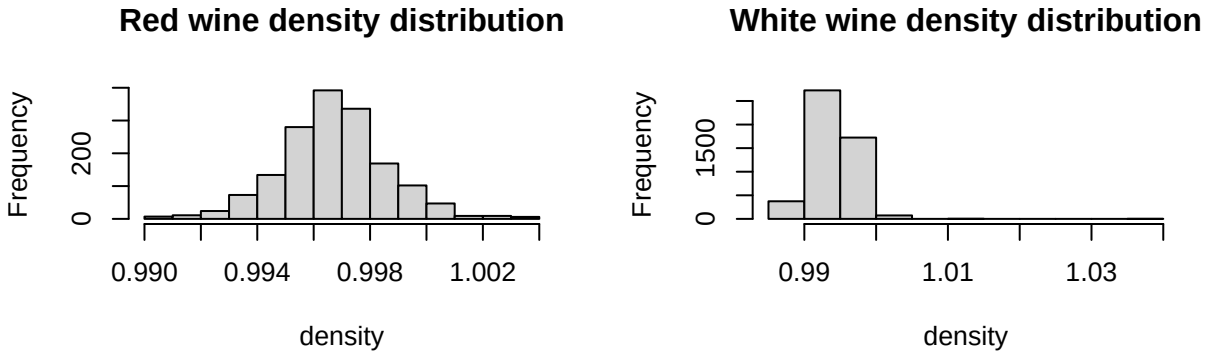
STA302 Project

#histogram for normality for red wine and white wine

```
red_wine <- read.table("./wine.csv", header=T, sep = ";")
white_wine <- read.table("./winequality-white.csv", header=T, sep = ";")
par(mfrow=c(2,2))

hist(red_wine$density, main = "Red wine density distribution", xlab = "density", ylab = "Frequency")
hist(white_wine$density, main = "White wine density distribution", xlab = "density", ylab = "Frequency")
```



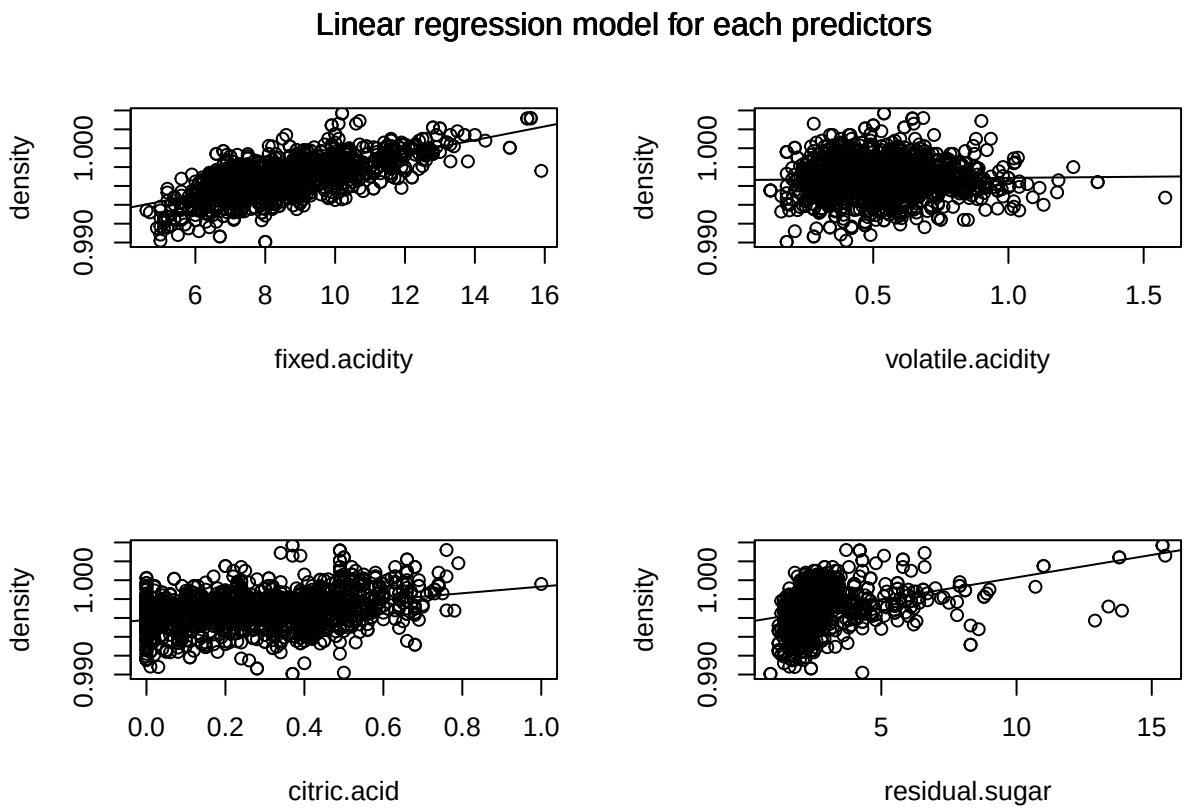
Reading data and fixing data frame

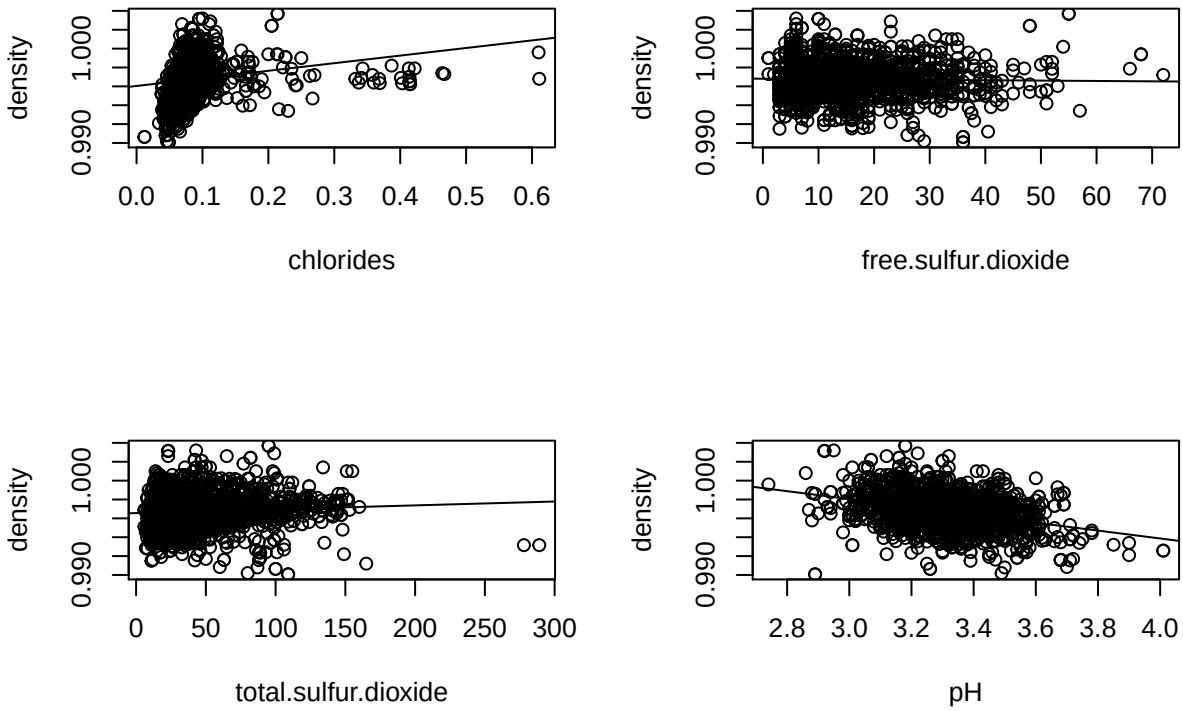
```
# reading table
red_wine <- read.table("./wine.csv", header=T, sep = ";")

# removing "quality" column since it is unnecessary for the research question
red_wine <- subset(red_wine, select = -c(quality) )

# reordering column order for convenience
```

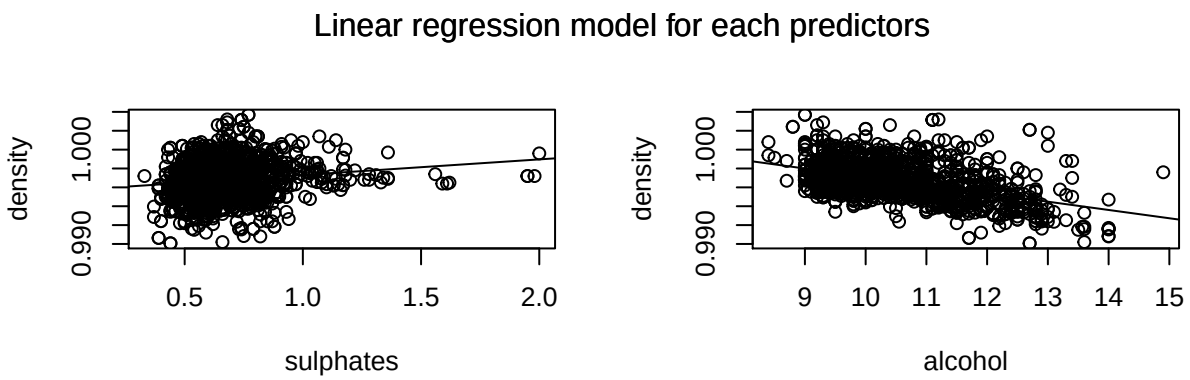
```
red_wine <- red_wine[, c(8,1,2,3,4,5,6,7,9,10,11)]
```

#simple linear regression for each parameter

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

for (i in c(2:11)){
  plot(red_wine[,i], red_wine$density, xlab = names(red_wine)[i],ylab = "density")
  abline(lsfit(red_wine[,i], red_wine$density))
mtext("Linear regression model for each predictors", side = 3, line = -2, outer = TRUE)

}
```



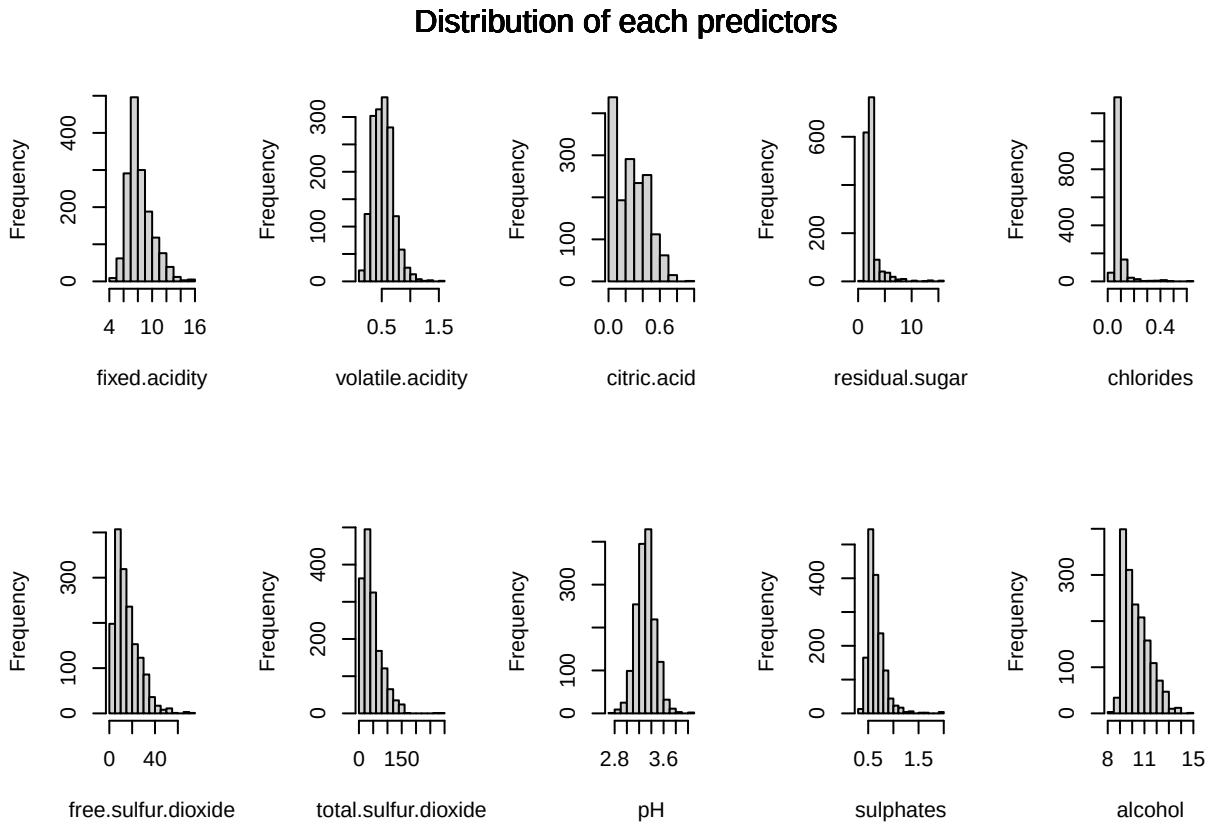

Linear regression model for each predictors
density
0.990
1.000
0.0
0.1
0.2
0.3
0.4
0.5
0.6
chlorides
density
0.990
1.000
0
10
20
30
40
50
60
70
free.sulfur.dioxide
density
0.990
1.000
0
50
100
150
200
250
300
total.sulfur.dioxide
density
0.990
1.000
2.8
3.0
3.2
3.4
3.6
3.8
4.0
pH



There are predictors that could fit a linear model based on simple linear regression model for each predictor with a response variable, density.

Distribution of each predictors

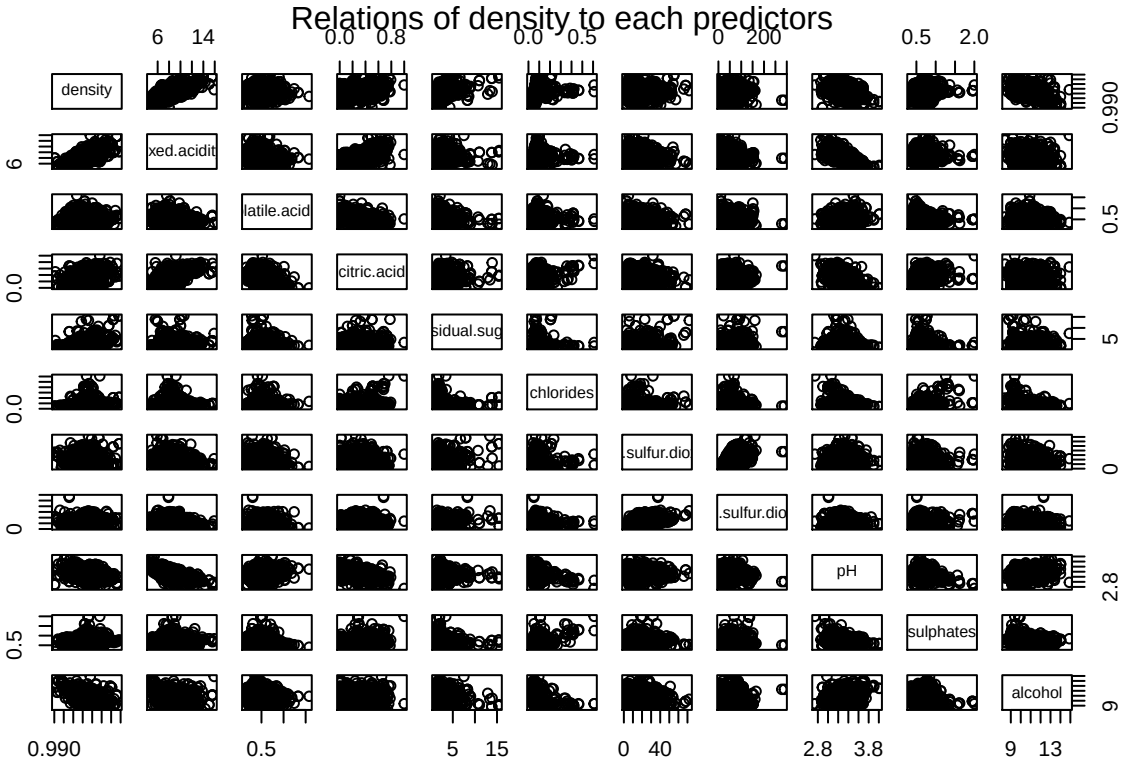
```
par(mfrow=c(2,5))
for (i in c(2:11)){
  hist(red_wine[,i], xlab = names(red_wine)[i], main = "")
  mtext("Distribution of each predictors", side = 3, line = -2, outer = TRUE)

}
```



Skewed predictors indicate a problem with normality violation, linearity, and poorly fitted model.

Relations of density to each predictors

```
pairs(density ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides + free.sulfu
mtext("Relations of density to each predictors", side = 3, line = -2, outer = TRUE)
```



Relation of density to each predictors indicate a problem with non-constant variance. However, we continue to find a fit model by reducing using ANOVA tests.

Full model

```
full_model = lm(density ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar + chlorides +
summary(full_model)
```

```
##
## Call:
## lm(formula = density ~ fixed.acidity + volatile.acidity + citric.acid +
##     residual.sugar + chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = red_wine)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.830e-03 -4.256e-04 -2.078e-05  4.240e-04  2.884e-03
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           9.793e-01  7.107e-04 1378.045  < 2e-16 ***
## fixed.acidity         9.421e-04  1.863e-05   50.571  < 2e-16 ***
## volatile.acidity      7.189e-04  1.393e-04    5.160 2.78e-07 ***
## citric.acid           5.849e-05  1.707e-04    0.343 0.731944
## residual.sugar        4.128e-04  1.398e-05   29.516  < 2e-16 ***
```

```
## chlorides             1.818e-03  4.842e-04    3.755 0.000180 ***
## free.sulfur.dioxide  -8.575e-06  2.509e-06   -3.417 0.000649 ***
## total.sulfur.dioxide  2.654e-06  8.427e-07    3.149 0.001667 **
## pH                    5.068e-03  1.823e-04   27.806  < 2e-16 ***
## sulphates             1.311e-03  1.285e-04   10.201  < 2e-16 ***
## alcohol              -9.253e-04  2.012e-05  -45.998  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0007517 on 1588 degrees of freedom
## Multiple R-squared:  0.8424, Adjusted R-squared:  0.8414
## F-statistic: 848.6 on 10 and 1588 DF,  p-value: < 2.2e-16
```

```
summary(full_model)$r.squared
```

```
## [1] 0.8423648
```

From the model output, linear relationship does exist between density and the parameters since p value is smaller than 0.05. Also, we can say that this model explains approximately 84.2% of variation originally present in the response density (84% of the total variation can be explained by these predictors in this way while 16% is still unexplained).

Removing insignificant parameter

```
# what if we decide to remove all non-significant predictors
reduced_model1 <- lm(density ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sul

# running an F-test
anova(reduced_model1, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
## Model 2: density ~ fixed.acidity + volatile.acidity + citric.acid + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1   1589 0.00089735
## 2   1588 0.00089728  1 6.6321e-08 0.1174 0.7319
```

Since citric.acid is not significant, we check if it is okay to remove the predictor. Based on the partial F-test, we see that full model is not so significant compared to reduced_model1. This indicates that it is okay to remove the predictor citric.acid.

Summary of reduced_model1

```
summary(reduced_model1)
```

```
##
## Call:
## lm(formula = density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
```

```
##     pH + sulphates + alcohol, data = red_wine)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.833e-03 -4.292e-04 -1.729e-05  4.232e-04  2.886e-03
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           9.793e-01  7.091e-04 1381.024  < 2e-16 ***
## fixed.acidity         9.456e-04  1.556e-05   60.767  < 2e-16 ***
## volatile.acidity      6.932e-04  1.173e-04    5.908 4.23e-09 ***
## residual.sugar        4.131e-04  1.394e-05   29.628  < 2e-16 ***
## chlorides             1.862e-03  4.667e-04    3.990 6.90e-05 ***
## free.sulfur.dioxide  -8.716e-06  2.475e-06   -3.522 0.000441 ***
## total.sulfur.dioxide  2.731e-06  8.120e-07    3.363 0.000789 ***
## pH                    5.066e-03  1.821e-04   27.819  < 2e-16 ***
## sulphates             1.312e-03  1.284e-04   10.219  < 2e-16 ***
## alcohol              -9.238e-04  1.962e-05  -47.072  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0007515 on 1589 degrees of freedom
## Multiple R-squared:  0.8424, Adjusted R-squared:  0.8415
## F-statistic: 943.4 on 9 and 1589 DF,  p-value: < 2.2e-16
```

```
summary(reduced_model1)$r.squared
```

```
## [1] 0.8423531
```

```
anova(reduced_model1)
```

```
## Analysis of Variance Table
##
## Response: density
##                        Df     Sum Sq    Mean Sq   F value    Pr(>F)
## fixed.acidity           1 0.00254032 0.00254032 4498.3472 < 2.2e-16 ***
## volatile.acidity        1 0.00022723 0.00022723  402.3682 < 2.2e-16 ***
## residual.sugar          1 0.00042762 0.00042762  757.2279 < 2.2e-16 ***
## chlorides               1 0.00006865 0.00006865  121.5662 < 2.2e-16 ***
## free.sulfur.dioxide     1 0.00000642 0.00000642   11.3699 0.0007644 ***
## total.sulfur.dioxide    1 0.00002828 0.00002828   50.0864 2.199e-12 ***
## pH                      1 0.00024216 0.00024216  428.8102 < 2.2e-16 ***
## sulphates               1 0.00000280 0.00000280    4.9664 0.0259841 *
## alcohol                 1 0.00125128 0.00125128 2215.7474 < 2.2e-16 ***
## Residuals            1589 0.00089735 0.00000056
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Only a small amount of information was lost based on the r squared value of reduced_model1. P-value of reduced_model1 is smaller than 0.05 so we can say it is significant.

Can other predictors be removed?

```
# getting rid of sulphates
reduced_model2 <- lm(density ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sul
```

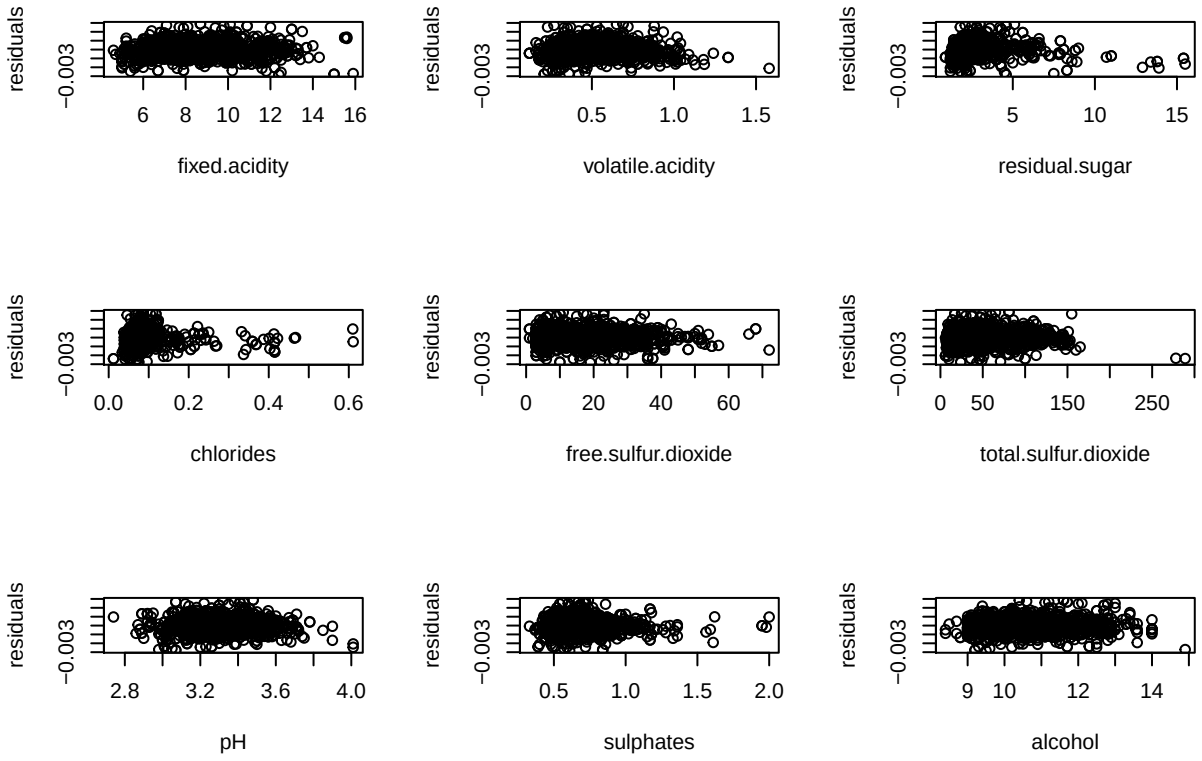
```r
anova(reduced_model2, reduced_model1)
```

```
## Analysis of Variance Table
##
## Model 1: density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + alcohol
## Model 2: density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   1590 0.00095631
## 2   1589 0.00089735  1 5.8967e-05 104.42 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

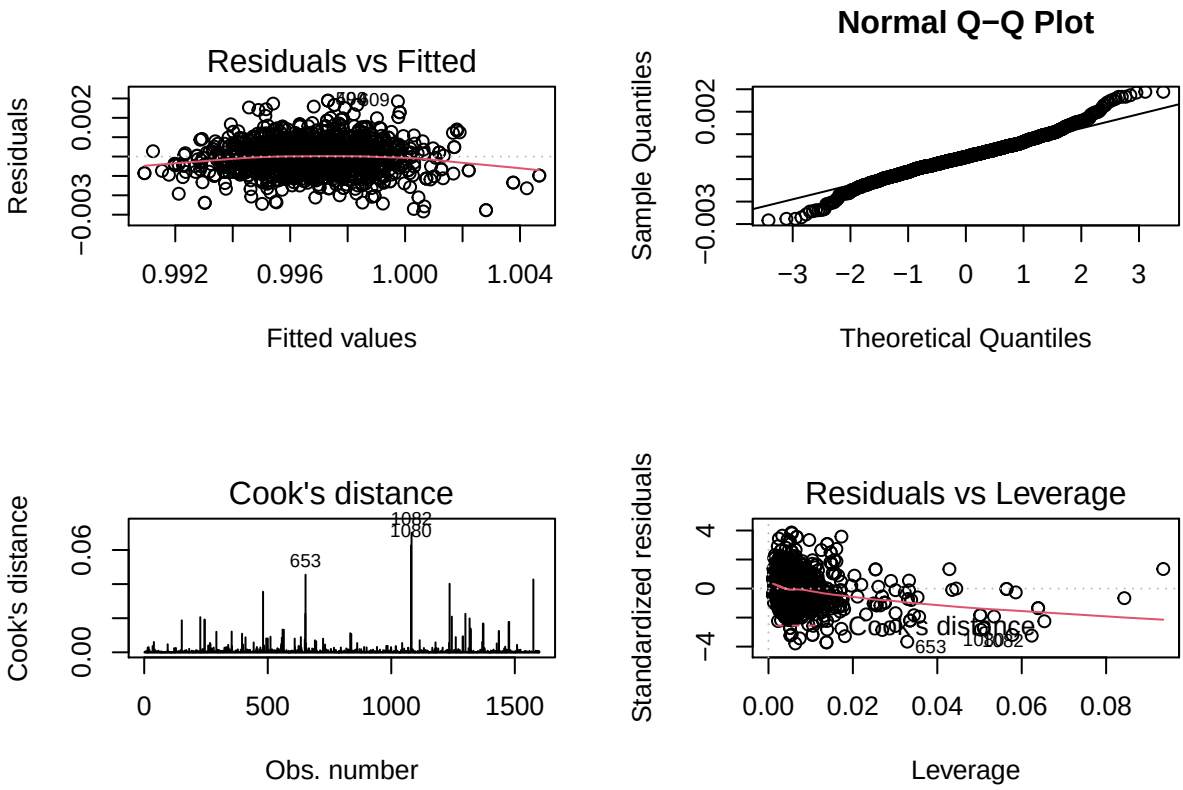
Since sulphate's variance in ANOVA test is non-significant, we check if it is okay to remove. However, the first reduced model is significant than the one removed. So we stick with the first reduced model.

Diagnosis tests on reduced_model1

```r
# checking to see linearity, uncorrelated errors, constant variance
# residual vs predictors
par(mfrow=c(3,3))
plot(reduced_model1$residuals ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sul

mtext("Residuals of reduced model vs Predictors", side = 3, line = -1.5, outer = TRUE)
```




```r
par(mfrow=c(2,2))

# residual vs fitted values
plot(reduced_model1, which = 1)

# qq plot
qqnorm(reduced_model1$residuals)
qqline(reduced_model1$residuals)

# cook's distance and leverage point graph
plot(reduced_model1, which = 4)
plot(reduced_model1, which = 5)
```



Linearity is broken for some predictors such as residual.sugar, chlorides, total.sulfur.dioxide, and sulphates. For some predictors such as volatile.acidity, and free.sulfur.dioxide, linearity is a slightly broken. We can also see some that some residual values is separated from the cluster. So we can say that there are uncorrelated errors.

Fitted value and residual plot shows that there might be a problem with the constant variance since the red line is curved and not extremely close to the horizontal line, residual = 0.

The qq plot is close to a straight line, though we can see it is fat-tailed. So we can say that there is more data at the extremes of the distribution and less data in the centre.

Removing outliers and large residuals

```
# using cook's distance and leverage point graph
cooksd <- cooks.distance(reduced_model1)
influential <- as.numeric(names(cooks.distance(reduced_model1))[(cooksd > (4/nrow(red_wine)))])
largeresiduals <- rstudent(reduced_model1) > 3

red_wine2 <- red_wine[-influential,]
red_wine3 <- red_wine2[-largeresiduals,]
```

Summary of fixed_model1

```
# residuals vs predictors
par(mfrow=c(3,3))
```

```
fixed_model1 <- lm(density ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfu
summary(fixed_model1)
```

```
##
## Call:
## lm(formula = density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = red_wine3)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.769e-03 -3.983e-04 -9.340e-06  3.736e-04  2.087e-03
##
## Coefficients:
##                        Estimate Std. Error  t value Pr(>|t|)
## (Intercept)           9.805e-01  6.231e-04 1573.504  < 2e-16 ***
## fixed.acidity         9.193e-04  1.390e-05   66.156  < 2e-16 ***
## volatile.acidity      5.117e-04  1.031e-04    4.962 7.78e-07 ***
## residual.sugar        4.921e-04  1.640e-05   30.015  < 2e-16 ***
## chlorides             2.388e-03  4.587e-04    5.206 2.20e-07 ***
## free.sulfur.dioxide  -5.629e-06  2.146e-06   -2.623  0.00881 **
## total.sulfur.dioxide  1.898e-06  7.170e-07    2.648  0.00818 **
## pH                    4.887e-03  1.576e-04   31.012  < 2e-16 ***
## sulphates             1.276e-03  1.135e-04   11.240  < 2e-16 ***
## alcohol              -9.676e-04  1.719e-05  -56.291  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005992 on 1479 degrees of freedom
## Multiple R-squared:  0.8807, Adjusted R-squared:  0.8799
## F-statistic:  1213 on 9 and 1479 DF,  p-value: < 2.2e-16
```

```
summary(fixed_model1)$r.squared
```

```
## [1] 0.8806516
```

```
anova(fixed_model1)
```

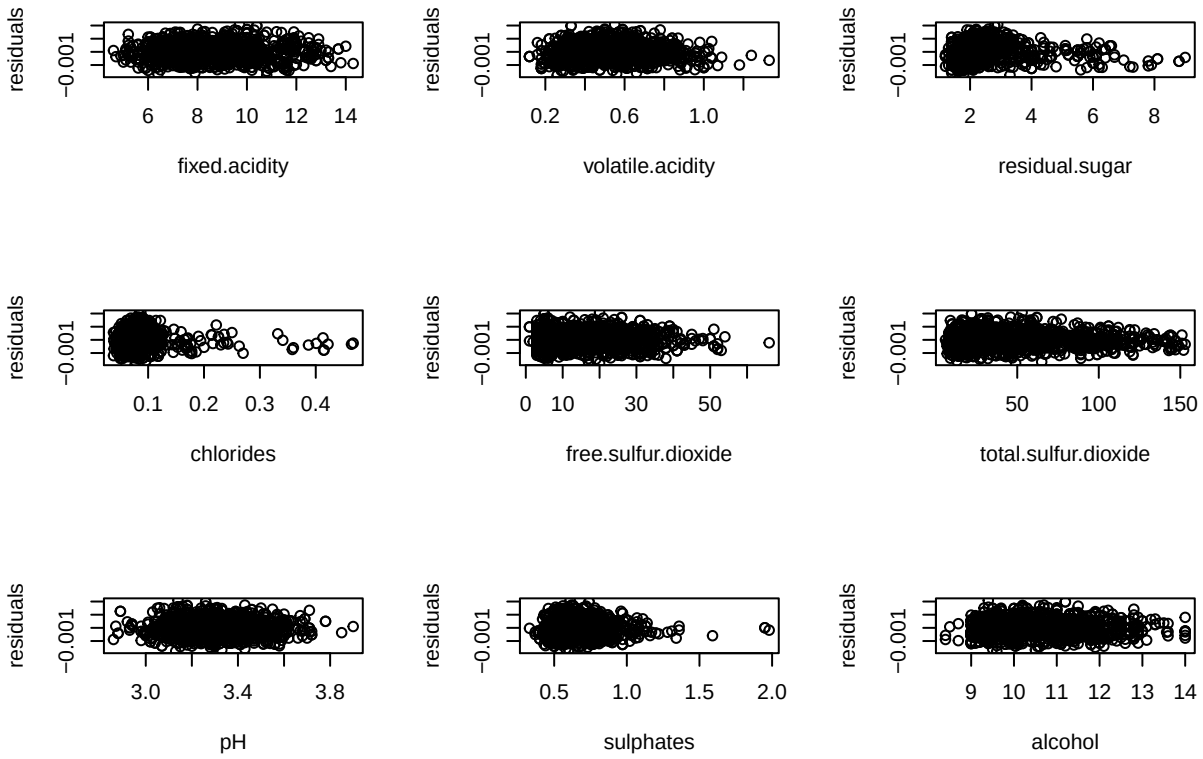
```
## Analysis of Variance Table
##
## Response: density
##                        Df     Sum Sq    Mean Sq   F value    Pr(>F)
## fixed.acidity           1 0.00201712 0.00201712 5618.2431 < 2.2e-16 ***
## volatile.acidity        1 0.00018079 0.00018079  503.5504 < 2.2e-16 ***
## residual.sugar          1 0.00023697 0.00023697  660.0242 < 2.2e-16 ***
## chlorides               1 0.00007423 0.00007423  206.7526 < 2.2e-16 ***
## free.sulfur.dioxide     1 0.00001251 0.00001251   34.8422 4.427e-09 ***
## total.sulfur.dioxide    1 0.00005765 0.00005765  160.5708 < 2.2e-16 ***
## pH                      1 0.00020123 0.00020123  560.4710 < 2.2e-16 ***
## sulphates               1 0.00000004 0.00000004    0.1135    0.7363
## alcohol                 1 0.00113766 0.00113766 3168.7185 < 2.2e-16 ***
## Residuals            1479 0.00053101 0.00000036
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the p-value, it is valid. Also, r-square value has increased which tells us that it is significant to

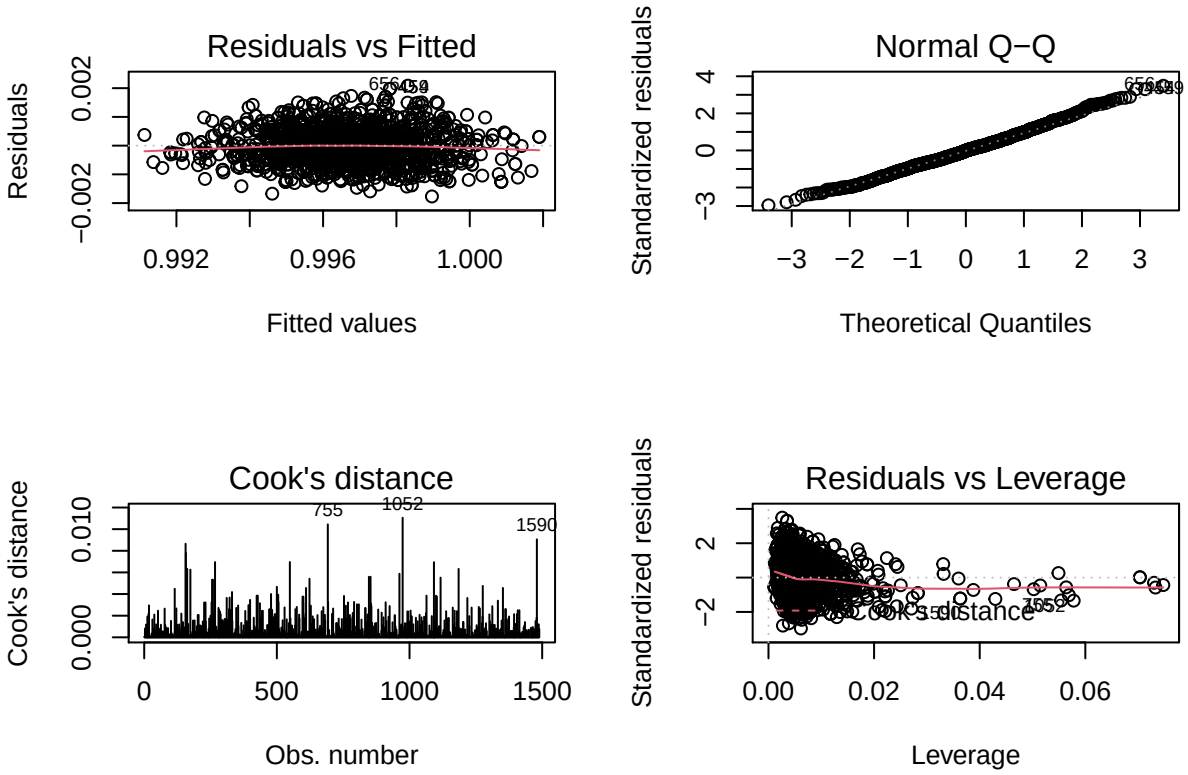
remove outliers. Let's check on the plots.

Diagnostic tests on fixed_model1

```
par(mfrow=c(3,3))
plot(fixed_model1$residuals ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfu
```




```
par(mfrow=c(2,2))
# residual vs fitted values
plot(fixed_model1, which = 1)

# qq plot
plot(fixed_model1, which = 2)

# cook's distance and leverage point graph
plot(fixed_model1, which = 4)
plot(fixed_model1, which = 5)
```

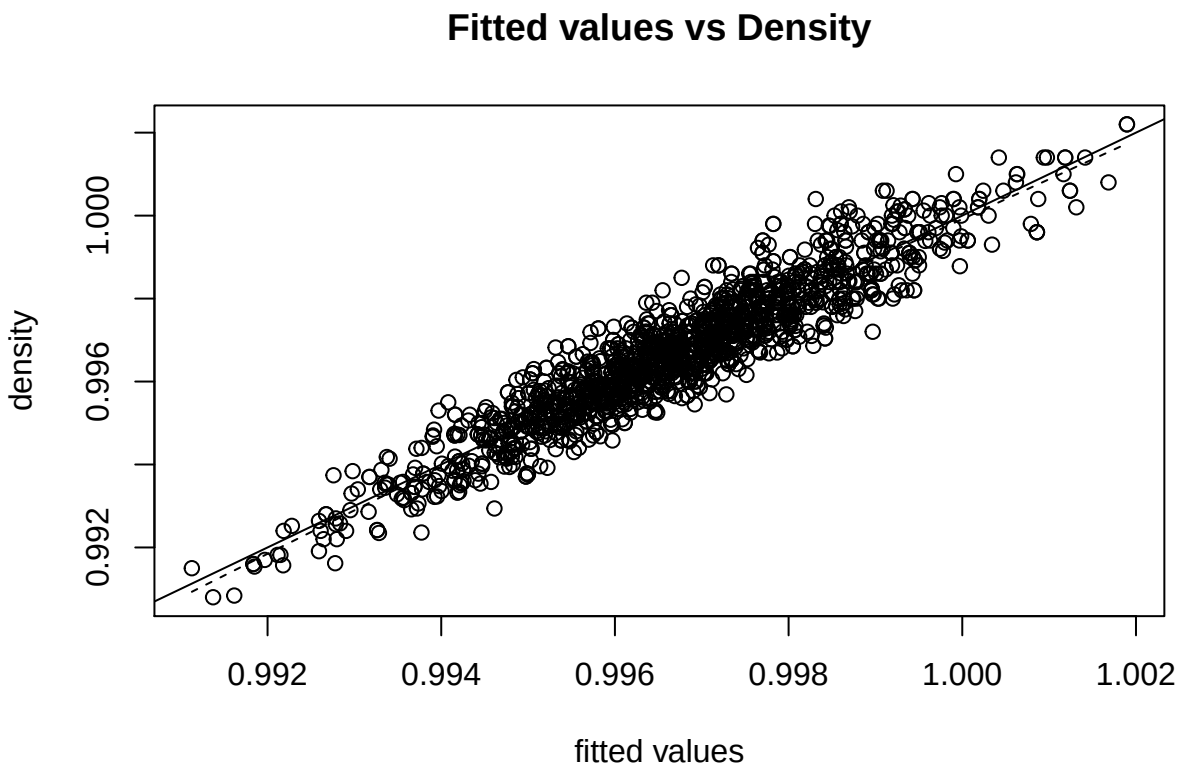


After removing outliers, high leverage points, and large residual values, it is clearly shown that the red line of residuals vs fitted graph is much closer to the horizontal line. From here, we can say that the it has constant variance. From the qq plot, the problem with the fat tailed has been solved.

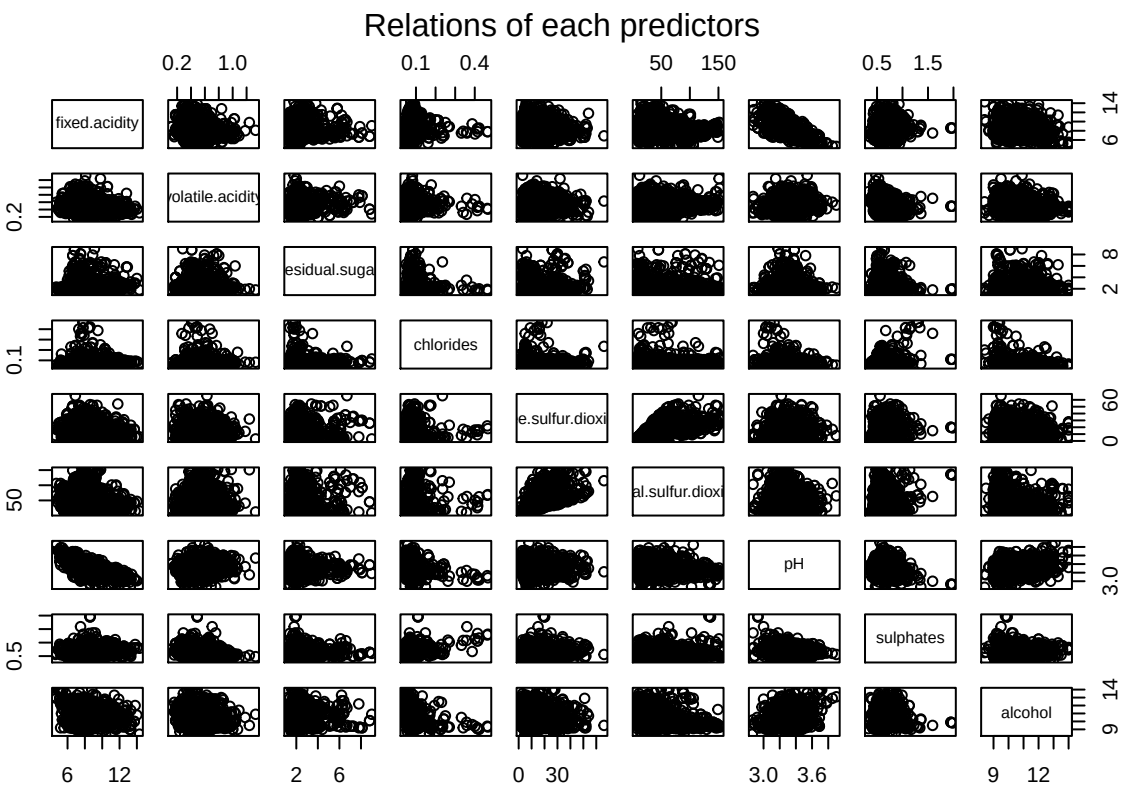
To conclude, we have fixed the problem of constant variance and uncorrelated errors. However, the plot of residuals vs predictors still shows that there is a problem with linearity. Let's fix that now.

Checking conditions of fixed_model

```
# check if conditional mean response is a single function of a linear combination of the predictors
fit <- fixed_model1$fitted.values
plot(red_wine3$density ~ fit, main = "Fitted values vs Density", xlab = "fitted values", ylab = "densit
abline(a = 0, b = 1)
lines(lowess(red_wine3$density ~ fit), lty=2)
```




```
# check if conditional mean of each predictor is a linear function with another predictor
fit <- fixed_model1$fitted.values
pairs(~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfur.dioxide + total.sul

mtext("Relations of each predictors", side = 3, line = -1.5, outer = TRUE)
```



No problem with the 2 additional conditions.

Summary of final_model

```
red_wine4 <- subset(red_wine3, select = -c(citric.acid) )

for (i in c(2:10)){
  red_wine4[,i] <- log(red_wine4[,i])
}

final_model <- lm(density ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfur
summary(final_model)
```

```
##
## Call:
## lm(formula = density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = red_wine4)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.605e-03 -4.094e-04 -2.415e-05  3.686e-04  1.859e-03
##
## Coefficients:
##                         Estimate Std. Error  t value Pr(>|t|)
```

```
## (Intercept)           9.848e-01  9.096e-04 1082.656  < 2e-16 ***
## fixed.acidity         7.683e-03  1.224e-04   62.774  < 2e-16 ***
## volatile.acidity      3.343e-04  5.184e-05    6.449 1.52e-10 ***
## residual.sugar        1.625e-03  5.341e-05   30.431  < 2e-16 ***
## chlorides             3.083e-04  6.061e-05    5.086 4.12e-07 ***
## free.sulfur.dioxide  -1.648e-04  3.781e-05   -4.359 1.40e-05 ***
## total.sulfur.dioxide  1.419e-04  3.846e-05    3.688 0.000234 ***
## pH                    1.590e-02  5.203e-04   30.558  < 2e-16 ***
## sulphates             1.116e-03  8.152e-05   13.695  < 2e-16 ***
## alcohol              -9.894e-03  1.909e-04  -51.833  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005918 on 1479 degrees of freedom
## Multiple R-squared:  0.8836, Adjusted R-squared:  0.8829
## F-statistic:  1247 on 9 and 1479 DF,  p-value: < 2.2e-16
```

```
summary(final_model)$r.squared
```

```
## [1] 0.8835757
```

```
anova(final_model)
```

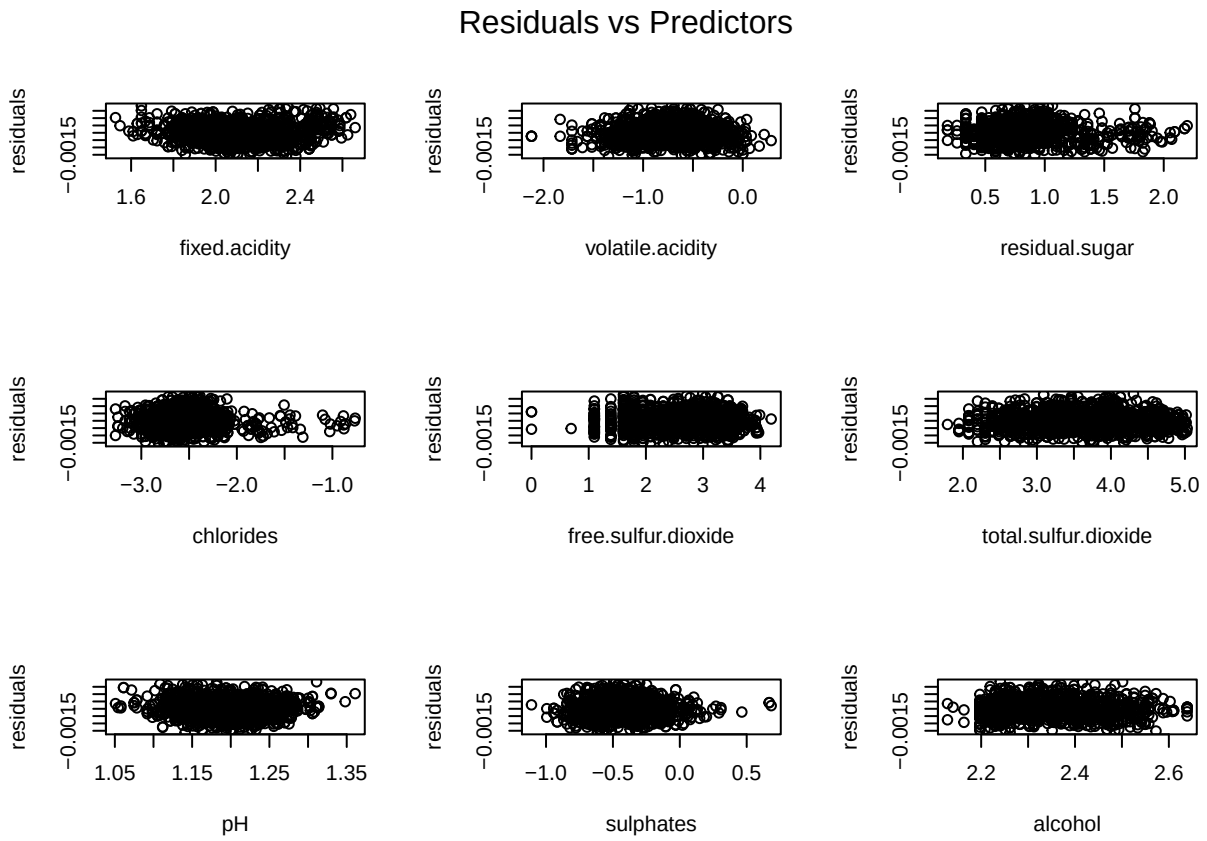
```
## Analysis of Variance Table
##
## Response: density
##                        Df     Sum Sq    Mean Sq   F value    Pr(>F)
## fixed.acidity           1 0.00205730 0.00205730 5874.0892 < 2.2e-16 ***
## volatile.acidity        1 0.00021856 0.00021856  624.0513 < 2.2e-16 ***
## residual.sugar          1 0.00025034 0.00025034  714.7665 < 2.2e-16 ***
## chlorides               1 0.00011817 0.00011817  337.4058 < 2.2e-16 ***
## free.sulfur.dioxide     1 0.00001119 0.00001119   31.9564 1.889e-08 ***
## total.sulfur.dioxide    1 0.00009227 0.00009227  263.4619 < 2.2e-16 ***
## pH                      1 0.00024103 0.00024103  688.2111 < 2.2e-16 ***
## sulphates               1 0.00000136 0.00000136    3.8842   0.04893 *
## alcohol                 1 0.00094097 0.00094097 2686.7047 < 2.2e-16 ***
## Residuals            1479 0.00051800 0.00000035
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
confint(final_model, conf.level=0.95)
```

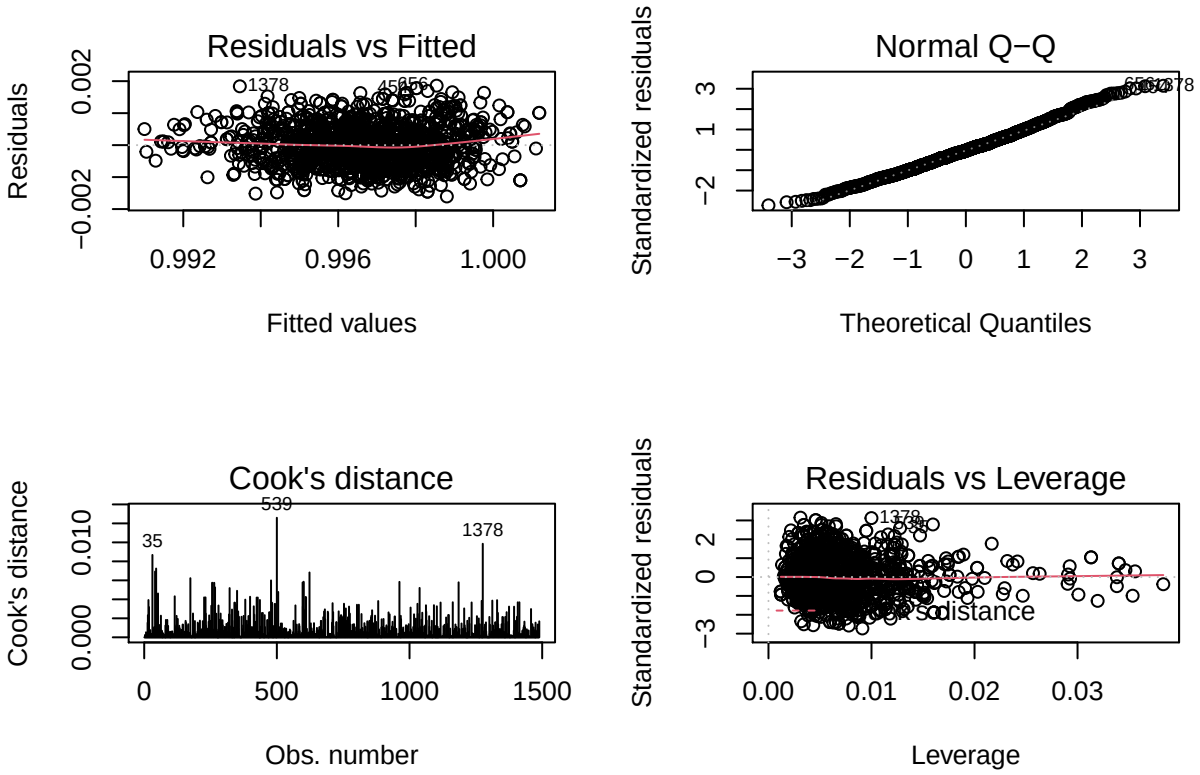
```
##                             2.5 %        97.5 %
## (Intercept)           9.829824e-01  9.865508e-01
## fixed.acidity         7.443151e-03  7.923323e-03
## volatile.acidity      2.326366e-04  4.360103e-04
## residual.sugar        1.520590e-03  1.730130e-03
## chlorides             1.893703e-04  4.271467e-04
## free.sulfur.dioxide  -2.389864e-04 -9.063813e-05
## total.sulfur.dioxide  6.642193e-05  2.173222e-04
## pH                    1.487753e-02  1.691859e-02
## sulphates             9.565314e-04  1.276340e-03
## alcohol              -1.026839e-02 -9.519545e-03
```

Diagnostic tests of final_model

```
# residuals vs predictors
par(mfrow=c(3,3))
plot(final_model$residuals ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulfu
mtext("Residuals vs Predictors", side = 3, line = -1.5, outer = TRUE)
```

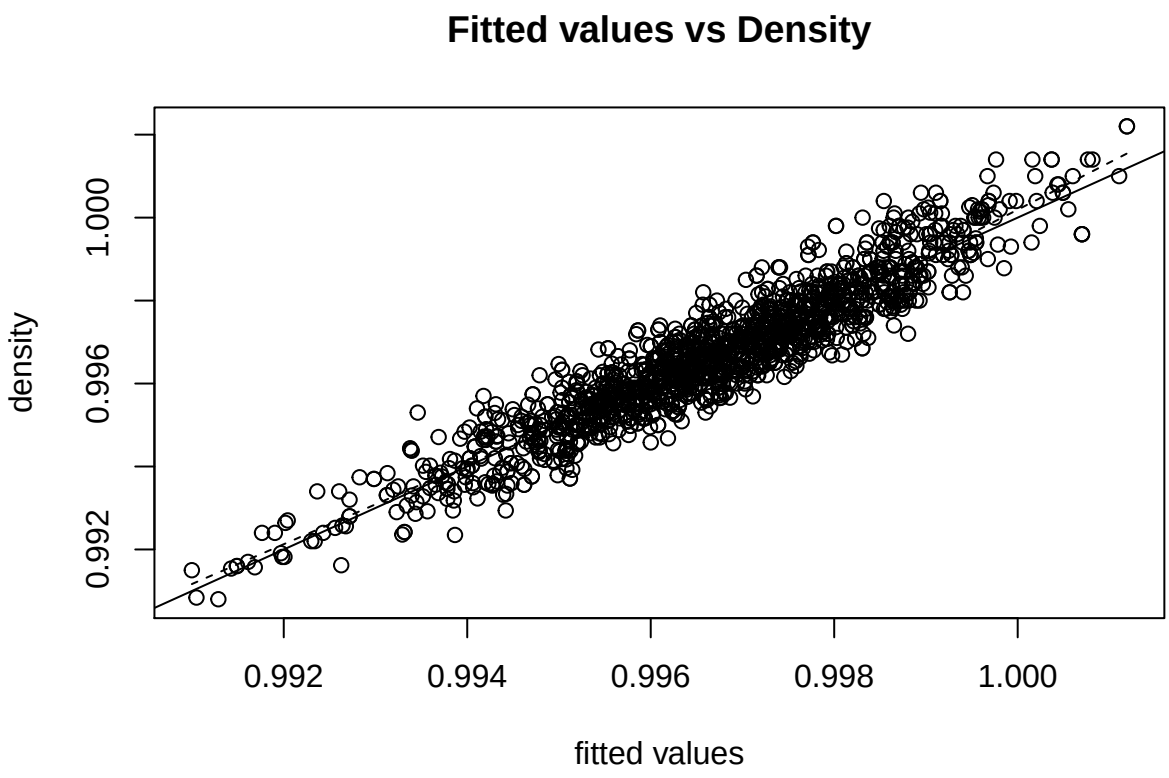



```
par(mfrow=c(2,2))
# residual vs fitted values
plot(final_model, which = 1)

# qq plot
plot(final_model, which = 2)

# cook's distance and leverage point graph
plot(final_model, which = 4)
plot(final_model, which = 5)
```

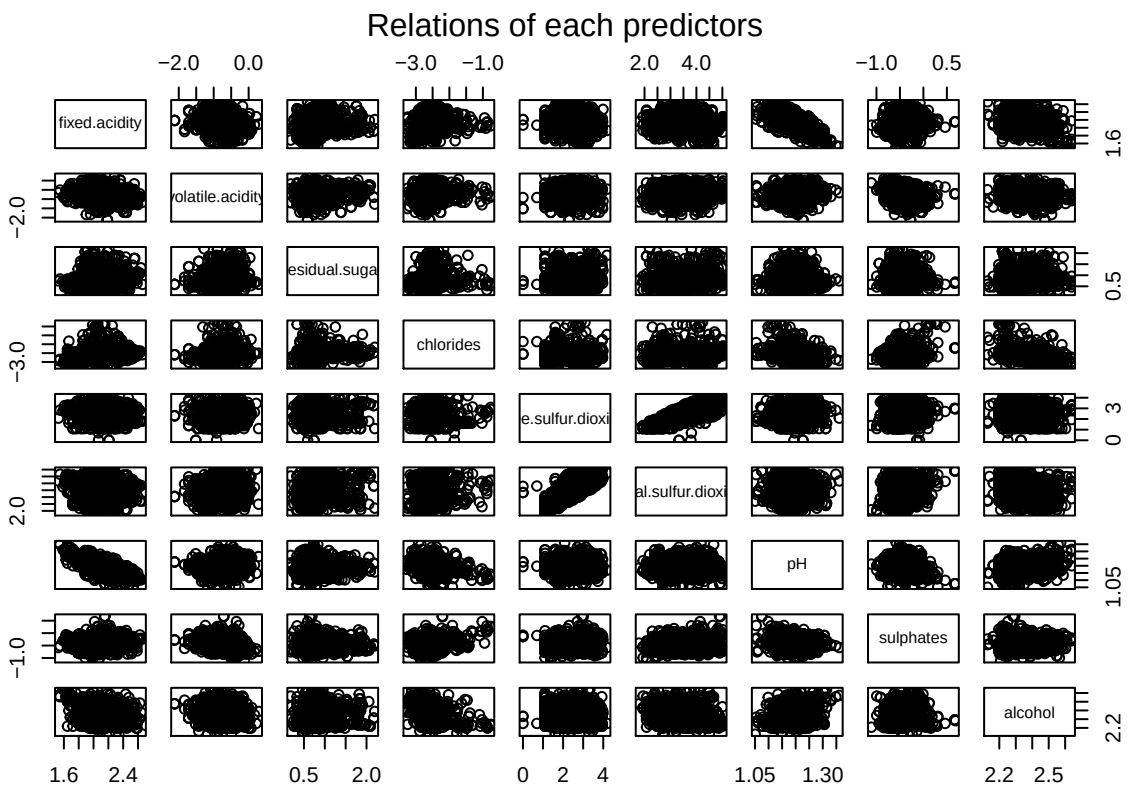


Checking conditions of final_model

```
# check if conditional mean response is a single function of a linear combination of the predictors
fit <- final_model$fitted.values
plot(red_wine4$density ~ fit, main = "Fitted values vs Density", xlab = "fitted values", ylab = "density
abline(a = 0, b = 1)
lines(lowess(red_wine4$density ~ fit), lty=2)
```




```
# check if conditional mean of each predictor is a linear function with another predictor
pairs(red_wine4[,2:10])
mtext("Relations of each predictors", side = 3, line = -1.5, outer = TRUE)
```



Validating model with split data

```
# creating two independent dataset (training set, test dataset by 75:25 ratio)

splitting_data <- sort(sample(nrow(red_wine4), nrow(red_wine4)*.75))
train <- red_wine4[splitting_data,]
test <- red_wine4[-splitting_data,]
```

Comparing two data set's summary

```
# comparing training dataset and test data set's explanatory analysis

train_validate <- lm(density ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sul
test_validate <- lm(density ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sulf

summary(train_validate)
```

```
##
## Call:
## lm(formula = density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = train)
##
## Residuals:
```

```
##        Min         1Q     Median         3Q        Max
## -1.521e-03 -4.169e-04 -2.646e-05  3.704e-04  1.865e-03
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.836e-01  1.052e-03 935.379  < 2e-16 ***
## fixed.acidity         7.791e-03  1.413e-04  55.149  < 2e-16 ***
## volatile.acidity      3.271e-04  5.949e-05   5.498 4.76e-08 ***
## residual.sugar        1.642e-03  6.191e-05  26.527  < 2e-16 ***
## chlorides             2.879e-04  7.131e-05   4.038 5.77e-05 ***
## free.sulfur.dioxide  -1.662e-04  4.316e-05  -3.851 0.000125 ***
## total.sulfur.dioxide  1.518e-04  4.400e-05   3.451 0.000580 ***
## pH                    1.661e-02  5.996e-04  27.704  < 2e-16 ***
## sulphates             1.075e-03  9.448e-05  11.374  < 2e-16 ***
## alcohol              -9.896e-03  2.236e-04 -44.265  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005942 on 1106 degrees of freedom
## Multiple R-squared:  0.8822, Adjusted R-squared:  0.8812
## F-statistic: 920.4 on 9 and 1106 DF,  p-value: < 2.2e-16
```

```
summary(test_validate)
```

```
##
## Call:
## lm(formula = density ~ fixed.acidity + volatile.acidity + residual.sugar +
##     chlorides + free.sulfur.dioxide + total.sulfur.dioxide +
##     pH + sulphates + alcohol, data = test)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -1.524e-03 -3.836e-04 -3.711e-05  3.668e-04  1.636e-03
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.886e-01  1.846e-03 535.637  < 2e-16 ***
## fixed.acidity         7.318e-03  2.502e-04  29.242  < 2e-16 ***
## volatile.acidity      3.711e-04  1.078e-04   3.443 0.000643 ***
## residual.sugar        1.565e-03  1.092e-04  14.342  < 2e-16 ***
## chlorides             3.315e-04  1.163e-04   2.850 0.004625 **
## free.sulfur.dioxide  -1.531e-04  8.059e-05  -1.900 0.058223 .
## total.sulfur.dioxide  1.059e-04  8.082e-05   1.311 0.190716
## pH                    1.357e-02  1.062e-03  12.772  < 2e-16 ***
## sulphates             1.291e-03  1.643e-04   7.854 4.61e-14 ***
## alcohol              -9.877e-03  3.691e-04 -26.760  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0005849 on 363 degrees of freedom
## Multiple R-squared:  0.8901, Adjusted R-squared:  0.8874
## F-statistic: 326.7 on 9 and 363 DF,  p-value: < 2.2e-16
```

```
summary(train_validate)$r.squared
```

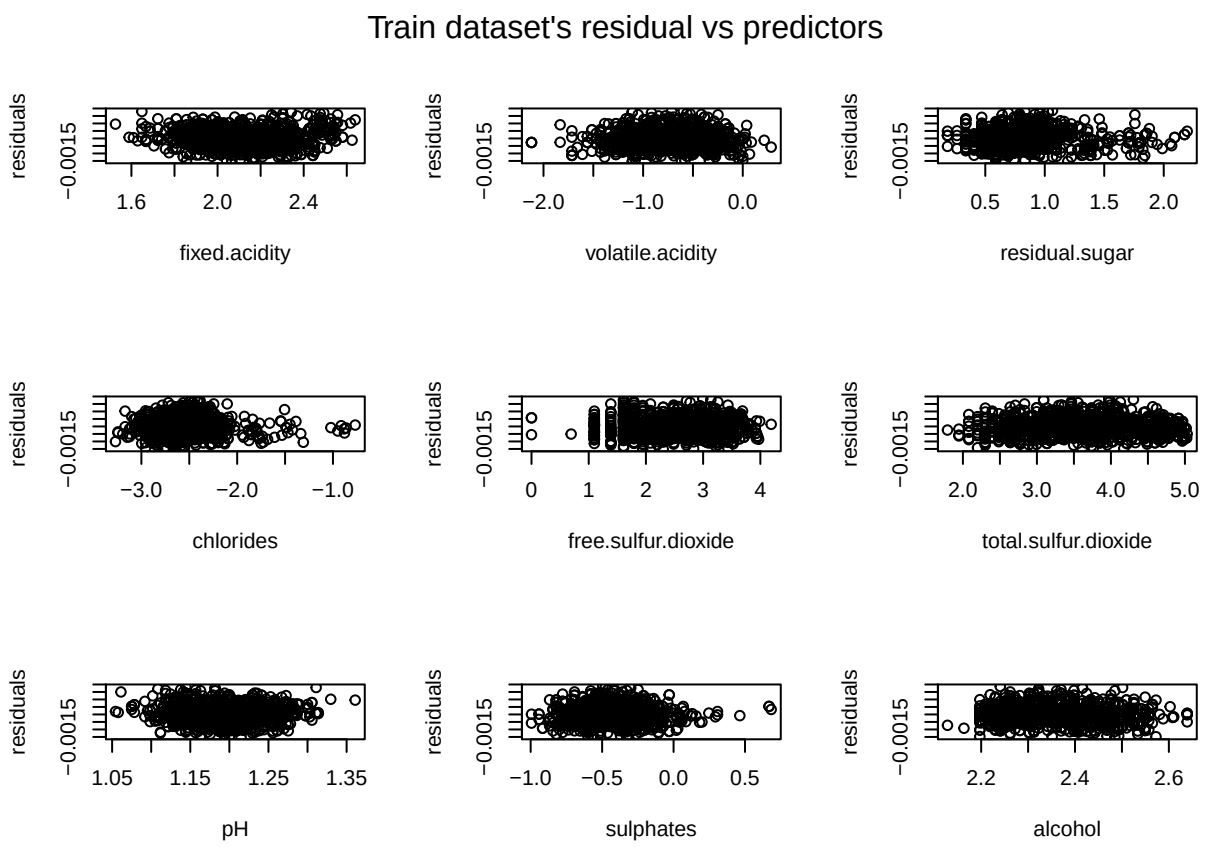
```
## [1] 0.8822059
```

```
summary(test_validate)$r.squared
```

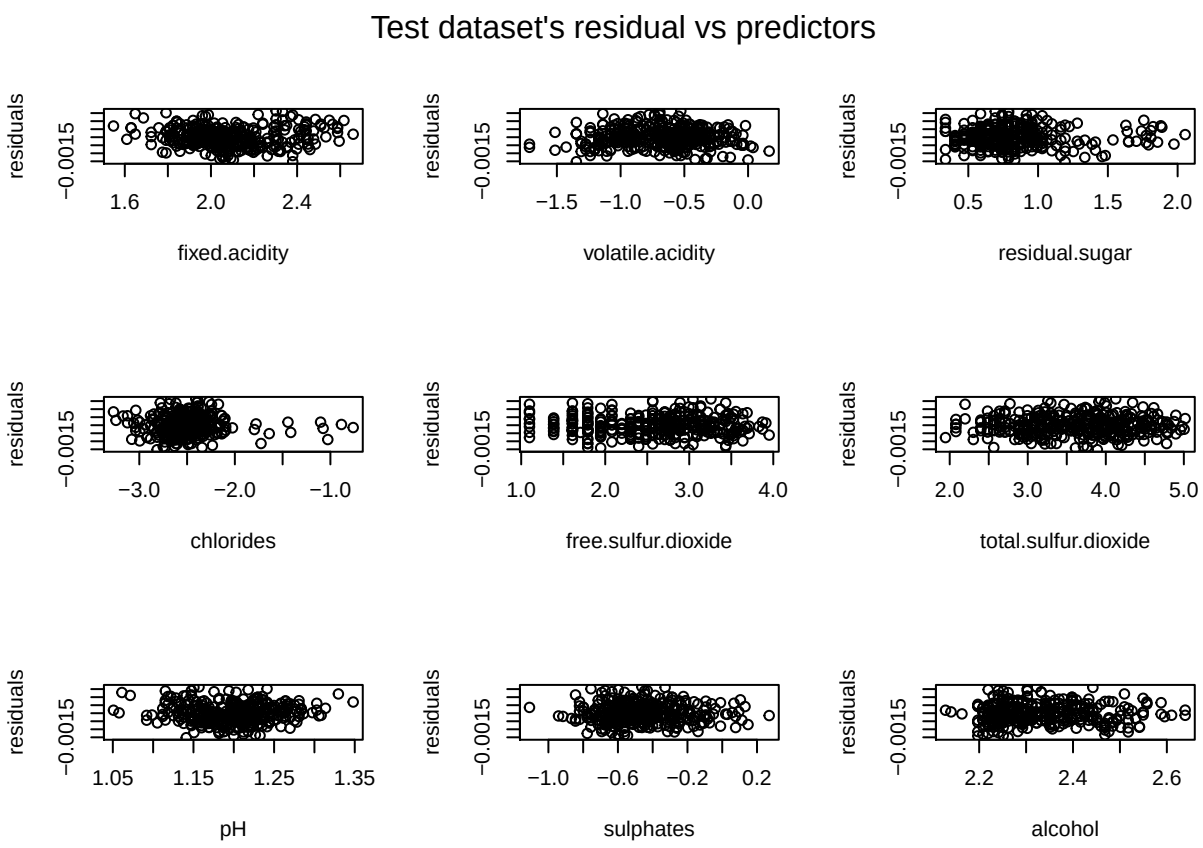
```
## [1] 0.8901155
```

#Comparing two data sets diagnosis test

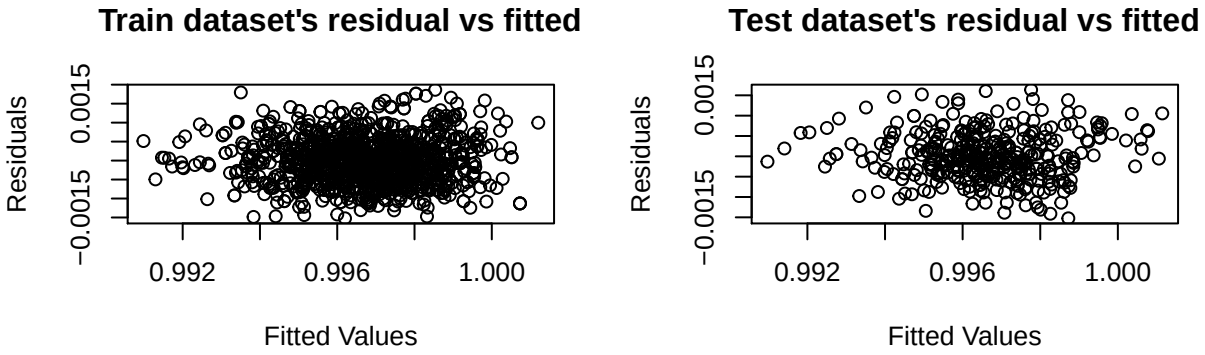
```
# comparing plots of residual vs predictors
par(mfrow=c(3,3))
plot(train_validate$residuals ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sul
mtext("Train dataset's residual vs predictors", side = 3, line = -1.5, outer = TRUE)
```

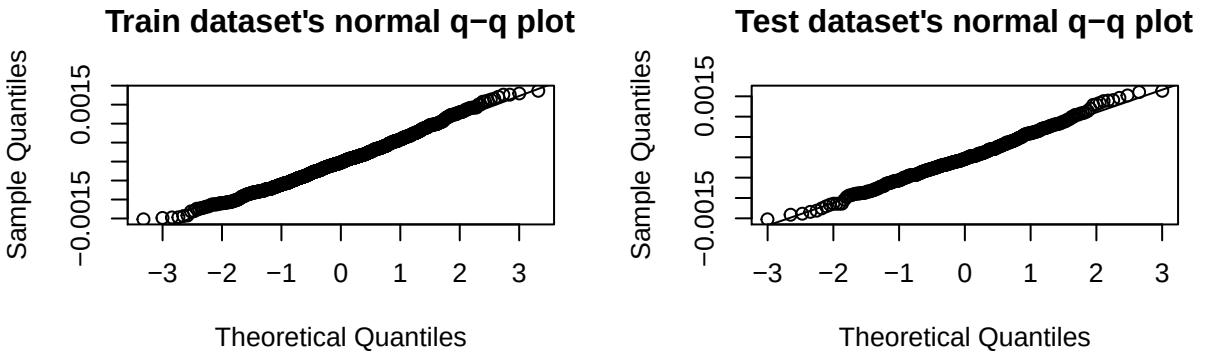



```
plot(test_validate$residuals ~ fixed.acidity + volatile.acidity + residual.sugar + chlorides + free.sul
mtext("Test dataset's residual vs predictors", side = 3, line = -1.5, outer = TRUE)
```



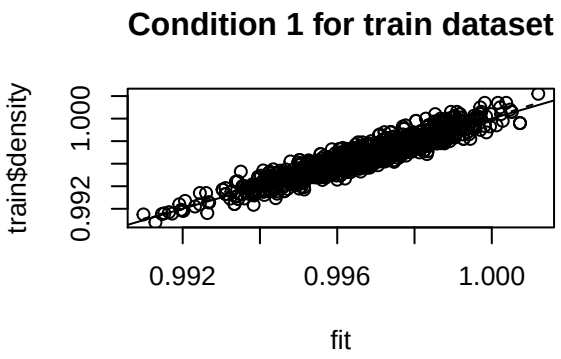

```
# comparing assumption checks
par(mfrow=c(2,2))
# residual vs fitted values
plot(train_validate$residuals ~ train_validate$fitted.values, main="Train dataset's residual vs fitted"
plot(test_validate$residuals ~ test_validate$fitted.values, main="Test dataset's residual vs fitted", x

# qq plot
qqnorm(train_validate$residuals, main = "Train dataset's normal q-q plot")
qqline(train_validate$residuals)

qqnorm(test_validate$residuals, main = "Test dataset's normal q-q plot")
qqline(test_validate$residuals)
```

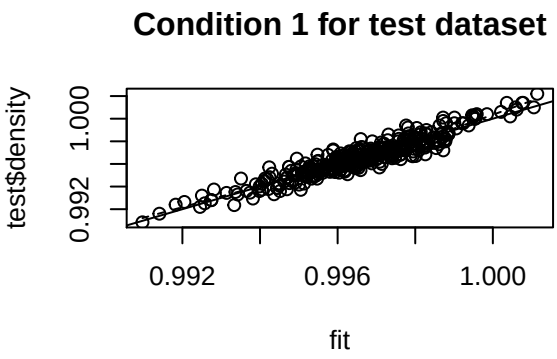




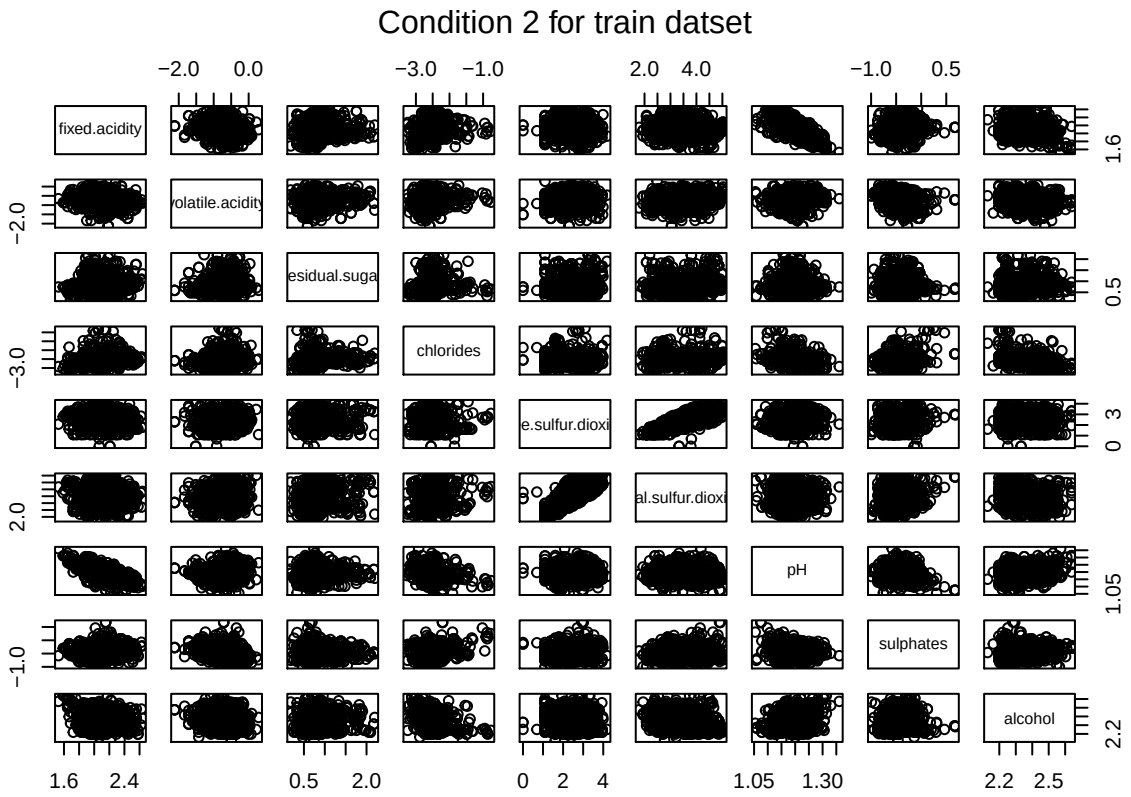
#Comparing conditions

```
# check condition 1
par(mfrow=c(2,2))
fit <- train_validate$fitted.values
plot(train$density ~ fit, main = "Condition 1 for train dataset")
abline(a = 0, b = 1)
lines(lowess(train$density ~ fit), lty=2)

fit <- test_validate$fitted.values
plot(test$density ~ fit, main = "Condition 1 for test dataset")
abline(a = 0, b = 1)
lines(lowess(test$density ~ fit), lty=2)

# check condition 2
pairs(train[,2:10])
```

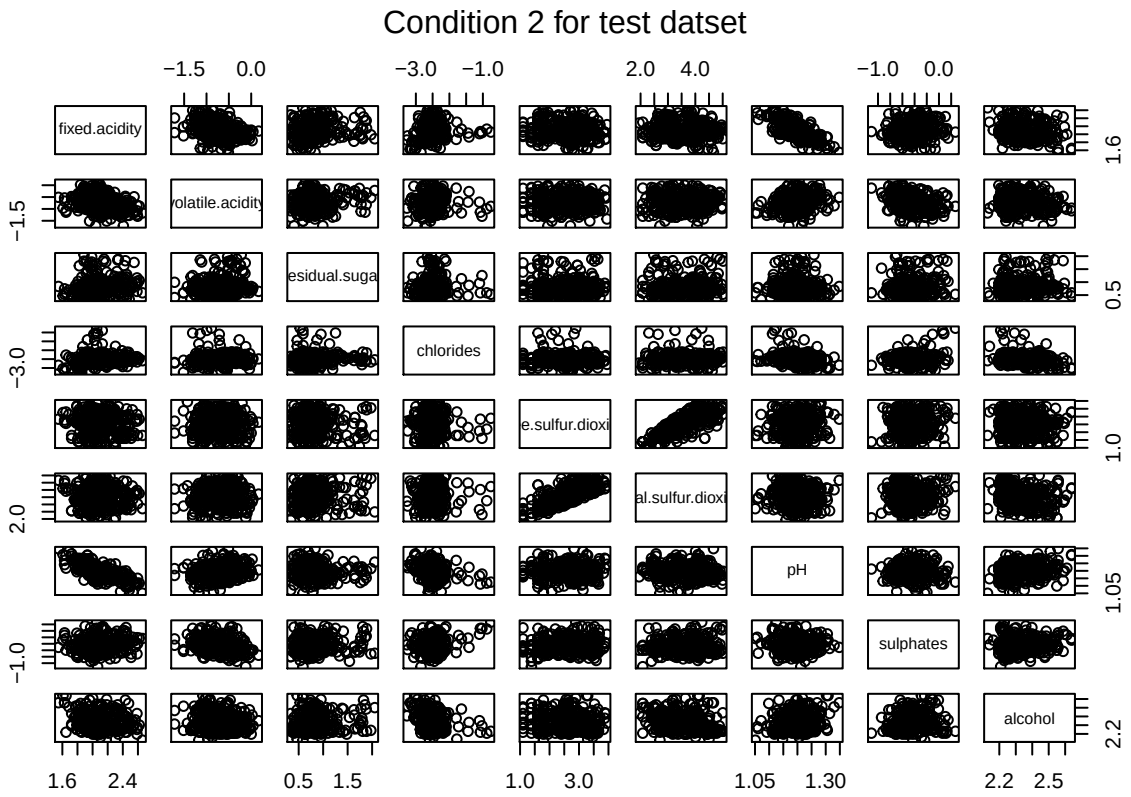





```
mtext("Condition 2 for train datset", side = 3, line = -1.5, outer = TRUE)
```

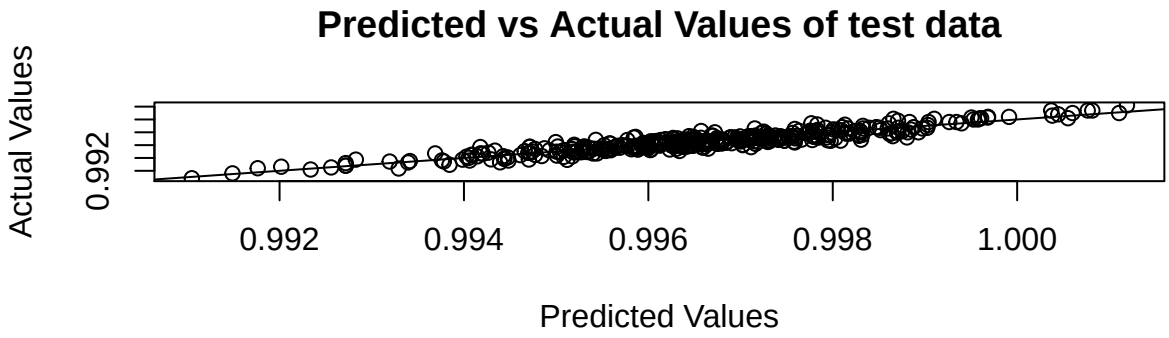


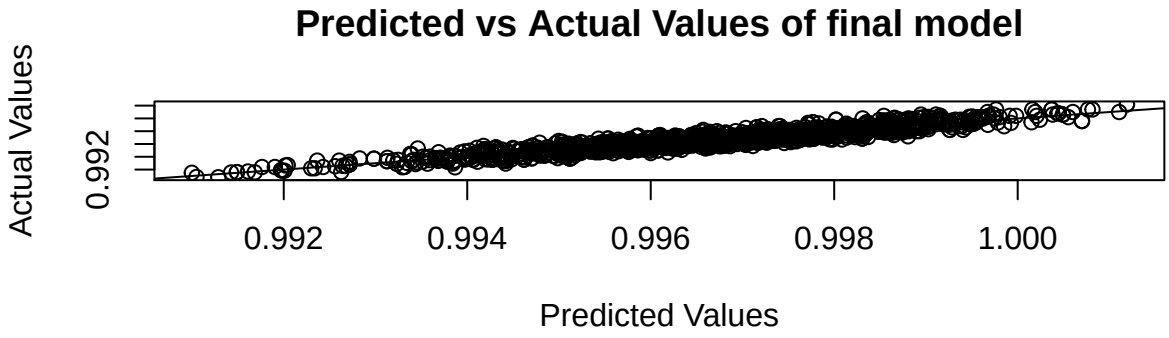

```
pairs(test[,2:10])
mtext("Condition 2 for test datset", side = 3, line = -1.5, outer = TRUE)
```



Predictions

```
# use final model(s) with test data to predict, and the predictions should be equal or their prediction
# If you have several models, of similar results, you can test and decide as the best will have the low
par(mfrow=c(2,1))

plot(x=predict(final_model, newdata = test), y= test$density,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Predicted vs Actual Values of test data')
abline(a=0, b=1)

plot(x=predict(final_model), y= red_wine4$density,
     xlab='Predicted Values',
     ylab='Actual Values',
     main='Predicted vs Actual Values of final model')
abline(a=0, b=1)
```


Predicted vs Actual Values of test data
Actual Values
0.992
0.992
0.994
0.996
0.998
1.000
Predicted Values


Predicted vs Actual Values of final model
Actual Values
0.992
0.992
0.994
0.996
0.998
1.000
Predicted Values