# Risidual Analysis

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

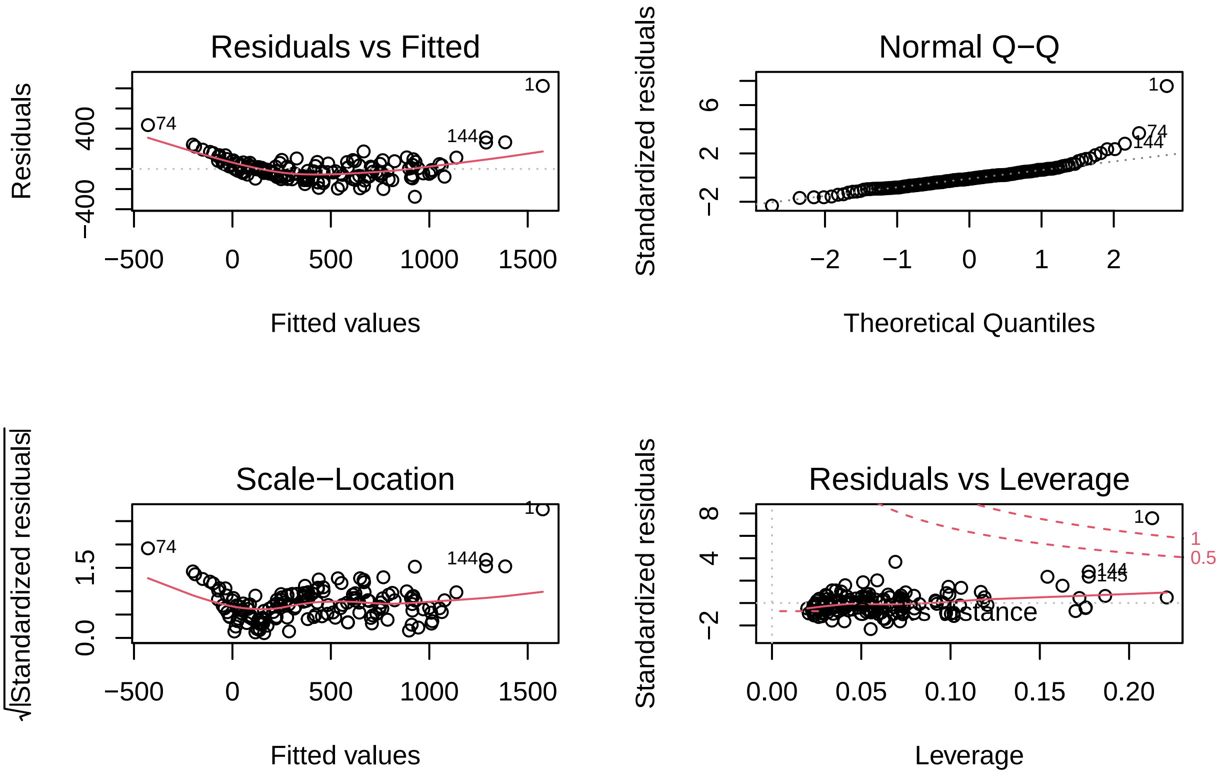
1. The relationship between the response y and the regression is linear (at least approximately).
2. The error term $\epsilon$ has zero mean.
3. The error term $\epsilon$ has constant variance $\sigma^2$.
4. The errors are uncorrelated.
5. The errors are normally distributed.

```
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 ~ ., data = data)

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



## Addressing Outliers

Observation 1 (the unique data point we added) immediately strikes us as having a huge risidual that makes it an outliar from all other data points. It also has significant leverage, pulling the linear regression line away from the best fit for the other data points. In this case, the fish Weight is disproportionately large from its other attributes like Height and Length. There are many possible causes for this (a faulty measurement scale, a fisherman disproportionately inflating the weight to get a better price, simple data entry error, or even the fish swallowing some object). However, since we have seen how consistent the proportionality of fish dimensions is to weight, it is almost impossible that this could be a legitimate observation. For this reason, removing the observation is justified.

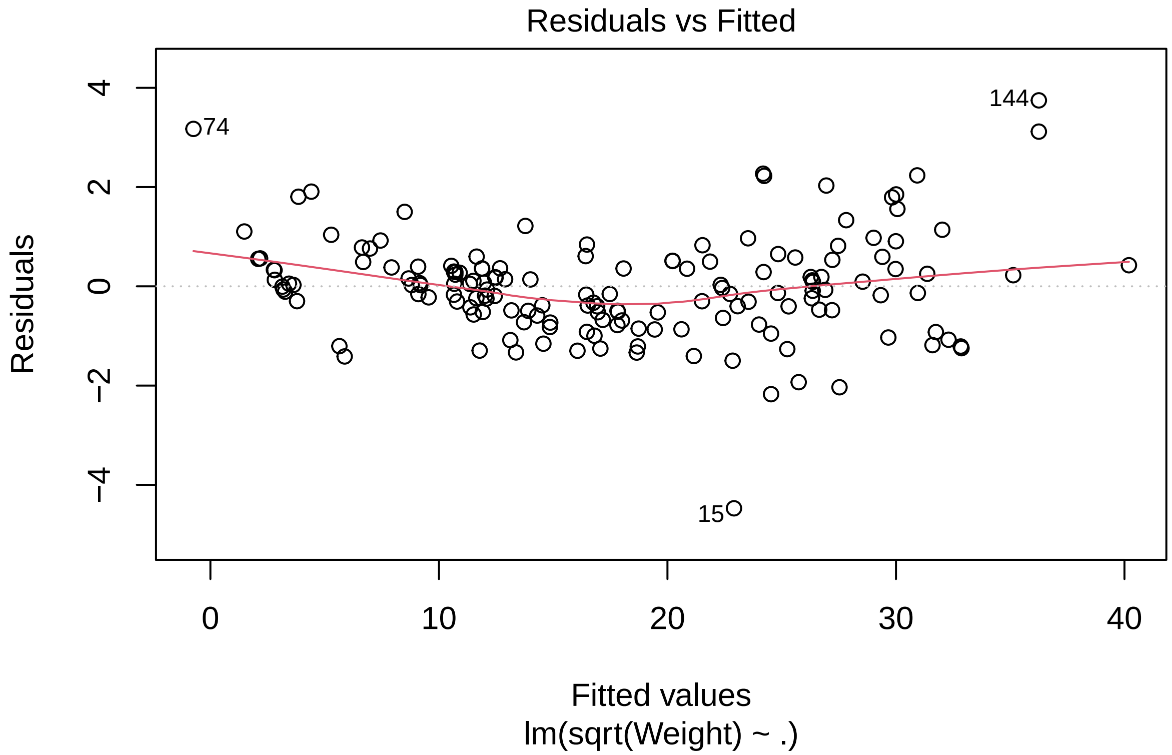
```r
data <- data[-c(1),]
model <- lm(Weight ~ ., data = data)
```

## 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.

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

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.

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

```
## 
## Call:
## lm(formula = Weight ~ ., data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -229.21  -55.05   -7.04   33.25  397.14 
## 
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -746.4762    86.2013  -8.660 7.19e-15 ***
## Species_Parkki      64.2364    48.8466   1.315  0.19051    
## Species_Perch       40.6409    82.9492   0.490  0.62489    
## Species_Pike      -254.4729   126.0128  -2.019  0.04524 *  
## Species_Roach       26.4221    77.7535   0.340  0.73447    
## Species_Smelt      299.6009    92.3743   3.243  0.00146 ** 
## Species_Whitefish   42.4582    79.4437   0.534  0.59383    
## Length1             37.8338     4.0004   9.457  < 2e-16 ***
## Height              14.0338    13.2228   1.061  0.29026    
## Width                0.8559    24.4172   0.035  0.97208    
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 96.67 on 149 degrees of freedom
## Multiple R-squared:  0.931,  Adjusted R-squared:  0.9268 
## F-statistic: 223.2 on 9 and 149 DF,  p-value: < 2.2e-16
```

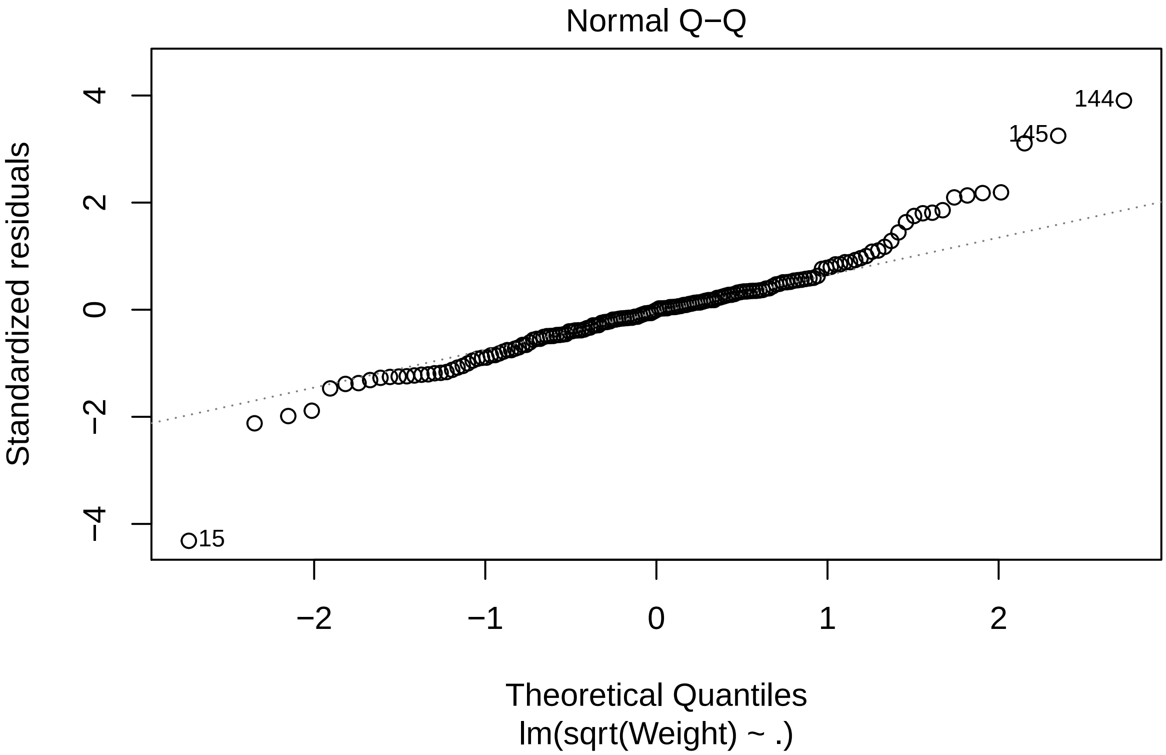
```r
summary(model_transformed)
```

```
## 
## Call:
## lm(formula = sqrt(Weight) ~ ., data = data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.4731 -0.5234 -0.0057  0.4190  3.7476 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)       -8.88830    0.94432  -9.412  < 2e-16 ***
## Species_Parkki     0.69560    0.53511   1.300 0.195635    
## Species_Perch      0.73702    0.90870   0.811 0.418618    
## Species_Pike      -2.35110    1.38045  -1.703 0.090628 .  
## Species_Roach      0.42972    0.85178   0.504 0.614656    
## Species_Smelt      2.37680    1.01195   2.349 0.020151 *  
## Species_Whitefish  1.55532    0.87029   1.787 0.075951 .  
## Length1            0.64245    0.04382  14.660  < 2e-16 ***
## Height             0.55811    0.14485   3.853 0.000173 ***
## Width              1.00205    0.26749   3.746 0.000256 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.059 on 149 degrees of freedom
## Multiple R-squared:  0.9873, Adjusted R-squared:  0.9865
## F-statistic:  1287 on 9 and 149 DF,  p-value: < 2.2e-16
```

Performing this transformation has significantly improved the $R^2$ of the model from 0.9309509 to 0.9872994. Overall, we can see that **1. The relationship between the response y and the regression is linear** and **2. The error term $\epsilon$ has zero mean.** have been satisfied.

## Normal Q-Q

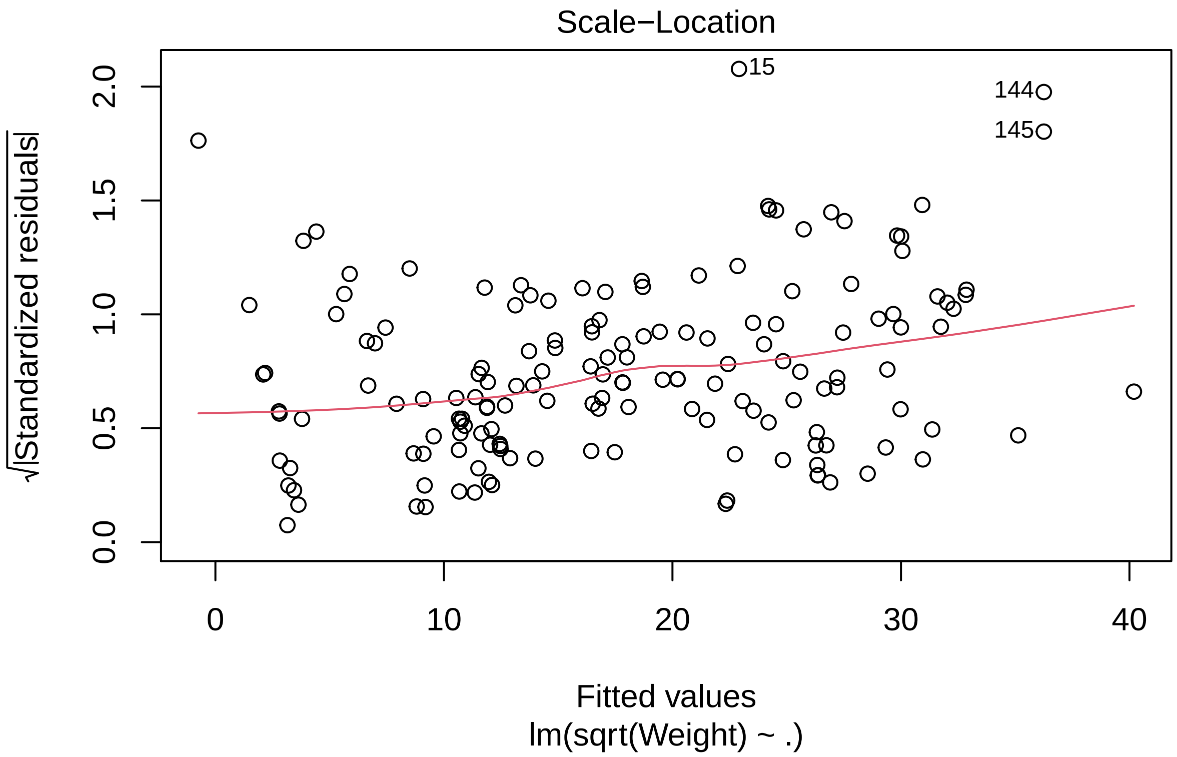
```
plot(model_transformed, which = 2)
```



The normality assumption is is adequately met after the transformation. The majority of the data follows an x=y diagonal line from +/-2 SD. The one negative result of the square root transformation is that the risidual of observation 15 was magnified. We can see later whether this is a cause for concern based on its Cook's D value. Overall, **5. The errors are normally distributed** is satisfied.

## Scale-Location

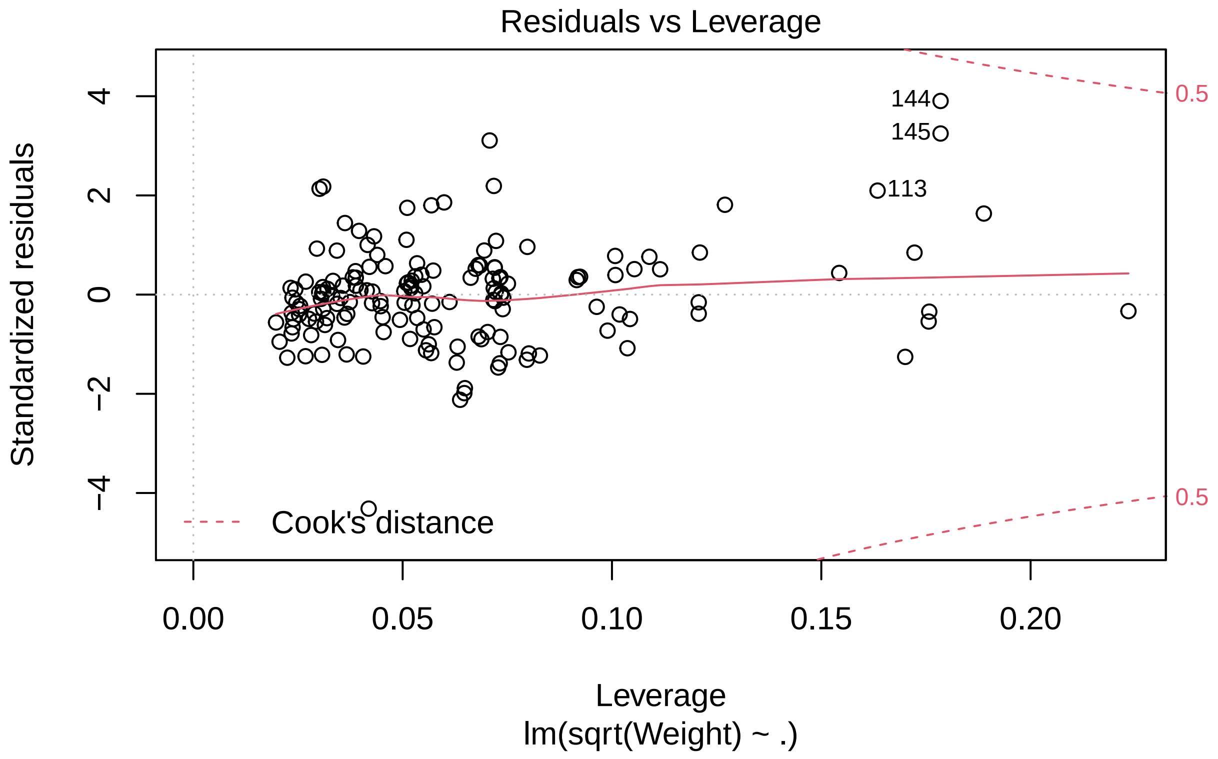
```r
plot(model_transformed, which = 3)
```



The variance is fairly constant. Before the transormation, there was a definite pattern that would lead us to conclude non-constant variance. **3. The error term $\epsilon$ has constant variance $\sigma^2$.** is satisfied.

## Residuals vs. Leverage

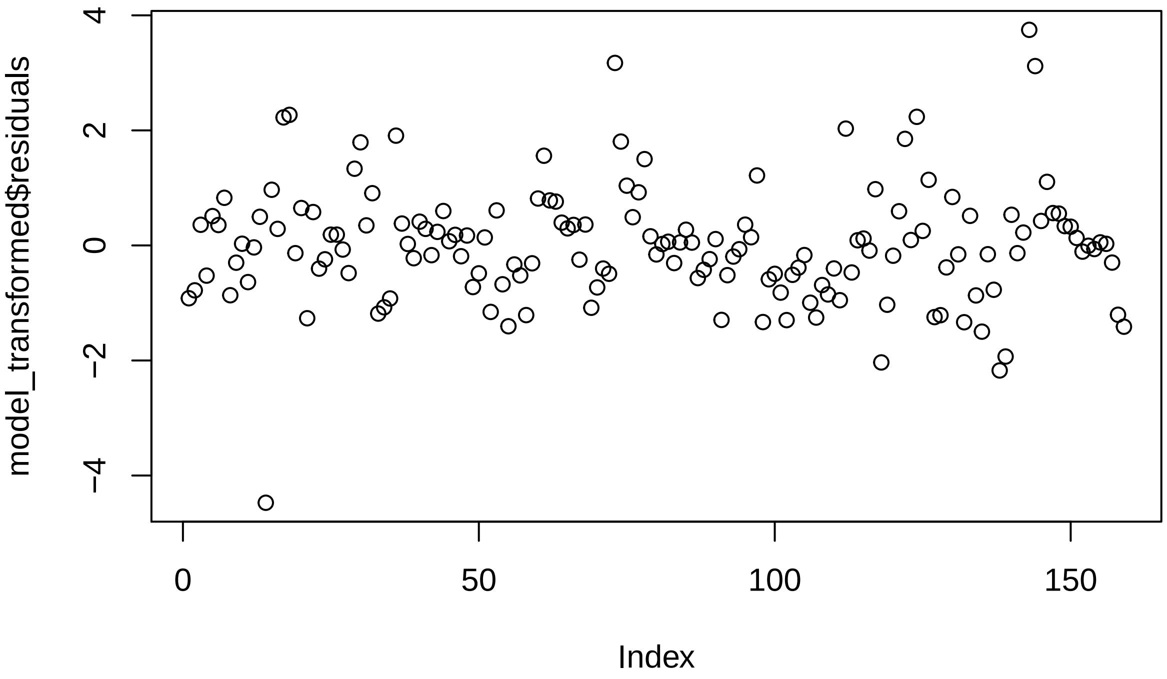
```r
plot(model_transformed, which = 5)
```



Observation 1 was an influential observation negatively impacting the explanatory power of the model. Since it was deemed to be invalid, it was removed. While observation 15 is an outlier, it does not carry much leverage over the regression and should be fine to leave in. It is notable that there are a few observations which exhibit significantly more leverage. However, their Cook's distance still suggest that they are not influential.

## Residual vs. Index

```
plot(model_transformed$residuals)
```



The Risidual vs. Index plot shows the observations index on the x-axis and its risidual on the y-axis. We want a random scattering of residuals around $\epsilon = 0$ (i.e. no correlation of the errors). we can clearly see that this is the case, so **4. The errors are uncorrelated.** is satisfied.

## Conclusion

Our model fully satisfies all linear regression assumptions, and almost fully explains the variance 0.9872994. Although little in explanatory power can be gained from a model selection process, it will be conducted for completeness.

```
data <- transform(data, Weight = sqrt(Weight))
write.csv(data, 'cleaned_data_4.csv', row.names = FALSE)
```