Fishing for Insights: Statistical Modeling for Fish Weight Prediction

Mark He

2025-12-13

Libraries

```r
library(faraway)
```

```
## Warning: package ’faraway’ was built under R version 4.4.3
```

```r
library(MASS)
```

```
## Warning: package ’MASS’ was built under R version 4.4.3
```

```r
library(car)
```

```
## Warning: package ’car’ was built under R version 4.4.3

## Loading required package: carData

## Warning: package ’carData’ was built under R version 4.4.3

##
## Attaching package: ’car’

## The following objects are masked from ’package:faraway’:
##
##     logit, vif
```

```r
library(lmtest)
```

```
## Warning: package ’lmtest’ was built under R version 4.4.3

## Loading required package: zoo

## Warning: package ’zoo’ was built under R version 4.4.3

##
## Attaching package: ’zoo’

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

```
fish = read.csv("Fish.csv", stringsAsFactors = TRUE)
head(fish)
```

```
##   Species Weight Length1 Length2 Length3  Height  Width
## 1   Bream    242    23.2    25.4    30.0 11.5200 4.0200
## 2   Bream    290    24.0    26.3    31.2 12.4800 4.3056
## 3   Bream    340    23.9    26.5    31.1 12.3778 4.6961
## 4   Bream    363    26.3    29.0    33.5 12.7300 4.4555
## 5   Bream    430    26.5    29.0    34.0 12.4440 5.1340
## 6   Bream    450    26.8    29.7    34.7 13.6024 4.9274
```

Part 1: Descriptive study of the dataset

In Part 1, I conduct summary analysis of the original dataset, and delete a datapoint with the weight of zero. Then, I draw a box plot about weights by species, a histogram of weights, and scatter plots of every pair of numeric variables. I also print out the correlation matrix of all numeric variables. From the box plot, I can see that species has an influence on weights. For example, Bream has the largest median weight in all 7 types of fish. The variance of the weights of Perch, Pike and Whitefish are noticeably higher than the rest. In the histogram, I can see that the weights of most of the fish are below 1000 grams. 68 out of 158 fish have an weight that is below 200 grams. This distribution of weights suggests that we may need to use a log transformation on the weights when we are doing statistical modeling. As for the scatter plots and the correlation matrix, I can see that some pairs of numeric variables have strong linear correlation. Pairs that involve any two of Length1, Length2 and Length3 have correlation coefficients that are bigger than 0.99. Therefore, it is a good idea to keep only one of the three variables in my model to avoid collinearity. The shapes of other scatter plots are more towards curves, which indicates that we may need to use transformations on some of our variables. According to the finding in the histogram, it is likely that we need to use a log transformation on the 'Weight' variable.

```
str(fish)
```

```
## 'data.frame':    159 obs. of  7 variables:
##  $ Species: Factor w/ 7 levels "Bream","Parkki",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Weight : num  242 290 340 363 430 450 500 390 450 500 ...
##  $ Length1: num  23.2 24 23.9 26.3 26.5 26.8 26.8 27.6 27.6 28.5 ...
##  $ Length2: num  25.4 26.3 26.5 29 29 29.7 29.7 30 30 30.7 ...
##  $ Length3: num  30 31.2 31.1 33.5 34 34.7 34.5 35 35.1 36.2 ...
##  $ Height : num  11.5 12.5 12.4 12.7 12.4 ...
##  $ Width  : num  4.02 4.31 4.7 4.46 5.13 ...
```

```
cat("\n")
```

```
cat("Here is a summary of the dataset:\n\n")
```

```
## Here is a summary of the dataset:
```

```
summary(fish)
```

```
##       Species       Weight          Length1         Length2     
##  Bream    :35   Min.   :   0.0   Min.   : 7.50   Min.   : 8.40  
##  Parkki   :11   1st Qu.: 120.0   1st Qu.:19.05   1st Qu.:21.00  
##  Perch    :56   Median : 273.0   Median :25.20   Median :27.30  
##  Pike     :17   Mean   : 398.3   Mean   :26.25   Mean   :28.42  
##  Roach    :20   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.50  
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40  
##  Whitefish: 6                                                   
##     Length3          Height           Width      
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048  
##  1st Qu.:23.15   1st Qu.: 5.945   1st Qu.:3.386  
##  Median :29.40   Median : 7.786   Median :4.248  
##  Mean   :31.23   Mean   : 8.971   Mean   :4.417  
##  3rd Qu.:39.65   3rd Qu.:12.366   3rd Qu.:5.585  
##  Max.   :68.00   Max.   :18.957   Max.   :8.142  
## 
```

```
sum(is.na(fish))
```

```
## [1] 0
```

```
cat("As shown in the summary, the min value of Weight is 0.0,\nso we need to delete that data point.\n\
```

```
## As shown in the summary, the min value of Weight is 0.0,
## so we need to delete that data point.
```

```
print(paste("The index of the fish with the weight of zero: ",which(fish$Weight == 0)))
```

```
## [1] "The index of the fish with the weight of zero:  41"
```

```
cat("\n The summary of the dataset after deleting that data point: \n")
```

```
## 
##  The summary of the dataset after deleting that data point:
```

```
fish_c = fish[fish$Weight != 0, ]
summary(fish_c)
```

```
##       Species       Weight          Length1         Length2     
##  Bream    :35   Min.   :   5.9   Min.   : 7.50   Min.   : 8.40  
##  Parkki   :11   1st Qu.: 121.2   1st Qu.:19.15   1st Qu.:21.00  
##  Perch    :56   Median : 281.5   Median :25.30   Median :27.40  
##  Pike     :17   Mean   : 400.8   Mean   :26.29   Mean   :28.47  
##  Roach    :19   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.75  
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40  
##  Whitefish: 6                                                   
##     Length3          Height           Width      
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048  
##  1st Qu.:23.20   1st Qu.: 5.941   1st Qu.:3.399  
##  Median :29.70   Median : 7.789   Median :4.277  
##  Mean   :31.28   Mean   : 8.987   Mean   :4.424  
##  3rd Qu.:39.67   3rd Qu.:12.372   3rd Qu.:5.587  
##  Max.   :68.00   Max.   :18.957   Max.   :8.142  
## 
```

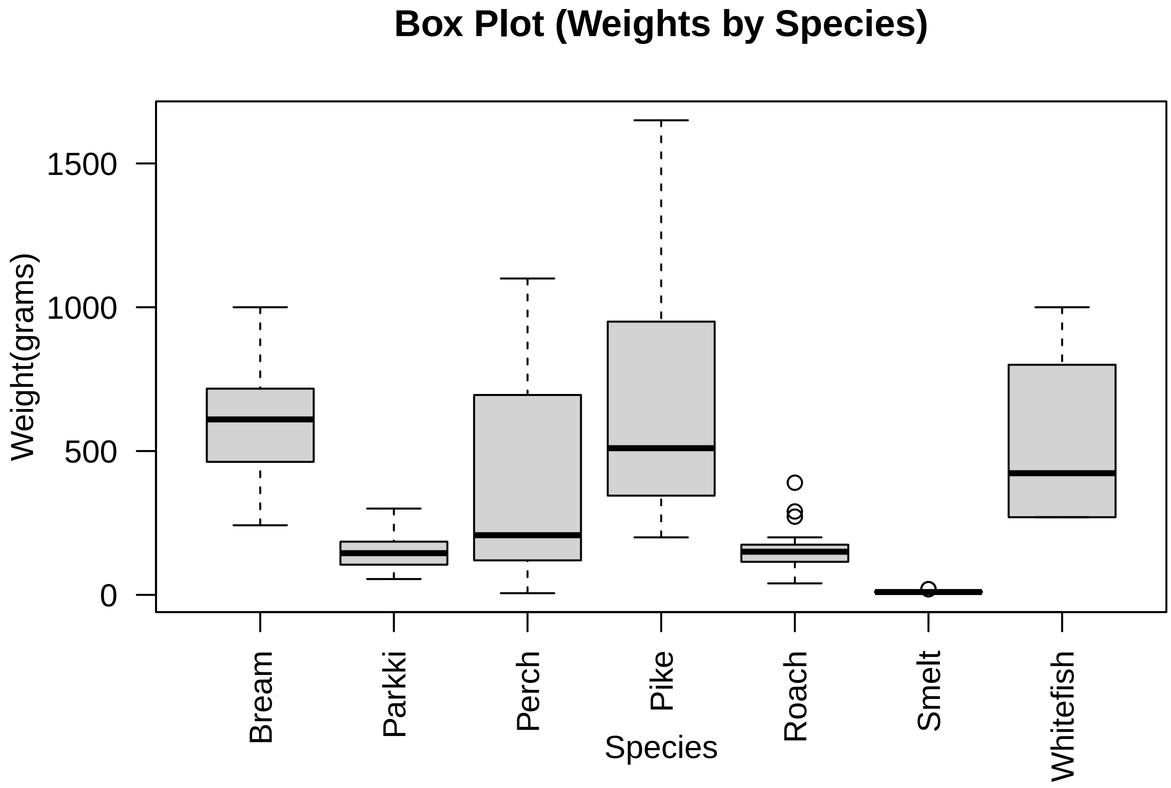
```
table(fish_c$Species)
```

```
##
##      Bream     Parkki      Perch       Pike      Roach      Smelt  Whitefish
##         35         11         56         17         19         14          6
```

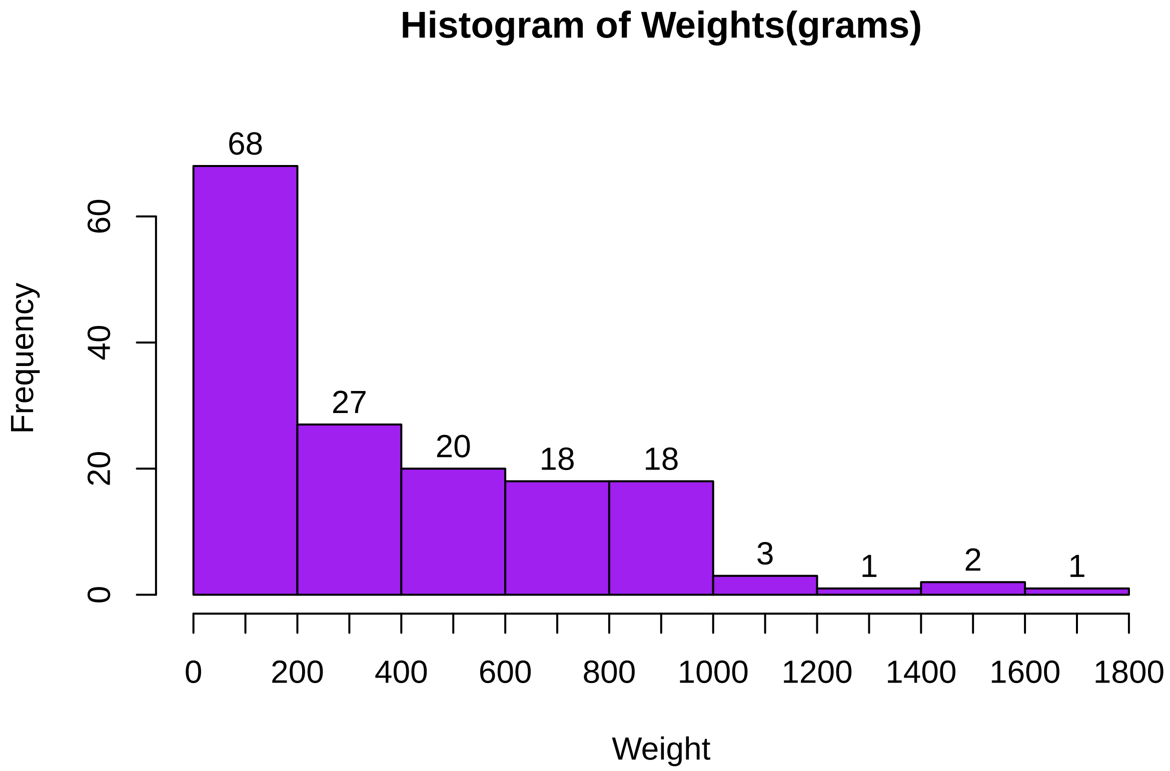
```
nrow(fish_c)
```

```
## [1] 158
```

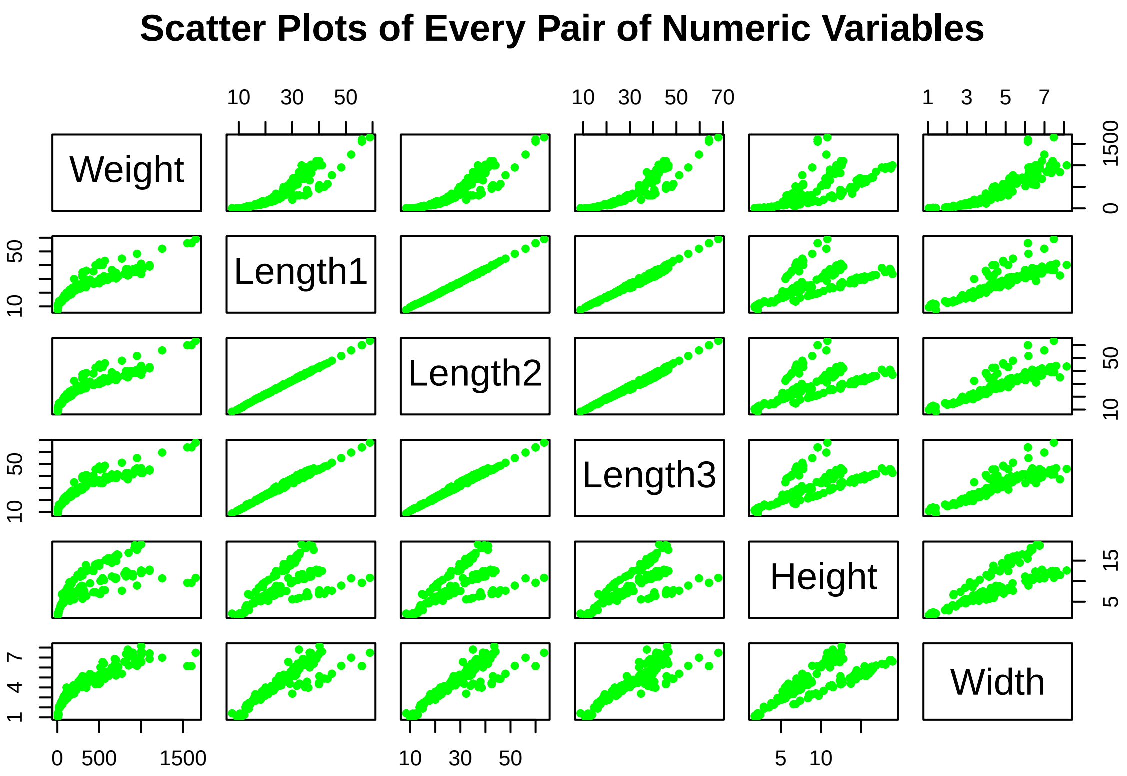
```
boxplot(Weight~Species, data = fish_c, xlab = "Species", ylab = "Weight(grams)", main = "Box Plot (Weig
```



```
hist(fish_c$Weight, xlab = "Weight", col = "purple", border = "black", main = "Histogram of Weights(gra
axis(side = 1, at = seq(0, 1800, by = 100))
```



```r
pairs(fish_c[, -1], main = "Scatter Plots of Every Pair of Numeric Variables", pch = 20, col = "green")
```


Scatter Plots of Every Pair of Numeric Variables

```r
corr = cor(fish_c[, -1])
corr
```

```
##            Weight   Length1   Length2   Length3    Height     Width
## Weight  1.0000000 0.9157195 0.9186031 0.9230903 0.7238573 0.8866536
## Length1 0.9157195 1.0000000 0.9995162 0.9920042 0.6244087 0.8666843
## Length2 0.9186031 0.9995162 1.0000000 0.9940830 0.6395032 0.8732011
## Length3 0.9230903 0.9920042 0.9940830 1.0000000 0.7026548 0.8781887
## Height  0.7238573 0.6244087 0.6395032 0.7026548 1.0000000 0.7924005
## Width   0.8866536 0.8666843 0.8732011 0.8781887 0.7924005 1.0000000
```

Part 2: Statistical Modeling

In part 2, we need to develop a model that can best estimate the weight of a fish when given its species and all the tape measurements. For the majority of this section, I only consider the numeric variables, except for the case when I evaluate the original model that contains all the variables. I will select the proper numeric variables to stay in my model. After the numeric variables are determined, I will proceed to explore and discuss the influence of Species, the only categorical variable in the dataset, to my model.

From the diagnostic plot of the original model, where we consider all the explanatory variables including Species, we can see that it does not satisfy the assumptions of linear model. From its Residuals vs Fitted plot, we can see that there is a noticeable curve in the plot, and the residuals are not randomly scattered around zero. The data points in the Q-Q plot also tend to move away from the dotted line as the theoretical quantile increases, which suggests that the error terms do not follow a normal distribution. Therefore, the original model does not meet the assumptions of linear models, and we may need to do some transformations on its variables.

First of all, I test the performance of two types of transformation: box-cox and log. After looking at the diagnostic plots and the Breusch-Pagan test, I choose to use the log transformation on all the numeric variables. When comparing the "Residuals vs Fitted" plots of the two transformations, the residuals in the plot for log transformation are more randomly scattered around zero than that of the box-cox transformation when the fitted values change. Also, while the scale-location plot of the box-cox transformation has a "U-shape" curve, the scale-location plot of the log transformation has a more stable red line. Besides, from the studentized Breusch-Pagan test, the p-value of the log transformation is 0.7686, while the p-value of box-cox transformation is around 0.008 (<0.05). Therefore, we can reject the null hypothesis that the error terms of the model after box-cox transformation are homoscedastic. However, we cannot reject the null hypothesis that the error terms of the model after log transformation are homoscedastic. Therefore, the model after the log transformation is more appropriate than the one with the Box-Cox transformation, as it better satisfies the linearity assumptions. Overall, based on the test and the diagnostic plots, I choose to use log transformation on the numeric variables.

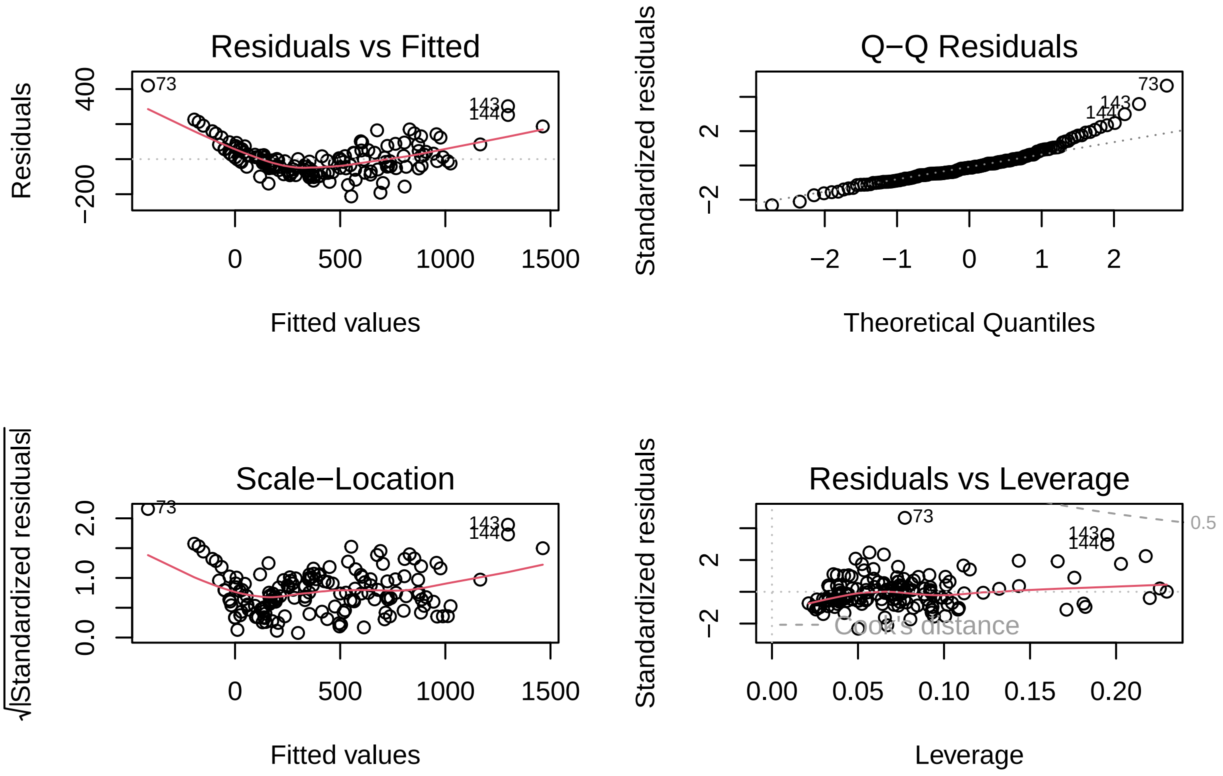
```
basic = lm(Weight ~ ., data = fish_c)
summary(basic)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish_c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -212.56  -52.52  -11.70   36.55  419.97
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -912.7110   127.4597  -7.161 3.64e-11 ***
## SpeciesParkki     160.9212    75.9591   2.119 0.035824 *
## SpeciesPerch      133.5542   120.6083   1.107 0.269969
## SpeciesPike      -209.0262   135.4911  -1.543 0.125061
## SpeciesRoach      104.9243    91.4636   1.147 0.253188
## SpeciesSmelt      442.2125   119.6944   3.695 0.000311 ***
## SpeciesWhitefish   91.5688    96.8338   0.946 0.345901
## Length1           -79.8443    36.3322  -2.198 0.029552 *
## Length2            81.7091    45.8395   1.783 0.076746 .
## Length3            30.2726    29.4837   1.027 0.306233
## Height              5.8069    13.0931   0.444 0.658057
## Width              -0.7819    23.9477  -0.033 0.974000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.95 on 146 degrees of freedom
## Multiple R-squared:  0.9358, Adjusted R-squared:  0.931
## F-statistic: 193.6 on 11 and 146 DF,  p-value: < 2.2e-16
```

```
vif(basic)
```

```
##              GVIF Df GVIF^(1/(2*Df))
## Species 1518.2163  6        1.841243
## Length1 2353.1440  1       48.509216
## Length2 4304.1801  1       65.606250
## Length3 2090.3415  1       45.720253
## Height     56.2501  1         7.500007
## Width      29.0982  1         5.394275
```

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

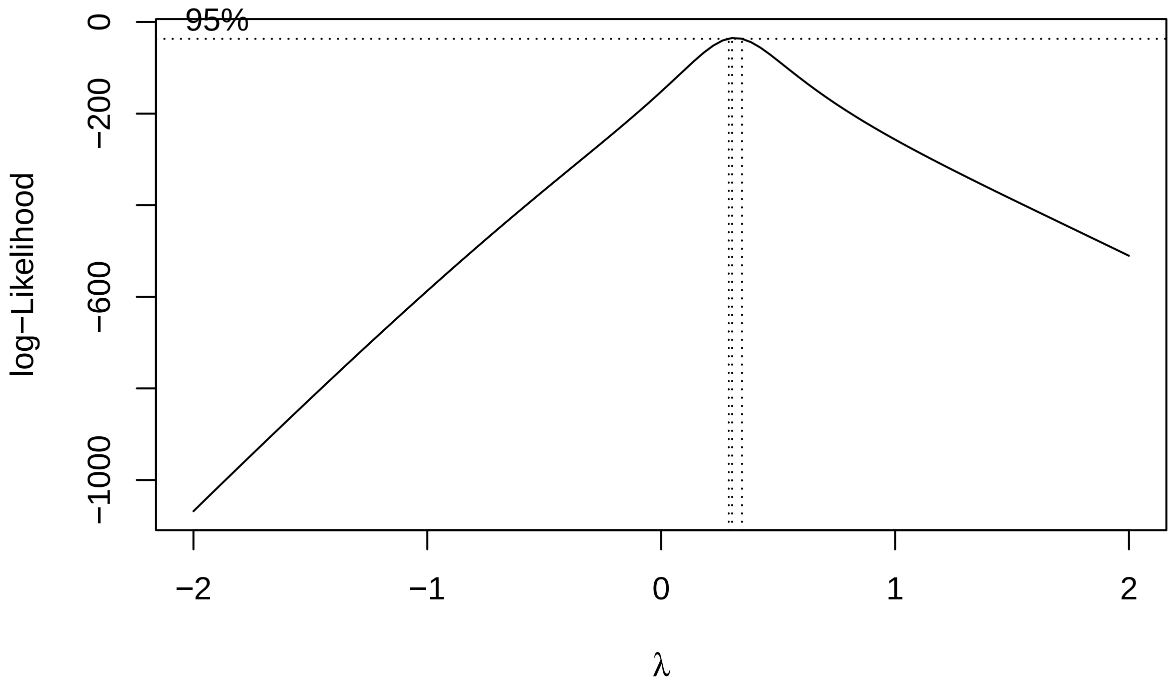


```
par(mfrow = c(1,1))

bp_res_orig = bptest(basic)
print(bp_res_orig)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  basic
## BP = 16.937, df = 11, p-value = 0.1097
```

```
bcr = boxcox(basic)
```



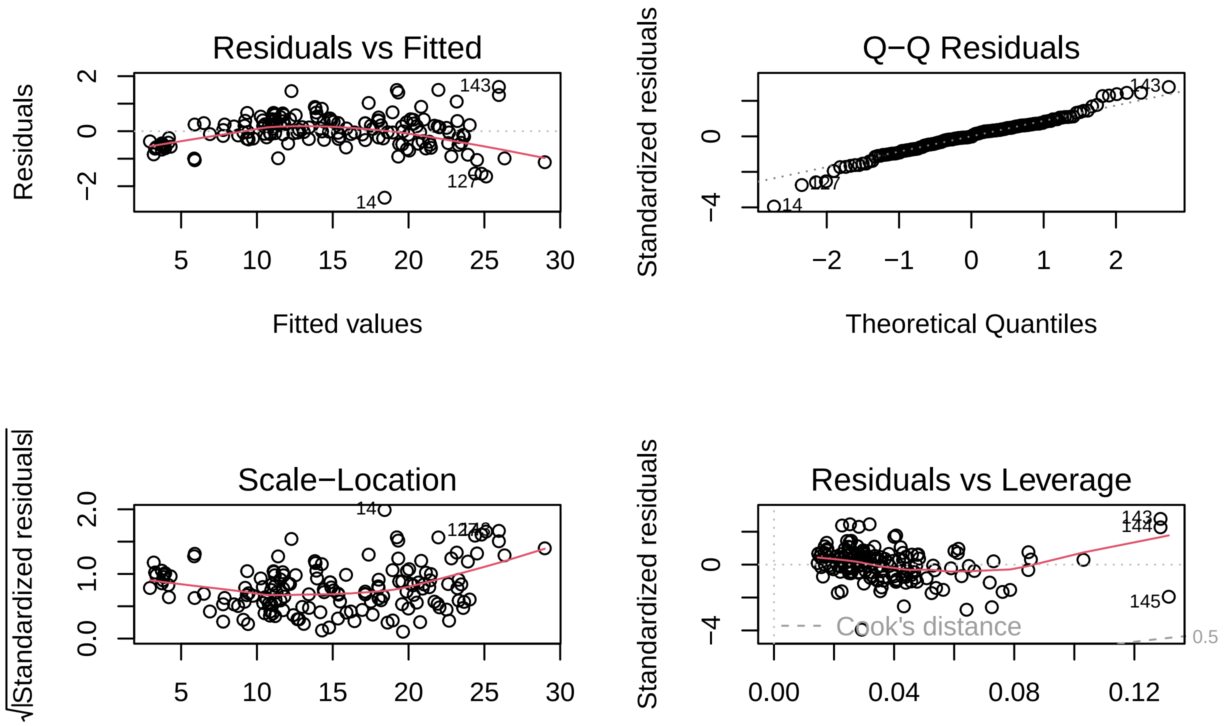
```
opti_l = bcr$x[which.max(bcr$y)]
opti_l
```

```
## [1] 0.3030303
```

```
# Try Box-Cox
fish_bc = fish_c
fish_bc$trans_W1 = (fish_bc$Weight^opti_l - 1)/opti_l
trans_1 = lm(trans_W1 ~ . - Weight- Species, data = fish_bc)
summary(trans_1)
```

```
##
## Call:
## lm(formula = trans_W1 ~ . - Weight - Species, data = fish_bc)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41809 -0.35267 -0.00825  0.35831  1.60886
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.66009    0.14940 -11.111  < 2e-16 ***
## Length1     -0.01663    0.20264  -0.082   0.9347
## Length2      0.35561    0.21061   1.688   0.0934 .
## Length3     -0.06345    0.08749  -0.725   0.4694
## Height       0.37635    0.04400   8.554 1.19e-14 ***
## Width        1.24644    0.10269  12.138  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6206 on 152 degrees of freedom
## Multiple R-squared:  0.9902, Adjusted R-squared:  0.9898
## F-statistic:  3059 on 5 and 152 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2,2))
plot(trans_1)
```



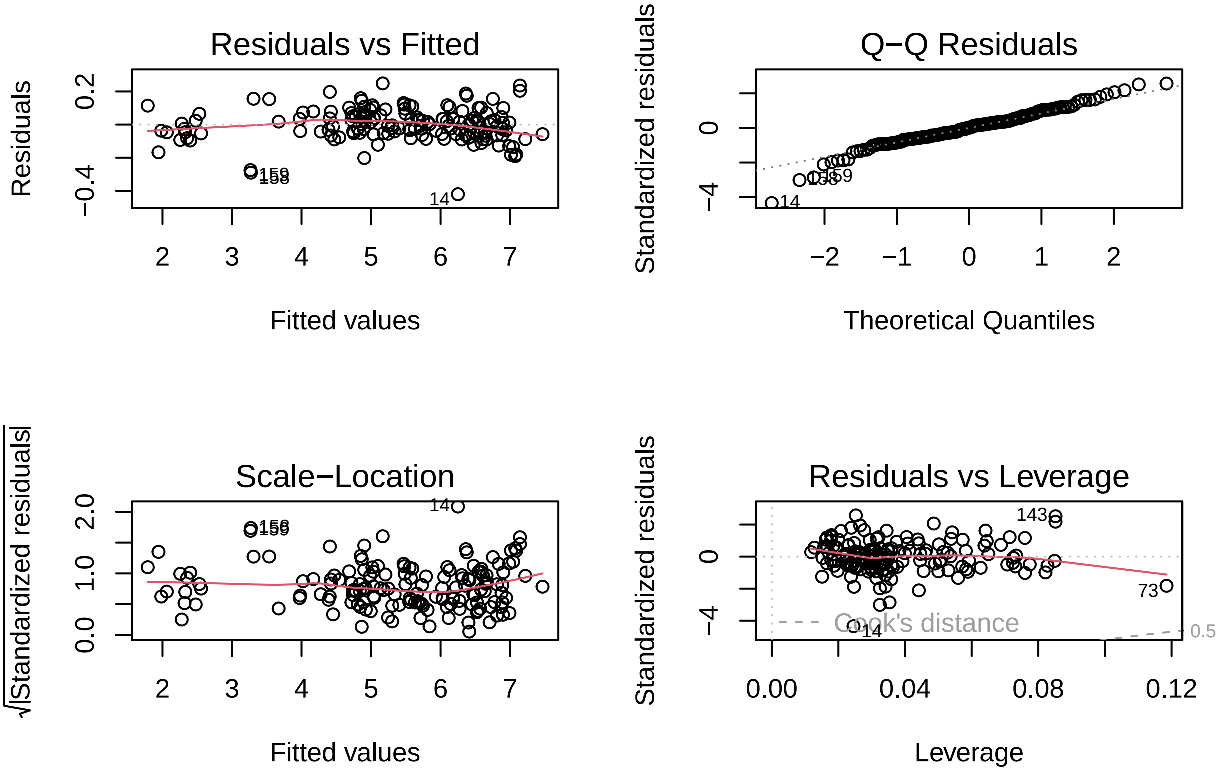
```
par(mfrow = c(1,1))
bp_res_t1 = bptest(trans_1)
print(bp_res_t1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trans_1
## BP = 15.594, df = 5, p-value = 0.008105
```

```r
# Try Log
fish_log = fish_c
fish_log$trans_W2 = log(fish_log$Weight)
trans_2 = lm(trans_W2 ~ log(Length1)+log(Length2)+log(Length3)+log(Width) + log(Height), data = fish_log
summary(trans_2)
```

```
##
## Call:
## lm(formula = trans_W2 ~ log(Length1) + log(Length2) + log(Length3) +
##     log(Width) + log(Height), data = fish_log)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42071 -0.05312  0.00182  0.05491  0.24862
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.94224    0.16375 -11.861   <2e-16 ***
## log(Length1)  0.40077    0.69865   0.574   0.5671
## log(Length2)  1.59553    0.76512   2.085   0.0387 *
## log(Length3) -0.51316    0.37899  -1.354   0.1777
## log(Width)    0.84464    0.08095  10.434   <2e-16 ***
## log(Height)   0.68039    0.05880  11.572   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09805 on 152 degrees of freedom
## Multiple R-squared:  0.9947, Adjusted R-squared:  0.9945
## F-statistic:  5714 on 5 and 152 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(trans_2)
```



```r
par(mfrow = c(1,1))
bp_res_t2 = bptest(trans_2)
print(bp_res_t2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trans_2
## BP = 2.5523, df = 5, p-value = 0.7686
```

Next, I check whether there are outliers in the data set and delete them if necessary. I use cook's distance to evaluate data points' influence to the model. If an observation's cook distance is larger than 0.5, I will delete it. After checking the Residuals vs Leverage plot of the current model, I find that there is no observations whose cook's distance is above 0.5, so I do not need to remove any observations here.

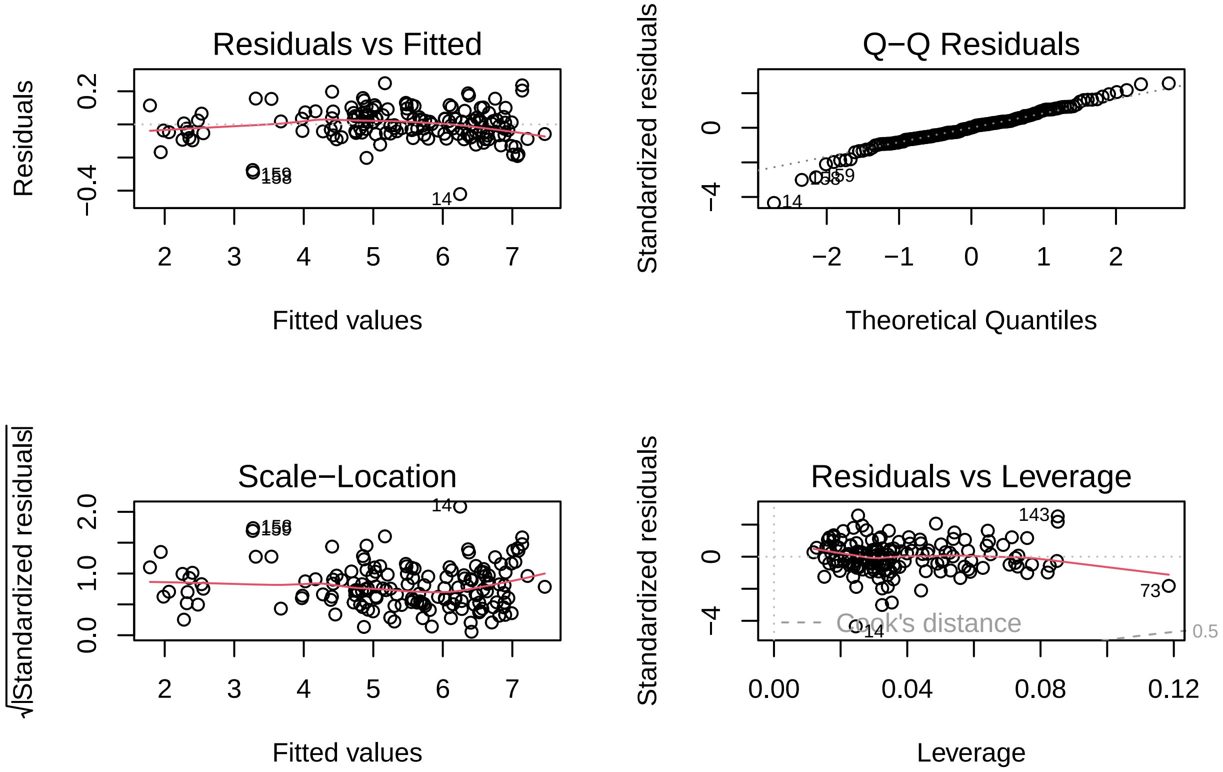
```r
fish_trans = fish_c
fish_trans$trans_W = log(fish_trans$Weight)
summary(fish_trans)
```

```
##      Species        Weight          Length1         Length2
##  Bream    :35   Min.   :   5.9   Min.   : 7.50   Min.   : 8.40
##  Parkki   :11   1st Qu.: 121.2   1st Qu.:19.15   1st Qu.:21.00
##  Perch    :56   Median : 281.5   Median :25.30   Median :27.40
##  Pike     :17   Mean   : 400.8   Mean   :26.29   Mean   :28.47
##  Roach    :19   3rd Qu.: 650.0   3rd Qu.:32.70   3rd Qu.:35.75  
##  Smelt    :14   Max.   :1650.0   Max.   :59.00   Max.   :63.40  
##  Whitefish: 6                                                   
##     Length3          Height           Width           trans_W     
##  Min.   : 8.80   Min.   : 1.728   Min.   :1.048   Min.   :1.775  
##  1st Qu.:23.20   1st Qu.: 5.941   1st Qu.:3.399   1st Qu.:4.798  
##  Median :29.70   Median : 7.789   Median :4.277   Median :5.640  
##  Mean   :31.28   Mean   : 8.987   Mean   :4.424   Mean   :5.410  
##  3rd Qu.:39.67   3rd Qu.:12.372   3rd Qu.:5.587   3rd Qu.:6.477  
##  Max.   :68.00   Max.   :18.957   Max.   :8.142   Max.   :7.409  
## 
```

```
trans = lm(trans_W ~ log(Length1)+log(Length2)+log(Length3)+log(Width) + log(Height), data = fish_trans
summary(trans)
```

```
## 
## Call:
## lm(formula = trans_W ~ log(Length1) + log(Length2) + log(Length3) + 
##     log(Width) + log(Height), data = fish_trans)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.42071 -0.05312  0.00182  0.05491  0.24862 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  -1.94224    0.16375 -11.861   <2e-16 ***
## log(Length1)  0.40077    0.69865   0.574   0.5671    
## log(Length2)  1.59553    0.76512   2.085   0.0387 *  
## log(Length3) -0.51316    0.37899  -1.354   0.1777    
## log(Width)    0.84464    0.08095  10.434   <2e-16 ***
## log(Height)   0.68039    0.05880  11.572   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.09805 on 152 degrees of freedom
## Multiple R-squared:  0.9947, Adjusted R-squared:  0.9945 
## F-statistic:  5714 on 5 and 152 DF,  p-value: < 2.2e-16
```

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



```
par(mfrow = c(1,1))
bp_res_t = bptest(trans)
print(bp_res_t)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  trans
## BP = 2.5523, df = 5, p-value = 0.7686
```

The third step is to do variable selection to develop the current model. Here, I use both stepwise AIC and stepwise BIC method to evaluate and select the variables. I compare the output of the two methods and find that they give the same selection: log(Weight), log(Length2) and log(Height). The p-value of each of them is less than 2e-16. Therefore, I use these three variables to create a new model, whose response variable is log(Weight). Then, I use Breusch-Pagan test on this model, and I find that its p-value is 0.9466, which means that its error terms are homoscedastic. Its Q-Q plot also shows that the residuals approximately follow normal distribution. In its Residual vs Fitted plot, I find that the red line is relatively flat and does not show a noticeable curve shape. The Scale-Location and Residuals vs Leverage plot also present a relatively flat red line. Overall, the plots and the BP test show that the current model is an appropriate linear model.

The output of this part is the model with the three variables I select in step 3. I will this model to do ANCOVA analysis and discuss Species's influence to further develop the model.

```
library(lmtest)
library(alr4)
```

```
## Warning: package 'alr4' was built under R version 4.4.3
```

```
## Loading required package: effects
```

```
## Warning: package 'effects' was built under R version 4.4.3
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
##
## Attaching package: 'alr4'
```

```
## The following objects are masked from 'package:faraway':
##
##     cathedral, pipeline, twins
```

```
empty_m = lm(trans_W ~ 1, data = fish_trans)

full = lm(trans_W ~ log(Length1)+log(Length2)+log(Length3)+log(Width)+log(Height), data = fish_trans)

step_aic = step(empty_m, scope = list(lower = empty_m, upper = full),
                direction = "both", k = 2, trace = TRUE)
```

```
## Start:  AIC=90.2
## trans_W ~ 1
##
##                  Df Sum of Sq     RSS     AIC
## + log(Width)      1    265.17  10.950 -417.74
## + log(Length3)    1    261.10  15.019 -367.82
## + log(Length2)    1    257.66  18.455 -335.27
## + log(Length1)    1    254.72  21.402 -311.86
## + log(Height)     1    233.45  42.665 -202.86
## <none>                        276.118   90.20
##
## Step:  AIC=-417.74
## trans_W ~ log(Width)
##
##                  Df Sum of Sq     RSS     AIC
## + log(Length3)    1     7.499   3.451 -598.19
## + log(Length2)    1     6.063   4.887 -543.21
## + log(Length1)    1     5.866   5.084 -536.97
## + log(Height)     1     1.918   9.032 -446.17
## <none>                         10.950 -417.74
## - log(Width)      1   265.168 276.118   90.20
##
## Step:  AIC=-598.19
## trans_W ~ log(Width) + log(Length3)
##
##                  Df Sum of Sq     RSS     AIC
## + log(Height)     1    1.7250  1.7257 -705.68
## + log(Length1)    1    0.7004  2.7503 -632.04
## + log(Length2)    1    0.6256  2.8252 -627.80
## <none>                          3.4507 -598.19
## - log(Length3)  1    7.4995 10.9502 -417.74
## - log(Width)    1   11.5680 15.0187 -367.82
##
## Step:  AIC=-705.68
## trans_W ~ log(Width) + log(Length3) + log(Height)
##
##                 Df Sum of Sq    RSS     AIC
## + log(Length2)  1    0.2612 1.4645 -729.61
## + log(Length1)  1    0.2225 1.5032 -725.49
## <none>                      1.7257 -705.68
## - log(Height)   1    1.7250 3.4507 -598.19
## - log(Width)    1    3.8376 5.5633 -522.73
## - log(Length3)  1    7.3063 9.0320 -446.17
##
## Step:  AIC=-729.61
## trans_W ~ log(Width) + log(Length3) + log(Height) + log(Length2)
##
##                 Df Sum of Sq    RSS     AIC
## - log(Length3)  1   0.01564 1.4802 -729.93
## <none>                      1.4645 -729.61
## + log(Length1)  1   0.00316 1.4614 -727.95
## - log(Length2)  1   0.26116 1.7257 -705.68
## - log(Width)    1   1.05041 2.5149 -646.18
## - log(Height)   1   1.36064 2.8252 -627.80
##
## Step:  AIC=-729.93
## trans_W ~ log(Width) + log(Height) + log(Length2)
##
##                 Df Sum of Sq    RSS     AIC
## <none>                      1.4802 -729.93
## + log(Length3)  1    0.0156 1.4645 -729.61
## + log(Length1)  1    0.0012 1.4790 -728.06
## - log(Width)    1    2.0036 3.4838 -596.69
## - log(Height)   1    3.4069 4.8871 -543.21
## - log(Length2)  1    7.5519 9.0320 -446.17
```

```
summary(step_aic)
```

```
##
## Call:
## lm(formula = trans_W ~ log(Width) + log(Height) + log(Length2),
##     data = fish_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42835 -0.05739  0.00406  0.05504  0.26420
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.01044    0.12178  -16.51   <2e-16 ***
## log(Width)    0.90246    0.06251   14.44   <2e-16 ***
## log(Height)   0.61206    0.03251   18.83   <2e-16 ***
## log(Length2)  1.49774    0.05343   28.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09804 on 154 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9945
## F-statistic:  9525 on 3 and 154 DF,  p-value: < 2.2e-16
```

```
n = nrow(fish_trans)
step_bic = step(empty_m, scope = list(lower = empty_m, upper = full),
                direction = "both", k = log(n), trace = TRUE)
```

```
## Start:  AIC=93.26
## trans_W ~ 1
##
##                 Df Sum of Sq     RSS     AIC
## + log(Width)     1    265.17  10.950 -411.61
## + log(Length3)   1    261.10  15.019 -361.70
## + log(Length2)   1    257.66  18.455 -329.15
## + log(Length1)   1    254.72  21.402 -305.73
## + log(Height)    1    233.45  42.665 -196.73
## <none>                       276.118   93.26
##
## Step:  AIC=-411.61
## trans_W ~ log(Width)
##
##                 Df Sum of Sq     RSS     AIC
## + log(Length3)   1     7.499   3.451 -589.01
## + log(Length2)   1     6.063   4.887 -534.02
## + log(Length1)   1     5.866   5.084 -527.79
## + log(Height)    1     1.918   9.032 -436.98
## <none>                        10.950 -411.61
## - log(Width)     1   265.168 276.118   93.26
##
## Step:  AIC=-589.01
## trans_W ~ log(Width) + log(Length3)
##
##                 Df Sum of Sq     RSS     AIC
## + log(Height)    1    1.7250  1.7257 -693.43
## + log(Length1)   1    0.7004  2.7503 -619.79
## + log(Length2)   1    0.6256  2.8252 -615.55
## <none>                        3.4507 -589.01
## - log(Length3)   1    7.4995 10.9502 -411.61
## - log(Width)     1   11.5680 15.0187 -361.70
##
## Step:  AIC=-693.43
## trans_W ~ log(Width) + log(Length3) + log(Height)
##
##                 Df Sum of Sq    RSS     AIC
## + log(Length2)   1    0.2612 1.4645 -714.30
## + log(Length1)   1    0.2225 1.5032 -710.18
## <none>                       1.7257 -693.43
## - log(Height)    1    1.7250 3.4507 -589.01
## - log(Width)     1    3.8376 5.5633 -513.54
## - log(Length3)   1    7.3063 9.0320 -436.98
```

```
##
## Step:  AIC=-714.3
## trans_W ~ log(Width) + log(Length3) + log(Height) + log(Length2)
##
##                 Df Sum of Sq    RSS     AIC
## - log(Length3)  1   0.01564 1.4802 -717.68
## <none>                      1.4645 -714.30
## + log(Length1)  1   0.00316 1.4614 -709.57
## - log(Length2)  1   0.26116 1.7257 -693.43
## - log(Width)    1   1.05041 2.5149 -633.92
## - log(Height)   1   1.36064 2.8252 -615.55
##
## Step:  AIC=-717.68
## trans_W ~ log(Width) + log(Height) + log(Length2)
##
##                 Df Sum of Sq    RSS     AIC
## <none>                      1.4802 -717.68
## + log(Length3)  1    0.0156 1.4645 -714.30
## + log(Length1)  1    0.0012 1.4790 -712.74
## - log(Width)    1    2.0036 3.4838 -587.50
## - log(Height)   1    3.4069 4.8871 -534.02
## - log(Length2)  1    7.5519 9.0320 -436.98
```

```
summary(step_bic)
```

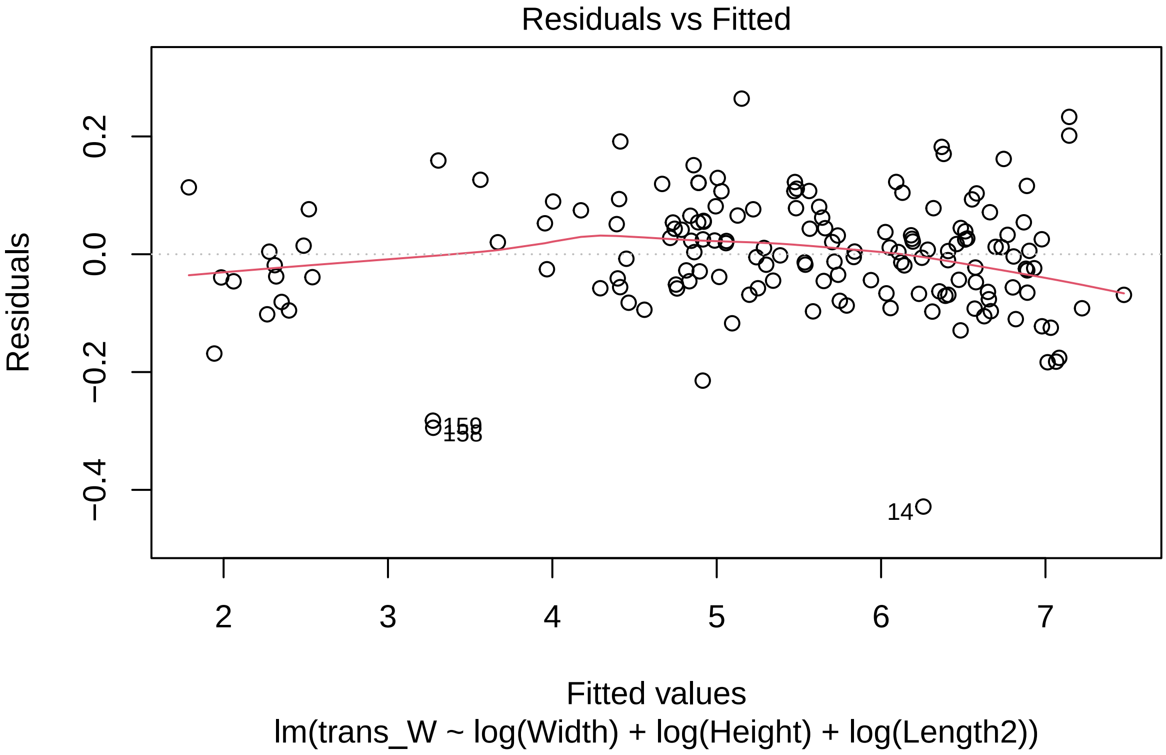
```
##
## Call:
## lm(formula = trans_W ~ log(Width) + log(Height) + log(Length2),
##     data = fish_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42835 -0.05739  0.00406  0.05504  0.26420
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.01044    0.12178  -16.51   <2e-16 ***
## log(Width)    0.90246    0.06251   14.44   <2e-16 ***
## log(Height)   0.61206    0.03251   18.83   <2e-16 ***
## log(Length2)  1.49774    0.05343   28.03   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09804 on 154 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9945
## F-statistic:  9525 on 3 and 154 DF,  p-value: < 2.2e-16
```

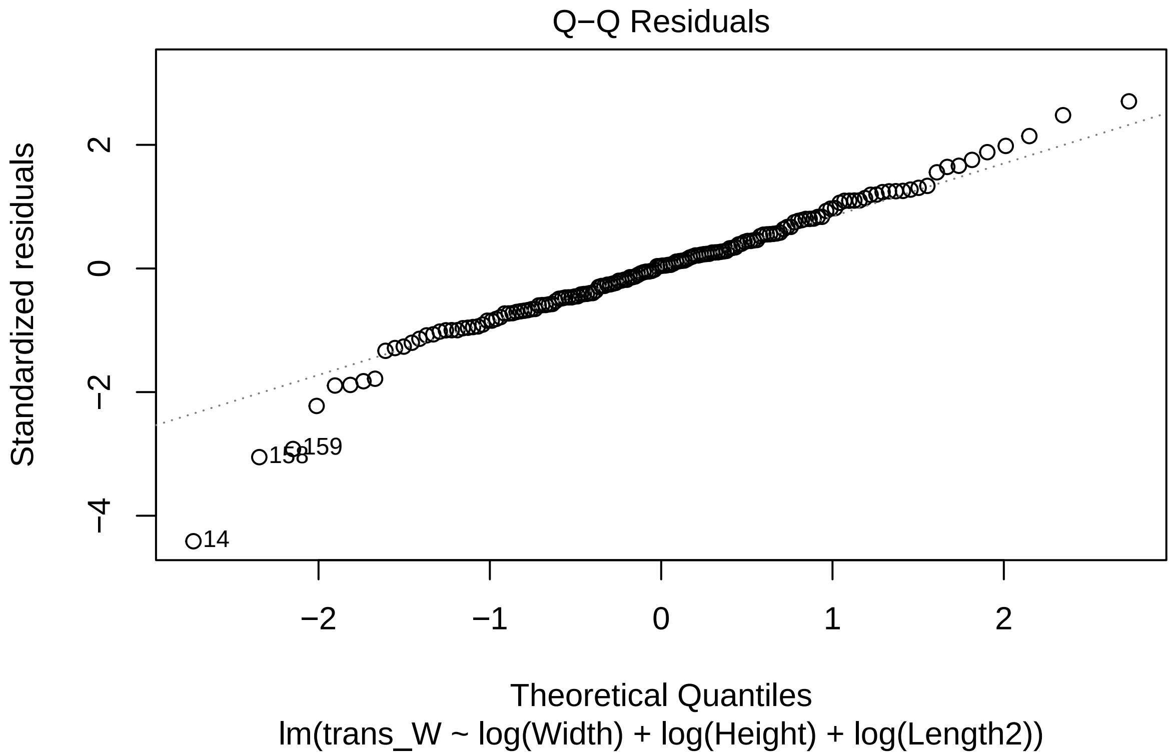
```
bp_res_aic = bptest(step_aic)
print(bp_res_aic)
```

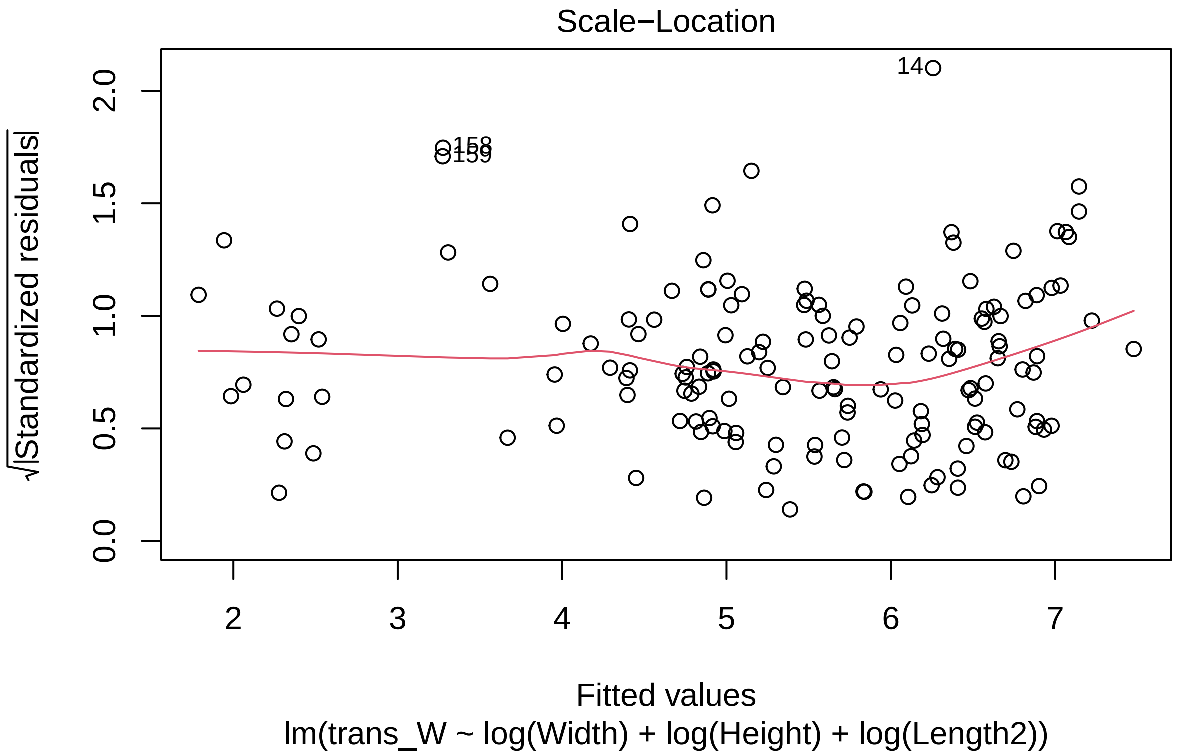
```
##
##  studentized Breusch-Pagan test
##
```

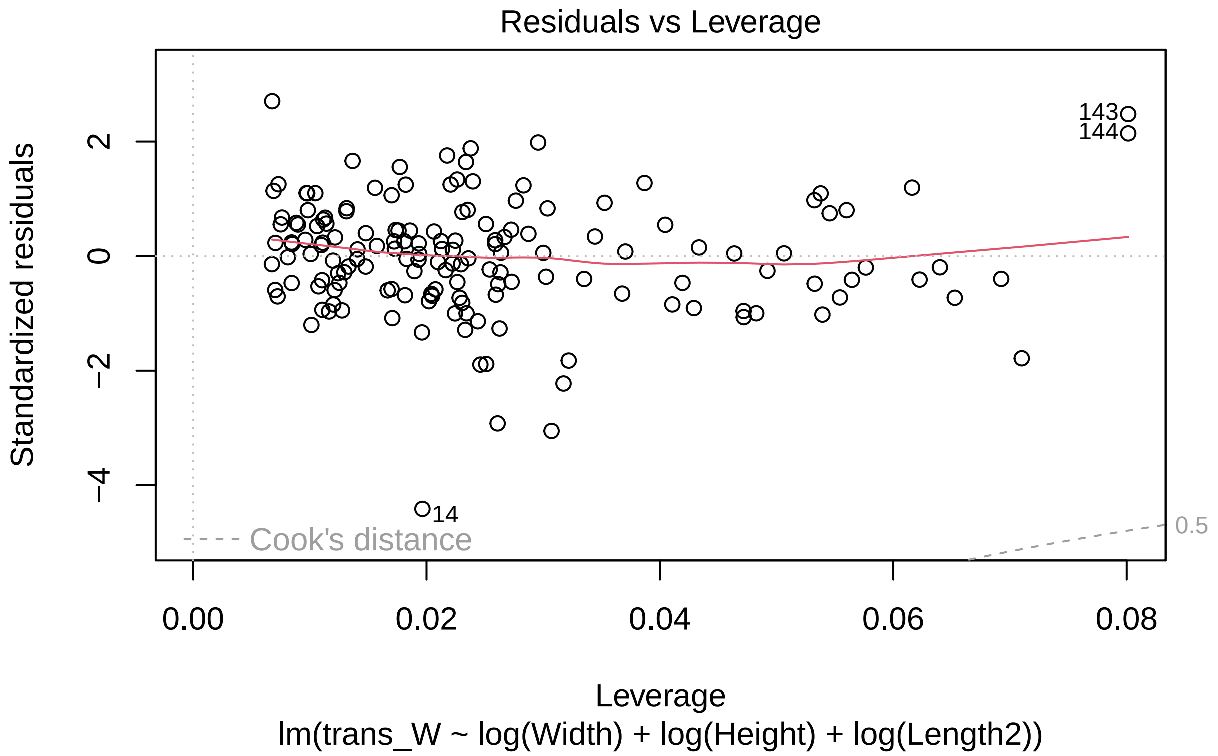
```
## data:  step_aic
## BP = 0.3688, df = 3, p-value = 0.9466
```

```
plot(step_aic)
```




Q−Q Residuals
Standardized residuals
2
0
−2
−4
158
159
14
−2
−1
0
1
2
Theoretical Quantiles
lm(trans_W ~ log(Width) + log(Height) + log(Length2))


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2
3
4
5
6
7
Fitted values
lm(trans_W ~ log(Width) + log(Height) + log(Length2))
14
158
159



```
bp_res_bic = bptest(step_bic)
print(bp_res_bic)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  step_bic
## BP = 0.3688, df = 3, p-value = 0.9466
```

```
anova(step_aic)
```

```
## Analysis of Variance Table
##
## Response: trans_W
##               Df  Sum Sq Mean Sq  F value    Pr(>F)
## log(Width)     1 265.168 265.168 27588.74 < 2.2e-16 ***
## log(Height)    1   1.918   1.918   199.57 < 2.2e-16 ***
## log(Length2)   1   7.552   7.552   785.71 < 2.2e-16 ***
## Residuals    154   1.480   0.010
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Part 3: ANCOVA and Species

In this part, I will use the output of part 2 and discuss the influence of Species, the only categorical variable. I start with the full model, where Species produces an additive change in log(Weight) and also changes the effect of X (log(Length2), log(Width), and log(Height)) on log(Weight). Then, I define the “zero” model as the model that doesn’t have any explanatory variables. After that, I define the additive model, where Species only changes the intercept, and the numeric model, where Species has no effect on log(Weight), as well as the categorical model, where X (log(Length2), log(Width), and log(Height)) has no effect on log(Weights). After that, I start to do a sequence of F-tests. I compare between the full model and the additive model, and find that the p-value of the F-test is 0.1507 (>0.05), so the additive model should be sufficient, when compared with the full model. The rest of the F-tests show that all the explanatory variables in the additive model are necessary. Therefore, I determine that the additive model is the final model to use for this project.

After that, I generate the diagnostic plots of the additive model, and also do a Breusch-Pagan test on the model. In its Residuals vs Fitted plot, the residual values are randomly scattered around zero. Its Q-Q plot shows that most of the points are very close to the dotted line, which means the residuals approximately follow normal distribution. The red lines in its Scale-Location and Residuals vs Leverage plot are also quite stable and do not show noticeable signs of curves, which means the error terms have constant variance, and the model is robust against outliers. The p-value of the studentized Breusch-Pagan test, 0.776, further confirms that the error terms of this model are homoscedastic. Therefore, this model satisfies the assumptions of linear models.

To test its performance, I use the model to predict the weight of fish given the information in the data set, and compare them with the actual weights. The root mean square error of the current model is around 52.1472 grams. This root mean square error is slightly higher than that of the full model(45.966 grams), but it has a stronger ability to handle new data. So I still choose this additive model to be the final model. The R^2 of this model is 0.9957, which means that this model can explain 99.57% of the total variation in the data set.

Answer to the Questions: 1. Is species needed at all? Yes, Species is needed in the final model, which is the additive model. In this model, Species only changes the intercept. 2.Does it enter additively or are interactions with other predictors needed? It enters additively in the final model. 3.Does each species need an entirely separate analysis and model? No, it doesn’t. The additive model is sufficient enough for this data set. As the full model that considers interactions is not our final choice, we don’t need to run an entirely separate analysis and model for each species.

```
full_model = lm(trans_W ~ Species * (log(Length2)+log(Width)+log(Height)), data = fish_trans)

model_add = lm(trans_W ~ Species + log(Length2)+log(Width)+log(Height), data = fish_trans)

anova(model_add, full_model)
```

```
## Analysis of Variance Table
##
## Model 1: trans_W ~ Species + log(Length2) + log(Width) + log(Height)
## Model 2: trans_W ~ Species * (log(Length2) + log(Width) + log(Height))
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    148 1.2004
## 2    130 1.0075 18   0.19287 1.3825 0.1507
```

```
model_num = lm(trans_W ~ log(Length2)+log(Width)+log(Height), data = fish_trans)

anova(model_num, model_add)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: trans_W ~ log(Length2) + log(Width) + log(Height)
## Model 2: trans_W ~ Species + log(Length2) + log(Width) + log(Height)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    154 1.4802
## 2    148 1.2004  6    0.2798 5.7497 2.107e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
model_cat = lm(trans_W ~ Species, data = fish_trans)

anova(model_cat, model_add)
```

```
## Analysis of Variance Table
##
## Model 1: trans_W ~ Species
## Model 2: trans_W ~ Species + log(Length2) + log(Width) + log(Height)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    151 88.177
## 2    148  1.200  3    86.977 3574.6 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
model_zero = lm(trans_W ~ 1, data = fish_trans)

anova(model_zero, model_num)
```

```
## Analysis of Variance Table
##
## Model 1: trans_W ~ 1
## Model 2: trans_W ~ log(Length2) + log(Width) + log(Height)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    157 276.12
## 2    154   1.48  3    274.64 9524.7 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
anova(model_zero, model_cat)
```

```
## Analysis of Variance Table
##
## Model 1: trans_W ~ 1
## Model 2: trans_W ~ Species
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1    157 276.118
## 2    151  88.177  6    187.94 53.64 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

```
summary(full_model)
```

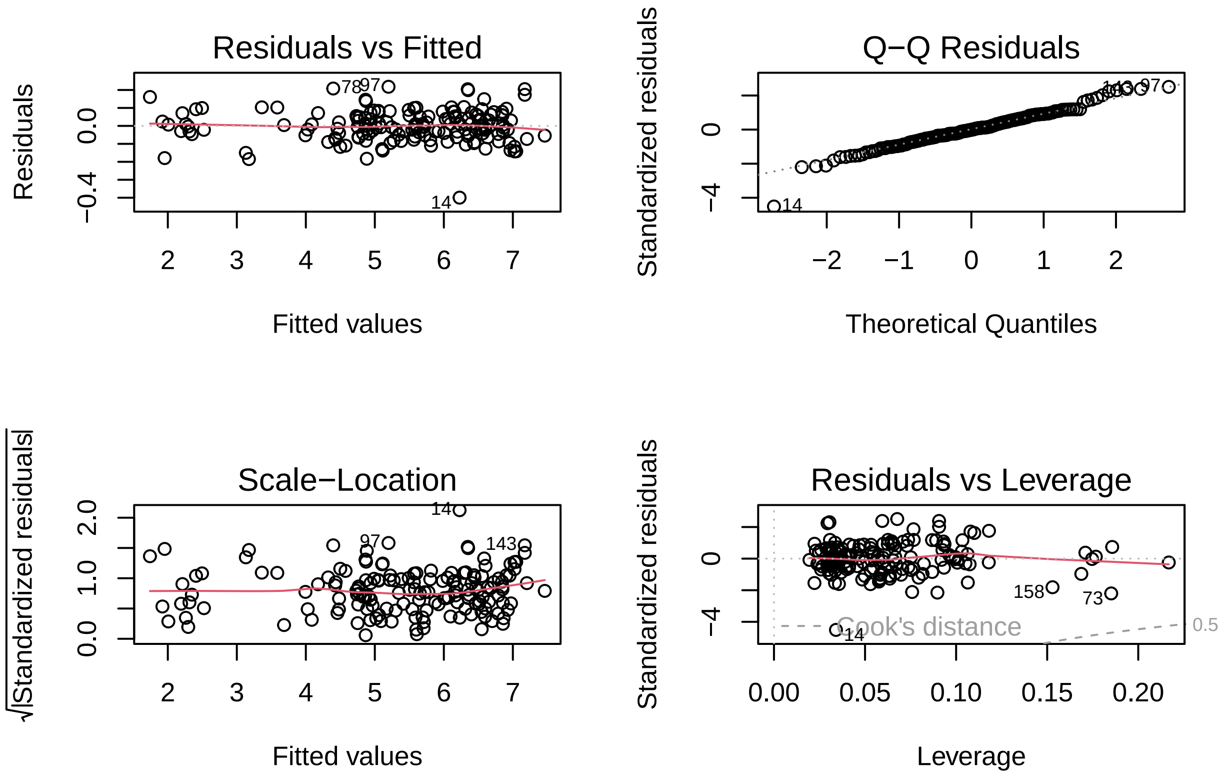
```
##
## Call:
## lm(formula = trans_W ~ Species * (log(Length2) + log(Width) +
##     log(Height)), data = fish_trans)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.37243 -0.04827 -0.00232  0.04811  0.21156
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1.54658    0.70119  -2.206  0.02917 *
## SpeciesParkki                  1.35359    1.85285   0.731  0.46637
## SpeciesPerch                  -0.76993    0.80477  -0.957  0.34049
## SpeciesPike                   -2.21135    1.05560  -2.095  0.03812 *
## SpeciesRoach                   0.93425    1.19836   0.780  0.43704
## SpeciesSmelt                  -0.22798    1.37280  -0.166  0.86836
## SpeciesWhitefish              -1.30099    1.52933  -0.851  0.39650
## log(Length2)                   0.95385    0.43901   2.173  0.03161 *
## log(Width)                     0.45619    0.30150   1.513  0.13268
## log(Height)                    1.40585    0.42353   3.319  0.00117 **
## SpeciesParkki:log(Length2)    -0.10265    1.10339  -0.093  0.92602
## SpeciesPerch:log(Length2)      0.66587    0.49556   1.344  0.18140
## SpeciesPike:log(Length2)       1.35979    0.58371   2.330  0.02137 *
## SpeciesRoach:log(Length2)     -0.09265    0.69217  -0.134  0.89372
## SpeciesSmelt:log(Length2)      0.45572    0.73827   0.617  0.53813
## SpeciesWhitefish:log(Length2)  0.32460    1.73200   0.187  0.85163
## SpeciesParkki:log(Width)       1.23280    0.83145   1.483  0.14057
## SpeciesPerch:log(Width)        0.10603    0.35267   0.301  0.76417
## SpeciesPike:log(Width)         0.22548    0.46570   0.484  0.62908
## SpeciesRoach:log(Width)        0.95281    0.52941   1.800  0.07422 .
## SpeciesSmelt:log(Width)       -0.09105    0.39013  -0.233  0.81582
## SpeciesWhitefish:log(Width)   -0.58903    1.28456  -0.459  0.64733
## SpeciesParkki:log(Height)     -1.12500    0.88510  -1.271  0.20598
## SpeciesPerch:log(Height)      -0.58325    0.47725  -1.222  0.22387
## SpeciesPike:log(Height)       -1.28184    0.62661  -2.046  0.04281 *
## SpeciesRoach:log(Height)      -0.82847    0.59340  -1.396  0.16506
## SpeciesSmelt:log(Height)      -0.70336    0.63444  -1.109  0.26963
## SpeciesWhitefish:log(Height)   0.69700    2.96641   0.235  0.81461
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.08803 on 130 degrees of freedom
## Multiple R-squared:  0.9964, Adjusted R-squared:  0.9956
## F-statistic:  1315 on 27 and 130 DF,  p-value: < 2.2e-16
```

```
summary(model_add)
```

```
##
## Call:
## lm(formula = trans_W ~ Species + log(Length2) + log(Width) +
##     log(Height), data = fish_trans)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -0.39946 -0.05151 -0.00078  0.05610  0.21781 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      -2.54077    0.24661 -10.303  < 2e-16 ***
## SpeciesParkki     0.08161    0.03454   2.363  0.01945 *  
## SpeciesPerch      0.10741    0.07837   1.371  0.17258    
## SpeciesPike       0.03375    0.14148   0.239  0.81180    
## SpeciesRoach      0.09416    0.06808   1.383  0.16870    
## SpeciesSmelt     -0.06505    0.11984  -0.543  0.58810    
## SpeciesWhitefish  0.18585    0.07007   2.652  0.00886 ** 
## log(Length2)      1.71536    0.15548  11.032  < 2e-16 ***
## log(Width)        0.54828    0.11579   4.735 5.08e-06 ***
## log(Height)       0.73451    0.15204   4.831 3.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.09006 on 148 degrees of freedom
## Multiple R-squared:  0.9957, Adjusted R-squared:  0.9954 
## F-statistic:  3766 on 9 and 148 DF,  p-value: < 2.2e-16
```

```r
par(mfrow = c(2,2))
plot(model_add)
```



```
par(mfrow = c(1,1))
bp_res_add = bptest(model_add)
print(bp_res_add)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  model_add
## BP = 5.6335, df = 9, p-value = 0.776
```

```
fish_ft = fish_c
lg_p = predict(model_add, fish_ft)
fish_ft $ pred_res = exp(lg_p)
fish_ft $ residual = fish_ft $ pred_res - fish_ft $Weight
head(fish_ft[c("Weight", "pred_res", "residual")], 22)
```

```
##    Weight pred_res    residual
## 1     242 261.3911   19.391065
## 2     290 305.5671   15.567068
## 3     340 322.7002  -17.299783
## 4     363 373.5823   10.582280
## 5     430 397.0903  -32.909669
## 6     450 431.7881  -18.211891
## 7     500 462.2902  -37.709794
## 8     390 405.8341   15.834097
## 9     450 444.6165   -5.383518
## 10    500 474.0179  -25.982141
## 11    475 490.5718   15.571751
## 12    500 477.7645  -22.235507
## 13    500 450.8866  -49.113445
## 14    340 506.9461  166.946124
## 15    600 540.1893  -59.810723
## 16    600 576.5453  -23.454679
## 17    700 573.6912 -126.308797
## 18    700 570.6191 -129.380855
## 19    610 601.3390   -8.660986
## 20    650 603.3986  -46.601384
## 21    575 629.6345   54.634450
## 22    685 642.9234  -42.076572
```

```
rmse = sqrt(mean(fish_ft$residual ˆ 2))
print(paste("The RMSE (root mean square error) in grams: ", rmse))
```

```
## [1] "The RMSE (root mean square error) in grams:  52.1472235342515"
```

```
fish_ft1 = fish_c
lg_p = predict(full_model, fish_ft1)
fish_ft1 $ pred_res = exp(lg_p)
fish_ft1 $ residual = fish_ft1 $ pred_res - fish_ft1 $Weight
head(fish_ft1[c("Weight", "pred_res", "residual")], 22)
```

```
##    Weight pred_res     residual
## 1     242 273.0357   31.0357148
## 2     290 325.9207   35.9207188
## 3     340 337.6214   -2.3786269
## 4     363 373.6678   10.6677604
## 5     430 386.0958  -43.9041837
## 6     450 439.3209  -10.6791375
## 7     500 480.6044  -19.3955908
## 8     390 392.4707    2.4707215
## 9     450 458.5211    8.5210752
## 10    500 484.3690  -15.6309745
## 11    475 497.1164   22.1164412
## 12    500 489.2343  -10.7656510
## 13    500 446.9712  -53.0287863
## 14    340 493.4265  153.4264885
## 15    600 550.9202  -49.0798162
## 16    600 596.4989   -3.5011077
## 17    700 567.9747 -132.0252509
## 18    700 567.7882 -132.2117754
## 19    610 610.5736    0.5736479
## 20    650 575.9479  -74.0520994
## 21    575 613.7418   38.7418174
## 22    685 652.7175  -32.2824594
```

```r
rmse1 = sqrt(mean(fish_ft1$residual ^ 2))
print(paste("The RMSE (root mean square error) in grams: ", rmse1))
```

```
## [1] "The RMSE (root mean square error) in grams:  45.9663977021086"
```

Part 4: Recommendations

Based on the output in part 3, my model can be written as the formula below:

$$\hat{W} = \exp\left(-2.541 + \beta_s + 1.715 \cdot \ln(Length2) + 0.548 \cdot \ln(Width) + 0.735 \cdot \ln(Height)\right)$$

, where s represents a specific species of fish. Here is the full list of β_s given s. $\beta_{Parkki} = 0.082$, $\beta_{Perch} = 0.107$,$\beta_{Pike} = 0.034$,$\beta_{Roach} = 0.094$,$\beta_{Smelt} = -0.065$, $\beta_{Whitefish} = 0.186$. The model is expected to predict the weight of a fish with a root mean square error of around 52.147 grams. To make things easier for everyone to predict the weight of a fish given its species and other tape measurements, I will implement a small computer program that takes the fish's Species, Length2, Height and Width as input, and prints out the prediction. All the calculations are made inside the program using the formula. This way, people can conveniently get the prediction in seconds, instead of plugging all the values in the formula and calculate the predicted weight.