Investigating the relationship between alcohol content in wine and its physicochemical properties

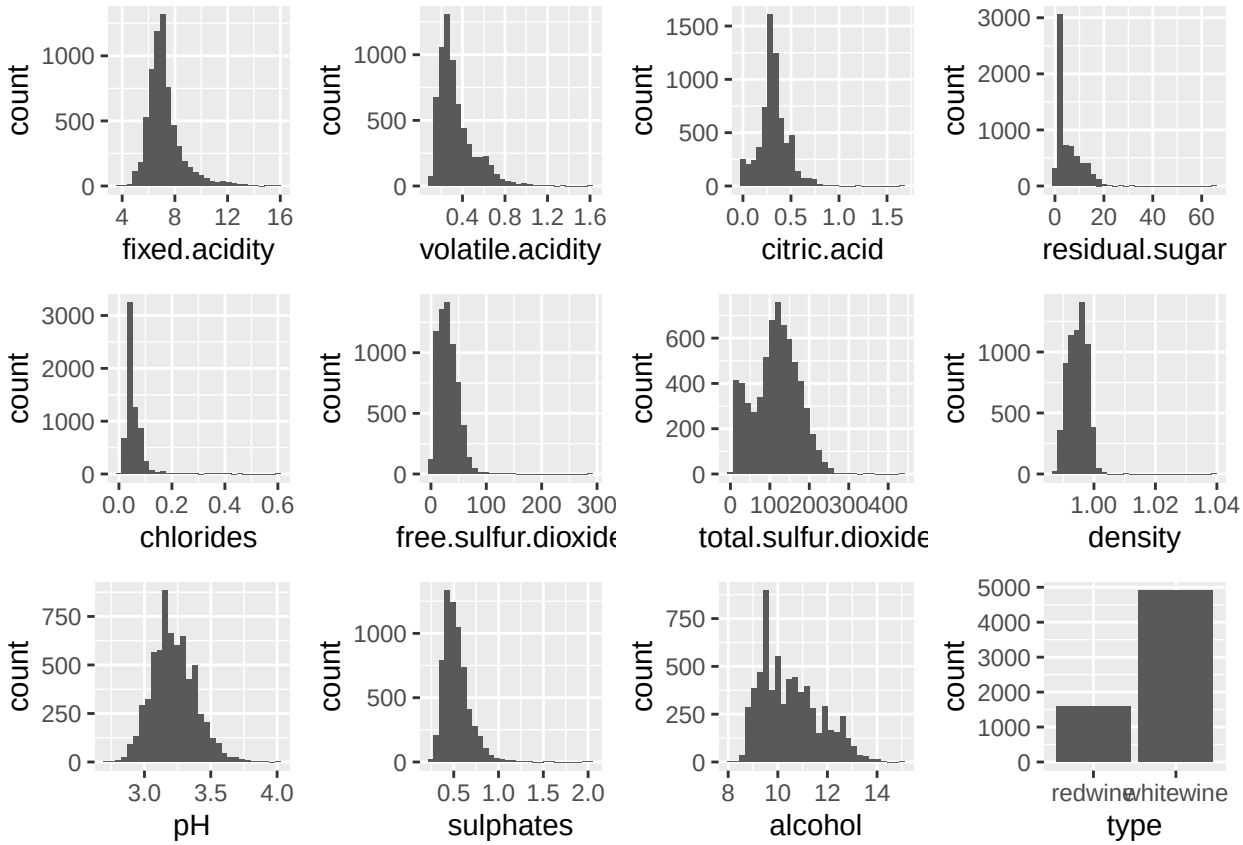
Anav Vora and Lavanya Kudli

12/16/2024

Introduction The relationship between alcohol content in wine and other physicochemical properties is of particular interest for brewers. In this report, we detail steps for the selection of the best model for predicting the alcohol content and also the best model for estimating the alcohol content.

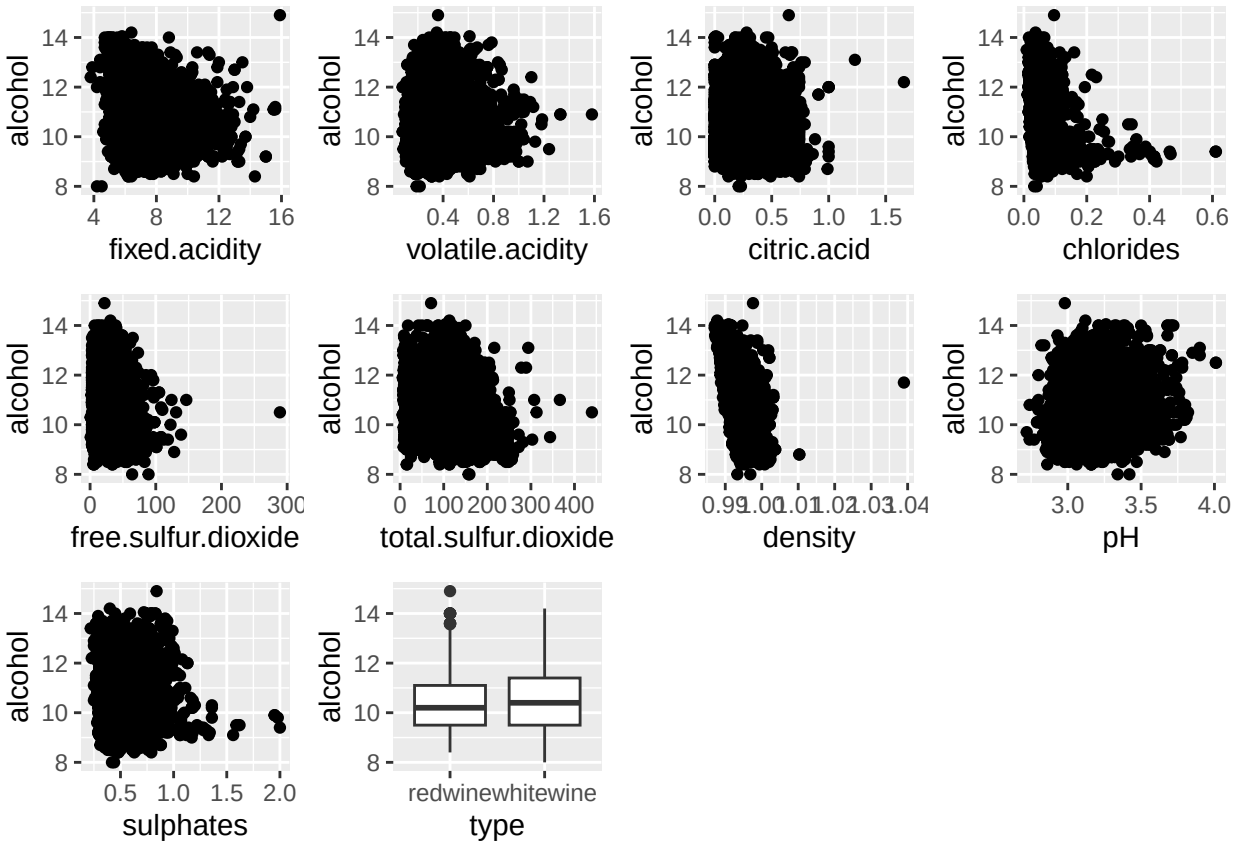
Data Description

We begin our analysis by conducting exploratory data analysis.



From the historgam we see that some predictors appear to have a normal distribution such as fixed.acidity, citric.acid, pH, alcohol, free.sulphur.dioxide, and density. Total.sulphur.dioxide actually seems to have a bimodal distribution. The number of whitewines are more than the redwines in the dataset.The rest of the predictors seem to have a skewed distribution (volatile.acidity,residual.sugar,chlorides, and sulphates).

Next, we use scatter plots and box and whisker plots to get a rough idea regarding the influence of various predictors on the alcohol percentage.



We observe that all the scatter plots contain clouds of data and no particular trend can be seen. Perhaps, 2-D scatter plots do not capture the complex trends in the data. Additionally, the median alcohol contents of both redwine and white wine are close enough. There seem to be some outliers in the redwine.

Model A: Building for Prediction

Part 1) AIC criterion

```
## Start:  AIC=-6871.24
## alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + type
##
##                         Df Sum of Sq    RSS     AIC
## - total.sulfur.dioxide  1       0.5 1379.4 -6871.3
## <none>                              1378.9 -6871.2
## - chlorides             1       4.4 1383.3 -6856.7
## - free.sulfur.dioxide   1       6.0 1384.9 -6850.7
## - citric.acid           1      22.9 1401.9 -6787.5
## - volatile.acidity      1      28.4 1407.3 -6767.4
## - sulphates             1      98.4 1477.4 -6514.9
## - type                  1     234.9 1613.9 -6055.6
## - pH                    1     541.1 1920.1 -5152.7
## - fixed.acidity         1     838.2 2217.1 -4405.3
## - residual.sugar        1    1566.9 2945.8 -2928.4
## - density               1    3778.0 5157.0   -18.2
##
```

```
## Step:  AIC=-6871.33
## alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     type
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               1379.4 -6871.3
## + total.sulfur.dioxide  1       0.5 1378.9 -6871.2
## - chlorides             1       4.3 1383.7 -6857.2
## - free.sulfur.dioxide   1      12.8 1392.2 -6825.3
## - citric.acid           1      22.5 1401.9 -6789.3
## - volatile.acidity      1      27.9 1407.3 -6769.3
## - sulphates             1      97.9 1477.4 -6516.9
## - type                  1     368.0 1747.4 -5644.5
## - pH                    1     544.3 1923.7 -5144.9
## - fixed.acidity         1     860.4 2239.9 -4354.1
## - residual.sugar        1    1604.9 2984.3 -2862.8
## - density               1    4257.2 5636.7   442.1
```

We see that eliminating total sulfur dioxide reduces the AIC value from -6871.24 to -6871.3. Therefore, based on AIC criterion we select a model with 10 predictors and exclude total.sulfur.dioxide.

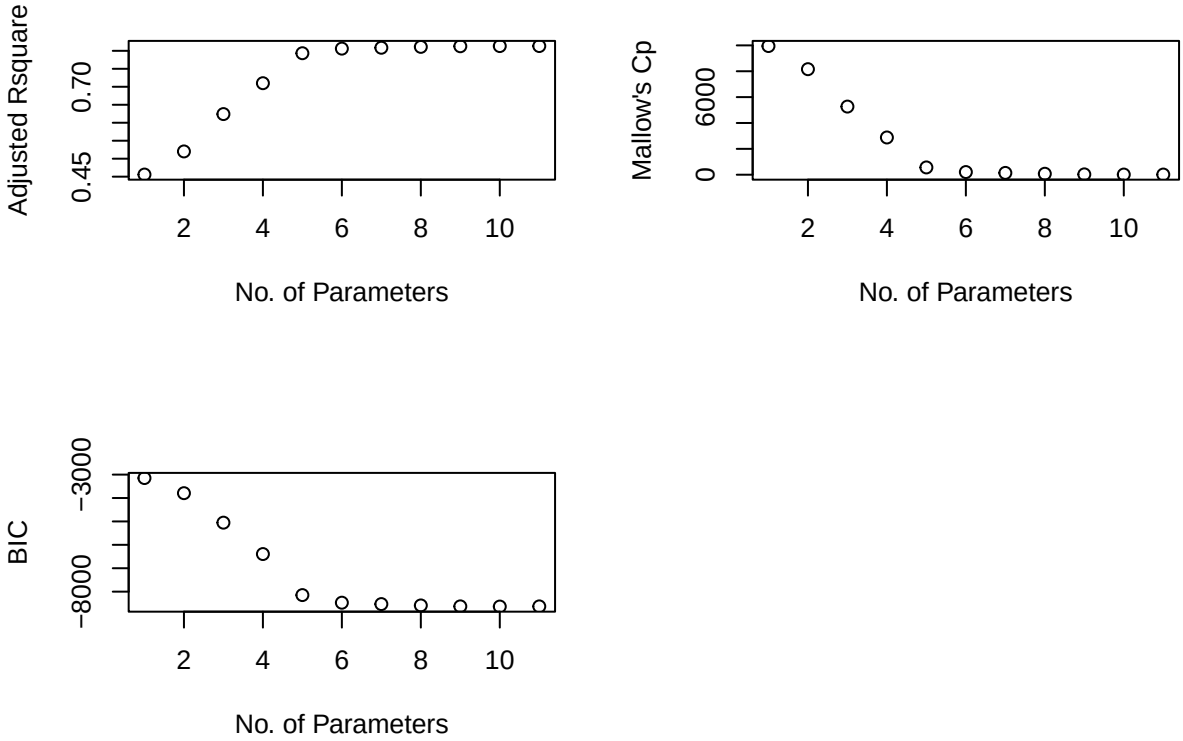
Part 2) BIC criterion

```
## Start:  AIC=-6789.89
## alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + type
##
##                         Df Sum of Sq    RSS     AIC
## - total.sulfur.dioxide  1       0.5 1379.4 -6796.8
## <none>                               1378.9 -6789.9
## - chlorides             1       4.4 1383.3 -6782.1
## - free.sulfur.dioxide   1       6.0 1384.9 -6776.2
## - citric.acid           1      22.9 1401.9 -6713.0
## - volatile.acidity      1      28.4 1407.3 -6692.8
## - sulphates             1      98.4 1477.4 -6440.4
## - type                  1     234.9 1613.9 -5981.0
## - pH                    1     541.1 1920.1 -5078.2
## - fixed.acidity         1     838.2 2217.1 -4330.7
## - residual.sugar        1    1566.9 2945.8 -2853.8
## - density               1    3778.0 5157.0    56.4
##
## Step:  AIC=-6796.76
## alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
##     type
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               1379.4 -6796.8
## + total.sulfur.dioxide  1       0.5 1378.9 -6789.9
## - chlorides             1       4.3 1383.7 -6789.4
## - free.sulfur.dioxide   1      12.8 1392.2 -6757.5
## - citric.acid           1      22.5 1401.9 -6721.5
```

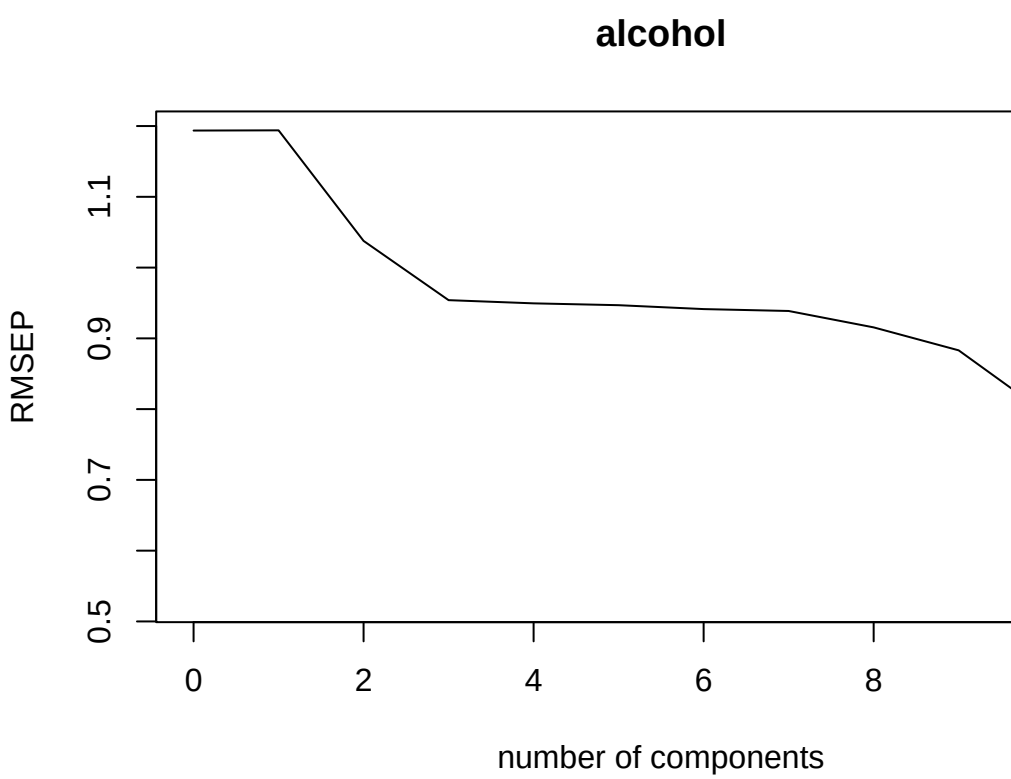
```
## - volatile.acidity      1      27.9 1407.3 -6701.5
## - sulphates             1      97.9 1477.4 -6449.1
## - type                  1     368.0 1747.4 -5576.7
## - pH                    1     544.3 1923.7 -5077.1
## - fixed.acidity         1     860.4 2239.9 -4286.3
## - residual.sugar        1    1604.9 2984.3 -2795.0
## - density               1    4257.2 5636.7   509.9
```

We see that eliminating total sulfur dioxide reduces the AIC value from -6789.89 to -6796.8. Therefore, based on BIC criterion we select a model with 10 predictors and exclude total.sulfur.dioxide.

Part 3) Leaps and bound based methods. We used nvmax as 11 to include all the predictors



From the leaps and bound adjusted R-square we see that the number of parameters chosen is 11 parameters. However from leaps and bound BIC and Cp mallows criterion we see that the chosen value is 10 parameters.

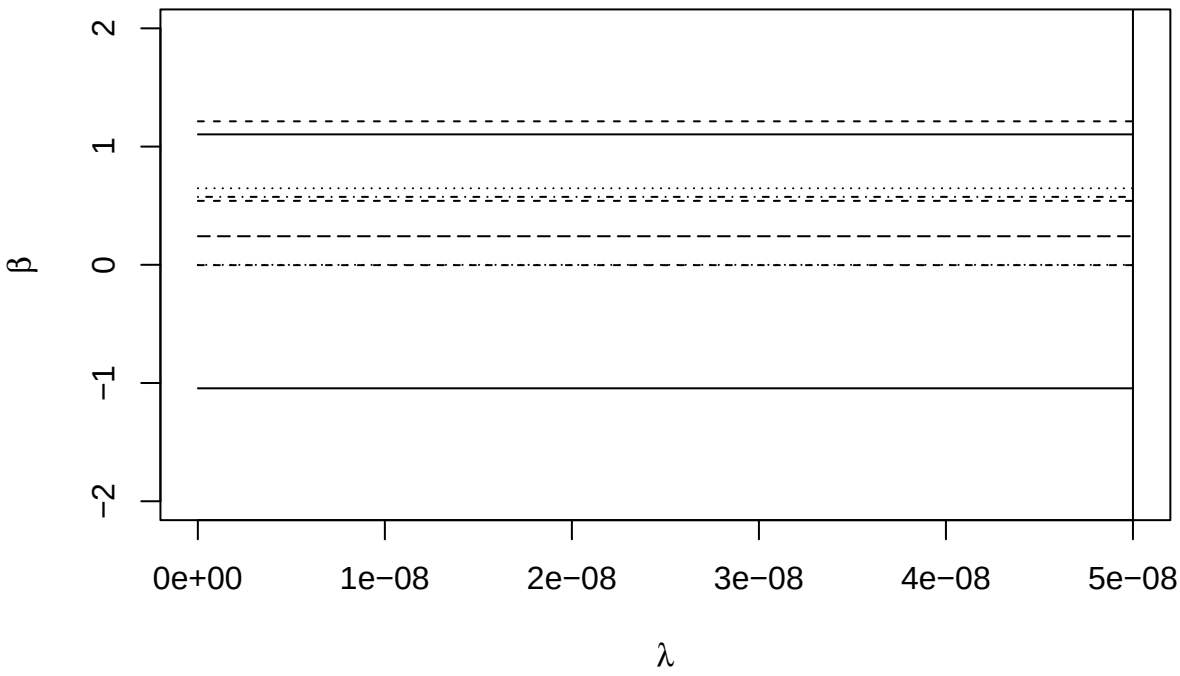


Part 4) Principal component analysis

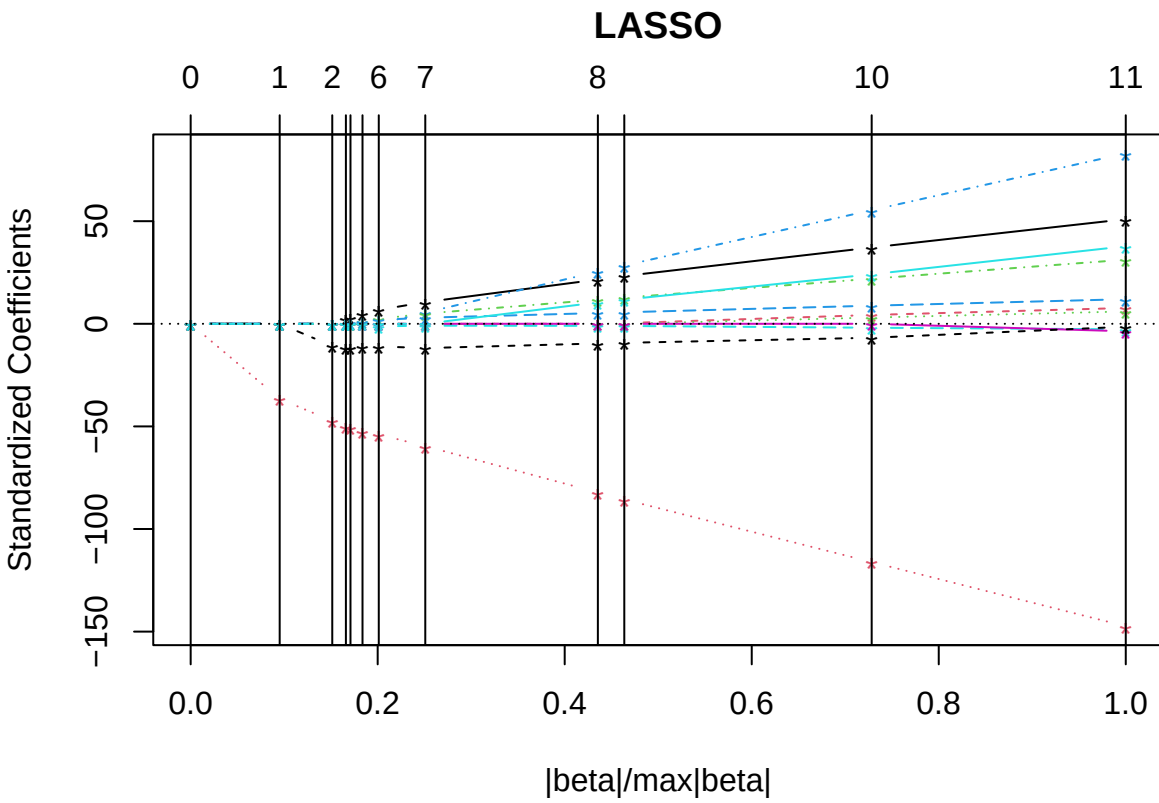
PCR CV suggests we should use 11 components, which is as many predictors as we have. Basically, we don't need to do PCR then, just use all the predictors

Part 5) Ridge regression

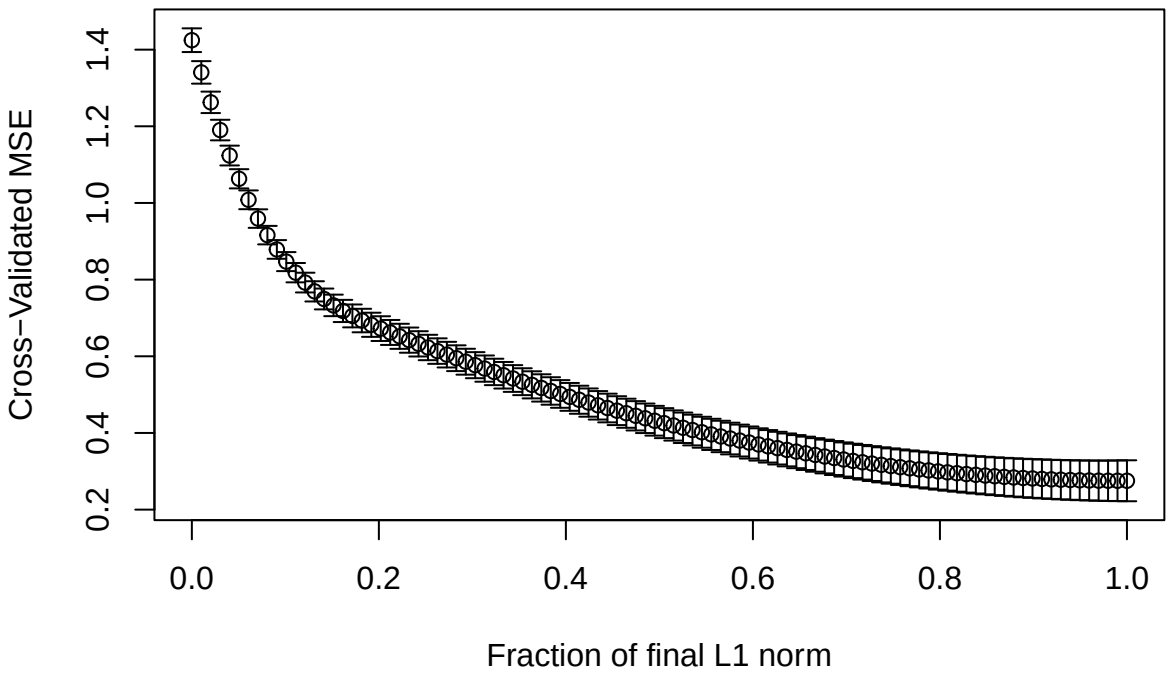
```
## 5.000000e-08
##           31
```



We construct a ridge trace plot to determine the appropriate range of λ, using generalized cross validation (GCV). We obtain an optimal lambda of 5.000000e-08. Our plot is insensitive to lambda meaning the penalty does not seem to have much of an effect on the model. This means we can select the entire model with all the 11 predictors.



Part 6) Lasso regression
The x-axis is the value of the L1 norm of the coefficients relative to the norm of the LS solution t. As this value increases more predictors enter into the model. Initially at lower t-values, only predictor 8 enters the model and at the highest t-value all the predictors enter the model.




```
## [1] 99
```

```
## [1] 0.989899
```

```
##        fixed.acidity     volatile.acidity          citric.acid
##         5.351427e-01         6.364011e-01         5.639194e-01
##       residual.sugar            chlorides  free.sulfur.dioxide
##         2.384724e-01        -1.033595e+00        -2.723282e-03
## total.sulfur.dioxide              density                   pH
##        -4.003188e-04        -6.722773e+02         2.675966e+00
##            sulphates                 type
##         1.092417e+00         1.197269e+00
```

The optimal value of t can be selected by Cross-Validation, we see that the optimal t value is 0.989. None of the coefficients obtained by LASSO have a non-zero value, all the predictors are retained.

Model A Testing

Overall we have 4 models for testing. The first is using all the predictors for regression as suggested by PCR and the leap-and-bounds algorithm with adjusted R^2. Next, the greedy algorithms based on AIC and BIC, and the leap-and-bounds algorithm with BIC and Mallow's Cp suggest that all the predictors except total sulphur dioxide should be used for regression giving us the second model. The final two models correspond to ridge and lasso regression.

```
## [1] 0.4545454
```

```
## [1] 0.4543447
```

```
## [1] 0.4545454
```

```
## [1] 0.4546584
```

The smallest RMSE is (0.4543) for the model made without total sulphur dioxide obtained from the greedy algorithms based on AIC and BIC, and the leap-and-bounds algorithm with BIC and Mallow's Cp.

We will do diagnostics for that model. First looking at influential observations

```
## [1] 298
```

```
## [1] 21
```

We observe that there are 298 high leverage points of which 21 are bad high leverage points. Next, we examine if there are some points which do not fit the model as well as others points, i.e., if there are any outliers.

```
##      4381      3126      3263      3253       560       565       396       354
## 34.484852  6.961415  5.694930  5.694930  5.513303  5.513303  5.333045  5.153819
##       500       494       268       557       559       609      1427       511
##  4.960021  4.960021  4.631542  4.602966  4.602966  4.344343  4.302043  4.175661
##       727       634       654       610
##  4.127996  4.005790  4.005570  3.960728
```

```
## [1] -4.429918
```

```
## [1] 12
```

We see that the bonferroni cv is -4.43 and there are 12 outliers since 12 studentized residuals have a value large than 4.43.

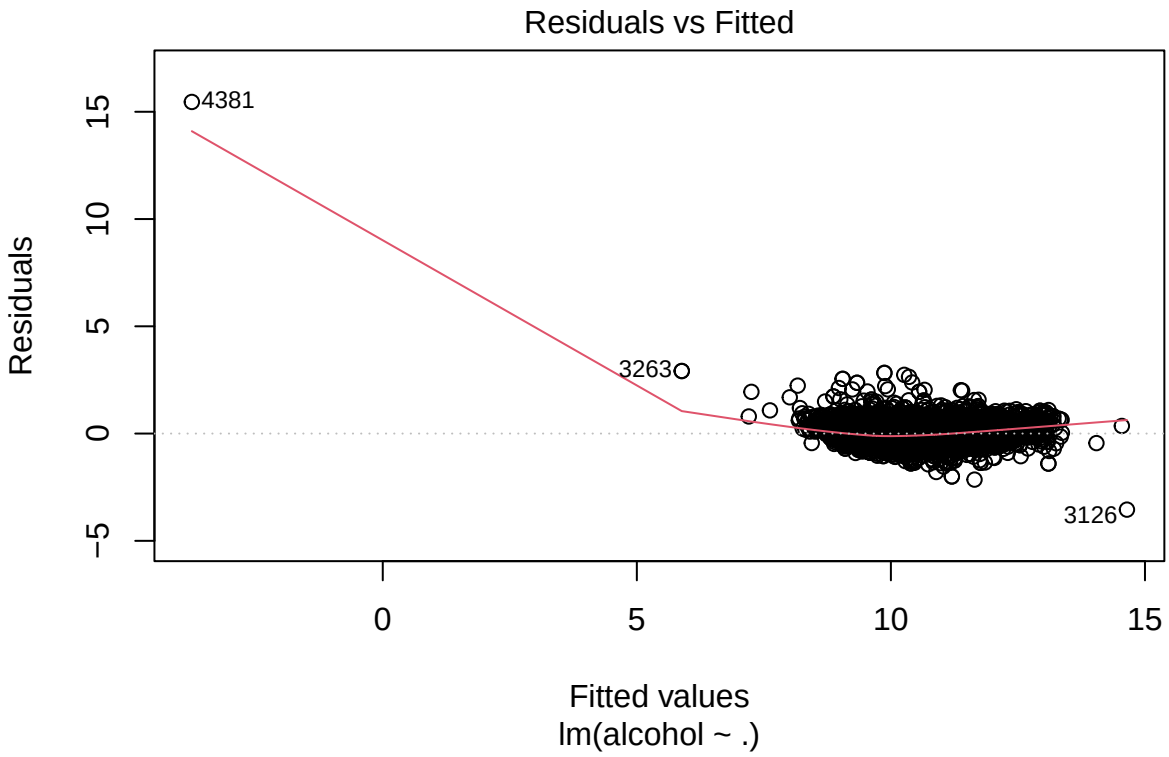
Finally, we check if there are any individual points that affect the model parameters, i.e., influential points

```
##       4381       3126       3263       3253       1320        227        354
## 6.78274632 0.07040062 0.02545111 0.02545111 0.01832832 0.01826108 0.01544016
##       1373       1371        560
## 0.01510671 0.01510671 0.01429681
```

There is 1 observation has cooks distance above 1, so there is one influential point in the data.

We now test if MLR model assumptions are satisfied. Note that, we do not need to check for independence of error terms (i.e., correlation) as the data we have is not collected in an ordered sequence.

First we check if the variance is constant using the visual test and the bp test.




```
##
##  studentized Breusch-Pagan test
##
## data:  Leaps_Based_BIC_Model
## BP = 408.38, df = 10, p-value < 2.2e-16
```

We see some extra points on the left.

The hypothesis for the BP test is,

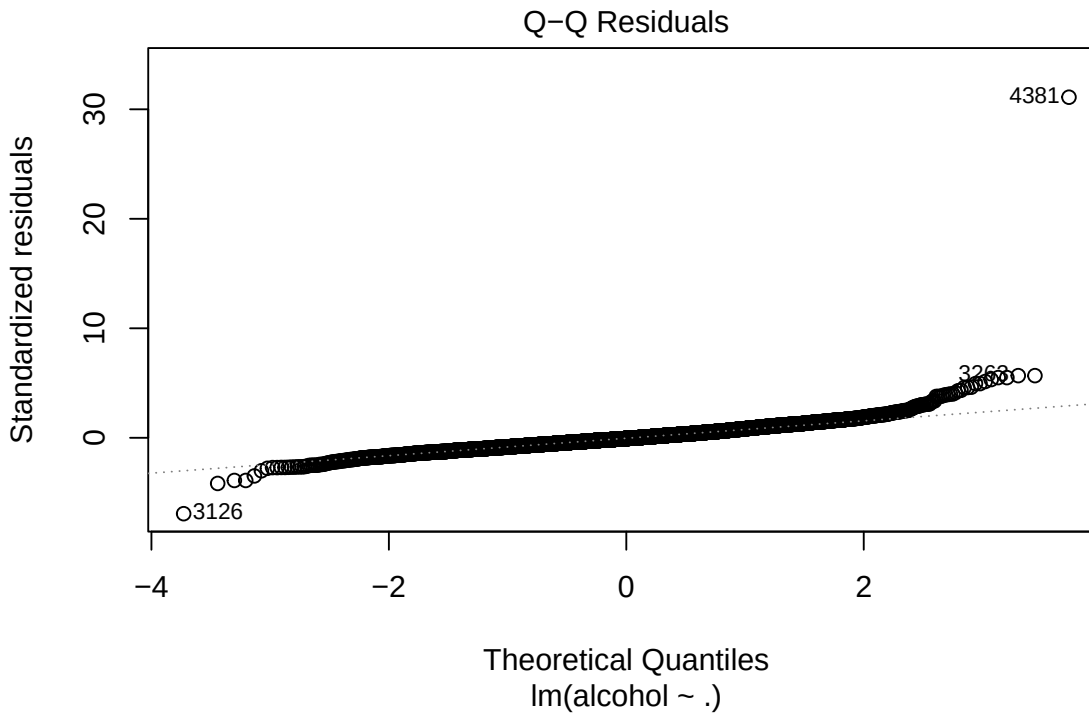
H0 :the variance is constant

Hα: the variance is not constant

Since the p-value is less than 0.05, we reject the null and conclude that the variance is not constant.

Next, we check if the error terms are normally distributed by using the following two approaches:-

1) a Q-Q plot



2) a KS test since our n>50

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  Leaps_Based_BIC_Model$residuals
## D = 0.20101, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

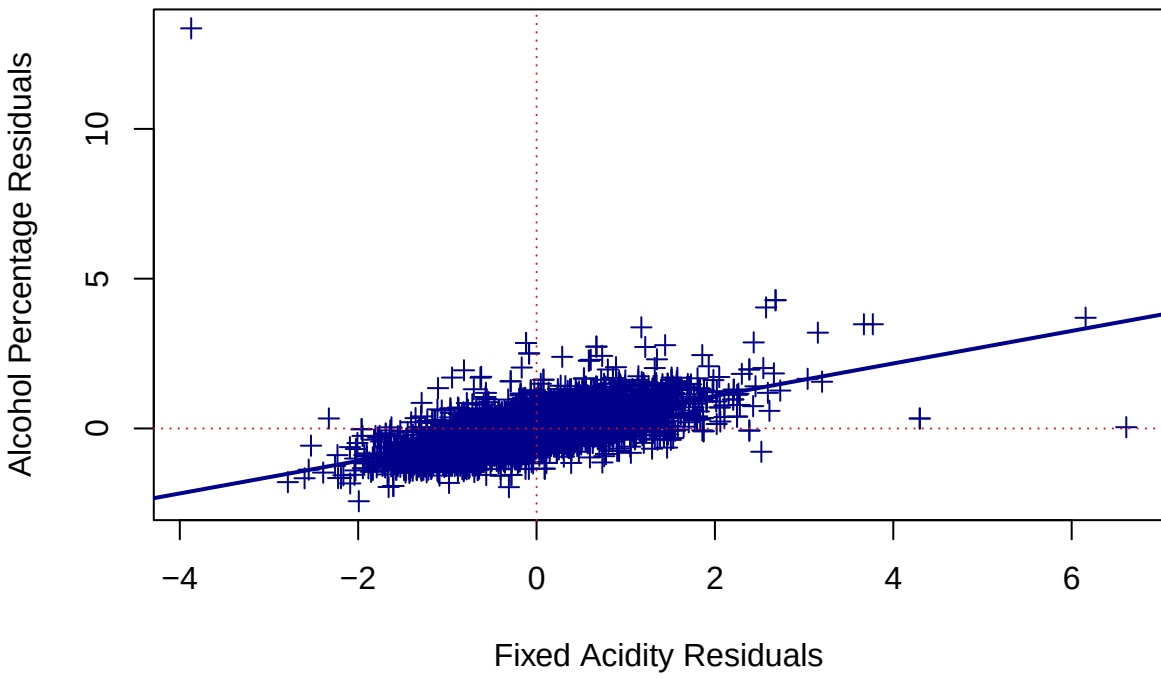
From the Q-Q plot we see that there is a departure from normality at the lower end and the upper end.

We also do the KS test with the below hypothesis:

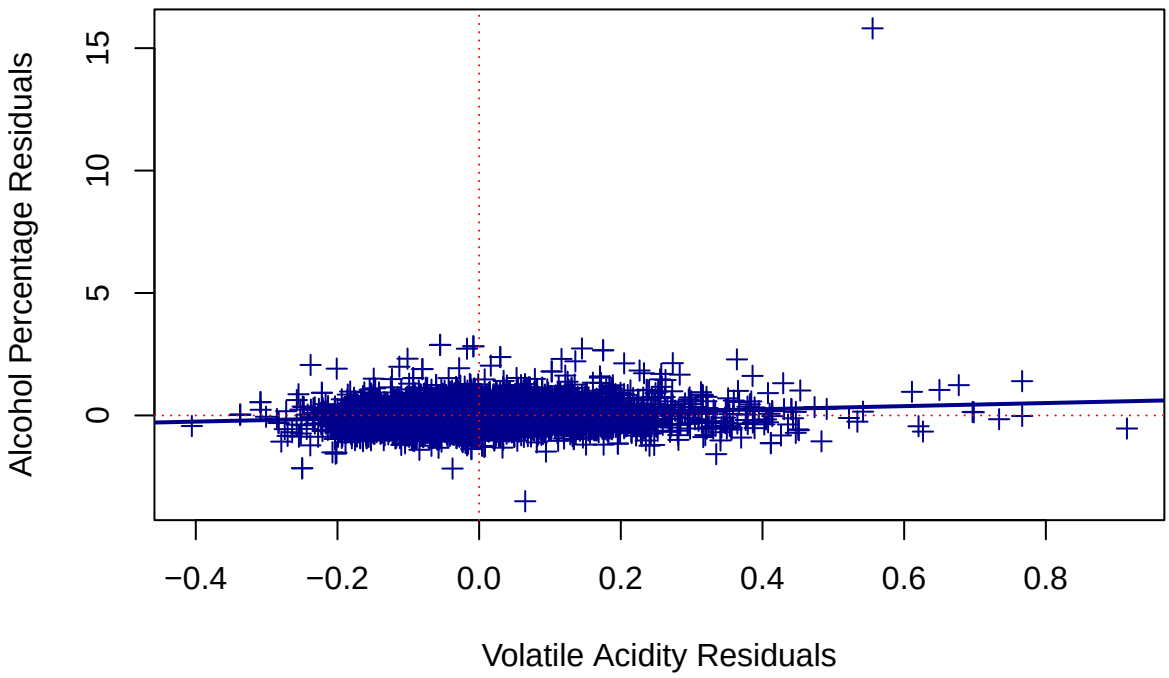
H0: the distribution is normal

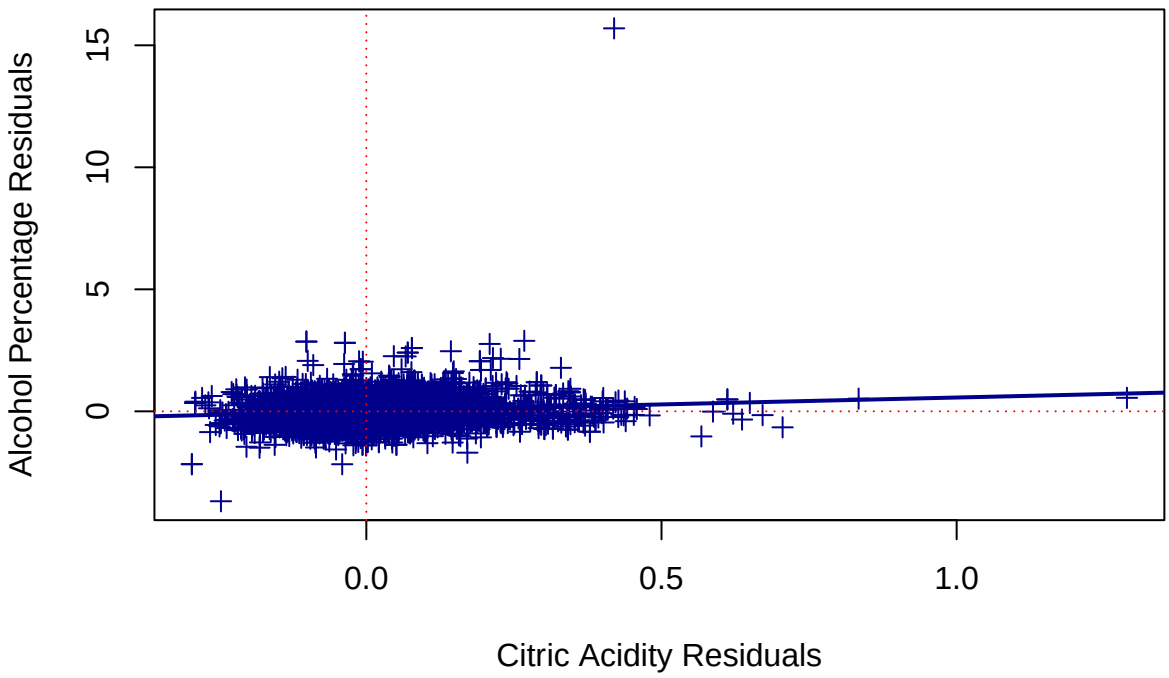
Hα: the distribution is not normal

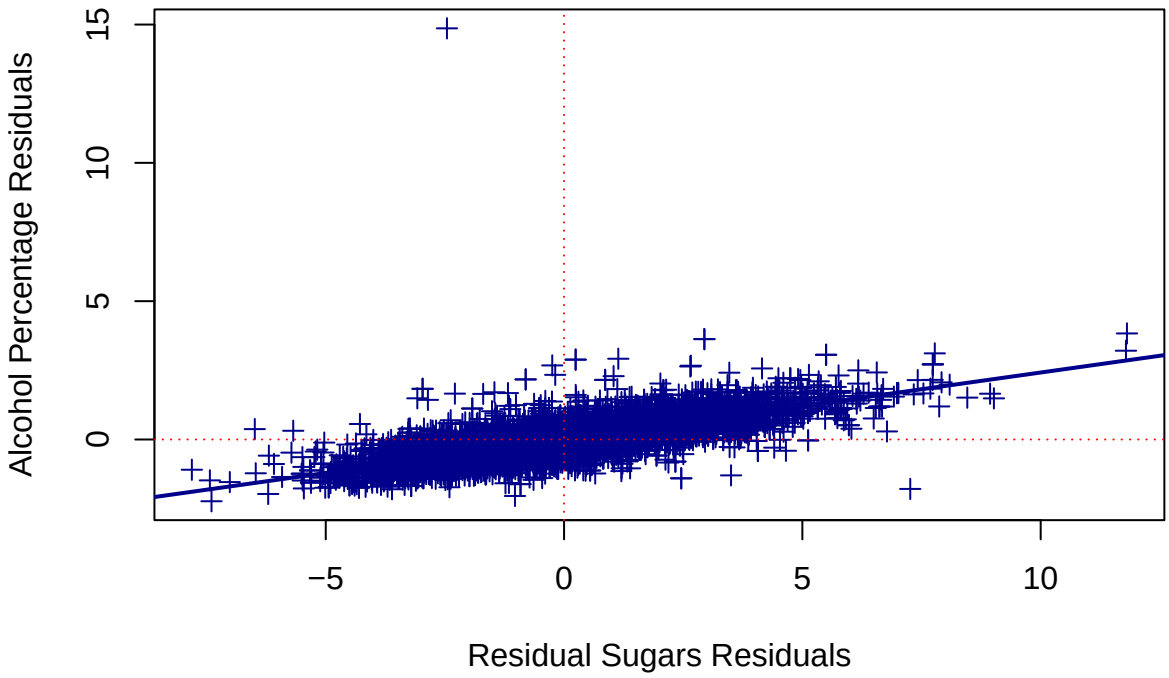
The p-value of 2.2e-16 is less than 0.05. So, we reject the null hypotheses of normality and conclude that the normality assumption is not satisfied.

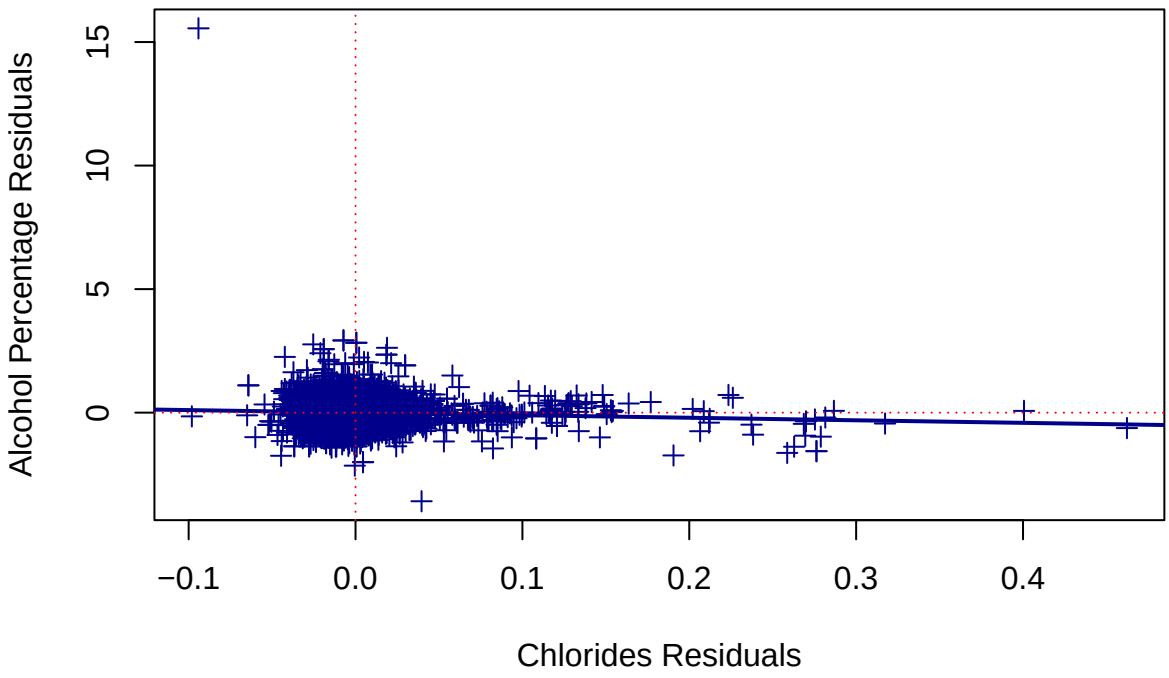


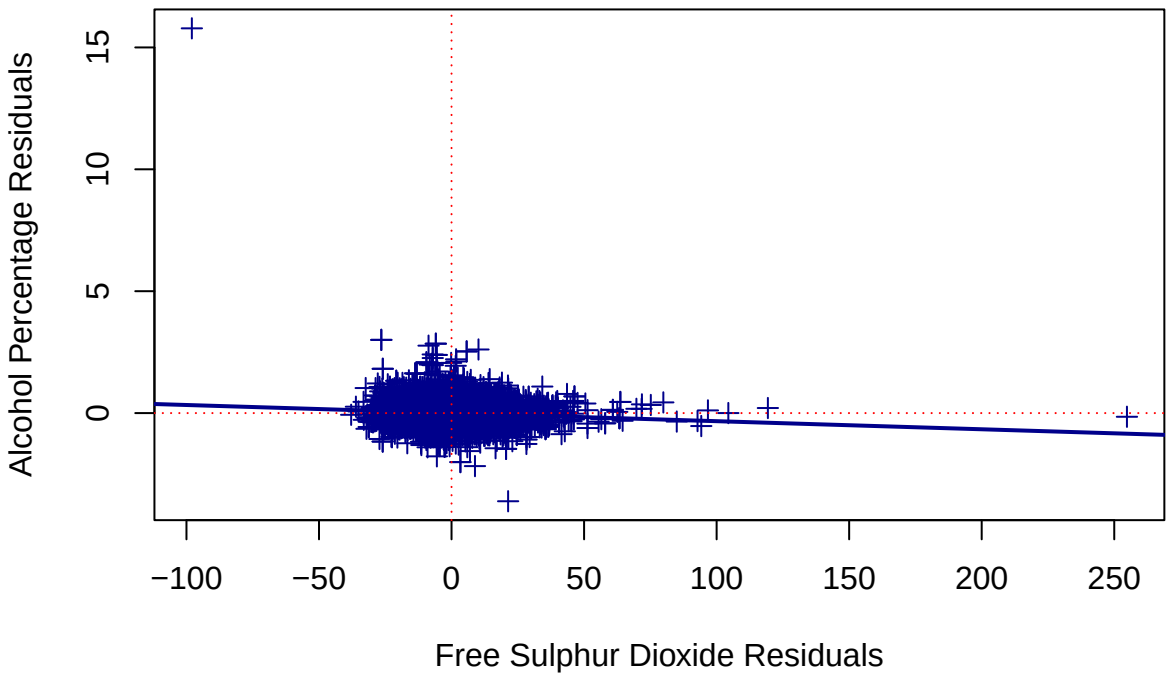
Now checking linearity:

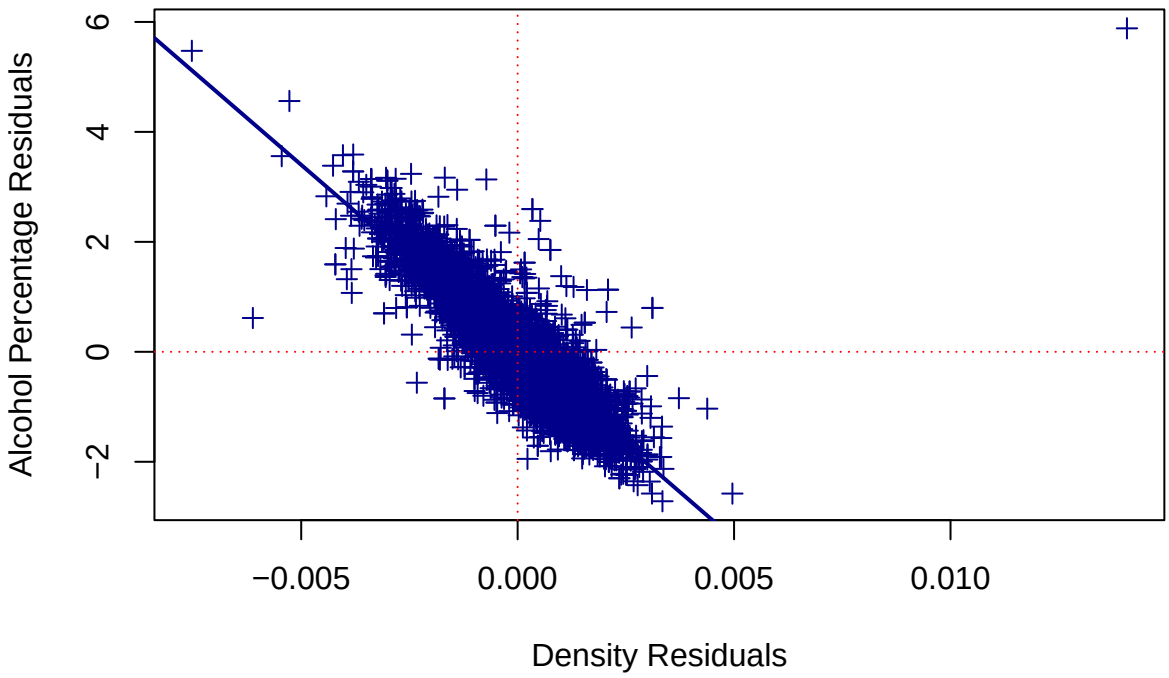

Alcohol Percentage Residuals
15
10
5
0
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Volatile Acidity Residuals

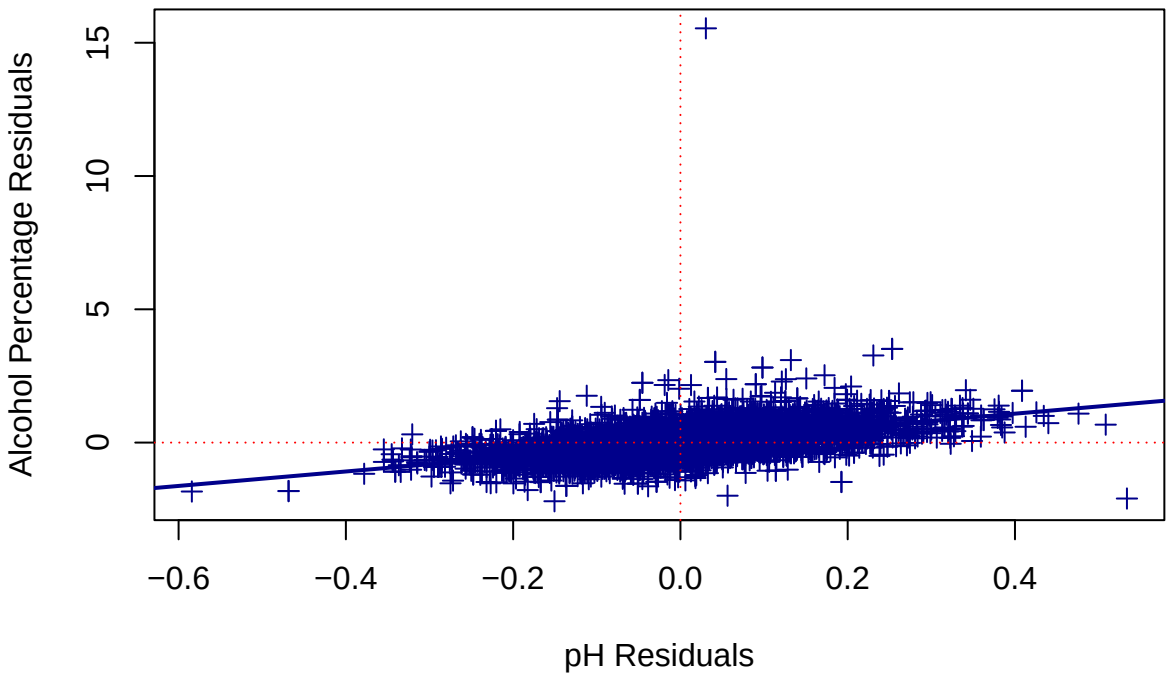

Alcohol Percentage Residuals
15
10
5
0
0.0
0.5
1.0
Citric Acidity Residuals

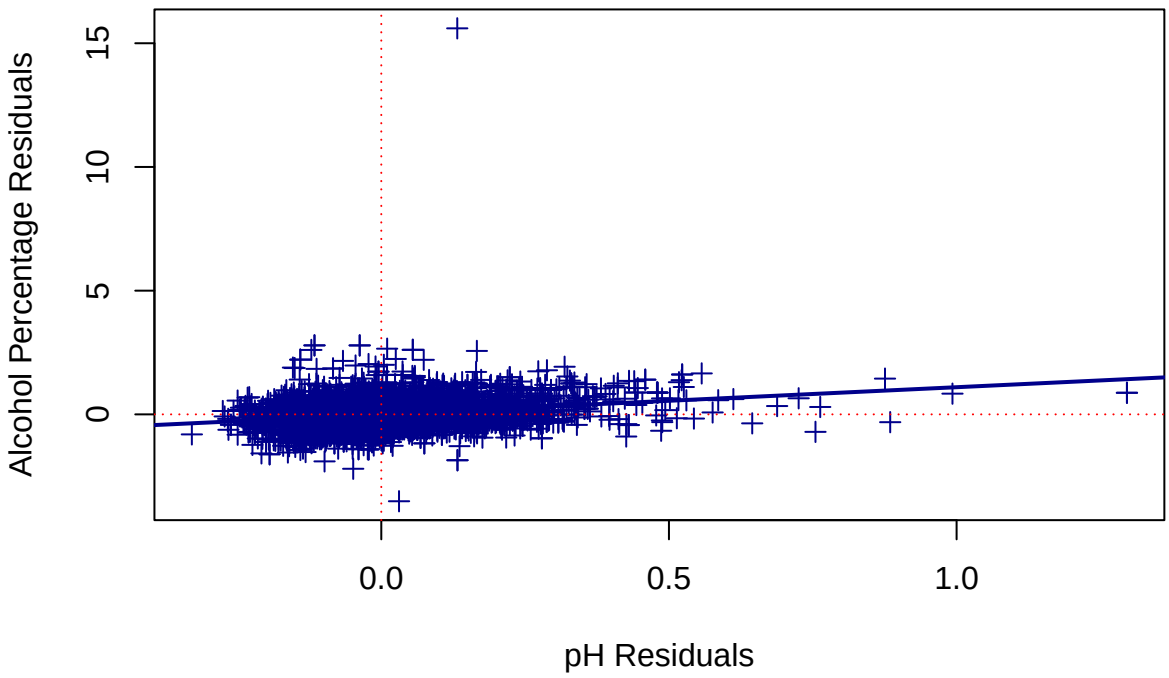

Alcohol Percentage Residuals
15
10
5
0
−5
0
5
10
Residual Sugars Residuals


Alcohol Percentage Residuals
15
10
5
0
−0.1
0.0
0.1
0.2
0.3
0.4
Chlorides Residuals


Alcohol Percentage Residuals
15
10
5
0
−100
−50
0
50
100
150
200
250
Free Sulphur Dioxide Residuals


Alcohol Percentage Residuals
6
4
2
0
−2
−0.005
0.000
0.005
0.010
Density Residuals


Alcohol Percentage Residuals
15
10
5
0
−0.6
−0.4
−0.2
0.0
0.2
0.4
pH Residuals



From the linearity plot we see that the points appear to be equally scattered around the line.

Looking at correlation of predictors

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                  1.00             0.22        0.33          -0.11
## volatile.acidity               0.22             1.00       -0.37          -0.19
## citric.acid                    0.33            -0.37        1.00           0.14
## residual.sugar                -0.11            -0.19        0.14           1.00
## chlorides                      0.30             0.37        0.05          -0.13
## free.sulfur.dioxide           -0.28            -0.35        0.13           0.40
## density                        0.46             0.28        0.10           0.55
## pH                            -0.26             0.26       -0.33          -0.27
## sulphates                      0.30             0.23        0.06          -0.18
##                       chlorides free.sulfur.dioxide density    pH sulphates
## fixed.acidity              0.30               -0.28    0.46 -0.26      0.30
## volatile.acidity           0.37               -0.35    0.28  0.26      0.23
## citric.acid                0.05                0.13    0.10 -0.33      0.06
## residual.sugar            -0.13                0.40    0.55 -0.27     -0.18
## chlorides                  1.00               -0.19    0.36  0.03      0.41
## free.sulfur.dioxide       -0.19                1.00    0.02 -0.15     -0.19
## density                    0.36                0.02    1.00  0.01      0.27
## pH                         0.03               -0.15    0.01  1.00      0.19
## sulphates                  0.41               -0.19    0.27  0.19      1.00
```

We do not see any strong correlations between the predictors selected.

In summary, we note that the model selected for prediction has issues of unusual observations and departures from the assumptions of constant variance and normality required for linear models.

Model B - model selection

```
##
## Call:
## lm(formula = alcohol ~ ., data = Wine_Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5559 -0.2892 -0.0361  0.2549 15.6752
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           6.745e+02  4.874e+00  138.392  < 2e-16 ***
## fixed.acidity         5.432e-01  8.474e-03   64.109  < 2e-16 ***
## volatile.acidity      6.502e-01  5.531e-02   11.756  < 2e-16 ***
## citric.acid           5.320e-01  5.437e-02    9.784  < 2e-16 ***
## residual.sugar        2.404e-01  2.776e-03   86.606  < 2e-16 ***
## chlorides            -1.013e+00  2.294e-01   -4.415 1.03e-05 ***
## free.sulfur.dioxide  -2.954e-03  5.252e-04   -5.625 1.93e-08 ***
## total.sulfur.dioxide -2.499e-04  2.224e-04   -1.124    0.261
## density              -6.827e+02  5.014e+00 -136.159  < 2e-16 ***
## pH                    2.721e+00  5.226e-02   52.058  < 2e-16 ***
## sulphates             1.095e+00  5.059e-02   21.645  < 2e-16 ***
## type                  1.210e+00  3.598e-02   33.645  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5037 on 6485 degrees of freedom
## Multiple R-squared:  0.822,  Adjusted R-squared:  0.8217
## F-statistic:  2722 on 11 and 6485 DF,  p-value: < 2.2e-16
```

We see that the p-value of the full model is less than 0.05 which means that atleast one of the predictors is statistically significant.

We now apply a backward elimination approach starting with the predictors with the highest p-value.

We perform partial F-tests for dropping each predictor. We use the following generic form of the hypothesis test for all fine tuning.

Hypotheses:

H0: $\beta_i = \beta_j = ... = \beta_k = 0$

Hα: Atleast one of β_i, β_j, ... β_k is not equal to zero

where i, j,..., k are the predictors being dropped together. One or more predictors can be dropped together from the full model.

Decision rule: If the p-value is less than 0.05 then reject the null hypothesis.

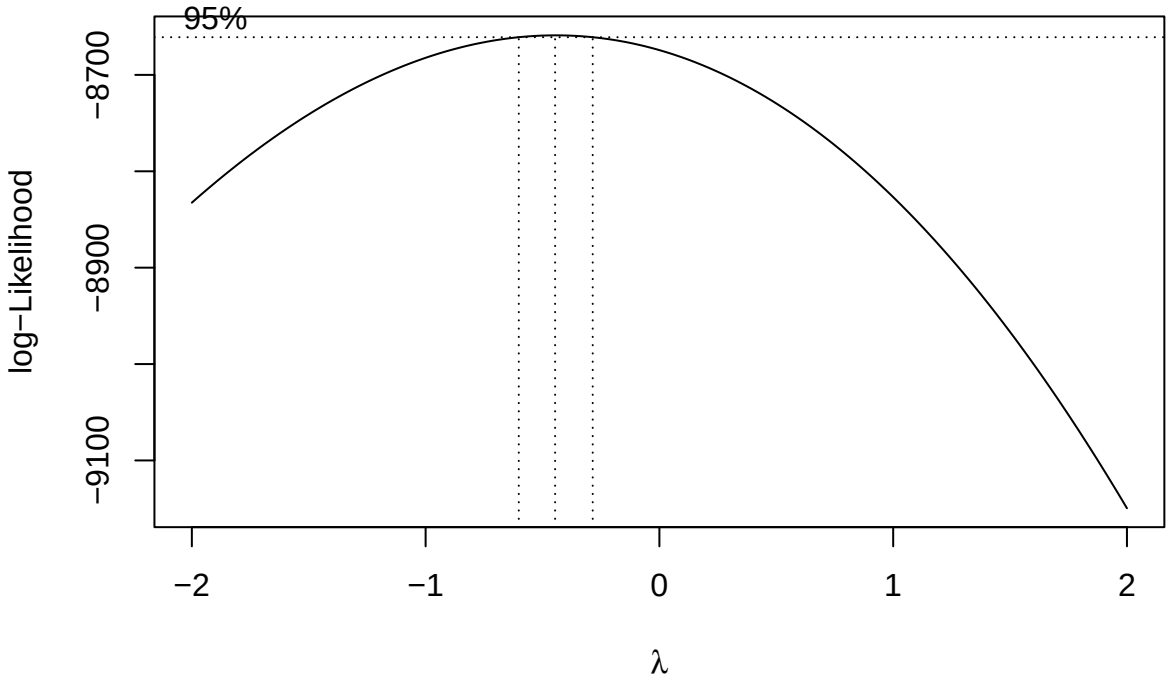
```
## Analysis of Variance Table
##
## Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + density + pH + sulphates +
```

```
##     type
## Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1   6486 1645.5
## 2   6485 1645.2  1   0.32052 1.2634 0.2611
```

Since p-value is > 0.05, we fail to reject the null and say that the reduced model is adequate. Let us now examine if we can drop the next predictor with high p-value i.e chlorides

```
## Analysis of Variance Table
##
## Model 1: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     free.sulfur.dioxide + density + pH + sulphates + type
## Model 2: alcohol ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     density + pH + sulphates + type
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   6487 1650.4
## 2   6485 1645.2  2    5.1924 10.234 3.653e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value is < 0.05, we can not accept the reduced model since we reject the null. We do not need to perform the model diagnostics again as the selected model A and B are the same. Now we perform box-cox as a remedial measure to fix normality deviations.



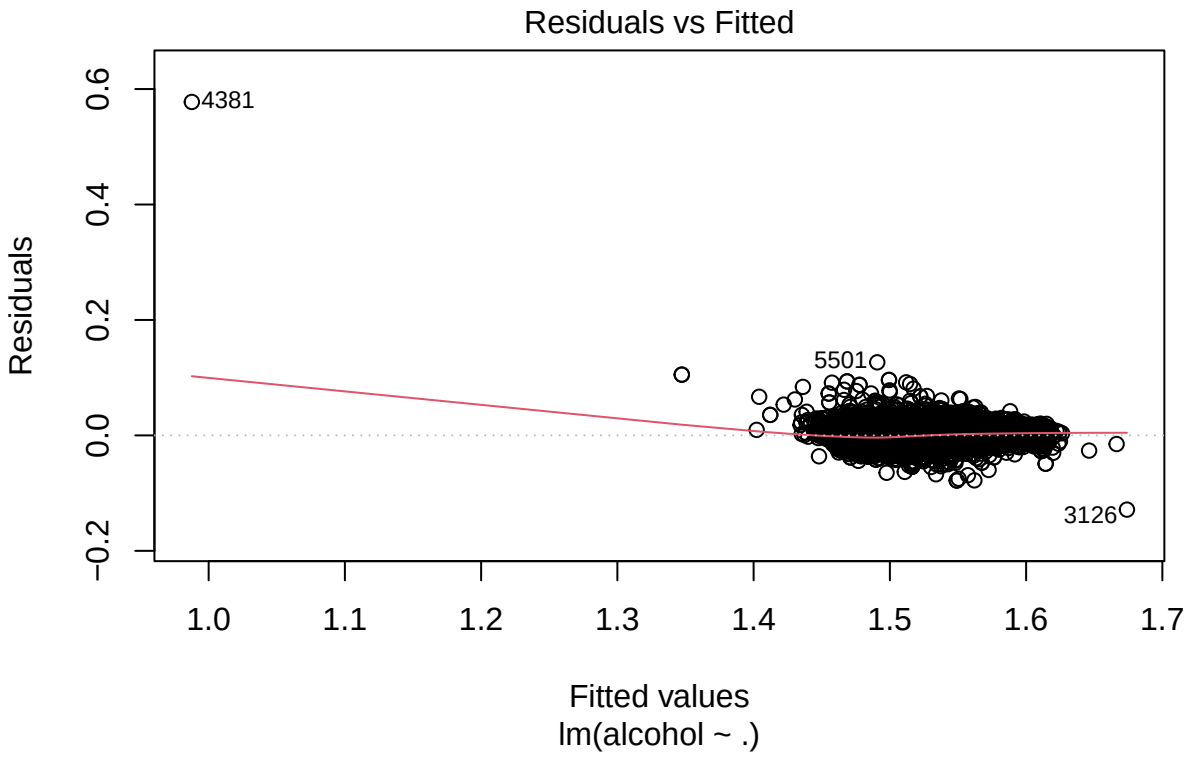
```
## [1] -0.4461153
```

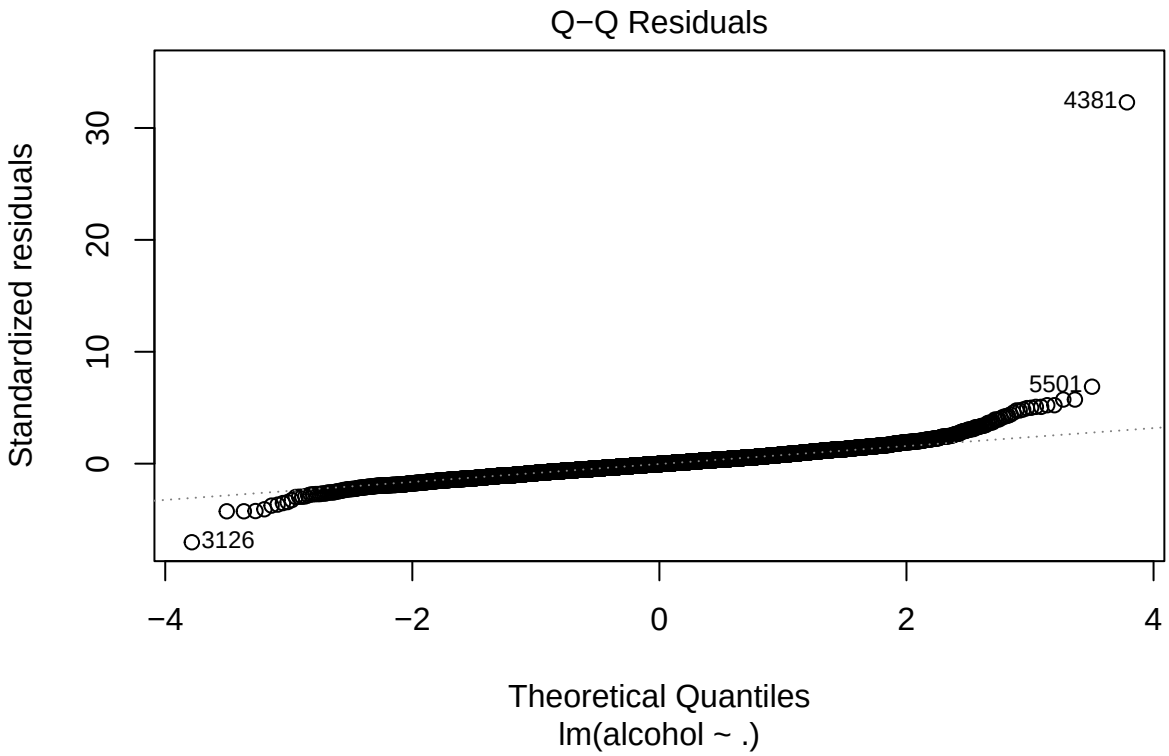
Since 1 is not in the interval, we proceed to perform transformations. From the box-cox results, we see that the optimal lambda -0.446, we round the value and choose -0.4. Notably, 1 and 0 are not in the confidence interval.

Fitting a new model with box-cox transformed "alcohol". We use the box-cox formula to transform our model: $g(Y) = \frac{Y^{\lambda}-1}{\lambda}$ where Y is "alcohol" and λ = -0.4

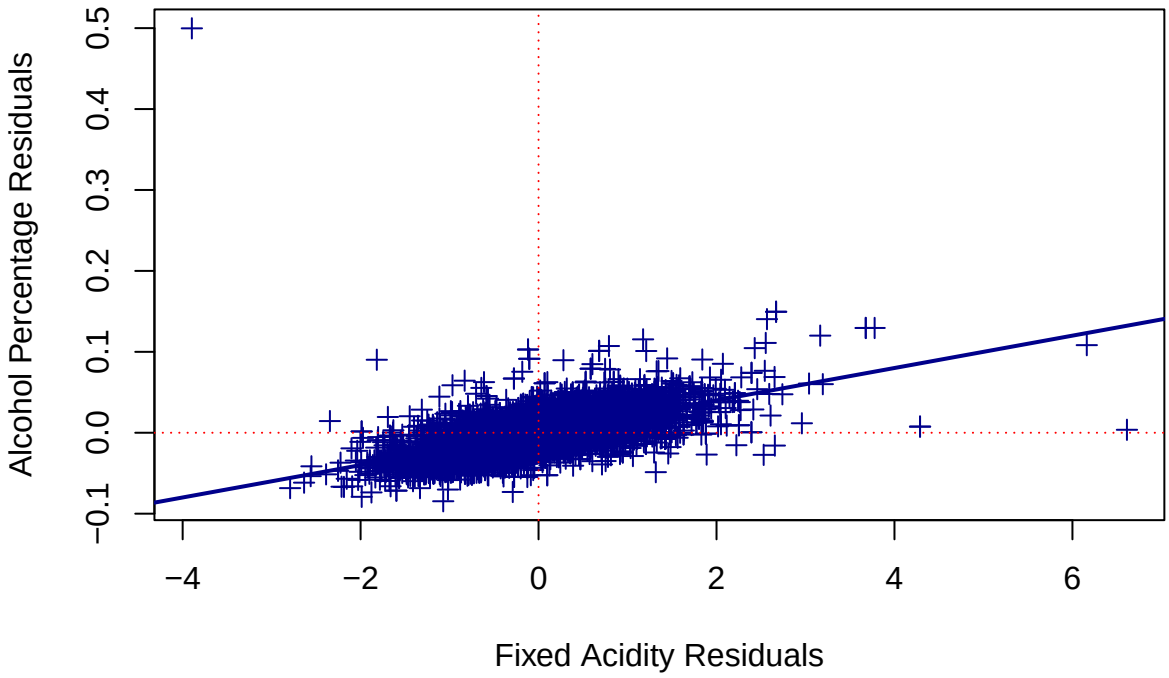
```
##
## Call:
## lm(formula = alcohol ~ ., data = Wine_Data.red1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12859 -0.01052 -0.00056  0.00942  0.57765
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)          2.529e+01  1.688e-01  149.829  < 2e-16 ***
## fixed.acidity        2.000e-02  3.074e-04   65.066  < 2e-16 ***
## volatile.acidity     1.940e-02  1.995e-03    9.726  < 2e-16 ***
## citric.acid          1.616e-02  1.981e-03    8.155 4.15e-16 ***
## residual.sugar       8.400e-03  1.007e-04   83.428  < 2e-16 ***
## chlorides           -4.151e-02  8.392e-03   -4.946 7.76e-07 ***
## free.sulfur.dioxide -1.240e-04  1.554e-05   -7.982 1.69e-15 ***
## density             -2.445e+01  1.737e-01 -140.774  < 2e-16 ***
## pH                   1.006e-01  1.910e-03   52.652  < 2e-16 ***
## sulphates            3.957e-02  1.847e-03   21.426  < 2e-16 ***
## type                 4.356e-02  1.079e-03   40.370  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01843 on 6486 degrees of freedom
## Multiple R-squared:  0.8187, Adjusted R-squared:  0.8184
## F-statistic:  2928 on 10 and 6486 DF,  p-value: < 2.2e-16
```

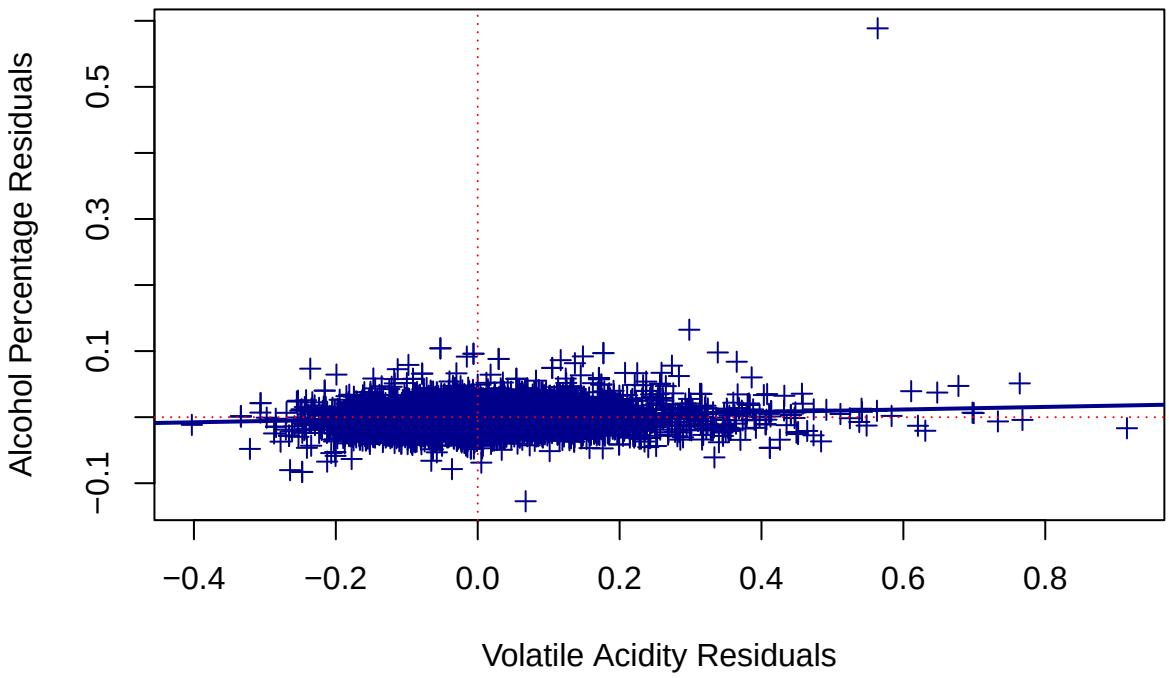
Now we check again for normality and constant variance assumptions again for the new box-cox based model.

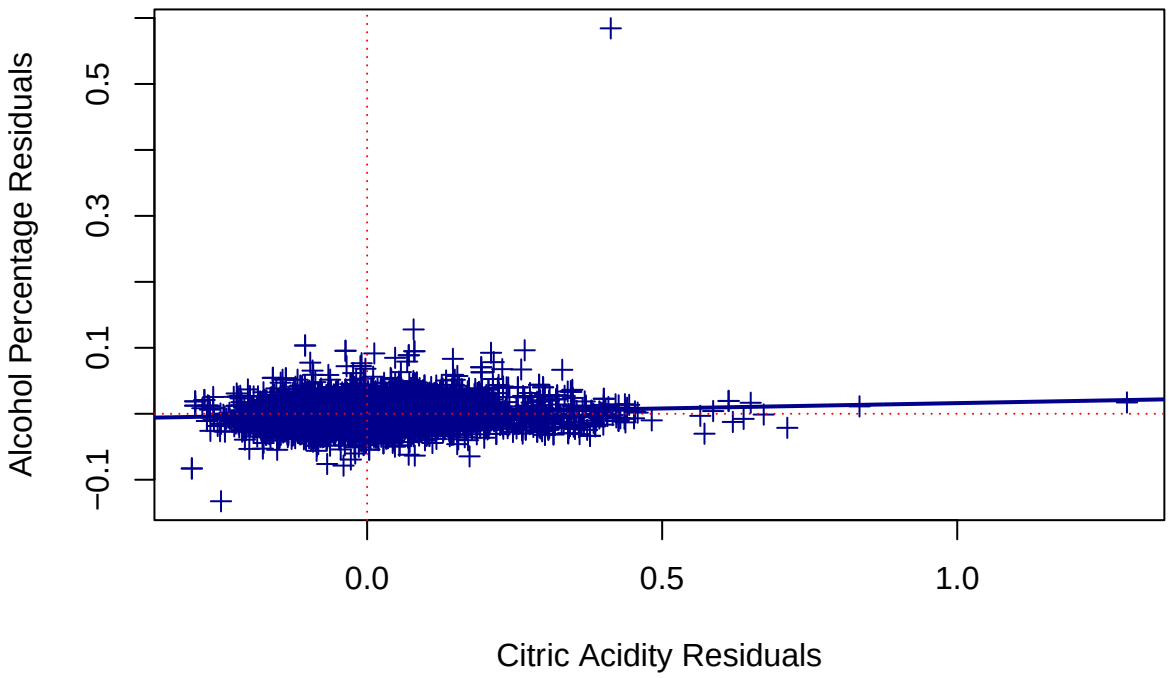

Residuals vs Fitted
Residuals
Fitted values
lm(alcohol ~ .)
4381
5501
3126
0.6
0.4
0.2
0.0
−0.2
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7

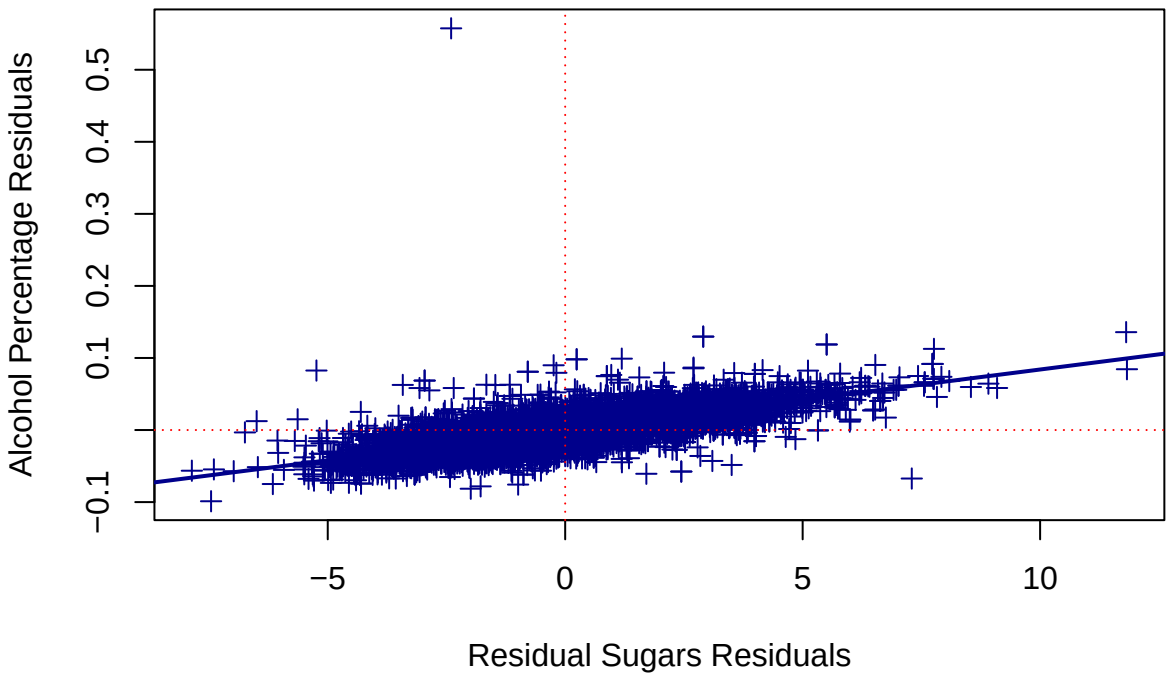


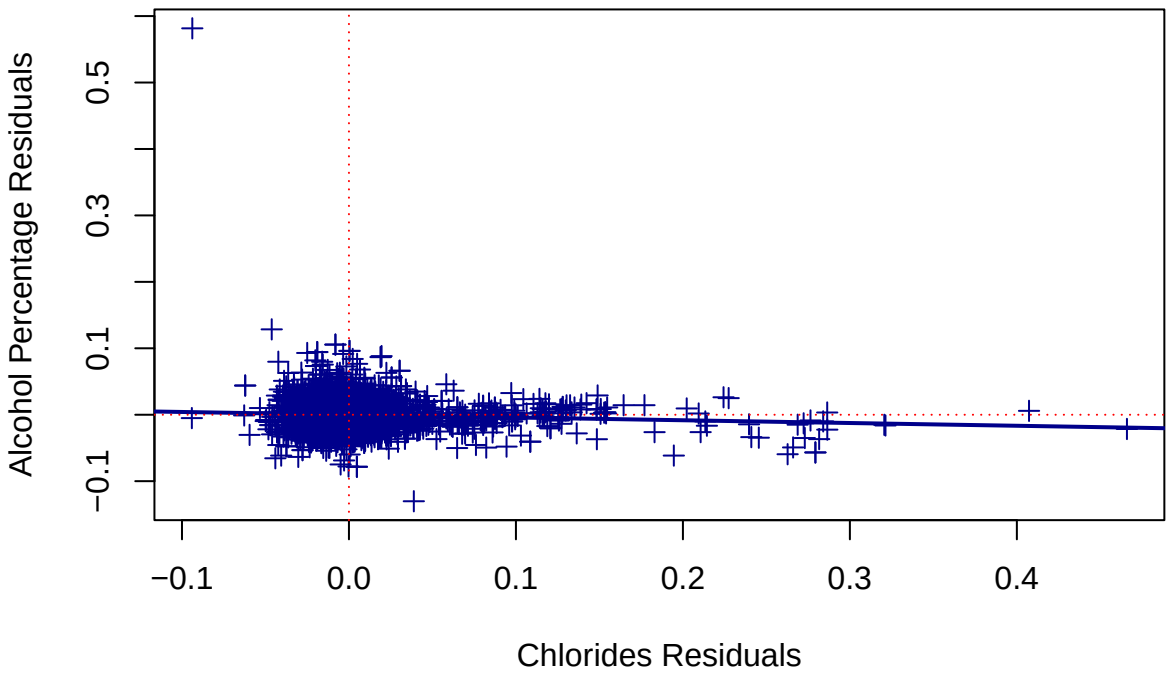
The Q-Q plot of the transformed model still appears to be similar to the previous Q-Q plot of the un-transformed model. The variance plot also does not seem to have changed.

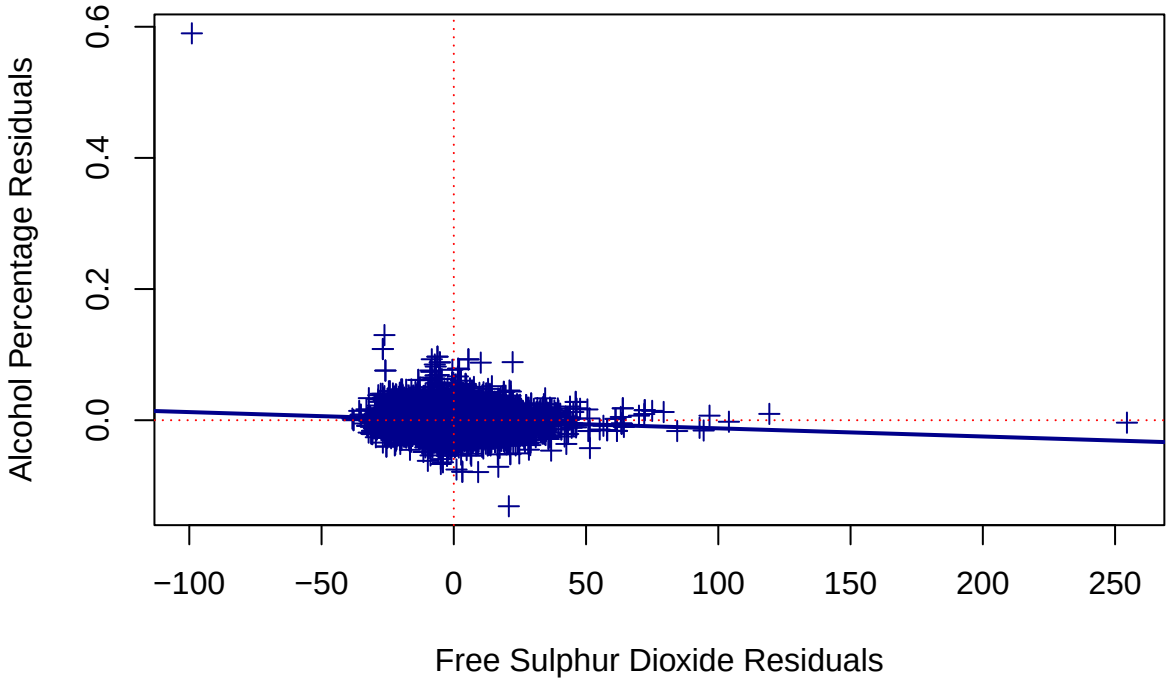

Alcohol Percentage Residuals
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
−4
−2
0
2
4
6
Fixed Acidity Residuals

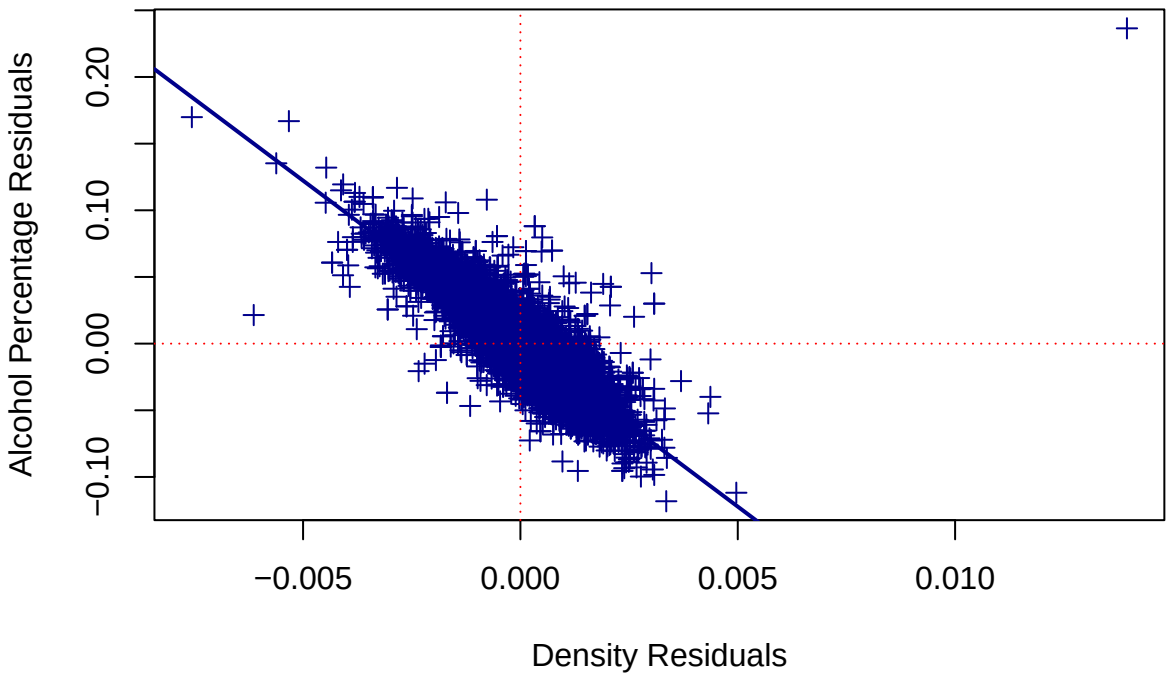

Alcohol Percentage Residuals
−0.1
0.1
0.3
0.5
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Volatile Acidity Residuals

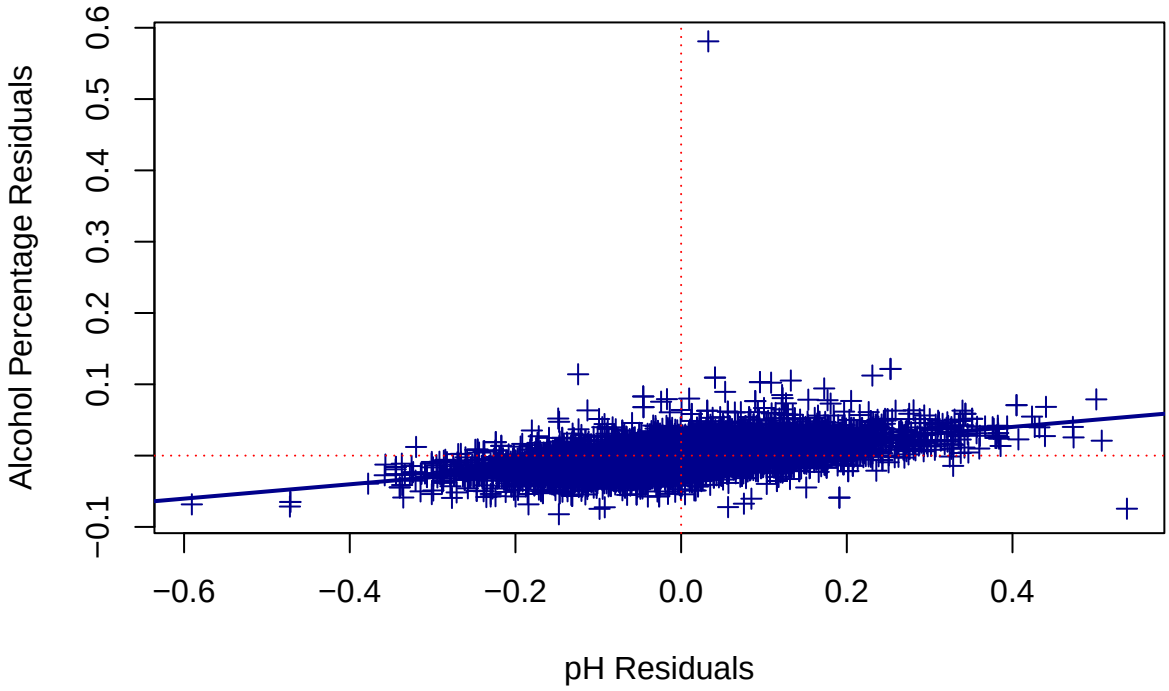

Alcohol Percentage Residuals
−0.1
0.1
0.3
0.5
0.0
0.5
1.0
Citric Acidity Residuals

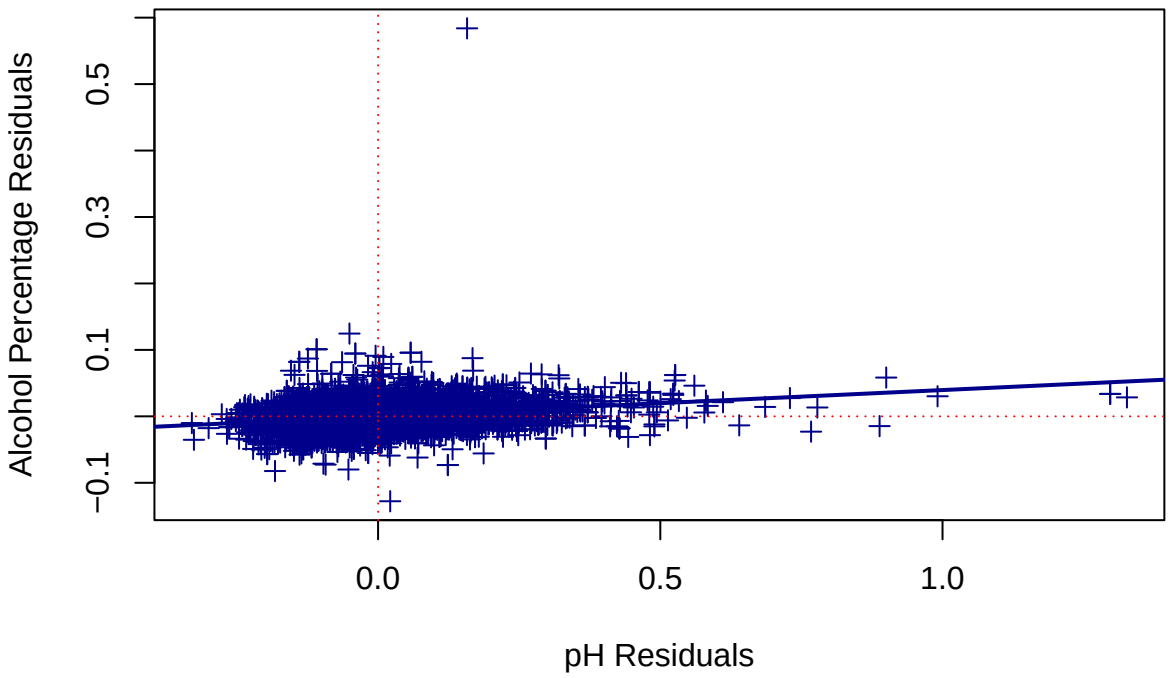

Alcohol Percentage Residuals
−0.1
0.1
0.2
0.3
0.4
0.5
−5
0
5
10
Residual Sugars Residuals


Alcohol Percentage Residuals
0.5
0.3
0.1
−0.1
−0.1
0.0
0.1
0.2
0.3
0.4
Chlorides Residuals


Alcohol Percentage Residuals
0.0
0.2
0.4
0.6
−100
−50
0
50
100
150
200
250
Free Sulphur Dioxide Residuals


Alcohol Percentage Residuals
0.20
0.10
0.00
−0.10
−0.005
0.000
0.005
0.010
Density Residuals


Alcohol Percentage Residuals
−0.1
0.1
0.2
0.3
0.4
0.5
0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
pH Residuals



The linearity plots also seems to be similar to what was observed before, the points still seem to be scattered equally around the line. The transformation did not remedy departure from normality and constant variance.

Conclusion

The best model obtained from prediction and selection does not include total sulfur dioxide content. We selected the best model for prediction based on the lowest RMSE and we picked the best model for selection based on backward elimination method using partial F-tests. We observed departure from normality and constant variance in model. These departures did not get fixed upon doing a box-cox transformation. We would be more concerned about these departures if we wanted to develop confidence-prediction intervals around our point estimates. If we only cared about getting an accurate point estimate, which is generally the case for prediction models, these departures may not be that concerning.