Regression Analysis Project

Regression Analysis using Red wine quality data

JUNE 9, 2019

Mujeer M

Contents

Code

Red wine quality

Regression Analysis Project

MUJEER M.

Introduction

It is quite challenging to identify the quality of wine by just looking at the bottle. The objective of this report is to evaluate the relationship between variables in our wine quality dataset and identify through Regression analysis as to which variables have a linear relationship with the dependant variable (quality). This dataset consists of different variables that define the quality of the wine.

Methodology

We first checked our dataset if there are any issues before proceeding with the analysis. We then fit classic regression model and check if there is any relationship between the independant and dependant variables. We further do a stepwise regression to select the best model based on AIC scores. We also fit a GLM model and then used transformation on linear model to see if we get better results.

Data Retrieving

The dataset was sourced from Kaggle at https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-(https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-) 2009. There are 12 numerical data attributes and 1599 observations in this dataset. The variables are fixed.acidity, volatile.acidity citric.acid residual.sugar chlorides free.sulfur.dioxide, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality.

Setup

The necessary packages are loaded.

Hide

```
library(readr)
library(analytics)
library(ggplot2)
library(GGally)
library(tidyr)
library(tidyverse)
library(readxl)
library(car)
library(MASS)
library(caret)
library(matlib)
library(broom)
library(ggfortify)
library(survey)
library(gvlma)
library(trafo) #for model assumptions and transformation
library(rcompanion) #for calculating pseudo R-squared
library(lmtest)
```

Reading Data

The read.csv function is used to read the data into R.

Hide

```
redwine <- read_csv("winequality-red.csv")

head(redwine)
```

Data Preprocessing

The required pre-processing is performed on the dataset for better analysis and evaluation.

Hide

```
#renaming columns due to column name errors
names(redwine)[names(redwine) == "fixed acidity"] <- "fixed_acidity"
names(redwine)[names(redwine) == "volatile acidity"] <- "volatile_acidity"
names(redwine)[names(redwine) == "citric acid"] <- "citric_acid"
names(redwine)[names(redwine) == "residual sugar"] <- "residual_sugar"
names(redwine)[names(redwine) == "free sulfur dioxide"] <- "free_sulfur_dioxide"
names(redwine)[names(redwine) == "total sulfur dioxide"] <- "total_sulfur_dioxide"
colnames(redwine)
```

```
 [1] "fixed_acidity"        "volatile_acidity"     "citric_acid"
 [4] "residual_sugar"       "chlorides"            "free_sulfur_dioxide"
 [7] "total_sulfur_dioxide" "density"              "pH"
[10] "sulphates"            "alcohol"              "quality"
```

Model Fitting

The model is fitted with all the variables

Hide

```
redwine_reg = lm(formula = redwine$quality ~ fixed_acidity+volatile_acidity+citric_acid+resid
ual_sugar+chlorides+free_sulfur_dioxide+total_sulfur_dioxide+density+pH+sulphates+alcohol, da
ta = redwine)
summary(redwine_reg)
```

```
Call:
lm(formula = redwine$quality ~ fixed_acidity + volatile_acidity +
    citric_acid + residual_sugar + chlorides + free_sulfur_dioxide +
    total_sulfur_dioxide + density + pH + sulphates + alcohol,
    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
```

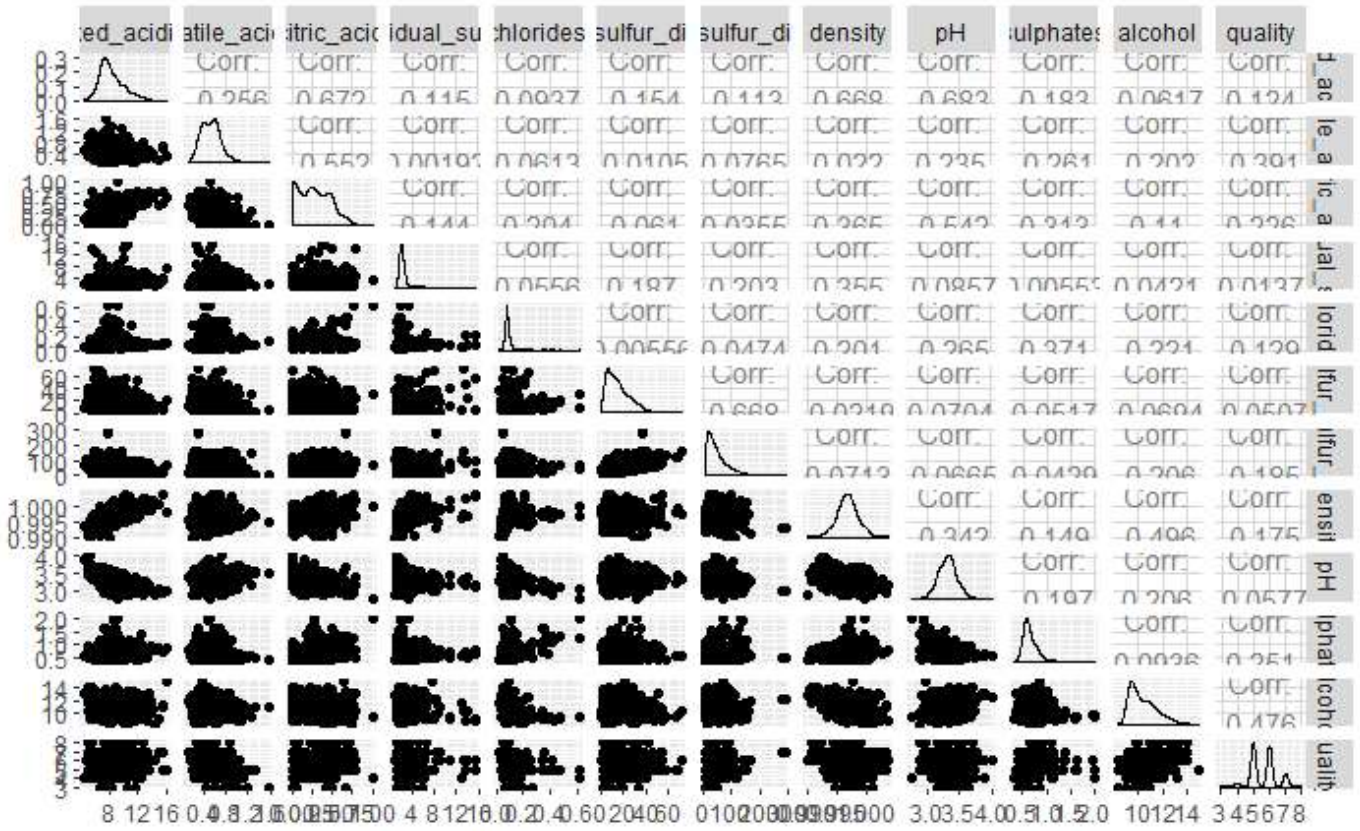
1. Volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol are significant.
2. 36% of the variability in the data is explained by the full model

Correlation plot

The below code creates the correlation plot for all variables.It displays the linear relationship between 2 variables. The correlation value ranges from -1(negative correlation) to 1(positive correlation). The value closer to -1 or 1 indicates a stronger positive or negative correlation.

Hide

```
# correlation plot of all variables (-1[negative correlation] and 1[positive correlation])
ggpairs(redwine)
```



Anova

Analysis of Variance (ANOVA) is a methodology used for calculations that provide information about levels of variability within a regression model and form a basis for tests of significance.The Anova gives the overall fit of the model.

Hide

```
anova(redwine_reg)
```

```
Analysis of Variance Table

Response: redwine$quality
                       Df Sum Sq Mean Sq  F value    Pr(>F)
fixed_acidity           1  16.04  16.038  38.1924 8.132e-10 ***
volatile_acidity        1 143.57 143.573 341.9062 < 2.2e-16 ***
citric_acid             1   0.02   0.024   0.0581  0.809535
residual_sugar          1   0.16   0.158   0.3764  0.539600
chlorides               1  13.06  13.062  31.1057 2.868e-08 ***
free_sulfur_dioxide     1   2.97   2.974   7.0828  0.007861 **
total_sulfur_dioxide    1  30.09  30.093  71.6631 < 2.2e-16 ***
density                 1  61.31  61.310 146.0054 < 2.2e-16 ***
pH                      1   7.15   7.154  17.0358 3.859e-05 ***
sulphates               1  55.70  55.697 132.6366 < 2.2e-16 ***
alcohol                 1  45.67  45.672 108.7643 < 2.2e-16 ***
Residuals            1587 666.41   0.420
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Based on the overall fit, all the variables are significant apart from citric acid and residual sugar.

Model Building

Full model contains all the variables:

Hide

```
full_redwine=lm(formula = redwine$quality ~ fixed_acidity+volatile_acidity+citric_acid+residu
al_sugar+chlorides+free_sulfur_dioxide+total_sulfur_dioxide+density+pH+sulphates+alcohol, dat
a = redwine)
```

Output has not been discussed as this is the same as our basic multi linear model.

Null model contains no variable:

Hide

```
null_redwine=lm(redwine$quality~1, data=redwine)
```

Forward Selection:

This method is used to provide an initial screening of variables.Forward selection begins with no candidate variables in the model. Initially it selects the variable with the highest R-squared value. At each step the varible that increases the R-squared is included to the required set of variables. It stops adding variable when none of the variablea are further significant for the model building.

Hide

```
#forward selection using AIC values
step(null_redwine, scope=list(lower=null_redwine, upper=full_redwine), direction="forward")
```

```
Start:  AIC=-682.5
redwine$quality ~ 1

                      Df Sum of Sq     RSS      AIC
+ alcohol              1   236.295  805.87 -1091.65
+ volatile_acidity     1   158.967  883.20  -945.14
+ sulphates            1    65.865  976.30  -784.89
+ citric_acid          1    53.405  988.76  -764.61
+ total_sulfur_dioxide 1    35.707 1006.46  -736.24
+ density              1    31.887 1010.28  -730.19
+ chlorides            1    17.318 1024.85  -707.29
+ fixed_acidity        1    16.038 1026.13  -705.29
+ pH                    1     3.473 1038.69  -685.84
+ free_sulfur_dioxide  1     2.674 1039.49  -684.61
<none>                             1042.17  -682.50
+ residual_sugar       1     0.197 1041.97  -680.80

Step:  AIC=-1091.65
redwine$quality ~ alcohol

                      Df Sum of Sq    RSS     AIC
+ volatile_acidity     1    94.074 711.80 -1288.1
+ sulphates            1    44.977 760.89 -1181.5
+ citric_acid          1    31.953 773.92 -1154.3
+ pH                   1    26.362 779.51 -1142.8
+ fixed_acidity        1    24.623 781.25 -1139.3
+ total_sulfur_dioxide 1     8.270 797.60 -1106.2
+ density              1     5.203 800.67 -1100.0
<none>                             805.87 -1091.7
+ chlorides            1     0.611 805.26 -1090.9
+ free_sulfur_dioxide  1     0.325 805.55 -1090.3
+ residual_sugar       1     0.041 805.83 -1089.7

Step:  AIC=-1288.14
redwine$quality ~ alcohol + volatile_acidity

                      Df Sum of Sq    RSS     AIC
+ sulphates            1   19.6916 692.10 -1331.0
+ total_sulfur_dioxide 1    6.3730 705.42 -1300.5
+ pH                   1    5.9515 705.84 -1299.6
+ fixed_acidity        1    5.7061 706.09 -1299.0
+ density              1    1.9410 709.86 -1290.5
<none>                             711.80 -1288.1
+ free_sulfur_dioxide  1    0.6621 711.13 -1287.6
+ chlorides            1    0.3762 711.42 -1287.0
+ citric_acid          1    0.1936 711.60 -1286.6
+ residual_sugar       1    0.0101 711.79 -1286.2

Step:  AIC=-1331
redwine$quality ~ alcohol + volatile_acidity + sulphates

                      Df Sum of Sq    RSS     AIC
+ total_sulfur_dioxide 1    8.2176 683.89 -1348.1
+ chlorides            1    7.4925 684.61 -1346.4
+ fixed_acidity        1    3.3282 688.78 -1336.7
+ pH                   1    3.0454 689.06 -1336.0
+ free_sulfur_dioxide  1    1.1129 690.99 -1331.6
```

```
<none>                              692.10 -1331.0
+ citric_acid           1    0.2522 691.85 -1329.6
+ density               1    0.2222 691.88 -1329.5
+ residual_sugar        1    0.0143 692.09 -1329.0

Step:  AIC=-1348.1
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide

                      Df Sum of Sq    RSS     AIC
+ chlorides            1    8.0370 675.85 -1365.0
+ pH                   1    3.3094 680.58 -1353.8
+ fixed_acidity        1    2.1037 681.78 -1351.0
+ free_sulfur_dioxide  1    1.3557 682.53 -1349.3
<none>                             683.89 -1348.1
+ residual_sugar       1    0.2634 683.62 -1346.7
+ density              1    0.1077 683.78 -1346.3
+ citric_acid          1    0.0730 683.81 -1346.3

Step:  AIC=-1365
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide +
    chlorides

                      Df Sum of Sq    RSS     AIC
+ pH                   1    5.9189 669.93 -1377.1
+ fixed_acidity        1    2.4065 673.44 -1368.7
+ free_sulfur_dioxide  1    1.2403 674.61 -1365.9
<none>                             675.85 -1365.0
+ residual_sugar       1    0.5531 675.30 -1364.3
+ citric_acid          1    0.1615 675.69 -1363.4
+ density              1    0.1526 675.70 -1363.4

Step:  AIC=-1377.06
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide +
    chlorides + pH

                      Df Sum of Sq    RSS     AIC
+ free_sulfur_dioxide  1   2.39413 667.54 -1380.8
<none>                             669.93 -1377.1
+ citric_acid          1   0.80525 669.13 -1377.0
+ residual_sugar       1   0.28390 669.65 -1375.7
+ density              1   0.04468 669.89 -1375.2
+ fixed_acidity        1   0.01040 669.92 -1375.1

Step:  AIC=-1380.79
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide +
    chlorides + pH + free_sulfur_dioxide

                 Df Sum of Sq    RSS     AIC
<none>                        667.54 -1380.8
+ citric_acid     1   0.47480 667.06 -1379.9
+ residual_sugar  1   0.16673 667.37 -1379.2
+ density         1   0.03079 667.51 -1378.9
+ fixed_acidity   1   0.00663 667.53 -1378.8

Call:
lm(formula = redwine$quality ~ alcohol + volatile_acidity + sulphates +
    total_sulfur_dioxide + chlorides + pH + free_sulfur_dioxide,
    data = redwine)
```

```
Coefficients:
         (Intercept)              alcohol     volatile_acidity            sulphates
            4.430099             0.289303            -1.012753             0.882665
total_sulfur_dioxide            chlorides                   pH   free_sulfur_dioxide
           -0.003482            -2.017814            -0.482661             0.005077
```

Forward elimination is suggesting the model with variables **volatile_acidity + sulphates + total_sulfur_dioxide + chlorides + pH + free_sulfur_dioxide with highest negative AIC=-1380.79**

Backward Selection:

Backward modelling begins with a model where all variables are included. Hence it always retains a high R-squared value.this model strats with all the variables. At each step, the least significant one is removed from the model. This process continuous until no significant variables remains. The user sets the significance level of a variable entering into the model.

Hide

```
step(full_redwine, data=redwine, direction="backward")
```

```
Start:  AIC=-1375.49
redwine$quality ~ fixed_acidity + volatile_acidity + citric_acid +
    residual_sugar + chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
    density + pH + sulphates + alcohol

                      Df Sum of Sq    RSS     AIC
- density              1     0.287 666.70 -1376.8
- fixed_acidity        1     0.389 666.80 -1376.5
- residual_sugar       1     0.498 666.91 -1376.3
- citric_acid          1     0.646 667.06 -1375.9
<none>                             666.41 -1375.5
- free_sulfur_dioxide  1     1.694 668.10 -1373.4
- pH                   1     1.957 668.37 -1372.8
- chlorides            1     8.391 674.80 -1357.5
- total_sulfur_dioxide 1     8.427 674.84 -1357.4
- sulphates            1    26.971 693.38 -1314.0
- volatile_acidity     1    33.620 700.03 -1298.8
- alcohol              1    45.672 712.08 -1271.5

Step:  AIC=-1376.8
redwine$quality ~ fixed_acidity + volatile_acidity + citric_acid +
    residual_sugar + chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
    pH + sulphates + alcohol

                      Df Sum of Sq    RSS     AIC
- fixed_acidity        1     0.108 666.81 -1378.5
- residual_sugar       1     0.231 666.93 -1378.2
- citric_acid          1     0.654 667.35 -1377.2
<none>                             666.70 -1376.8
- free_sulfur_dioxide  1     1.829 668.53 -1374.4
- pH                   1     4.325 671.02 -1368.5
- total_sulfur_dioxide 1     8.728 675.43 -1358.0
- chlorides            1     8.761 675.46 -1357.9
- sulphates            1    27.287 693.98 -1314.7
- volatile_acidity     1    35.000 701.70 -1297.0
- alcohol              1   119.669 786.37 -1114.8

Step:  AIC=-1378.54
redwine$quality ~ volatile_acidity + citric_acid + residual_sugar +
    chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
    pH + sulphates + alcohol

                      Df Sum of Sq    RSS     AIC
- residual_sugar       1     0.257 667.06 -1379.9
- citric_acid          1     0.565 667.37 -1379.2
<none>                             666.81 -1378.5
- free_sulfur_dioxide  1     1.901 668.71 -1376.0
- pH                   1     7.065 673.87 -1363.7
- chlorides            1     9.940 676.75 -1356.9
- total_sulfur_dioxide 1    10.031 676.84 -1356.7
- sulphates            1    27.673 694.48 -1315.5
- volatile_acidity     1    36.234 703.04 -1295.9
- alcohol              1   120.633 787.44 -1114.7

Step:  AIC=-1379.93
redwine$quality ~ volatile_acidity + citric_acid + chlorides +
    free_sulfur_dioxide + total_sulfur_dioxide + pH + sulphates +
```

```
    alcohol

                       Df Sum of Sq    RSS     AIC
- citric_acid           1     0.475 667.54 -1380.8
<none>                              667.06 -1379.9
- free_sulfur_dioxide   1     2.064 669.13 -1377.0
- pH                    1     7.138 674.20 -1364.9
- total_sulfur_dioxide  1     9.828 676.89 -1358.5
- chlorides             1     9.832 676.89 -1358.5
- sulphates             1    27.446 694.51 -1317.5
- volatile_acidity      1    35.977 703.04 -1297.9
- alcohol               1   122.667 789.73 -1112.0

Step:  AIC=-1380.79
redwine$quality ~ volatile_acidity + chlorides + free_sulfur_dioxide +
    total_sulfur_dioxide + pH + sulphates + alcohol

                       Df Sum of Sq    RSS     AIC
<none>                              667.54 -1380.8
- free_sulfur_dioxide   1     2.394 669.93 -1377.1
- pH                    1     7.073 674.61 -1365.9
- total_sulfur_dioxide  1    10.787 678.32 -1357.2
- chlorides             1    10.809 678.35 -1357.1
- sulphates             1    27.060 694.60 -1319.2
- volatile_acidity      1    42.318 709.85 -1284.5
- alcohol               1   124.483 792.02 -1109.4

Call:
lm(formula = redwine$quality ~ volatile_acidity + chlorides +
    free_sulfur_dioxide + total_sulfur_dioxide + pH + sulphates +
    alcohol, data = redwine)

Coefficients:
         (Intercept)      volatile_acidity             chlorides   free_sulfur_dioxide
            4.430099             -1.012753             -2.017814              0.005077
total_sulfur_dioxide                    pH             sulphates               alcohol
           -0.003482             -0.482661              0.882665              0.289303
```

Backward elimination is suggesting the model with variables **volatile_acidity + chlorides + free_sulfur_dioxide + total_sulfur_dioxide + pH + sulphates + alcohol with highest negative AIC=-1380.79**

Stepwise Selection:

Stepwise selction is a combination of both the forward and the backward selection methods. At each step when a variable is added, then all the candidate variable for the model is checked to see if the significance is reduced below a particular tolerant level. Stepwise regression required to significance levels , one for adding varible and one for removing the variable.

Hide

```
step(null_redwine, scope = list(upper=full_redwine), data=redwine, direction="both")
```

```
Start:  AIC=-682.5
redwine$quality ~ 1

                      Df Sum of Sq     RSS      AIC
+ alcohol              1   236.295  805.87 -1091.65
+ volatile_acidity     1   158.967  883.20  -945.14
+ sulphates            1    65.865  976.30  -784.89
+ citric_acid          1    53.405  988.76  -764.61
+ total_sulfur_dioxide 1    35.707 1006.46  -736.24
+ density              1    31.887 1010.28  -730.19
+ chlorides            1    17.318 1024.85  -707.29
+ fixed_acidity        1    16.038 1026.13  -705.29
+ pH                    1     3.473 1038.69  -685.84
+ free_sulfur_dioxide  1     2.674 1039.49  -684.61
<none>                             1042.17  -682.50
+ residual_sugar       1     0.197 1041.97  -680.80

Step:  AIC=-1091.65
redwine$quality ~ alcohol

                      Df Sum of Sq     RSS     AIC
+ volatile_acidity     1    94.074  711.80 -1288.1
+ sulphates            1    44.977  760.89 -1181.5
+ citric_acid          1    31.953  773.92 -1154.3
+ pH                   1    26.362  779.51 -1142.8
+ fixed_acidity        1    24.623  781.25 -1139.3
+ total_sulfur_dioxide 1     8.270  797.60 -1106.2
+ density              1     5.203  800.67 -1100.0
<none>                             805.87 -1091.7
+ chlorides            1     0.611  805.26 -1090.9
+ free_sulfur_dioxide  1     0.325  805.55 -1090.3
+ residual_sugar       1     0.041  805.83 -1089.7
- alcohol              1   236.295 1042.17  -682.5

Step:  AIC=-1288.14
redwine$quality ~ alcohol + volatile_acidity

                      Df Sum of Sq    RSS      AIC
+ sulphates            1    19.692 692.10 -1331.00
+ total_sulfur_dioxide 1     6.373 705.42 -1300.52
+ pH                   1     5.952 705.84 -1299.56
+ fixed_acidity        1     5.706 706.09 -1299.01
+ density              1     1.941 709.86 -1290.50
<none>                             711.80 -1288.14
+ free_sulfur_dioxide  1     0.662 711.13 -1287.63
+ chlorides            1     0.376 711.42 -1286.98
+ citric_acid          1     0.194 711.60 -1286.57
+ residual_sugar       1     0.010 711.79 -1286.16
- volatile_acidity     1    94.074 805.87 -1091.65
- alcohol              1   171.402 883.20  -945.14

Step:  AIC=-1331
redwine$quality ~ alcohol + volatile_acidity + sulphates

                      Df Sum of Sq    RSS      AIC
+ total_sulfur_dioxide 1     8.218 683.89 -1348.10
+ chlorides            1     7.493 684.61 -1346.40
```

```
+ fixed_acidity         1     3.328 688.78 -1336.70
+ pH                    1     3.045 689.06 -1336.05
+ free_sulfur_dioxide   1     1.113 690.99 -1331.57
<none>                              692.10 -1331.00
+ citric_acid           1     0.252 691.85 -1329.58
+ density               1     0.222 691.88 -1329.51
+ residual_sugar        1     0.014 692.09 -1329.03
- sulphates             1    19.692 711.80 -1288.14
- volatile_acidity      1    68.789 760.89 -1181.48
- alcohol               1   166.109 858.21  -989.03

Step:  AIC=-1348.1
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide

                      Df Sum of Sq    RSS     AIC
+ chlorides            1     8.037 675.85 -1365.0
+ pH                   1     3.309 680.58 -1353.8
+ fixed_acidity        1     2.104 681.78 -1351.0
+ free_sulfur_dioxide  1     1.356 682.53 -1349.3
<none>                              683.89 -1348.1
+ residual_sugar       1     0.263 683.62 -1346.7
+ density              1     0.108 683.78 -1346.3
+ citric_acid          1     0.073 683.81 -1346.3
- total_sulfur_dioxide 1     8.218 692.10 -1331.0
- sulphates            1    21.536 705.42 -1300.5
- volatile_acidity     1    66.047 749.93 -1202.7
- alcohol              1   145.552 829.44 -1041.6

Step:  AIC=-1365
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide +
    chlorides

                      Df Sum of Sq    RSS     AIC
+ pH                   1     5.919 669.93 -1377.1
+ fixed_acidity        1     2.407 673.44 -1368.7
+ free_sulfur_dioxide  1     1.240 674.61 -1365.9
<none>                              675.85 -1365.0
+ residual_sugar       1     0.553 675.30 -1364.3
+ citric_acid          1     0.162 675.69 -1363.4
+ density              1     0.153 675.70 -1363.4
- chlorides            1     8.037 683.89 -1348.1
- total_sulfur_dioxide 1     8.762 684.61 -1346.4
- sulphates            1    29.201 705.05 -1299.4
- volatile_acidity     1    58.869 734.72 -1233.5
- alcohol              1   119.894 795.74 -1105.9

Step:  AIC=-1377.06
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide +
    chlorides + pH

                      Df Sum of Sq    RSS     AIC
+ free_sulfur_dioxide  1     2.394 667.54 -1380.8
<none>                              669.93 -1377.1
+ citric_acid          1     0.805 669.13 -1377.0
+ residual_sugar       1     0.284 669.65 -1375.7
+ density              1     0.045 669.89 -1375.2
+ fixed_acidity        1     0.010 669.92 -1375.1
- pH                   1     5.919 675.85 -1365.0
```

```
- total_sulfur_dioxide  1     9.233 679.16 -1357.2
- chlorides             1    10.647 680.58 -1353.8
- sulphates             1    27.445 697.38 -1314.9
- volatile_acidity      1    44.972 714.90 -1275.2
- alcohol               1   125.812 795.74 -1103.9

Step:  AIC=-1380.79
redwine$quality ~ alcohol + volatile_acidity + sulphates + total_sulfur_dioxide +
    chlorides + pH + free_sulfur_dioxide

                      Df Sum of Sq    RSS     AIC
<none>                             667.54 -1380.8
+ citric_acid          1     0.475 667.06 -1379.9
+ residual_sugar       1     0.167 667.37 -1379.2
+ density              1     0.031 667.51 -1378.9
+ fixed_acidity        1     0.007 667.53 -1378.8
- free_sulfur_dioxide  1     2.394 669.93 -1377.1
- pH                   1     7.073 674.61 -1365.9
- total_sulfur_dioxide  1    10.787 678.32 -1357.2
- chlorides             1    10.809 678.35 -1357.1
- sulphates             1    27.060 694.60 -1319.2
- volatile_acidity      1    42.318 709.85 -1284.5
- alcohol               1   124.483 792.02 -1109.4

Call:
lm(formula = redwine$quality ~ alcohol + volatile_acidity + sulphates +
    total_sulfur_dioxide + chlorides + pH + free_sulfur_dioxide,
    data = redwine)

Coefficients:
         (Intercept)               alcohol      volatile_acidity             sulphates
            4.430099              0.289303              -1.012753              0.882665
total_sulfur_dioxide             chlorides                    pH   free_sulfur_dioxide
           -0.003482             -2.017814             -0.482661              0.005077
```

Stepwise regression is suggesting the model with variables **alcohol + volatile_acidity + sulphates + total_sulfur_dioxide + chlorides + pH + free_sulfur_dioxide AIC=-1380.79**

Manual F-test-based backward selection

Hide

```
drop1(full_redwine,test="F")
```

```
Single term deletions

Model:
redwine$quality ~ fixed_acidity + volatile_acidity + citric_acid +
    residual_sugar + chlorides + free_sulfur_dioxide + total_sulfur_dioxide +
    density + pH + sulphates + alcohol
                     Df Sum of Sq    RSS     AIC  F value    Pr(>F)
<none>                             666.41 -1375.5
fixed_acidity         1     0.389 666.80 -1376.5   0.9275   0.33565
volatile_acidity      1    33.620 700.03 -1298.8  80.0632 < 2.2e-16 ***
citric_acid           1     0.646 667.06 -1375.9   1.5387   0.21499
residual_sugar        1     0.498 666.91 -1376.3   1.1850   0.27650
chlorides             1     8.391 674.80 -1357.5  19.9815 8.374e-06 ***
free_sulfur_dioxide   1     1.694 668.10 -1373.4   4.0346   0.04474 *
total_sulfur_dioxide  1     8.427 674.84 -1357.4  20.0689 8.005e-06 ***
density               1     0.287 666.70 -1376.8   0.6832   0.40861
pH                    1     1.957 668.37 -1372.8   4.6612   0.03100 *
sulphates             1    26.971 693.38 -1314.0  64.2290 2.127e-15 ***
alcohol               1    45.672 712.08 -1271.5 108.7643 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Removing insignificant variables found by partial F test

Hide

```
drop1(update(full_redwine, ~ . -citric_acid-residual_sugar-density), test = "F")
```

```
Single term deletions

Model:
redwine$quality ~ fixed_acidity + volatile_acidity + chlorides +
    free_sulfur_dioxide + total_sulfur_dioxide + pH + sulphates +
    alcohol
                     Df Sum of Sq    RSS     AIC  F value    Pr(>F)
<none>                             667.53 -1378.8
fixed_acidity         1     0.007 667.54 -1380.8   0.0158   0.90000
volatile_acidity      1    42.192 709.72 -1282.8 100.4977 < 2.2e-16 ***
chlorides             1    10.693 678.22 -1355.4  25.4689 5.012e-07 ***
free_sulfur_dioxide   1     2.390 669.92 -1375.1   5.6936   0.01714 *
total_sulfur_dioxide  1    10.637 678.17 -1355.5  25.3353 5.366e-07 ***
pH                    1     4.186 671.72 -1370.8   9.9709   0.00162 **
sulphates             1    26.888 694.42 -1317.7  64.0454 2.324e-15 ***
alcohol               1   124.487 792.02 -1107.4 296.5183 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

After citric acid, residual sugar, density variable is dropped. Fixed acidity has become insignificant.

Excluding citric acid, residual sugar, density variable and fixed acidity

Hide

```
drop1(update(full_redwine, ~ . -citric_acid-residual_sugar-density-fixed_acidity), test = "F"
)
```

```
Single term deletions

Model:
redwine$quality ~ volatile_acidity + chlorides + free_sulfur_dioxide +
    total_sulfur_dioxide + pH + sulphates + alcohol
                     Df Sum of Sq    RSS     AIC  F value    Pr(>F)
<none>                             667.54 -1380.8
volatile_acidity      1    42.318 709.85 -1284.5 100.8593 < 2.2e-16 ***
chlorides             1    10.809 678.35 -1357.1  25.7630 4.314e-07 ***
free_sulfur_dioxide   1     2.394 669.93 -1377.1   5.7061   0.01702 *
total_sulfur_dioxide  1    10.787 678.32 -1357.2  25.7087 4.435e-07 ***
pH                    1     7.073 674.61 -1365.9  16.8570 4.235e-05 ***
sulphates             1    27.060 694.60 -1319.2  64.4956 1.865e-15 ***
alcohol               1   124.483 792.02 -1109.4 296.6910 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Now all other variables shows as significant

Manual F-test-based forward selection

Hide

```
add1(null_redwine, scope =full_redwine, test = "F")
```

```
Single term additions

Model:
redwine$quality ~ 1
                     Df Sum of Sq     RSS      AIC  F value    Pr(>F)
<none>                             1042.17  -682.50
fixed_acidity         1    16.038 1026.13  -705.29  24.9600 6.496e-07 ***
volatile_acidity      1   158.967  883.20  -945.14 287.4444 < 2.2e-16 ***
citric_acid           1    53.405  988.76  -764.61  86.2577 < 2.2e-16 ***
residual_sugar        1     0.197 1041.97  -680.80   0.3012   0.58322
chlorides             1    17.318 1024.85  -707.29  26.9856 2.313e-07 ***
free_sulfur_dioxide   1     2.674 1039.49  -684.61   4.1085   0.04283 *
total_sulfur_dioxide  1    35.707 1006.46  -736.24  56.6578 8.622e-14 ***
density               1    31.887 1010.28  -730.19  50.4052 1.875e-12 ***
pH                    1     3.473 1038.69  -685.84   5.3405   0.02096 *
sulphates             1    65.865  976.30  -784.89 107.7404 < 2.2e-16 ***
alcohol               1   236.295  805.87 -1091.65 468.2670 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Residual sugar is insignificant and all other variables are significant

From all above tests we can see that residual sugar is insignificant

Hide

```
add1(update(null_redwine, ~ . +residual_sugar), scope = full_redwine, test = "F")
```

```
Single term additions

Model:
redwine$quality ~ residual_sugar
                     Df Sum of Sq     RSS      AIC  F value    Pr(>F)
<none>                           1041.97  -680.80
fixed_acidity         1    15.841 1026.13  -703.30  24.6392 7.653e-07 ***
volatile_acidity      1   158.989  882.98  -943.54 287.3754 < 2.2e-16 ***
citric_acid           1    53.584  988.38  -763.22  86.5245 < 2.2e-16 ***
chlorides             1    17.578 1024.39  -706.00  27.3860 1.887e-07 ***
free_sulfur_dioxide   1     3.059 1038.91  -683.50   4.6998   0.03031 *
total_sulfur_dioxide  1    38.372 1003.60  -738.80  61.0226 1.015e-14 ***
density               1    38.557 1003.41  -739.09  61.3282 8.743e-15 ***
pH                    1     3.358 1038.61  -683.96   5.1601   0.02324 *
sulphates             1    65.828  976.14  -783.15 107.6288 < 2.2e-16 ***
alcohol               1   236.140  805.83 -1089.73 467.6908 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

Hide

```
add1(update(null_redwine, ~ . +residual_sugar+free_sulfur_dioxide+pH), scope = full_redwine,
test = "F")
```

```
Single term additions

Model:
redwine$quality ~ residual_sugar + free_sulfur_dioxide + pH
                     Df Sum of Sq     RSS      AIC F value    Pr(>F)
<none>                           1036.07  -685.87
fixed_acidity         1    12.360 1023.71  -703.06  19.245 1.225e-05 ***
volatile_acidity      1   158.505  877.57  -949.37 287.906 < 2.2e-16 ***
citric_acid           1    55.348  980.72  -771.66  89.958 < 2.2e-16 ***
chlorides             1    23.149 1012.92  -720.01  36.429 1.965e-09 ***
total_sulfur_dioxide  1    48.918  987.15  -761.21  78.990 < 2.2e-16 ***
density               1    54.319  981.75  -769.98  88.194 < 2.2e-16 ***
sulphates             1    64.640  971.43  -786.88 106.066 < 2.2e-16 ***
alcohol               1   257.056  779.02 -1139.85 525.981 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

1. All the variables are significant (fixed acidity, volatile acidity, citric acid, chlorides,
2. Total sulfur dioxide, density, sulphates and alcohol) after removing residual sugar- free sulfur dioxide and pH

Candidate models

1. volatile acidity, chlorides, total sulfur dioxide, sulphates, alcohol
2. volatile acidity, chlorides, total sulfur dioxide, sulphates, alcohol, free sulfur dioxide, pH
3. our 3rd model will be using GLM

Hide

```
redwine_c1=lm(quality ~ volatile_acidity+chlorides+total_sulfur_dioxide+sulphates+alcohol, da
ta = redwine)
summary(redwine_c1)
```

```

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
```

Hide

```
redwine_c2=lm(quality ~ volatile_acidity+chlorides+total_sulfur_dioxide+sulphates+alcohol+fre
e_sulfur_dioxide+pH, data = redwine)
summary(redwine_c2)
```

```
Call:
lm(formula = quality ~ volatile_acidity + chlorides + total_sulfur_dioxide +
    sulphates + alcohol + free_sulfur_dioxide + pH, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
volatile_acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
total_sulfur_dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
free_sulfur_dioxide   0.0050774  0.0021255   2.389    0.017 *
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16
```

The model redwine_c2 is better than model redwine_c1.Hence,we will proceed with redwine_c2 as it could make better predictions.

Multicollinearity

Hide

```
vif(redwine_reg)
```

```
      fixed_acidity     volatile_acidity          citric_acid       residual_sugar
           7.767512             1.789390             3.128022             1.702588
          chlorides  free_sulfur_dioxide total_sulfur_dioxide              density
           1.481932             1.963019             2.186813             6.343760
                 pH            sulphates              alcohol
           3.329732             1.429434             3.031160
```

Hide

```
vif(redwine_c1)
```

```
   volatile_acidity           chlorides total_sulfur_dioxide           sulphates
           1.134865            1.280047             1.052686            1.315932
            alcohol
           1.162223
```

Hide

```
vif(redwine_c2)
```

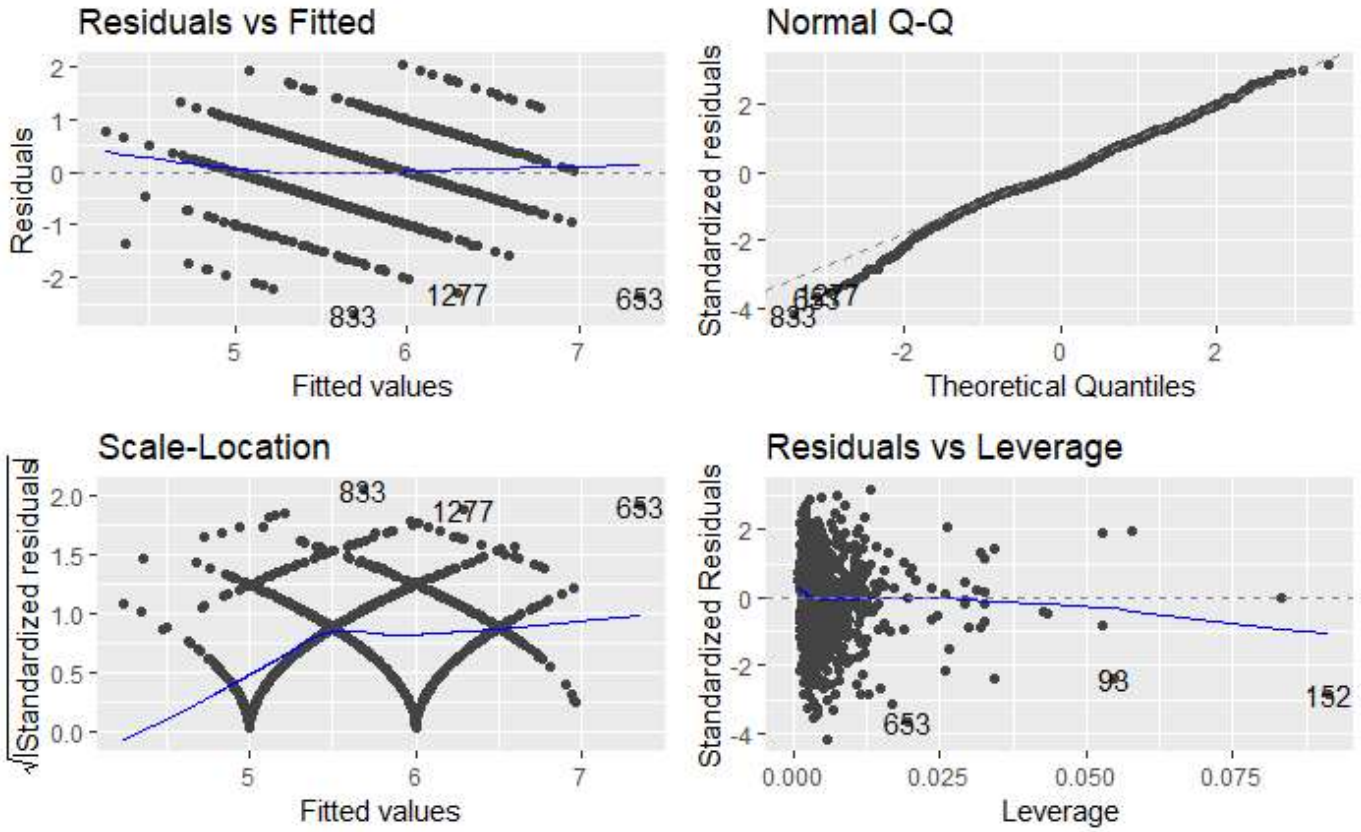
```
    volatile_acidity           chlorides total_sulfur_dioxide           sulphates
            1.241819            1.333333             1.943920            1.321931
             alcohol  free_sulfur_dioxide                   pH
            1.220157            1.882706             1.254570
```

1. The full model has vif scores for fixed acidity and density above 5.
2. Model redwine_c1 and redwine_c2 have vif scores for all variables below 5 which shows that there is no multicollinearity.
3. The next step is to proceed with redwine_c2 as our final model.

Diagnostic plots

Hide

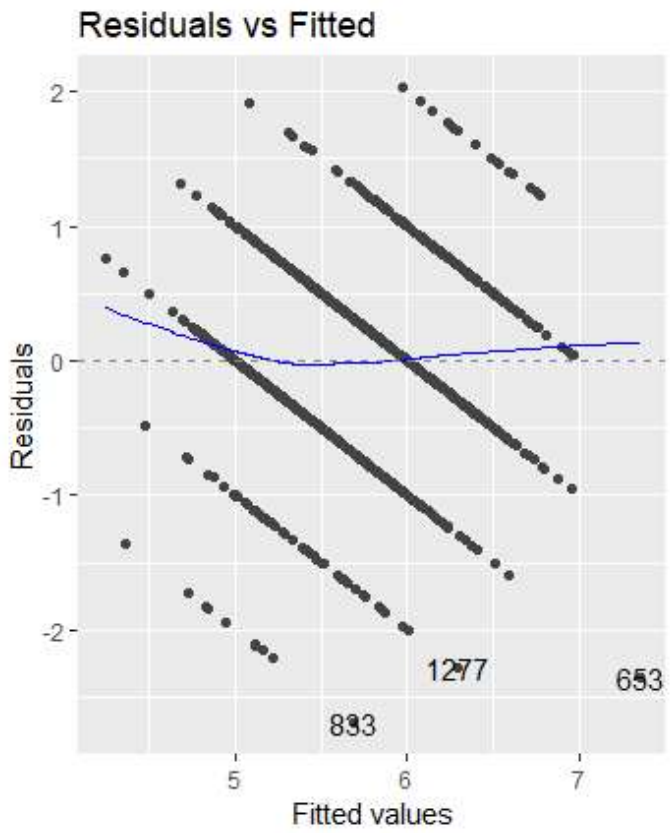
```
par(mfrow = c(2, 2))
autoplot(redwine_c2)
```



a) Linearity of the data

Hide

```
autoplot(redwine_c2,1)
```

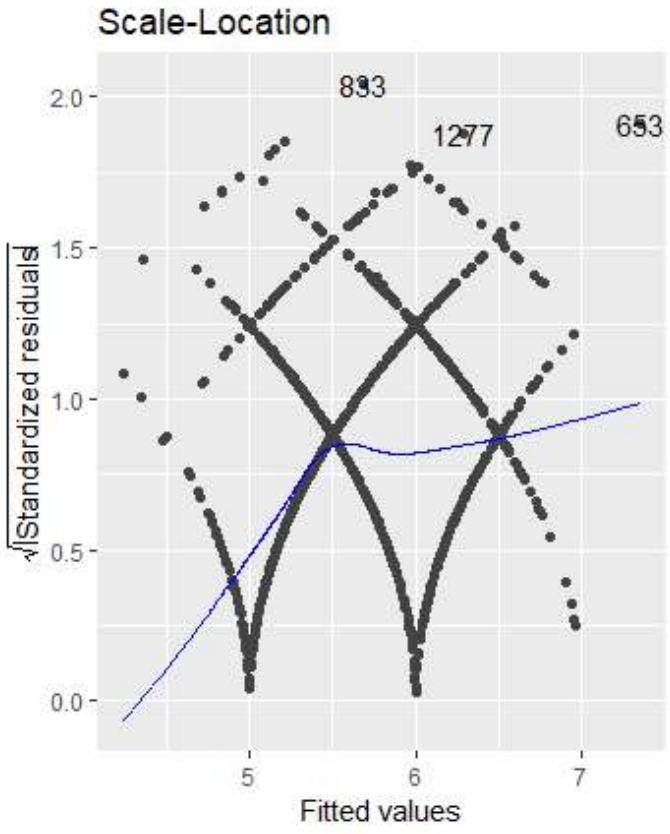


The residual plot indicates a linear relationship in the data with three outliers over the -2 range

b) Homogeneity of variance

Hide

```
autoplot(redwine_c2,3)
```

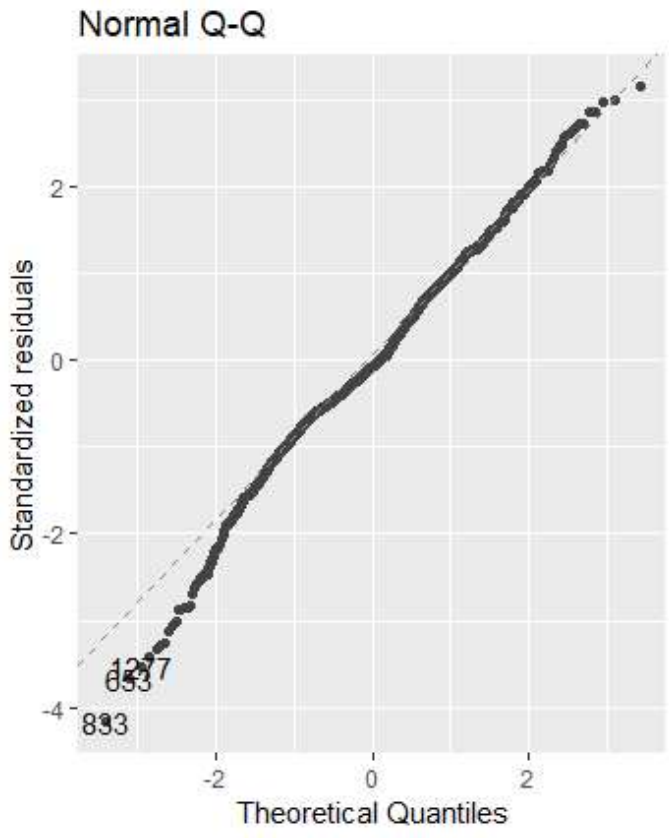


1. It can be seen that the variability (variances) of the residual points increases with the value of the fitted outcome variable, suggesting non-constant variances in the residual errors (or heteroscedasticity).
2. A possible solution to reduce the heteroscedasticity problem is to use a log or square root transformation of the outcome variable (quality).Points 833, 1277, 653 are detected as outliers, which can severely affect normality and homogeneity of variance. It could be useful to remove outliers to meet the test assumptions.

c) Normality of residuals

Hide

```
autoplot(redwine_c2,2)
```



Most of the points fall approximately along the reference line and the plot shows three outliers among which 833 is outside the -3 bound. Which could be highly influencing the data.

Shapiro-Wilk test

Hide

```
shapiro.test(redwine_c2$residuals)
```

```
    Shapiro-Wilk normality test

data:  redwine_c2$residuals
W = 0.99137, p-value = 4.321e-08
```

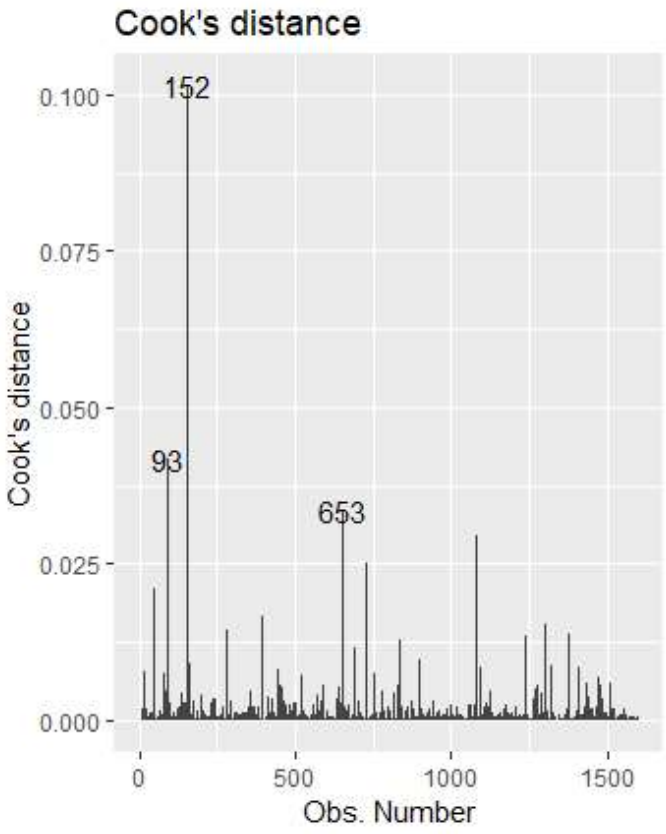
H0: Errors are normally distributed H1: Errors are not normally distributed

The conclusion above in homogeneity of variance is supported by the Shapiro-Wilk test on the residuals (W = 0.99137, p-value = 4.321e-08) with chosen alpha level 0.05, the p-value is less than 0.05, Hence, the null hypothesis that the data is normally distributed is rejected. However, this could be because of the extreme outlier as stated above.

d) Independence of residuals error terms

Hide

```
autoplot(redwine_c2, 4)
```



The plot above highlights the top 3 most extreme points (#93, #152 and #653), with standardized residuals above 0.025 range.

Non-constant variance test for Homoscedasticity

Hide

```
ncvTest(redwine_c2)
```

```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 20.10007, Df = 1, p = 7.3494e-06
```

H0: Errors have a constant variance H1: Errors have a non-constant variance

Since the p-value is < 0.05 we reject the H0. This implies that constant error variance assumption is violated.

Checking auto-correlation

Hide

```
durbinWatsonTest(redwine_c2)
```

```
 lag Autocorrelation D-W Statistic p-value
   1       0.1250136      1.749967       0
 Alternative hypothesis: rho != 0
```

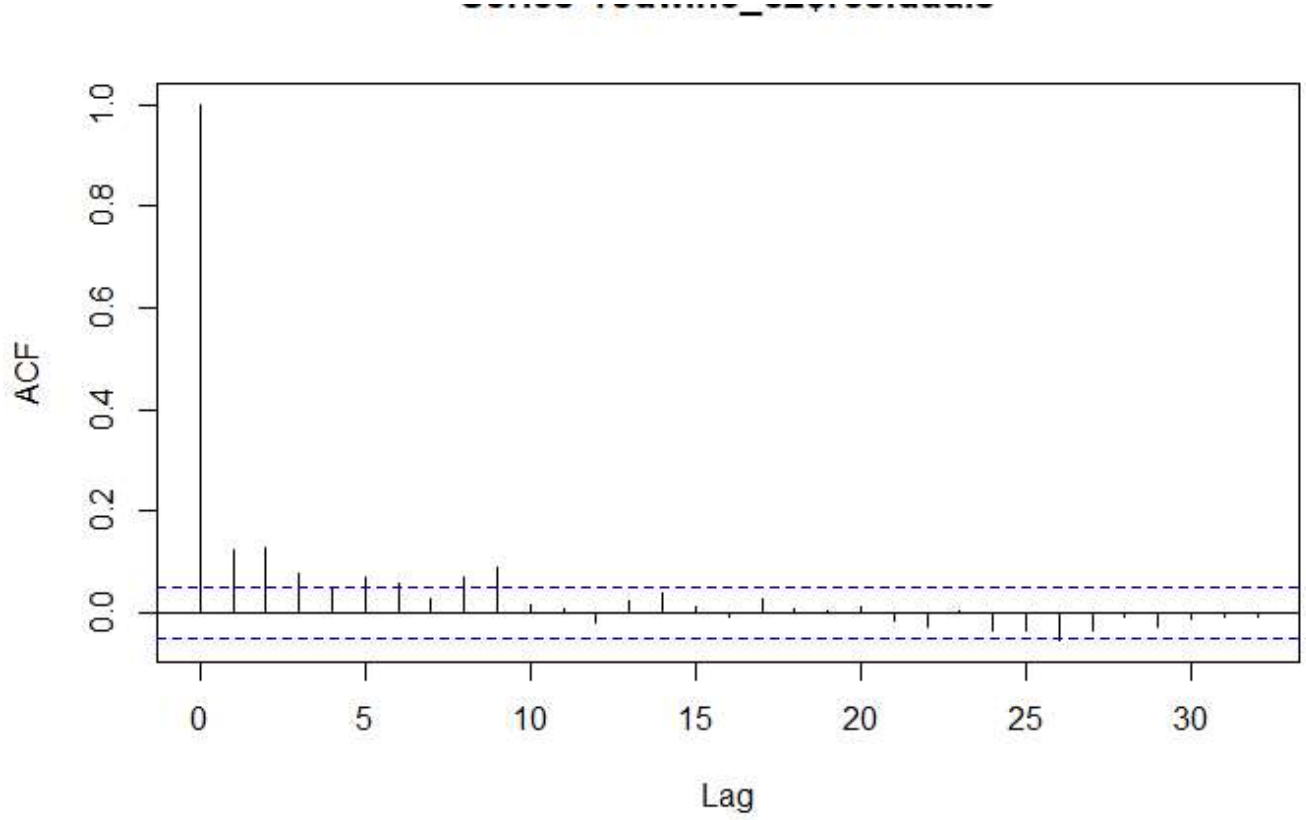
H0: Errors are uncorrelated H1: Errors are correlated

Since the p-value is < 0.05 we reject H0. This implies that uncorrelated error assumption is violated. However, the test statistic value is 1.749967 which is in the range of 1.5 to 2.5 and are relatively normal.

Test for Autocorrelated Errors

Hide

```
acf(redwine_c2$residuals)
```



There is one significant correlation at lag 0 and another slightly significant correlation at lag 2 in the residual series

Results:

Global Test of Model Assumptions: gvlma()

Model redwine_c2

Hide

```
gvmodel.redwine_c2 <- gvlma(redwine_c2)
summary(gvmodel.redwine_c2)
```

```
Call:
lm(formula = quality ~ volatile_acidity + chlorides + total_sulfur_dioxide +
    sulphates + alcohol + free_sulfur_dioxide + pH, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
volatile_acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
total_sulfur_dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
free_sulfur_dioxide   0.0050774  0.0021255   2.389    0.017 *
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = redwine_c2)
```

	Value <dbl>	p-value <dbl>	Decision <chr>
Global Stat	37.99895194	1.121117e-07	Assumptions NOT satisfied!
Skewness	6.45921911	1.103781e-02	Assumptions NOT satisfied!
Kurtosis	28.78544223	8.085660e-08	Assumptions NOT satisfied!
Link Function	2.70102751	1.002836e-01	Assumptions acceptable.
Heteroscedasticity	0.05326308	8.174795e-01	Assumptions acceptable.

5 rows

1. High positive kurtosis value indicates that the distribution has heavier tails than the normal distribution
2. Positive skew indicates that the tail is on the right which means that the distribution is asymmetrical

Model redwine_c1

Hide

```
gvmodel.redwine_c1 <- gvlma(redwine_c1)
summary(gvmodel.redwine_c1)
```

```
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


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = redwine_c1)
```

	Value <dbl>	p-value <dbl>	Decision <chr>
Global Stat	41.309014	2.319582e-08	Assumptions NOT satisfied!
Skewness	6.239798	1.249107e-02	Assumptions NOT satisfied!
Kurtosis	31.498912	1.995521e-08	Assumptions NOT satisfied!
Link Function	3.399982	6.519713e-02	Assumptions acceptable.
Heteroscedasticity	0.170322	6.798258e-01	Assumptions acceptable.

5 rows

1. High positive kurtosis value indicates that the distribution has heavier tails than the normal distribution
2. Positive skew indicates that the tail is on the right which means that the distribution is asymmetrical

GLM - 3rd model (redwine_c2.glm)

Hide

```
redwine_c2.glm <- glm(quality ~ volatile_acidity+chlorides+total_sulfur_dioxide+sulphates+alcohol+free_sulfur_dioxide+pH, data = redwine, family=gaussian(link = "log"))
summary(redwine_c2.glm)
```

```
Call:
glm(formula = quality ~ volatile_acidity + chlorides + total_sulfur_dioxide +
    sulphates + alcohol + free_sulfur_dioxide + pH, family = gaussian(link = "log"),
    data = redwine)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.68190  -0.36964  -0.05655   0.45908   2.09009

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.5210871  0.0703801  21.612  < 2e-16 ***
volatile_acidity     -0.1819815  0.0182697  -9.961  < 2e-16 ***
chlorides            -0.3623432  0.0729211  -4.969 7.46e-07 ***
total_sulfur_dioxide -0.0006913  0.0001253  -5.515 4.05e-08 ***
sulphates             0.1533076  0.0185816   8.250 3.27e-16 ***
alcohol               0.0492878  0.0028272  17.434  < 2e-16 ***
free_sulfur_dioxide   0.0009949  0.0003782   2.630  0.00861 **
pH                   -0.0803629  0.0205913  -3.903 9.91e-05 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

(Dispersion parameter for gaussian family taken to be 0.4188726)

    Null deviance: 1042.17  on 1598  degrees of freedom
Residual deviance:  666.43  on 1591  degrees of freedom
AIC: 3156.3

Number of Fisher Scoring iterations: 4
```

1. We have a value of 1042.17 on 1598 degrees of freedom. Including the independent variables (volatile_acidity, chlorides, total_sulfur_dioxide, sulphates, alcohol, free_sulfur_dioxide, pH) decreased the deviance to 666.43 points on 1591 degrees of freedom, a significant reduction in deviance.
2. Fisher's Scoring Algorithm needed four iterations to perform the fit.

Pseudo-R-squared

Hide

```
nagelkerke(redwine_c2.glm)
```

```
$Models

Model: "glm, quality ~ volatile_acidity + chlorides + total_sulfur_dioxide + sulphates + alco
hol + free_sulfur_dioxide + pH, gaussian(link = \"log\"), redwine"
Null:  "glm, quality ~ 1, gaussian(link = \"log\"), redwine"

$Pseudo.R.squared.for.model.vs.null
                            Pseudo.R.squared
McFadden                           0.185545
Cox and Snell (ML)                 0.360537
Nagelkerke (Cragg and Uhler)       0.396122

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq     p.value
      -7     -357.48 714.95 4.1134e-150

$Number.of.observations

Model: 1599
Null:  1599

$Messages
[1] "Note: For models fit with REML, these statistics are based on refitting with ML"

$Warnings
[1] "None"
```

The Pseudo R-squared with maximum likelyhood is 36%, not much of difference against our multiple linear regression model.

Update here produces null model for comparison

Hide

```
anova(redwine_c2.glm, update(redwine_c2.glm, ~1),test="Chisq")
```

```
Analysis of Deviance Table

Model 1: quality ~ volatile_acidity + chlorides + total_sulfur_dioxide +
    sulphates + alcohol + free_sulfur_dioxide + pH
Model 2: quality ~ 1
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1591     666.43
2      1598    1042.17 -7  -375.74 < 2.2e-16 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1
```

To get the significance for the overall model we use the following:

Hide

```
1-pchisq(1042.17-666.43, df=(1598-1591))
```

```
[1] 0
```

The value 0 is suggesting that the model is a significant improvement over the base model

Computing pseudo R-square

Hide

```
modelChi <- redwine_c2.glm$null.deviance - redwine_c2.glm$deviance
pseudo.R2 <- modelChi / redwine_c2.glm$null.deviance
pseudo.R2
```

```
[1] 0.3605367
```

36% of the variability in the data is explained by our model which is the same as multiple linear model.

Checking with trafo function as to which transformation suits better for our model

Hide

```
assumptions(object = redwine_c2)
```

```
The default lambdarange for the Log shift opt transformation is calculated dependent on the d
ata range. The lower value is set to -2 and the upper value to 5

The default lambdarange for the Square-root shift transformation is calculated dependent on t
he data range. The lower value is set to -2 and the upper value to 5

Test normality assumption
```

	Skewness <dbl>	Kurtosis <dbl>	Shapiro_W <dbl>	Shapiro_p <dbl>
untransformed	-0.1557	3.6573	0.9914	0
boxcox	-0.1966	3.7115	0.9904	0
dual	-0.1932	3.7223	0.9904	0
log	-0.7105	5.0791	0.9676	0
bickeldoksum	-0.1966	3.7115	0.9904	0
gpower	-0.1977	3.7089	0.9904	0
manly	-0.2219	3.7340	0.9898	0
modulus	-0.2014	3.7166	0.9903	0
logshiftopt	-0.4095	4.1010	0.9833	0
sqrtshift	-0.6198	4.8395	0.9725	0

1-10 of 14 rows Previous **1** 2 Next

```
Test homoscedasticity assumption
```

	BreuschPagan_V <dbl>	BreuschPagan_p <dbl>
untransformed	54.8155	0
boxcox	52.6210	0
dual	52.9668	0
log	48.8527	0
bickeldoksum	52.6209	0
gpower	52.5167	0
manly	51.2817	0
modulus	52.3714	0
logshiftopt	46.7420	0
sqrtshift	47.6205	0

1-10 of 14 rows Previous **1** 2 Next

```
Test linearity assumption
Press [enter] to continue
```

1. The untransformed model shows better in all segments such as skewness, kurtosis and normality test.
2. Followed by boxcox and dual transformation. Hence, we will try to boxcox transformation and see if we get better results

Boxcox transformation with skewness method

Hide

```
linMod_trafo2 <- trafo_lm(object = redwine_c2, trafo = "boxcox",method = "skew")
plot(linMod_trafo2)
```

The distribution of pearson residuals of data has improved slightly as seen in the histogram. However, there no visible difference in all other residual plots

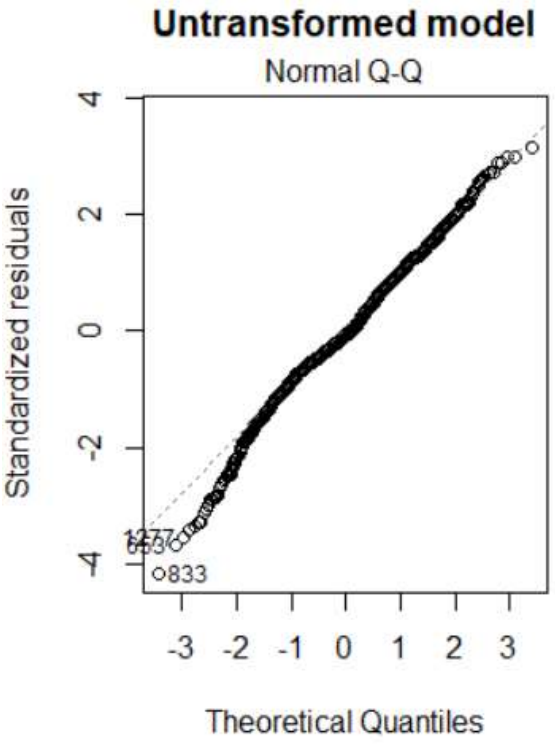
Transformed v. Untransformed summary

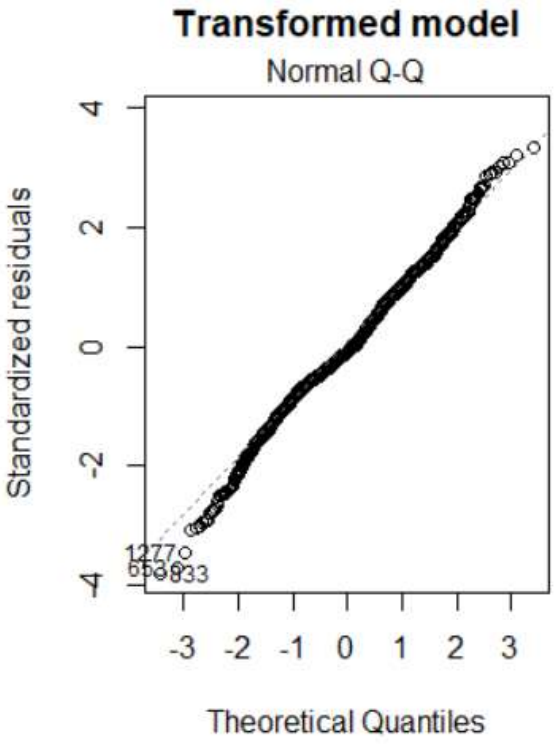
Hide

```
summary(linMod_trafo2)
```

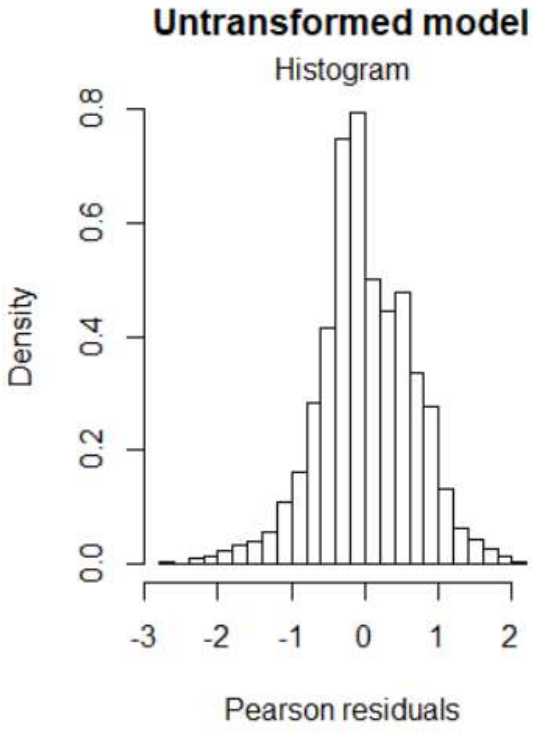
Transformed V. Untransformed plots:

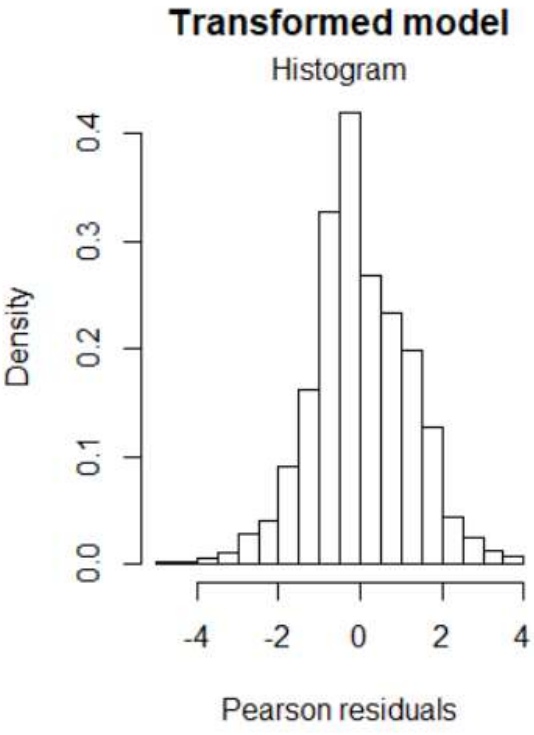
1. QQ Plot



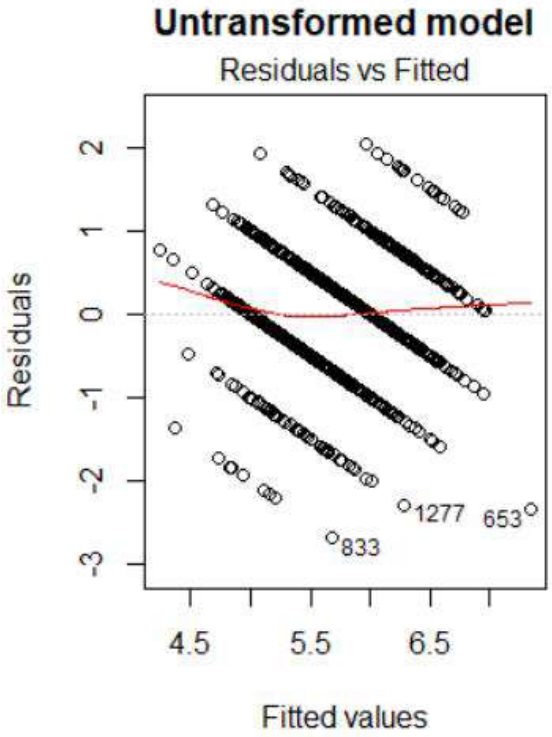


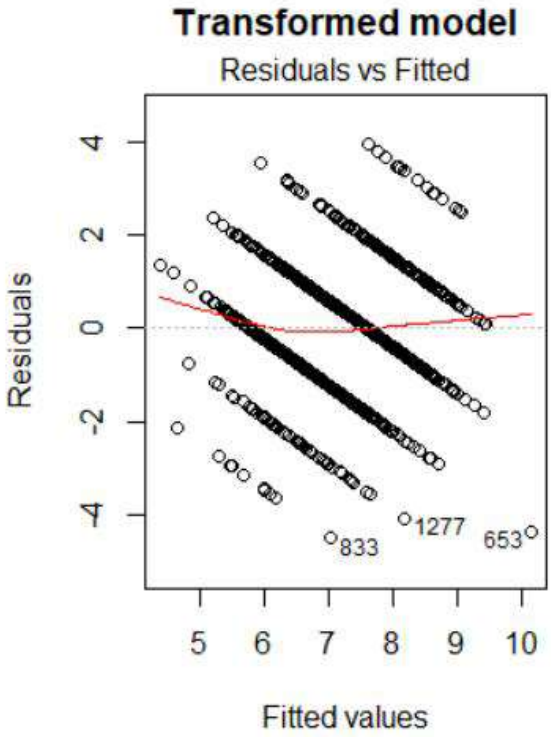
2. Histogram of Pearson residuals





3. Residual V. Fitted





```
Summary of untransformed model

Call:
lm(formula = quality ~ volatile_acidity + chlorides + total_sulfur_dioxide +
    sulphates + alcohol + free_sulfur_dioxide + pH, data = redwine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.68918 -0.36757 -0.04653  0.46081  2.02954

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.4300987  0.4029168  10.995  < 2e-16 ***
volatile_acidity     -1.0127527  0.1008429 -10.043  < 2e-16 ***
chlorides            -2.0178138  0.3975417  -5.076 4.31e-07 ***
total_sulfur_dioxide -0.0034822  0.0006868  -5.070 4.43e-07 ***
sulphates             0.8826651  0.1099084   8.031 1.86e-15 ***
alcohol               0.2893028  0.0167958  17.225  < 2e-16 ***
free_sulfur_dioxide   0.0050774  0.0021255   2.389    0.017 *
pH                   -0.4826614  0.1175581  -4.106 4.23e-05 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 0.6477 on 1591 degrees of freedom
Multiple R-squared:  0.3595,    Adjusted R-squared:  0.3567
F-statistic: 127.6 on 7 and 1591 DF,  p-value: < 2.2e-16


Summary of transformed model:  boxcox transformation

Formula in call:  qualityt ~ volatile_acidity + chlorides + total_sulfur_dioxide + sulphates
+ alcohol + free_sulfur_dioxide + pH

Call:
lm(formula = formula, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5192 -0.7003 -0.1064  0.8443  3.9523

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.652887   0.740490   6.284 4.26e-10 ***
volatile_acidity     -1.825208   0.185332  -9.848  < 2e-16 ***
chlorides            -3.714372   0.730612  -5.084 4.13e-07 ***
total_sulfur_dioxide -0.006515   0.001262  -5.162 2.76e-07 ***
sulphates             1.642192   0.201992   8.130 8.55e-16 ***
alcohol               0.542320   0.030868  17.569  < 2e-16 ***
free_sulfur_dioxide   0.009158   0.003906   2.344   0.0192 *
pH                   -0.901030   0.216051  -4.170 3.20e-05 ***
---
Signif. codes:  0 □***□ 0.001 □**□ 0.01 □*□ 0.05 □.□ 0.1 □ □ 1

Residual standard error: 1.19 on 1591 degrees of freedom
Multiple R-squared:  0.3635,    Adjusted R-squared:  0.3607
F-statistic: 129.8 on 7 and 1591 DF,  p-value: < 2.2e-16
```

There is very slight improvement in R-squared value against untransformed data, rest of the statistics don't have much difference. Hence, we will consider the untransformed model.

Discussion

1. Although the base model and our final model have similar R-square values, the variables in our final model redwine_c2 are highly significant compared against the base model.
2. Almost all the regression model selection methods suggested the variables in our final model.
3. Kurtosis and skewness is high in both multiple linear models redwine_c1 and redwine_c2.
4. The 3rd model redwine_c2.glm using the same variables with GLM also has high Kurtosis and skewness values.
5. The boxcox or any other transformation method shown in Trafo results doesn't improve our final untransformed model.
6. We could further remove outliers which are highly influencing the data as shown in our diagnostic plots.

Conclusion

In this report we evaluated the regression relationship between variables in our wine quality dataset and tried to identify through Regression analysis as to which variables have a relationship with the dependant variable (quality). The data is not normally distributed and is also skewed, to which even data transformations on our model didn't have any improvement. We could try fitting and testing censored regression model (tobit regression) using 'censReg' package.