Homework 5

Giang Le (gianghl2)

Part II: Homework Questions – to be submitted

If $n = p$ and the $\mathbf{X}$ matrix is invertible, show that the hat matrix $\mathbf{H}$ is given by the $p \times p$ identity matrix. In this case, what are h_{ii} and $\hat{Y}_i$?

$$\mathbf{H} = \mathbf{X}(X^\intercal X)^{-1}X^\intercal = \mathbf{X}(X^{-1}(X^\intercal)^{-1})X^\intercal = (\mathbf{X}X^{-1})((X^\intercal)^{-1})X^\intercal) = (\mathbf{I})(\mathbf{I}) \tag{1}$$

as X is invertible.

The identity matrix I has dimension pxp or n x n because $\mathbf{X}, \mathbf{X^{-1}}, \mathbf{X}^\intercal$ and $\mathbf{X}^{\intercal -1}$ all have dimension p x p. $h_{ii} = 1$ because $\mathbf{H}$ is the identity matrix and so all diagonal values are 1.

$$\hat{Y} = \mathbf{HY} \rightarrow \hat{Y}_i = Y_i \tag{2}$$

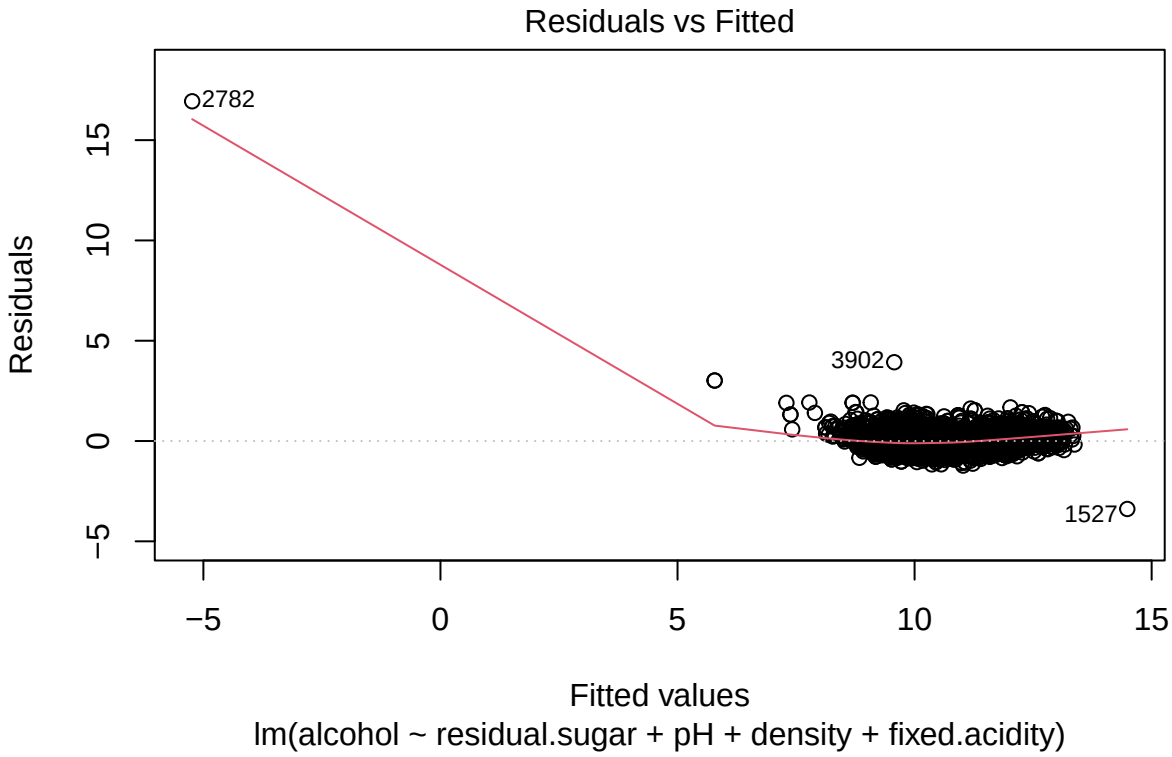
The `whitewines.csv` data set contains information related to white variants of the Portuguese "Vinho Verde" wine. Specifically, we have recorded the following information:

(a) fixed acidity, (b) volatile acidity, (c) citric acid , (d) residual sugar, (e) chlorides, (f) free sulfur dioxide, (g) total sulfur dioxide, (h) density, (i) pH, (j) sulphates, (k) alcohol, (l) quality (score between 0 and 10)

In this homework, our goal is to explain the relationship between alcohol level (dependent variable) and residual sugar, pH, density and fixed acidity.

Check the constant variance assumption.

```
# read in the data
wines = read.csv('whitewines.csv',header=TRUE,sep=";")

# Fit a model with the alcohol level as the dependent variable and residual sugar,
# pH, density, and fixed acidity as the predictors.

wines.lm = lm(alcohol ~ residual.sugar + pH + density + fixed.acidity, data=wines)


# Check the constant variance assumption by plotting the residuals against
# the fitted values
plot(wines.lm, which=1)
```



It looks like the variances are not constant, due to a outlier. We confirm by performing a Breush-Pagan test:

```
install.packages("lmtest",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpoOwTin/downloaded_packages
```

```
install.packages("MASS",repos = "http://cran.us.r-project.org")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9c/3_mgdyf12z7dvb8rt4d60nt80000gn/T//RtmpoOwTin/downloaded_packages
```

```
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(wines.lm)
```

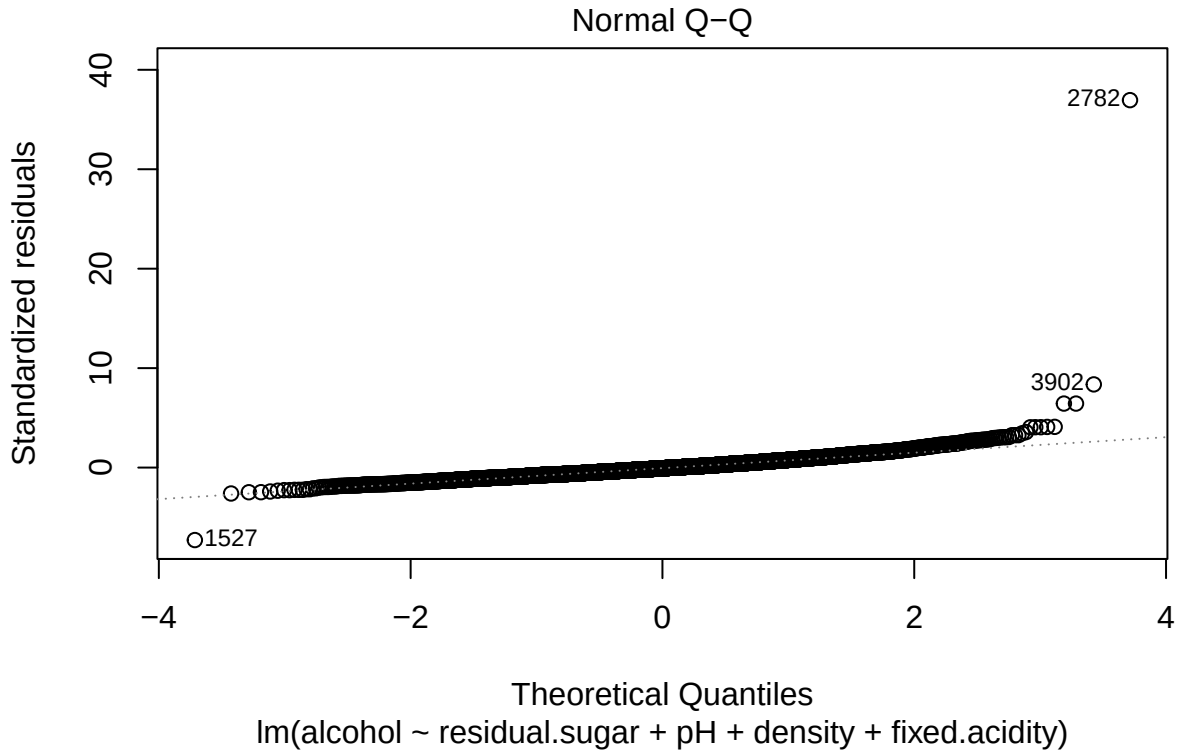
```
##
##  studentized Breusch-Pagan test
##
## data:  wines.lm
## BP = 250.65, df = 4, p-value < 2.2e-16
```

The p-value of the test is very low. Our decision is that we reject the null hypothesis of constant variance

and concludes that our model does not have homoscedasticity. It's possible that removing the outlier would improve the model and make it fit better with this assumption.

Check the normality assumption.

```
# Plotting the QQ plot
plot(wines.lm, which=2)
```




```
# Plotting the histogram of the residuals
hist(wines.lm$residuals)
```

The normality assumption looks satisfied according to the QQ plot. According to the histogram, the residuals distribution looks rightly skewed, probably due to outliers.

Check for the structure of the relationship between the predictors and the response.

I plot the Added Variable plots to check the model structure.

```
# residual sugar removed
r.ysugar = update(wines.lm, ~  pH + density + fixed.acidity)$res;
r.sugar = lm(residual.sugar ~  pH + density + fixed.acidity, data=wines)$res;
tmp1=lm(r.ysugar ~ r.sugar);
summary(tmp1)
```

```
##
## Call:
## lm(formula = r.ysugar ~ r.sugar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3867 -0.2735 -0.0334  0.2200 16.9366
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.129e-17  6.716e-03    0.00        1
## r.sugar     2.367e-01  2.702e-03   87.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.47 on 4896 degrees of freedom
## Multiple R-squared:  0.6105, Adjusted R-squared:  0.6105
## F-statistic:  7675 on 1 and 4896 DF,  p-value: < 2.2e-16
```

```r
# pH removed
r.ypH = update(wines.lm, ~ residual.sugar + density + fixed.acidity)$res;
r.pH = lm(pH ~ residual.sugar + density + fixed.acidity, data=wines)$res;
tmp2=lm(r.ypH ~ r.pH);
summary(tmp2)
```

```
##
## Call:
## lm(formula = r.ypH ~ r.pH)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3867 -0.2735 -0.0334  0.2200 16.9366
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.250e-16  6.716e-03    0.00        1
## r.pH        2.535e+00  5.279e-02   48.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.47 on 4896 degrees of freedom
## Multiple R-squared:  0.3202, Adjusted R-squared:  0.3201
## F-statistic:  2307 on 1 and 4896 DF,  p-value: < 2.2e-16
```

```r
# density removed
r.ydensity = update(wines.lm, ~ residual.sugar + pH +  fixed.acidity)$res;
r.density = lm(pH ~ residual.sugar + pH +  fixed.acidity, data=wines)$res;
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 2 in
## model.matrix: no columns are assigned
```

```r
tmp3=lm(r.ydensity ~ r.density);
summary(tmp3)
```

```
##
## Call:
## lm(formula = r.ydensity ~ r.density)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3789 -0.8054 -0.1790  0.7099  7.6851
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.078e-16  1.563e-02       0        1
## r.density   2.358e-16  1.162e-01       0        1
##
## Residual standard error: 1.094 on 4896 degrees of freedom
## Multiple R-squared:  2.098e-31,  Adjusted R-squared:  -0.0002042
## F-statistic: 1.027e-27 on 1 and 4896 DF,  p-value: 1
```

```
# fixed acidity removed
r.yacidity = update(wines.lm, ~ residual.sugar + pH +  density)$res;
r.acidity = lm(pH ~ residual.sugar + pH +  density, data=wines)$res;
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

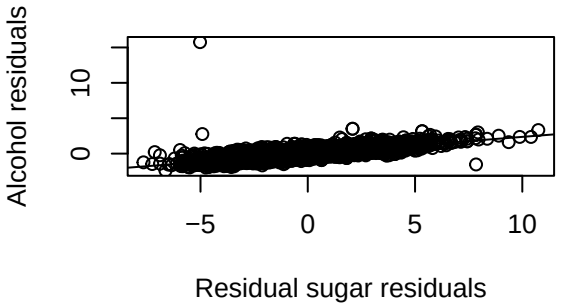
```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 2 in
## model.matrix: no columns are assigned
```

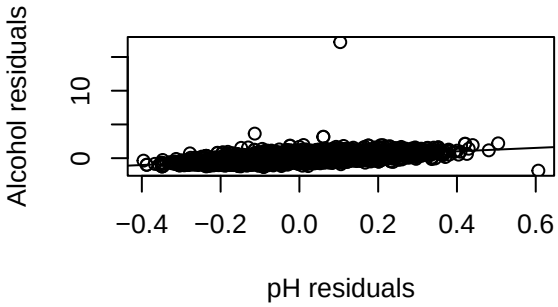
```
tmp4=lm(r.yacidity ~ r.acidity);
summary(tmp4)
```

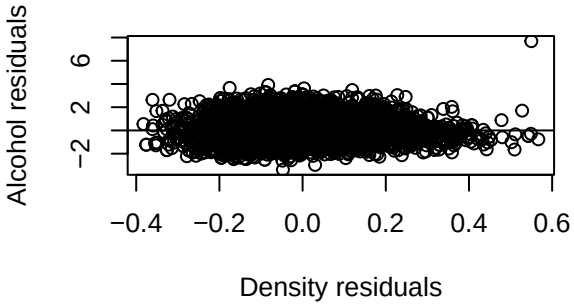
```
##
## Call:
## lm(formula = r.yacidity ~ r.acidity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.1073 -0.4064 -0.0506  0.3372 15.9763
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6.940e-17  8.501e-03       0        1
## r.acidity   -2.322e-16  5.789e-02       0        1
##
## Residual standard error: 0.595 on 4896 degrees of freedom
## Multiple R-squared:  1.618e-32,  Adjusted R-squared:  -0.0002042
## F-statistic: 7.919e-29 on 1 and 4896 DF,  p-value: 1
```

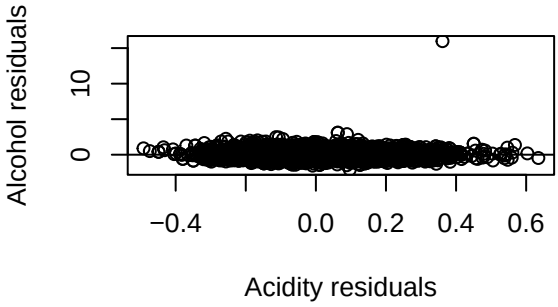
Now I add the variables plot

```
# Added variables plots
par(mfrow=c(2,2))
plot(r.sugar, r.ysugar, xlab="Residual sugar residuals", ylab="Alcohol residuals"); abline(tmp1)
plot(r.pH, r.ypH, xlab="pH residuals", ylab="Alcohol residuals"); abline(tmp2)
plot(r.density, r.ydensity, xlab="Density residuals", ylab="Alcohol residuals"); abline(tmp3)
plot(r.acidity, r.yacidity, xlab="Acidity residuals", ylab="Alcohol residuals"); abline(tmp4)
```









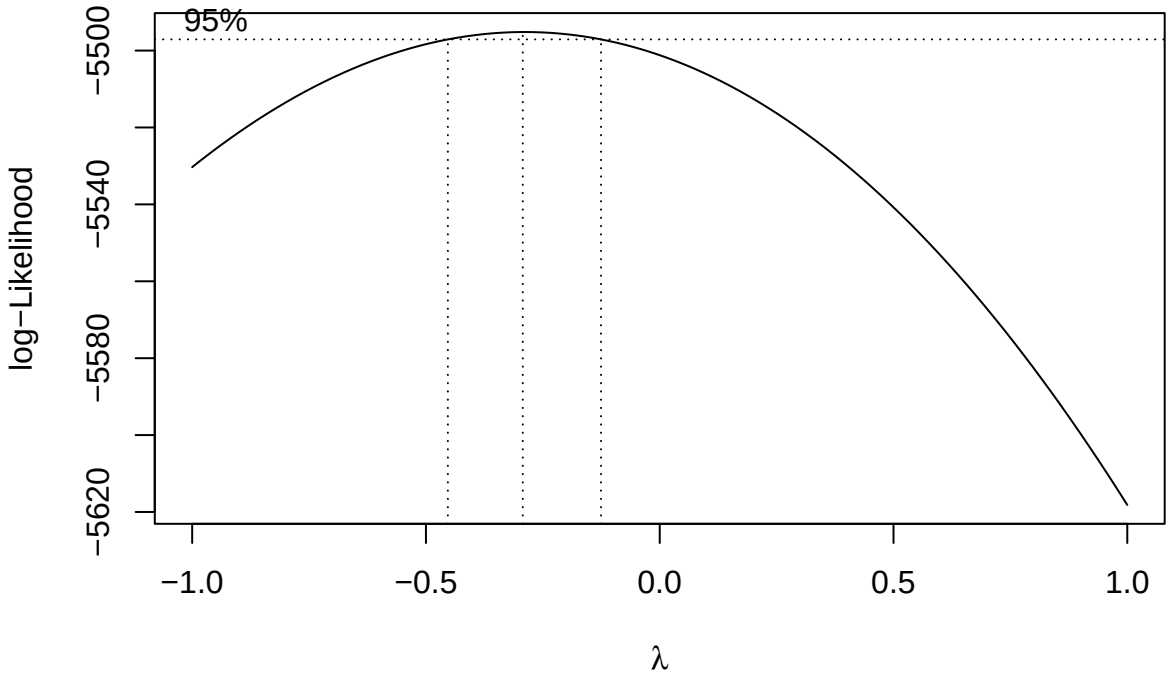
According to these plots, it seems linearity is not a problem for this set of data because all the points are scattered evenly without no clear trend.

Is any transformation of the predictors suggested?

Residual sugar and pH variables might need to be transformed, according to what the plots above suggest.

Use the Box-Cox method to determine an optimal transformation of the response. Would it be reasonable to leave the response untransformed?

```
library("MASS")

wines.transformation = boxcox(wines.lm, lambda=seq(-1, 1, by=0.05))
```



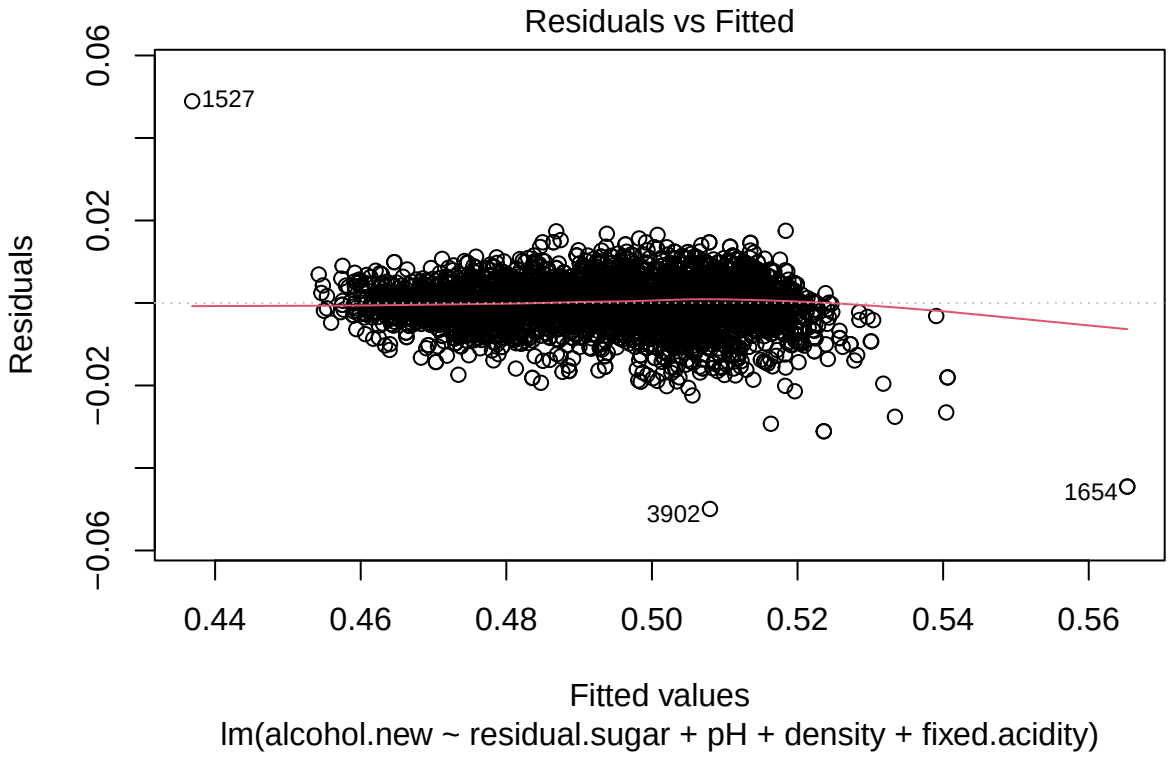
It looks like the optimal value for the response is about -0.3. I will transform the response by raising it to -0.3 power.

Use the optimal transformation of the response and refit the additive model. Does this make any difference to the transformations suggested for the predictors?

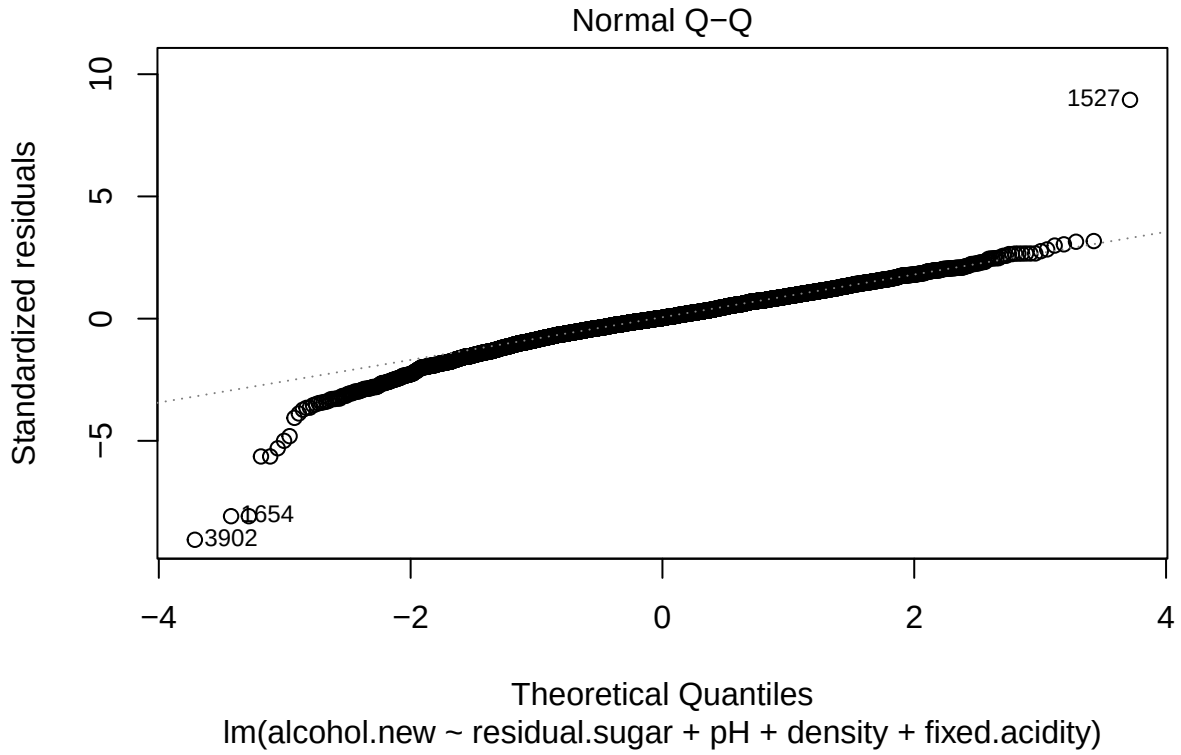
```
# Remove the outlier data point identified in the previous homework
wines <- wines[-c(2782),]

# Transform Y
wines$alcohol.new = 1/(wines$alcohol^0.3)
wines.lm.new = lm(alcohol.new ~ residual.sugar + pH + density + fixed.acidity, data=wines)
```

According to the plots, it looks like the constant variance assumption is met. The QQ plot and normality checks are also better. However, it is unclear if this is due to removing the outlier datapoint or the transformation of the response. I didn't apply any transformation of the predictors, because I don't think linearity is an issue here.

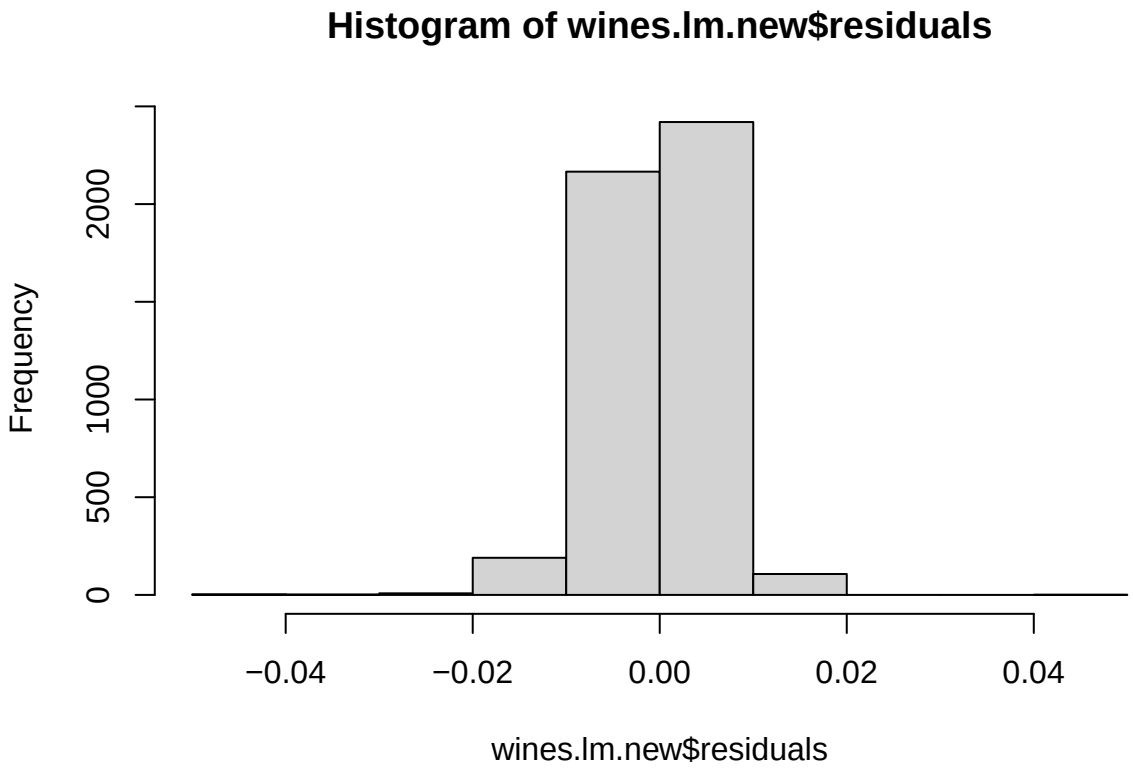
```
plot(wines.lm.new, which=1)
```




```
plot(wines.lm.new, which=2)
```




```
hist(wines.lm.new$residuals)
```


Histogram of wines.lm.new$residuals
Frequency
0
500
1000
2000
−0.04
−0.02
0.00
0.02
0.04
wines.lm.new$residuals