB, 19 Aug. 2019, datauab.github.io/red_wine_quality/.

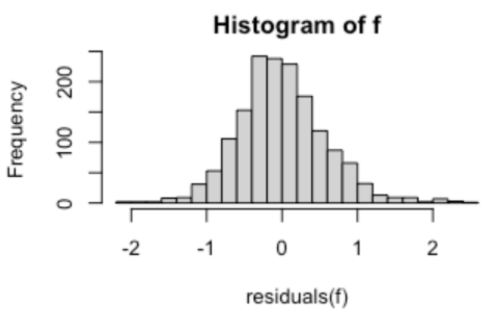
Onukogu, Chinwendu. “Red Wine Quality Prediction Using Classification and Regression Model.” Medium, Dev Genius, 9 July 2020, medium.com/dev-genius/red-wine-quality-prediction-using-classification-and-\ regression-model-f19337821b71#:~:text=The%20main%20aim%20of%20the%20red%20wine%20quality,are%20non-volatile%20acids%20that%20do%20not%20evaporate%20readily.

Individual Contributions:

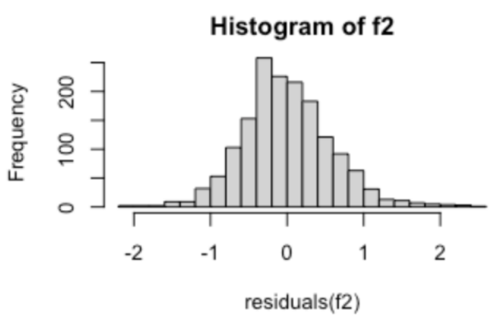
Isabel Heard: “Introduction” code, analysis, and presentation slides.
Corinne Steuk: “Methods & Analysis” code, analysis, presentation slides.; edited and ran “Results” code
Taylor Hartman: wrote “Results” code, analysis, and presentation slides

Appendix:

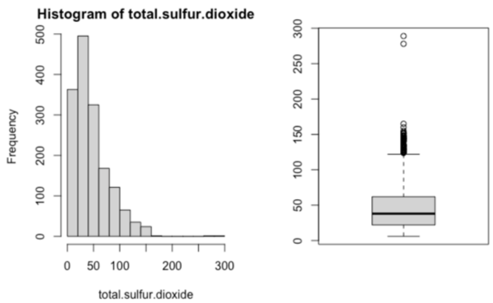
Histogram of f



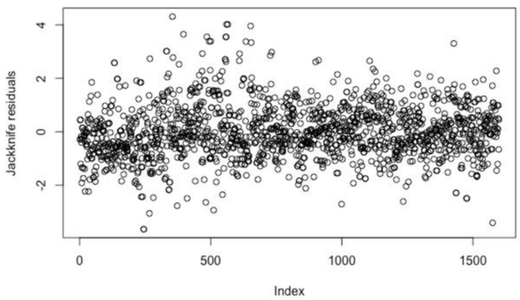
Histogram of f2



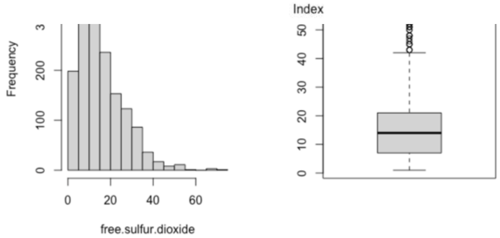
Histogram and boxplot of total sulfur dioxide



Jackknife residual plot



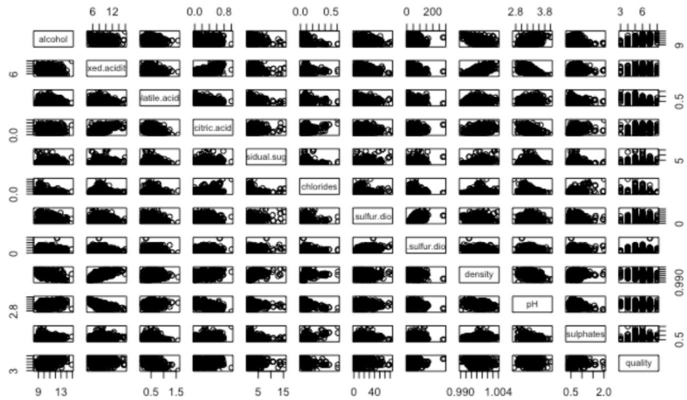
Histogram and boxplot of free sulfur dioxide



Summary table of f

```
Call:
lm(formula = alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
    residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + sulphates + quality, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2.15912 -0.36953 -0.04903  0.34211  2.51405

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           5.632e+02  1.334e+01  42.201  < 2e-16 ***
fixed.acidity         4.925e-01  2.034e-02  24.215  < 2e-16 ***
volatile.acidity      5.893e-01  1.128e-01   5.223 2.00e-07 ***
citric.acid           8.197e-01  1.334e-01   6.143 1.02e-09 ***
residual.sugar        2.624e-01  1.208e-02  21.715  < 2e-16 ***
chlorides            -9.330e-01  3.862e-01  -2.416   0.0158 *
free.sulfur.dioxide  -3.019e-03  1.992e-03  -1.515   0.1299
total.sulfur.dioxide -1.390e-03  6.715e-04  -2.071   0.0386 *
density              -5.736e+02  1.365e+01 -42.027  < 2e-16 ***
pH                    3.617e+00  1.507e-01  24.001  < 2e-16 ***
sulphates             9.540e-01  1.042e-01   9.154  < 2e-16 ***
quality               2.322e-01  2.227e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5942 on 1587 degrees of freedom
Multiple R-squared:  0.6913,	Adjusted R-squared:  0.6891
F-statistic:   323 on 11 and 1587 DF,  p-value: < 2.2e-16
```

Scatter matrix plot



Standard deviations

Standard deviations

```{r}
sd(alcohol)
sd(fixed.acidity)
sd(volatile.acidity)
sd(citric.acid)
sd(residual.sugar)
sd(chlorides)
sd(free.sulfur.dioxide)
sd(total.sulfur.dioxide)
sd(density)
sd(pH)
sd(sulphates)
sd(quality)
```

```
[1] 1.065668
[1] 1.741096
[1] 0.1790597
[1] 0.1948011
[1] 1.409928
[1] 0.0470653
[1] 10.46016
[1] 32.89532
[1] 0.001887334
[1] 0.1543865
[1] 0.169507
[1] 0.8075694
```