Wine Alcohol Linear Regression Analysis

Math 261 Project II

Qian Meng
December, 2020

Outline

- Dataset
- Fit Full Model & Analysis
- Variable Selection
- Final Model
- Adequacy Check
- Extreme Points
- Scaled Model
- Conclusion

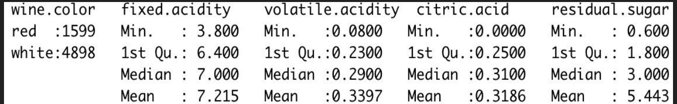
Data Summary

Data set:

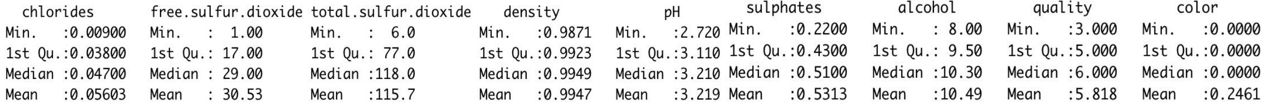
Wine (6,497 obs. Red wine 1,599 obs. White wine 4,898 obs.)

Variables:

- 2 discrete (wine color : 0 - white、1 - red, quality: 3~9)
- 11 continuous (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol)



```
wine.color   fixed.acidity    volatile.acidity  citric.acid      residual.sugar
red  :1599   Min.   : 3.800   Min.   :0.0800    Min.   :0.0000   Min.   : 0.600
white:4898   1st Qu.: 6.400   1st Qu.:0.2300    1st Qu.:0.2500   1st Qu.: 1.800
             Median : 7.000   Median :0.2900    Median :0.3100   Median : 3.000
             Mean   : 7.215   Mean   :0.3397    Mean   :0.3186   Mean   : 5.443
```

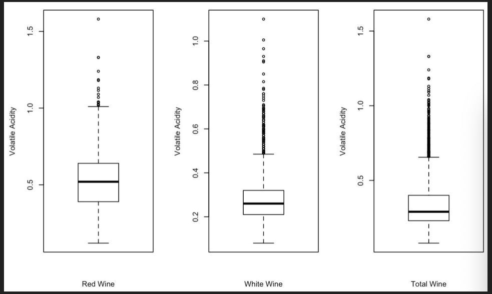


```
  chlorides        free.sulfur.dioxide total.sulfur.dioxide    density             pH          sulphates          alcohol          quality           color
Min.   :0.00900   Min.   :  1.00      Min.   :  6.0        Min.   :0.9871   Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000   Min.   :0.0000
1st Qu.:0.03800   1st Qu.: 17.00      1st Qu.: 77.0        1st Qu.:0.9923   1st Qu.:3.110   1st Qu.:0.4300   1st Qu.: 9.50   1st Qu.:5.000   1st Qu.:0.0000
Median :0.04700   Median : 29.00      Median :118.0        Median :0.9949   Median :3.210   Median :0.5100   Median :10.30   Median :6.000   Median :0.0000
Mean   :0.05603   Mean   : 30.53      Mean   :115.7        Mean   :0.9947   Mean   :3.219   Mean   :0.5313   Mean   :10.49   Mean   :5.818   Mean   :0.2461
```

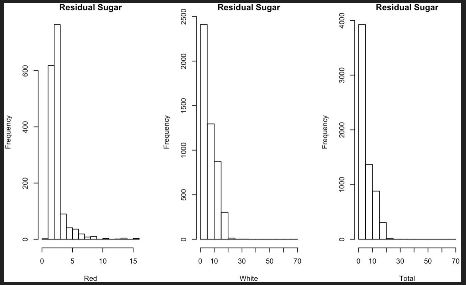
- Response: ~~quality~~ -> alcohol

Data Exploration

Box plots

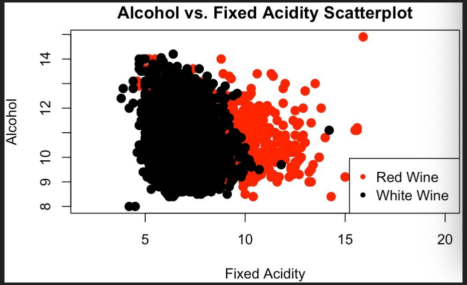


Histograms



Scatterplots

black dot for white wine, Red for red



Findings:

- ❖ Distribution of ‘Total’ and ‘White wine’ seems to be more similar to each other than ‘Red wine’.
- ❖ Several variables(e.g:residual sugar) are slight right-skewed (Mean > median)
- ❖ ‘Red wines’ points are more widely spread.

Multicollinearity

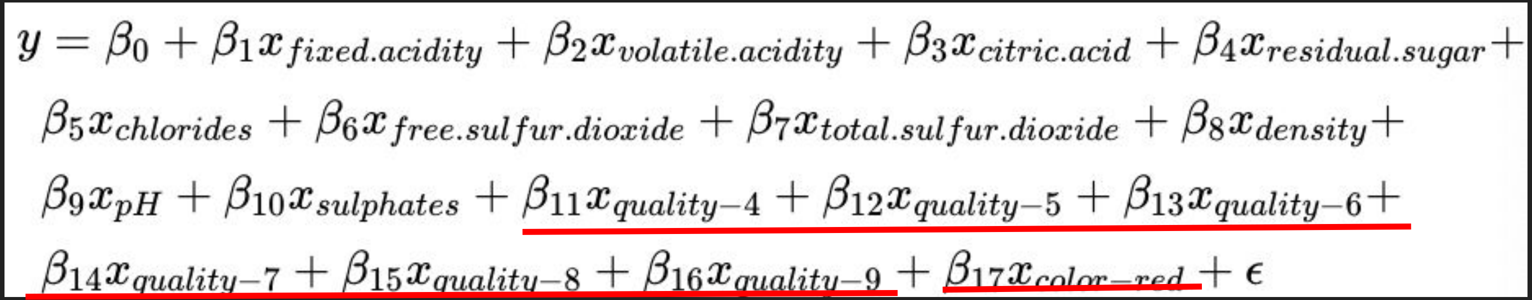
Findings:

- High correlation: total sulfur dioxide & free sulfur dioxide (0.72), volatile acidity & color (0.65), total sulfur dioxide & color (-0.7).
- High VIFs: residual sugar(5.0) , density(6.9), & color(6.3).

Conclusion:

- No strong correlation between variables.
- Using Variable Selection to eliminate redundant predictors.

Fit Full Model



$$\begin{aligned} y = {} & \beta_0 + \beta_1 x_{fixed.acidity} + \beta_2 x_{volatile.acidity} + \beta_3 x_{citric.acid} + \beta_4 x_{residual.sugar} + \\ & \beta_5 x_{chlorides} + \beta_6 x_{free.sulfur.dioxide} + \beta_7 x_{total.sulfur.dioxide} + \beta_8 x_{density} + \\ & \beta_9 x_{pH} + \beta_{10} x_{sulphates} + \underline{\beta_{11} x_{quality-4} + \beta_{12} x_{quality-5} + \beta_{13} x_{quality-6} +} \\ & \underline{\beta_{14} x_{quality-7} + \beta_{15} x_{quality-8} + \beta_{16} x_{quality-9}} + \underline{\beta_{17} x_{color-red}} + \epsilon \end{aligned}$$

- ❖ wine color : 0 - white, 1 - red
- ❖ quality: 3 (low quality) ~ 9 (high quality)

Full Model Analysis

Findings:

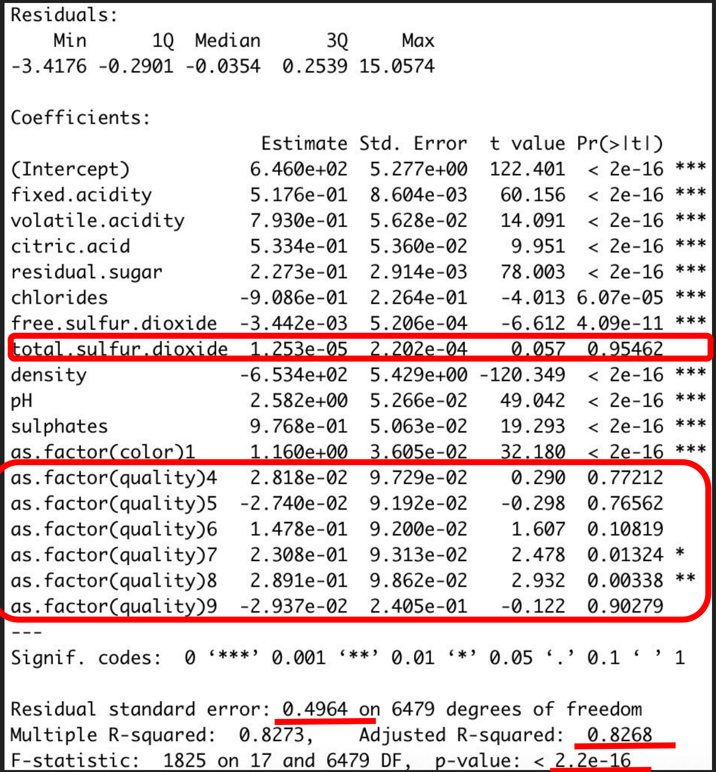
- MSres= 0.25
- Adj R-square = 0.8268
- p-value of the F-statistic is less than 2.2e-16 (significant)

- t-statistic of total sulfur dioxide is not significant.

  > eliminate total sulfur dioxide from model
- t-statistic of quality (level = 3~6,9) is not significant

  > maybe eliminate quality from model *

* Removing quality from model is also supported by ‘Variables Selection’.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.4176 -0.2901 -0.0354  0.2539 15.0574 

Coefficients:
                       Estimate Std. Error  t value Pr(>|t|)    
(Intercept)           6.460e+02  5.277e+00  122.401  < 2e-16 ***
fixed.acidity         5.176e-01  8.604e-03   60.156  < 2e-16 ***
volatile.acidity      7.930e-01  5.628e-02   14.091  < 2e-16 ***
citric.acid           5.334e-01  5.360e-02    9.951  < 2e-16 ***
residual.sugar        2.273e-01  2.914e-03   78.003  < 2e-16 ***
chlorides            -9.086e-01  2.264e-01   -4.013 6.07e-05 ***
free.sulfur.dioxide  -3.442e-03  5.206e-04   -6.612 4.09e-11 ***
total.sulfur.dioxide  1.253e-05  2.202e-04    0.057  0.95462    
density              -6.534e+02  5.429e+00 -120.349  < 2e-16 ***
pH                    2.582e+00  5.266e-02   49.042  < 2e-16 ***
sulphates             9.768e-01  5.063e-02   19.293  < 2e-16 ***
as.factor(color)1     1.160e+00  3.605e-02   32.180  < 2e-16 ***
as.factor(quality)4   2.818e-02  9.729e-02    0.290  0.77212    
as.factor(quality)5  -2.740e-02  9.192e-02   -0.298  0.76562    
as.factor(quality)6   1.478e-01  9.200e-02    1.607  0.10819    
as.factor(quality)7   2.308e-01  9.313e-02    2.478  0.01324 *  
as.factor(quality)8   2.891e-01  9.862e-02    2.932  0.00338 ** 
as.factor(quality)9  -2.937e-02  2.405e-01   -0.122  0.90279    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4964 on 6479 degrees of freedom
Multiple R-squared:  0.8273,	Adjusted R-squared:  0.8268 
F-statistic:  1825 on 17 and 6479 DF,  p-value: < 2.2e-16
```

Variable Selection

Methods:

- Select the best subsets of the variables. (Mallow's Cp statistic)

  R function: leaps() {leaps package} * similar with regsubsets()

- Stepwise regression. (backward, forward, both)

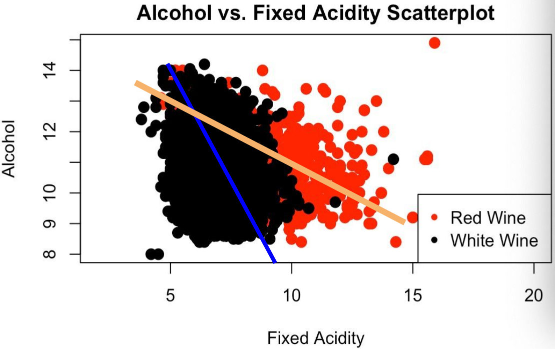
  R function: Step() {stats package}

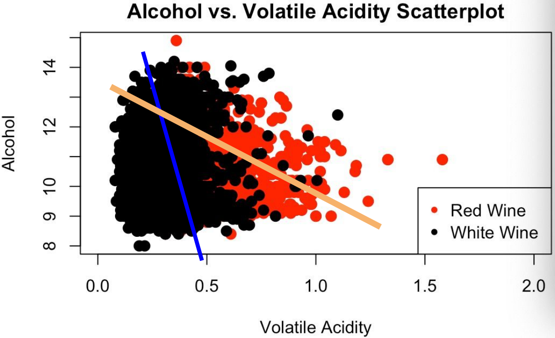
Results:

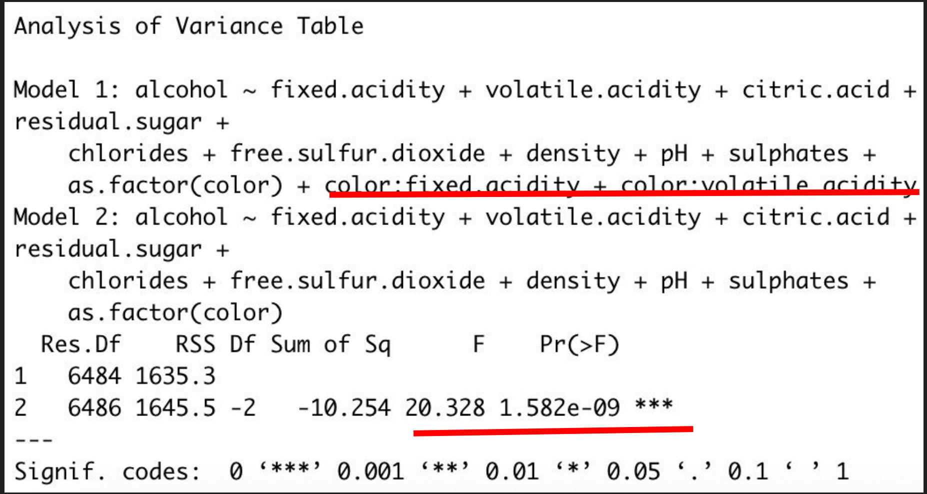
- Eliminate total sulfur dioxide (recall high correlation (0.72) between total sulfur dioxide & free sulfur dioxide)
- Eliminate quality (recall the poor t-test performance)

Interaction Terms

Interaction terms with Indicator Variable: color-red:fixed.acidity, color-red:volatile.acidity





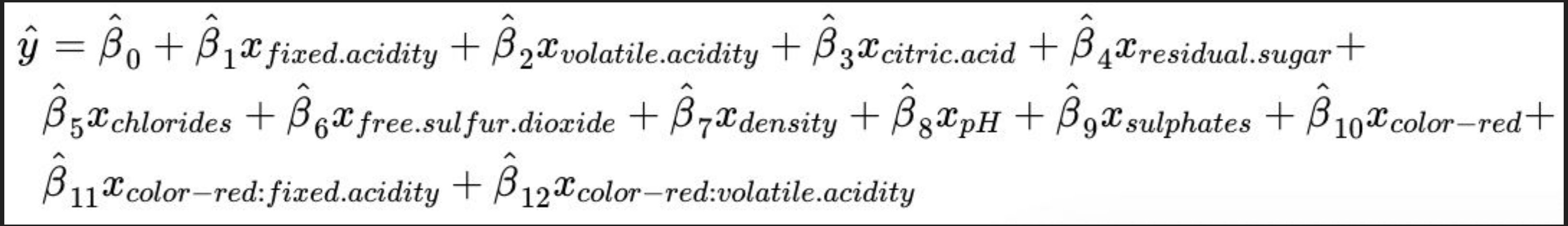


```
Analysis of Variance Table

Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
residual.sugar +
    chlorides + free.sulfur.dioxide + density + pH + sulphates +
    as.factor(color) + color:fixed.acidity + color:volatile.acidity
Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid +
residual.sugar +
    chlorides + free.sulfur.dioxide + density + pH + sulphates +
    as.factor(color)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   6484 1635.3
2   6486 1645.5 -2   -10.254 20.328 1.582e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- ❖ Finding: Model performs better with interactions.

Final model



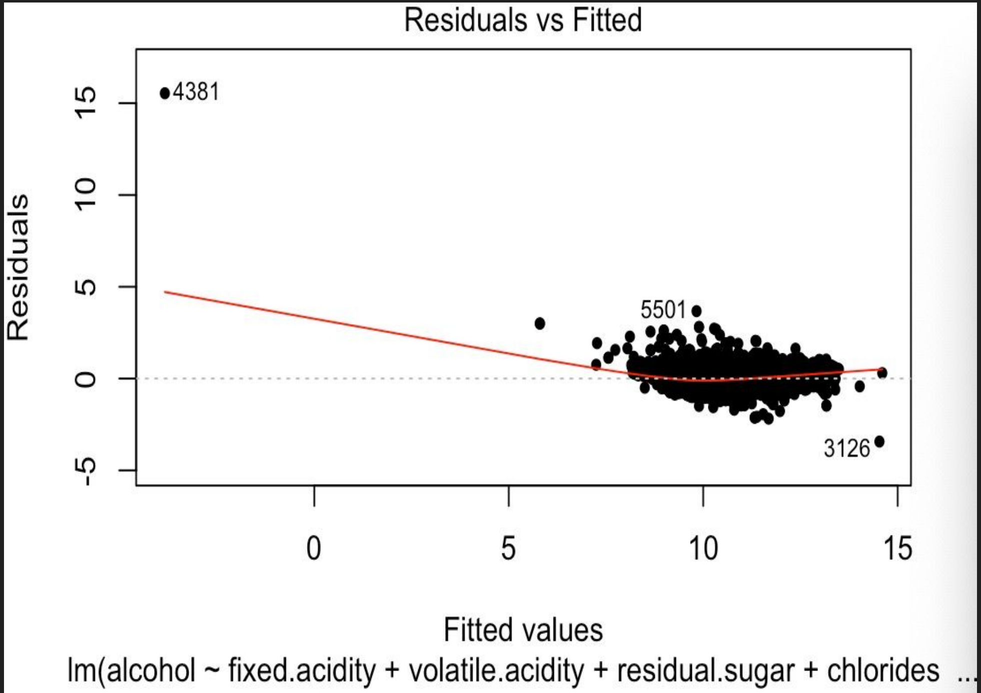
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_{fixed.acidity} + \hat{\beta}_2 x_{volatile.acidity} + \hat{\beta}_3 x_{citric.acid} + \hat{\beta}_4 x_{residual.sugar} + \hat{\beta}_5 x_{chlorides} + \hat{\beta}_6 x_{free.sulfur.dioxide} + \hat{\beta}_7 x_{density} + \hat{\beta}_8 x_{pH} + \hat{\beta}_9 x_{sulphates} + \hat{\beta}_{10} x_{color-red} + \hat{\beta}_{11} x_{color-red:fixed.acidity} + \hat{\beta}_{12} x_{color-red:volatile.acidity}$$

Check Adequacy (Residual plot)

Findings:

- ❖ Outlier: point 4381
- ❖ Constant variance, no clear curvature.

> Does not require higher order terms or linearity transformations.



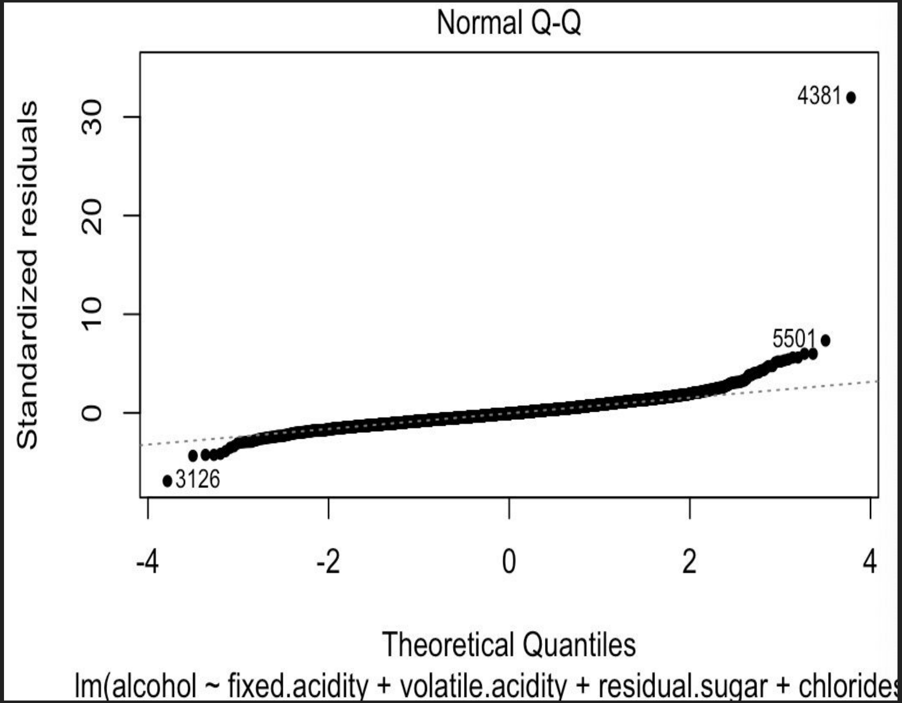
Check Adequacy (Q-Q plot)

Findings:

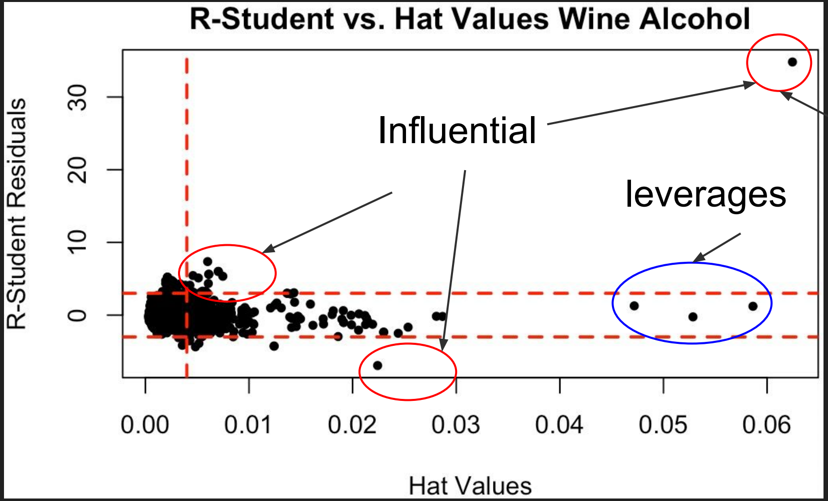
❖ Outlier : point 4381

❖ Heavy tails

Different transformations (y’ = sqrt(y), log(y),1/y) can’t improve.

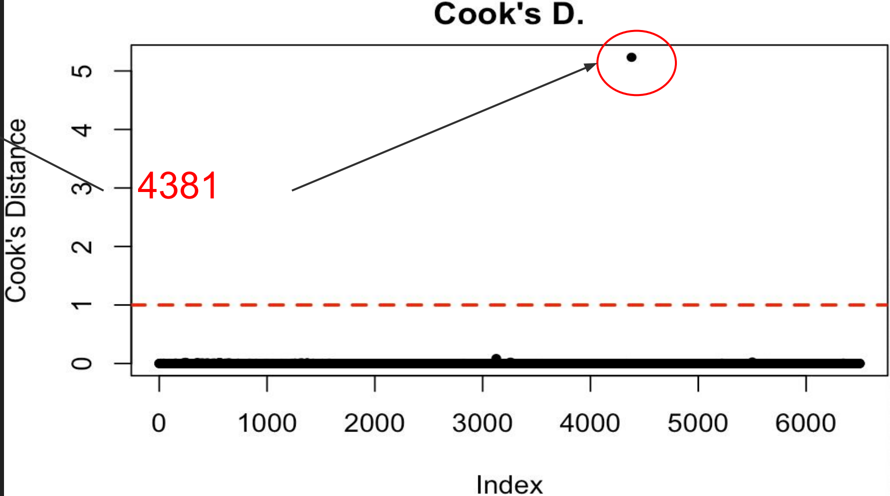


Outlier, Leverage, Influential point



Influential

leverages



4381

Findings:

- ❖ point 4381 : most leverage, largest residual.
  - > Point 4381 is a highly influential point.

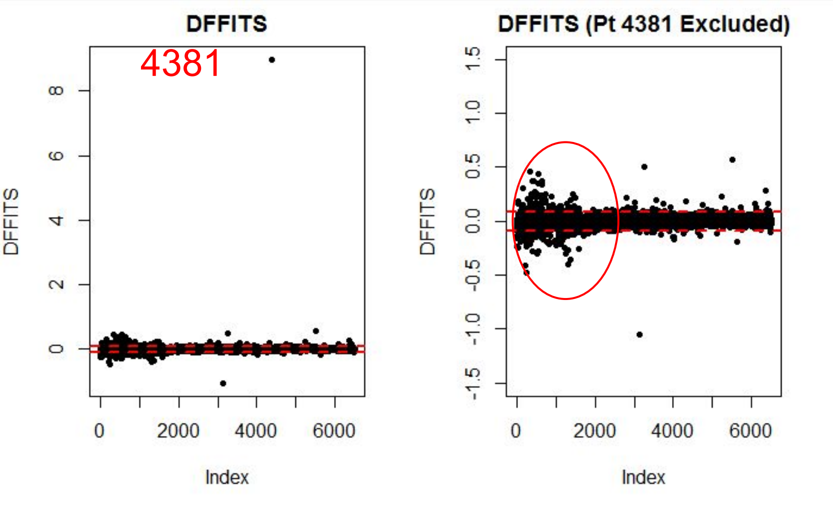
DFBETAS & DFFITS

Findings:

- DFBETAS: 2 points (4381 and 5501) influential in calculating each model parameter.

- DFFITS: Points with lower index values (more Red wine) appear to be more influential.
- White wines appear to exceed the DFFITS threshold less often than red.

*Left 1,599 points ‘**red wine**’,
*Right 4,898 points ‘**white wine**’.

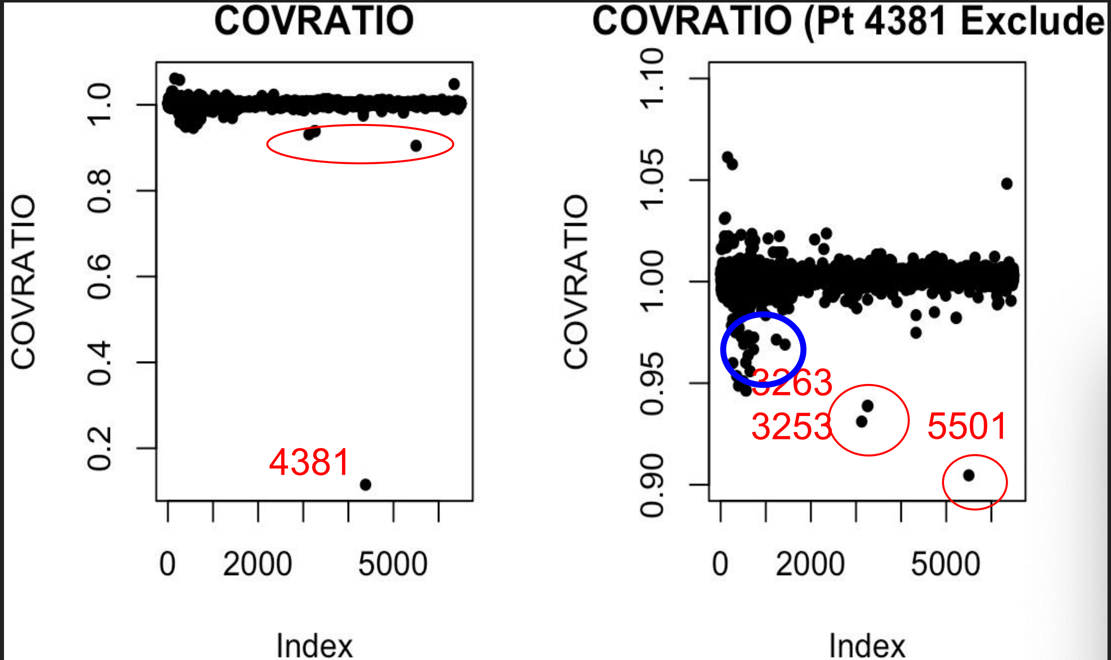


COVRATIO

Findings:

- point 4381 greatly reduces the precision of the model.
- 3 other points that degrade model precision as well. (3253, 3263, and 5501)
- a large cluster of points on the left-hand side that fall below the lower threshold.
  > the model is trained largely to fit white wine
- 118 points degrade model precision with their inclusion in the data set, and 205 points improve model precision.

*Left 1,599 points ‘**red wine**’,
*Right 4,898 points ‘**white wine**’.



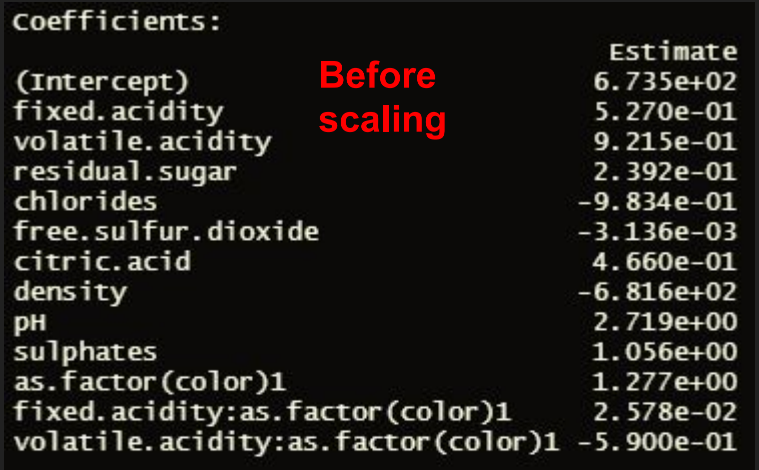
Scaled Model

Findings:

- Coefficient for density is too high.
  (point 4381 has a considerably different density from the other obs.)
  > Unit Normal Scale the response & predictors

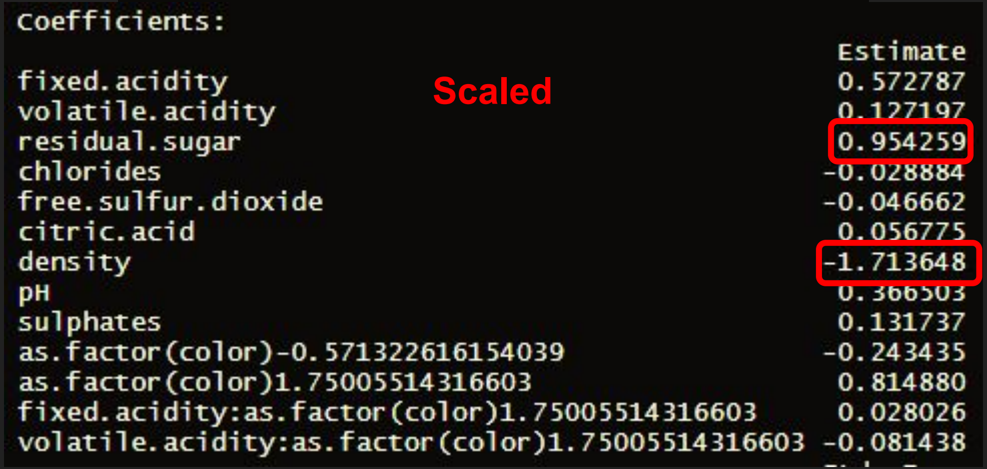
After scaled:

- Residual sugar plays a crucial role.
- Density also plays a crucial role.



Before scaling

```
Coefficients:
                                        Estimate
(Intercept)                            6.735e+02
fixed.acidity                          5.270e-01
volatile.acidity                       9.215e-01
residual.sugar                         2.392e-01
chlorides                             -9.834e-01
free.sulfur.dioxide                   -3.136e-03
citric.acid                            4.660e-01
density                               -6.816e+02
pH                                     2.719e+00
sulphates                              1.056e+00
as.factor(color)1                      1.277e+00
fixed.acidity:as.factor(color)1        2.578e-02
volatile.acidity:as.factor(color)1 -5.900e-01
```



Scaled

```
Coefficients:
                                                   Estimate
fixed.acidity                                      0.572787
volatile.acidity                                   0.127197
residual.sugar                                     0.954259
chlorides                                         -0.028884
free.sulfur.dioxide                               -0.046662
citric.acid                                        0.056775
density                                           -1.713648
pH                                                 0.366503
sulphates                                          0.131737
as.factor(color)-0.571322616154039                -0.243435
as.factor(color)1.75005514316603                   0.814880
fixed.acidity:as.factor(color)1.75005514316603     0.028026
volatile.acidity:as.factor(color)1.75005514316603 -0.081438
```

Conclusion & Further Direction

Differences in properties between red and white wine.

- ❖ Scatterplots, interaction terms indicates a difference.
- ❖ Red appeared to have a higher frequency of points that both enhancing and degrading precision.

  > Future direction: to use the split data sets for red and white wine, and observe how well the final model fits the two data sets separately.

Extreme points

- ❖ The final model produced satisfactory model statistics, the few identified extreme points (eg., 4381) give undue influence on the model.
- ❖ Examined the model fit with the specified point omitted. -> adj R-square increase from .8227 to .8412.

  > Future direction: fit a model based on dataset that do not possess extreme values of density and residual sugar.
- ❖ Another possible reason the model does not fit particularly well is due to a lack of sampling wines with high density and high residual sugar.

  > Future direction: gather a larger sample size and pay special attention to these variables, ensuring an adequate representation of various wines.

References

[1] 261 Course Slides (By Dr. Guangliang Chen)

[2] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.