Statistical Modeling of Wine Quality
Shiraz vs. Cabernet Sauvignon

1. Introduction

This paper discusses the statistical analysis and methods used to derive the quality factor from Shiraz and Cabernet wine brands. The overall quality factor was measured, along with several dependent variables, some of which were ultimately used in the building of the model used to predict quality:

1. Variety: Categorical variable used to delineate either Shiraz or Cabernet Sauvignon
2. pH: Used to measure ripeness in relation to acidity. Lower pH wines will taste tart and crisp, while higher pH wines are more susceptible to bacterial growth. All values in this dataset range fall between 3.4 and 4.0.
3. Sulfites: A preservative widely used in winemaking for its antioxidant and antibacterial properties.
4. Density: Determined by the concentration of alcohol, sugar, glycerol, and other dissolved solids.
5. Color: Numerical representation of wine color. Data falls between 2 and 8 in these samples.
6. Polymeric pigment color: Anthocyanin red pigment compound colors. Primary contributor to the red blend in both of these wines.
7. Anthocyanin color: Reflect red-blue hues dependent on pH. Includes polymeric and monomeric Anthocyanins.
8. Total Anthocyanins: Measured in Grams per Liter
9. Degree of ionization of Anthocyanins (percent): Degree of ionized Anthocyanins
10. Ionized anthocyanins: Percent of ionized Anthocyanins

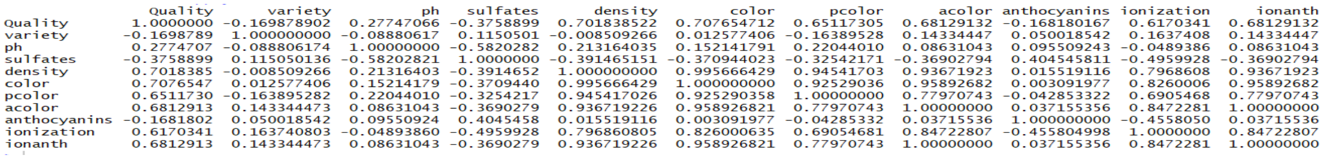
2. Statistical Techniques for Model Development

The following techniques will be employed to develop the linear model.

1. Examine the correlation between different variables to determine if there is any gross multi-collinearity issues
2. Conduct hypothesis testing of the entire model to determine if any of the coefficients are statistically significant.
3. Conduct individual coefficient testing using the extra sum of squares method.
4. Evaluate individual coefficients using the Analysis Of Variance testing (ANOVA)
5. Conduct model adequacy checking
6. Find optimal regressor variables by running advanced statistical analysis such as adjusted R^2 and AIC testing.
7. Split data into training and test sets to verify model accuracy
8. Re-run any steps after modifying model parameters

3. Basic Regressor Variable Analysis

The first step in the model building phase involves a complete analysis of the regressor variables. To accomplish this, a correlation table is generated within R and shown below in table 1.1.



```
               Quality      variety          ph   sulfates      density        color      pcolor      acolor anthocyanins ionization     ionanth
Quality      1.0000000 -0.169878902  0.27747066 -0.3758899  0.701838522  0.707654712  0.65117305  0.68129132 -0.168180167  0.6170341  0.68129132
variety     -0.1698789  1.000000000 -0.08880617  0.1150501 -0.008509266  0.012577406 -0.16389528  0.14334447  0.050018542  0.1637408  0.14334447
ph           0.2774707 -0.088806174  1.00000000 -0.5820282  0.213164035  0.152141791  0.22044010  0.08631043  0.095509243 -0.0489386  0.08631043
sulfates    -0.3758899  0.115050136 -0.58202821  1.0000000 -0.391465151 -0.370944023 -0.32542171 -0.36902794  0.404545811 -0.4959928 -0.36902794
density      0.7018385 -0.008509266  0.21316403 -0.3914652  1.000000000  0.995666429  0.94541703  0.93671923  0.015519116  0.7968608  0.93671923
color        0.7076547  0.012577406  0.15214179 -0.3709440  0.995666429  1.000000000  0.92529036  0.95892682  0.003091977  0.8260006  0.95892682
pcolor       0.6511730 -0.163895282  0.22044010 -0.3254217  0.945417026  0.925290358  1.00000000  0.77970743 -0.042853322  0.6905468  0.77970743
acolor       0.6812913  0.143344473  0.08631043 -0.3690279  0.936719226  0.958926821  0.77970743  1.00000000  0.037155356  0.8472281  1.00000000
anthocyanins -0.1681802  0.050018542  0.09550924  0.4045458  0.015519116  0.003091977 -0.04285332  0.03715536  1.000000000 -0.4558050  0.03715536
ionization   0.6170341  0.163740803 -0.04893860 -0.4959928  0.796860805  0.826000635  0.69054681  0.84722807 -0.455804998  1.0000000  0.84722807
ionanth      0.6812913  0.143344473  0.08631043 -0.3690279  0.936719226  0.958926821  0.77970743  1.00000000  0.037155356  0.8472281  1.00000000
```

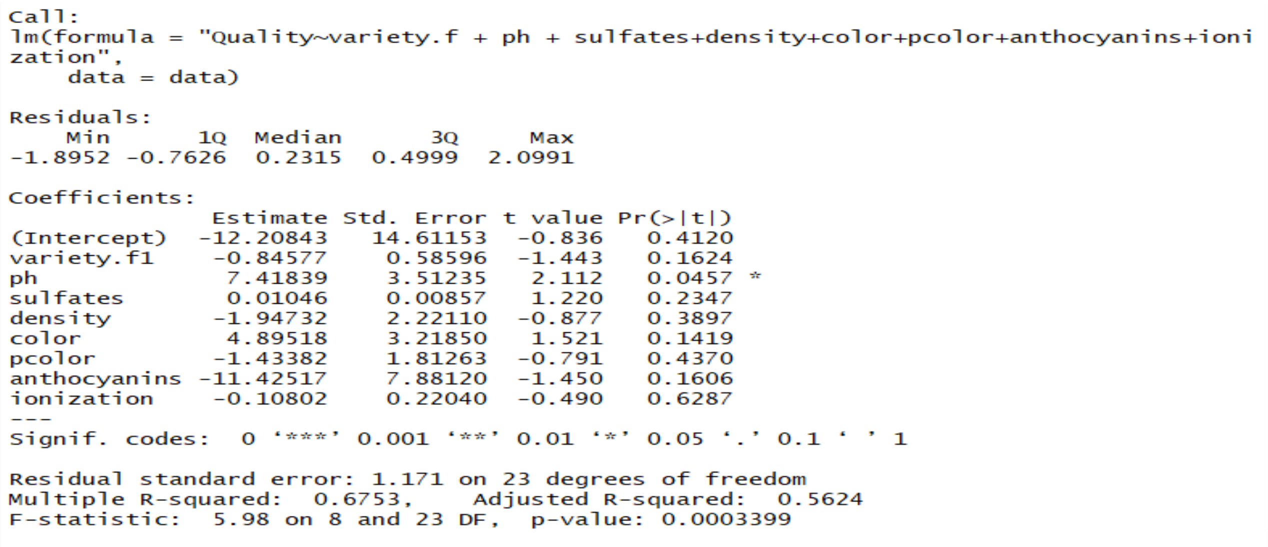
Figure 1: Regressor Variable Analysis

Ionized Anthocyanins are completely collinear with Anthocyanin Color, which leads me to believe that the color is entirely driven from the percentage of ionized molecules in solution. Density is highly correlated with color, polymeric color, anthocyanin color, ionization and ionized anthocyanin. Color is highly correlated with density, polymeric color, anthocyanin color, ionization and ionized anthocyanin. Correlation with the response variable (Quality) is strong among most of the regressors, with the possible exception of variety and total anthocyanins.

Because of the complete correlation between ionized anthocyanins and anthocyanin color, one of these will be removed from the model. I have chosen to remove both variables due to the existence of multicollinearity between not only these two values, but also the values mentioned above. Without the removal of anthocyanin color, R returns a singularity error, and displays acolor as NA. The model after the initial regressor correlation analysis is as follows:

Quality = β0 + β1* Variety + β2*pH + β3*sulfites + β4*density + β5*color + β6*pcolor + β7* anthocyanins +β8*ionization

The resulting linear model summary is shown below:



```
Call:
lm(formula = "Quality~variety.f + ph + sulfates+density+color+pcolor+anthocyanins+ioni
zation",
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8952 -0.7626  0.2315  0.4999  2.0991 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept) -12.20843   14.61153  -0.836   0.4120  
variety.f1   -0.84577    0.58596  -1.443   0.1624  
ph            7.41839    3.51235   2.112   0.0457 *
sulfates      0.01046    0.00857   1.220   0.2347  
density      -1.94732    2.22110  -0.877   0.3897  
color         4.89518    3.21850   1.521   0.1419  
pcolor       -1.43382    1.81263  -0.791   0.4370  
anthocyanins -11.42517   7.88120  -1.450   0.1606  
ionization   -0.10802    0.22040  -0.490   0.6287  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.171 on 23 degrees of freedom
Multiple R-squared:  0.6753,	Adjusted R-squared:  0.5624 
F-statistic:  5.98 on 8 and 23 DF,  p-value: 0.0003399
```

Figure 2: Model Summary with 8 regressors

This model explains about 68% of the variance in quality, and has an F-statistic of approximately 6, which indicates that at least one of the coefficients is statistically significant in the model. The only coefficient that is listed as significant (given that the model controls for all of the above listed values) is ph, with a .0457 level of significance. The lack of significance among the other factors indicates that there are some multicollinearity issues left to address.

Regressor variable analysis continues by calculating some of the SSR values that occur when going from the reduced model to the full model. Figure 2 below shows us the graphical depiction of the reduced model with the average value plotted (mean of the response variable).

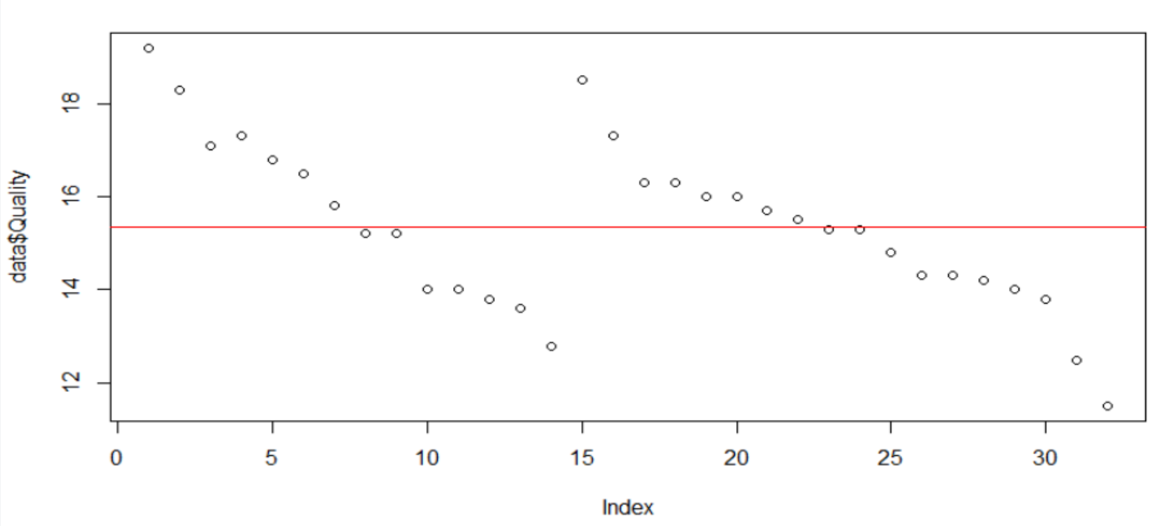
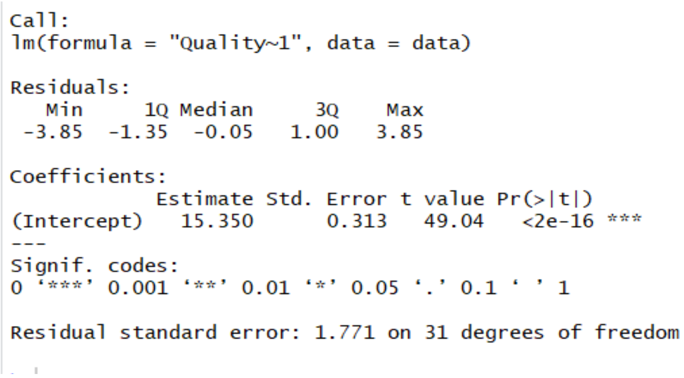


Figure 3: Response Variable with Mean Line

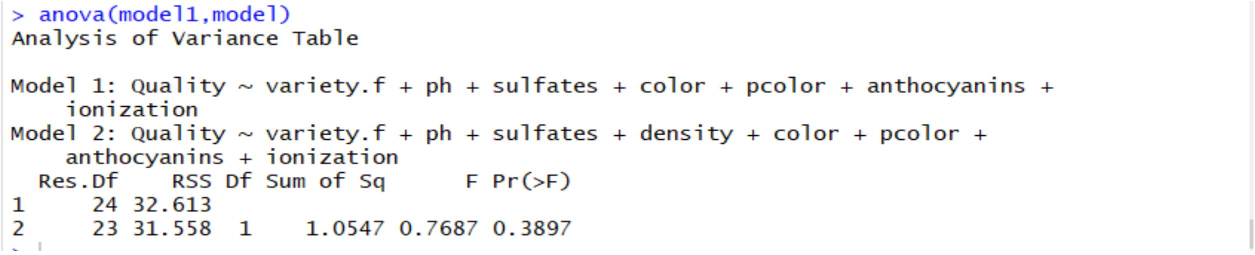
From this plot it is clear that there is some unknown categorical interaction going on. The one categorical variable in the data is wine type (Shiraz vs. Cab), and I hope that this will be the explanatory value for the two groups of data in this graph. I expect that the reduced model will have a large sum of squared error value based on amount of observations that fall above and below the average value line. The reduced model summary from regressing on the mean is shown below in Figure 3.



```
Call:
lm(formula = "Quality~1", data = data)

Residuals:
   Min     1Q Median     3Q    Max
 -3.85  -1.35  -0.05   1.00   3.85

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   15.350      0.313   49.04   <2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.771 on 31 degrees of freedom
```

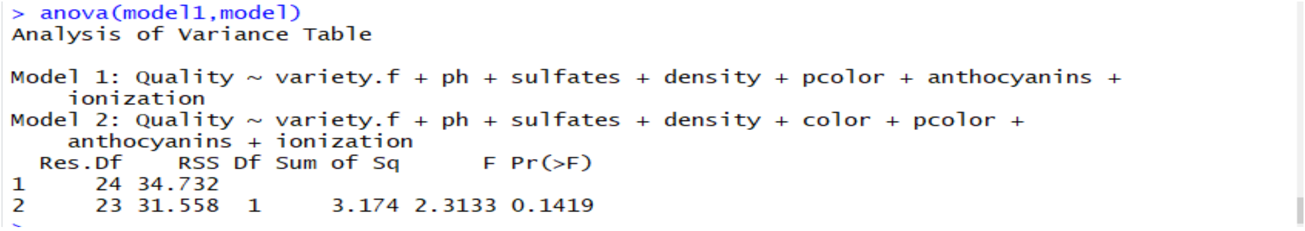
Figure 4: Reduced Model Summary (mean only)

The next test to perform is on the necessity of density as a regressor. From Figure 1 it is shown that it has a 99% correlation ratio with color. To test the β4 coefficient for significance given all other regressors included in the model, an Analysis of Variance will be conducted. This analysis compares the reduced model that does not include density, to the full model that does include it. I will also perform a similar test on color to see which factor is more significant since the high collinearity means only one of these variables is necessary.



```
> anova(model1,model)
Analysis of Variance Table

Model 1: Quality ~ variety.f + ph + sulfates + color + pcolor + anthocyanins +
    ionization
Model 2: Quality ~ variety.f + ph + sulfates + density + color + pcolor +
    anthocyanins + ionization
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     24 32.613
2     23 31.558  1    1.0547 0.7687 0.3897
```

Figure 5: Reduced Model with Color compared to Full Model



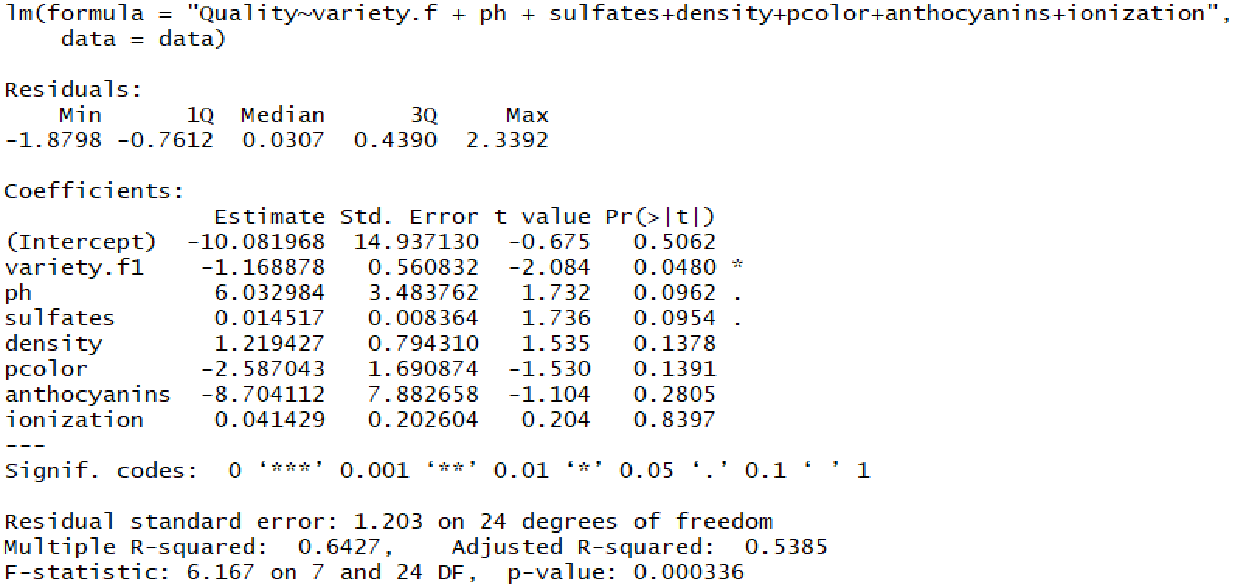
```
> anova(model1,model)
Analysis of Variance Table

Model 1: Quality ~ variety.f + ph + sulfates + density + pcolor + anthocyanins +
    ionization
Model 2: Quality ~ variety.f + ph + sulfates + density + color + pcolor +
    anthocyanins + ionization
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     24 34.732
2     23 31.558  1     3.174 2.3133 0.1419
```

Figure 6: Reduced Model with Density compared to Full Model

From the two ANOVA tests run above, neither density or color are significant when conducting the sequential sum of squares test. Assuming either of these was added to a model with all the regressors, there is no statistically significant increase in explained sum of square regression explanatory power of this model. I will elect to discard color since density does have the higher F value in this case (Figure 5). The new regression equation becomes:

Quality = β0 + β1* Variety + β2*pH + β3*sulfites + β4*density + β5*pcolor + β6* anthocyanins +β7*ionization

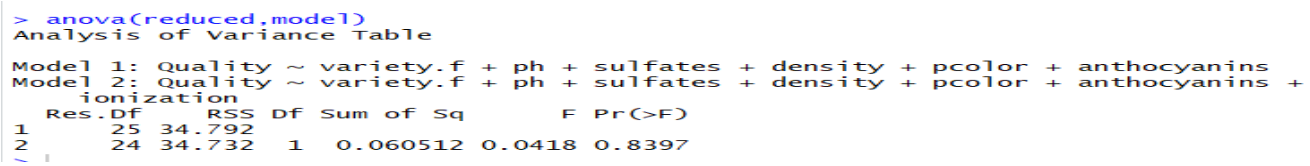
Figure 6 shows the results of the new model. There is a decrease in explained variation (R^2), but the coefficients are starting to become more and more significant to the model. Now, variety is significant (as expected by viewing Figure 2) and 2 others are significant at an alpha of 0.1. Based on the low statistical significance indicated by the ionization variable, the next test will be an extra sum of squares test for this variable.



```
lm(formula = "Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins+ionization",
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8798 -0.7612  0.0307  0.4390  2.3392 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -10.081968  14.937130  -0.675   0.5062  
variety.f1    -1.168878   0.560832  -2.084   0.0480 *
ph             6.032984   3.483762   1.732   0.0962 .
sulfates       0.014517   0.008364   1.736   0.0954 .
density        1.219427   0.794310   1.535   0.1378  
pcolor        -2.587043   1.690874  -1.530   0.1391  
anthocyanins  -8.704112   7.882658  -1.104   0.2805  
ionization     0.041429   0.202604   0.204   0.8397  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.203 on 24 degrees of freedom
Multiple R-squared:  0.6427,	Adjusted R-squared:  0.5385 
F-statistic: 6.167 on 7 and 24 DF,  p-value: 0.000336
```

Figure 7: Summary of Full Model after removing Color

Figure 7 shows the results of the extra sum of squares test assuming the reduced model does not include ionization.



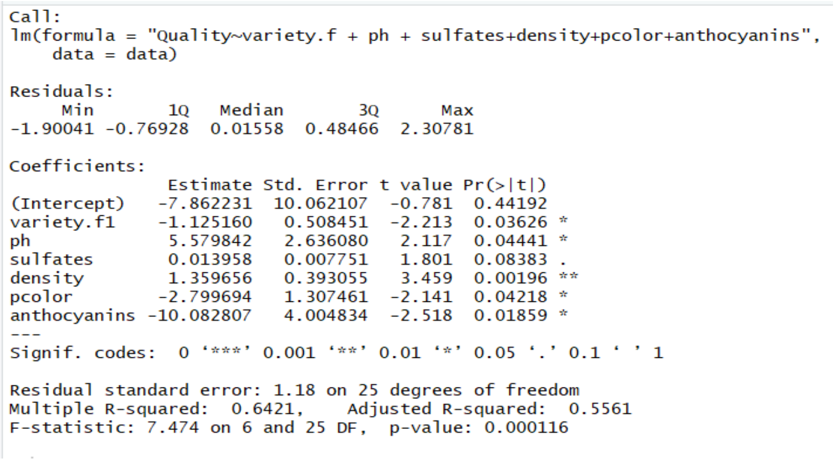
```
> anova(reduced,model)
Analysis of Variance Table

Model 1: Quality ~ variety.f + ph + sulfates + density + pcolor + anthocyanins
Model 2: Quality ~ variety.f + ph + sulfates + density + pcolor + anthocyanins + 
    ionization
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     25 34.792                           
2     24 34.732  1  0.060512 0.0418 0.8397
```

Figure 8: Results of ANOVA without Ionization

These results show a very small reduction in residual sum of squares with the addition of the ionization variable. There is also a very small F value, and a high p-value which indicates that this variable does little to explain the variation in quality. The new model without this variable becomes:

Quality = β0 + β1* Variety + β2*pH + β3*sulfites + β4*density + β5*pcolor +β6* anthocyanins

An examination of Figure 8 reveals that the model coefficients are all significant at the alpha level of 0.1 with 5 of them significant at the 0.05 level. I am now ready to move on to more advanced regressor analysis techniques.



```
Call:
lm(formula = "Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins",
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.90041 -0.76928  0.01558  0.48466  2.30781 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -7.862231  10.062107  -0.781  0.44192   
variety.f1    -1.125160   0.508451  -2.213  0.03626 * 
ph             5.579842   2.636080   2.117  0.04441 * 
sulfates       0.013958   0.007751   1.801  0.08383 . 
density        1.359656   0.393055   3.459  0.00196 **
pcolor        -2.799694   1.307461  -2.141  0.04218 * 
anthocyanins -10.082807   4.004834  -2.518  0.01859 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.18 on 25 degrees of freedom
Multiple R-squared:  0.6421,	Adjusted R-squared:  0.5561 
F-statistic: 7.474 on 6 and 25 DF,  p-value: 0.000116
```

Figure 9: Model Summary without Ionizations

4. Model Adequacy Analysis

Here we examine some of the assumptions behind the OLS modeling by examining residual values.

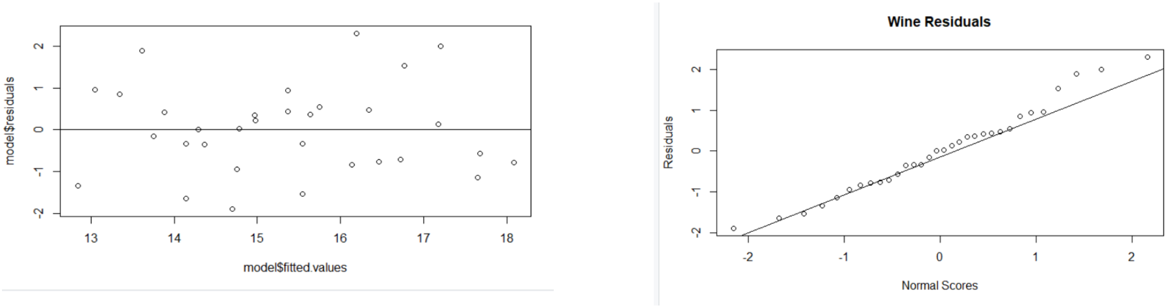


Figure 10: Plot of model residuals

These residuals show a mostly normal distribution pattern on the QQ plot, and a fairly randomized plot of residuals. I am comfortable with the assumption that the errors are normally distributed in this model.

Because the variance of the different residuals can change depending on the different input values (not a violation of homoskedacity in OLS since we are talking variance and not error), the studentized residuals will give us a proper perspective for discovering outliers. The studentized residuals in Figure 10 show that all values are within the plus or minus 2 threshold, and the conclusion from this graph is that all the data is usable for model prediction.

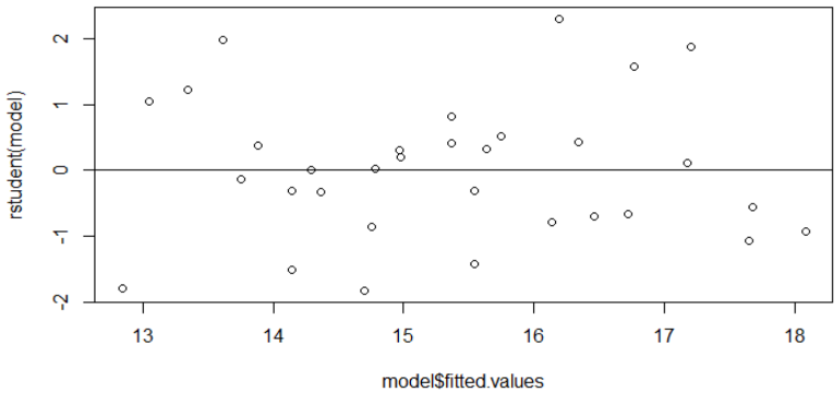


Figure 11: Studentized Model Residuals

Because I do not see any pattern in the residuals, and there are no potential outliers to deal with, there is no need to transform our input or the output. The conclusion from the adequacy checking is that the OLS assumptions of heteroskedacity and normality of error distribution are satisfied.

5. Advanced Regressor Variable Analysis

This section will return to the original list of regressor variables in order to see if certain combinations of them result in statistically significant and meaningful models. The section above removed ionizations because it did not add anything in the extra sum of squares analysis, but this was assuming all other regressors were present. It is possible that this regressor would have been significant given another set of initial regressor variables. I will try to examine all of these relationships with the advanced techniques in order to determine the final optimal combination of variables for the best model.

The model I used for the advanced techniques looks like this:

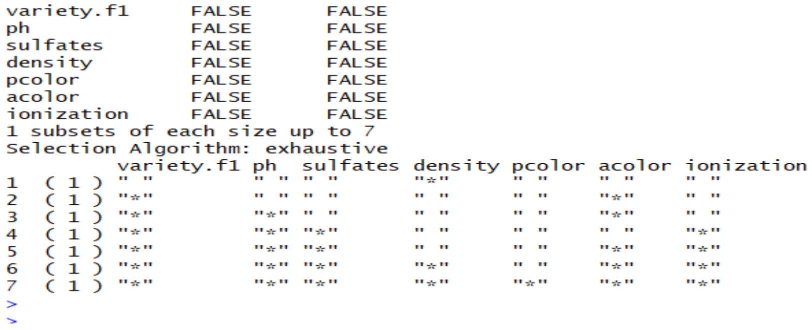
$$\textbf{Quality} = \beta 0 + \beta 1^{*}\ \textbf{Variety} + \beta 2^{*}\textbf{pH} + \beta 3^{*}\textbf{sulfites} + \beta 4^{*}\textbf{density} + \beta 5^{*}\textbf{pcolor} + \beta 6^{*}\ \textbf{acolor} + \beta 7^{*}\textbf{ionizations}$$

There are 3 terms that had to be removed due to singularity errors in R. This error was caused by the high degree of multicollinearity between numerous variables. This model mostly conforms to our earlier model, but ionizations is reintroduced since I determined it was a logical error to remove it in the first place.

	B0	B1	B2	B3	B4	B5	B6	B7
x_1	15.69	-.597						
x_12	.8884	-.5143	3.87					
x_123	11.89	-.44	1.23	-.011				
x_1234	6.607	-.53	1.44	-.001	.5			
x_12345	4.55	-.8	1.88	0.0004	.86	-1.27		
x_123456	-11.2	-.7	6.57	.001	-3.47	6.62	5.96	
x_1234567	-21	-1.1	8.7	.01	-2.21	3.82	3.83	0.16
x_134567	12.5	-.79		-.002	-.15	.64	1.18	.03
x_124567	-9.3	-.72	6.06		-3.01	5.71	5.1	.05
x_12456	-9.38	-.66	6.15		-3.46	6.61	5.89	
x_1456	12.1	-.81			-.03	.433	1.23	
x_1245	4.95	-.80	1.80		.85	-1.2		

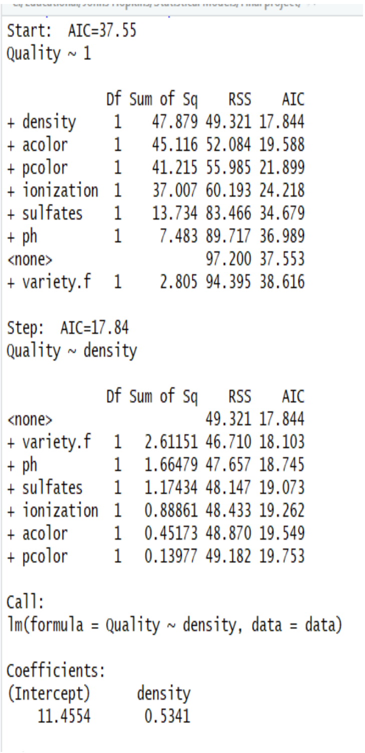
Observations: Beta5 and Beta 6 vary wildly if compared to the instance where they are both in the model with when only one of them is in the model. This makes sense based on the high correlation between the two values. I suspect the final model will only include one of these variables.

Below we run an exhaustive subset search, along with a forward and backward search on the model parameters.

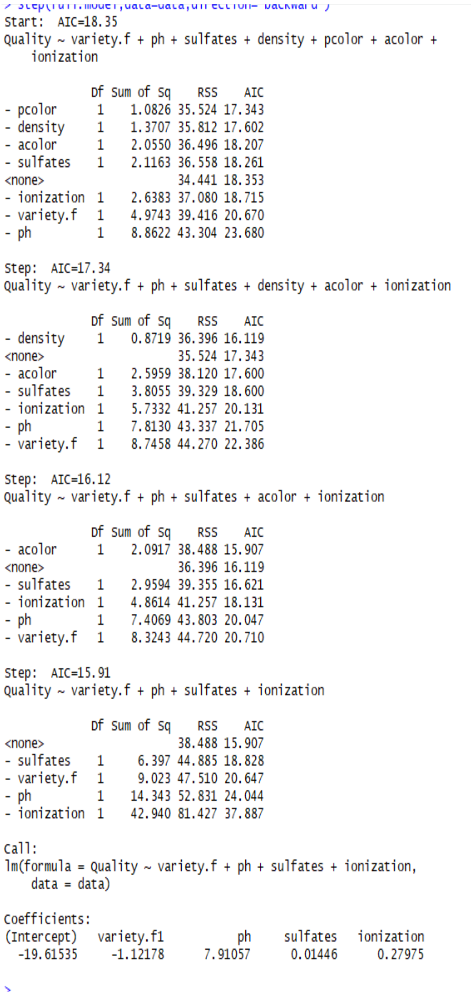


```
variety.f1      FALSE      FALSE
ph              FALSE      FALSE
sulfates        FALSE      FALSE
density         FALSE      FALSE
pcolor          FALSE      FALSE
acolor          FALSE      FALSE
ionization      FALSE      FALSE
1 subsets of each size up to 7
Selection Algorithm: exhaustive
         variety.f1 ph  sulfates density pcolor acolor ionization
1  ( 1 ) " "        " " " "      "*"     " "    " "    " "
2  ( 1 ) "*"        " " " "      " "     " "    "*"    " "
3  ( 1 ) "*"        "*" " "      " "     " "    "*"    " "
4  ( 1 ) "*"        "*" "*"      " "     " "    " "    "*"
5  ( 1 ) "*"        "*" "*"      " "     " "    "*"    "*"
6  ( 1 ) "*"        "*" "*"      "*"     " "    "*"    "*"
7  ( 1 ) "*"        "*" "*"      "*"     "*"    "*"    "*"
>
>
```

Figure 12: Exhaustive Search Subset Selection



```
C:/Educational/Johns Hopkins/Statistical Models/Final project/
Start:  AIC=37.55
Quality ~ 1

             Df Sum of Sq    RSS    AIC
+ density     1    47.879 49.321 17.844
+ acolor      1    45.116 52.084 19.588
+ pcolor      1    41.215 55.985 21.899
+ ionization  1    37.007 60.193 24.218
+ sulfates    1    13.734 83.466 34.679
+ ph          1     7.483 89.717 36.989
<none>                    97.200 37.553
+ variety.f   1     2.805 94.395 38.616

Step:  AIC=17.84
Quality ~ density

             Df Sum of Sq    RSS    AIC
<none>                    49.321 17.844
+ variety.f   1   2.61151 46.710 18.103
+ ph          1   1.66479 47.657 18.745
+ sulfates    1   1.17434 48.147 19.073
+ ionization  1   0.88861 48.433 19.262
+ acolor      1   0.45173 48.870 19.549
+ pcolor      1   0.13977 49.182 19.753

Call:
lm(formula = Quality ~ density, data = data)

Coefficients:
(Intercept)      density  
    11.4554       0.5341  
```

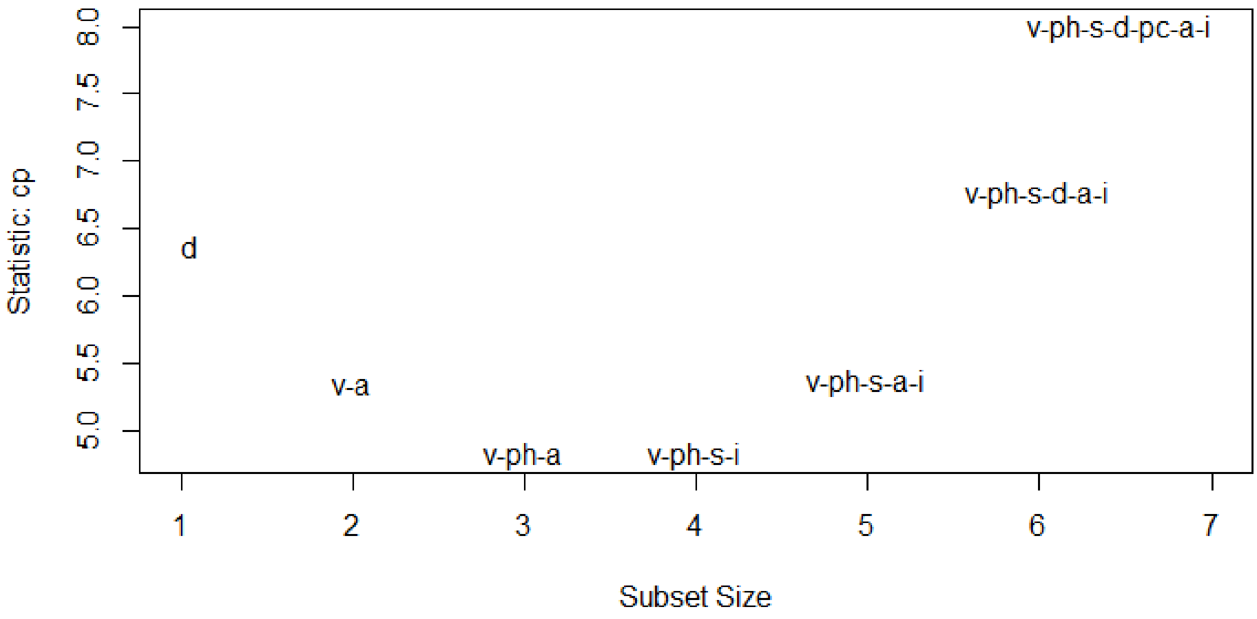


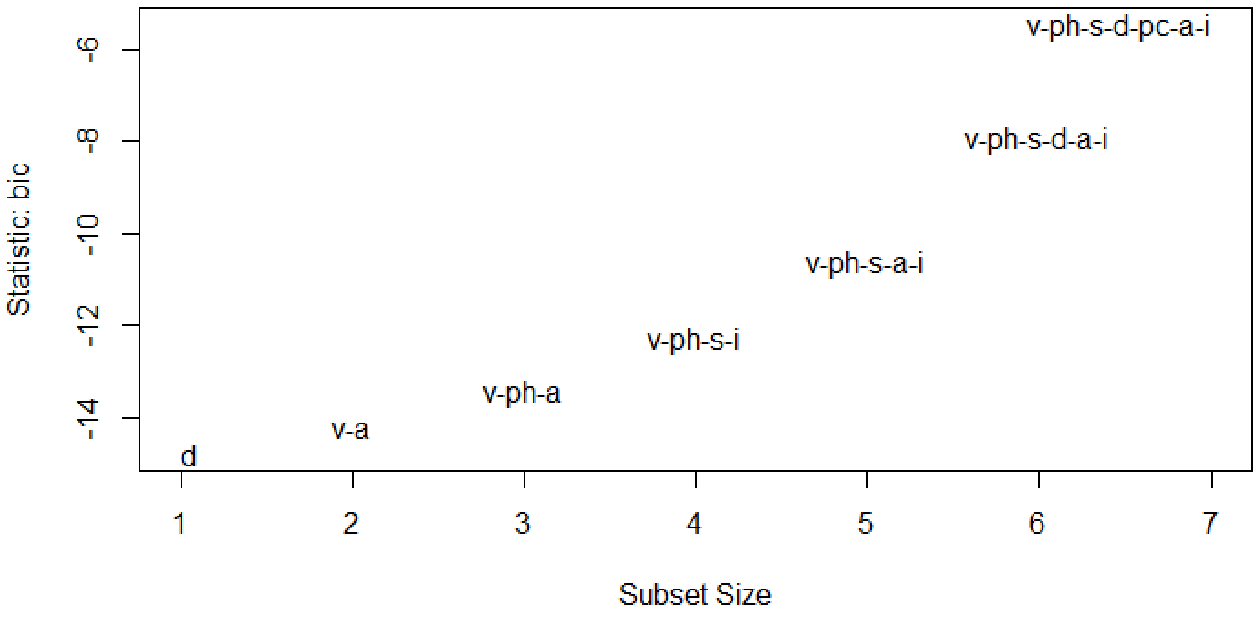
```
> step(full.model,data=data,direction="backward")
Start:  AIC=18.35
Quality ~ variety.f + ph + sulfates + density + pcolor + acolor + 
    ionization

             Df Sum of Sq    RSS    AIC
- pcolor      1    1.0826 35.524 17.343
- density     1    1.3707 35.812 17.602
- acolor      1    2.0550 36.496 18.207
- sulfates    1    2.1163 36.558 18.261
<none>                    34.441 18.353
- ionization  1    2.6383 37.080 18.715
- variety.f   1    4.9743 39.416 20.670
- ph          1    8.8622 43.304 23.680

Step:  AIC=17.34
Quality ~ variety.f + ph + sulfates + density + acolor + ionization

             Df Sum of Sq    RSS    AIC
- density     1    0.8719 36.396 16.119
<none>                    35.524 17.343
- acolor      1    2.5959 38.120 17.600
- sulfates    1    3.8055 39.329 18.600
- ionization  1    5.7332 41.257 20.131
- ph          1    7.8130 43.337 21.705
- variety.f   1    8.7458 44.270 22.386

Step:  AIC=16.12
Quality ~ variety.f + ph + sulfates + acolor + ionization

             Df Sum of Sq    RSS    AIC
- acolor      1    2.0917 38.488 15.907
<none>                    36.396 16.119
- sulfates    1    2.9594 39.355 16.621
- ionization  1    4.8614 41.257 18.131
- ph          1    7.4069 43.803 20.047
- variety.f   1    8.3243 44.720 20.710

Step:  AIC=15.91
Quality ~ variety.f + ph + sulfates + ionization

             Df Sum of Sq    RSS    AIC
<none>                    38.488 15.907
- sulfates    1     6.397 44.885 18.828
- variety.f   1     9.023 47.510 20.647
- ph          1    14.343 52.831 24.044
- ionization  1    42.940 81.427 37.887

Call:
lm(formula = Quality ~ variety.f + ph + sulfates + ionization, 
    data = data)

Coefficients:
(Intercept)   variety.f1           ph     sulfates   ionization  
  -19.61535     -1.12178      7.91057      0.01446      0.27975  
```

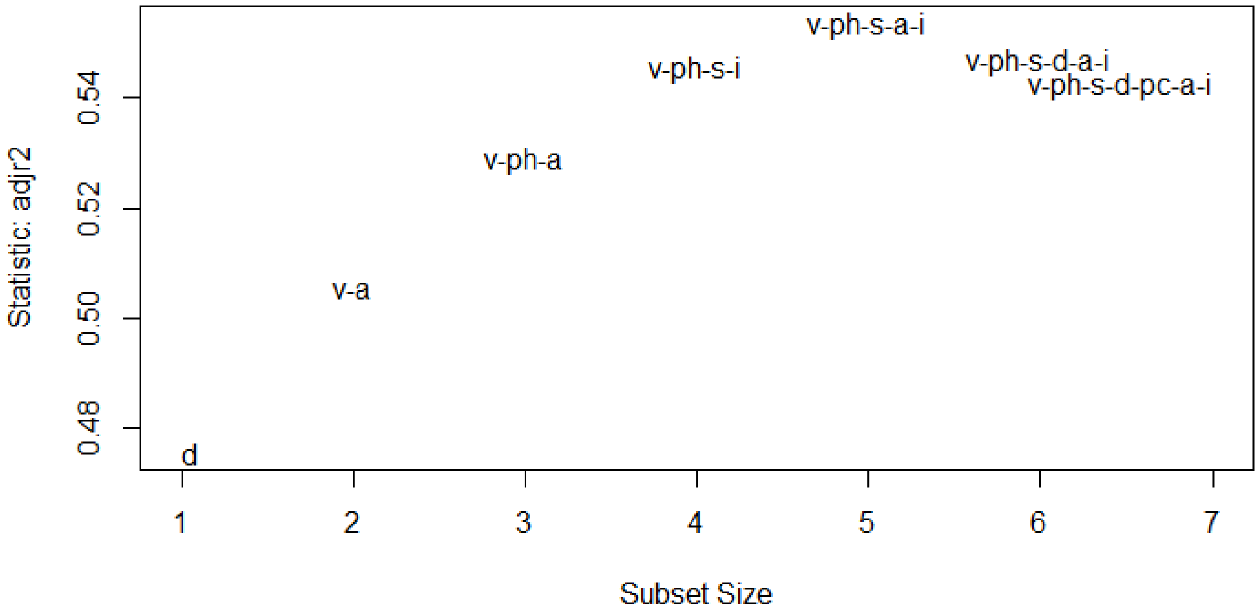
Figure 13: Forward (left) and Backward (right) Search for Best Regressor Subset

The results of these operations are very interesting. The forward selection and stepwise regression techniques both chose the model that contains just a single density term. The density regressor does have the highest correlation with quality, so it is no surprise that this would be included in the model, but the presence of it alone was unexpected. Backwards regression chose a model that has variety, ph, sulfates and ionization.

Let’s examine the different parameters by model subset.


Statistic: cp
5.0
5.5
6.0
6.5
7.0
7.5
8.0
d
v-a
v-ph-a
v-ph-s-i
v-ph-s-a-i
v-ph-s-d-a-i
v-ph-s-d-pc-a-i
1
2
3
4
5
6
7
Subset Size


Statistic: bic
-14
-12
-10
-8
-6
d
v-a
v-ph-a
v-ph-s-i
v-ph-s-a-i
v-ph-s-d-a-i
v-ph-s-d-pc-a-i
1
2
3
4
5
6
7
Subset Size



After examining all this data, I believe the best choice is the model with variety, ph, sulfites and ionizations. This model has the lowest cp score, is the result of the backward selection model and is nearly the highest on adjusted r^2. It falls a little short in the BIC test, but is only about 10% off from the best model, which is a single factor formula of just density. The next best model is the single factor model with just density, and it has a much lower R^2 parameter when compared to the 4 factor model. For even more analysis, let’s examine all the key values for the 7 different best models found in our subset regression analysis.

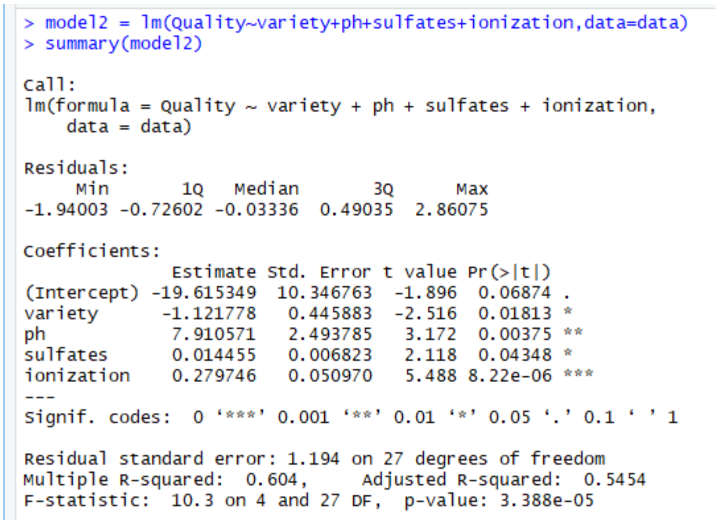
Regress	Regressors	SSres	MSres	Rsq	Rsq-adj	Cp	Press STAT
1	d	49.32	1.64	.49	.47	6.4	55.23
2	v-a	44.9	1.55	.53	.50	5.4	53.8
3	v-ph-a	41.3	1.48	.57	.53	4.8	54.3
4	v-ph-s-i	38.48	1.43	.6	.554	4.8	54.12
5	v-ph-s-i-a	36.4	1.4	.63	.553	5.4	64.15
6	v-ph-s-d-a-i	35.5	1.42	.63	.550	6.6	67
7	v-ph-s-d-pc-a-i	34.44	1.44	.65	.54	7.8	73.1

This table shows that the most well rounded model is model 4, the one we selected above. It explains the most of the variance when adjusted by the number of variables, has the lowest Cp score, was selected as optimal by backward regression, and is very close to the lowest press statistic.

Our final selection is thus:

Quality = β0 + β1* Variety + β2*pH + β3*sulfites + β4*ionizations

This model has the following parameters:



```
> model2 = lm(Quality~variety+ph+sulfates+ionization,data=data)
> summary(model2)

Call:
lm(formula = Quality ~ variety + ph + sulfates + ionization,
    data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.94003 -0.72602 -0.03336  0.49035  2.86075

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.615349  10.346763  -1.896  0.06874 .
variety      -1.121778   0.445883  -2.516  0.01813 *
ph            7.910571   2.493785   3.172  0.00375 **
sulfates      0.014455   0.006823   2.118  0.04348 *
ionization    0.279746   0.050970   5.488 8.22e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.194 on 27 degrees of freedom
Multiple R-squared:  0.604,     Adjusted R-squared:  0.5454
F-statistic:  10.3 on 4 and 27 DF,  p-value: 3.388e-05
```

6. Interaction and transformations

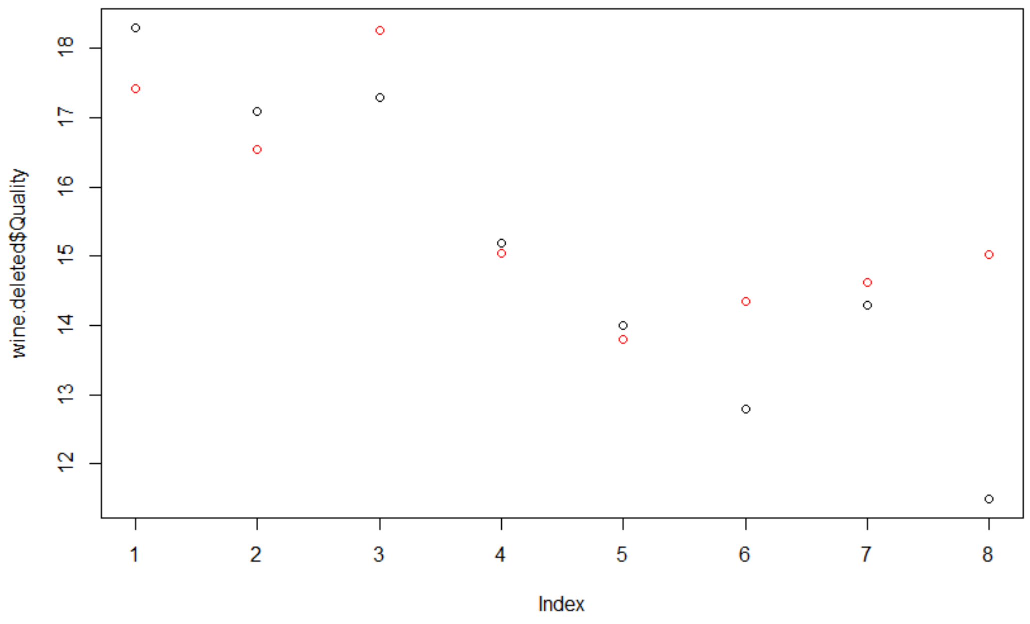
Before finalizing the model, various parameters were tested for an interaction effect. For example, here is a short list of the many parameters that were checked:

variety * ph
variety * ionization
variety * pcolor
density*pcolor
density*sulfates

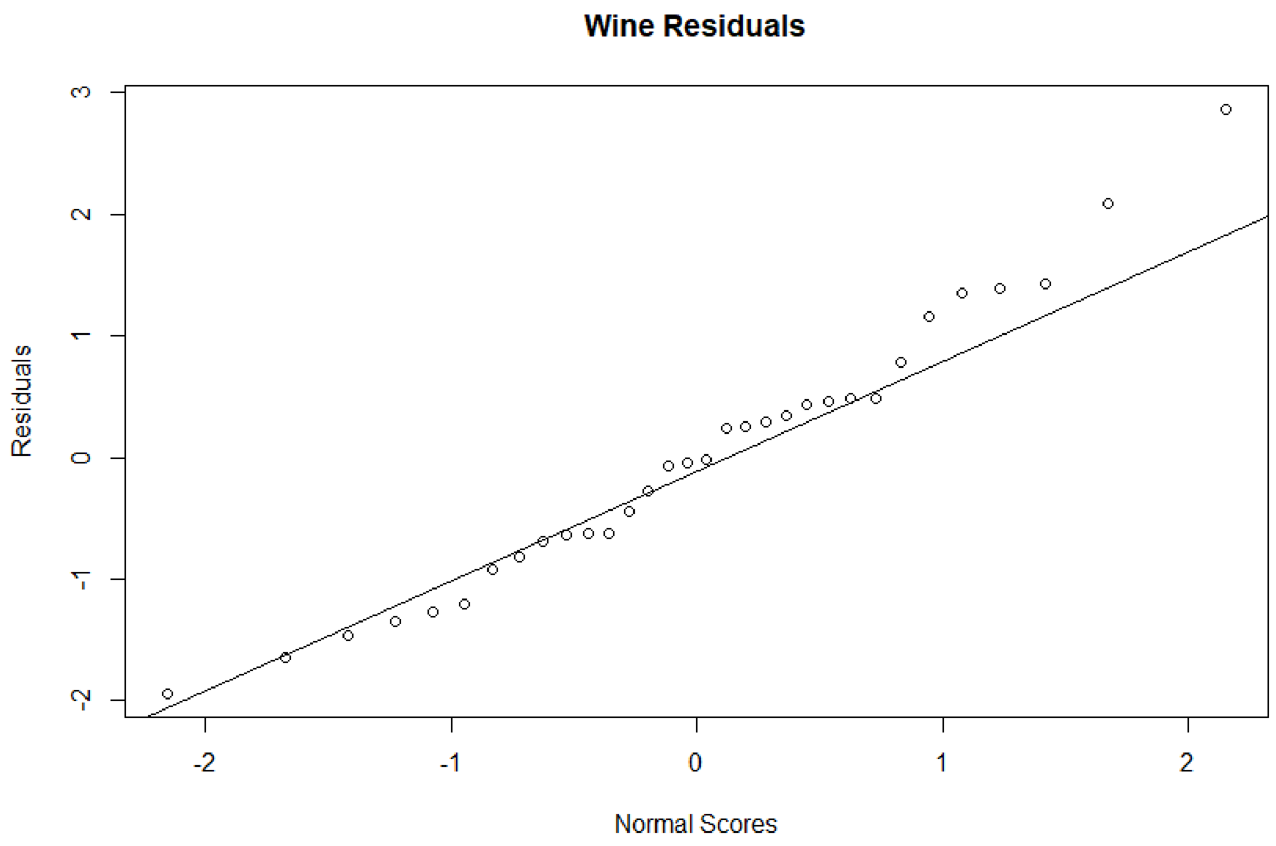
I also attempted to transform various parameters in order to increase model explanatory power. All regressor values were squared, log values were tried, and ph was exponentiated. None of these transformations had a measurable effect on R-Squared.

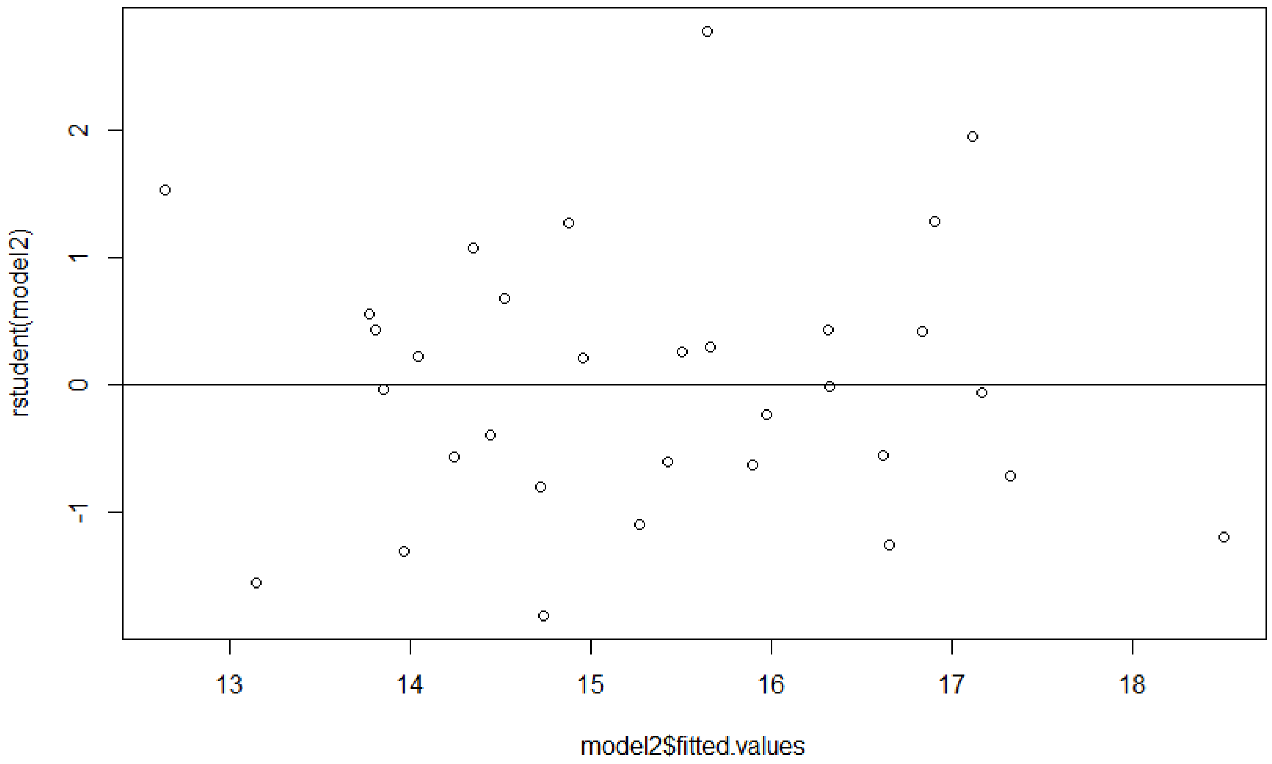
7. Final Model Validation

As a final check, I will break the limited amount of samples into 2 different groups to perform cross validation. The first group will consist of the training set, which will be responsible for building the model parameters. The second group of data will be the test set, where I will run a prediction routine with these values and calculate the errors. The groups were broken down into a training set of 24 data points, and a test set of 8 data points. The resulting graph shows the difference in predicted values versus actual values for the chosen model (predicted in red):



Since the model parameters have changed, I will verify one last time that the residuals satisfy normality by examining the QQ plot, and a plot of studentized residuals.





8. Summary

The final model with parameters is:

Quality = -19.61 - 1.12* Variety + 7.91*pH + .01*sulfites + .28*ionizations

There were no issues with residuals, and all OLS assumptions held for this model. I was not able to discover any interaction terms through trial and error, nor do I expect any of these to logically have a relationship.

R code used for this project:

```
data = read.csv("B19.csv")

colnames(data) <- c("Quality","variety","ph","sulfates", "density",
                   "color","pcolor","acolor","anthocyanins",
                   "ionization","ionanth")

data$variety.f = factor(data$variety)

model = lm("Quality~variety.f + ph + sulfates+density+color+pcolor+acolor + anthocyanins+ionization
+
           ionanth",data=data)
summary(model)

#run correlation test on all values
cor(data[,unlist(lapply(data, is.numeric))])

#Found a completely collinear value - removed ionanth

model = lm("Quality~variety.f + ph + sulfates+density+color+pcolor+acolor +
anthocyanins+ionization",data=data)
summary(model)

#remove acolor
model = lm("Quality~variety.f + ph +
sulfates+density+color+pcolor+anthocyanins+ionization",data=data)
summary(model)


#let's look at the dependent variable on a plot
plot(data$Quality)
abline(h=mean(data$Quality),col="Red")


reduced = lm("Quality~1",data=data)
summary(reduced)

model1 = lm("Quality~variety.f + ph + sulfates+color+pcolor+anthocyanins+ionization",data=data)
anova(model1,model)

model1 = lm("Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins+ionization",data=data)
anova(model1,model)

#new model with color gone
model = lm("Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins+ionization",data=data)
summary(model)
```

```
reduced = lm("Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins",data=data)
summary(reduced)

anova(reduced,model)

model = lm("Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins",data=data)
summary(model)

#Adequacy checking

qqnorm(model$residuals,
      ylab = "Residuals",
      xlab = "Normal Scores",
      main = "Wine Residuals")
qqline(model$residuals)

plot(model$fitted.values,model$residuals)
abline(0,0)

plot(model$fitted.values,rstudent(model))
abline(0,0)

model = lm("Quality~variety.f + ph + sulfates+density+pcolor+acolor+ionization",data=data)

test =  lm("Quality~ variety.f +ph+density+pcolor",data)
summary(test)

library("leaps")

#bring back all factors
model7 = regsubsets(Quality~variety.f + ph + sulfates+density+pcolor+acolor + ionization,data=data)
reg.summary = summary(model7)
par(mfrow=c(2,2))
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS",type="l")
plot(reg.summary$adjr2 ,xlab="Number of Variables ", ylab="Adjusted RSq",type="l")
plot(reg.summary$bic ,xlab="Number of Variables ",ylab="BIC",type='l')
plot(reg.summary$cp ,xlab="Number of Variables ",ylab="Cp", type='l')

library("car")
subsets(model7, statistic="rss")
subsets(model7, statistic="adjr2")
subsets(model7, statistic="cp")
subsets(model7, statistic="bic")

#6 factor model
model = regsubsets(Quality~variety.f + ph + sulfates+density+pcolor+anthocyanins,data=data)
reg.summary = summary(model)
par(mfrow=c(2,2))
plot(reg.summary$rss ,xlab="Number of Variables ",ylab="RSS",type="l")
plot(reg.summary$adjr2 ,xlab="Number of Variables ", ylab="Adjusted RSq",type="l")
plot(reg.summary$bic ,xlab="Number of Variables ",ylab="BIC",type='l')
plot(reg.summary$cp ,xlab="Number of Variables ",ylab="Cp", type='l')

library("car")
#subset options are bic, cp, adjr2, and rss
subsets(model, statistic="adjr2")

#stepwise regression with 7 regressor model

null.model = lm(Quality~1,data=data)
full.model = lm(Quality~variety.f + ph + sulfates+density + pcolor+acolor + ionization,data=data)
test6 = lm(Quality~variety.f + ph + sulfates+density+pcolor + anthocyanins,data=data)

step(null.model,scope=list(lower=null.model,upper=full.model),direction="forward")
step(full.model,data=data,direction="backward")
step(null.model,scope=list(upper=full.model),data=data,direction="both")

#exhaustive search
regsubsets.out <-
  regsubsets(Quality~variety.f + ph + sulfates+density+pcolor+acolor + ionization,
             data = data,
             nbest = 1,       # 1 best model for each number of predictors
             nvmax = NULL,    # NULL for no limit on number of variables
             force.in = NULL, force.out = NULL,
             method = "exhaustive")
a = summary(regsubsets.out)

#final model search
model1 = lm(Quality~variety.f + ph + sulfates+ionization,data = data)
summary(model1)
model2 = lm(Quality~variety+ph+sulfates+ionization,data=data)
summary(model2)

#calculate press statistic and other parameters for table

qpcR::PRESS(model2)
ssres = sum(model2$residuals^2)
msres = ssres/model2$df.residual
ssres
msres

#Cross validation
set.seed(300)
wine.samples <- sample(1:32, 24, replace=F)
wine.new = data[wine.samples,]

b = lm(Quality~variety+ph+sulfates+ionization,data=wine.new)
c = lm(Quality~density,data=wine.new)
```

```
ind = seq(1:32)
wine.deleted_ind = setdiff(ind,wine.samples)
wine.deleted = data[wine.deleted_ind,]

b$fitted.values #gives us fitted values for model b observations
pred_vals = predict.lm(b,wine.deleted)

tss_4 = sum((wine.deleted$Quality - pred_vals)^2)

pred_vals = predict.lm(c,wine.deleted)
tss_1 = sum((wine.deleted$Quality - pred_vals)^2)

plot(wine.deleted$Quality)
points(pred_vals,col="Red")

#examine residuals one last time
plot(model2$fitted.values,rstudent(model2))
abline(h=0)

qqnorm(model2$residuals,
       ylab = "Residuals",
       xlab = "Normal Scores",
       main = "Wine Residuals")
qqline(model2$residuals)

#interaction terms

model.int = lm(Quality~variety.f + ph + sulfates + ionization**2 + density*acolor,data = data)
summary(model.int)
```