Risidual Analysis

Five important assumptions need to hold so that the regression model can be useful hypothesis testing and predication. These are:

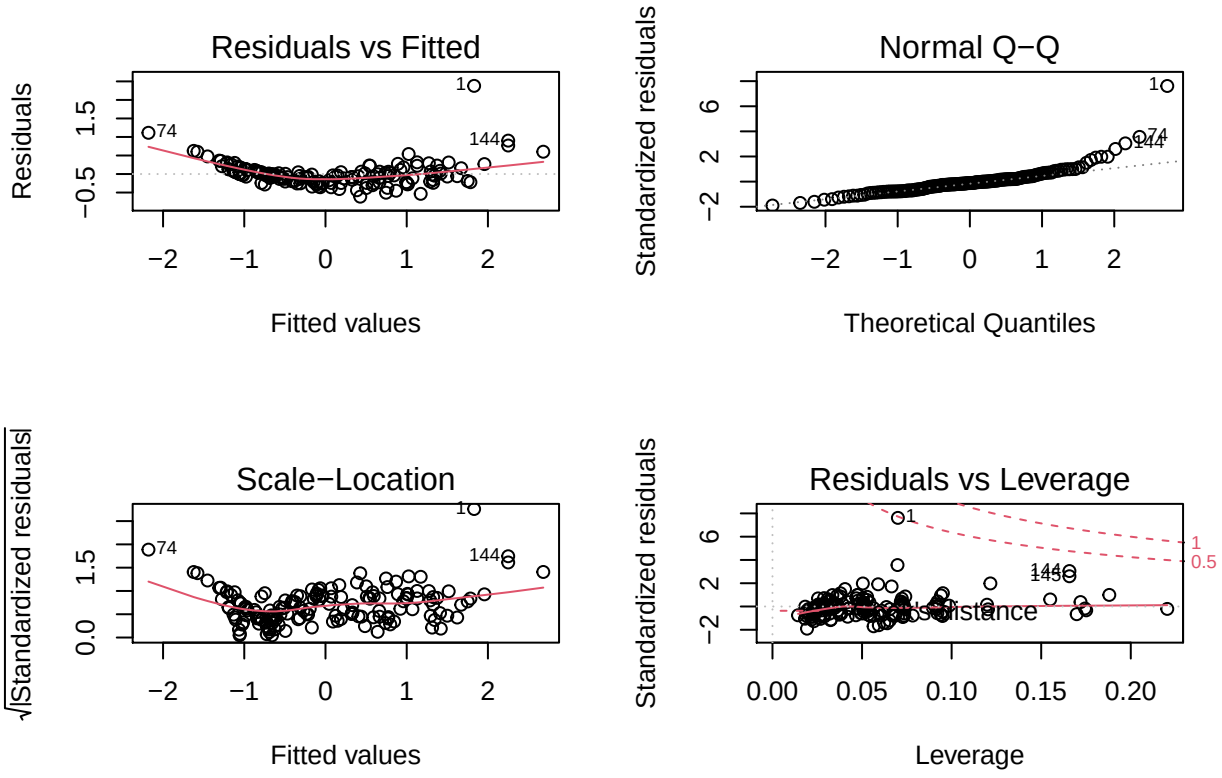
1. The relationship between the response y and the regression is linear (at least approximately).
2. The error term ϵ has zero mean.
3. The error term ϵ has constant variance σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed.

```
data <- read.csv("cleaned_data_3.csv", fileEncoding="UTF-8-BOM")
```

Lets use the risidual plots using standardized residuals so that we can compare the current state of the model with these assumptions.

```
model <- lm(Weight ~ . - 1, data = data)

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



Risiduals vs Fitted

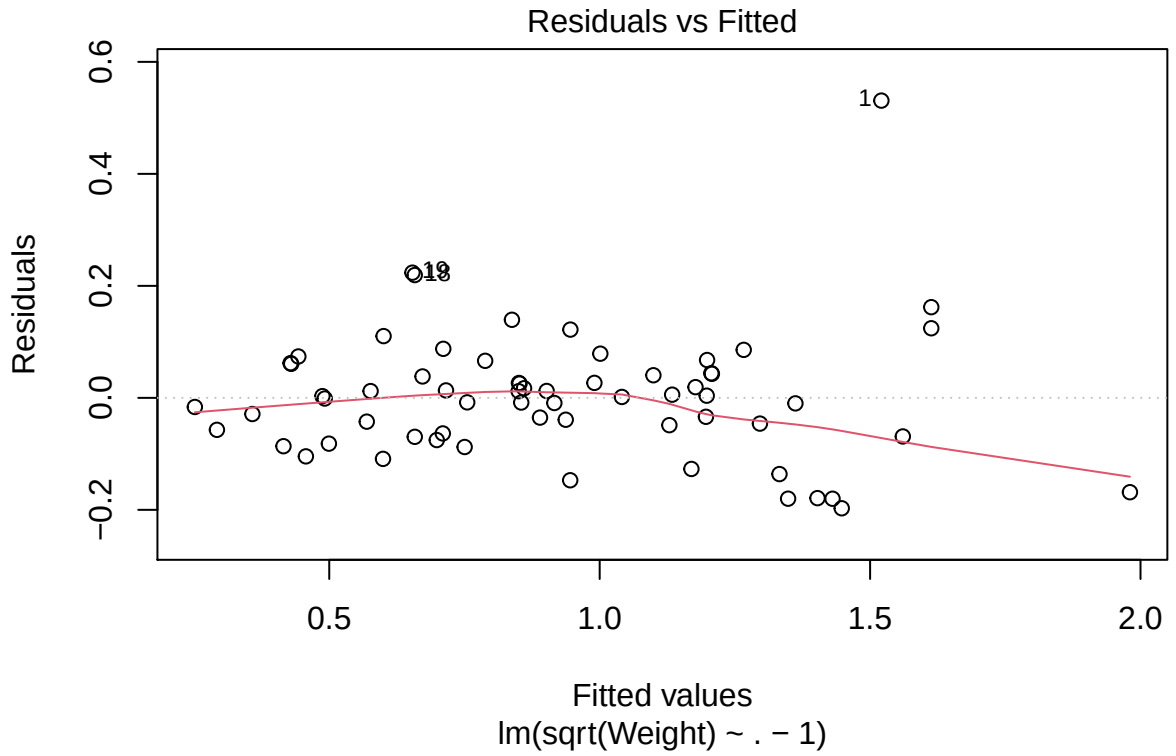
Here, assumption **1. The relationship between the response y and the regression is linear** is violated (we want to see a linear pattern between the Risiduals and Fitted values). This is not surprising

since we observed a non linear pattern between the predictors and response in the pairs plot.

```
model_transformed <- lm(sqrt(Weight) ~ . - 1, data = data)
```

```
## Warning in sqrt(Weight): NaNs produced
```

```
plot(model_transformed, which = 1)
```



Taking the square root of the response seems to produce the best result compared to other transformations of the response like the natural logarithm. Drawing a horizontal line at 0 seems reasonable. This satisfies 1. The relationship between the response y and the regression is linear. To confirm that we now have an improved linear relationship, we can compare the R^2 of the model.

```
model <- lm(Weight ~ . - 1, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ . - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61102 -0.18494 -0.03696  0.07822  2.37901
##
## Coefficients:
```

```
##                   Estimate Std. Error t value Pr(>|t|)
## Species_Parkki     0.17835    0.10755   1.658   0.0993 .
## Species_Perch      0.01546    0.05602   0.276   0.7830
## Species_Pike      -0.78434    0.18033  -4.349 2.50e-05 ***
## Species_Roach     -0.01042    0.07825  -0.133   0.8942
## Species_Smelt      0.78443    0.10548   7.437 7.21e-12 ***
## Species_Whitefish -0.02601    0.13825  -0.188   0.8510
## Length1            0.96916    0.12333   7.858 6.79e-13 ***
## Height             0.05500    0.05263   1.045   0.2976
## Width              0.16137    0.11791   1.369   0.1732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.324 on 151 degrees of freedom
## Multiple R-squared:  0.9003, Adjusted R-squared:  0.8944
## F-statistic: 151.6 on 9 and 151 DF,  p-value: < 2.2e-16
```

```
summary(model_transformed)
```

```
##
## Call:
## lm(formula = sqrt(Weight) ~ . - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.19720 -0.06489  0.00235  0.04813  0.53079
##
## Coefficients: (3 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## Species_Parkki          NA         NA      NA       NA
## Species_Perch      0.12828    0.06873   1.866   0.0671 .
## Species_Pike      -0.12988    0.10141  -1.281   0.2054
## Species_Roach           NA         NA      NA       NA
## Species_Smelt           NA         NA      NA       NA
## Species_Whitefish  0.22008    0.08497   2.590   0.0121 *
## Length1            0.51752    0.05799   8.924 1.77e-12 ***
## Height             0.26415    0.02821   9.364 3.35e-13 ***
## Width              0.17614    0.06621   2.660   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.119 on 58 degrees of freedom
##   (96 observations deleted due to missingness)
## Multiple R-squared:  0.9873, Adjusted R-squared:  0.986
## F-statistic: 751.7 on 6 and 58 DF,  p-value: < 2.2e-16
```

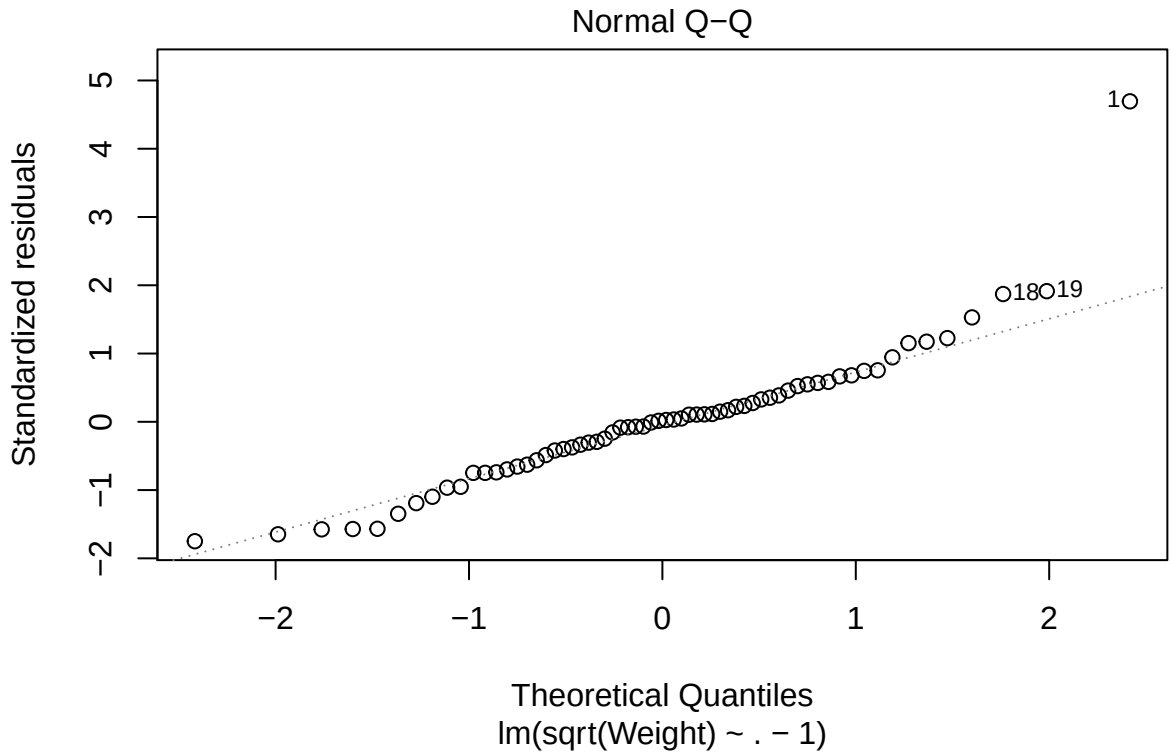
Performing this transformation has significantly improved the R^2 of the model from 0.900334 to 0.987303. Note that after the transformation we get NA values for Species_Parkki, Species_Roach, and Species_Smelt. Given the R^2 value after the transformation, it is the case that these variables are predicted perfectly by another (co-linear), and as a result add no value to the model. At this stage these predictors can be removed.

```
data$Species_Parkki <- NULL
data$Species_Roach <- NULL
data$Species_Smelt <- NULL
```

Overall, we can see that **1. The relationship between the response y and the regression is linear** and **2. The error term ϵ has zero mean.** have been satisfied.

Normal Q-Q

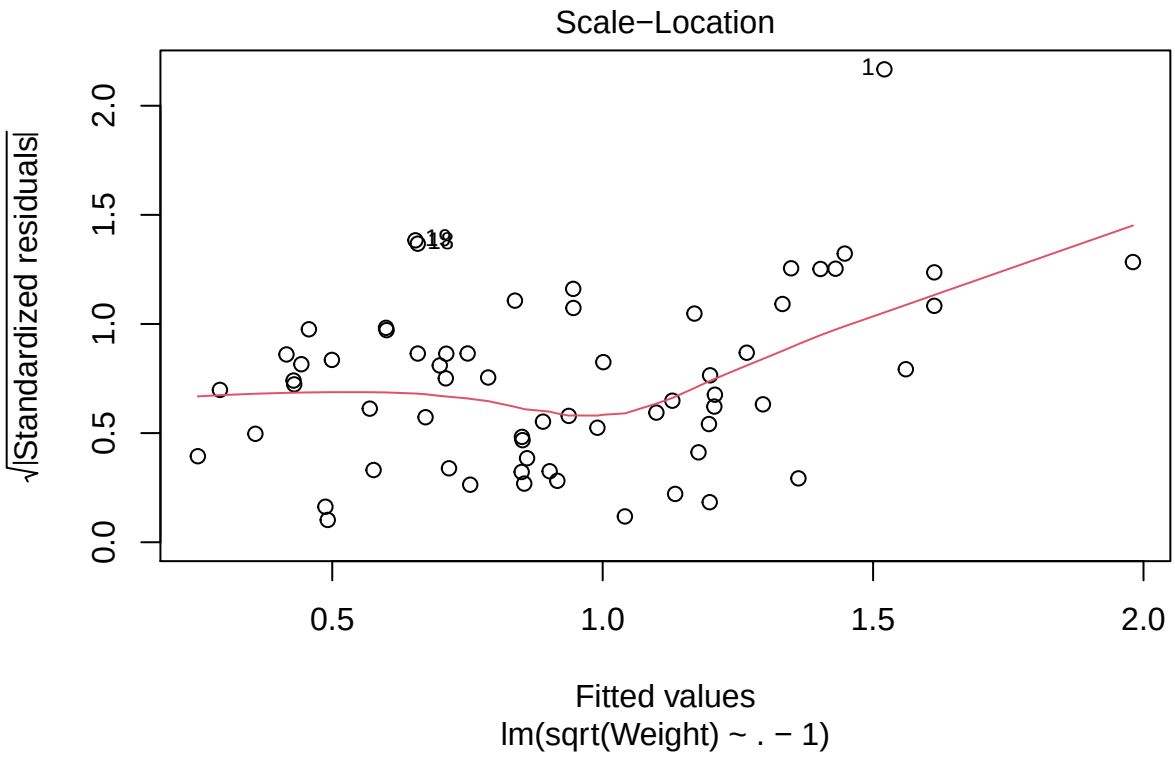
```
plot(model_transformed, which = 2)
```



The normality assumption is is adequately met after the transformation. That is **5. The errors are normally distributed**. Previously, you can see that the data was not normal and havily skewed to the right.

Scale-Location

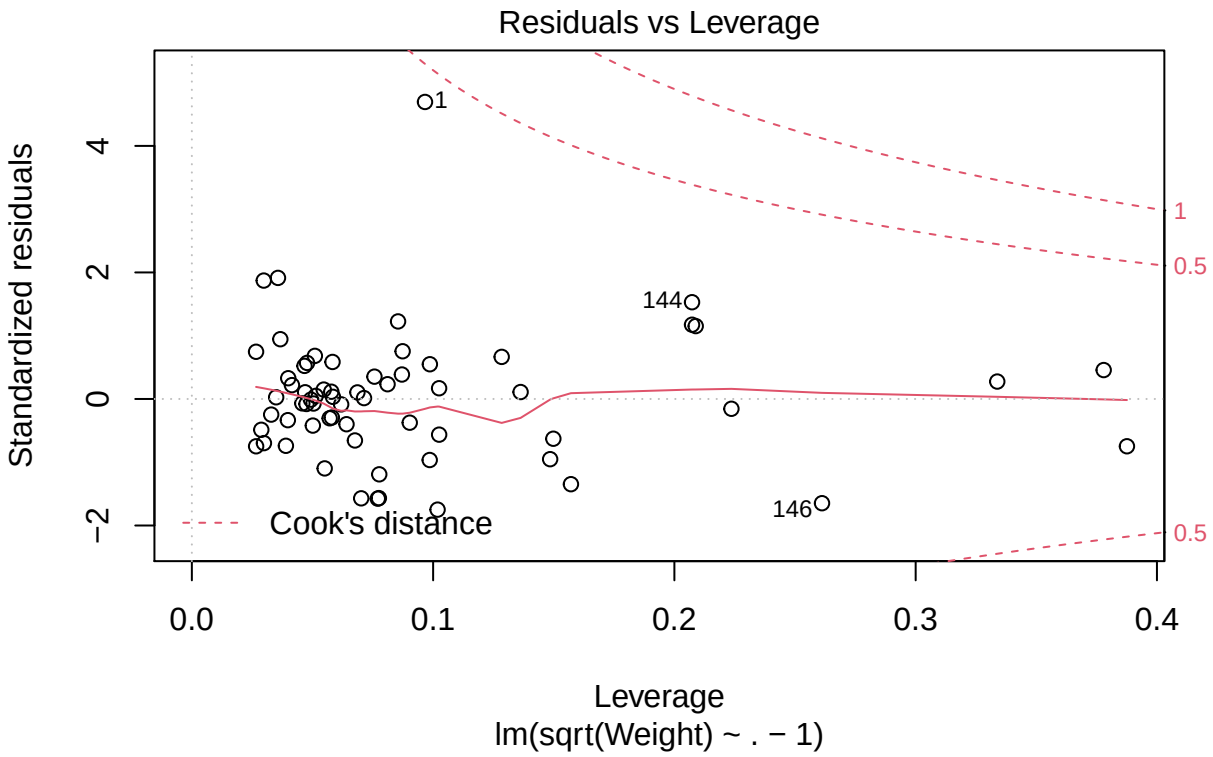
```
plot(model_transformed, which = 3)
```



The variance is fairly constant. Before the transormation, there was a defnitie pattern that would lead us to conclude non-constant variance. **3. The error term ϵ has constant variance σ^2.** is satisfied.

Residuals vs. Leverage

```
plot(model_transformed, which = 5)
```


Residuals vs Leverage
Standardized residuals
Leverage
lm(sqrt(Weight) ~ . − 1)
Cook's distance
1
144
146
0.5
1
0.5
−2
0
2
4
0.0
0.1
0.2
0.3
0.4