Analysis about Red Wine Quality

Yuhan Zhang

Shiyi Zhou

Contents

Abstract

In this project, we analyze the red wine data set from the UCI Machine Learning Repository and aim to determine the factors responsible for the overall quality of red wine. The dataset includes information on the chemical properties of different types of wine. Total 1,599 observations are presented in the data set, and twelve kinds of chemical properties such as fixed acidity, residual sugar, free sulfur dioxide, alcohol, and density are listed. Most of those variables are continuous numerical independent variables, which are physicochemical attributes, some of which may be correlated. The quality of red wine is our response variable, which is sensory data on a 1-to-10 scale. We will analyze distribution and collinearity between those variables how they relate to overall quality.

Keywords: red wine, overall quality of red wine

Introduction

Red wine quality is the result of a complex set of factors and interactions, which might include geological and soil quality, climate change, and viticultural decisions. Red wine quality and style are highly influenced by the qualitative and quantitative composition of aromatic compounds having various chemical structures and properties and their interaction within different red wine matrices. The understanding of interactions between the wine matrix and volatile compounds, the density of water used, and the quantity of acid and sugar is getting increasingly important for determining the quality rank of red wine. In this paper, we use statistical methods to build regression models helping people understand the factors that play more prominent roles in determining red wine quality and present several models from perspectives as a conclusion to describe the relationship between red wine quality and those factors.

Data Information

explanatory variables:

1 - fixed acidity: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) (tartaric acid - g / dm^3)

2 - volatile acidity: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant vinegar taste (acetic acid - g / dm^3)

3 - citric acid: found in small quantities, citric acid can add ‘freshness’ and flavour to wines (g / dm^3)

4 - residual sugar: the amount of sugar remaining after fermentation stops; it is rare to find wines with less than 1 gram/litre, and wines with greater than 45 grams/litre are considered sweet (g / dm^3)

5 - chlorides: the amount of salt in the wine (sodium chloride - g / dm^3

6 - free sulfur dioxide: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfide ion; it prevents microbial growth and the oxidation of wine (mg / dm^3)

7 - total sulfur dioxide: the amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine (mg / dm^3)

8 - density: the density of water is close to that of water depending on the percent alcohol and sugar content (g / cm^3)

9 - pH: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10 - sulphates: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant (potassium sulphate - g / dm3)

11 - alcohol: the percent alcohol content of the wine (% by volume)

Response variable:

quality: based on sensory data score between 0 and 10.

New variable:

sulphates & alcohol: two variables with the most considerable correlation to quality (response variable).

Methods

In this paper, we aim to find the most appropriate linear regression model to describe the relationship between the quality of red wine and the factors provided in the dataset.

Firstly, we build histograms (figure6) to detect the distributions of all independent variables and compute the correlations between those variables with the response variable, quality. As we notice, the correlations between particular variables and the response variable are too small, so we use data transformation to make the data more related and distributed normally (figure10). We also analyze variance inflation factors and diagnostic graphs of our data set.

Moreover, as listed below, we use model selection methods to build six linear regression models to describe the data.

Model 0: FULL MODEL Combination of all given variables

M0:β0 + β1*fixed acidity +β2* volatile acidity +β3*citric acid +β4*residual sugar+β5*chlorides +β6*free sulfur dioxide +β7*total sulfur dioxide +β8*density +β9*PH +β10*sulphates +β11*alcohol

Model1: using the stepwise method to select the FULL MODEL (Model 0).

M1: β0 +β2* volatile acidity + β5*chlorides + β7*total sulfur dioxide + β10*sulphates +β11*alcohol

Model2: FULL MODEL (Model 0) after data transformation for some significant variables

M2: β0 + β1*fixed acidity +β2* volatile acidity +β3*citric acid +β4*log(residual sugar) +β5*log(chlorides) +β6*free sulfur dioxide +β7*total sulfur dioxide +β8*density +β9*PH +β10*log(sulphates) +β11*alcohol

Model3: using the stepwise method to select Model 2

M3: β0 + β2* volatile acidity +β5*log(chlorides) +β7*total sulfur dioxide +β10*log(sulphates) +β11*alcohol

Model fit: using the stepwise method (both direction) to select Model 2

Mfit: β0 + β2* volatile acidity +β5*log(chlorides) + β7*total sulfur dioxide

+ β6*free sulfur dioxide + β9*PH+β10*log(sulphates) +β11*alcohol

Model4: FULL MODEL + log (volatile acidity & total sulfur acidity) + log (sulphates & alcohol) + (volatile acidity & alcohol)

M4: β0 + β1*fixed acidity +β2* volatile acidity +β3*citric acid +β4*residual sugar

+ β5*chlorides +β6*free sulfur dioxide +β7*total sulfur dioxide +β8*density +β9*PH

+ β10*sulphates +β11*alcohol +β12* log (volatile acidity & total sulfur acidity)

+ β13* log (sulphates & alcohol)+β14*(volatile acidity & alcohol)

Model 5: using the stepwise method (both direction) to select Model 4

M5: β0 +β2* volatile acidity+β5*chlorides +β7*total sulfur dioxide +β8*density

+β9*PH + β10*sulphates +β12* log (volatile acidity & total sulfur acidity)

+β13* log (sulphates & alcohol)

After building those modules, we use robust regression and cross-validation to check the models listed above. Under the above models, we make the following **assumptions**:

- The relationship between the independent variables and the response variable is approximately linear.
- The mean of the error term is zero.
- The variance of the error term is constant.
- The errors are uncorrelated and normally distributed.

Figure1 (diagnostic analysis of data)

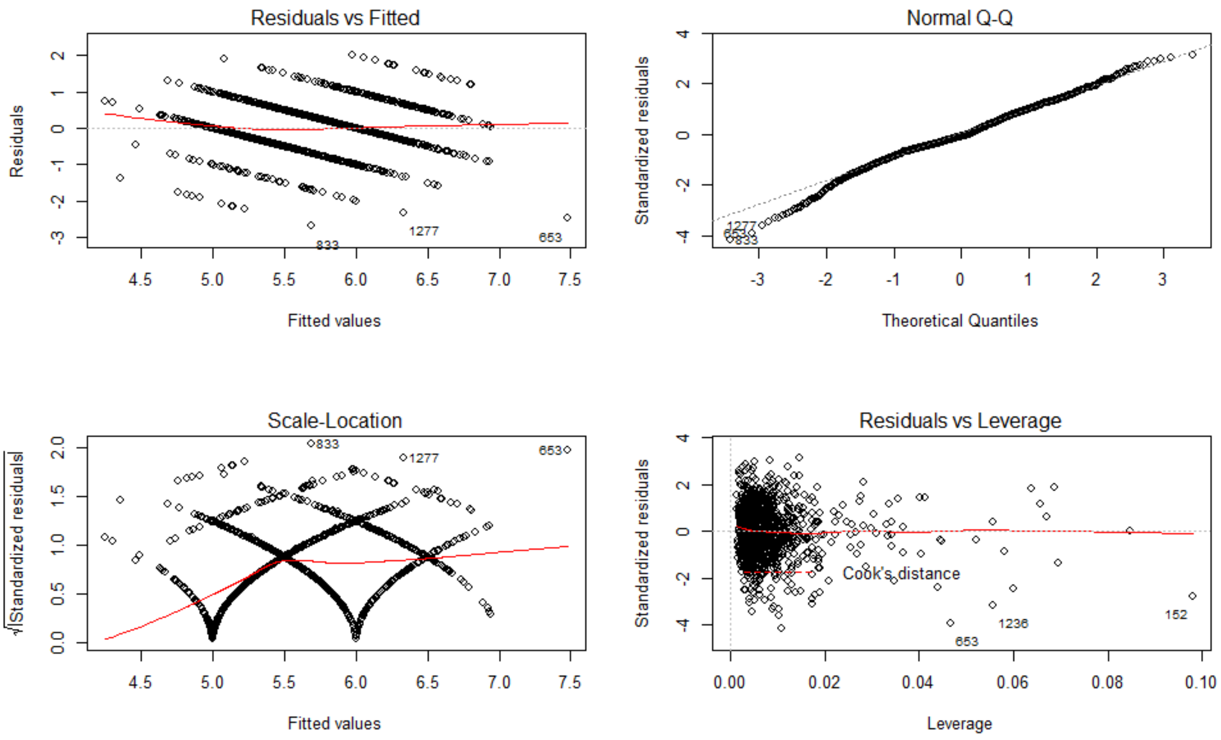
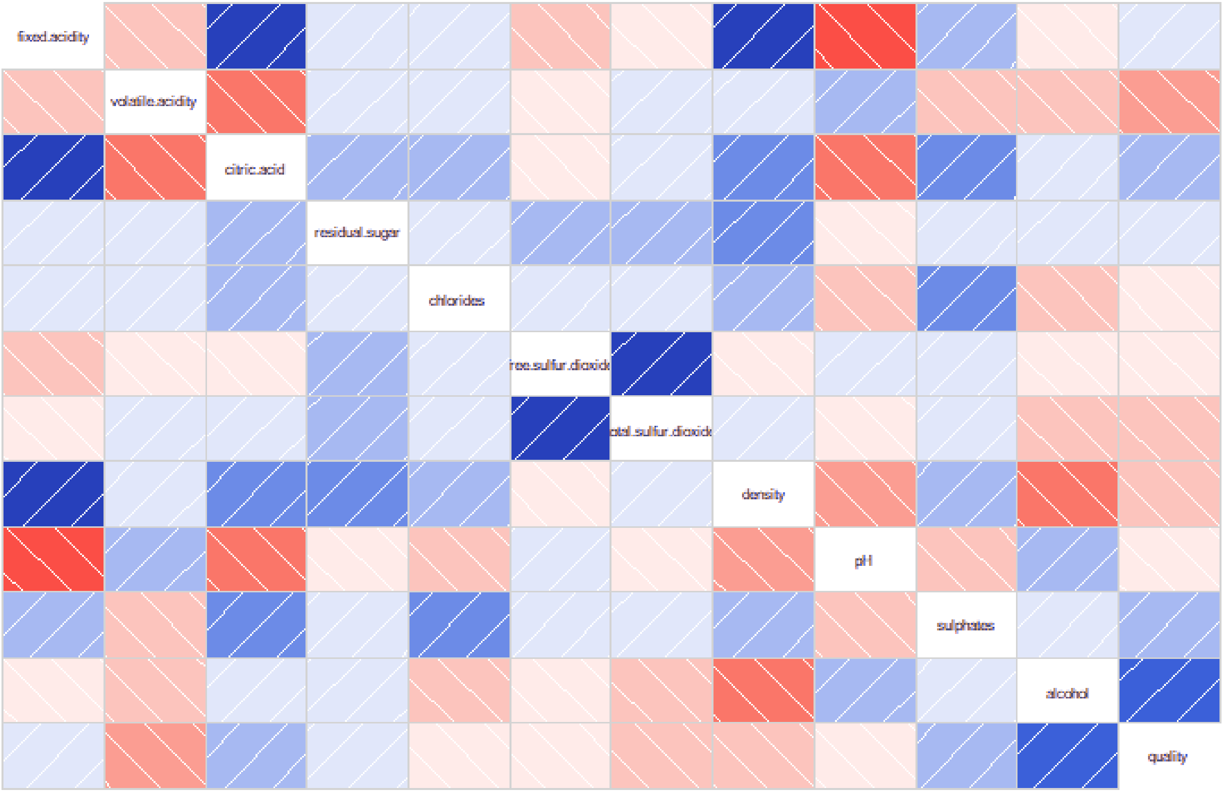


Figure 1

Residual vs Fitted plot shows data are distributed linearly as residuals are distributed around a horizontal line without a distinct pattern. QQ plot shows residual fit the line; as a result, we can conclude that residuals distributed normally. As there is a line with approximately equally spread points, the assumption of equal variance is met. As we can barely see the cook's distance lines, all cases are well inside the coke's distance, and there is no influential case.

Result

Figure2 (correlation plot of data)



From the correlation graph above, **alcohol** is strongly positively correlated with red wine quality, and **volatile acidity** is strongly negatively correlated with red wine quality. PH, chlorides, free sulfur dioxide, fixed acidity, and residual sugar do not significantly impact the quality of red wine. To be more specific, we can quickly notice except alcohol, sulphates and critical acid also have a relatively strong positive correlation with quality. Fixed acidity and residual sugar have a weaker but positive correlation with quality. Except for volatile acidity, density and total sulfur dioxide have a negative correlation with wine quality. PH, free sulfur dioxide, and chlorides play a smaller role in determining red wine quality.

Model 0 (FULL Model)

We construct model 0 with 12 given variables and then get the following summary table.

The full model can only explain 35.61% of variability of the data set.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.4315	0.3561	81.35	0.420	0.5910	3164.277	0.000365

Model 1

After using the stepwise method to select the FULL MODEL (Model 0), both correlation and AIC increases. Interestingly, R-squared decreases a little bit which might be the result of decreasing of variables included. Therefore, we conclude model 1 as a better model than model 0.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.4259	0.3495	172.7	0.424	0.5965	3174.767	0.0003742

Model 2

After doing data transformation for some variables with slighter correlation in model 0 with quality, our response variable, we get an even better model with smaller mean MSPE, smaller residuals, smaller AIC, smallest MSE, bigger R-squared, and bigger correlation.

Therefore, we classify model 2 as one of the ideal models we got from this research.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.4213	0.3659	174.84	0.414	0.6023	3139.716	0.0003641

Model3

In this model, we use backward selection to deal with variables in model 2. As we also use both forward and backward selection method, the analysis is presented in model fit part.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.5023	0.3575	178.9	0.419	0.4899	3154.782	0.0004366

Model fit

Compared with Model 3, model fit using" both direction" model selection method does a better job with smaller mean MSPE, ANOVA residual, and MSE and larger R-squared and correlation.

Therefore, to get a more desire from model selection, model fit stands out.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.4158	0.368	132.4	0.414	0.5981	3137.44	0.0003685

Model4

We are also interested in the multivariate relationships with the interaction of certain variables. In model 4, we add the interaction of variables with the strongest correlation with our response variable, quality and do data transformation on the new added variables. All statistical factors in this model become much better than others. We get the smallest mean MSPE, smallest residual, small MSE, biggest R-squared, biggest correlation, and biggest AIC in this model.

Therefore, we select model 4 as the other good model we got from this research.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.4108	0.3804	176.46	0.404	0.6149	3104.896	0.0003648

Model 5

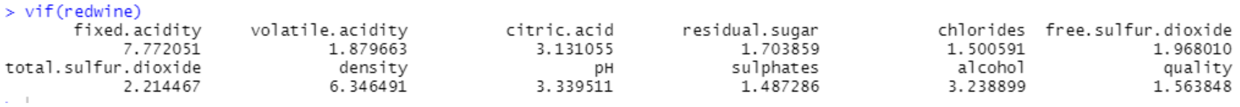
As model 4 is one of our best models, we continue using stepwise methods to select variables in mode 4 trying to make an even better model. Most data of model 5 become better, such as the smaller mean MSPE and AIC and bigger r^2 and correlation.

Therefore, after doing model selection, model 5 is our best model.

Mean MSPE	Adjusted R-squared	F-statistic	ANOVA (MSRes)	Correlation Of Quality	AIC	MSE for robust regression
0.4091	0.3804	110.4	0.404	0.6264	3100.4	0.000371

Final model analysis

Figure3 (variance inflation factors of data)



```
> vif(redwine)
       fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides  free.sulfur.dioxide
            7.772051             1.879663             3.131055             1.703859             1.500591             1.968010
total.sulfur.dioxide              density                   pH            sulphates              alcohol              quality
            2.214467             6.346491             3.339511             1.487286             3.238899             1.563848
```

As the variance inflation factors of the data are all between 1 and 10, there is no sign of multicollinearity in our model. Therefore, the assumption of multivariate distribution is met.

Figure4 (diagnostic analysis of model 5)

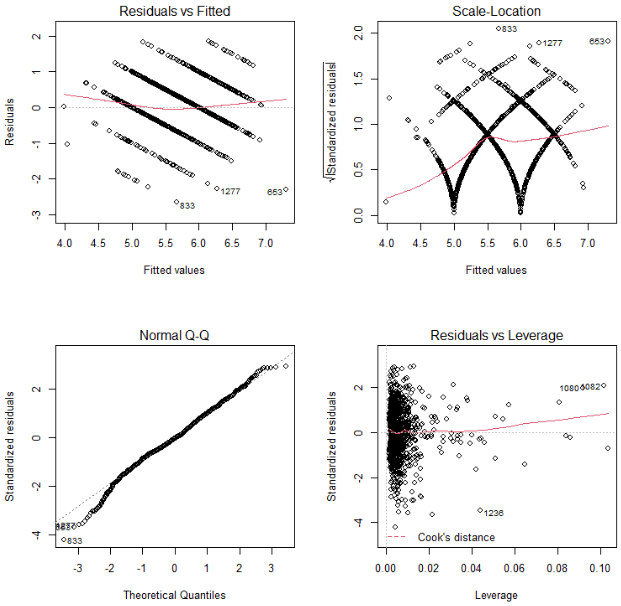
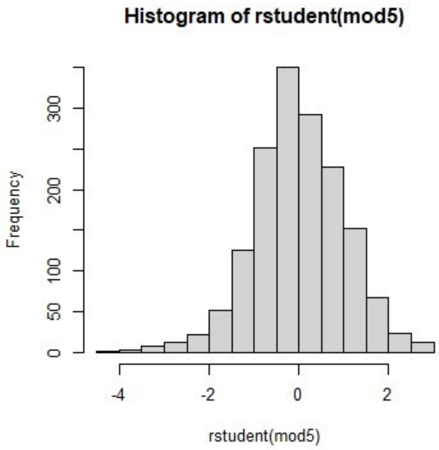


Figure5 (studentized R of model 5)



Based on the QQ plot and the distribution of the studentized residual, we conclude that the assumption of a normally distributed error term is apparently respected. Although there are some outliers, it is still insufficient to violate the assumption. The residual VS. fitted plot indicates that it is appropriate to assume the error term's mean is 0 and data is distributed linearly. Besides, the scale-location plot also shows the approximately equal variance of the error term. As shown in the histogram, some variables are not distributed linearly. To solve this problem, we did data transformation, such as log(x), to make those variables look better and be more correlated with response variable, quality

The model we suggest using is model 5 as analyzed in the method section. By robust regression, the final model we get is : Quality= 69.758-0.6360* volatile acidity-2.461*chlorides -0.00654*total sulfur dioxide -64.861*density -0.7418*PH +-2.238*sulphates +0.112* log (volatile acidity & total sulfur acidity) +2.601* log (sulphates & alcohol).

Conclusion

As the introduction says, our primary goal was to find the main factors which are closely linked with the red wine quality. From viewing the raw data, the quality is very centralized at [5,6], and its variance inflation factors of all variables are reasonable in between [1,10] for the multicollinearity among all the variables. First, we plot the residuals for the full linear model, which we constructed for all given variables. Then the plots show the model is not adequate with respect to the linear relationship. Since the given parameters are not all continuous variables, the data transformation method is used for processing the red wine data to be more linear. Second, we set up the new data points as dummy variables to help improve the model build. According to the model results, model_4 is most considerable from all its values: mean MSPE, adjusted R-squared, correlation with the response variable quality and Akaike Information Criterion.

From the above analysis in model_0 to model_4, we can tell that these five parameters: volatile acidity, chlorides, total sulfur dioxide, sulphates and alcohol which have a significant correlation with our analysis about red wine quality, therefore good quality wines have a dash of volatile acid, less total sulfur dioxide and sulphates concentration, and also with higher alcohol content; whereas the other parameters are not much bound up the quality of wine and hence will not be suitable for analysis. In conclusion, the relationship between all the independent variables and the response variable quality is linear, and the data is normally distributed.

Appendix

Find full R code on the following GitHub links:

https://github.com/carolzhangyh/stat350

https://github.com/litost11/finalproject350

figure6 (histogram of data)

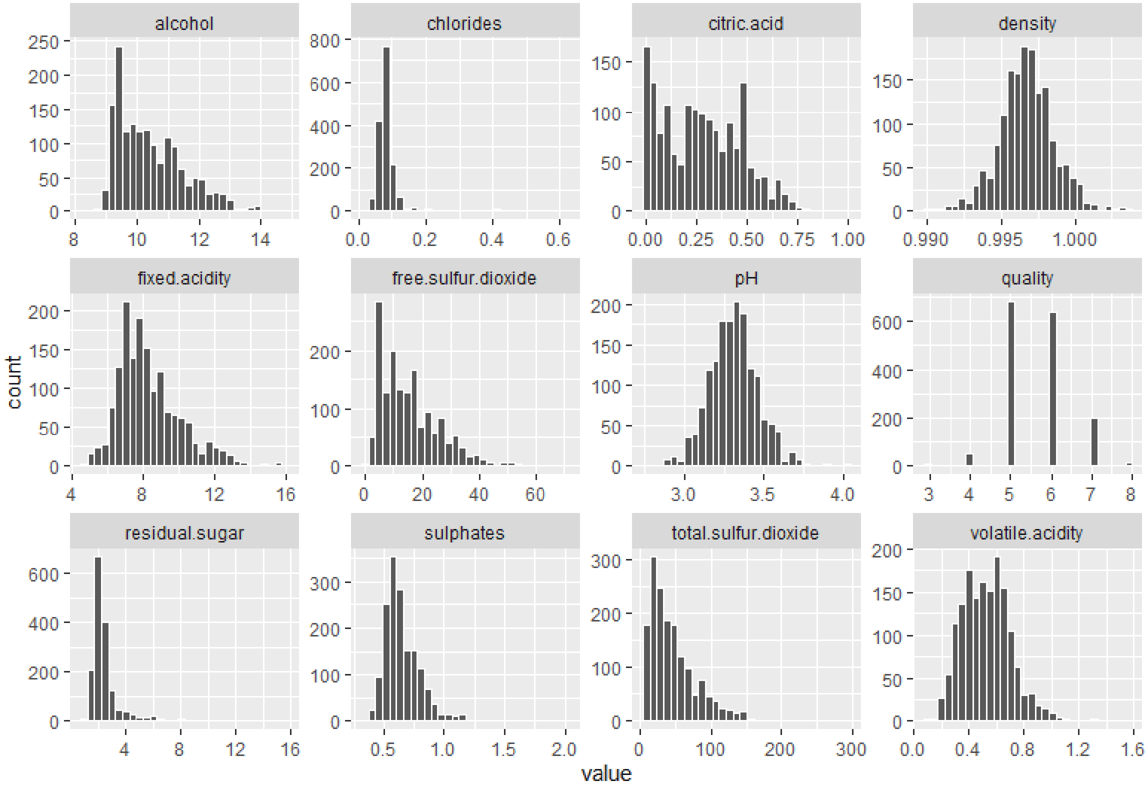


figure7 (residual plot for models)

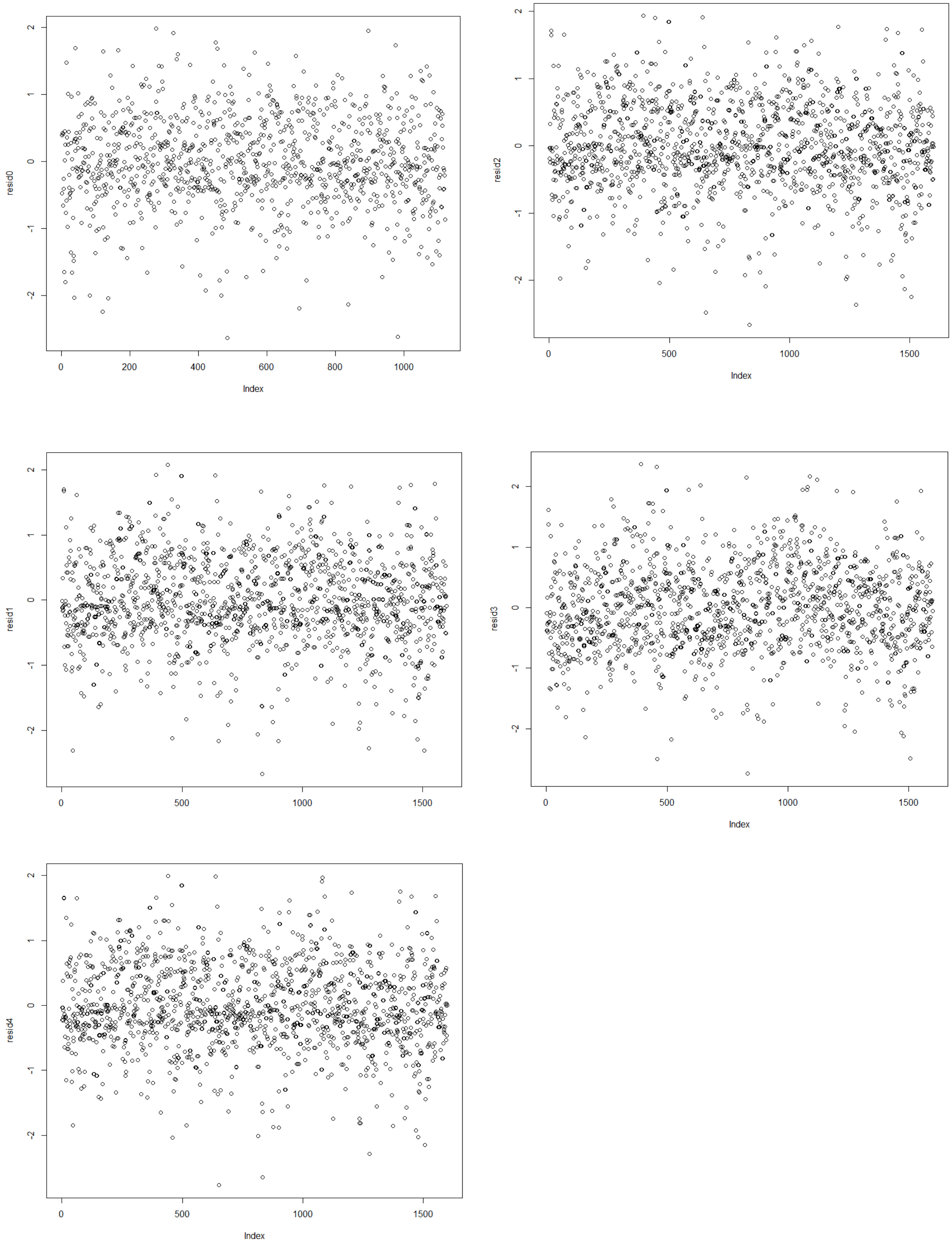


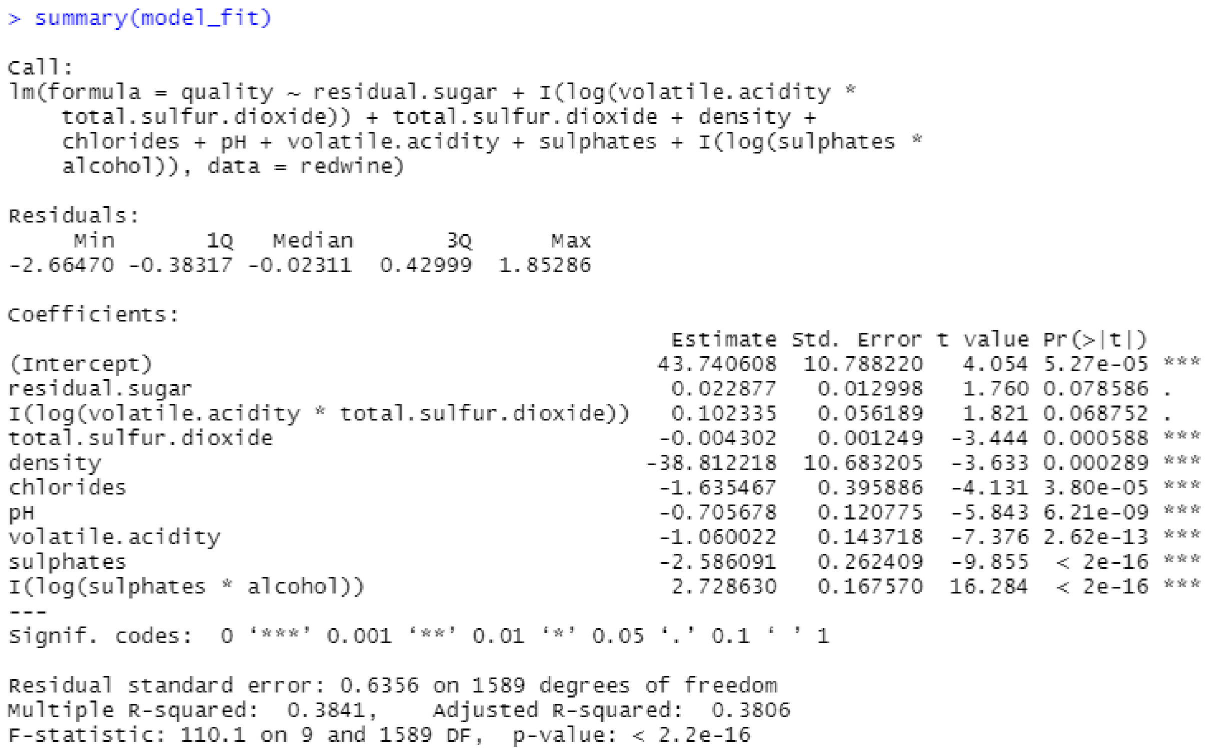
figure8 (summary for models)

```
> summary(model_0)

Call:
lm(formula = quality ~ ., data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.68911 -0.36652 -0.04699  0.45202  2.02498 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           2.197e+01  2.119e+01   1.036   0.3002    
fixed.acidity         2.499e-02  2.595e-02   0.963   0.3357    
volatile.acidity     -1.084e+00  1.211e-01  -8.948  < 2e-16 ***
citric.acid          -1.826e-01  1.472e-01  -1.240   0.2150    
residual.sugar        1.633e-02  1.500e-02   1.089   0.2765    
chlorides            -1.874e+00  4.193e-01  -4.470 8.37e-06 ***
free.sulfur.dioxide   4.361e-03  2.171e-03   2.009   0.0447 *  
total.sulfur.dioxide -3.265e-03  7.287e-04  -4.480 8.00e-06 ***
density              -1.788e+01  2.163e+01  -0.827   0.4086    
pH                   -4.137e-01  1.916e-01  -2.159   0.0310 *  
sulphates             9.163e-01  1.143e-01   8.014 2.13e-15 ***
alcohol               2.762e-01  2.648e-02  10.429  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.648 on 1587 degrees of freedom
Multiple R-squared:  0.3606,	Adjusted R-squared:  0.3561 
F-statistic: 81.35 on 11 and 1587 DF,  p-value: < 2.2e-16
```

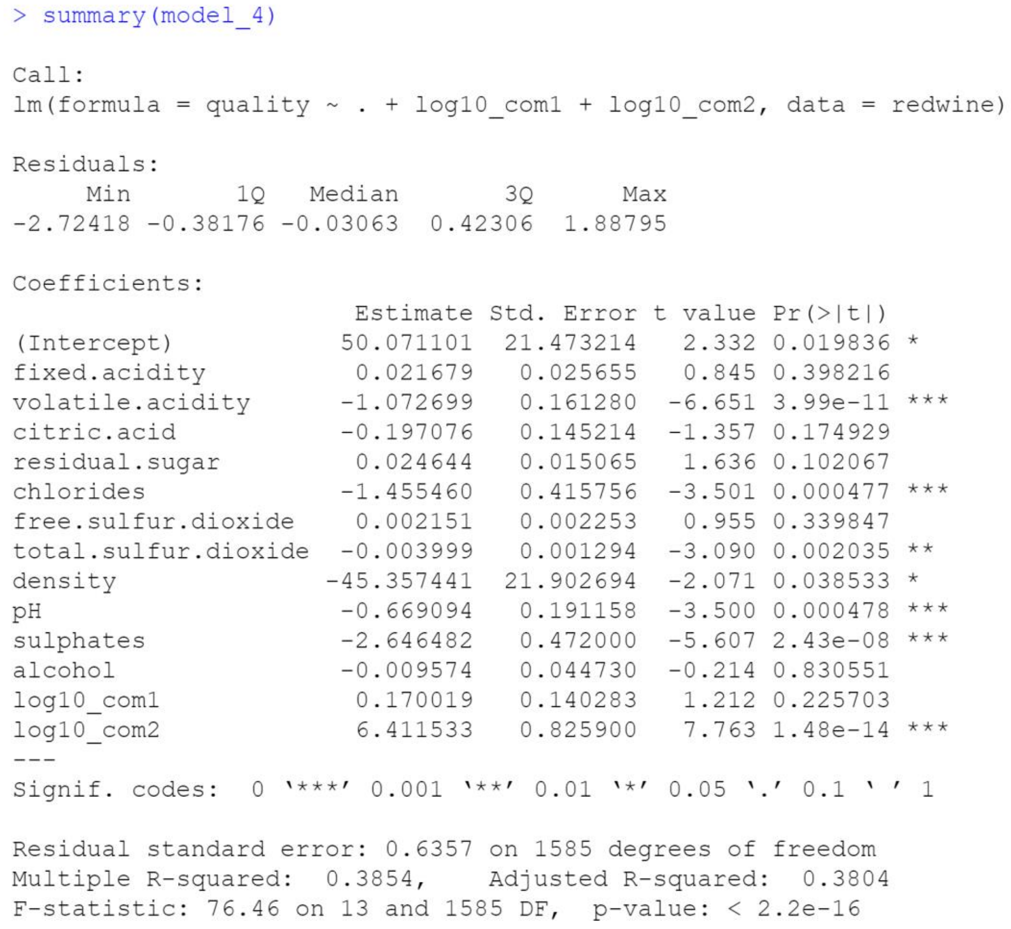
```
> summary(model_1)

Call:
lm(formula = quality ~ volatile.acidity + chlorides + total.sulfur.dioxide + 
    sulphates + alcohol, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.67443 -0.38254 -0.06368  0.44893  2.07310 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           3.0048920  0.2037663  14.747  < 2e-16 ***
volatile.acidity     -1.1419024  0.0969400 -11.779  < 2e-16 ***
chlorides            -1.7047871  0.3916886  -4.352 1.43e-05 ***
total.sulfur.dioxide -0.0023096  0.0005082  -4.544 5.92e-06 ***
sulphates             0.9148320  0.1102702   8.296 2.26e-16 ***
alcohol               0.2770979  0.0164836  16.811  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6514 on 1593 degrees of freedom
Multiple R-squared:  0.3515,	Adjusted R-squared:  0.3495 
F-statistic: 172.7 on 5 and 1593 DF,  p-value: < 2.2e-16
```

```
> summary(model_2)

Call:
lm(formula = quality ~ fixed.acidity + volatile.acidity + citric.acid +
    log10_rs + log10_chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
    density + pH + log10_sulphates + alcohol, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66553 -0.37063 -0.03908  0.43709  1.93395

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.287e+01  2.309e+01   1.424   0.1547
fixed.acidity        3.545e-02  2.596e-02   1.366   0.1722
volatile.acidity    -1.043e+00  1.201e-01  -8.689  < 2e-16 ***
citric.acid         -2.662e-01  1.434e-01  -1.856   0.0636 .
log10_rs             2.286e-01  1.477e-01   1.548   0.1219
log10_chlorides     -5.715e-01  1.362e-01  -4.196 2.87e-05 ***
free.sulfur.dioxide  3.803e-03  2.145e-03   1.773   0.0764 .
total.sulfur.dioxide -3.097e-03  7.227e-04  -4.286 1.93e-05 ***
density             -2.854e+01  2.347e+01  -1.216   0.2241
pH                  -4.223e-01  1.904e-01  -2.218   0.0267 *
log10_sulphates      1.833e+00  1.950e-01   9.400  < 2e-16 ***
alcohol              2.563e-01  2.828e-02   9.063  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6431 on 1587 degrees of freedom
Multiple R-squared:  0.3703,    Adjusted R-squared:  0.3659
F-statistic: 84.84 on 11 and 1587 DF,  p-value: < 2.2e-16
```

```
> summary(model_3)

Call:
lm(formula = quality ~ volatile.acidity + log10_chlorides + total.sulfur.dioxide
 +
    log10_sulphates + alcohol, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68642 -0.37338 -0.05292  0.43728  2.04034

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.3091377  0.2047212  16.164  < 2e-16 ***
volatile.acidity     -1.0764424  0.0974446 -11.047  < 2e-16 ***
log10_chlorides      -0.5056117  0.1268607  -3.986 7.04e-05 ***
total.sulfur.dioxide -0.0022540  0.0005043  -4.470 8.37e-06 ***
log10_sulphates       1.7633324  0.1860932   9.476  < 2e-16 ***
alcohol               0.2677367  0.0168023  15.935  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6473 on 1593 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3575
F-statistic: 178.9 on 5 and 1593 DF,  p-value: < 2.2e-16
```

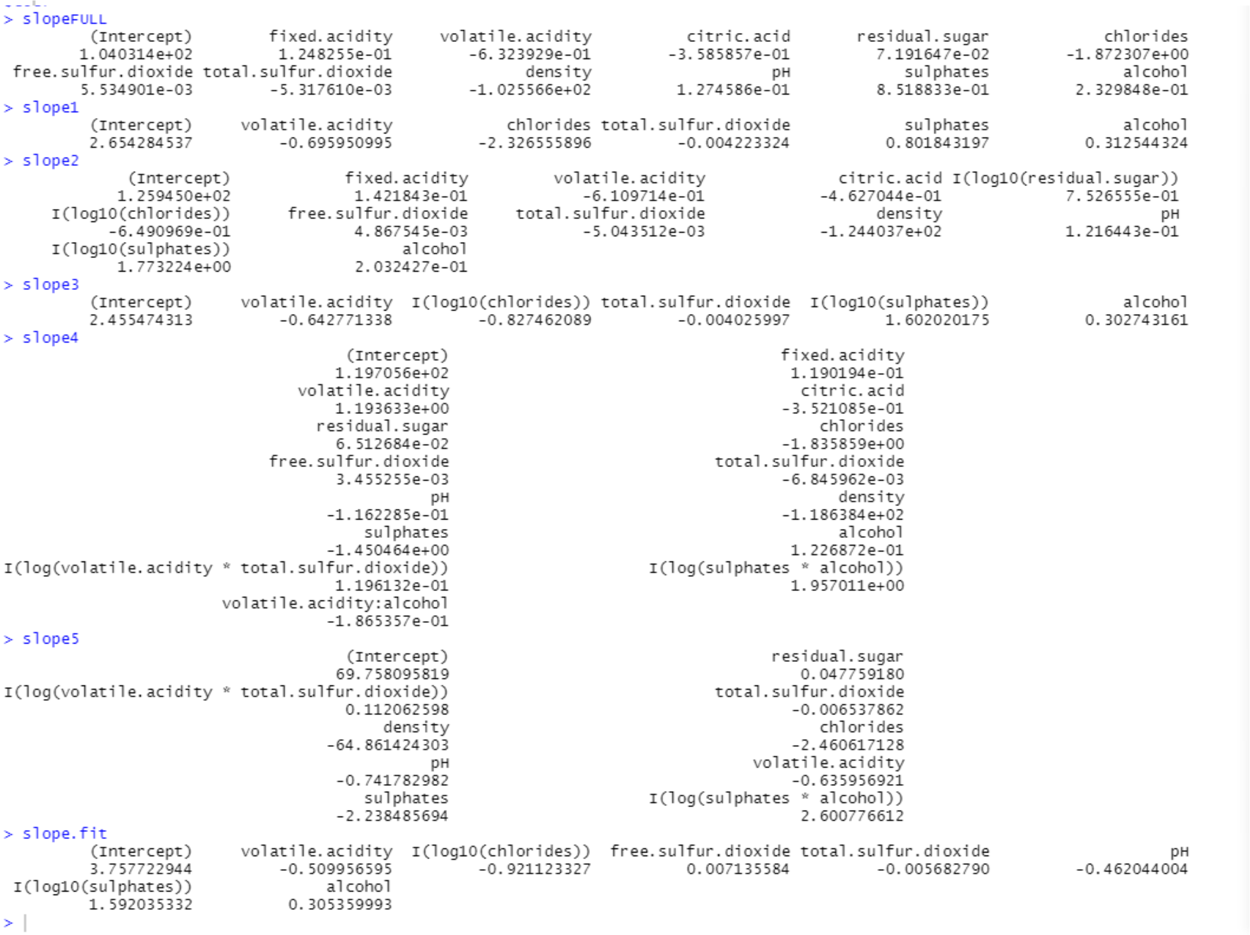


```
> summary(model_fit)

Call:
lm(formula = quality ~ residual.sugar + I(log(volatile.acidity *
    total.sulfur.dioxide)) + total.sulfur.dioxide + density +
    chlorides + pH + volatile.acidity + sulphates + I(log(sulphates *
    alcohol)), data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66470 -0.38317 -0.02311  0.42999  1.85286

Coefficients:
                                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                                  43.740608  10.788220   4.054 5.27e-05 ***
residual.sugar                                0.022877   0.012998   1.760 0.078586 .
I(log(volatile.acidity * total.sulfur.dioxide))  0.102335   0.056189   1.821 0.068752 .
total.sulfur.dioxide                         -0.004302   0.001249  -3.444 0.000588 ***
density                                     -38.812218  10.683205  -3.633 0.000289 ***
chlorides                                    -1.635467   0.395886  -4.131 3.80e-05 ***
pH                                           -0.705678   0.120775  -5.843 6.21e-09 ***
volatile.acidity                             -1.060022   0.143718  -7.376 2.62e-13 ***
sulphates                                    -2.586091   0.262409  -9.855  < 2e-16 ***
I(log(sulphates * alcohol))                   2.728630   0.167570  16.284  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6356 on 1589 degrees of freedom
Multiple R-squared:  0.3841,    Adjusted R-squared:  0.3806
F-statistic: 110.1 on 9 and 1589 DF,  p-value: < 2.2e-16
```



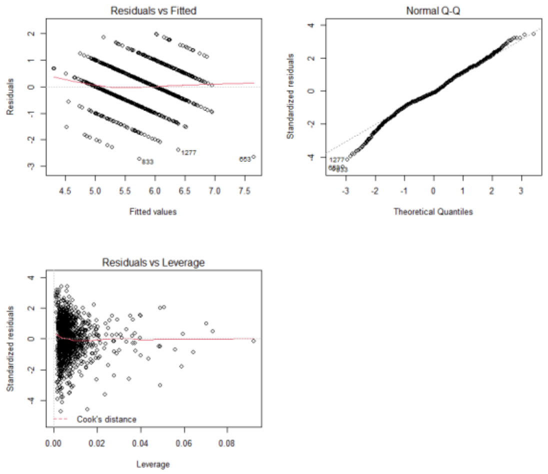
```
> summary(model_4)

Call:
lm(formula = quality ~ . + log10_com1 + log10_com2, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.72418 -0.38176 -0.03063  0.42306  1.88795

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           50.071101  21.473214   2.332 0.019836 *
fixed.acidity          0.021679   0.025655   0.845 0.398216
volatile.acidity      -1.072699   0.161280  -6.651 3.99e-11 ***
citric.acid           -0.197076   0.145214  -1.357 0.174929
residual.sugar         0.024644   0.015065   1.636 0.102067
chlorides             -1.455460   0.415756  -3.501 0.000477 ***
free.sulfur.dioxide    0.002151   0.002253   0.955 0.339847
total.sulfur.dioxide  -0.003999   0.001294  -3.090 0.002035 **
density              -45.357441  21.902694  -2.071 0.038533 *
pH                    -0.669094   0.191158  -3.500 0.000478 ***
sulphates             -2.646482   0.472000  -5.607 2.43e-08 ***
alcohol               -0.009574   0.044730  -0.214 0.830551
log10_com1             0.170019   0.140283   1.212 0.225703
log10_com2             6.411533   0.825900   7.763 1.48e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6357 on 1585 degrees of freedom
Multiple R-squared:  0.3854,    Adjusted R-squared:  0.3804
F-statistic: 76.46 on 13 and 1585 DF,  p-value: < 2.2e-16
```

```
> summary(mod5)

Call:
lm(formula = quality ~ residual.sugar + I(log(volatile.acidity *
    total.sulfur.dioxide)) + total.sulfur.dioxide + density +
    chlorides + pH + volatile.acidity + sulphates + I(log(sulphates *
    alcohol)), data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.66470 -0.38317 -0.02311  0.42999  1.85286

Coefficients:
                                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                                43.740608  10.788220   4.054 5.27e-05 ***
residual.sugar                              0.022877   0.012998   1.760 0.078586 .
I(log(volatile.acidity * total.sulfur.dioxide))  0.102335   0.056189   1.821 0.068752 .
total.sulfur.dioxide                       -0.004302   0.001249  -3.444 0.000588 ***
density                                   -38.812218  10.683205  -3.633 0.000289 ***
chlorides                                  -1.635467   0.395886  -4.131 3.80e-05 ***
pH                                         -0.705678   0.120775  -5.843 6.21e-09 ***
volatile.acidity                           -1.060022   0.143718  -7.376 2.62e-13 ***
sulphates                                  -2.586091   0.262409  -9.855  < 2e-16 ***
I(log(sulphates * alcohol))                 2.728630   0.167570  16.284  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6356 on 1589 degrees of freedom
Multiple R-squared:  0.3841,    Adjusted R-squared:  0.3806
F-statistic: 110.1 on 9 and 1589 DF,  p-value: < 2.2e-16
```

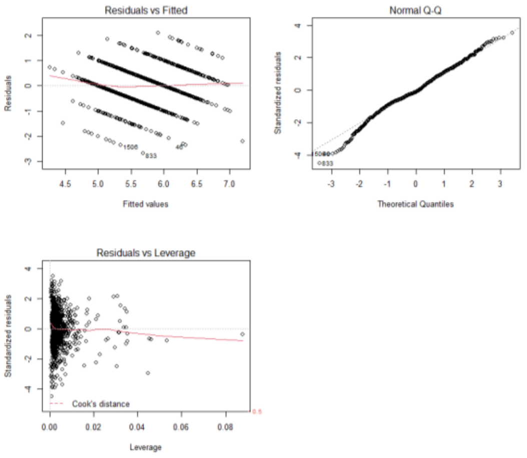


```
> slopeFULL
         (Intercept)        fixed.acidity     volatile.acidity          citric.acid       residual.sugar            chlorides
        1.040314e+02         1.248255e-01        -6.323929e-01        -3.585857e-01         7.191647e-02        -1.872307e+00
 free.sulfur.dioxide total.sulfur.dioxide              density                   pH            sulphates              alcohol
        5.534901e-03        -5.317610e-03        -1.025566e+02         1.274586e-01         8.518833e-01         2.329848e-01
> slope1
         (Intercept)     volatile.acidity            chlorides total.sulfur.dioxide            sulphates              alcohol
         2.654284537         -0.695950995         -2.326555896         -0.004223324          0.801843197          0.312544324
> slope2
             (Intercept)            fixed.acidity         volatile.acidity              citric.acid I(log10(residual.sugar))
            1.259450e+02             1.421843e-01            -6.109714e-01            -4.627044e-01             7.526555e-01
     I(log10(chlorides))      free.sulfur.dioxide     total.sulfur.dioxide                  density                       pH
           -6.490969e-01             4.867545e-03            -5.043512e-03            -1.244037e+02             1.216443e-01
     I(log10(sulphates))                  alcohol
            1.773224e+00             2.032427e-01
> slope3
         (Intercept)     volatile.acidity  I(log10(chlorides)) total.sulfur.dioxide  I(log10(sulphates))              alcohol
         2.455474313         -0.642771338         -0.827462089         -0.004025997          1.602020175          0.302743161
> slope4
                                    (Intercept)                                    fixed.acidity
                                   1.197056e+02                                     1.190194e-01
                               volatile.acidity                                      citric.acid
                                   1.193633e+00                                    -3.521085e-01
                                 residual.sugar                                        chlorides
                                   6.512684e-02                                    -1.835859e+00
                            free.sulfur.dioxide                             total.sulfur.dioxide
                                   3.455255e-03                                    -6.845962e-03
                                             pH                                          density
                                  -1.162285e-01                                    -1.186384e+02
                                      sulphates                                          alcohol
                                  -1.450464e+00                                     1.226872e-01
I(log(volatile.acidity * total.sulfur.dioxide))                      I(log(sulphates * alcohol))
                                   1.196132e-01                                     1.957011e+00
                      volatile.acidity:alcohol
                                  -1.865357e-01
> slope5
                                    (Intercept)                                   residual.sugar
                                   69.758095819                                      0.047759180
I(log(volatile.acidity * total.sulfur.dioxide))                             total.sulfur.dioxide
                                    0.112062598                                     -0.006537862
                                        density                                        chlorides
                                  -64.861424303                                     -2.460617128
                                             pH                                 volatile.acidity
                                   -0.741782982                                     -0.635956921
                                      sulphates                      I(log(sulphates * alcohol))
                                   -2.238485694                                      2.600776612
> slope.fit
         (Intercept)     volatile.acidity  I(log10(chlorides))  free.sulfur.dioxide total.sulfur.dioxide                   pH
         3.757722944         -0.509956595         -0.921123327          0.007135584         -0.005682790         -0.462044004
 I(log10(sulphates))              alcohol
         1.592035332          0.305359993
> |
```

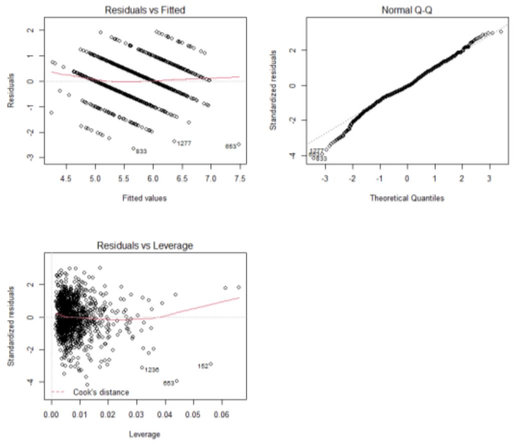
figure9 (diagnostic analysis for robust regression)



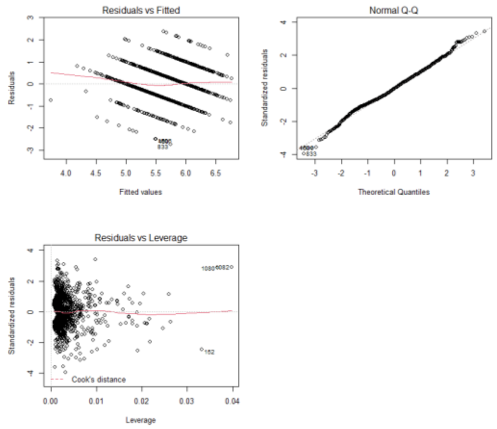
mod0



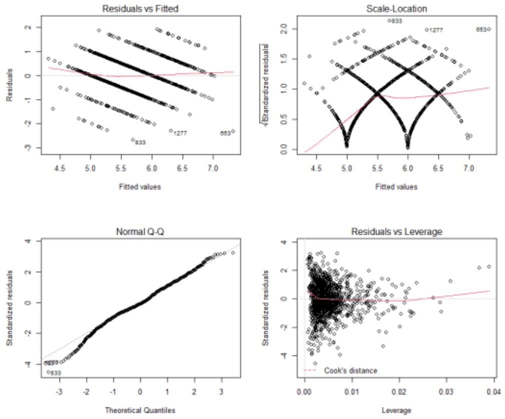
mod1



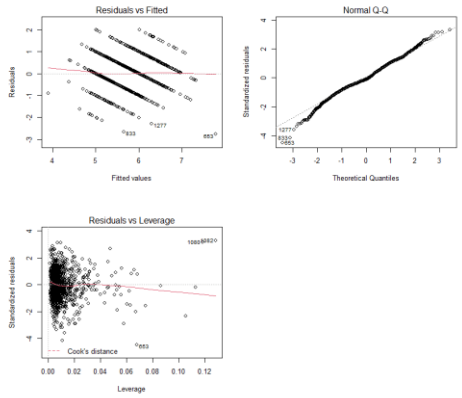
mod2



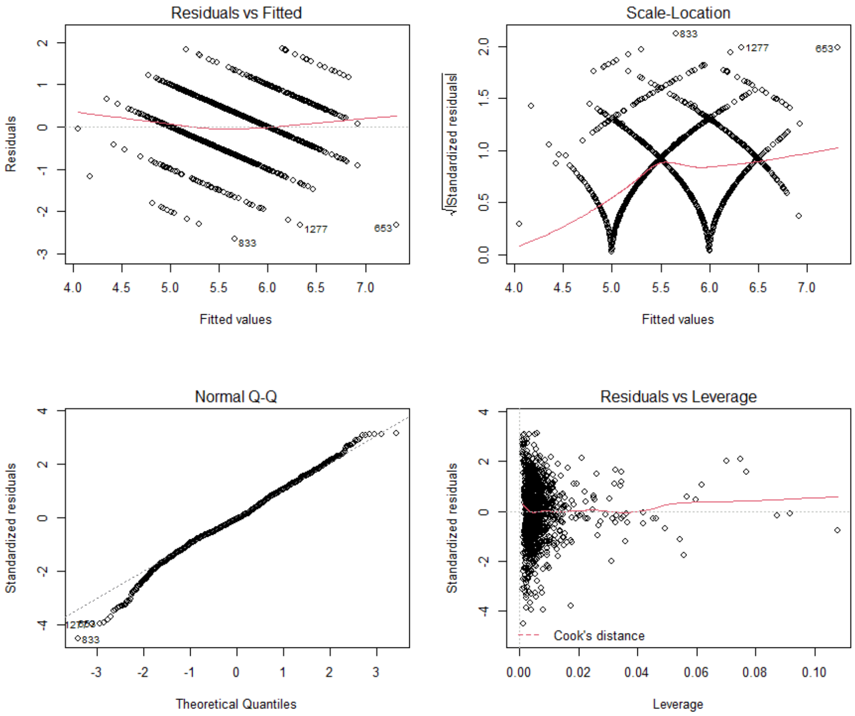
mod3



mod fit

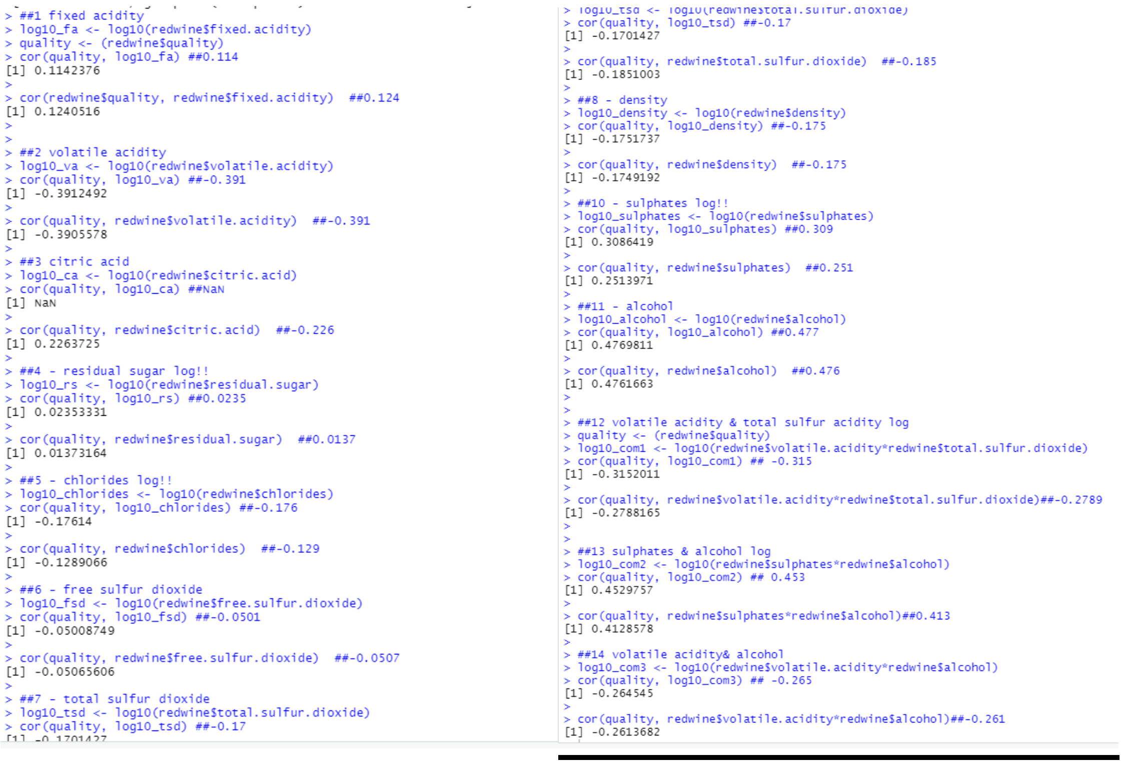


mod4



Mod5

Figure10 (correlation between variables and response variable before and after doing data transformation)



```
> ##1 fixed acidity
> log10_fa <- log10(redwine$fixed.acidity)
> quality <- (redwine$quality)
> cor(quality, log10_fa) ##0.114
[1] 0.1142376
>
> cor(redwine$quality, redwine$fixed.acidity)  ##0.124
[1] 0.1240516
>
>
> ##2 volatile acidity
> log10_va <- log10(redwine$volatile.acidity)
> cor(quality, log10_va) ##-0.391
[1] -0.3912492
>
> cor(quality, redwine$volatile.acidity)  ##-0.391
[1] -0.3905578
>
> ##3 citric acid
> log10_ca <- log10(redwine$citric.acid)
> cor(quality, log10_ca) ##NaN
[1] NaN
>
> cor(quality, redwine$citric.acid)  ##-0.226
[1] 0.2263725
>
> ##4 - residual sugar log!!
> log10_rs <- log10(redwine$residual.sugar)
> cor(quality, log10_rs) ##0.0235
[1] 0.02353331
>
> cor(quality, redwine$residual.sugar)  ##0.0137
[1] 0.01373164
>
> ##5 - chlorides log!!
> log10_chlorides <- log10(redwine$chlorides)
> cor(quality, log10_chlorides) ##-0.176
[1] -0.17614
>
> cor(quality, redwine$chlorides)  ##-0.129
[1] -0.1289066
>
> ##6 - free sulfur dioxide
> log10_fsd <- log10(redwine$free.sulfur.dioxide)
> cor(quality, log10_fsd) ##-0.0501
[1] -0.05008749
>
> cor(quality, redwine$free.sulfur.dioxide)  ##-0.0507
[1] -0.05065606
>
> ##7 - total sulfur dioxide
> log10_tsd <- log10(redwine$total.sulfur.dioxide)
> cor(quality, log10_tsd) ##-0.17
[1] -0.1701427
```

```
> log10_tsd <- log10(redwine$total.sulfur.dioxide)
> cor(quality, log10_tsd) ##-0.17
[1] -0.1701427
>
> cor(quality, redwine$total.sulfur.dioxide)  ##-0.185
[1] -0.1851003
>
> ##8 - density
> log10_density <- log10(redwine$density)
> cor(quality, log10_density) ##-0.175
[1] -0.1751737
>
> cor(quality, redwine$density)  ##-0.175
[1] -0.1749192
>
> ##10 - sulphates log!!
> log10_sulphates <- log10(redwine$sulphates)
> cor(quality, log10_sulphates) ##0.309
[1] 0.3086419
>
> cor(quality, redwine$sulphates)  ##0.251
[1] 0.2513971
>
> ##11 - alcohol
> log10_alcohol <- log10(redwine$alcohol)
> cor(quality, log10_alcohol) ##0.477
[1] 0.4769811
>
> cor(quality, redwine$alcohol)  ##0.476
[1] 0.4761663
>
>
> ##12 volatile acidity & total sulfur acidity log
> quality <- (redwine$quality)
> log10_com1 <- log10(redwine$volatile.acidity*redwine$total.sulfur.dioxide)
> cor(quality, log10_com1) ## -0.315
[1] -0.3152011
>
> cor(quality, redwine$volatile.acidity*redwine$total.sulfur.dioxide)##-0.2789
[1] -0.2788165
>
>
> ##13 sulphates & alcohol log
> log10_com2 <- log10(redwine$sulphates*redwine$alcohol)
> cor(quality, log10_com2) ## 0.453
[1] 0.4529757
>
> cor(quality, redwine$sulphates*redwine$alcohol)##0.413
[1] 0.4128578
>
> ##14 volatile acidity& alcohol
> log10_com3 <- log10(redwine$volatile.acidity*redwine$alcohol)
> cor(quality, log10_com3) ## -0.265
[1] -0.264545
>
> cor(quality, redwine$volatile.acidity*redwine$alcohol)##-0.261
[1] -0.2613682
```

Reference

Wine Quality. (n.d.). Retrieved December 06, 2020, from https://www.sciencedirect.com/topics/food-science/wine-quality