Physicochemical determinants of subjective wine quality

Lloyd Kim, 207861834

Huiseong Yoo*, 215048747

1. Introduction

Alcoholic beverage consumption is a common point of social engagement in various situations, and people enjoy it worldwide (Ritchie & Roser, 2018). It has been demonstrated to, in moderation, have certain health benefits, including better cognitive functioning, lower risk of cardiovascular disease, and reduced mortality among middle-aged and older adults (Artero et al., 2015). Furthermore, wine has been singled out as potentially the most beneficial of all alcoholic drinks. Due to its relatively high contents of resveratrol and flavonoids, it is deemed as a healthy addition to a balanced diet, of course, when taken in moderation (Artero et al., 2015). It is even an integral part of one of the most prominent food regimens recommended for weight loss and general health – the Mediterranean diet (Mezzano et al., 2001).

Aside from having health benefits, wine is consumed due to its pleasant taste and smell. The quality of a wine can be determined by tasting, but also through physicochemical tests (Ebeler, 1999). The relationship between various physicochemical characteristics of a wine and its taste is not completely understood (Legin et al., 2003). However, the objective characteristics of a wine that contribute to its taste and quality are important, since objective measurements are a superior way to determine standards of regularization and certification, in comparison to needing to rely on subjective opinions of experts.

Thus, in this report, the data set created by Cortez et al. (2009) will be analyzed. These authors measured various physicochemical characteristics of many Portuguese wines and recorded the quality of these wines, as determined by experts who tasted it. The goal of this report will be to determine the most important physicochemical determinants of subjective wine quality.

2. Methodology

This data set, provided by Cortez et al. (2009) contains 12 variables, and is formatted as a .csv file. The values were separated by semicolons. The 12 variables all relate to various characteristics of the wines, and each of the rows (4898) relates to one sample of the wines. The variable quality will be used as the response variable, since it is in the focus of the study. It is a numerical variable, measured at the interval level. It was determined by at least three expert tasters, who each gave a grade for the taste ranging from 0 to 10. The median of the three grades was entered into the dataset. The remaining 11 variables are physicochemical characteristics of the sample, including the wine's fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol contents. All of them were continuous and measured at interval or ratio levels. In order to determine the best model, an all-possible regressions algorithm was used, through the R package leaps (Lumley, 2020). The leaps package utilizes an algorithm to efficiently determine the best model for each number of predictors. The function returns a set of predictors that is the best fitting for each number of predictors, along with their Mallow's C values. Each of the models was constructed, and their PRESS statistics and adjusted R^2 values were assessed as well (pages 1-4 of the appendix). The results of this analysis may be seen in Table 1.

Table 1

The optimal models with various numbers of predictors generated through the all-possible regression algorithm, along with their respective measures of model adequacy

N	Included predictors	Adj. R2	PRESS	Mallow's C
1	alcohol	0.1895	3114.842	618.9350
2	alcohol, volatile acidity	0.2399	2922.024	277.3040
3	alcohol, volatile acidity, residual sugar	0.2581	2852.789	154.8290
4	alcohol, volatile acidity, residual sugar, free sulfur dioxide	0.2634	2835.081	119.6254
5	alcohol, volatile acidity, residual sugar, density, pH	0.2703	2808.287	73.4614
6	alcohol, volatile acidity, residual sugar, density, pH, sulphates	0.2758	2788.672	37.4649
7	alcohol, volatile acidity, residual sugar, density, pH, sulphates, free sulfur dioxide	0.2791	**2778.756**	15.9119
8	alcohol, volatile acidity, residual sugar, density, pH, sulphates, free sulfur dioxide, fixed acidity	**0.2806**	*2782.164*	**6.8056**
9	alcohol, volatile acidity, residual sugar, density, pH, sulphates, free sulfur dioxide, fixed acidity, total sulfur dioxide	*0.2805*	2783.621	*8.2383*
10	alcohol, volatile acidity, residual sugar, density, pH, sulphates, free sulfur dioxide, fixed acidity, total sulfur dioxide, chlorides	0.2804	2784.448	10.0532
11	alcohol, volatile acidity, residual sugar, density, pH, sulphates, free sulfur dioxide, fixed acidity, total sulfur dioxide, chlorides, citric acid	0.2802	2785.419	12.0000

Note. Bolded values refer to the optimal values in regards to each criterion, italicized values refer to the second best (highest for Adj. R2, lowest for PRESS and Mallow's C)

As the table shows, the model with 8 predictors seems to be the most favoured one, as it has the highest value of adjusted R2, and the lowest values of Mallow's C It has the second lowest value of PRESS. While the model with 7 predictors has the lowest PRESS value, it has a much higher Mallow's C than the model with 8 predictors. However, the model with 9 predictors

may also be well-fitting, as it has very close values of all three statistics, in comparison to the model with 8 predictors. Thus, the models with 8 and 9 predictors will be assessed before determining the best one.

Model 8 may be expressed through a regression equation as follows:

$$Y_{Quality} = \beta_0 + \beta_{fixed\ acidity} * x_{fixed\ acidity} + \beta_{volatile\ acidity} * x_{volatile\ acidity} + \beta_{residual\ sugar} * x_{residual\ sugar} + \beta_{free\ sulfur\ dioxide} * x_{free\ sulfur\ dioxide} + \beta_{density} * x_{density} + \beta_{pH} * x_{pH} + \beta_{sulphates} * x_{sulphates} + \beta_{alcohol} * x_{alcohol} + \varepsilon$$

Model 9 may be expressed as follows:

$$Y_{Quality} = \beta_0 + \beta_{fixed\ acidity} * x_{fixed\ acidity} + \beta_{volatile\ acidity} * x_{volatile\ acidity} + \beta_{residual\ sugar} * x_{residual\ sugar} + \beta_{free\ sulfur\ dioxide} * x_{free\ sulfur\ dioxide} + \beta_{density} * x_{density} + \beta_{pH} * x_{pH} + \beta_{sulphates} * x_{sulphates} + \beta_{alcohol} * x_{alcohol} + \beta_{total\ sulfur\ dioxide} * x_{total\ sulfur\ dioxide} + \varepsilon$$

The goal of this study is to explore and determine how well both of these models fit to the data, how successful they are in explaining the response variable, and which of the predictor variables have the strongest and weakest impacts on it, as well as in what direction.

In order to investigate the models, it is necessary to first check if all of the multiple linear regression assumptions are met. To this end, first an analysis of outliers will be done, including the assessment of leverage and influence points. These will be conducted through assessing residual plots, as well as various statistics, such as the probability of residual values, h-hat values and PRESS residuals. The values that show anomalous values that point to a probable measurement error will be discarded. Furthermore, the entries that are deemed as leverage or influential points, but do not have anomalous values will be examined, but not removed from the model, as they may be informational.

Then, analyses of multicollinearity will be conducted, utilizing three methods. The correlations between pairs of predictor variables will be assessed first, in order to determine an existence of eventual extremely high bivariate correlations which may be a source of multicollinearity. Furthermore, variance inflation factors (VIF) of each predictor will be assessed, with any VIF values over 10 indicating multicollinearity issues. Lastly, eigensystem analysis will be conducted, and any factors with a condition index over 100 will be assessed in order to determine which variables contribute to it the most. Based on the cumulate findings from the three methods, it will be determined whether or not there is a need for additional editing of the variables present in the model.

The determination of the normal distribution and independence of residuals will be analyzed through assessing the normal probability plot of the residuals, as well as the plot of externally studentized residuals fitted against the predicted values of the model. It is expected that the probability plot will approximately follow a straight line, which will indicate normality. Furthermore, the points in the plot of externally studentized residuals against predicted values should be relatively equally dispersed around the null line, indicating independence of the residuals.

The same scatterplot (externally studentized residuals against fitted values) will be utilized to determine if there is a deviation from a linear function between the set of the predictors and the response variable. Again, a relatively straight-band form of the scatter points is expected. Furthermore, the plot will be used to determine the equality of error variance across values of predictors. The residuals should vary in a similar manner across all values of the fitted model. Lastly, to make sure that the mean of the error is (close to) zero, it will be calculated as well.

To test for outliers, the externally studentized residuals of both models were computed. Then, normal QQ plots of them were created, in order to visually assess the existence of outliers. As can be seen in Figures 1 and 2 in the appendix, there were a few values that were disconnected from the rest of the distribution. The studentized residuals were also assessed by fitting them against the predicted values of y. The plots that were constructed (Figure 3 and 4) for both models showed a very distant value in the upper left corner of the plot, which had a low fitted value and a high externally studentized residual.

A test of highest and lowest studentized residuals with the Bonferroni-adjusted p values showed that there are five values which may be outliers: values 4746, 2782, 3308, 254, and 446. The same entries were deemed as significantly outlying by this test for both the 8-predictors model and the 9-predictors model. In order to further investigate these values, h-hat and PRESS residuals values were examined as well. The h-hat values were first plotted by themselves, and, as can be seen in Figures 5 and 6 of the appendix, there was a single entry that had a very high h-hat value. Upon further inspection, it was determined that the entry in question was entry 2782, with a h-hat value of 0.33. Thus, this entry was definitely a very influential observation, which was already observed from the scatterplot in Figures 3 and 4. The assessment of PRESS residuals on a plot (Figures 7 and 8) also showed an entry with a clearly very high value of this statistic, along with several which may also be deviating from the majority of the points. As could be expected, the high point was entry 2782, with the following highest PRESS being the point 776. The lowest points were entries 4746 and 3308.

All of the mentioned entries were inspected in the datafile, in order to determine which should be removed. Points 2782, 3308, and 4746 had anomalous values (clearly much higher than all other entries) on some of the researched variables, and were thus discarded from the model before further analyses. The rest of the values were not removed. After the removal of these values, previously created plots were created again (Figures 9 to 14 in the appendix), and no visible outliers could now be detected. It should be noted that a few values still had slightly higher h hat values than the rest (up to .04), but this was seen as natural fluctuation in the influence of different data points and nothing was done to change it.

The next step in model checking was to check for issues with multicollinearity. The first way of checking this was to assess the correlations amongst the predictors. While there were some significant correlations, the highest of them was 0.79, which is not indicative for bivariate multicollinearity. However, further tests were conducted to check for other indicators for multivariate multicollinearity. The first of those was determining the variance inflation factors of the predictors. As can be seen in page 35 of the appendix, the highest VIF was 35.474 in the model with 8 predictors and 38.105 in the model with 9 predictors. In both models, this highest value was that of the density variable. These values are high and point towards a multicollinearity issue.

As the final check for multicollinearity, the eigenvalues and condition indices were calculated. As can be seen in pages 35-38 of the appendix, the last condition index was 8235.954 for the eight-predictor model and 9007.297 for the nine-predictor model, which is well above 100, the cut-off value in regards to which the existence of multicollinearity is judged in this technique. As the factor had the highest salience on the predictor density, it needed to be removed from the models.

After the removal of the variable, both models showed satisfactory multicollinearity diagnostic indicators: all VIF values were below 3 (pages 37-38 of the appendix), and the last conditioning index only slightly above 100 (105.021 for the smaller model, 111.699 for the larger one; pages 38-40 of the appendix).

After the removal of density from the models, their residuals were again assessed for normality. As can be seen in Figures 17 and 19, they still showed distributions close to normal, with slight deviations at the edges of the distributions. Furthermore, as can be seen in figures 18 and 20, both models showed satisfactory equality of variance of residuals across values of fitted values. Furthermore, most of the residuals fit within a relatively straight band above and below null. The means of the residuals of both models were very close to 0.

Therefore, all assumptions of the multiple linear regression were satisfied for both models. In order to attempt to compare the two models again (after the removal of density), they were compared on the PRESS and adjusted R^2 again (page 43 of the appendix). The model "8b", now the seven-predictor model, had a PRESS of 2781.477 and an adjusted R^2 of 0.274, while the model "9b", now the eight-predictor model, had a PRESS of 2778.672 and adjusted R^2 of 0.275. As the model 9b showed slightly better indicators of fit, the analysis of variance procedure was used to determine whether or not such a difference is statistically significant.

The ANOVA test (page 43 of the appendix) showed that the difference between the predictive power of the two models was significant ($F(1, 4884) = 7.1331$, $p = .0076$). Thus, the "9b", or the model with eight predictors, was chosen as the best fitting model for the data. This model can be expressed as follows:

$$\begin{aligned} Y_{Quality} = \beta_0 &+ \beta_{fixed\ acidity} * x_{fixed\ acidity} + \beta_{volatile\ acidity} * x_{volatile\ acidity} \\ &+ \beta_{residual\ sugar} * x_{residual\ sugar} + \beta_{free\ sulfur\ dioxide} * x_{free\ sulfur\ dioxide} \\ &+ \beta_{total\ sulfur\ dioxide} * x_{total\ sulfur\ dioxide} + \beta_{sulphates} * x_{sulphates} + \beta_{pH} \\ &* x_{pH} + \beta_{alcohol} * x_{alcohol} + \varepsilon \end{aligned}$$

In mathematical terms, the main question of interest is whether or not this model is able to predict the data, and to what extent. The null hypothesis is that all β coefficients equal 0. The alternative hypothesis is that at least one of the β coefficients is different from 0.

3. Results

The final model significantly predicted 27.52% of the response variable ($F(8, 4884) = 233.2$, $p < .001$). All the predictors were significant (page 44 of the appendix). Most of the predictors were significant at the $\alpha = .0001$ level, total sulfur dioxide and fixed acidity were

significant at the α = .001 level, while the pH was significant at the α = .01 level. The final model may be summarized in the following equation:

$$Y_{Quality} = 1.79 - 0.045 * x_{fixed\ acidity} - 1.94 * x_{volatile\ acidity} + 0.025 * x_{residual\ sugar} + 0.006 * x_{free\ sulfur\ dioxide} - 0.001 * x_{total\ sulfur\ dioxide} + 0.413 * x_{sulphates} + 0.186 * x_{pH} + 0.371 * x_{alcohol} + \varepsilon$$

It may also be presented in the standardized form, which makes the regression coefficients more comparable amongst one another:

$$Y_{Quality} = -0.043 * x_{fixed\ acidity} - 0.221 * x_{volatile\ acidity} + 0.144 * x_{residual\ sugar} + 0.111 * x_{free\ sulfur\ dioxide} - 0.047 * x_{total\ sulfur\ dioxide} + 0.053 * x_{sulphates} + 0.031 * x_{pH} + 0.517 * x_{alcohol} + \varepsilon$$

In this form, all the variables are standardized – transformed so that their mean is 0 and their standard deviation 1. The strongest predictor, as determined by the standardized regression coefficients (page 45 of the appendix), was the alcohol content, followed by volatile acidity, residual sugar, and free sulfur dioxide. The more alcohol, residual sugar, and free sulfur dioxide, and the less volatile acid the wine contained, the higher its quality rating was. The rest of the predictors had lower regression coefficients, but were also significant, which indicates that they have also have some impact.

4. Discussion of the results

The regression analysis that was conducted in this report showed that the provided physicochemical data can explain 30% of variance in subjectively graded wine quality. The model that was the most successful and best fitting in predicting the quality included fixed acidity, volatile acidity, residual sugar, free sulfur dioxide, total sulfur dioxide, sulphates, pH, and alcohol. The strongest positive predictor of quality was alcohol content, and the strongest negative predictor was volatile acidity. Furthermore, other variables that had a positive impact on wine quality were (in order of prediction strength): residual sugar, free sulfur dioxide, sulphates, and pH. Negative effects were also exhibited by total sulfur dioxide and fixed acidity.

There are a few notes that should be made about the final model. While the null hypothesis that all β coefficients equal 0 was rejected, there is room for the model to improve. This mainly relates to the range in which the data was collected: there were no wines which had grades of quality below 3 or over 9. Furthermore, the number of wines with these exact grades (3 and 9) was relatively low (as can be seen in page 46 of the appendix). While this is probably a natural consequence of the quality of wines, it also means that this model is realistically only applicable for wines with quality grades ranging from 4 to 8, and that it performs best in the range of 5 to 7. Thus, the possibility of extrapolating beyond these grades is limited. The relationship of the regression variables that were used may be non-linear beyond the created model, as it is sensible that “too much” or “too little” of any physicochemical variable could ruin the quality of the wine.

Furthermore, although it was a very comprehensive measurement of physicochemical properties of the wines, the model captured less than a third of variance of the wine quality. There may be various reasons for this fact. One is the previously described limitation in the range of the response variable. Another is that it may be that the subjective aspects of wine tasting are much more important than objective ones in the determination of how good a wine is, and that the objective characteristics simply amount for this much. Thirdly, it may be that there are other physicochemical properties that were not taken into account in this study, which would improve the strength of prediction.

A last improvement that could have been done to improve the quality of this dataset and the model was to use the mean instead of the median of the three grades given by the experts as the response variable. While mean is more sensitive to extreme values, in the circumstances in which experts are the ones grading the wines, it would be sensible to take into account the grades of all three people, not only the middle one. This would lead to more precise measurements of the response variable and would potentially make the model better.

5. Conclusion

In this study, the quality of white wine, one of the most socially and medically well-liked alcoholic beverages was examined. Through a rigorous analysis and model selection, it was determined that it could be predicted by a model which had seven regressors. The model showed good signs of fit, no significant issues with any of the assumptions of regression, and predicted close to a third of the variance of wine quality.

Future studies may want to expand upon these results in multiple ways. It would be interesting to assess both subjective and objective measurements of wine characteristics and determine which ones are more useful in predicting wine quality. It would also be useful to understand their mutual relationship. Furthermore, future studies may want to utilize databases with more diverse wines, including those with very low and very high grades of quality.

6. References

Artero, A., Artero, A., Tarín, J. J., & Cano, A. (2015). The impact of moderate wine consumption on health. *Maturitas*, *80*(1), 3–13.

Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). *Using data mining for wine quality assessment*. 66–79.

Ebeler, S. (1999). *Flavor Chemistry—Thirty Years of Progress*. Kluwer Academic Publishers.

Legin, A., Rudnitskaya, A., Lvova, L., Vlasov, Y., Di Natale, C., & D'amico, A. (2003). Evaluation of Italian wine by the electronic tongue: Recognition, quantitative analysis and correlation with human sensory perception. *Analytica Chimica Acta*, *484*(1), 33–44.

Lumley, T. (2020). *Leaps: Regression Subset Selection. R package version 3.1. 2020*.

Mezzano, D., Leighton, F., Martinez, C., Marshall, G., Cuevas, A., Castillo, O., Panes, O., Munoz, B., Perez, D., & Mizon, C. (2001). Complementary effects of Mediterranean diet and moderate red wine intake on haemostatic cardiovascular risk factors. *European Journal of Clinical Nutrition*, *55*(6), 444–451.

Ritchie, H., & Roser, M. (2018). Alcohol consumption. *Our World in Data*.

7. Appendix

#reading the data

```
winequality.white <-
  read.csv("C:/MATH3330/Project/R/winequality-white.csv", sep = ";")
View(winequality.white)
```

#required libraries

```
library(leaps)
library(MASS)
library(car)
library(qpcR)
library(Hmisc)
library(olsrr)
library(asbio)
library(QuantPsyc)
```

#model selection

```
x <- model.matrix(quality ~ . - 1, data = winequality.white)
y <- winequality.white$quality

bestmods <- leaps(x, y, nbest = 1)
print(bestmods)
```

```
$which
      1     2     3     4     5     6     7     8     9     A     B
1  FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
```

```
2  FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
3  FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE TRUE
4  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE TRUE
5  FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE TRUE
6  FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE  TRUE  TRUE  TRUE TRUE
7  FALSE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE TRUE
8   TRUE  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE TRUE
9   TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE TRUE
10  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE TRUE
11  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE TRUE

$label
 [1] "(Intercept)" "1"           "2"           "3"           "4"           "5"           "6"           "7"
 [9] "8"           "9"           "A"           "B"

$size
 [1]  2  3  4  5  6  7  8  9 10 11 12

$Cp
 [1] 618.935028 277.304045 154.828971 119.625443  73.461400  37.464938  15.911941
6.805571   8.238314  10.053204
[11]  12.000000
```

#creating all models

```
modela <- lm(quality ~ alcohol, data=winequality.white)
modelb <- lm(quality ~ alcohol + volatile.acidity, data=winequality.white)
modelc <- lm(quality ~ alcohol + volatile.acidity + residual.sugar, data=winequality.white)
modeld <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + free.sulfur.dioxide,
data=winequality.white)
```

```
modele <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH,
data=winequality.white)

modelf <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates,
data=winequality.white)

modelg <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide, data=winequality.white)

modelh <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide + fixed.acidity, data=winequality.white)

modeli <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide, data=winequality.white)

modelj <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide + chlorides, data=winequality.white)

modelk <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide + chlorides + citric.acid,
data=winequality.white)
```

#calculating press and adjusted R^2 for all models

```
press(modela, as.R2 = FALSE)

[1] 3114.842

press(modelb, as.R2 = FALSE)

[1] 2922.024

press(modelc, as.R2 = FALSE)

[1] 2852.789

press(modeld, as.R2 = FALSE)

[1] 2835.081

press(modele, as.R2 = FALSE)

[1] 2808.287

press(modelf, as.R2 = FALSE)

[1] 2788.672

press(modelg, as.R2 = FALSE)

[1] 2778.756

press(modelh, as.R2 = FALSE)
```

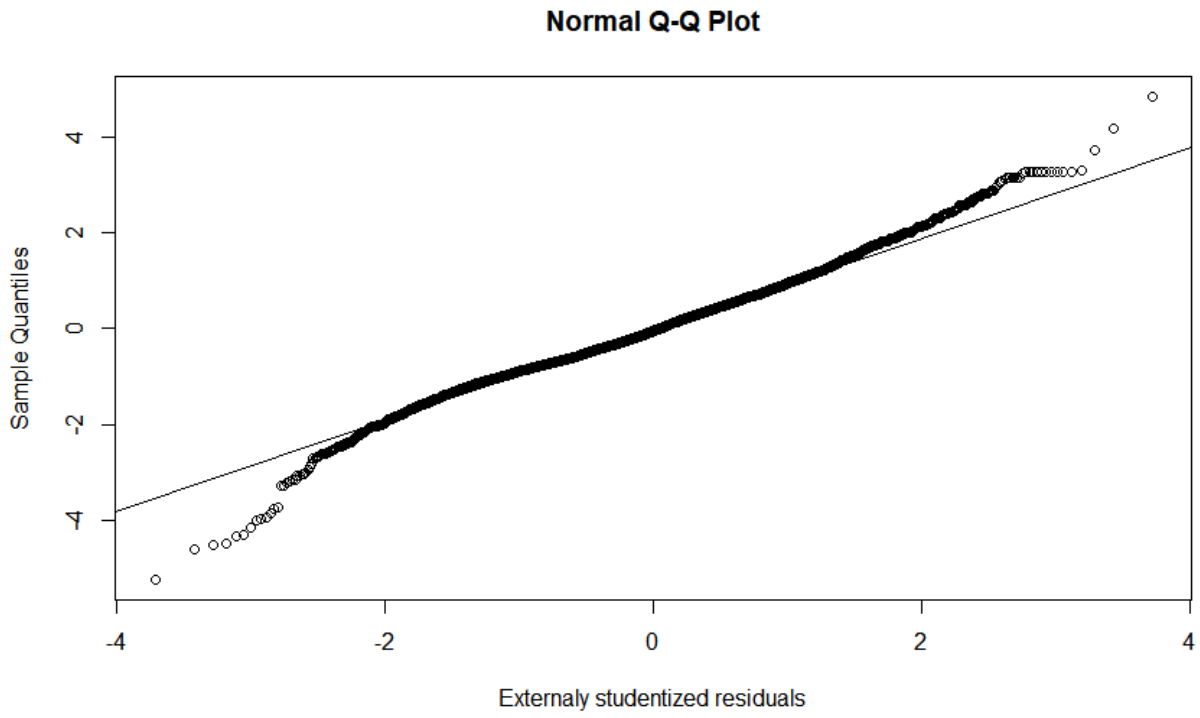
```
[1] 2782.164

press(modeli, as.R2 = FALSE)

[1] 2783.621

press(modelj, as.R2 = FALSE)

[1] 2784.448

press(modelk, as.R2 = FALSE)

[1] 2785.419

summary(modela)$adj.r.squared

[1] 0.1895598

summary(modelb)$adj.r.squared

[1] 0.2399208

summary(modelc)$adj.r.squared

[1] 0.2580716

summary(modeld)$adj.r.squared

[1] 0.2633925

summary(modele)$adj.r.squared

[1] 0.2703282

summary(modelf)$adj.r.squared

[1] 0.2757705

summary(modelg)$adj.r.squared

[1] 0.2790891

summary(modelh)$adj.r.squared

[1] 0.2805767

summary(modeli)$adj.r.squared

[1] 0.280513

summary(modelj)$adj.r.squared

[1] 0.2803931

summary(modelk)$adj.r.squared
```

[1] 0.2802536

#creating and assessing residuals of the two best models

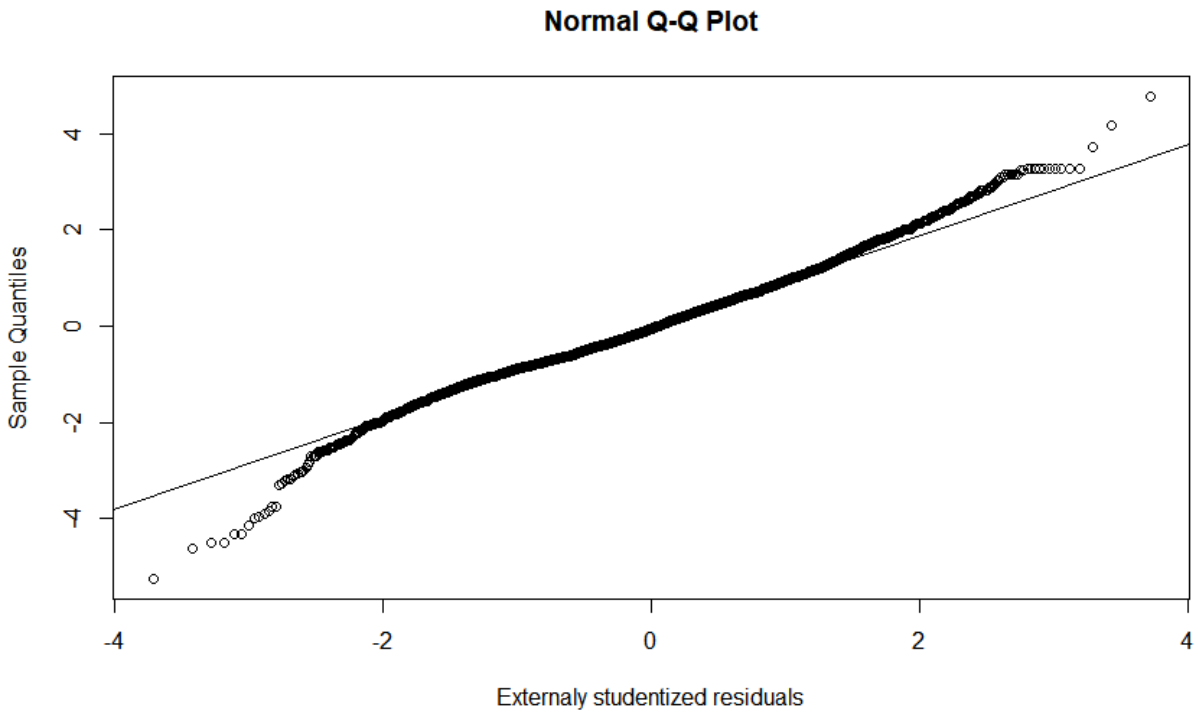
```
model8 <- modelh
model9 <- modeli

stres8 <- rstudent(model8)
qqnorm(stres8,
    xlab = "Externaly studentized residuals")
qqline(stres8)
```

Figure 1



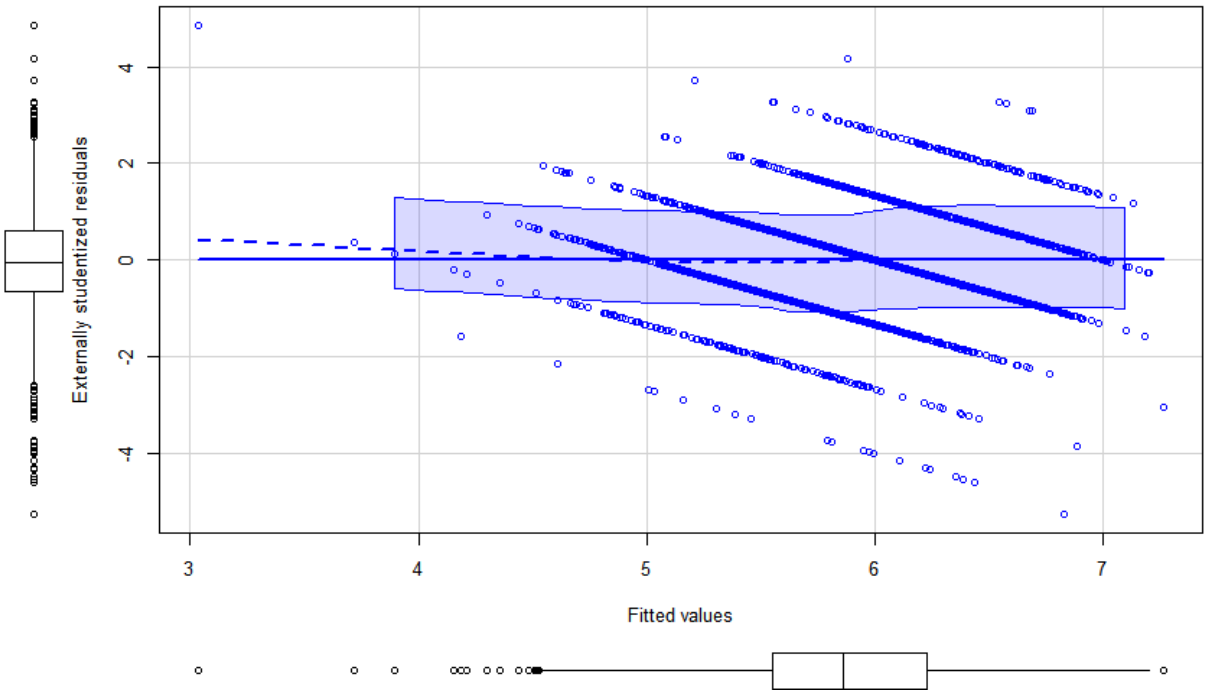

```
stres9 <- rstudent(model9)
qqnorm(stres9,
    xlab = "Externaly studentized residuals")
qqline(stres9)
```

Figure 2



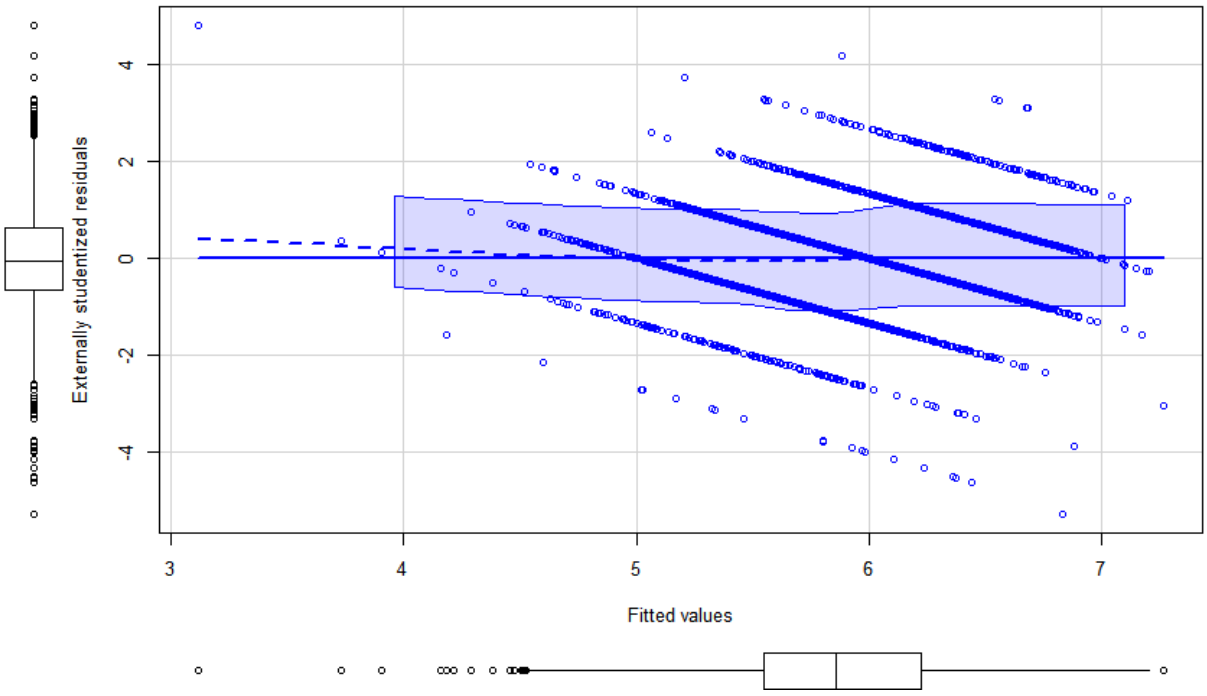

```
fv8 <- fitted.values(model8)

fv9 <- fitted.values(model9)

scatterplot(fv8, stres8,
        xlab = "Fitted values",
        ylab = "Externally studentized residuals")
```

Figure 3




```
scatterplot(fv9, stres9,
        xlab = "Fitted values",
        ylab = "Externally studentized residuals")
```

Figure 4



#Testing for outliers

outlierTest(model8)

4746	-5.246380	1.6167e-07	0.00079185
2782	4.844262	1.3101e-06	0.00641690
3308	-4.601124	4.3085e-06	0.02110300
254	-4.522712	6.2497e-06	0.03061100
446	-4.477447	7.7261e-06	0.03784200

outlierTest(model9)

	rstudent	unadjusted p-value	Bonferroni p
4746	-5.262167	1.4845e-07	0.00072712
2782	4.785201	1.7585e-06	0.00861310

3308 -4.615441 4.0231e-06 0.01970500

254 -4.511264 6.5953e-06 0.03230400

446 -4.490605 7.2657e-06 0.03558700

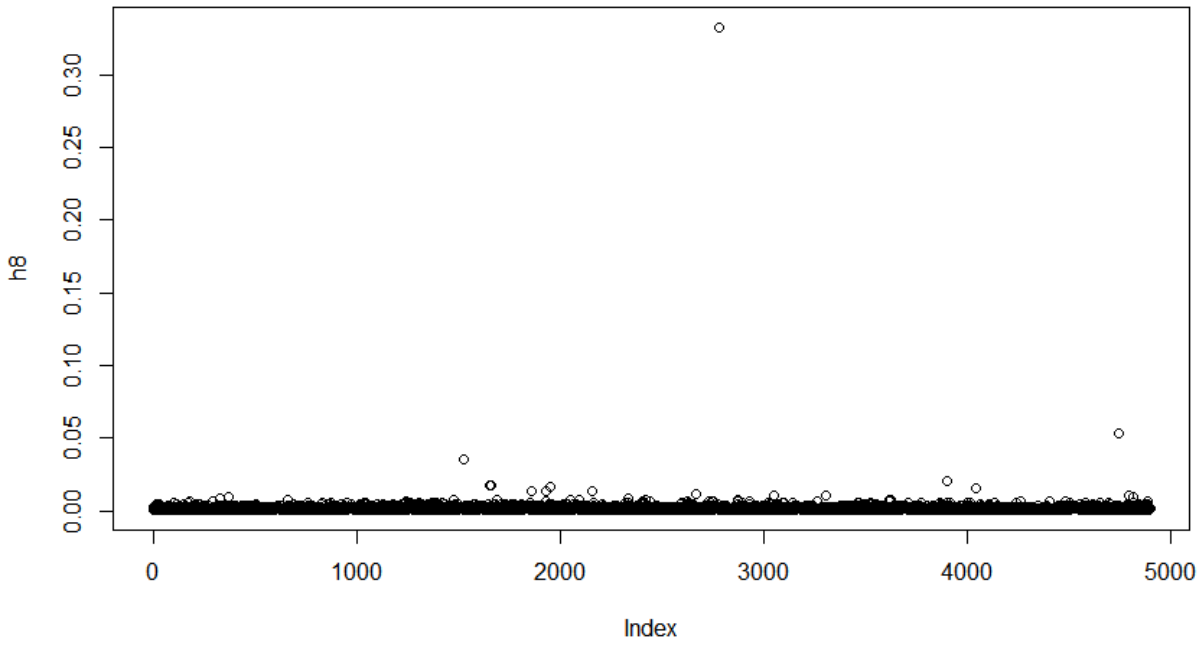
#Assessing hat-values and PRESS residuals

h8 <- hatvalues(model8)

h9 <- hatvalues(model9)

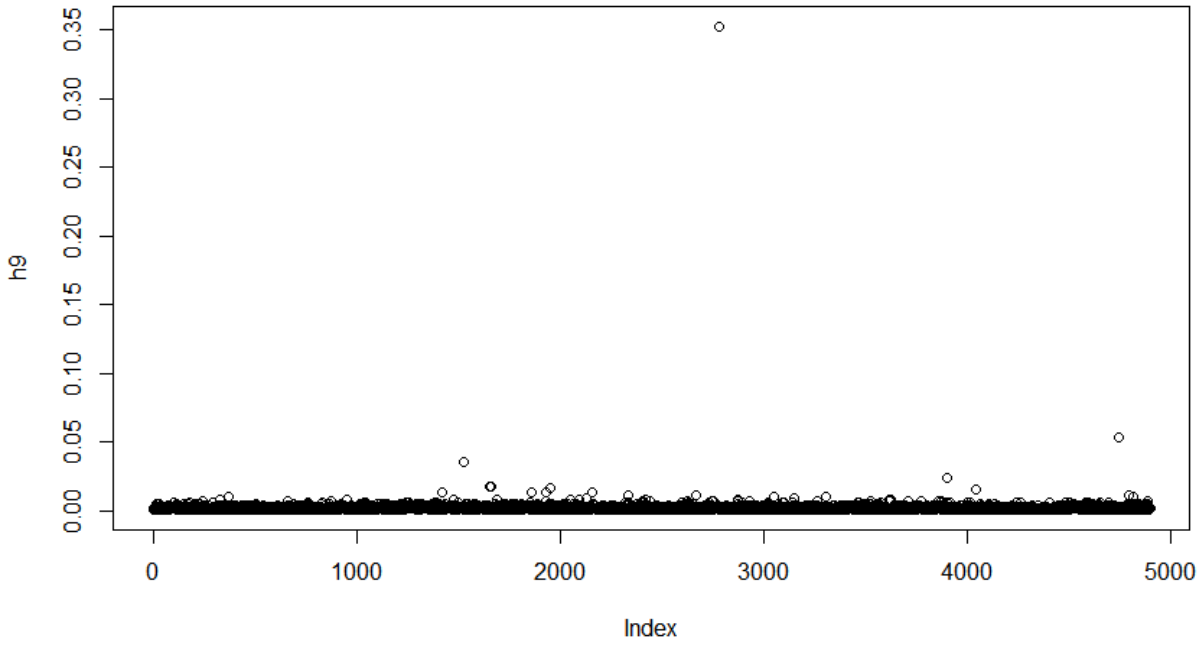
plot(h8)

Figure 5



plot(h9)

Figure 6



boxplot(h8, plot=FALSE)

$stats

[,1]

[1,] 0.0003172086

[3,] 0.0014983560

[4,] 0.0020754325

[5,] 0.0035435984

$n

[1] 4898

$conf

[,1]

[1,] 0.001476141

[2,] 0.001520571

$out

18 21 24 32 73 99 116 148 170 179

0.004145512 0.004145512 0.004852827 0.004263439 0.003594924 0.005510804 0.004301744 0.004248638 0.005124168 0.006073555

208 209 222 231 251 272 295 312 326 339

0.004182518 0.003622356 0.004321368 0.004441739 0.003953362 0.003584609 0.006090535 0.003904209 0.008046919 0.003850804

359 373 445 479 509 660 688 702 758 759

0.003610810 0.009502948 0.003926842 0.003576830 0.004481776 0.007387979 0.004733249 0.003582149 0.003834838 0.005043991

760 767 822 831 835 853 855 867 874 927

0.005210451 0.003723988 0.004032222 0.005499783 0.005499783 0.004603939 0.004603939 0.004603939 0.005667523 0.004134996

949 975 1015 1017 1035 1037 1041 1054 1100 1124

0.005727325 0.003998868 0.003883023 0.004055420 0.005541808 0.004619190 0.005949239 0.004210392 0.004050577 0.003960145

1127 1153 1172 1173 1179 1181 1191 1215 1218 1229

0.004447483 0.003788297 0.003906847 0.003872680 0.004199028 0.003906847 0.003872680 0.003870696 0.004248549 0.004428430

1240 1246 1251 1256 1264 1294 1295 1305 1308 1321

0.006361711 0.005688490 0.004242059 0.005421538 0.005116611 0.004661518 0.004652154 0.005119692 0.004011113 0.004459153

1336 1351 1353 1370 1373 1374 1386 1387 1395 1402

0.003834203 0.004069679 0.003834203 0.003872810 0.005176938 0.005176938 0.004324919 0.005856706 0.005856706 0.004144191

1408 1418 1437 1477 1483 1497 1527 1576 1578 1581

0.004627059 0.005881378 0.004523478 0.007777831 0.004116045 0.004543095 0.035284777 0.004187799 0.004573171 0.004668688

1650 1654 1664 1682 1689 1709 1723 1728 1732 1760

0.004025946 0.017119840 0.017119840 0.004630545 0.007715163 0.004179927 0.003798032 0.004474719 0.004752837 0.004776711

1776 1802 1808 1810 1843 1844 1849 1857 1863 1887

0.004810610 0.003694384 0.004380159 0.004380159 0.003565907 0.003851743 0.005390693 0.013592629 0.003916165 0.003899657

1901 1932 1933 1945 1948 1952 1959 1962 1963 1964

0.003950174 0.013547012 0.006453741 0.003920882 0.004417838 0.016476094 0.003783971 0.004451391 0.003572835 0.003783971

1996 1998 1999 2007 2015 2018 2031 2037 2051 2064

0.003649742 0.003649742 0.003649742 0.003649742 0.004384238 0.004384238 0.004029057 0.004361032 0.007826631 0.004134405

2076 2093 2155 2163 2165 2207 2212 2322 2335 2337

0.004134405 0.007113641 0.013170020 0.005022971 0.003875968 0.004024639 0.003581651 0.005831508 0.008792613 0.005399276

2395 2404 2405 2409 2418 2442 2590 2595 2626 2630

0.005347202 0.006687028 0.004018943 0.005652499 0.007652429 0.006237288 0.004253347 0.005770596 0.005985134 0.005667101

2635 2638 2647 2652 2669 2712 2731 2732 2749 2751

0.004699440 0.004699440 0.003932082 0.004810893 0.011023161 0.003817026 0.006213403 0.006156287 0.006746863 0.006746863

2772 2782 2819 2873 2874 2875 2894 2927 2931 2932

0.004054743 0.333082194 0.003674834 0.007878029 0.004592685 0.006601364 0.005052917 0.004023665 0.006282257 0.004023665

3015 3023 3024 3026 3051 3073 3095 3096 3098 3140

0.003685867 0.005886393 0.003685867 0.004462236 0.010467144 0.004626169 0.005043954 0.005043954 0.005727441 0.005125894

3166 3245 3266 3308 3418 3437 3459 3462 3471
3498

0.003980462 0.003561562 0.006142003 0.010077047 0.003877774 0.003670117 0.003669771
0.005188779 0.005188779 0.004906839

3521 3524 3529 3557 3561 3565 3572 3620 3621
3624

0.004628725 0.004628725 0.005505238 0.004165598 0.003976413 0.003976413 0.003882941
0.007514344 0.006047141 0.007514344

3711 3755 3765 3774 3862 3863 3864 3869 3870
3872

0.005830719 0.003709106 0.003709106 0.005871314 0.004605288 0.003825536 0.003825536
0.005254416 0.004605288 0.003825536

3880 3902 3905 3916 3982 3999 4000 4001 4013
4040

0.004902023 0.020096444 0.005449718 0.005869052 0.003778682 0.005539064 0.005539064
0.005539064 0.005539064 0.015222590

4066 4108 4110 4137 4240 4260 4402 4447 4471
4481

0.004409146 0.004300494 0.003917443 0.005016504 0.005079965 0.006033996 0.006246702
0.003754034 0.003859490 0.006061612

4504 4553 4581 4583 4598 4609 4610 4618 4649
4650

0.005598149 0.004108109 0.003587484 0.005353349 0.003893911 0.003598984 0.003569888
0.004301277 0.003831102 0.005200000

4651 4697 4702 4703 4746 4788 4790 4793 4816
4819

0.005193402 0.005499204 0.004050594 0.004050594 0.053099006 0.004010745 0.004224955
0.010807544 0.009827714 0.004996037

4842 4848 4868 4878 4887 4888

0.004501144 0.003966270 0.003955419 0.004185890 0.006795682 0.004181187

$group

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

[59] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

[117] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

[175] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

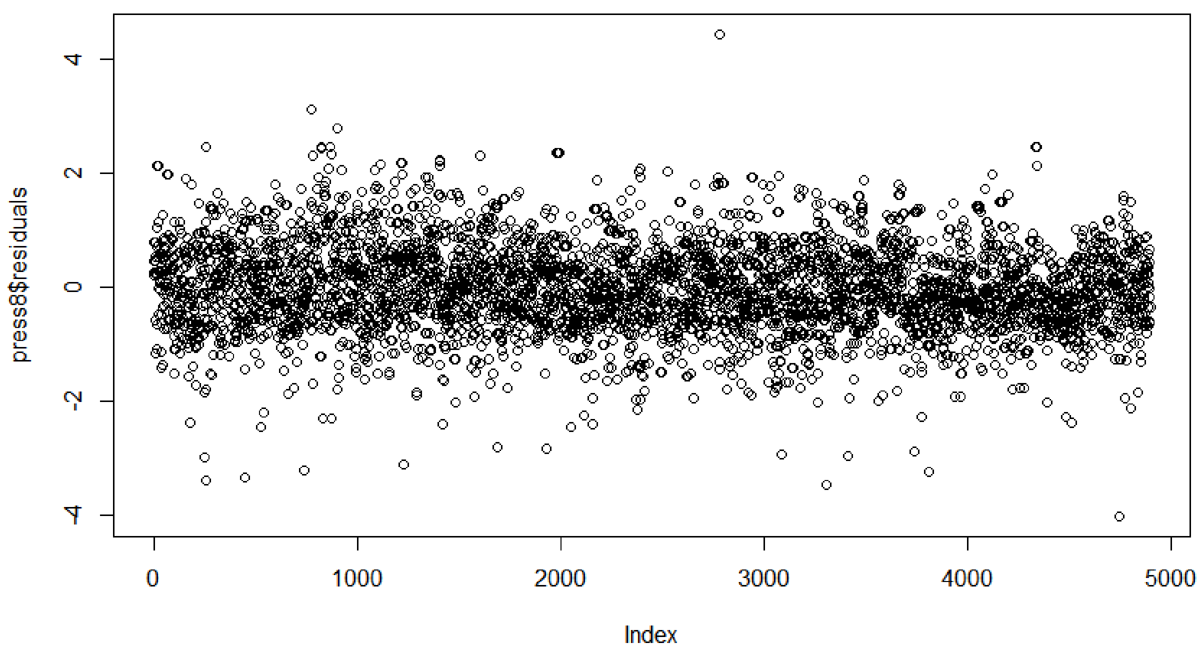
[233] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

$names

[1] "1"

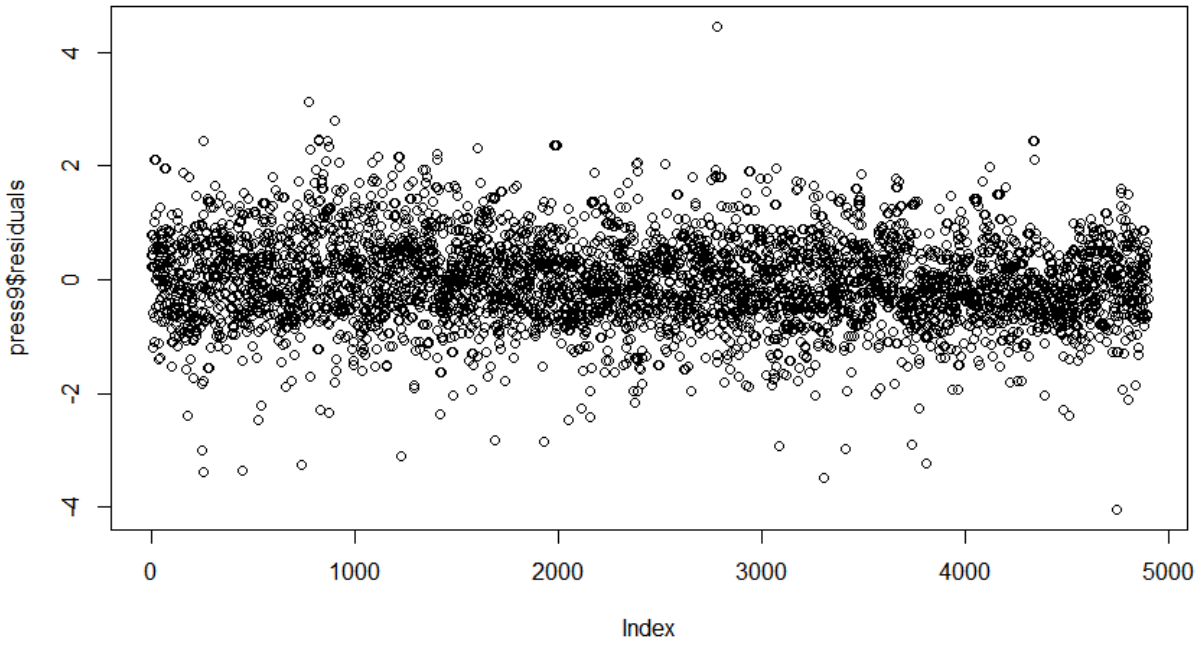
plot(press8$residuals)

Figure 7



plot(press9$residuals)

Figure 8



boxplot(press8$residuals, plot = FALSE)

$stats

[,1]

[1,] -1.93029273

[3,] -0.03968604

[4,] 0.46667406

[5,] 1.89346299

$n

[1] 4898

$conf

[,1]

[1,] -0.06138999

[2,] -0.01798208

$out

[1] 2.125551 2.125551 2.118565 1.967628 1.967628 -2.374024 -3.002687 -3.394640 2.453914 -3.360478 -2.457056

[12] -2.219106 -3.228147 3.128497 2.292912 2.292912 1.928828 2.465005 2.429463 -2.302701 2.089909 2.449609

[23] -2.319132 2.332039 2.797986 2.059265 2.062794 2.163307 2.167592 1.981029 2.167592 -3.110533 1.926994

[34] 1.926994 2.212128 2.117187 2.220835 -2.399194 -2.032317 -1.941854 2.318027 -2.809212 -2.847424 2.351379

[45] 2.351379 2.351379 2.351379 2.351379 2.351379 -2.473214 -2.250106 -1.962399 -2.414335 -1.968091 -2.160717

[56] 2.034159 2.034159 -1.978132 1.924328 2.065320 2.024752 2.024752 -1.955222 1.921543 4.445746 1.927547

[67] 1.927547 1.927547 1.960697 -2.948166 -2.021779 -3.466714 -2.974649 -1.959120 -2.009424 -2.892342 -2.289621

[78] -3.242101 1.967885 2.451487 2.451487 2.451487 2.451487 2.451487 2.451487 2.451487 2.123332 2.451487

[89] -2.025071 -2.277889 -2.385255 -4.039069 -1.952217 -2.121689

$group

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[59] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

$names

[1] "1"

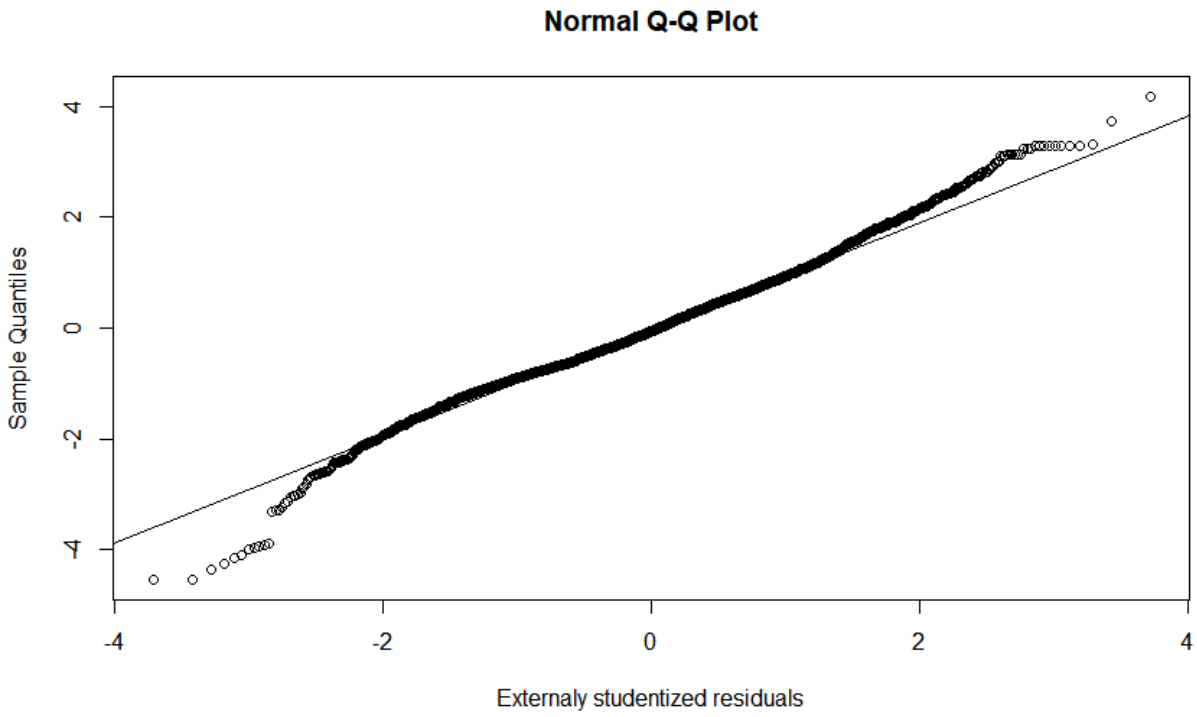
#Removing outliers

winequality.white <- winequality.white[-c(2782, 3308, 4746),]

#Assessing the residuals after the removal of outliers

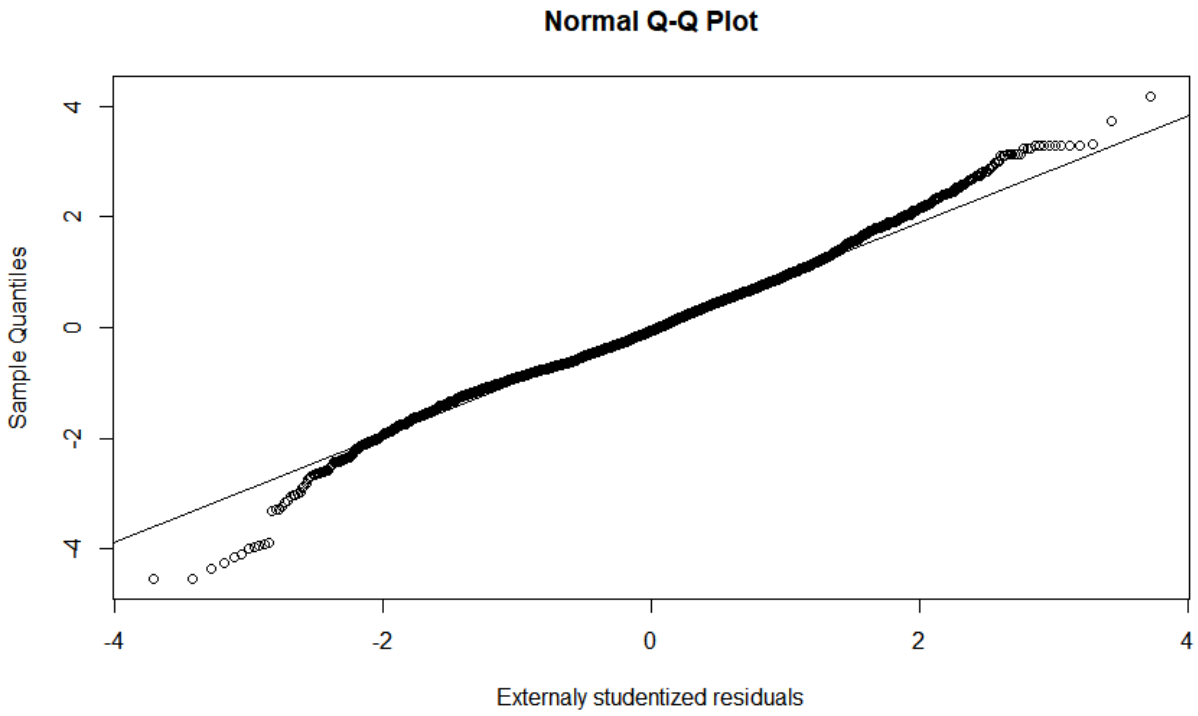
```
model8 <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide + fixed.acidity, data=winequality.white)

model9 <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + density + pH + sulphates +
free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide, data=winequality.white)

stres8 <- rstudent(model8)

qqnorm(stres8,
    xlab = "Externaly studentized residuals")

qqline(stres8)
```

Figure 9



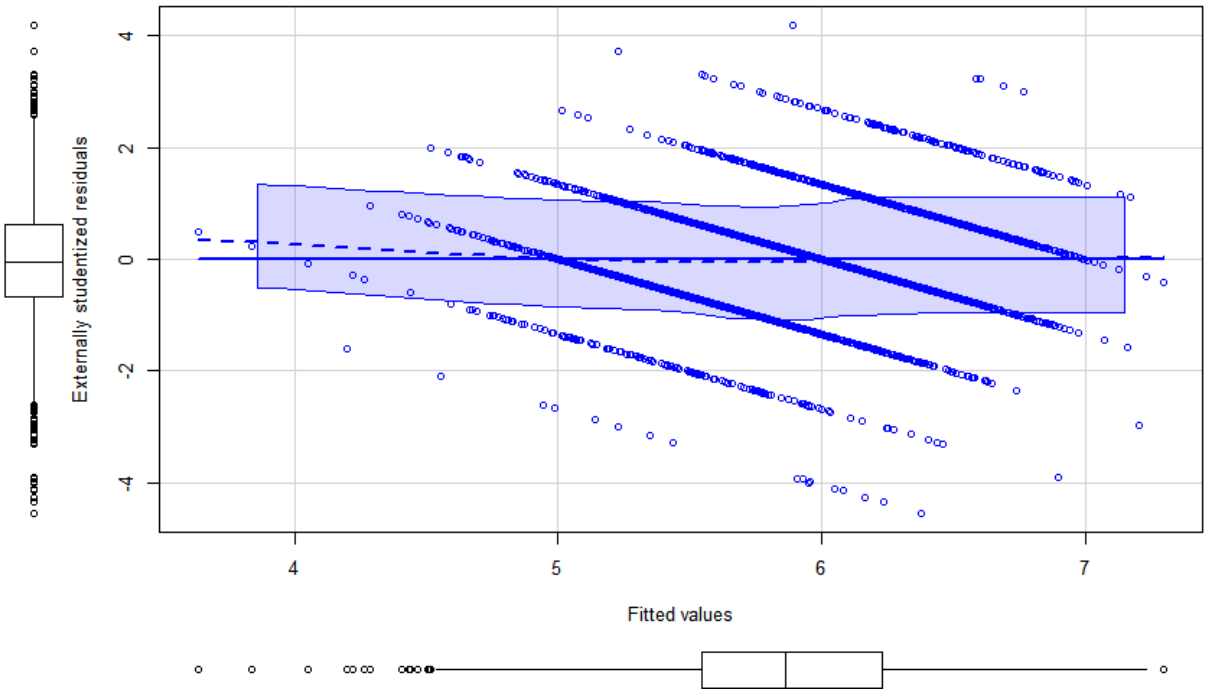

```
stres9 <- rstudent(model9)

qqnorm(stres9,
    xlab = "Externaly studentized residuals")
```

Figure 10



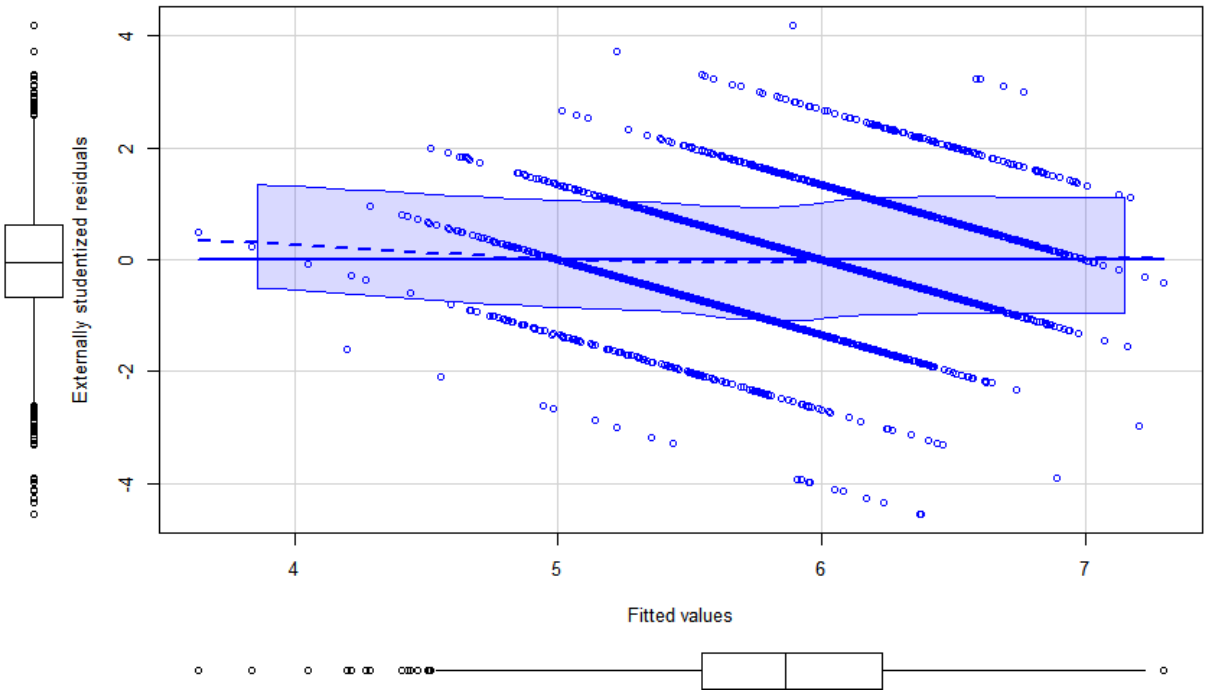

```
fv8 <- fitted.values(model8)

fv9 <- fitted.values(model9)

scatterplot(fv8, stres8,
        xlab = "Fitted values",
        ylab = "Externally studentized residuals")
```

Figure 11




```
scatterplot(fv9, stres9,
        xlab = "Fitted values",
```

Figure 12



#Assessing outliers after the removal of outliers

outlierTest(model8)

446 -4.543989 5.6531e-06 0.027672

254 -4.542769 5.6858e-06 0.027832

outlierTest(model9)

rstudent unadjusted p-value Bonferroni p

446 -4.545392 5.6157e-06 0.027489

254 -4.541436 5.7217e-06 0.028008

h8 <- hatvalues(model8)

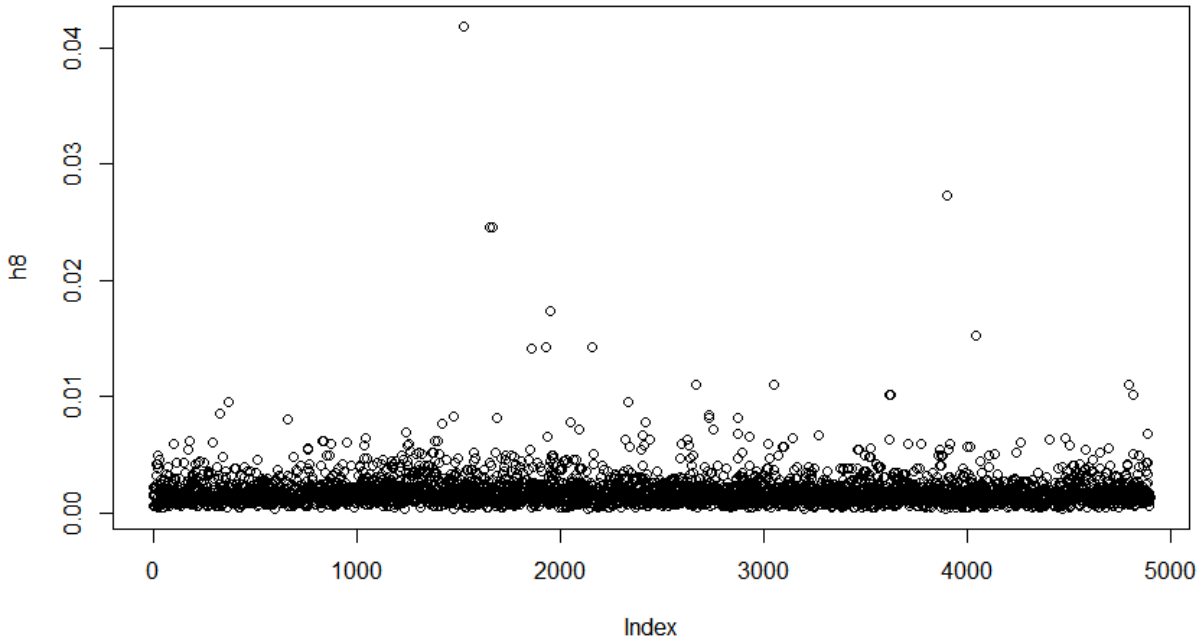
h9 <- hatvalues(model9)

press8<- PRESS(model8)

press9<- PRESS(model9)

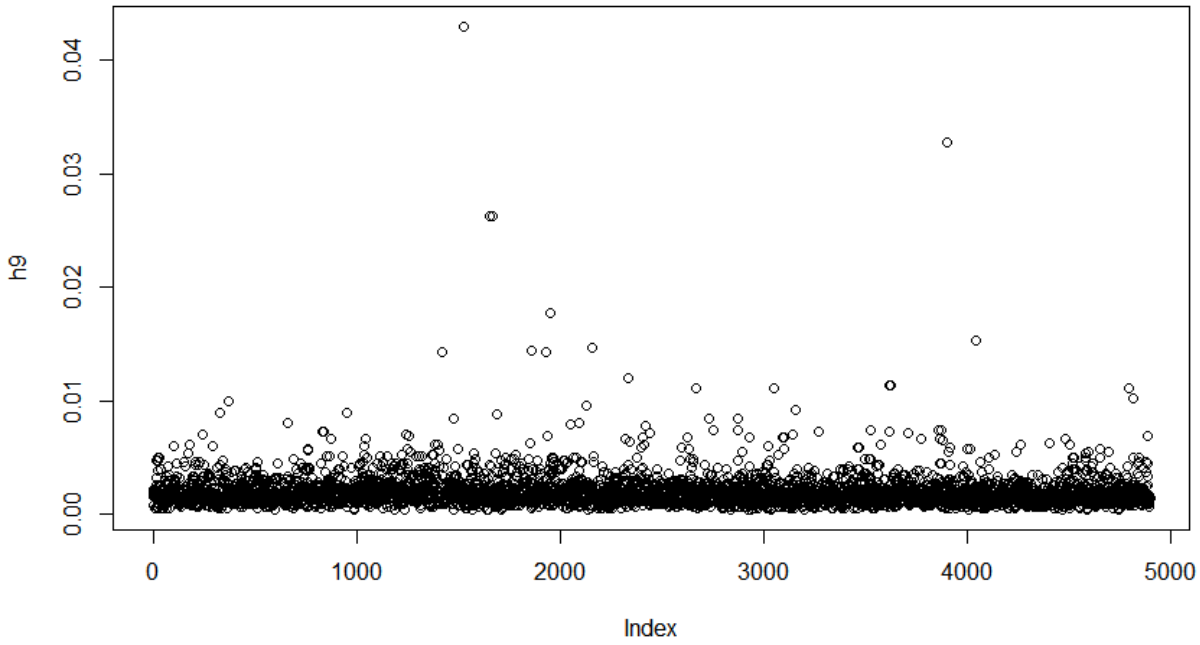
plot(h8)

Figure 13



plot(h9)

Figure 14



boxplot(h8, plot=FALSE)

$stats

[,1]

[1,] 0.0003223479

[3,] 0.0015596101

[4,] 0.0021549009

[5,] 0.0036800382

$n

[1] 4895

$conf

[,1]

[1,] 0.001536515

[2,] 0.001582705

$out

18	21	24	25	32	73	99	116	148	170
0.004150585	0.004150585	0.004913444	0.003859238	0.004557363	0.003904742	0.005967580	0.004364512	0.004329181	0.005383526

179	208	222	231	251	295	312	326	339	373
0.006149610	0.004193234	0.004373218	0.004493455	0.004289090	0.006097809	0.003906449	0.008543536	0.004818619	0.009509382

396	406	445	509	610	660	688	702	722	758
0.003775644	0.003775644	0.003972417	0.004609146	0.003748226	0.008042568	0.004794735	0.003717834	0.004024180	0.003872720

759	760	767	822	831	835	853	855	867	874
0.005450504	0.005619659	0.004162577	0.004070973	0.006137411	0.006137411	0.004940110	0.004940110	0.004940110	0.005955832

906	927	949	975	1015	1017	1035	1037	1041	1054
0.003876673	0.004200543	0.006113516	0.004063749	0.003958614	0.004081331	0.005755658	0.004713865	0.006395019	0.004660502

1079	1100	1124	1127	1142	1153	1172	1173	1179	1181
0.004094252	0.004285548	0.004193720	0.004554356	0.003704663	0.003804268	0.003918885	0.004415432	0.004251397	0.003918885

1191	1215	1218	1229	1240	1246	1250	1251	1256	1264
0.004415432	0.003887495	0.004938510	0.004510673	0.006983660	0.005761217	0.004032257	0.004407031	0.005975833	0.005374194

1294	1295	1305	1308	1309	1321	1336	1351	1353	1370
0.005018401	0.005033275	0.005163640	0.004059563	0.003694527	0.005085566	0.004074493	0.004070527	0.004074493	0.004004507

1373	1374	1386	1387	1395	1402	1408	1418	1437	1477

0.005188347 0.005188347 0.004390349 0.006179139 0.006179139 0.004269886 0.005116303
0.007664690 0.004533974 0.008332670

1483 1497 1527 1542 1562 1565 1576 1578 1581
1650

0.004264472 0.004665673 0.041870795 0.003718923 0.003734429 0.003734429 0.004415244
0.005148935 0.004705303 0.004350034

1654 1658 1664 1682 1689 1709 1723 1728 1732
1760

0.024574676 0.004125237 0.024574676 0.005155242 0.008139417 0.004688214 0.003984815
0.004527701 0.004958094 0.004894944

1776 1782 1802 1808 1810 1843 1844 1849 1857
1863

0.004833284 0.004176487 0.003886780 0.004436185 0.004436185 0.003754444 0.004606032
0.005400939 0.014159925 0.003920040

1887 1901 1932 1933 1945 1948 1952 1959 1962
1963

0.004261949 0.003992178 0.014302558 0.006496113 0.003968277 0.004491452 0.017398162
0.004980278 0.004683029 0.004252708

1964 1996 1998 1999 2007 2015 2018 2031 2037
2051

0.004980278 0.003714062 0.003714062 0.003714062 0.003714062 0.004488325 0.004488325
0.004515389 0.004375396 0.007857811

2064 2076 2093 2094 2095 2099 2109 2155 2163
2165

0.004520164 0.004520164 0.007175198 0.003815014 0.003815014 0.003828315 0.003828315
0.014210452 0.005054923 0.004156554

2207 2322 2334 2335 2337 2395 2404 2405 2409
2418

0.004067176 0.006297242 0.003726446 0.009501512 0.005717385 0.005406290 0.006739025
0.004032111 0.005846917 0.007794120

2442 2590 2595 2626 2630 2635 2638 2647 2652
2669

0.006252039 0.004641856 0.005946560 0.006309596 0.005800737 0.004711552 0.004711552
0.004075914 0.004925571 0.011041297

2712 2731 2732 2749 2751 2772 2819 2873 2874 2875

0.003933884 0.008435454 0.008190453 0.007219321 0.007219321 0.004059179 0.003761643 0.008185799 0.004660181 0.006747594

2894 2927 2931 2932 3015 3023 3024 3026 3051 3073

0.005219142 0.004085715 0.006546901 0.004085715 0.003761837 0.005944503 0.003761837 0.004717762 0.011043024 0.004879243

3095 3096 3098 3140 3166 3222 3245 3266 3380 3388

0.005738784 0.005738784 0.005734614 0.006486533 0.003990718 0.003977353 0.003804535 0.006643165 0.003863725 0.003863725

3390 3414 3418 3437 3459 3462 3471 3498 3521 3524

0.003727125 0.003799595 0.003878101 0.003854135 0.003860561 0.005459183 0.005459183 0.004911332 0.004801146 0.004801146

3529 3557 3561 3565 3572 3620 3621 3624 3711 3755

0.005556952 0.004224733 0.004000382 0.004000382 0.003912525 0.010179427 0.006323590 0.010179427 0.005890992 0.003842159

3765 3774 3862 3863 3864 3869 3870 3872 3880 3902

0.003842159 0.005964763 0.004904724 0.004115324 0.004115324 0.005492097 0.004904724 0.004115324 0.004941966 0.027330561

3905 3916 3982 3999 4000 4001 4013 4040 4066 4108

0.005468965 0.005929491 0.003884394 0.005653256 0.005653256 0.005653256 0.005653256 0.015249772 0.004409371 0.004999253

4110 4137 4240 4260 4402 4447 4471 4481 4504 4520

0.003935229 0.005099167 0.005227582 0.006057844 0.006249526 0.003768479 0.003937790 0.006413008 0.005860418 0.004117757

4526 4553 4580 4581 4583 4598 4618 4649 4650 4651

0.004117757 0.004206003 0.003693255 0.003884123 0.005377731 0.003906405 0.004539567 0.003839911 0.005223659 0.005218225

4697 4702 4703 4788 4790 4793 4816 4819 4842 4848

0.005509816 0.004053208 0.004053208 0.004016740 0.004238124 0.011053533 0.010201240 0.005004142 0.004885072 0.003992209

4868 4878 4887 4888

0.004026349 0.004366082 0.006836932 0.004274167

$group

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[59] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[117] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[175] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

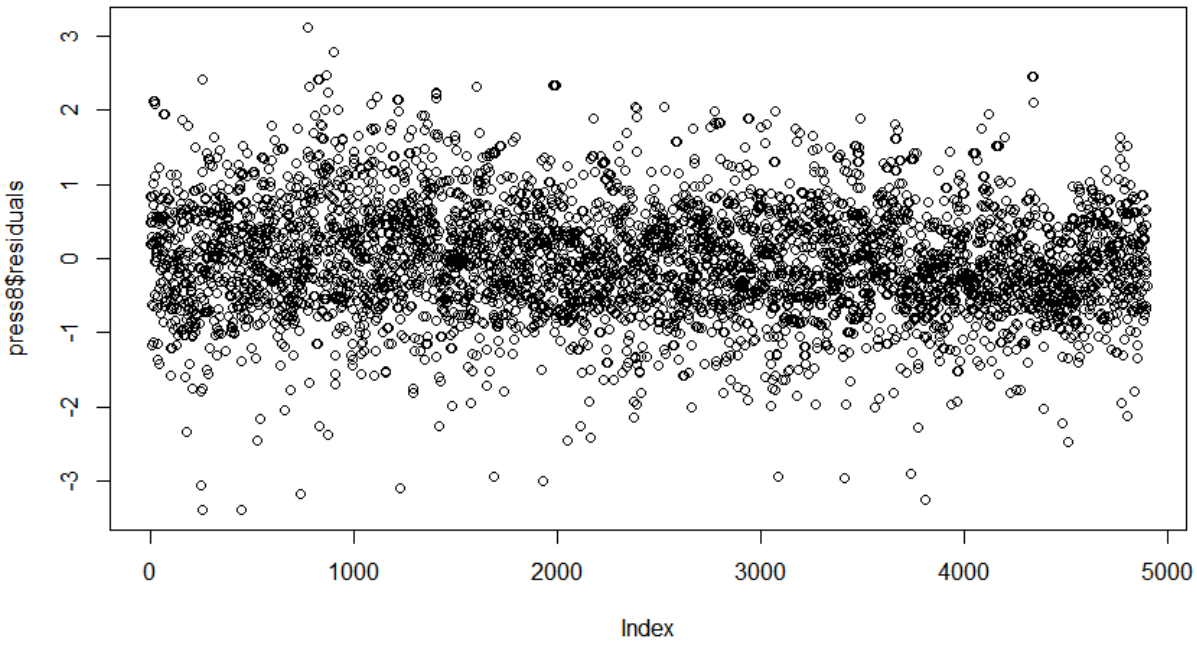
[233] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

$names

[1] "1"

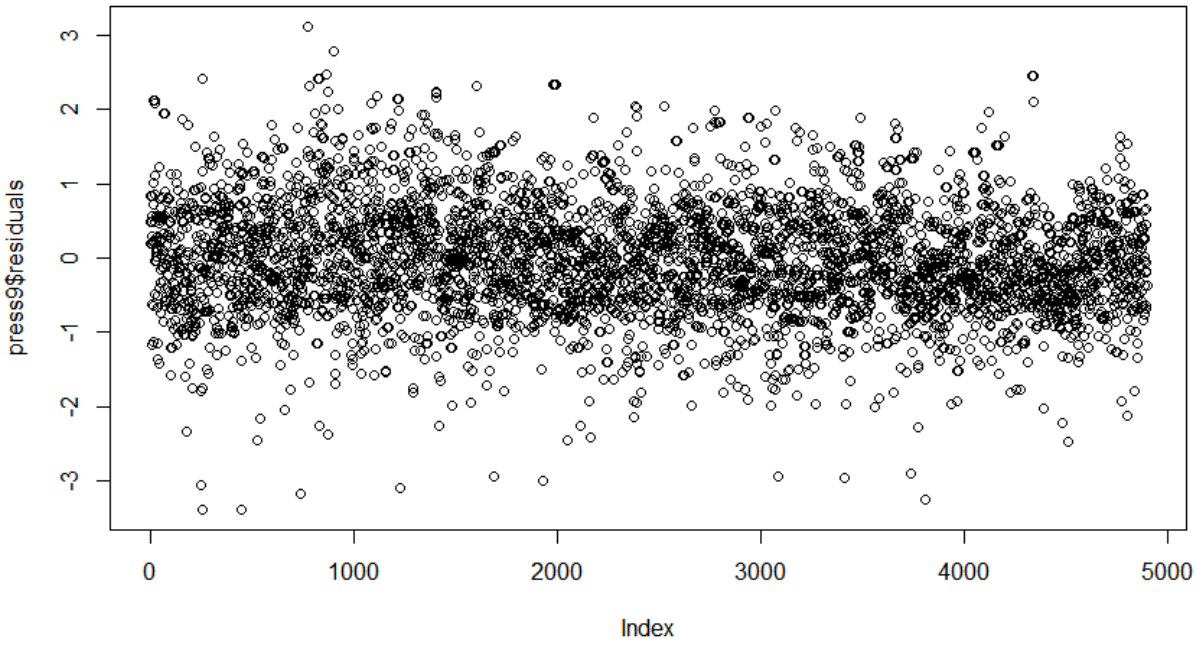
plot(press8$residuals)

Figure 15



plot(press9$residuals)

Figure 16



```
boxplot(press8$residuals, plot = FALSE)
$stats
          [,1]
[1,] 0.0003223479
[2,] 0.0011322232
[3,] 0.0015596101
[4,] 0.0021549009
[5,] 0.0036800382

$n
[1] 4895

$conf
          [,1]
[1,] 0.001536515
[2,] 0.001582705

$out
       18        21        24        25        32        73        99       116       148       170
0.004150585 0.004150585 0.004913444 0.003859238 0.004557363 0.003904742 0.005967580
0.004364512 0.004329181 0.005383526
      179       208       222       231       251       295       312       326       339       373
0.006149610 0.004193234 0.004373218 0.004493455 0.004289090 0.006097809 0.003906449
0.008543536 0.004818619 0.009509382
      396       406       445       509       610       660       688       702       722       758
0.003775644 0.003775644 0.003972417 0.004609146 0.003748226 0.008042568 0.004794735
0.003717834 0.004024180 0.003872720
      759       760       767       822       831       835       853       855       867       874
```

0.005450504 0.005619659 0.004162577 0.004070973 0.006137411 0.006137411 0.004940110
0.004940110 0.004940110 0.005955832

906 927 949 975 1015 1017 1035 1037 1041 1054

0.003876673 0.004200543 0.006113516 0.004063749 0.003958614 0.004081331 0.005755658
0.004713865 0.006395019 0.004660502

1079 1100 1124 1127 1142 1153 1172 1173 1179
1181

0.004094252 0.004285548 0.004193720 0.004554356 0.003704663 0.003804268 0.003918885
0.004415432 0.004251397 0.003918885

1191 1215 1218 1229 1240 1246 1250 1251 1256
1264

0.004415432 0.003887495 0.004938510 0.004510673 0.006983660 0.005761217 0.004032257
0.004407031 0.005975833 0.005374194

1294 1295 1305 1308 1309 1321 1336 1351 1353
1370

0.005018401 0.005033275 0.005163640 0.004059563 0.003694527 0.005085566 0.004074493
0.004070527 0.004074493 0.004004507

1373 1374 1386 1387 1395 1402 1408 1418 1437
1477

0.005188347 0.005188347 0.004390349 0.006179139 0.006179139 0.004269886 0.005116303
0.007664690 0.004533974 0.008332670

1483 1497 1527 1542 1562 1565 1576 1578 1581
1650

0.004264472 0.004665673 0.041870795 0.003718923 0.003734429 0.003734429 0.004415244
0.005148935 0.004705303 0.004350034

1654 1658 1664 1682 1689 1709 1723 1728 1732
1760

0.024574676 0.004125237 0.024574676 0.005155242 0.008139417 0.004688214 0.003984815
0.004527701 0.004958094 0.004894944

1776 1782 1802 1808 1810 1843 1844 1849 1857
1863

0.004833284 0.004176487 0.003886780 0.004436185 0.004436185 0.003754444 0.004606032
0.005400939 0.014159925 0.003920040

1887 1901 1932 1933 1945 1948 1952 1959 1962 1963

0.004261949 0.003992178 0.014302558 0.006496113 0.003968277 0.004491452 0.017398162 0.004980278 0.004683029 0.004252708

1964 1996 1998 1999 2007 2015 2018 2031 2037 2051

0.004980278 0.003714062 0.003714062 0.003714062 0.003714062 0.004488325 0.004488325 0.004515389 0.004375396 0.007857811

2064 2076 2093 2094 2095 2099 2109 2155 2163 2165

0.004520164 0.004520164 0.007175198 0.003815014 0.003815014 0.003828315 0.003828315 0.014210452 0.005054923 0.004156554

2207 2322 2334 2335 2337 2395 2404 2405 2409 2418

0.004067176 0.006297242 0.003726446 0.009501512 0.005717385 0.005406290 0.006739025 0.004032111 0.005846917 0.007794120

2442 2590 2595 2626 2630 2635 2638 2647 2652 2669

0.006252039 0.004641856 0.005946560 0.006309596 0.005800737 0.004711552 0.004711552 0.004075914 0.004925571 0.011041297

2712 2731 2732 2749 2751 2772 2819 2873 2874 2875

0.003933884 0.008435454 0.008190453 0.007219321 0.007219321 0.004059179 0.003761643 0.008185799 0.004660181 0.006747594

2894 2927 2931 2932 3015 3023 3024 3026 3051 3073

0.005219142 0.004085715 0.006546901 0.004085715 0.003761837 0.005944503 0.003761837 0.004717762 0.011043024 0.004879243

3095 3096 3098 3140 3166 3222 3245 3266 3380 3388

0.005738784 0.005738784 0.005734614 0.006486533 0.003990718 0.003977353 0.003804535 0.006643165 0.003863725 0.003863725

3390 3414 3418 3437 3459 3462 3471 3498 3521 3524

0.003727125 0.003799595 0.003878101 0.003854135 0.003860561 0.005459183 0.005459183 0.004911332 0.004801146 0.004801146

3529 3557 3561 3565 3572 3620 3621 3624 3711 3755

0.005556952 0.004224733 0.004000382 0.004000382 0.003912525 0.010179427 0.006323590 0.010179427 0.005890992 0.003842159

3765 3774 3862 3863 3864 3869 3870 3872 3880 3902

0.003842159 0.005964763 0.004904724 0.004115324 0.004115324 0.005492097 0.004904724 0.004115324 0.004941966 0.027330561

3905 3916 3982 3999 4000 4001 4013 4040 4066 4108

0.005468965 0.005929491 0.003884394 0.005653256 0.005653256 0.005653256 0.005653256 0.015249772 0.004409371 0.004999253

4110 4137 4240 4260 4402 4447 4471 4481 4504 4520

0.003935229 0.005099167 0.005227582 0.006057844 0.006249526 0.003768479 0.003937790 0.006413008 0.005860418 0.004117757

4526 4553 4580 4581 4583 4598 4618 4649 4650 4651

0.004117757 0.004206003 0.003693255 0.003884123 0.005377731 0.003906405 0.004539567 0.003839911 0.005223659 0.005218225

4697 4702 4703 4788 4790 4793 4816 4819 4842 4848

0.005509816 0.004053208 0.004053208 0.004016740 0.004238124 0.011053533 0.010201240 0.005004142 0.004885072 0.003992209

4868 4878 4887 4888

0.004026349 0.004366082 0.006836932 0.004274167

$group

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[59] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

[117] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

[175] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

[233] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

$names

[1] "1"

#assessing bivariate correlations amongst predictors

```
predictors8 <- winequality.white[c(1, 2, 4, 6, 8, 9, 10, 11)]
predictors8 <- matrix(unlist(predictors), ncol = 8, nrow = 4893)

predictors9 <- winequality.white[c(1, 2, 4, 6, 7, 8, 9, 10, 11)]
predictors9 <- matrix(unlist(predictors), ncol = 9, nrow = 4893)

rcorr(predictors8)

      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]
[1,]  1.00 -0.03  0.01  0.01  0.01  0.01  0.02 -0.08
[2,] -0.03  1.00 -0.03 -0.03  0.36 -0.17  0.23 -0.12
[3,]  0.01 -0.03  1.00  0.04  0.00 -0.01  0.00  0.00
[4,]  0.01 -0.03  0.04  1.00 -0.06  0.03 -0.05  0.00
[5,]  0.01  0.36  0.00 -0.06  1.00 -0.46  0.57 -0.28
[6,]  0.01 -0.17 -0.01  0.03 -0.46  1.00 -0.36  0.16
[7,]  0.02  0.23  0.00 -0.05  0.57 -0.36  1.00 -0.30
```

```
[8,] -0.08 -0.12  0.00  0.00 -0.28  0.16 -0.30  1.00
```

n= 4893

P

```
     [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
[1,]       0.0503 0.6805 0.5456 0.3351 0.3285 0.1444 0.0000
[2,] 0.0503        0.0153 0.0789 0.0000 0.0000 0.0000 0.0000
[3,] 0.6805 0.0153        0.0055 0.9338 0.3392 0.8650 0.9263
[4,] 0.5456 0.0789 0.0055        0.0000 0.0281 0.0015 0.9843
[5,] 0.3351 0.0000 0.9338 0.0000        0.0000 0.0000 0.0000
[6,] 0.3285 0.0000 0.3392 0.0281 0.0000        0.0000 0.0000
[7,] 0.1444 0.0000 0.8650 0.0015 0.0000 0.0000        0.0000
[8,] 0.0000 0.0000 0.9263 0.9843 0.0000 0.0000 0.0000
```

rcorr(predictors9)

```
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]
 [1,]  1.00 -0.03  0.01  0.01  0.06  0.01  0.02  0.02 -0.07
 [2,] -0.03  1.00 -0.03 -0.03 -0.02  0.34 -0.18  0.22 -0.11
 [3,]  0.01 -0.03  1.00  0.04  0.07 -0.01  0.02 -0.01 -0.02
 [4,]  0.01 -0.03  0.04  1.00  0.06 -0.05  0.02 -0.04  0.03
 [5,]  0.06 -0.02  0.07  0.06  1.00 -0.09  0.07 -0.07 -0.03
 [6,]  0.01  0.34 -0.01 -0.05 -0.09  1.00 -0.53  0.65 -0.33
 [7,]  0.02 -0.18  0.02  0.02  0.07 -0.53  1.00 -0.41  0.19
 [8,]  0.02  0.22 -0.01 -0.04 -0.07  0.65 -0.41  1.00 -0.32
 [9,] -0.07 -0.11 -0.02  0.03 -0.03 -0.33  0.19 -0.32  1.00
```

n= 4893

P

```
     [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]   [,9]
 [1,]       0.0503 0.6805 0.5456 0.0000 0.4433 0.2024 0.0989 0.0000
 [2,] 0.0503        0.0153 0.0789 0.1732 0.0000 0.0000 0.0000 0.0000
 [3,] 0.6805 0.0153        0.0055 0.0000 0.5627 0.2677 0.4509 0.0982
 [4,] 0.5456 0.0789 0.0055        0.0001 0.0003 0.1350 0.0068 0.0180
 [5,] 0.0000 0.1732 0.0000 0.0001        0.0000 0.0000 0.0000 0.0244
 [6,] 0.4433 0.0000 0.5627 0.0003 0.0000        0.0000 0.0000 0.0000
 [7,] 0.2024 0.0000 0.2677 0.1350 0.0000 0.0000        0.0000 0.0000
 [8,] 0.0989 0.0000 0.4509 0.0068 0.0000 0.0000 0.0000        0.0000
 [9,] 0.0000 0.0000 0.0982 0.0180 0.0244 0.0000 0.0000 0.0000
```

#VIF

```
vif(model8)
        alcohol    volatile.acidity      residual.sugar             density                  pH           sulphates
       10.685809           1.048146          14.732942          35.456599           2.368221
1.155558
free.sulfur.dioxide       fixed.acidity
          1.161087           3.085107
vif(model9)
        alcohol     volatile.acidity      residual.sugar             density                  pH
       10.872452           1.090148          15.379550          38.090605           2.383991
       sulphates  free.sulfur.dioxide       fixed.acidity total.sulfur.dioxide
        1.157845           1.776772          3.118242           2.279134
```

#Eigensystem analysis

ols_eigen_cindex(model8)

Eigenvalue Condition Index intercept alcohol volatile.acidity residual.sugar density pH

1 8.315247e+00 1.000000 3.537025e-09 1.777351e-05 1.430277e-03 2.558972e-04 3.492778e-09 1.348453e-05

2 3.690609e-01 4.746665 9.306968e-09 1.560880e-04 5.539636e-03 5.709257e-02 8.568691e-09 4.528672e-05

3 1.570495e-01 7.276446 9.875552e-10 1.575757e-05 1.378255e-01 1.336825e-02 1.020923e-09 9.326872e-07

4 9.347380e-02 9.431758 4.900656e-08 2.506405e-04 7.710984e-01 3.886842e-03 4.854831e-08 1.930934e-04

5 4.264497e-02 13.963802 1.788470e-07 1.817729e-03 4.778225e-02 1.303962e-05 1.722615e-07 4.335438e-04

6 1.473845e-02 23.752628 8.470409e-08 2.058915e-02 1.163464e-02 7.321502e-03 6.453473e-08 3.672224e-03

7 6.827731e-03 34.897905 3.135815e-06 9.512796e-02 9.838378e-05 8.866288e-03 3.336307e-06 3.359017e-02

8 9.576600e-04 93.181972 5.885555e-05 3.225348e-03 4.755448e-03 5.488741e-03 5.542380e-05 5.166996e-01

9 1.226019e-07 8235.481815 9.999377e-01 8.787996e-01 1.983544e-02 9.037069e-01 9.999409e-01 4.453516e-01

sulphates free.sulfur.dioxide fixed.acidity

1 0.0006128013 1.935314e-03 6.839940e-05

2 0.0024016819 1.896133e-02 1.424653e-04

3 0.0003116584 7.090582e-01 8.522691e-05

4 0.0408009545 2.257400e-01 1.761735e-03

5 0.8064640683 3.373228e-05 7.613368e-03

6 0.0039214364 2.696801e-03 2.250236e-01

7 0.0339102584 3.929723e-02 3.824680e-02

8 0.0060502744 2.276147e-03 1.348229e-01

9 0.1055268664 1.220700e-06 5.922354e-01

ols_eigen_cindex(model9)

Eigenvalue Condition Index intercept alcohol volatile.acidity residual.sugar density pH

1 9.259451e+00 1.000000 2.649189e-09 1.401615e-05 0.001105614 0.0001988816 2.615949e-09 1.077802e-05

2 3.740020e-01 4.975717 1.080852e-08 1.759313e-04 0.006608066 0.0509431127 1.002817e-08 5.429877e-05

3 1.672041e-01 7.441649 4.328196e-09 5.021122e-05 0.104548945 0.0186837572 4.354693e-09 1.004265e-05

4 9.588089e-02 9.827127 5.345795e-08 3.358707e-04 0.747594190 0.0042767364 5.273916e-08 2.232251e-04

5 4.570300e-02 14.233781 1.292857e-07 2.683401e-03 0.001570396 0.0013878248 1.220267e-07 3.530482e-04

6 3.712236e-02 15.793371 3.665535e-08 1.050159e-04 0.086894245 0.0049355594 3.809037e-08 7.373275e-05

7 1.355394e-02 26.137248 1.594646e-07 1.305955e-02 0.031582803 0.0023824858 1.367405e-07 5.933648e-03

8 6.129691e-03 38.866311 3.266612e-06 1.144207e-01 0.006443995 0.0059666900 3.475692e-06 3.234516e-02

9 9.531134e-04 98.564458 5.432849e-05 2.172239e-03 0.006985892 0.0056546873 5.117190e-05 5.212650e-01

10 1.141304e-07 9007.245232 9.999420e-01 8.669831e-01 0.006665853 0.9055702647 9.999450e-01 4.397311e-01

sulphates free.sulfur.dioxide fixed.acidity total.sulfur.dioxide

1 4.923835e-04 0.001029690 5.446736e-05 0.0004155277

2 2.799043e-03 0.012651378 1.805173e-04 0.0013966278

3 3.141569e-06 0.374912480 1.834354e-04 0.0160359556

4 3.945154e-02 0.062940815 1.749852e-03 0.0125610770

5 4.637303e-01 0.156150249 2.469660e-03 0.2133883600

6 3.631501e-01 0.290394394 8.689941e-03 0.4719761131

7 2.179762e-02 0.067709522 2.469101e-01 0.0914955998

8 1.297028e-02 0.003875841 1.560446e-02 0.1189933949

9 4.743077e-03 0.006124446 1.379451e-01 0.0046393009

10 9.086257e-02 0.024211184 5.862124e-01 0.0690980433

#Creating models without density

model8b <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + pH + sulphates + free.sulfur.dioxide + fixed.acidity, data=winequality.white)

model9b <- lm(quality ~ alcohol + volatile.acidity + residual.sugar + pH + sulphates + free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide, data=winequality.white)

#VIF

vif(model8b)

alcohol volatile.acidity residual.sugar pH sulphates free.sulfur.dioxide

1.314242 1.027113 1.399655 1.297765 1.033551 1.161087

fixed.acidity

1.243051

vif(model9b)

alcohol volatile.acidity residual.sugar pH sulphates

1.464941 1.082747 1.433657 1.320540 1.052592

free.sulfur.dioxide fixed.acidity total.sulfur.dioxide

1.734096 1.275914 2.121530

#Eigensystem analysis

ols_eigen_cindex(model8b)

```
   Eigenvalue Condition Index    intercept     alcohol volatile.acidity residual.sugar          pH
sulphates
1 7.3253029221        1.000000 1.872939e-05 0.000185555     0.0018823567   3.512888e-03
3.161799e-05 0.0008819023
2 0.3636716698        4.488054 4.959326e-05 0.001490515     0.0075962706   6.025040e-01
1.043901e-04 0.0034543698
3 0.1569347310        6.832085 5.055208e-06 0.000149984     0.1465883867   1.380725e-01
2.615779e-06 0.0002667171
4 0.0915340864        8.945845 2.612674e-04 0.002776893     0.7636180470   4.066139e-02
4.597020e-04 0.0619038387
5 0.0411454055       13.342958 1.071827e-03 0.021675995     0.0601006686   8.171535e-08
1.227456e-03 0.8743868144
6 0.0146741555       22.342730 4.308830e-04 0.189755721     0.0129920339   8.304382e-02
7.101458e-03 0.0050139254
7 0.0060731183       34.730171 2.292723e-02 0.722720479     0.0006005886   7.503020e-02
1.104886e-01 0.0512412001
8 0.0006639114      105.040730 9.752354e-01 0.061244857     0.0066216479   5.717518e-02
8.805842e-01 0.0028512322
  free.sulfur.dioxide fixed.acidity
1        0.0025062493  0.0002182348
2        0.0161367130  0.0004709792
3        0.7030859405  0.0002345322
4        0.2294473513  0.0054637752
5        0.0000444301  0.0279563373
6        0.0023846703  0.5349068325
7        0.0414492925  0.0734501615
8        0.0049453530  0.3572991472
```

ols_eigen_cindex(model9b)

```
   Eigenvalue Condition Index    intercept     alcohol volatile.acidity residual.sugar          pH
sulphates
```

1 8.2719229314 1.000000 1.464757e-05 1.298187e-04 0.001394955 0.002702961 2.430594e-05 6.776828e-04

2 0.3671757601 4.746421 6.007786e-05 1.515037e-03 0.008698316 0.554114540 1.211504e-04 3.904853e-03

3 0.1666130968 7.046098 2.345304e-05 4.446213e-04 0.116331566 0.191991377 2.386361e-05 7.803037e-06

4 0.0935072006 9.405475 3.061570e-04 3.319951e-03 0.728863754 0.045982972 5.257813e-04 6.032012e-02

5 0.0444302834 13.644691 7.683129e-04 2.683281e-02 0.001693769 0.022872659 9.152163e-04 4.222713e-01

6 0.0368083442 14.990983 3.081147e-04 1.010854e-05 0.095558019 0.045454878 2.919881e-04 4.607684e-01

7 0.0134304868 24.817462 8.847315e-04 1.222723e-01 0.035229705 0.030117508 1.163956e-02 2.654124e-02

8 0.0054491670 38.961725 2.563183e-02 7.975579e-01 0.004313540 0.048175652 1.112598e-01 2.326655e-02

9 0.0006627297 111.721054 9.720027e-01 4.791738e-02 0.007916377 0.058587453 8.751983e-01 2.242037e-03

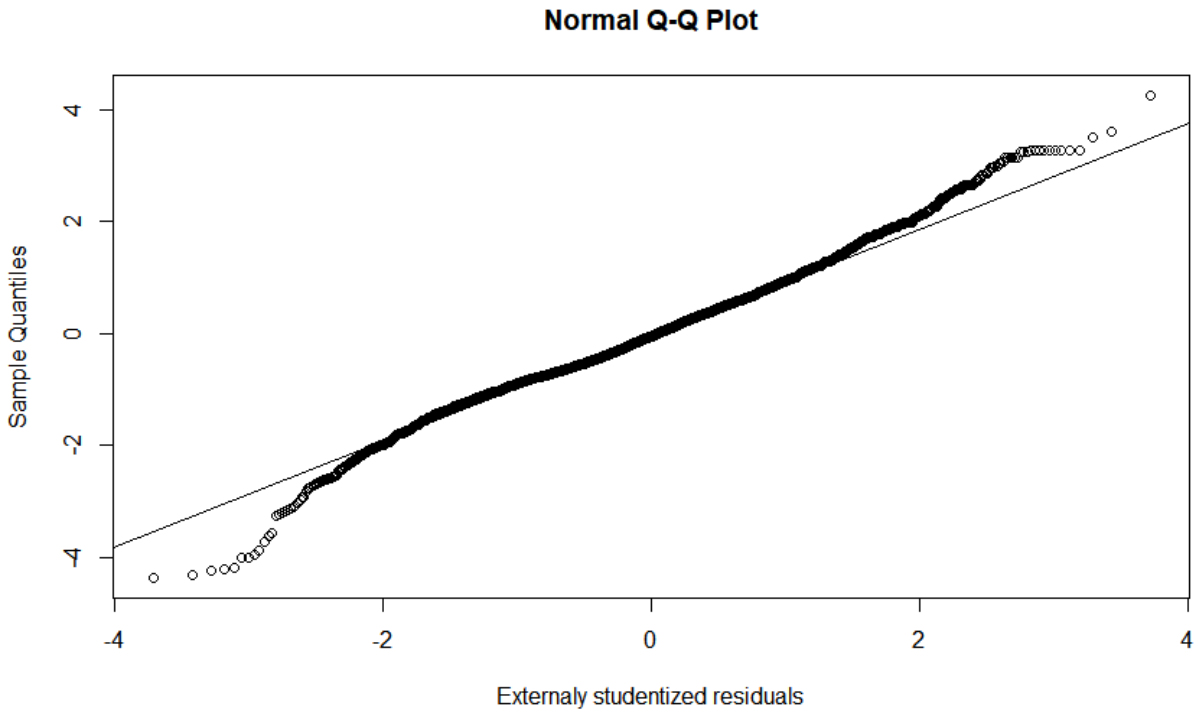
	free.sulfur.dioxide	fixed.acidity	total.sulfur.dioxide
1	0.001328955	0.0001663622	0.000561016
2	0.010895405	0.0005671099	0.001086921
3	0.380969715	0.0005179205	0.016302579
4	0.073140308	0.0054721658	0.011843202
5	0.185031655	0.0082103824	0.285542661
6	0.267686438	0.0281650065	0.452051229
7	0.071173348	0.5744171743	0.105336801
8	0.003048374	0.0269245825	0.125432191
9	0.006725803	0.3555592960	0.001843401

#Residuals after removing density from the models

```
stres8b <- rstudent(model8b)
```

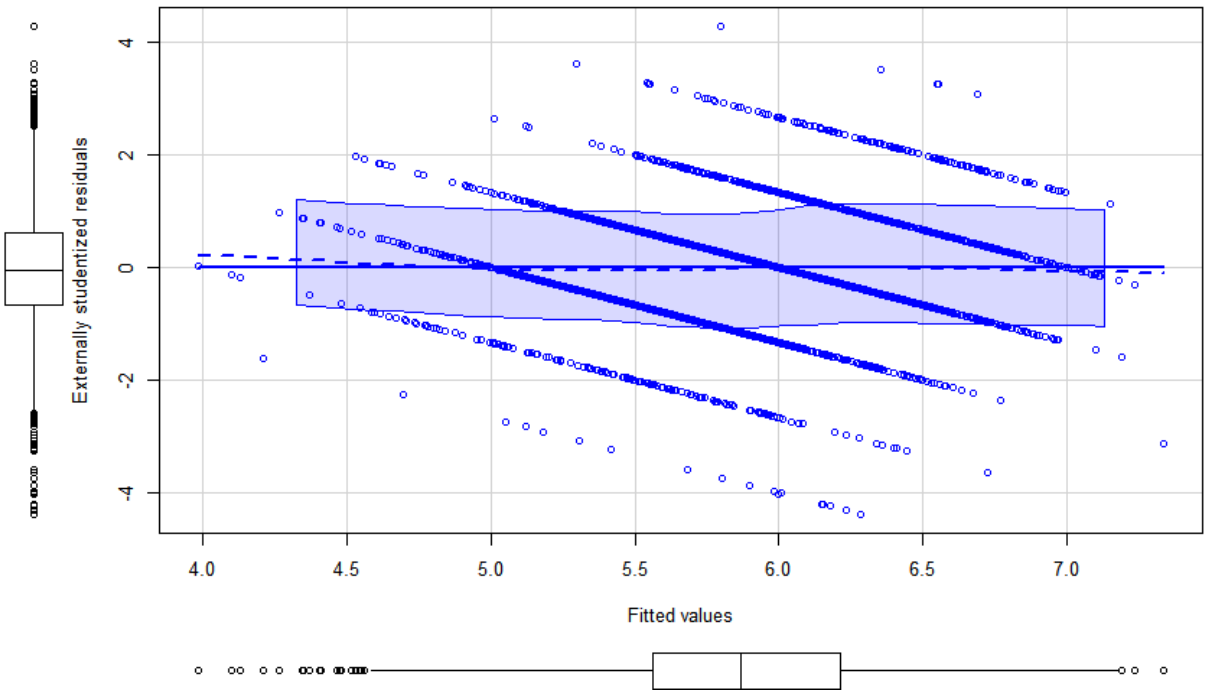
```
fv8b <- fitted.values(model8b)


stres9b <- rstudent(model9b)
fv9b <- fitted.values(model9b)


qqnorm(stres8b,
    xlab = "Externaly studentized residuals")
qqline(stres8b)
```

Figure 17



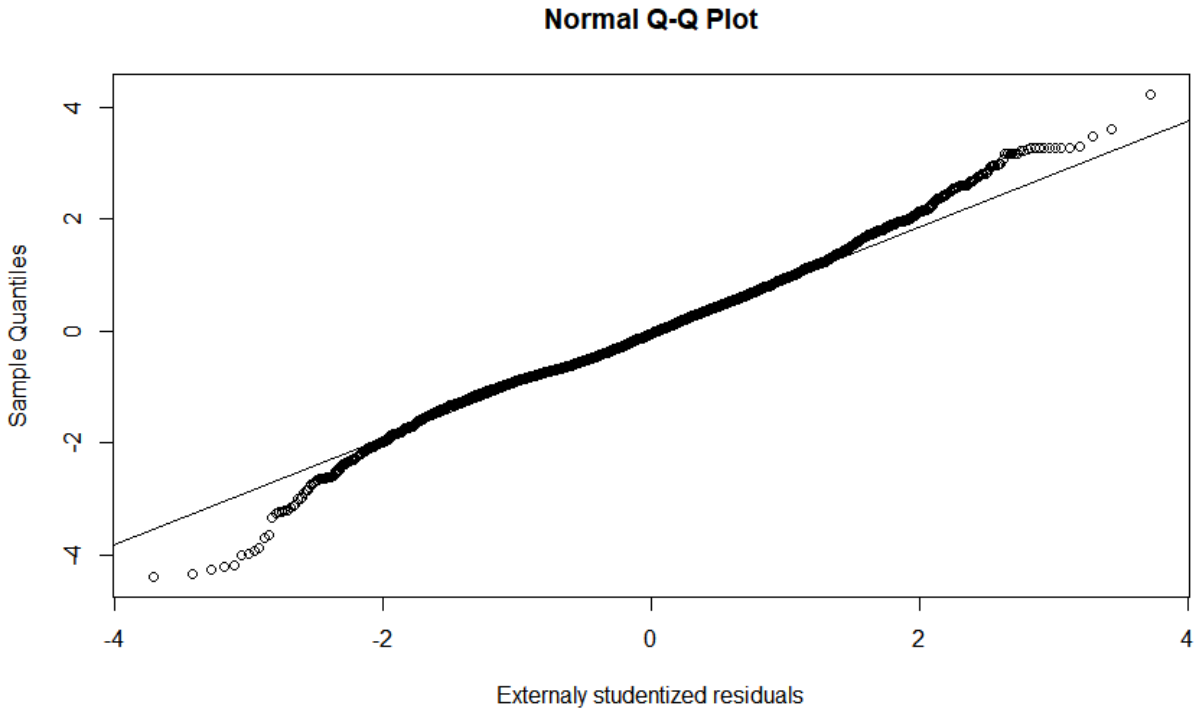

```
scatterplot(fv8b, stres8b,
        xlab = "Fitted values",
        ylab = "Externally studentized residuals")
```

Figure 18



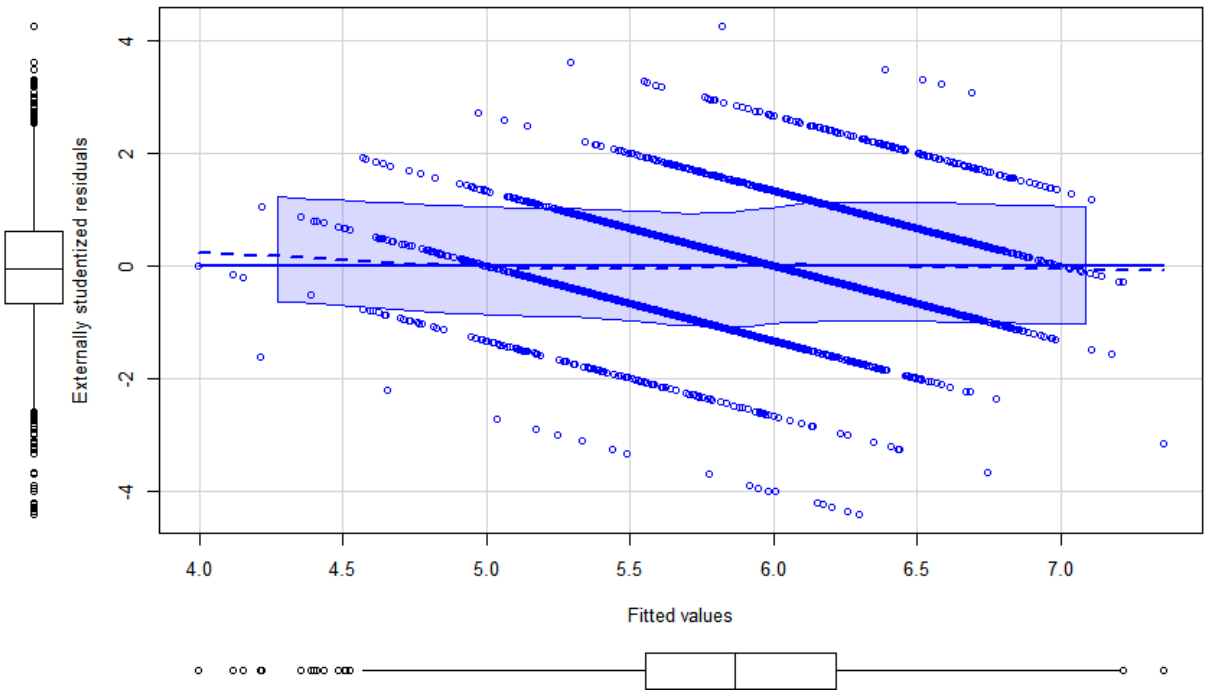

```
qqnorm(stres9b,
    xlab = "Externaly studentized residuals")
qqline(stres9b)
```

Figure 19




```
scatterplot(fv9b, stres9b,
        xlab = "Fitted values",
```

Figure 20



#Means of residuals

mean(model8b$residuals)

[1] -7.871087e-17

[1] -9.009259e-17

#Press and adjusted R^2 of models without density

press(model8b, as.R2 = FALSE)

[1] 2781.501

press(model9b, as.R2 = FALSE)

[1] 2778.7

summary(model8b)$adj.r.squared

```
[1] 0.2742739

summary(model9b)$adj.r.squared

[1] 0.2751827
```

#Comparing the two models

```
comparison <- anova(model8b, model9b)
comparison

Analysis of Variance Table

Model 1: quality ~ alcohol + volatile.acidity + residual.sugar + pH +
    sulphates + free.sulfur.dioxide + fixed.acidity
Model 2: quality ~ alcohol + volatile.acidity + residual.sugar + pH +
    sulphates + free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide
  Res.Df    RSS Df Sum of Sq     F    Pr(>F)
1   4887 2771.5
2   4886 2767.5  1    4.0373 7.128 0.007614 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

#Investigation of the model and regression coefficients

```
summary(model9b)
Call:
lm(formula = quality ~ alcohol + volatile.acidity + residual.sugar +
    pH + sulphates + free.sulfur.dioxide + fixed.acidity + total.sulfur.dioxide,
    data = winequality.white)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2995 -0.5002 -0.0366  0.4606  3.1809

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          1.7967566  0.3373143   5.327 1.05e-07 ***
alcohol              0.3713293  0.0105787  35.102  < 2e-16 ***
volatile.acidity    -1.9477040  0.1115584 -17.459  < 2e-16 ***
residual.sugar       0.0255348  0.0025754   9.915  < 2e-16 ***
pH                   0.1860988  0.0819145   2.272  0.02314 *
sulphates            0.4132175  0.0967385   4.271 1.98e-05 ***
free.sulfur.dioxide  0.0059497  0.0008551   6.958 3.90e-12 ***
fixed.acidity       -0.0454648  0.0144122  -3.155  0.00162 **
total.sulfur.dioxide -0.0009895  0.0003706  -2.670  0.00761 **
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7526 on 4886 degrees of freedom
Multiple R-squared:  0.2764,  Adjusted R-squared:  0.2752
F-statistic: 233.3 on 8 and 4886 DF,  p-value: < 2.2e-16
```

#Standardized regression coefficients

```
betas <- lm.beta(model9b)
betas
          alcohol     volatile.acidity       residual.sugar                   pH           sulphates
       0.51703330         -0.22108870          0.14447684          0.03177176          0.05333244
 free.sulfur.dioxide       fixed.acidity total.sulfur.dioxide
       0.11151149         -0.04336491         -0.04732496
```

#Frequencies of grades of quality

```
table(winequality.white$quality)

  3   4   5   6   7   8   9
 18 163 1457 2197 880 175   5
```