Risidual Analysis

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

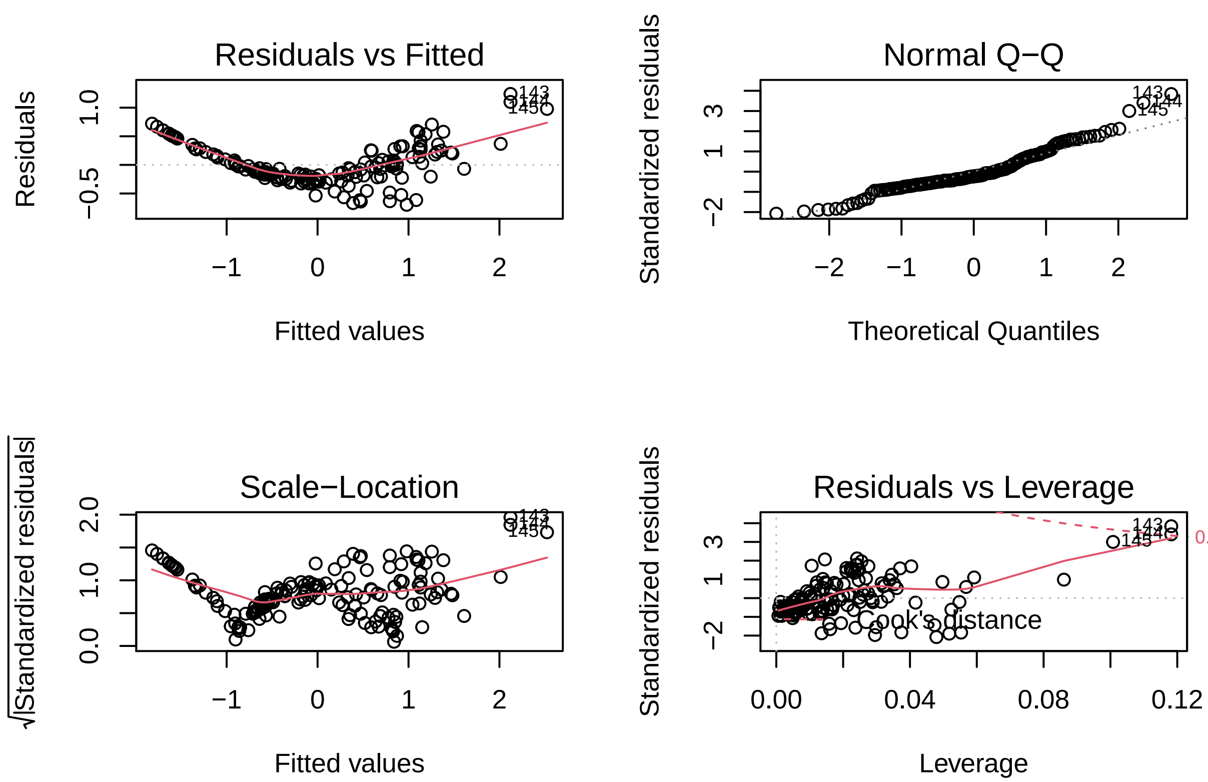
1. The relationship between the response y and the regression is linear (at least approximately).
2. The error term ϵ has zero mean.
3. The error term ϵ has constant variance σ^2.
4. The errors are uncorrelated.
5. The errors are normally distributed.

```
data <- read.csv("cleaned_data_scaled_only_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 ~ Length1 + Height + Width - 1, data = data)

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



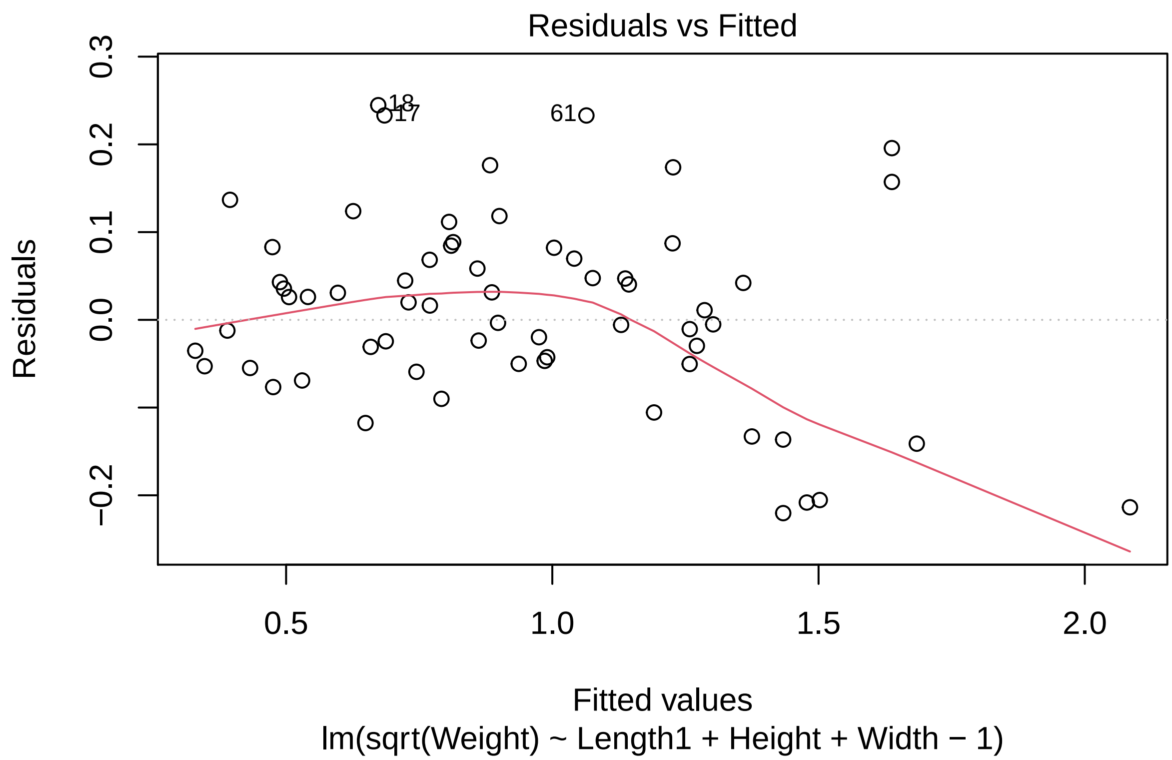
Risiduals vs Fitted

Here, assumption 1. 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) ~ Length1 + Height + Width - 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 ln(). 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 ~ Length1 + Height + Width - 1, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Height + Width - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69766 -0.20417 -0.08122  0.20661  1.23987
##
## Coefficients:
```

```
##          Estimate Std. Error t value Pr(>|t|)
## Length1  0.63298    0.05611  11.282  < 2e-16 ***
## Height   0.16325    0.04587   3.559 0.000494 ***
## Width    0.20868    0.07185   2.905 0.004212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3439 on 156 degrees of freedom
## Multiple R-squared:  0.8832, Adjusted R-squared:  0.881
## F-statistic: 393.2 on 3 and 156 DF,  p-value: < 2.2e-16
```

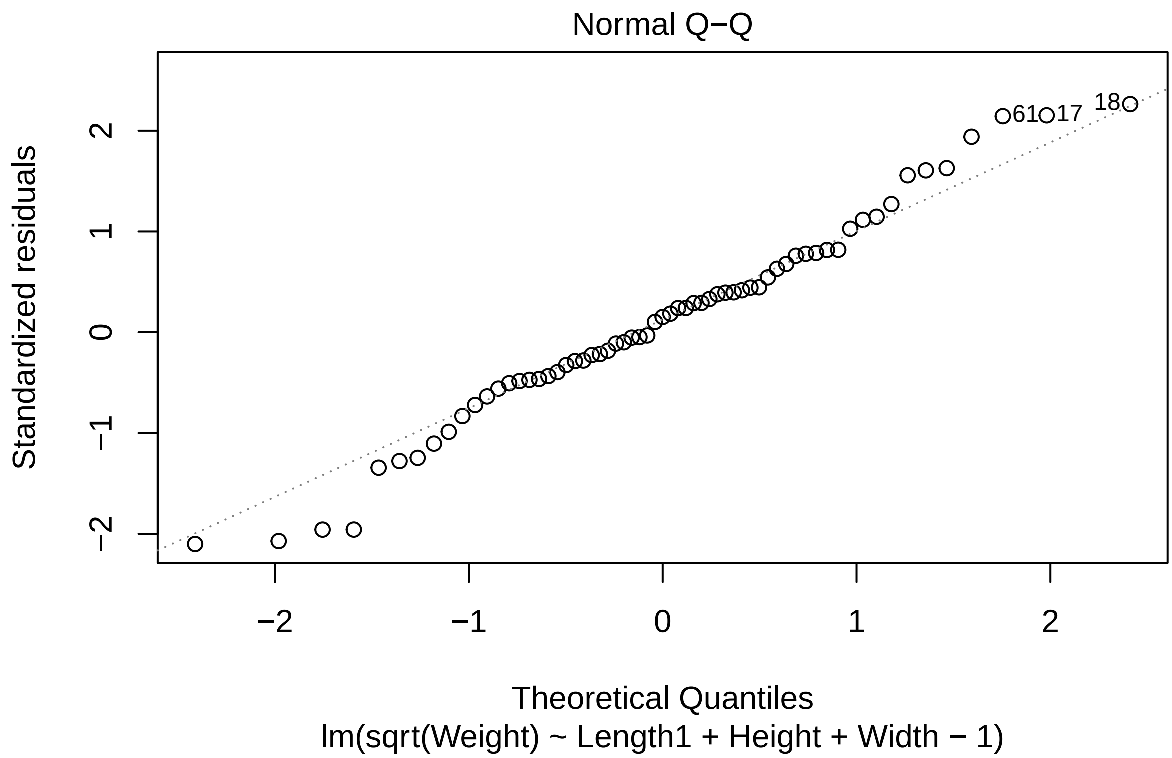
```r
summary(model_transformed)
```

```
##
## Call:
## lm(formula = sqrt(Weight) ~ Length1 + Height + Width - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.22030 -0.05028  0.01631  0.07599  0.24468
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## Length1  0.43068    0.02011   21.42   <2e-16 ***
## Height   0.25147    0.01701   14.79   <2e-16 ***
## Width    0.31142    0.02627   11.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1097 on 60 degrees of freedom
##   (96 observations deleted due to missingness)
## Multiple R-squared:  0.989,  Adjusted R-squared:  0.9885
## F-statistic:  1799 on 3 and 60 DF,  p-value: < 2.2e-16
```

Performing this transformation has significantly improved the R^2 of the model from 0.8831976 to 0.9890021.

Normal Q-Q

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


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(Weight) ~ Length1 + Height + Width − 1)
61
17
18
−2
−1
0
1
2