Aalto University
School of Science

MS-E2112
Multivariate Statistical Analysis
PROJECT

ROSMY SEBASTIAN (660424)
4 APRIL 2019

Wine dataset

Identify the wine based on cultivar

Apr 3, 2019

This data set was downloaded from *UCI ML datasets*[1]. This data have results of the various components or attributes in the wine that are obtained by chemical analysis. All the wines analysed are produced in Italy but are from three different cultivators which is denoted at levels 1,2 and 3 in the first column class. Thirteen attributes are determined for each of the three types of wines.Both red wine and white wine are included in the analysis. The 178 wine samples were analysed in this dataset.

The various attributes are:

1) Alcohol
2) Malic acid
3) Ash
4) Alcalinity of ash
5) Magnesium
6) Total phenols
7) Flavanoids
8) Nonflavanoid phenols
9) Proanthocyanins
10)Color intensity
11)Hue
12)OD280/OD315 of diluted wines
13)Proline

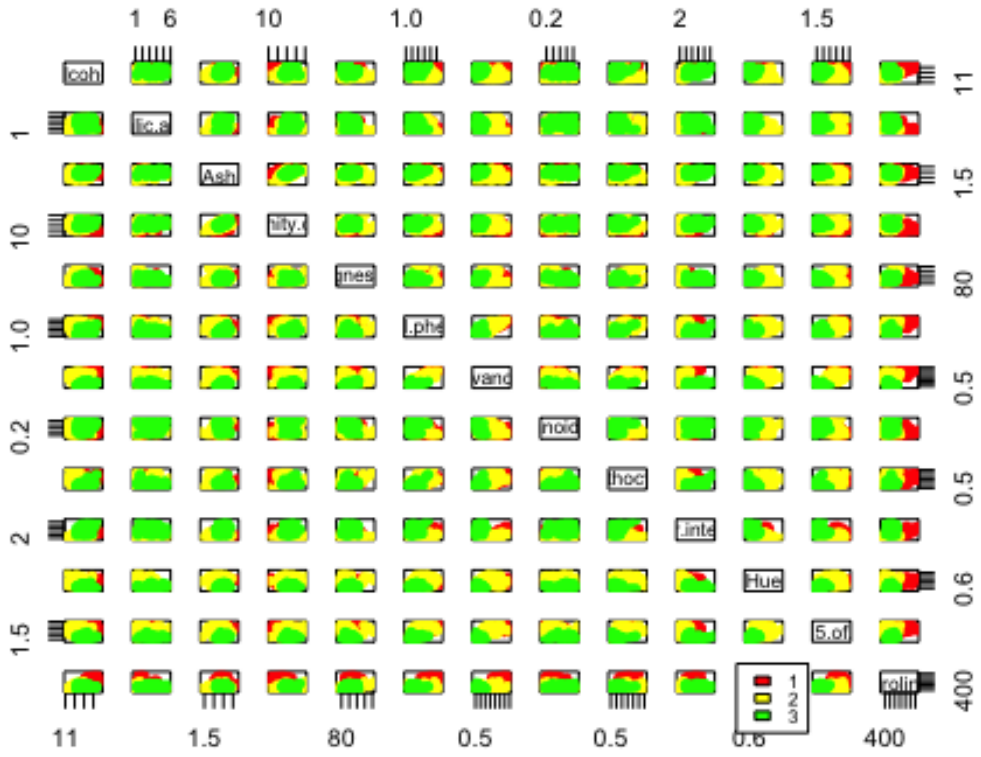
Description of the research question

The quality of wine is said to be affected based on the soil, climate and the region in which the wineyard is located. Hence the research question is:

1) Can the attributes be used to determine the area in which the wine was cultivated.
2) Can alcohol be used to determine the characteristic of the particular region.

Description of the dataset

The datapoints in the data was analysed for possible outliers and anomalies among the datapoints. Six datapoints were found to be influential based on cook distance, of which the datapoint *122* was found to have very high value for many attributes, when compared to the mean value for these attributes for the entire dataset (figure in Appendix. The data point 122 was hence removed from the dataset to be analysed.



The correlation between the class and various attributes were studied. Among the attributes Flavinoids had the highest correlation with class (-0.87), followed by OD280.OD315.of.diluted.wines (-0.79) and Total.phenols (-0.72).Correlation between Flavinoids and OD280.OD315.of.diluted.wines (-0.79) and Total.phenols is also evident from the plot. Clear separation between the cultivars in two dimension are evident in some of the cases, for e.g., alcohol and OD280.OD315.of.diluted.wines, alcohol and Flavinoids etc.

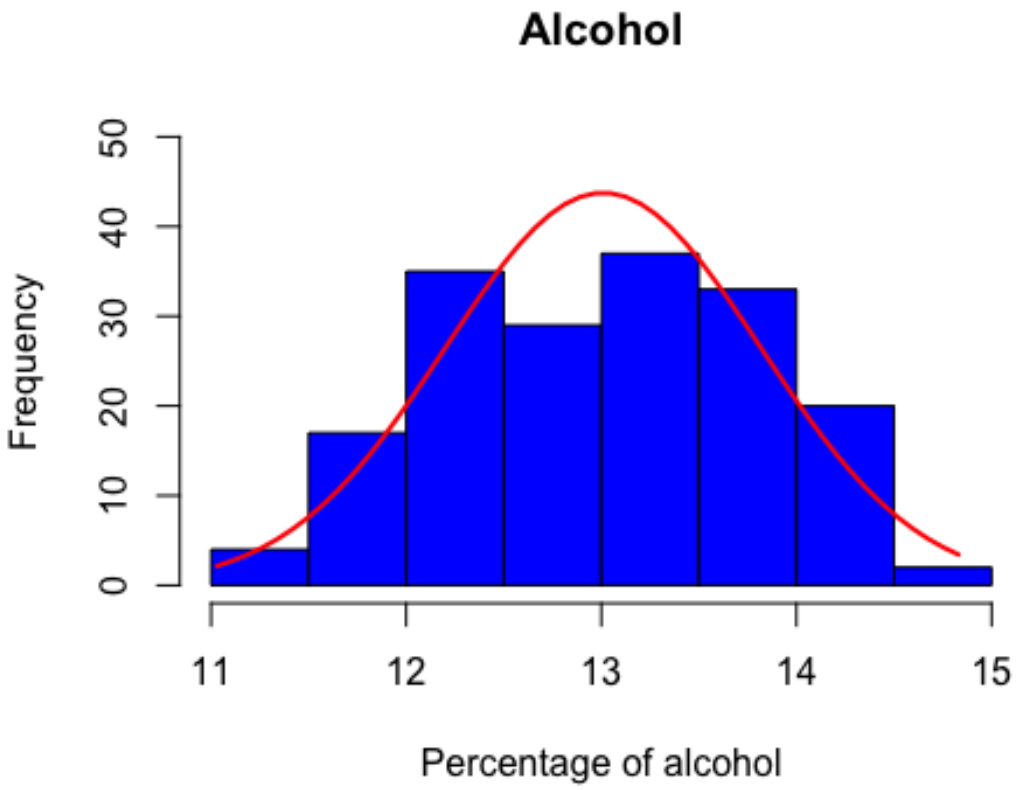
Among the attributes studied it was also seen that the corelation between Flavanoids and Total.phenols was very high (0.86). Similarly, Flavanoids and OD280.OD315.of.diluted.wines was also highly correlated (0.79). So the multicollinearity for the attributes were studied.

From the VIF values if can be seen that none of the attributes have multicollinearity. Hence all the attributes can be used for further analysis.

Univariate analysis

Inorder to determine the relation between percentage of alcohol and the three cultivars, univariate analysis of percentage of alcohol was done.

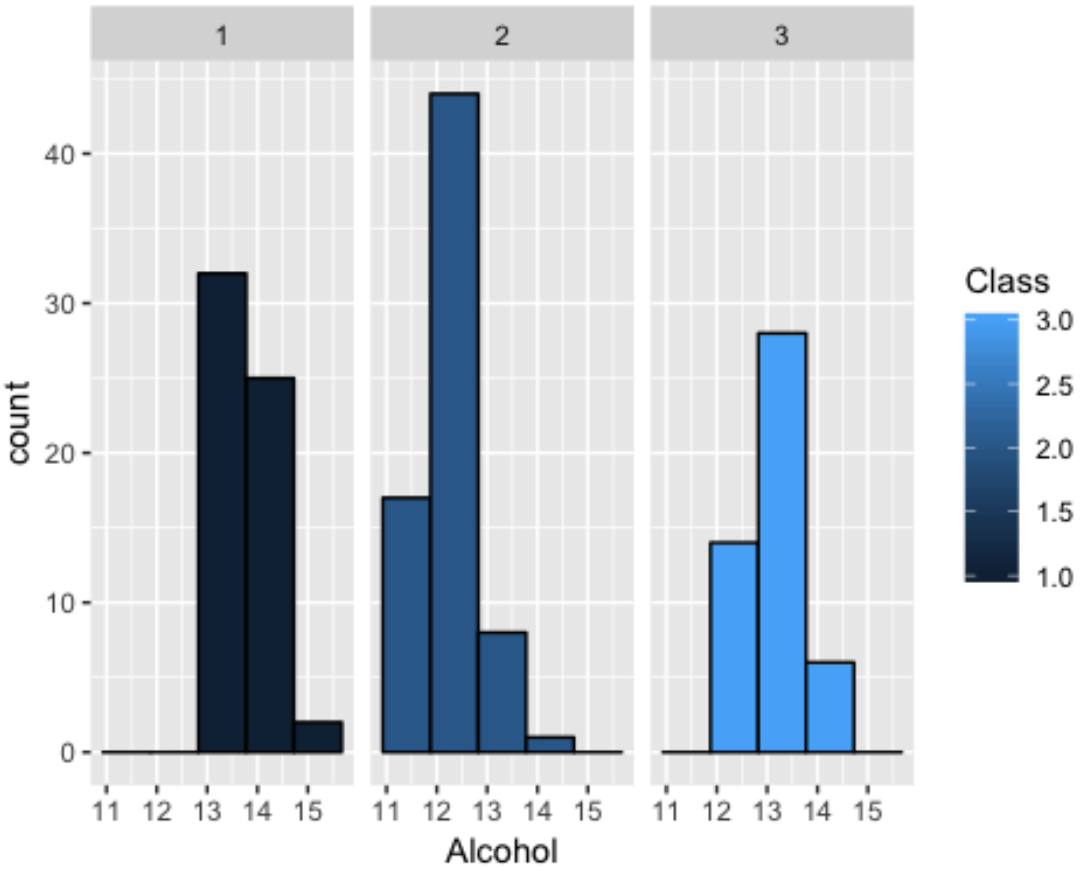
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.03   12.37   13.05   13.01   13.68   14.83
```



Above plot shows the distribution of percentage of alcohol among the entire datapoints. Summary of the distribution of percentage of alcohol was evaluated. Linear regression of class and percentage of alcohol was carried out. The summary statistics of the linear regression is given below. Summary of the linear regression showed significant relation between class and percentage of alcohol content. Percentage of alcohol explained 10 percentage of the class distinction among the data sample.

```
##
## Call:
## lm(formula = Class ~ Alcohol, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9884 -0.5905 -0.2048  0.8557  1.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.07820    0.89581   6.785 1.72e-10 ***
## Alcohol     -0.31827    0.06873  -4.631 7.08e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7357 on 175 degrees of freedom
```

```
## Multiple R-squared:  0.1092, Adjusted R-squared:  0.1041
## F-statistic: 21.44 on 1 and 175 DF,  p-value: 7.082e-06
```

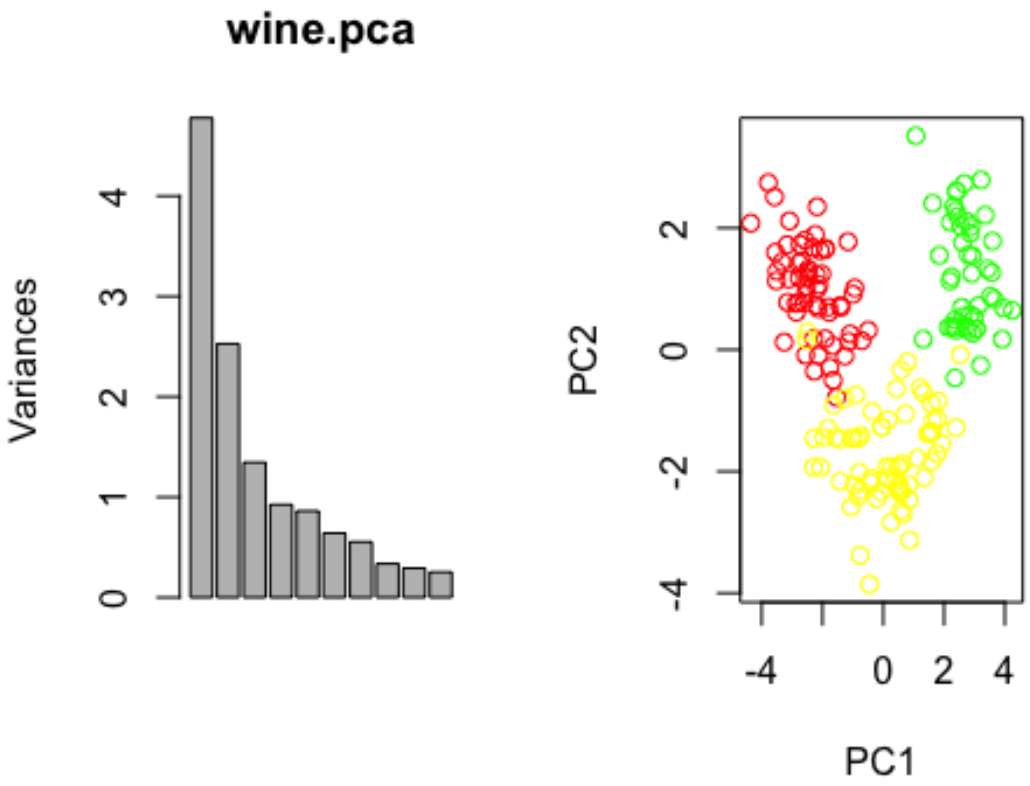


The distribution of the percentage of alcohol among the three classes of wines were plotted. It can be clearly seen from the plot that, for class 1, the percentage of alcohol ranges between 13 to ~16, with majority falling in the range 13 to 15. For class 2, the alcohol percentage range between 11 to 14. However, the majority of the wine has an alcohol percentage of 12-13. For class 3, on the other hand the alcohol percentage is distributed between 12 to 15. It can be seen from the above evaluations that the three cases have distinct distribution of percentage of acohol, however the distributions overlap each other. Hence, the attribute percentage of alcohol alone cannot be used to identify the cultivar from which the wine was obtained.

Multivariate analysis

A linear regression analysis of all the 13 attributes studied in this dataset, showed that 90 percentage of the characteristics in the class can be explained by the 13 independent attributes. However, it is tedious to understand the relations and interaction between these 13 attributes. Hence PCA can be used to reduce the dimensionality of the data and further investigate if the first research question can be answered. That is, can the attributes studied can segregate or identify the culivation from which the wine was produced.

PCA analysis of the dataset showed that the first two principle axis explains 56% of the variation in the data. On the other hand first eight principle components explain 92% of the data. Plotting the first two principle component shows that this components can clearly distinguish wines from the three cultivars.



Conclusion

In this analysis a wine dataset contaning 178 samples of wine from three different cultivars were analysed. Thirteen attributes of wine obtained by chemical analysis w as studied to answer two research questions.

Research question: Can alcohol be used to determine the characteristic of the particular region.

Although distribution of the percentage of alcohol is within a distinct range for each of the class of wine, it is not unique for the three class of wines. Hence, from the analysis it can be concluded that although percentage of alcohol can explain the distinction between the three classes of wine up to some level, it on itself cannot be used to distinguish the three classes of wine.

Research question: Can the thirteen attributes be used to determine the area in which the wine was cultivated.

In this exercise we have used PCA analysis to transform the thirteen attributes studied in this data-set into set of principle components. The first eight principle component explains 92% of the variance in the data. Thus in this case, PCA can be used to reduce the dimensionality of the datasets from thirteen to eight. Plotting the first two principle component clearly shows that, these two components are enough to distinguish the wine based on the cultivars. Hence, we can conclude that the attributes studied in this dataset clearly defines the distinct characteristics of wine from the cultivars studied in this dataset.

Critical evalution

Although the attributes defined in the dataset, clearly distinguish the cultivars from which the wine originated, the analysis should be further analysed to ascertain its validity. This is because some of the important attributes of the wine is not given in the dataset. For example, although the description says that the dataset contains both white and red wines, it is not given if the same cultivar produced both white and red wine. If the cultivar produced only one of two styles (i.e. white or red) of wine then these attributes might be determining the style of the wine rather than the distinct characteristics of the cultivars.

Reference

[1] https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.names

Appendix

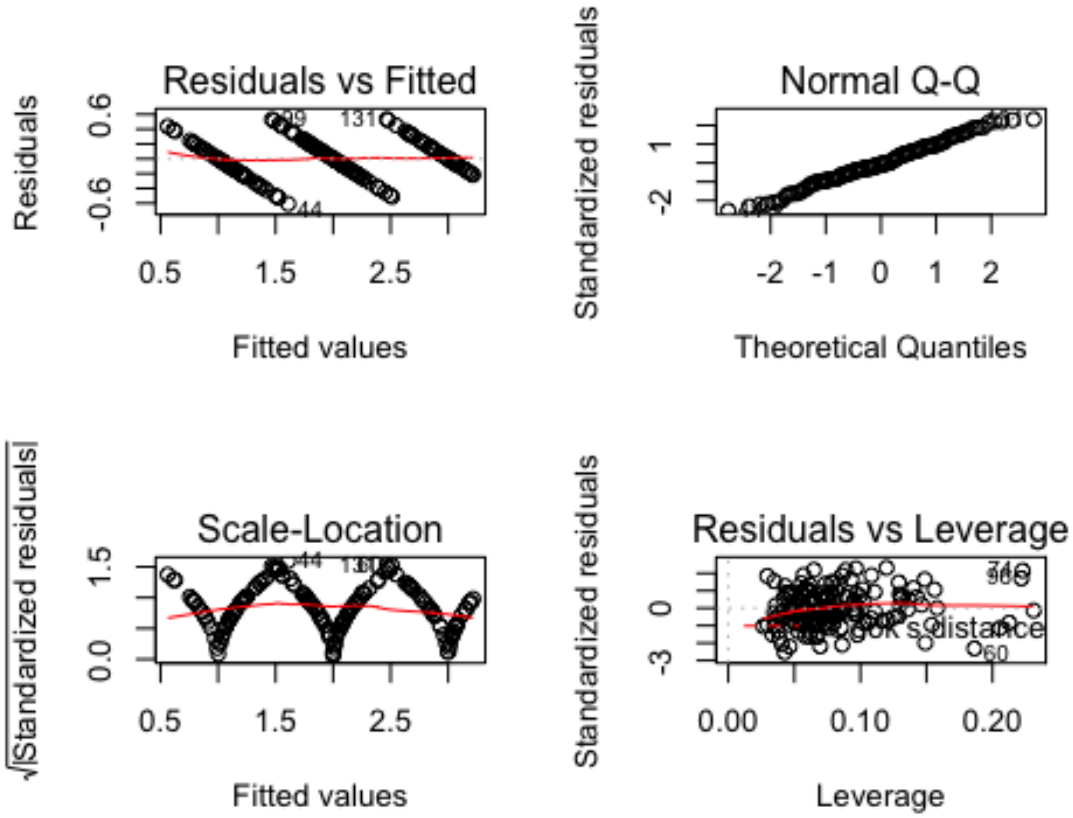
summary of the multivariate analysis

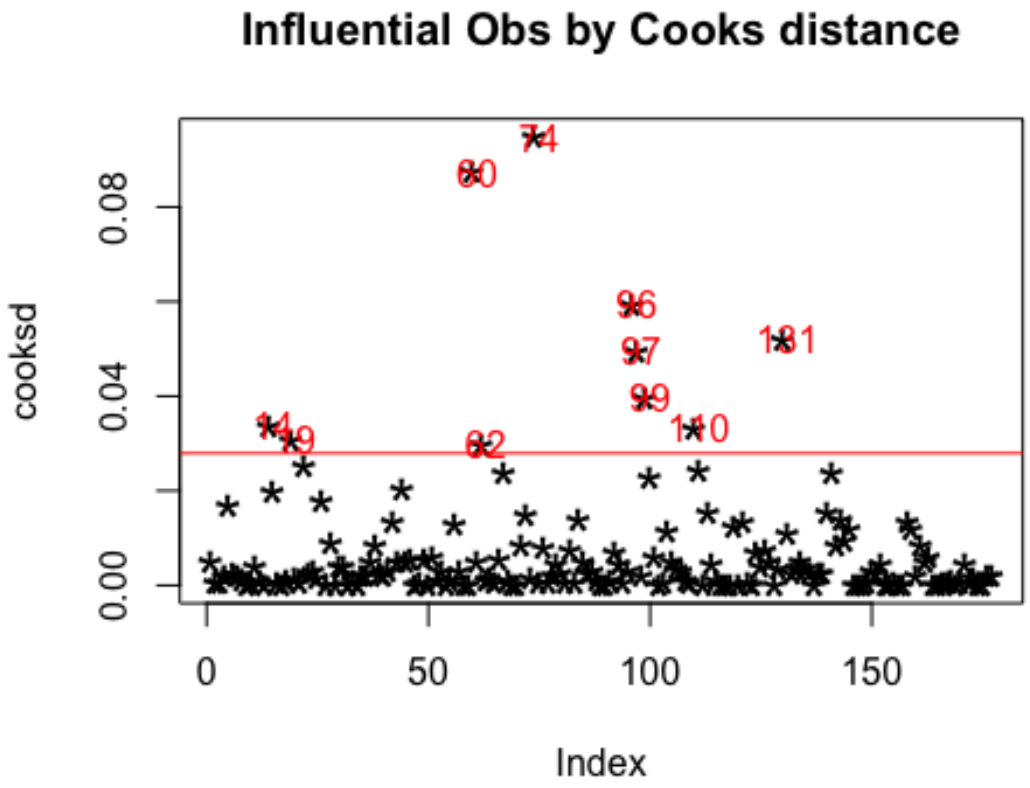
```
## Loading required package: carData

##
## Call:
## lm(formula = Class ~ ., data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61005 -0.15336 -0.01563  0.15922  0.53301
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.358e+00  4.810e-01   9.060 3.90e-16 ***
## Alcohol                      -8.858e-02  3.649e-02  -2.428  0.01628 *
## Malic.acid                    2.898e-02  2.126e-02   1.363  0.17476
## Ash                          -1.488e-01  9.942e-02  -1.496  0.13649
## Alcalinity.of.ash             3.595e-02  8.335e-03   4.313 2.78e-05 ***
## Magnesium                    -1.612e-03  1.569e-03  -1.027  0.30573
## Total.phenols                 1.899e-01  6.263e-02   3.032  0.00283 **
## Flavanoids                   -4.597e-01  5.451e-02  -8.434 1.70e-14 ***
```

```
## Nonflavanoid.phenols         -4.661e-01  2.036e-01  -2.289  0.02338 *
## Proanthocyanins               7.870e-02  4.657e-02   1.690  0.09299 .
## Color.intensity               6.570e-02  1.411e-02   4.656 6.64e-06 ***
## Hue                          -1.001e-01  1.296e-01  -0.772  0.44135
## OD280.OD315.of.diluted.wines -2.809e-01  5.065e-02  -5.546 1.16e-07 ***
## Proline                      -6.592e-04  9.912e-05  -6.651 4.21e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2455 on 163 degrees of freedom
## Multiple R-squared:  0.9076, Adjusted R-squared:  0.9003
## F-statistic: 123.2 on 13 and 163 DF,  p-value: < 2.2e-16
```

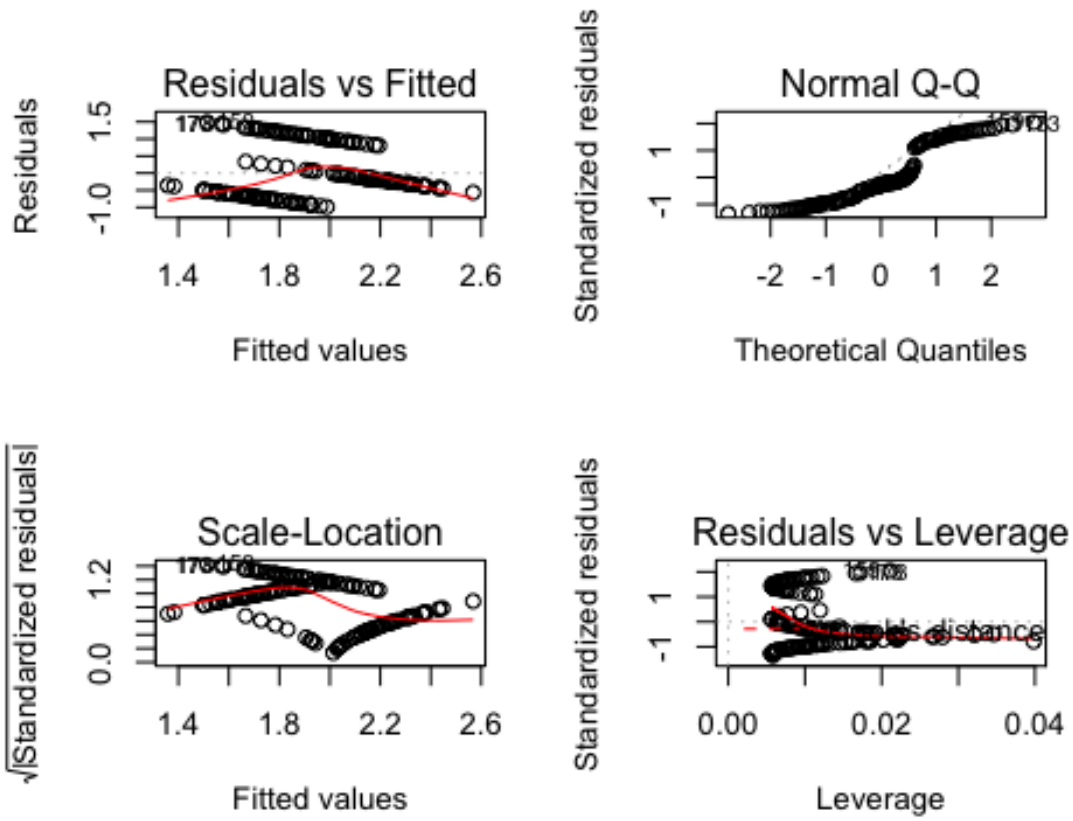
Plot of the multivariate analysis and cook's distance.


Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage




```
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 44 -2.583933           0.010651           NA
```

Plot of the univariate analysis class ~Alcohol.



Summary statistics of the PCA analysis.

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.1861 1.5899 1.1601 0.96171 0.92826 0.80043
## Proportion of Variance 0.3676 0.1945 0.1035 0.07114 0.06628 0.04928
## Cumulative Proportion  0.3676 0.5621 0.6656 0.73675 0.80303 0.85232
##                           PC7     PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.74398 0.57942 0.53979 0.50244 0.48134 0.40738
## Proportion of Variance 0.04258 0.02583 0.02241 0.01942 0.01782 0.01277
## Cumulative Proportion  0.89489 0.92072 0.94313 0.96255 0.98037 0.99314
##                           PC13
## Standard deviation     0.29864
## Proportion of Variance 0.00686
## Cumulative Proportion  1.00000
```

Plot of principle components 1) PC3 versus PC4 and 2) PC5 and PC6.