STAT6175: Linear Models Assignment 2

Predicting Fish Weight from Physical Measurements

Name: Do Nam Phong Phung
ID: 47828013

Questions

1. Read the data file Fish.csv into R calling it Fish. Obtain a histogram of each continuous variable. What transformations, if any, are appropriate? Give reasons for your answer. [5 MARKS]

- Aside from `Species`, all of the variables in the dataset are continuous. Therefore, we will plot a histogram for each of them.
- Square root/Log transformations are both appropriate for `Weight` variable as this variable is heavily right-skewed. Try to use a square root transformation first. If the skewness is too strong and can't be get rid of by a square root transformation, we can apply a log transformation.
- Light right-skewed can be observed in `Height` and `Length1`, `Length2`, `Length3`, depend on the real situation a square root or a log transform could be applied. Note that a square root transformation will result in a lighter transform compared to log transformation.

2. Create new variables sqrt_weight which is the square root of Weight. Obtain a scatter plot matrix of sqrt_weight and all other numerical variables except Weight. Comment. Report VIFs in a regression of sqrt_weight against Length1, Length2, Length3 and Width. Explain why you can only use one of Length1,Length2, Length3 and Width as a predictor in any model fitted for sqrt_weight. Refer to the scatterplot matrix and VIFs. [9 MARKS]

Comment on the scatter plot matrix:

- We can see the positive relationships between `sqrt_weight` and other numerical variables in the dataset. However, heteroscedasticity can be observed in these relationships.
- The correlation between `sqrt_weight` and other numerical variables are strong, most are higher than 0.93, only the correlation with `Height` is at 0.81.

The reason to only use one of `Length1`, `Length2`, `Length3` and `Width` as a predictor in any model fitted for sqrt_weight:

- Refer to the scatterplot matrix: `Length1`, `Length2`, `Length3` all extremely highly correlated to each other (all > 0.99). There is a strong linear relationship between each pair of these three variables.
- Refer to the VIFs: VIF values of three variables are also extremely high, indicates severe multicollinearity.
- Two reasons above show us that the three variables, when being used in a same model as predictors, are going to cover similar information, lead to multicolinearity, affect the estimation and make the model output less reiable.
- On the other hand, aside from a strong correlation with `sqrt_weight` at 0.948, `Width` has a VIF of 5.03, in the acceptable range. Therefore, it is most suitable to use `Weight` as our predictor in this case.

3. Define a new variable new_species that has following three meaningful levels:

- `Freshwater Predators: Pike, Perch`
- `Common Freshwater Species: Bream, Roach, Parkki, Whitefish`
- `Small Forage Fish: Smelt`

Obtain a frequency table of new_species and decide whether any levels should be grouped. Decide also what the reference category should be. [4 MARKS]

- "Small Forage Fish" should be grouped as it only has 14 observations, significantly smaller than the other two categories (has 72 and 73, respectively). (We can combine this label with Common Freshwater Species, which will still have reasonable amount of observations after merging).
- We can use Common Freshwater Species as the reference category, even though is has 1 observation less than the most observation category (but still having a high number of labels), it seems to be the category represents the 'normal' level, which is more commonly to be observed compared to other categories. (Still work and be a great choice even if merged with 'Small Forage Fish').

4. Obtain a graph of sqrt_weight versus Height using different symbols for the different levels of new_species. Add loess curves for the different levels of new_species and comment on the graph. Are Height and new_species promising predictors? Give reasons for your answer. [5 MARKS]

- We see linear positive trends of Small Forage Fish and Common Freshwater Species. Higher `height` seems to lead to a higher `sqrt_weight`. The trend is almost similar to Freshwater Predators, however, the loess curve show a non-linear (curved) relationship between two varibles in this category.
- `Height` and `new_species`, when are together, seem to be promising predictors. With the help of new_species, we can clearly see three pattern (With Freshwater Predators is a bit curved, other 2 categories are linear) of fish species in the relationship plot.
- Without the help of `new_species`, the distribution should look heteroschedastic and it is more difficult to interpret the data.

5. Now perform the regression of sqrt_weight on Height and Length3. List and check the model assumptions. Write down the equation fitted, define all terms. Comment on the regression results. [9 MARKS]

List and check the model assumptions

1. First assumption (VIF scores) - **No multicollinearity**: The VIF values are small (1.98 and 1.98), there is no sign of multicollinearity.
2. Second assumption (std vs fitted plot) - **Constant variance/Homoscedasticity**: Observing the standardized residuals vs fitted values plot, we can see that the variance is not constant around zero. There is a spreaded pattern observed in the plot, suggest heteroschedasticity and that the assumption could be violated.
3. Third assumption (QQ plot) - **Normality**: The plot is heavily tailed at both ends, the pattern does not look linear.
4. Fourth assumption - **Linearity**: Observing the scatter plot matrix in Appendix 2, we can see a linear relationship between these `sqrt_weight` and `Length3` and `Height`.

Conclusion: The model assumptions are not met, there is a sign of heteroschedasticity and the normality assumption is violated. Therefore, the model output could not be trusted, requires further observations.

Write down the equation fitted, define all terms

$$\hat{y_i} = -6.442 + 0.583x_{i1} + 0.606x_{i2}$$

Where (for Fish i):

- $\hat{y_i}$ = The square root of Weight or sqrt_weight
- x_{i1} = Height of the fish
- x_{i2} = Length3/ Cross length of the fish

Comment on the regression results

- The model is **significant** with p-value < 0.05. All predictors are signficant with p-value < 2e-16 < 0.05. However, they only have effect when the model assumptions are not violated.
- R-squared = 0.947, means that with the two predictors, the model is able to explain 94.7% of the variance. This means that the model is able to cover/explain an extremely large variability of the regression.
- For each unit change of `Height`, `sqrt_weight` increase by 0.583 unit. For each unit change of `Length3`, `sqrt_weight` increase by 0.606 unit.

6. Obtain the regression of sqrt_weight on Height, Length3, and new_species. Write down the equation fitted, define all terms. Obtain 95% confidence intervals for the regression coefficients. Is the overall regression significant, and what does this say about the situation? Are any of the predictors significant in the model? Are all model assumptions met? [12 MARKS] (Please select Freshwater Predators as the reference category for new_species.)

Write down the equation fitted, define all terms

$$\hat{y_i} = -5.786 + 0.872x_{i1} + 0.539x_{i2} - 2.549x_{i3} + 0.134x_{i4}$$

Where (for Fish i),

- $\hat{y_i}$ = The square root of Weight or sqrt_weight
- x_{i1} = Height of the fish
- x_{i2} = Length3/ Cross length of the fish
- x_{i3} = Common Freshwater Species
- x_{i4} = Small Forage Fish

Interpret the model

- The overall regression is **significant** with p-value < 0.05. This means that together, the predictors are signficant in predicting `sqrt_weight`.
- `Small Forage Fish` species has a p-value of $0.836 > 0.05$, which means that this predictor is insignificant in predicting the target variable.
- Aside from `Small Forage Fish` species, all of the predictors are significant as their p-values smaller than 0.05. This means that most of them have strong impact when predicting `sqrt_weight`. Note that the effect of the species `Common Freshwater Species` is weaker than `Height` and `Length3` as its p-value is larger compared to the other two.
- R-squared = 0.958, means that with the two predictors, the model is able to explain 95.8% of the variance. This means that the model is able to cover/explain an extremely large variability of the regression. Compared to model `m2`, this is a better result, means that this model could do a better work in predicting `sqrt_weight` and adding `new_species` definitely contributes to that.
- The model says that with every unit change of the predictors, the target `sqrt_weight` changes by the coefficient value. Notice that `Common Freshwater Species` has a negative impact, which is different from other predictors.
- Observing the 95% Confidence Intervals for the regression coefficients, we can see the range that the coefficients are likely to fall into, with 95% confidence. The range also provides information about how the coefficients illustrate the strength/significant of the predictors in the model. With higher coefficient range, the predictor is more likely to be stronger and more signficant in predicting the target variable. On the other hand, when observing the range of `Small Forage Fish`, we can see that it includes zero (means that the predictor has no impact), indicates the insignificant of the predictor.

Model assumptions

1. First assumption (VIF scores) - **No multicollinearity**: The VIF values are small and in the acceptable range. There is no sign of multicollinearity.
2. Second assumption (std vs fitted plot) - **Constant variance/Homoscedasticity**: Observing the standardized residuals vs fitted values plot, we can see that the variance is not constant around zero. There is a spreaded pattern observed in the plot, suggest heteroschedasticity and that the assumption could be violated.
3. Third assumption (QQ plot) - **Normality**: The plot is tailed at both ends and it doesn't seem to be linear.
4. Fourth assumption - **Linearity**: Observing the scatter plot matrix in Appendix 2, we can see a linear relationship between these `sqrt_weight` and `Length3` and `Height`. For `new_species`, as this is a categorical variable, the linearity assumptions always met.

Conclusion: The model assumptions are not met, there is a sign of heteroschedasticity and violation of normality. Therefore, the model output could not be trusted, requires further observations.

7. Using the model fitted in the previous question, which observation has the largest standardised residual, the highest leverage, and the largest value of Cook's distance? Do these values seem to be high enough to regard this observation as an outlier? [6 MARKS]

- The outlier and Leverage Diagnostics plot and the leverage results (Appendix 7) show that observation 145 has the highest leverage (but not an outlier). The Cook's distance bar plot and the calculation results show that observation 41 has the highest Cook's distance, is highly influential. The calculation results show that observation 41 also has the largest standardized residual. Note that observation 41, in the Outlier and Leverage Diagnostics plot, is an outlier.
- The values of 41 seems to be high enough for it to be considered as outliers (high residuals). Note that with low leverage and high residuals, we can confirm that the impact of this observation on the model fit is more likely due to its error.
- For observation 145, the values are not enough for it to be considered as an outlier.

Appendix

Appendix 1

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ---------------------------------------------------------------
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ---------------------------------------------------------------------------------------
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(gridExtra)
```

```
Attaching package: 'gridExtra'

The following object is masked from 'package:dplyr':

    combine
```

```r
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```r
library(car)
```

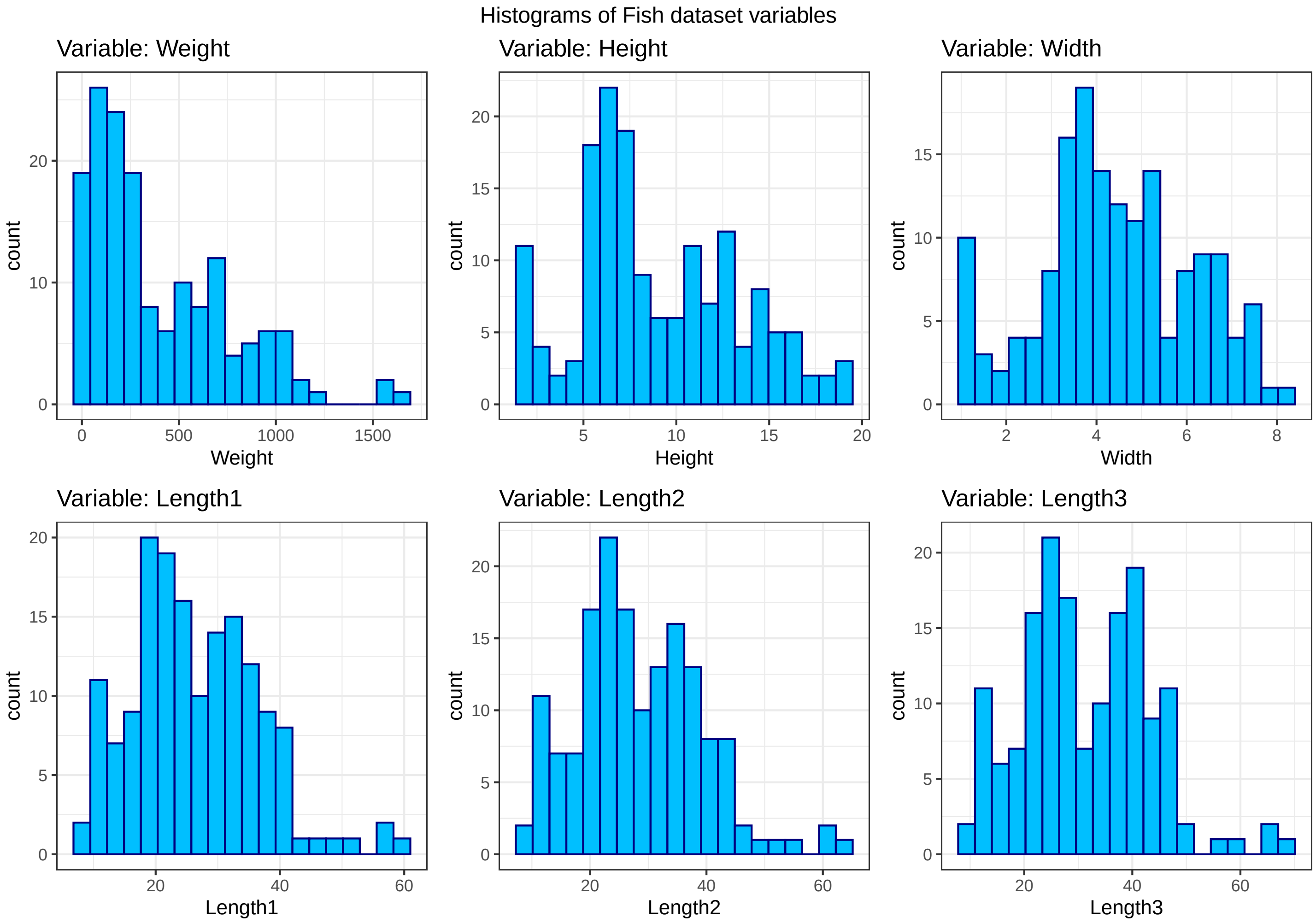
```
Loading required package: carData
```

```
Attaching package: 'car'

The following object is masked from 'package:dplyr':

    recode

The following object is masked from 'package:purrr':

    some
```

```
library(olsrr)
```

```
Attaching package: 'olsrr'

The following object is masked from 'package:datasets':

    rivers
```

```r
# Load datasets
Fish <- read.csv("Fish.csv")

# Create a function to plot the data dynamically
plot_histogram <- function(data, x) {
    # Plot the histogram of variable x of dataframe data
    # Parameters:
    #     data (DataFrame): A DataFrame dataset
    #     x (numeric): A continuous variable to be plotted
    # Return:
    #     A ggplot2 histogram plot count the value occurrence of 'x'
    data %>% ggplot(aes(x = .data[[x]]))+
        geom_histogram(
            bins = 20,
            color = 'navyblue',
            fill = 'deepskyblue'
        )+
        labs(
            title = paste('Variable:', x)
        )+
        theme_bw()
}

# Generate histograms of all `Fish` cont variables
p1 <- plot_histogram(Fish, "Weight")
p2a <- plot_histogram(Fish, "Length1")
p2b <- plot_histogram(Fish, "Length2")
p2c <- plot_histogram(Fish, "Length3")
p3 <- plot_histogram(Fish, "Height")
p4 <- plot_histogram(Fish, "Width")


# Arrange all plots into a grid then show it
grid.arrange(p1, p3, p4, p2a, p2b, p2c, ncol = 3, nrow = 2, top = 'Histograms of Fish dataset variabl
```



Histograms of Fish dataset variables

Variable: Weight

Variable: Height

Variable: Width

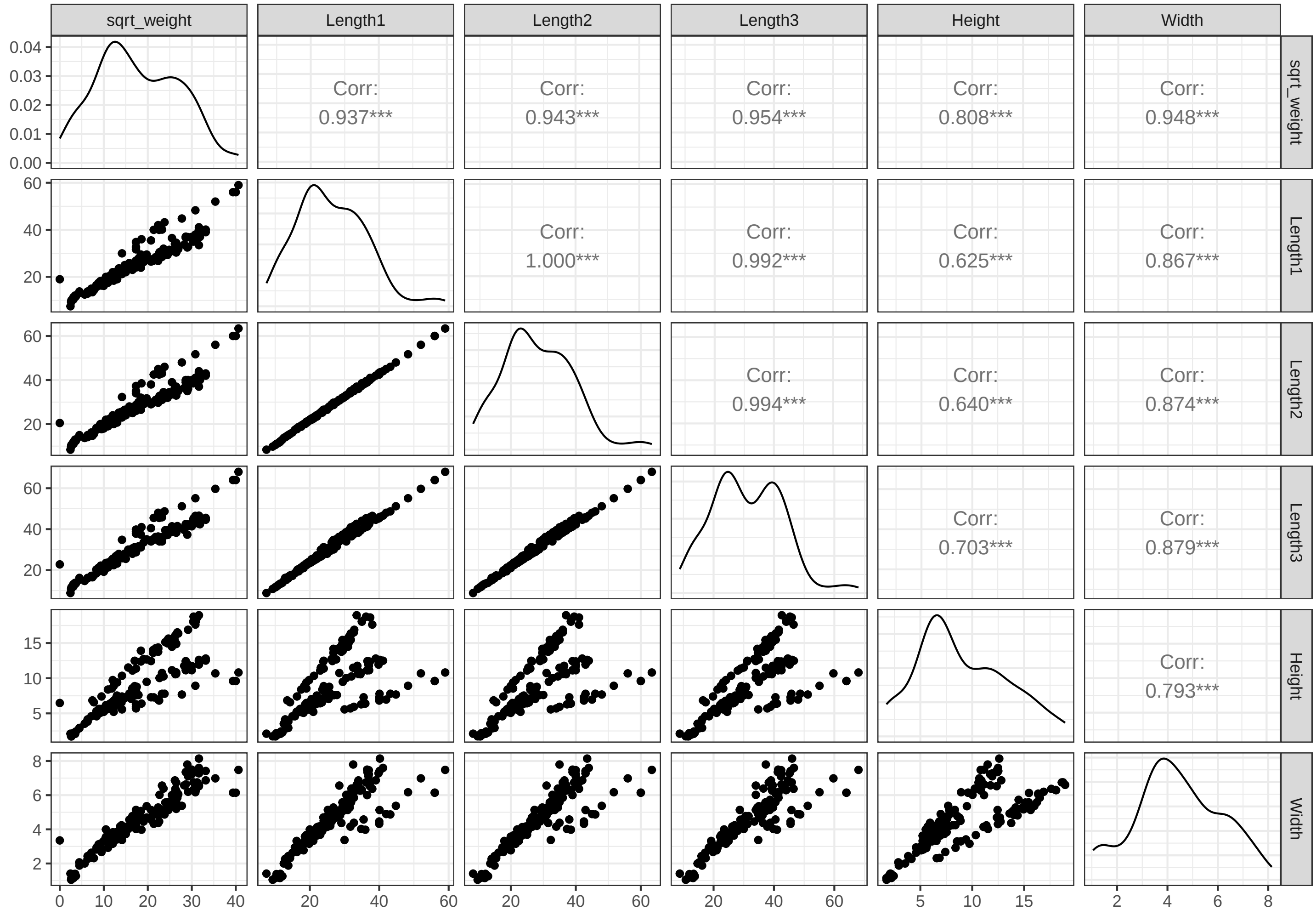
Variable: Length1

Variable: Length2

Variable: Length3

Appendix 2

```
# Create new variable `sqrt_weight`
Fish$sqrt_weight <- sqrt(Fish$Weight)

# Obtain a scatter plot matrix of `sqrt_weight` and other variables except `Weight`
ggpairs(Fish[,c('sqrt_weight', 'Length1', 'Length2', 'Length3', 'Height', 'Width')])+theme_bw()
```



```r
# Regression of `sqrt_weight` against `Length1`, `Length2`, `Length3` and `Width`
m1 <- lm(data = Fish, sqrt_weight ~ Length1 + Length2 + Length3 + Width)
vif(m1)
```

```
    Length1     Length2     Length3       Width 
1531.894896 2056.502202  109.469566    5.033301 
```

Appendix 3

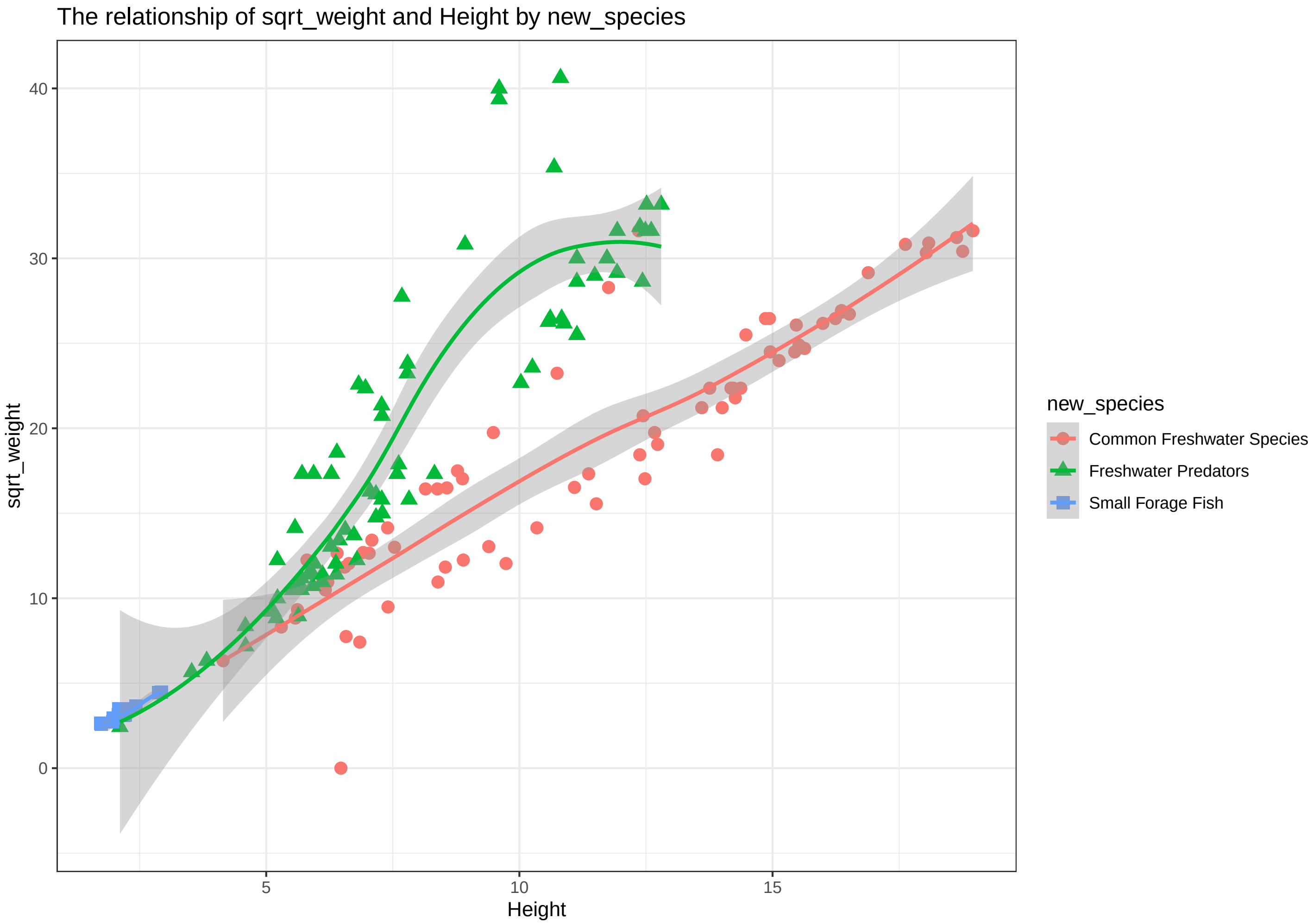
```r
# Create a new variable `new_species`
Fish <- Fish %>%
    mutate(new_species = case_when(
        # Assign each fish species with into its group
        # Check if `Species` is one of value in the set, then assign to the tag
        Species %in% c("Pike", "Perch") ~ "Freshwater Predators",
        Species %in% c("Bream", "Roach", "Parkki", "Whitefish") ~ "Common Freshwater Species",
        Species == "Smelt" ~ "Small Forage Fish"
    ))

# Obtain a frequency table of `new_species`
table(Fish$new_species)
```

```
Common Freshwater Species      Freshwater Predators         Small Forage Fish
                       72                        73                        14
```

Appendix 4

```r
# A graph of `sqrt_weight` ~ `Height`
Fish %>%
    ggplot(aes(
        x = Height,
        y = sqrt_weight,
        # Add color for easier intepretation
        color = as.factor(new_species),
        shape = as.factor(new_species)
    ))+
    geom_point(size = 3)+
    geom_smooth(method = loess)+
    labs(
        title = 'The relationship of sqrt_weight and Height by new_species',
        shape = 'new_species',
        color = 'new_species'
    )+
    theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
```


The relationship of sqrt_weight and Height by new_species
sqrt_weight
40
30
20
10
0
5
10
15
Height
new_species
Common Freshwater Species
Freshwater Predators
Small Forage Fish

Appendix 5

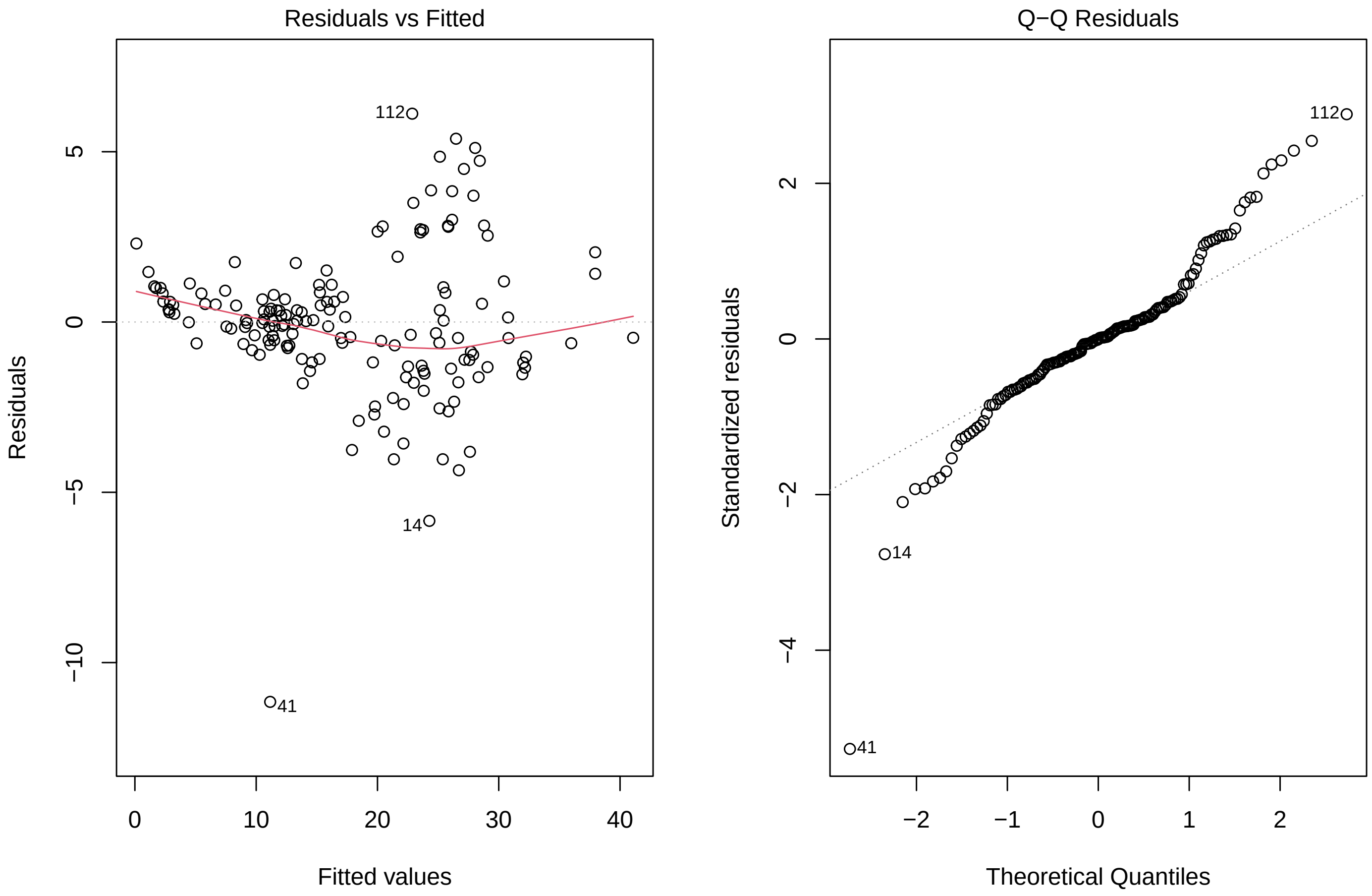
```
# Regression of `sqrt_weight` on `Height` and `Length3`
m2 <- lm(data = Fish, sqrt_weight ~ Height + Length3)
summary(m2)
```

```
Call:
lm(formula = sqrt_weight ~ Height + Length3, data = Fish)

Residuals:
     Min       1Q   Median       3Q      Max 
-11.1536  -0.9901  -0.0087   0.8395   6.1175 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.44184    0.48780  -13.21   <2e-16 ***
Height       0.58259    0.05555   10.49   <2e-16 ***
Length3      0.60627    0.02051   29.56   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.127 on 156 degrees of freedom
Multiple R-squared:  0.9474,    Adjusted R-squared:  0.9467 
F-statistic:  1405 on 2 and 156 DF,  p-value: < 2.2e-16
```

```
# Determine VIFs for multicollinearity check
vif(m2)
```

```
  Height  Length3 
1.979352 1.979352 
```

```r
# Diagnostic plot for constant variance & normality
par(mfrow = c(1,2))
plot(m2, which = 1:2)
```



Appendix 6

```r
# Re-level the reference category of new_species
Fish$new_species <- as.factor(Fish$new_species)
Fish$new_species <- relevel(Fish$new_species, ref = "Freshwater Predators")
# Regression of `sqrt_weight` on `Height`, `Length3` and `new_species`
m3 <- lm(data = Fish, sqrt_weight ~ Height + Length3 + new_species)
summary(m3)
```

```
Call:
lm(formula = sqrt_weight ~ Height + Length3 + new_species, data = Fish)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.5938 -0.9310 -0.0785  0.9800  5.7512 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -5.78594    0.56441 -10.251  < 2e-16 ***
Height                                0.87230    0.06931  12.585  < 2e-16 ***
Length3                               0.53863    0.02273  23.700  < 2e-16 ***
new_speciesCommon Freshwater Species -2.54941    0.43024  -5.926 1.97e-08 ***
new_speciesSmall Forage Fish          0.13382    0.64622   0.207    0.836    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.923 on 154 degrees of freedom
Multiple R-squared:  0.9576,	Adjusted R-squared:  0.9565 
F-statistic: 868.9 on 4 and 154 DF,  p-value: < 2.2e-16
```

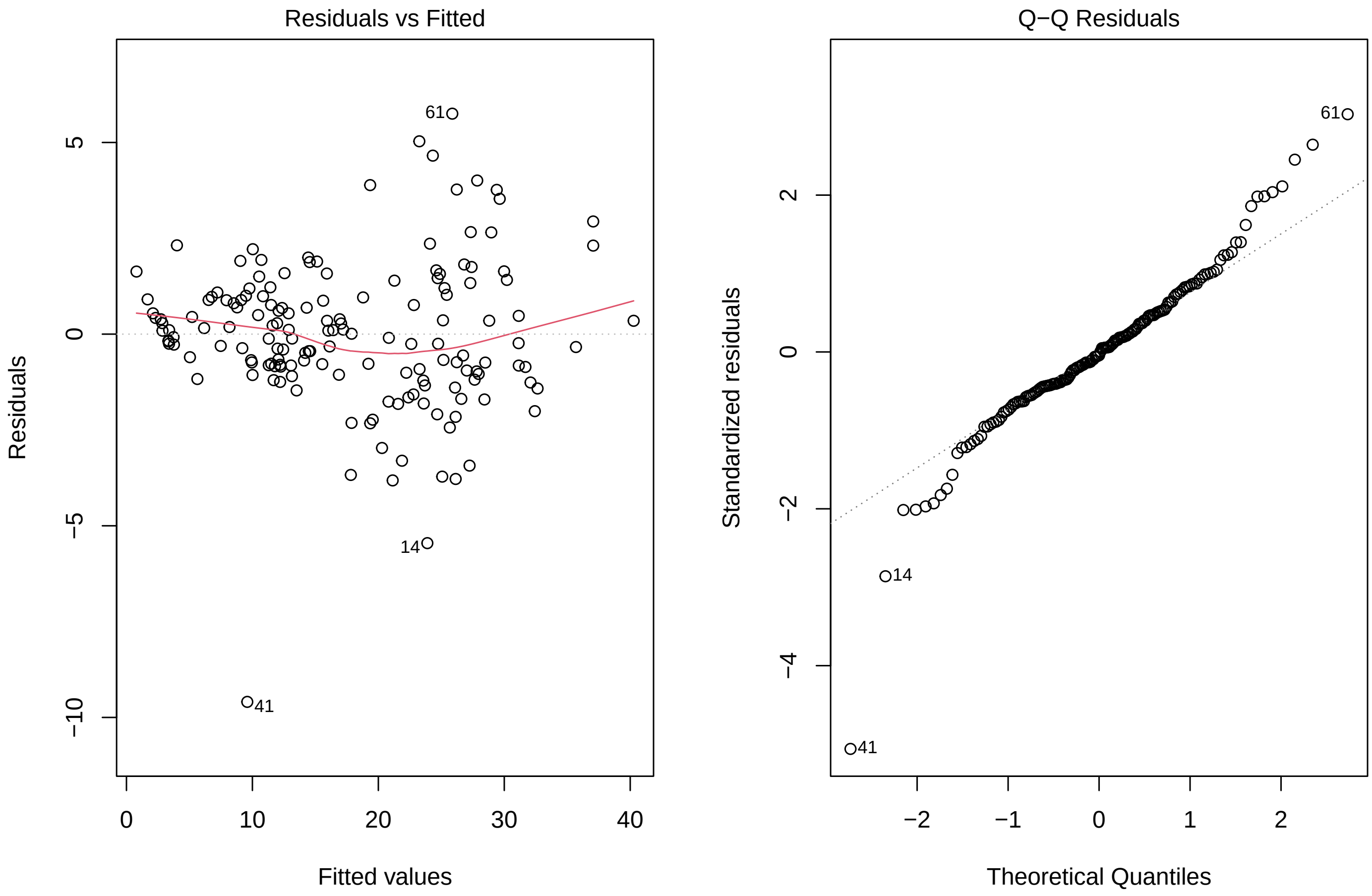
```
# Obtain 95% CI for the regression coefficients
confint(m3)
```

```
                                          2.5 %     97.5 %
(Intercept)                          -6.9009186 -4.6709638
Height                                0.7353700  1.0092233
Length3                               0.4937363  0.5835325
new_speciesCommon Freshwater Species -3.3993536 -1.6994714
new_speciesSmall Forage Fish         -1.1427846  1.4104246
```

```
# Determine VIFs for multicollinearity check
vif(m3)
```

```
                GVIF Df GVIF^(1/(2*Df))
Height      3.769993  1        1.941647
Length3     2.974109  1        1.724561
new_species 2.526390  2        1.260739
```

```
# Diagnostic plot for constant variance & normality
par(mfrow = c(1,2))
plot(m3, which = 1:2)
```



Appendix 7

```
options(pillar.sigfig = 10)

# Find the observation with the highest std res
# Find the observation with the largest absolute standardized residual
std_residuals <- rstandard(m3)
# Find the position of the max value
max_res_obs <- which.max(abs(std_residuals))
# Print out the result
std_residuals[max_res_obs]
```

```
      41
-5.06177
```

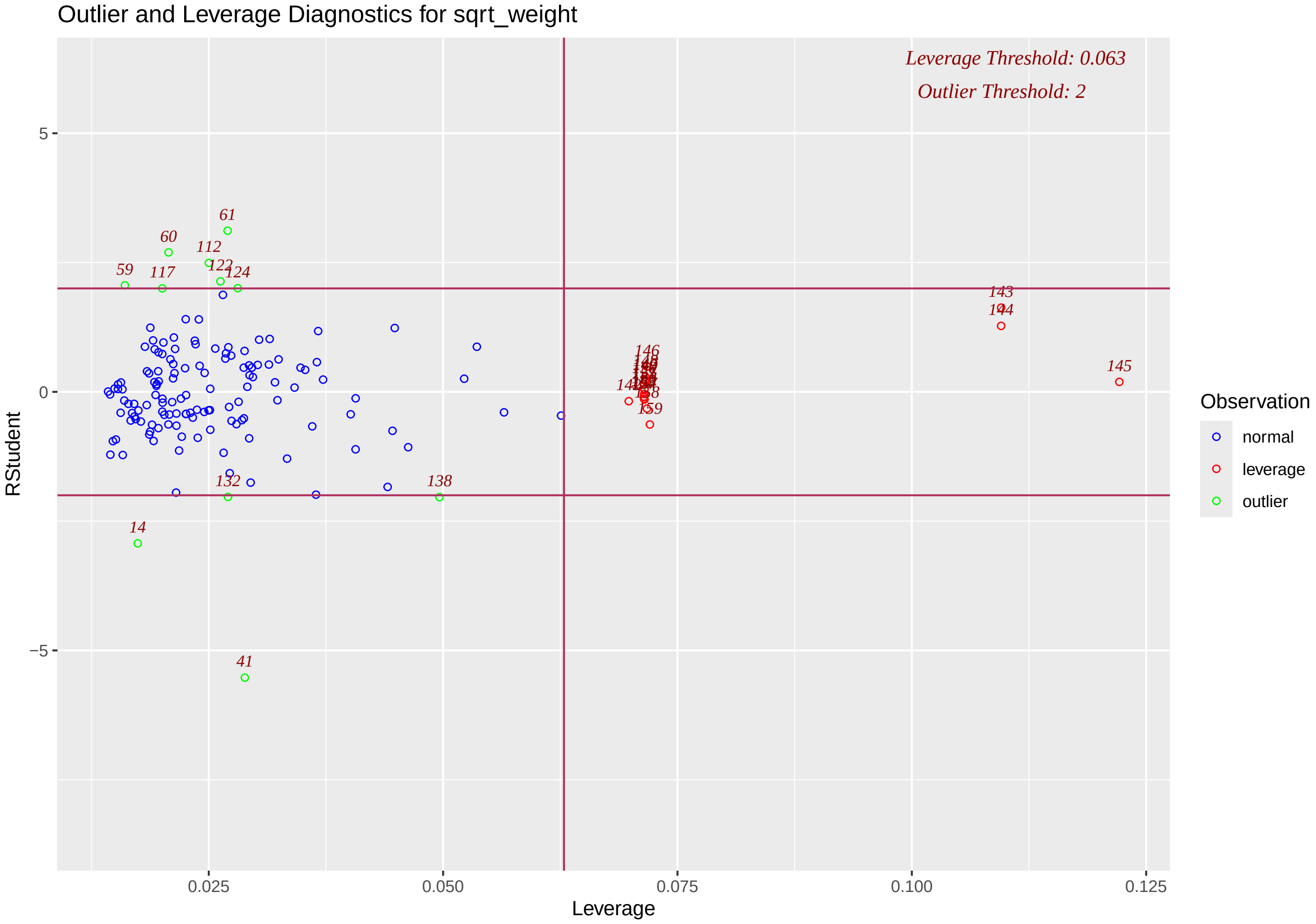
```
# Find the observation with the highest leverage
leverage <- hatvalues(m3)
max_lev_obs <- which.max(abs(leverage))
leverage[max_lev_obs]
```

```
      145
0.1221327
```

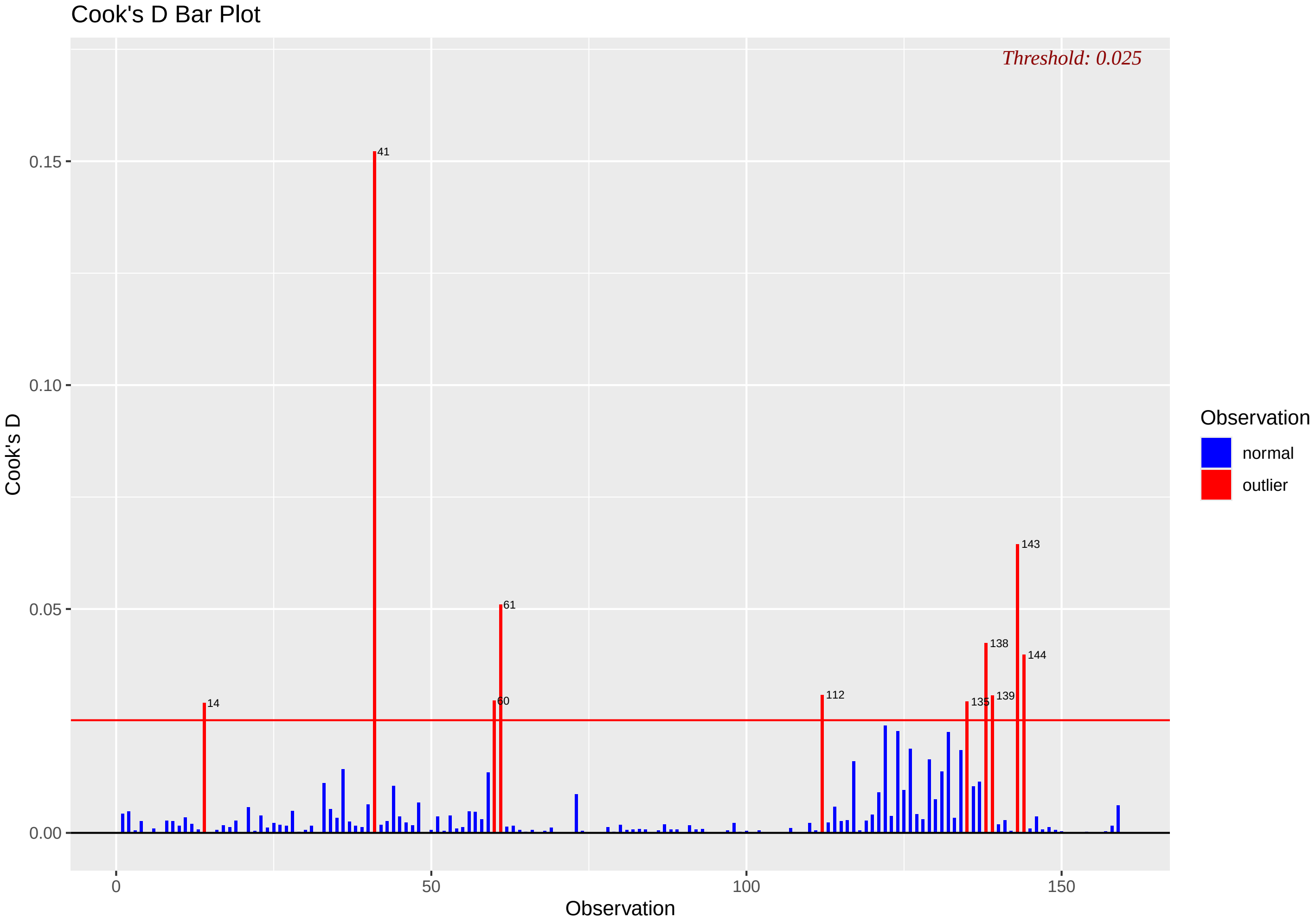
```
# Find the observation with the highest Cook's distance
cook_dist <- cooks.distance(m3)
max_cook_obs <- which.max(abs(cook_dist))
cook_dist[max_cook_obs]
```

```
       41
0.1522189
```

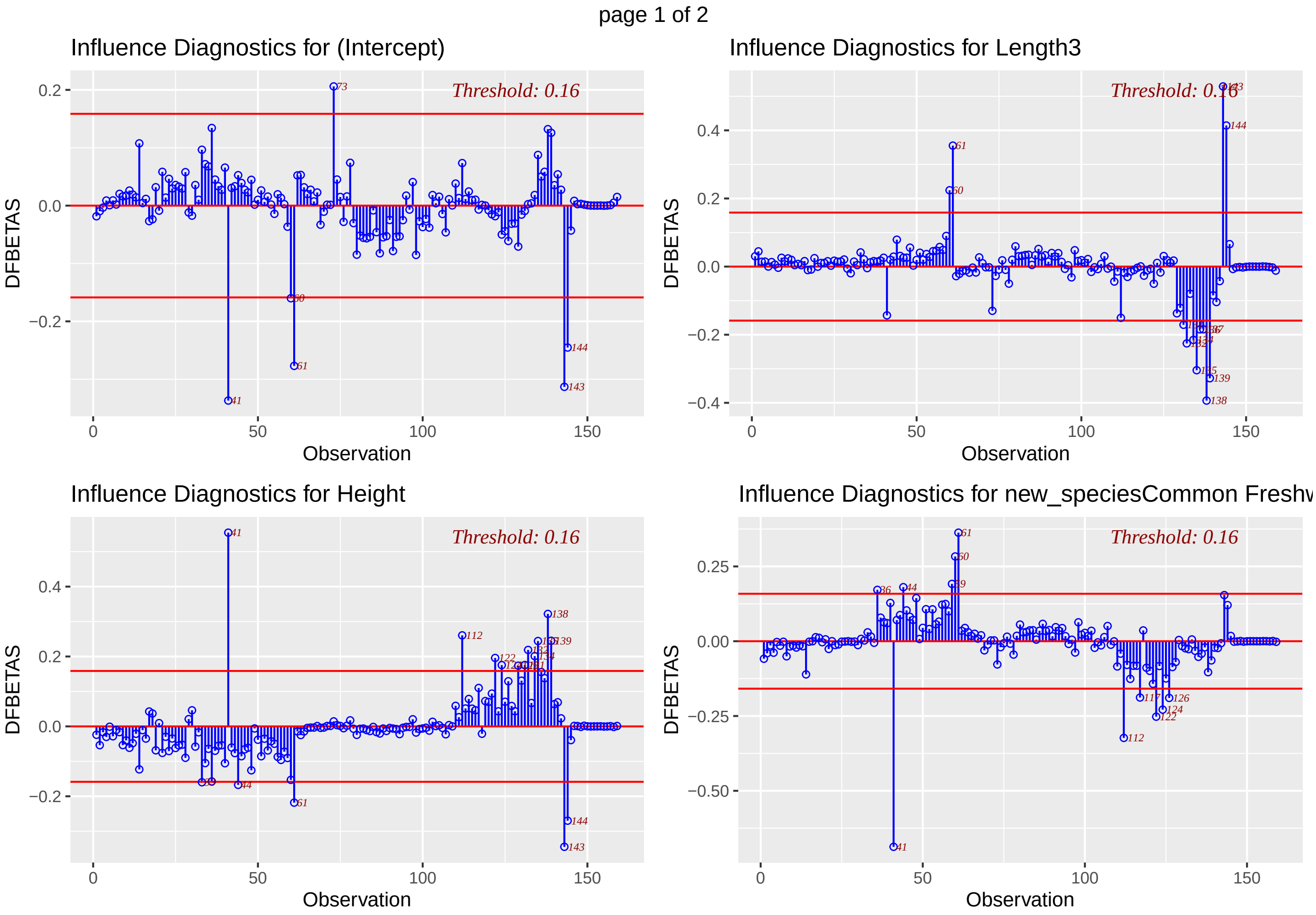
```
# Plot for a better intepretation
ols_plot_resid_lev(m3)
```

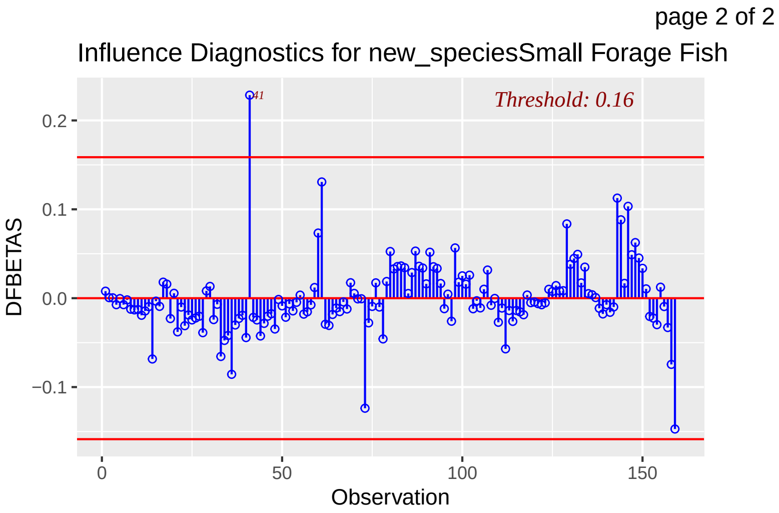


```
ols_plot_cooksd_bar(m3)
```



```
ols_plot_dfbetas(m3)
```





```r
# Determine the residual of observation 145 to double-check this point
std_residuals[145]
```

```
      145
0.1929626
```