Abstract

In this paper, we investigate the height of fish, and whether there is a relationship between the height and 6 variables. According to our research, we are going to build two models. One model also finds the interaction of category variable and non-category variables. we use ANOVA to choose one of these two models. After that, the one models we retained will be selected through the stepwise regression, transformation, including removing the outliers by studentized residuals method and robust regression to build our final model.

Introduction

Fish is a special animal that they cannot live without water. To learn more about the physical property, we assume whether there is a linear relationship between its height and the other variables that are its length in different degrees, weight, width, and species. According to the dataset we try to investigate. We focus on Bream, Parkki, Perch, Pike, Roach, Smelt, and White fish as sample of an analysis.

Data Description

Dataset:

There are 159 sample with 7 variables in this data, which is collected from the fish market.

Variables:

Species – species name of fish (‘Bream’, ‘Parkki’, ‘Perch’, ‘Pike’, ‘Roach’, ‘Smelt’, ‘White fish’)

Weight – weight of fish in Gram (g)

Length1 – vertical length in cm

Length2 – diagonal length in cm

Length3 – cross length in cm

Height – height in cm

Width – diagonal width in cm

New addition data point:

Species = Smelt, Weight = 8.0, Length1=10.2, Length2= 11.7

Length3= 11.8, Height = 1.9945, Width= 1.4152

Methods

According to the data, two kinds of full model we are going to build:

Model 1:

Height = β_0 + β_1*(Length1) + β_2*(Length2) + β_3*(Length3) + β_4*(Weight) + β_5*(Width) + β_6*(Species)

Model 2: (variable with interaction)

Height = β_0 + β_1*(Length1) + β_2*(Length2) + β_3*(Length3) + β_4*(Weight) + β_5*(Width) + β_6*(Species) + β_7*(Species) *(Length1) + β_8*(Species) *(Length2) + β_9*(Species) *(Length3) + β_{10}*(Species) *(Weight) + β_{11}*(Species) *(Width)

Where the variable ‘Species’ is a categorical variable

Due to this regression analysis, the models follow these five assumptions:

1. The relationship between the response variable (Height) and the regressors 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.

In two model, scatterplot matrices help us to determine whether the response variable is linear relationship with the other variable by looking. Then, we find out whether it is a significant evidence that there is a linear relationship between the response variable and each the regressors at a 0.05 significant level. Since both models are similar except for the interaction, the next step we will do is to detect if the interaction has a relationship with the response variable. We will choose the one model as our final model by ANOVA test. Regarding the multicollinearity, we try to center for each regressor which variance inflation factor (VIF) is larger than 10. After that, remake the models with the modified dataset, and using stepwise regression method to retain or reduce the regressor by the minimum AIC in order to find our best model. Then, we use fitted plot and normal Q-Q plot to determine both models are linear and normal. If not, we use the transformation method to improve the model to be more linear and normal. We use studentized residual and cook's distance, and check whether there are influence points and outliers. If we find the influence point and outliers, then remove outlier, but we may keep the influence points because that points may not severely interfere in change of slope. After we remake the model again, make a comparison to the same model with robust regression. We also observe the weight in robust regression to determine the problematic observations and consider if we need do remove it. Through these kinds of method, this model is confirmed as the final model for studying the response variable.

Result

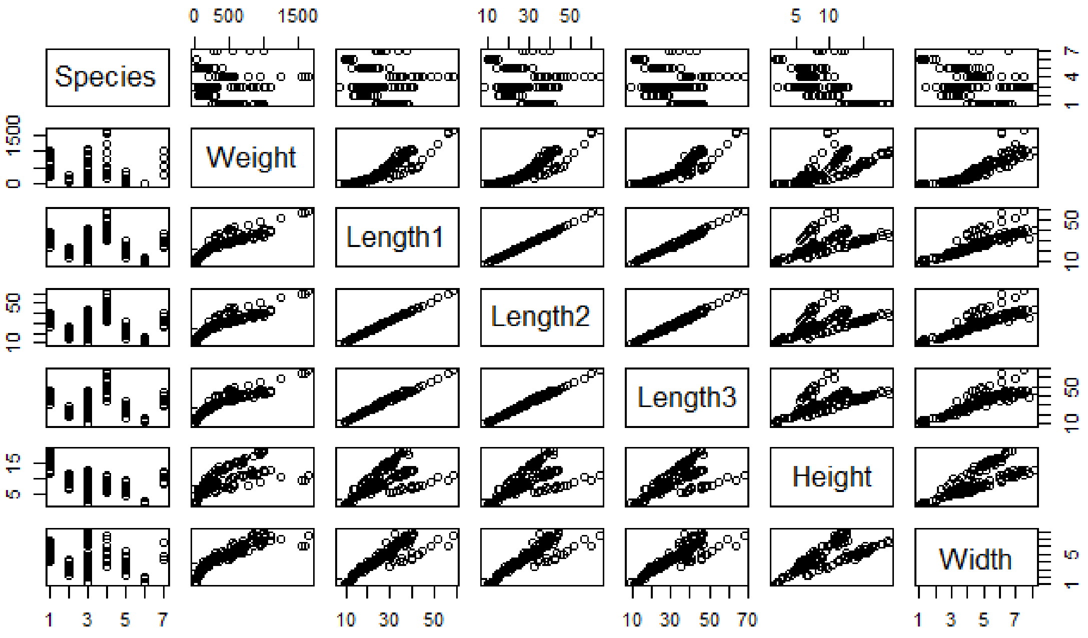
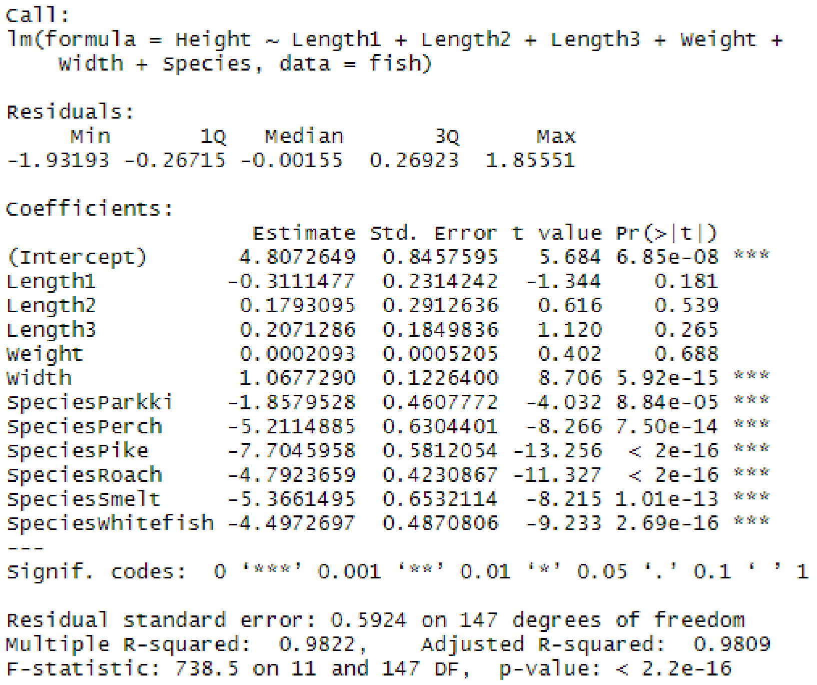


Figure 1: Scatterplot Matrices

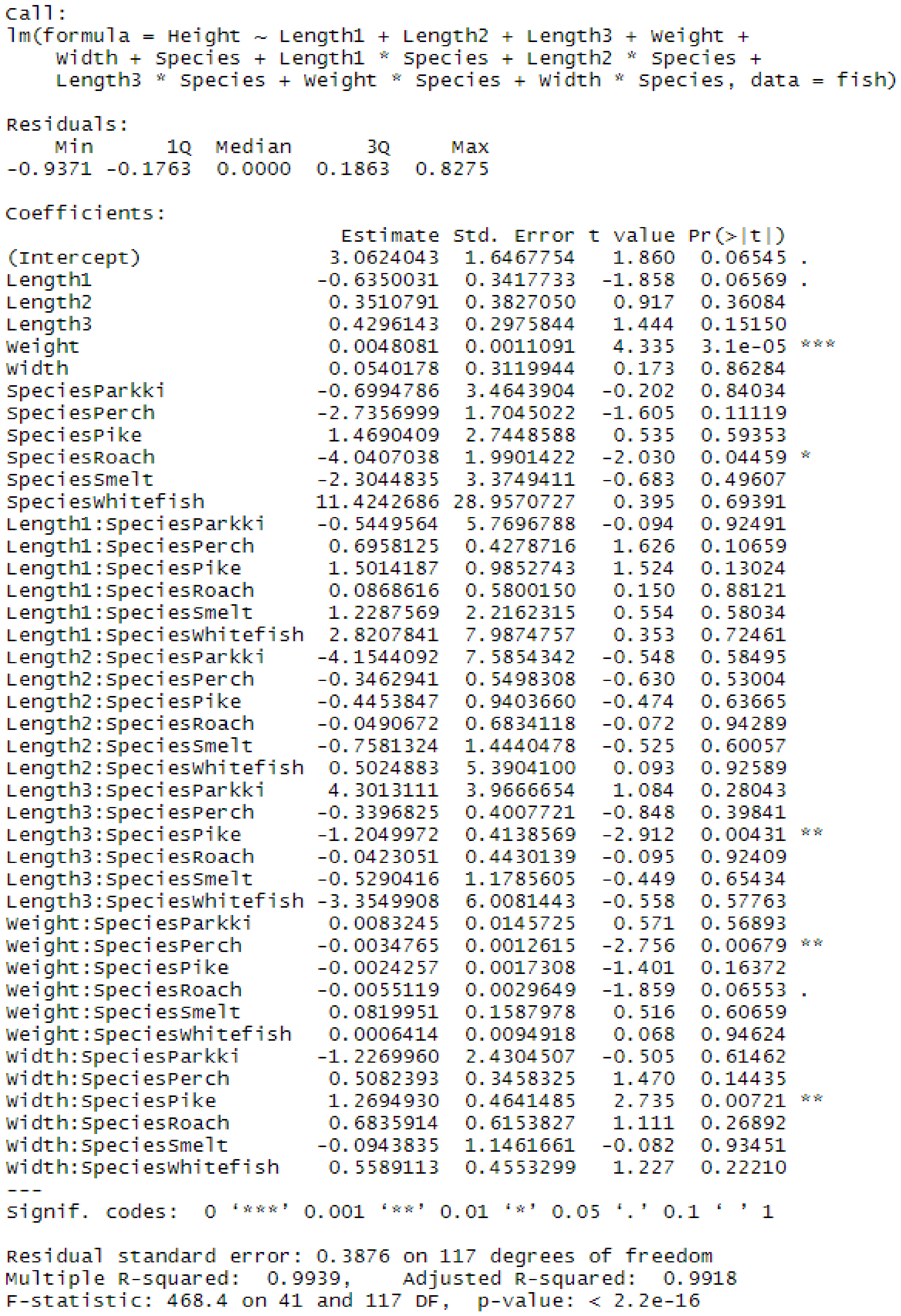
In figure 1, we find that the response variable (Height) has a linear relationship among the regressors (Length1, Length2, Length3, Weight, Width,). Since the regressor ‘Species’ is a category variable, we use analysis tools to figure it out soon.



```
Call:
lm(formula = Height ~ Length1 + Length2 + Length3 + Weight +
    Width + Species, data = fish)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.93193 -0.26715 -0.00155  0.26923  1.85551 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.8072649  0.8457595   5.684 6.85e-08 ***
Length1          -0.3111477  0.2314242  -1.344    0.181    
Length2           0.1793095  0.2912636   0.616    0.539    
Length3           0.2071286  0.1849836   1.120    0.265    
Weight            0.0002093  0.0005205   0.402    0.688    
Width             1.0677290  0.1226400   8.706 5.92e-15 ***
SpeciesParkki    -1.8579528  0.4607772  -4.032 8.84e-05 ***
SpeciesPerch     -5.2114885  0.6304401  -8.266 7.50e-14 ***
SpeciesPike      -7.7045958  0.5812054 -13.256  < 2e-16 ***
SpeciesRoach     -4.7923659  0.4230867 -11.327  < 2e-16 ***
SpeciesSmelt     -5.3661495  0.6532114  -8.215 1.01e-13 ***
SpeciesWhitefish -4.4972697  0.4870806  -9.233 2.69e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5924 on 147 degrees of freedom
Multiple R-squared:  0.9822,	Adjusted R-squared:  0.9809 
F-statistic: 738.5 on 11 and 147 DF,  p-value: < 2.2e-16
```

Figure 2: summary of model 1 (full model)

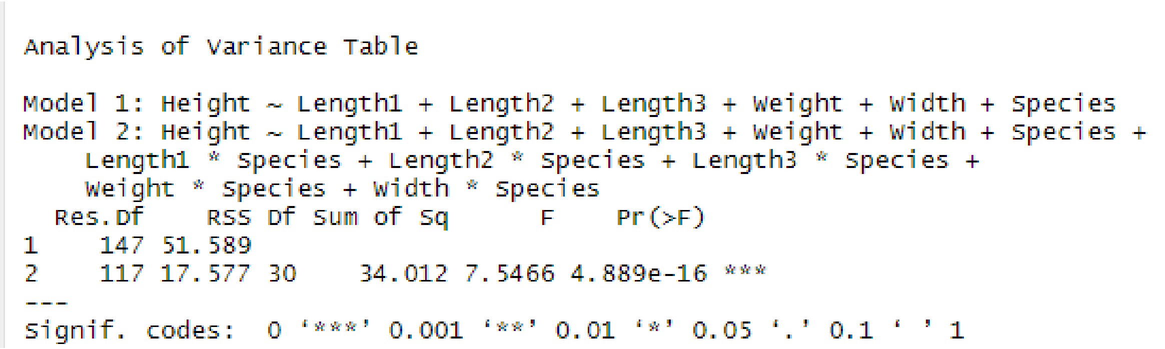
As we can see from figure 2, the p-value of F-statistic is < 2.2e-16, we conclude that at least one variable has a linear relationship with the response variable (Height). We also find that ‘Width’ regressors and all regressors in ‘species’ has a extremely smaller p-value than 0.001. This result reflects that height is affected by width and species.



```
Call:
lm(formula = Height ~ Length1 + Length2 + Length3 + Weight + 
    Width + Species + Length1 * Species + Length2 * Species + 
    Length3 * Species + Weight * Species + Width * Species, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9371 -0.1763  0.0000  0.1863  0.8275 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               3.0624043  1.6467754   1.860  0.06545 .  
Length1                  -0.6350031  0.3417733  -1.858  0.06569 .  
Length2                   0.3510791  0.3827050   0.917  0.36084    
Length3                   0.4296143  0.2975844   1.444  0.15150    
Weight                    0.0048081  0.0011091   4.335  3.1e-05 ***
Width                     0.0540178  0.3119944   0.173  0.86284    
SpeciesParkki            -0.6994786  3.4643904  -0.202  0.84034    
SpeciesPerch             -2.7356999  1.7045022  -1.605  0.11119    
SpeciesPike               1.4690409  2.7448588   0.535  0.59353    
SpeciesRoach             -4.0407038  1.9901422  -2.030  0.04459 *  
SpeciesSmelt             -2.3044835  3.3749411  -0.683  0.49607    
SpeciesWhitefish         11.4242686 28.9570727   0.395  0.69391    
Length1:SpeciesParkki    -0.5449564  5.7696788  -0.094  0.92491    
Length1:SpeciesPerch      0.6958125  0.4278716   1.626  0.10659    
Length1:SpeciesPike       1.5014187  0.9852743   1.524  0.13024    
Length1:SpeciesRoach      0.0868616  0.5800150   0.150  0.88121    
Length1:SpeciesSmelt      1.2287569  2.2162315   0.554  0.58034    
Length1:SpeciesWhitefish  2.8207841  7.9874757   0.353  0.72461    
Length2:SpeciesParkki    -4.1544092  7.5854342  -0.548  0.58495    
Length2:SpeciesPerch     -0.3462941  0.5498308  -0.630  0.53004    
Length2:SpeciesPike      -0.4453847  0.9403660  -0.474  0.63665    
Length2:SpeciesRoach     -0.0490672  0.6834118  -0.072  0.94289    
Length2:SpeciesSmelt     -0.7581324  1.4440478  -0.525  0.60057    
Length2:SpeciesWhitefish  0.5024883  5.3904100   0.093  0.92589    
Length3:SpeciesParkki     4.3013111  3.9666654   1.084  0.28043    
Length3:SpeciesPerch     -0.3396825  0.4007721  -0.848  0.39841    
Length3:SpeciesPike      -1.2049972  0.4138569  -2.912  0.00431 ** 
Length3:SpeciesRoach     -0.0423051  0.4430139  -0.095  0.92409    
Length3:SpeciesSmelt     -0.5290416  1.1785605  -0.449  0.65434    
Length3:SpeciesWhitefish -3.3549908  6.0081443  -0.558  0.57763    
Weight:SpeciesParkki      0.0083245  0.0145725   0.571  0.56893    
Weight:SpeciesPerch      -0.0034765  0.0012615  -2.756  0.00679 ** 
Weight:SpeciesPike       -0.0024257  0.0017308  -1.401  0.16372    
Weight:SpeciesRoach      -0.0055119  0.0029649  -1.859  0.06553 .  
Weight:SpeciesSmelt       0.0819951  0.1587978   0.516  0.60659    
Weight:SpeciesWhitefish   0.0006414  0.0094918   0.068  0.94624    
Width:SpeciesParkki      -1.2269960  2.4304507  -0.505  0.61462    
Width:SpeciesPerch        0.5082393  0.3458325   1.470  0.14435    
Width:SpeciesPike         1.2694930  0.4641485   2.735  0.00721 ** 
Width:SpeciesRoach        0.6835914  0.6153827   1.111  0.26892    
Width:SpeciesSmelt       -0.0943835  1.1461661  -0.082  0.93451    
Width:SpeciesWhitefish    0.5589113  0.4553299   1.227  0.22210    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3876 on 117 degrees of freedom
Multiple R-squared:  0.9939,	Adjusted R-squared:  0.9918 
F-statistic: 468.4 on 41 and 117 DF,  p-value: < 2.2e-16
```

Figure 3: summary of model 2 (interaction full model)

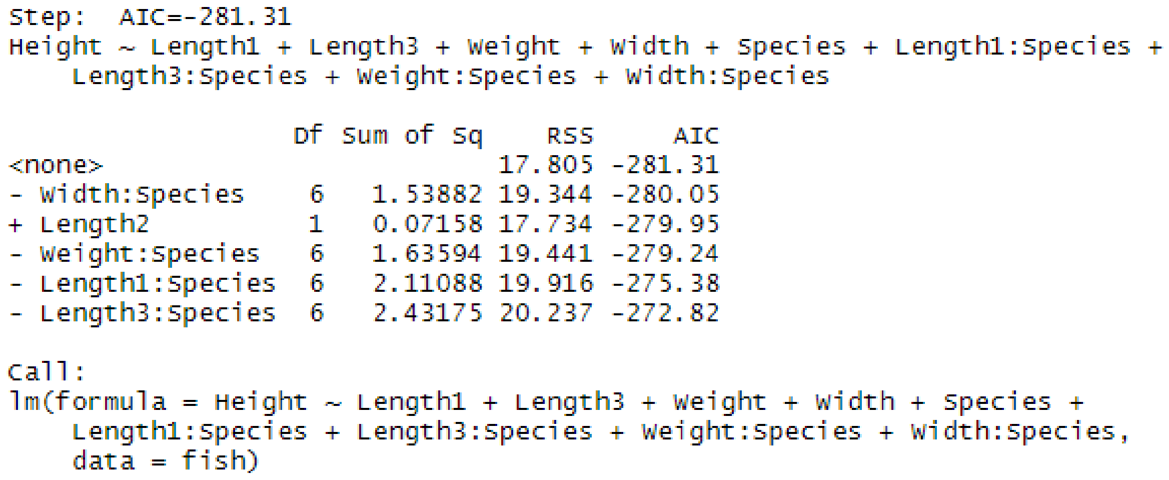
In figure 3, the p-value of F-statistic is < 2.2e-16, we conclude that at least one variable has a linear relationship with the response variable (Height) too. We also find that 'Weight', 'SpeciesRoach', 'Length3:SpeciesPike', 'Weight:SpeciesPerch', 'Width:SpeciesPike' has small p-value that is less then 0.05 significant level. Hence, we conclude that heigh is affected by weight, roach, interaction of length3 and pike, interaction of weight and perch, and interaction of width and pike.



```
Analysis of Variance Table

Model 1: Height ~ Length1 + Length2 + Length3 + Weight + Width + Species
Model 2: Height ~ Length1 + Length2 + Length3 + Weight + Width + Species +
    Length1 * Species + Length2 * Species + Length3 * Species +
    Weight * Species + Width * Species
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    147 51.589
2    117 17.577 30    34.012 7.5466 4.889e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Anova test for two model

Since the p-value is extremely small, we conclude that at least one variable which model 1 does not contain has a linear relationship with the response variable. Therefore, we will choose the model 2 to continue the rest of analysis.



```
Step:  AIC=-281.31
Height ~ Length1 + Length3 + Weight + Width + Species + Length1:Species +
    Length3:Species + Weight:Species + Width:Species

                  Df Sum of Sq    RSS     AIC
<none>                         17.805 -281.31
- Width:Species    6   1.53882 19.344 -280.05
+ Length2          1   0.07158 17.734 -279.95
- Weight:Species   6   1.63594 19.441 -279.24
- Length1:Species  6   2.11088 19.916 -275.38
- Length3:Species  6   2.43175 20.237 -272.82

Call:
lm(formula = Height ~ Length1 + Length3 + Weight + Width + Species +
    Length1:Species + Length3:Species + Weight:Species + Width:Species,
    data = fish)
```

Figure 5: stepwise regression

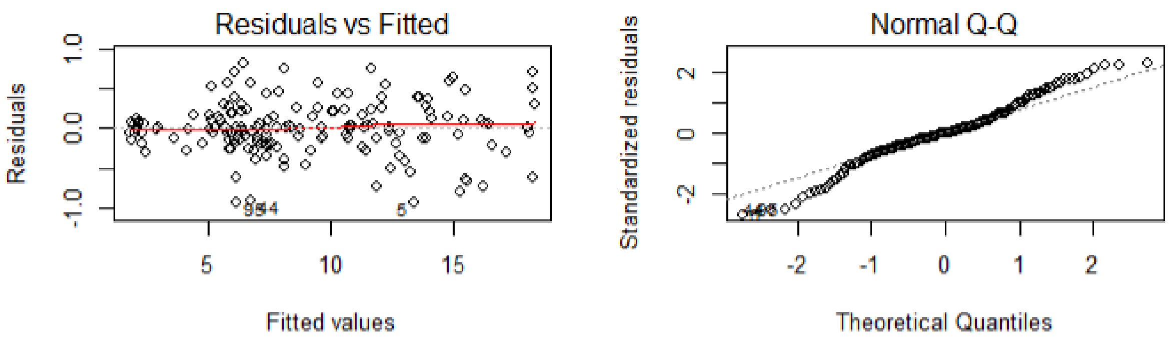


Figure 6: Fitted value plot and Normal Q-Q plot

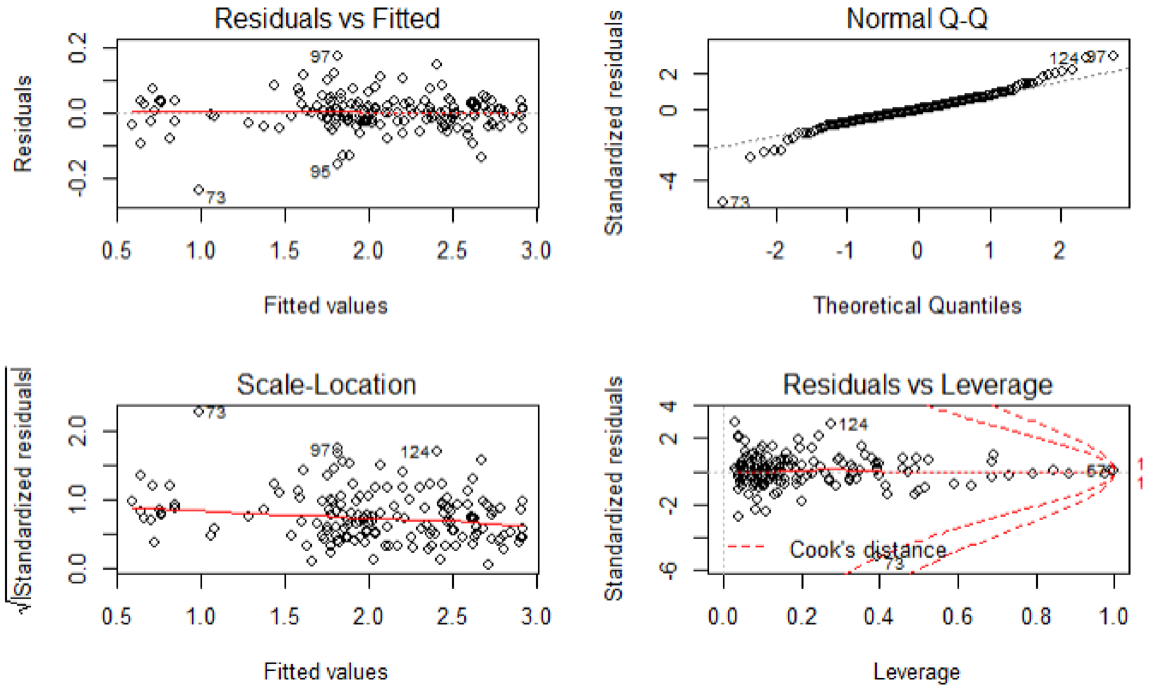


Figure 7: log-transformation for model 2

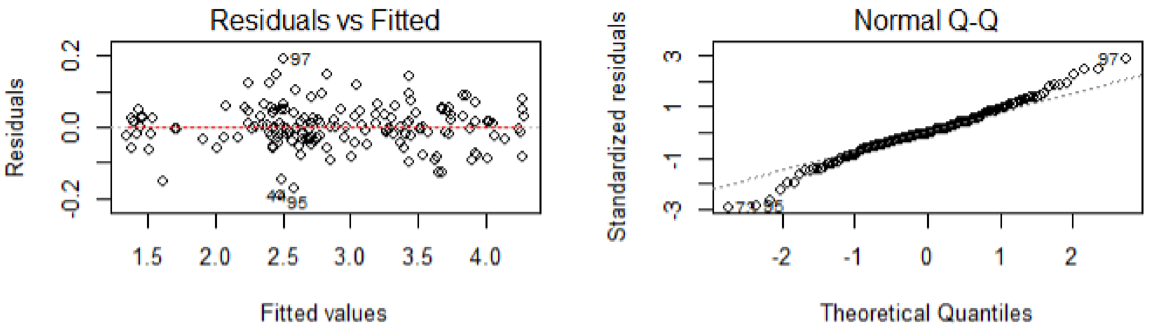


Figure 8: square root transformation for model 2

In figure 5, our variable selection is stepwise regression method. the minimum AIC we get is -281.31. In figure 6, we find that the model is linear, but it seems like non-normal. Hence, we will use transformation to try to make the model look better. Here are two kinds of transformation that I apply in figure 7 & 8. Compare log-transformation model to square root transformation model, we find that both models are linear. However, the model with log-transformation is more normally than the model with square root transformation. Therefore, we choose the log-transformation model as our part of final model.

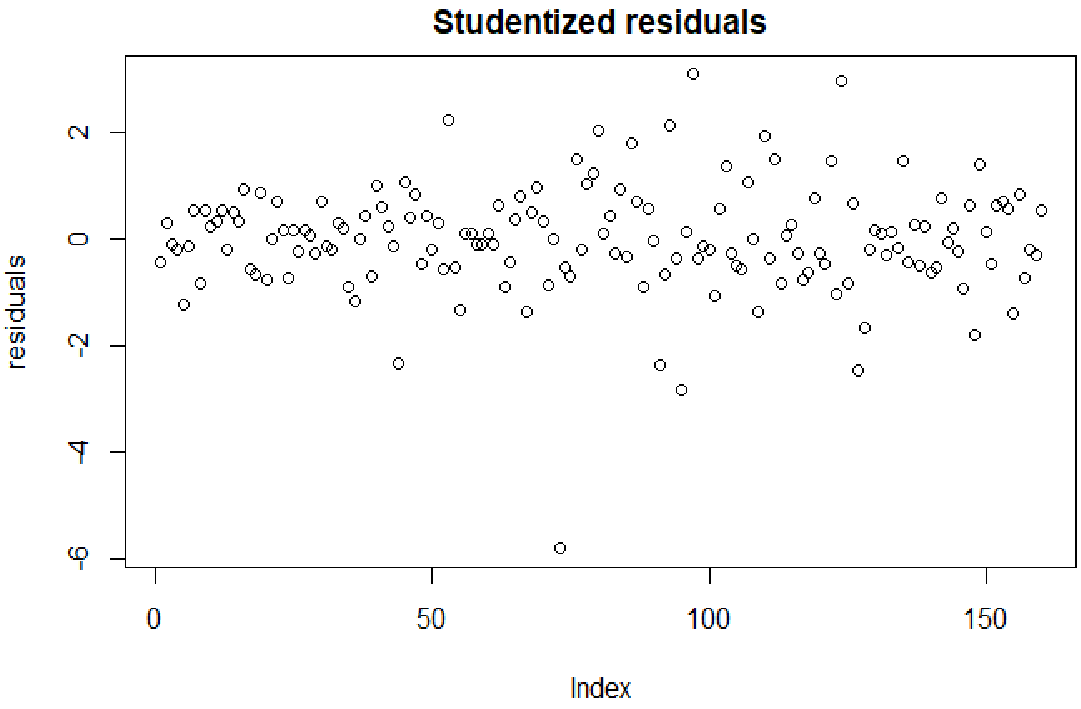


Figure 9: studentized residuals plot

In order to make the linear regression accurately, the next step we are going to do is to detect the outliers and remove them. In figure 9, we find that the residuals are distributed around zero. The mean of residual we acquire almost zero. To detect the outlier, we are considered the point is outlier which the absolute value of residual is larger than 3. After we remove those outliers, we repeat the detection until we cannot detect the outlier anymore. The result of the outliers we detect are 73, 97, 95.

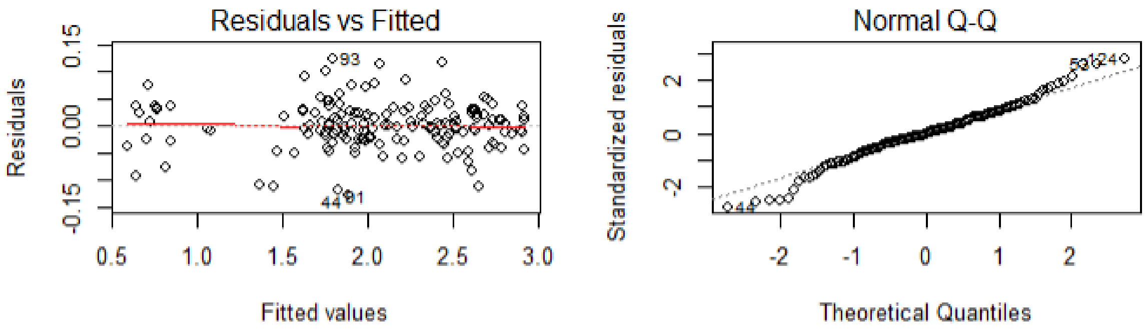


Figure 10: square root transformation for model 2 without some outlier

Since we remove some outlier that we detect, when we diagnostic the model, we still find that the model is linear but still non-normal. Therefore, we use robust regression method to analyze whether any problematic observation exists. According to robust regression model that we build, we want to find out those small weight so that a threshold that we set is 0.4. If the value of weight is less than 0.4, then this point regards as a problematic observation. As a result, observation 44, 53, 74, 75, 91, 93, 124, 127 are problematic observations. To consider making the model normal, we will remove these problematic observations.

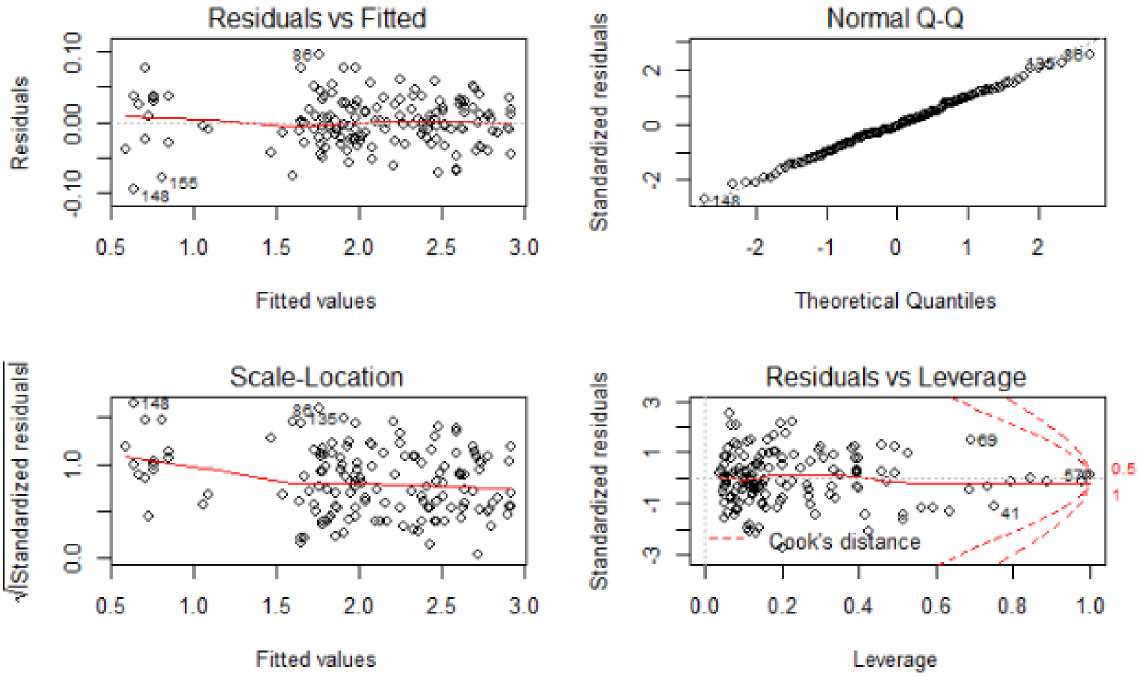
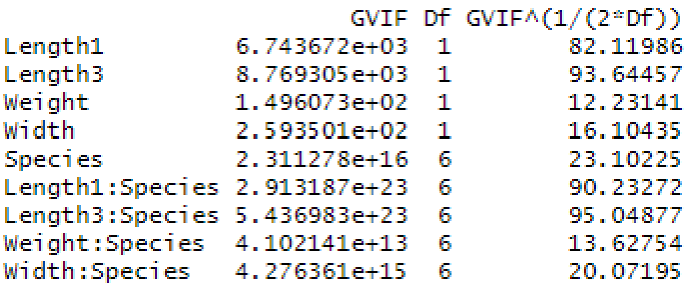


Figure 11: diagnostic plot for final modal of model 2

According to the diagnostic plot in figure 11, the model is linear and normal finally. However, in figure 12, we find that there is a serious multicollinearity for all variables.

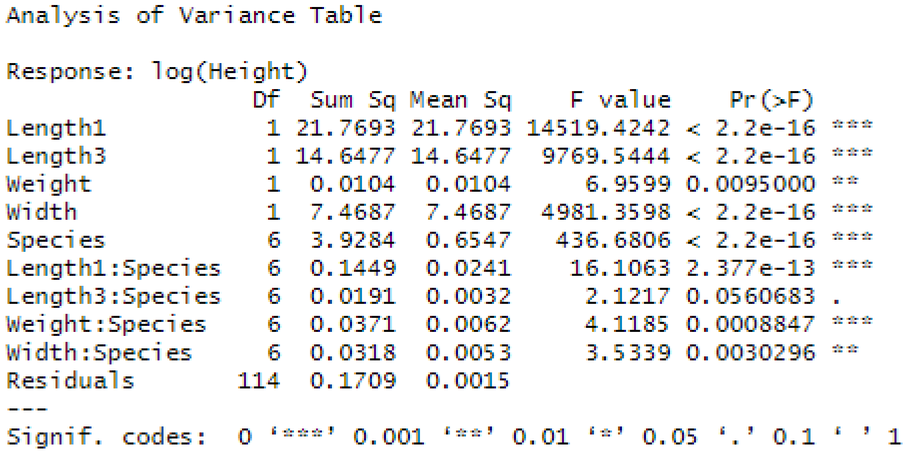


```
                          GVIF Df GVIF^(1/(2*Df))
Length1        6.743672e+03  1        82.11986
Length3        8.769305e+03  1        93.64457
Weight         1.496073e+02  1        12.23141
Width          2.593501e+02  1        16.10435
Species        2.311278e+16  6        23.10225
Length1:Species 2.913187e+23  6        90.23272
Length3:Species 5.436983e+23  6        95.04877
Weight:Species  4.102141e+13  6        13.62754
Width:Species   4.276361e+15  6        20.07195
```

Figure 12: variance inflation factor (VIF)

Conclusion

Based on the ANOVA test of final model (model 2), we conclude that height has a linear relationship among weight of fish in gram (Weight) ,vertical length in cm (Length1), cross length in cm (Length3), species (‘Bream’, ‘Parkki’, ‘Perch’, ‘Pike’, ‘Roach’, ‘Smelt’, ‘White fish’), diagonal width in cm (Width), the interaction of vertical length in cm and species (Length1: Species), the interaction of weight of fish in gram and Species (Weight: Species), and interaction of diagonal width in cm and species (Width:Species).



```
Analysis of Variance Table

Response: log(Height)
                 Df  Sum Sq Mean Sq    F value    Pr(>F)    
Length1           1 21.7693 21.7693 14519.4242 < 2.2e-16 ***
Length3           1 14.6477 14.6477  9769.5444 < 2.2e-16 ***
Weight            1  0.0104  0.0104     6.9599 0.0095000 ** 
Width             1  7.4687  7.4687  4981.3598 < 2.2e-16 ***
Species           6  3.9284  0.6547   436.6806 < 2.2e-16 ***
Length1:Species   6  0.1449  0.0241    16.1063 2.377e-13 ***
Length3:Species   6  0.0191  0.0032     2.1217 0.0560683 .  
Weight:Species    6  0.0371  0.0062     4.1185 0.0008847 ***
Width:Species     6  0.0318  0.0053     3.5339 0.0030296 ** 
Residuals       114  0.1709  0.0015                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

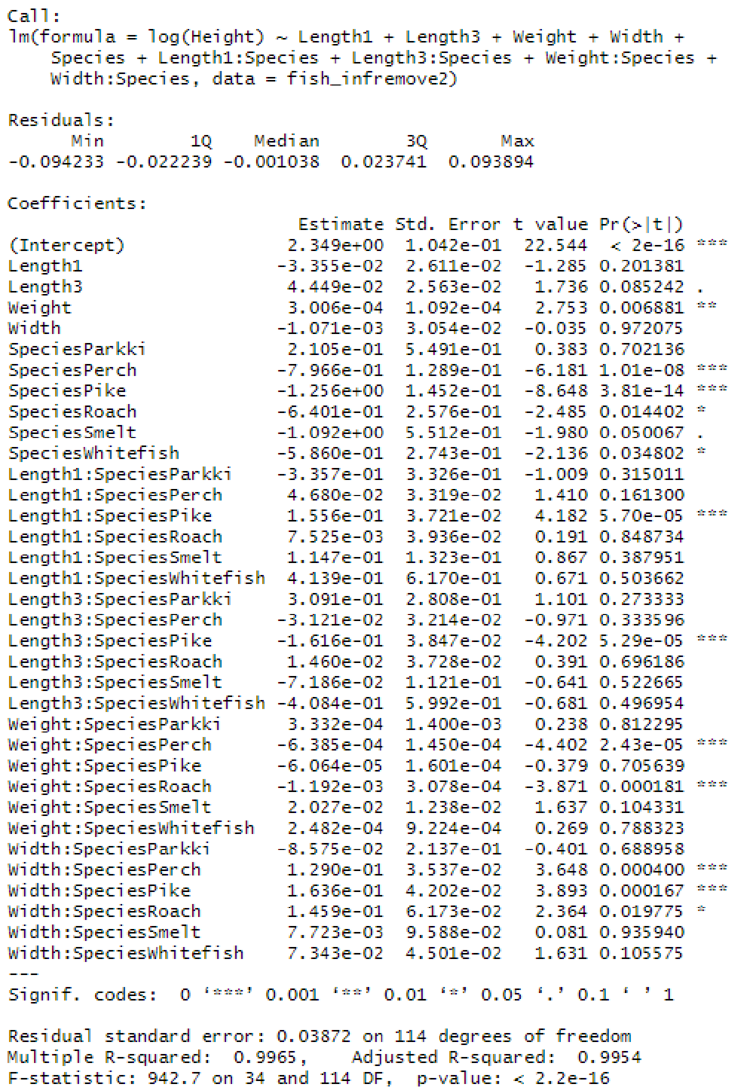
ANOVA test of Final model

According to the summary of final model, our estimator for the linear regression is

ln(Height) = 2.349 + 0.0003006*(Weight) – 0.7966*(Perch) – 1.256*(Pike) – 0.6401*(Roach) – 0.5860*(Whitefish) + 0.1556*(Pike)*(Length1) – 0.1616*(Pike)*(Length3)

– 0.0006385*(Perch)*(Weight) – 0.001192*(Roach)*(Weight) + 0.129 *(Width)*(Perch)

+ 0.1636*(Width)*(Pike) + 0.1459 *(Width)*(Roach)



```
Call:
lm(formula = log(Height) ~ Length1 + Length3 + Weight + Width + 
    Species + Length1:Species + Length3:Species + Weight:Species + 
    Width:Species, data = fish_infremove2)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.094233 -0.022239 -0.001038  0.023741  0.093894 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               2.349e+00  1.042e-01  22.544  < 2e-16 ***
Length1                  -3.355e-02  2.611e-02  -1.285 0.201381    
Length3                   4.449e-02  2.563e-02   1.736 0.085242 .  
Weight                    3.006e-04  1.092e-04   2.753 0.006881 ** 
Width                    -1.071e-03  3.054e-02  -0.035 0.972075    
SpeciesParkki             2.105e-01  5.491e-01   0.383 0.702136    
SpeciesPerch             -7.966e-01  1.289e-01  -6.181 1.01e-08 ***
SpeciesPike              -1.256e+00  1.452e-01  -8.648 3.81e-14 ***
SpeciesRoach             -6.401e-01  2.576e-01  -2.485 0.014402 *  
SpeciesSmelt             -1.092e+00  5.512e-01  -1.980 0.050067 .  
SpeciesWhitefish         -5.860e-01  2.743e-01  -2.136 0.034802 *  
Length1:SpeciesParkki    -3.357e-01  3.326e-01  -1.009 0.315011    
Length1:SpeciesPerch      4.680e-02  3.319e-02   1.410 0.161300    
Length1:SpeciesPike       1.556e-01  3.721e-02   4.182 5.70e-05 ***
Length1:SpeciesRoach      7.525e-03  3.936e-02   0.191 0.848734    
Length1:SpeciesSmelt      1.147e-01  1.323e-01   0.867 0.387951    
Length1:SpeciesWhitefish  4.139e-01  6.170e-01   0.671 0.503662    
Length3:SpeciesParkki     3.091e-01  2.808e-01   1.101 0.273333    
Length3:SpeciesPerch     -3.121e-02  3.214e-02  -0.971 0.333596    
Length3:SpeciesPike      -1.616e-01  3.847e-02  -4.202 5.29e-05 ***
Length3:SpeciesRoach      1.460e-02  3.728e-02   0.391 0.696186    
Length3:SpeciesSmelt     -7.186e-02  1.121e-01  -0.641 0.522665    
Length3:SpeciesWhitefish -4.084e-01  5.992e-01  -0.681 0.496954    
Weight:SpeciesParkki      3.332e-04  1.400e-03   0.238 0.812295    
Weight:SpeciesPerch      -6.385e-04  1.450e-04  -4.402 2.43e-05 ***
Weight:SpeciesPike       -6.064e-05  1.601e-04  -0.379 0.705639    
Weight:SpeciesRoach      -1.192e-03  3.078e-04  -3.871 0.000181 ***
Weight:SpeciesSmelt       2.027e-02  1.238e-02   1.637 0.104331    
Weight:SpeciesWhitefish   2.482e-04  9.224e-04   0.269 0.788323    
Width:SpeciesParkki      -8.575e-02  2.137e-01  -0.401 0.688958    
Width:SpeciesPerch        1.290e-01  3.537e-02   3.648 0.000400 ***
Width:SpeciesPike         1.636e-01  4.202e-02   3.893 0.000167 ***
Width:SpeciesRoach        1.459e-01  6.173e-02   2.364 0.019775 *  
Width:SpeciesSmelt        7.723e-03  9.588e-02   0.081 0.935940    
Width:SpeciesWhitefish    7.343e-02  4.501e-02   1.631 0.105575    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03872 on 114 degrees of freedom
Multiple R-squared:  0.9965,	Adjusted R-squared:  0.9954 
F-statistic: 942.7 on 34 and 114 DF,  p-value: < 2.2e-16
```

According to the formula of linear regression, we find there is a positive relation between height and weight for most of fish expect perch and pike. The height of Bream and smelt are affected by their weight. Perch, pike and roach have a positive relation between diagonal width and height. Only the height of pike has negative effect in cross length, but positive in vertical length.

Appendix

Summary of model 1:

```
> summary(m1)

Call:
lm(formula = Height ~ Length1 + Length2 + Length3 + Weight + 
    Width + Species, data = fish)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.93143 -0.25742 -0.01263  0.27204  1.86872 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)       4.7537079  0.8364704   5.683 6.81e-08 ***
Length1          -0.2783956  0.2209379  -1.260    0.210    
Length2           0.1224601  0.2663392   0.460    0.646    
Length3           0.2323394  0.1771879   1.311    0.192    
Weight            0.0002154  0.0005190   0.415    0.679    
Width             1.0626602  0.1218860   8.718 5.30e-15 ***
SpeciesParkki    -1.8203186  0.4531272  -4.017 9.34e-05 ***
SpeciesPerch     -5.1389241  0.6111284  -8.409 3.20e-14 ***
SpeciesPike      -7.6633827  0.5735744 -13.361  < 2e-16 ***
SpeciesRoach     -4.7605520  0.4169724 -11.417  < 2e-16 ***
SpeciesSmelt     -5.3437497  0.6499237  -8.222 9.35e-14 ***
SpeciesWhitefish -4.4472738  0.4749903  -9.363  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5909 on 148 degrees of freedom
Multiple R-squared:  0.9825,	Adjusted R-squared:  0.9812 
F-statistic: 754.9 on 11 and 148 DF,  p-value: < 2.2e-16
```

ANOVA test of model 1

```
> anova(m1)
Analysis of Variance Table

Response: Height
           Df  Sum Sq Mean Sq  F value    Pr(>F)    
Length1     1 1176.62 1176.62 3370.015 < 2.2e-16 ***
Length2     1  685.32  685.32 1962.862 < 2.2e-16 ***
Length3     1  600.33  600.33 1719.442 < 2.2e-16 ***
Weight      1   98.69   98.69  282.651 < 2.2e-16 ***
Width       1  203.69  203.69  583.385 < 2.2e-16 ***
Species     6  134.76   22.46   64.328 < 2.2e-16 ***
Residuals 148   51.67    0.35                       
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary of model 2

```
> summary(m2)

Call:
lm(formula = Height ~ Length1 + Length2 + Length3 + Weight + 
    Width + Species + Length1 * Species + Length2 * Species + 
    Length3 * Species + Weight * Species + Width * Species, data = fish)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9371 -0.1739  0.0000  0.1836  0.8275 
```

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               3.0624043  1.6405010   1.867  0.06442 .  
Length1                  -0.6350031  0.3404711  -1.865  0.06466 .  
Length2                   0.3510791  0.3812468   0.921  0.35900    
Length3                   0.4296143  0.2964506   1.449  0.14994    
Weight                    0.0048081  0.0011048   4.352 2.89e-05 ***
Width                     0.0540178  0.3108057   0.174  0.86232    
SpeciesParkki            -0.6994786  3.4511908  -0.203  0.83974    
SpeciesPerch             -2.7356999  1.6980080  -1.611  0.10982    
SpeciesPike               1.4690409  2.7344007   0.537  0.59211    
SpeciesRoach             -4.0407038  1.9825596  -2.038  0.04377 *  
SpeciesSmelt             -2.4262802  3.3406615  -0.726  0.46910    
SpeciesWhitefish         11.4242686 28.8467439   0.396  0.69280    
Length1:SpeciesParkki    -0.5449564  5.7476959  -0.095  0.92462    
Length1:SpeciesPerch      0.6958125  0.4262414   1.632  0.10525    
Length1:SpeciesPike       1.5014187  0.9815203   1.530  0.12877    
Length1:SpeciesRoach      0.0868616  0.5778051   0.150  0.88076    
Length1:SpeciesSmelt      0.7132029  1.5171580   0.470  0.63916    
Length1:SpeciesWhitefish  2.8207841  7.9570427   0.355  0.72360    
Length2:SpeciesParkki    -4.1544092  7.5565331  -0.550  0.58351    
Length2:SpeciesPerch     -0.3462941  0.5477359  -0.632  0.52846    
Length2:SpeciesPike      -0.4453847  0.9367831  -0.475  0.63535    
Length2:SpeciesRoach     -0.0490672  0.6808080  -0.072  0.94267    
Length2:SpeciesSmelt     -0.3379158  0.6002762  -0.563  0.57455    
Length2:SpeciesWhitefish  0.5024883  5.3698721   0.094  0.92561    
Length3:SpeciesParkki     4.3013111  3.9515521   1.089  0.27859    
Length3:SpeciesPerch     -0.3396825  0.3992452  -0.851  0.39660    
Length3:SpeciesPike      -1.2049972  0.4122801  -2.923  0.00416 ** 
Length3:SpeciesRoach     -0.0423051  0.4413260  -0.096  0.92380    
Length3:SpeciesSmelt     -0.4401571  1.1410410  -0.386  0.70038    
Length3:SpeciesWhitefish -3.3549908  5.9852528  -0.561  0.57617    
Weight:SpeciesParkki      0.0083245  0.0145170   0.573  0.56744    
Weight:SpeciesPerch      -0.0034765  0.0012567  -2.766  0.00658 ** 
Weight:SpeciesPike       -0.0024257  0.0017242  -1.407  0.16211    
Weight:SpeciesRoach      -0.0055119  0.0029536  -1.866  0.06450 .  
Weight:SpeciesSmelt       0.0516065  0.1268345   0.407  0.68483    
Weight:SpeciesWhitefish   0.0006414  0.0094557   0.068  0.94603    
Width:SpeciesParkki      -1.2269960  2.4211905  -0.507  0.61326    
Width:SpeciesPerch        0.5082393  0.3445149   1.475  0.14281    
Width:SpeciesPike         1.2694930  0.4623800   2.746  0.00699 ** 
Width:SpeciesRoach        0.6835914  0.6130380   1.115  0.26708    
Width:SpeciesSmelt       -0.0203541  1.1183303  -0.018  0.98551    
Width:SpeciesWhitefish    0.5589113  0.4535951   1.232  0.22033    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3861 on 118 degrees of freedom
Multiple R-squared:  0.994,	Adjusted R-squared:  0.992 
F-statistic: 479.9 on 41 and 118 DF,  p-value: < 2.2e-16
```

ANOVA test of model 2

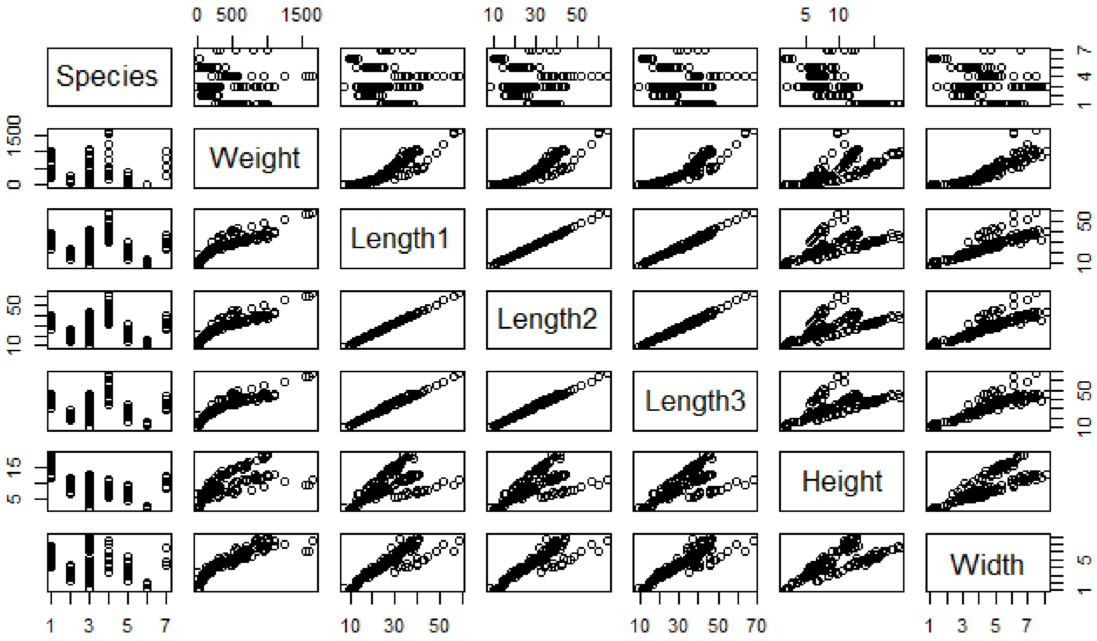
```
> anova(m2)
Analysis of Variance Table

Response: Height
                 Df  Sum Sq Mean Sq   F value Pr(>F)    
Length1           1 1176.62 1176.62 7892.0267 <2e-16 ***
Length2           1  685.32  685.32 4596.7029 <2e-16 ***
Length3           1  600.33  600.33 4026.6530 <2e-16 ***
Weight            1   98.69   98.69  661.9237 <2e-16 ***
Width             1  203.69  203.69 1366.1938 <2e-16 ***
Species           6  134.76   22.46  150.6455 <2e-16 ***
```

```
Length1:Species   6   29.06    4.84   32.4852 <2e-16 ***
Length2:Species   6    1.06    0.18    1.1852 0.3189
Length3:Species   6    1.55    0.26    1.7275 0.1205
Weight:Species    6    1.15    0.19    1.2865 0.2687
Width:Species     6    1.27    0.21    1.4142 0.2148
Residuals       118   17.59    0.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

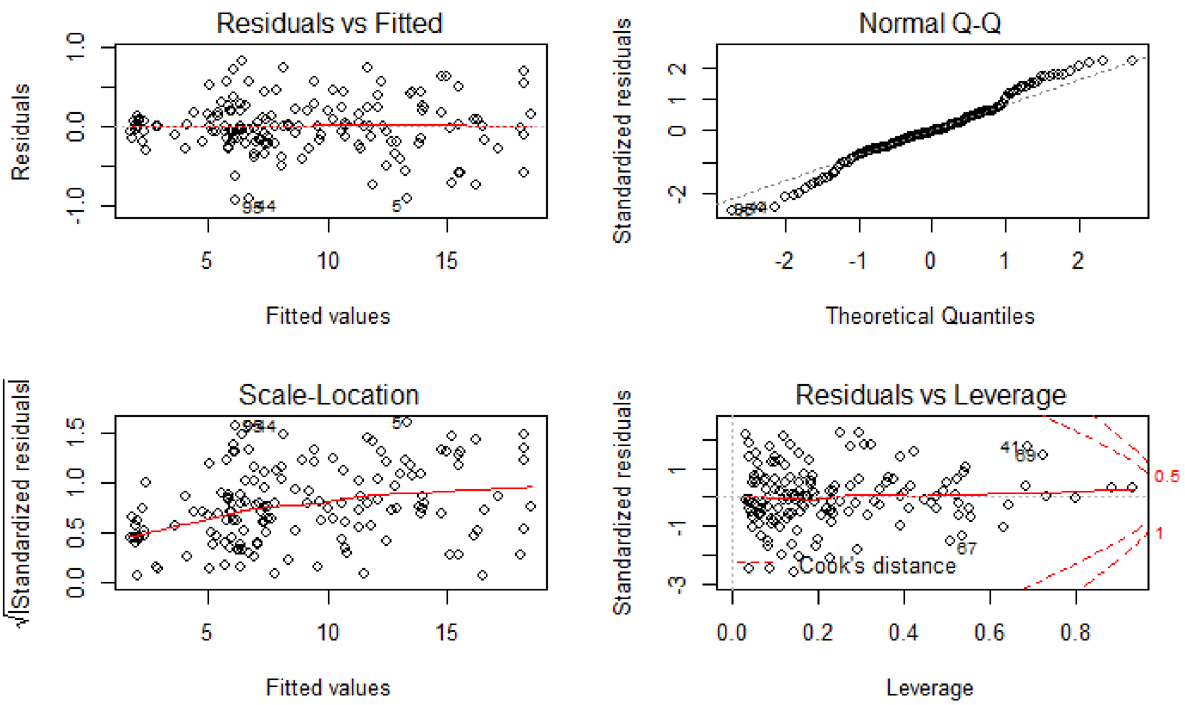
Model selection by ANOVA

```
> anova(m1,m2)
Analysis of Variance Table

Model 1: Height ~ Length1 + Length2 + Length3 + Weight + Width + Species
Model 2: Height ~ Length1 + Length2 + Length3 + Weight + Width + Species +

    Length1 * Species + Length2 * Species + Length3 * Species +
    Weight * Species + Width * Species
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    148 51.673
2    118 17.593 30    34.081 7.6197 3.049e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Scatterplot Matrices



Diagnostic plot of model 2(full model)



Stepwise regression of model 2

```
> step(m2, direction = 'both')
Start:  AIC=-269.23
Height ~ Length1 + Length2 + Length3 + Weight + Width + Species + 
    Length1 * Species + Length2 * Species + Length3 * Species + 
    Weight * Species + Width * Species

                  Df Sum of Sq    RSS     AIC
- Length2:Species  6   0.14119 17.734 -279.95
- Length1:Species  6   0.70363 18.296 -274.96
- Width:Species    6   1.26507 18.858 -270.12
<none>                         17.593 -269.23
- Weight:Species   6   1.43702 19.030 -268.67
- Length3:Species  6   1.88083 19.473 -264.98

Step:  AIC=-279.95
Height ~ Length1 + Length2 + Length3 + Weight + Width + Species + 
    Length1:Species + Length3:Species + Weight:Species + Width:Species

                  Df Sum of Sq    RSS     AIC
- Length2          1   0.07158 17.805 -281.31
<none>                         17.734 -279.95
- Width:Species    6   1.54194 19.276 -278.61
- Weight:Species   6   1.57419 19.308 -278.35
- Length1:Species  6   1.84468 19.578 -276.12
- Length3:Species  6   2.18655 19.920 -273.35
+ Length2:Species  6   0.14119 17.593 -269.23

Step:  AIC=-281.31
Height ~ Length1 + Length3 + Weight + Width + Species + Length1:Species + 
    Length3:Species + Weight:Species + Width:Species
```

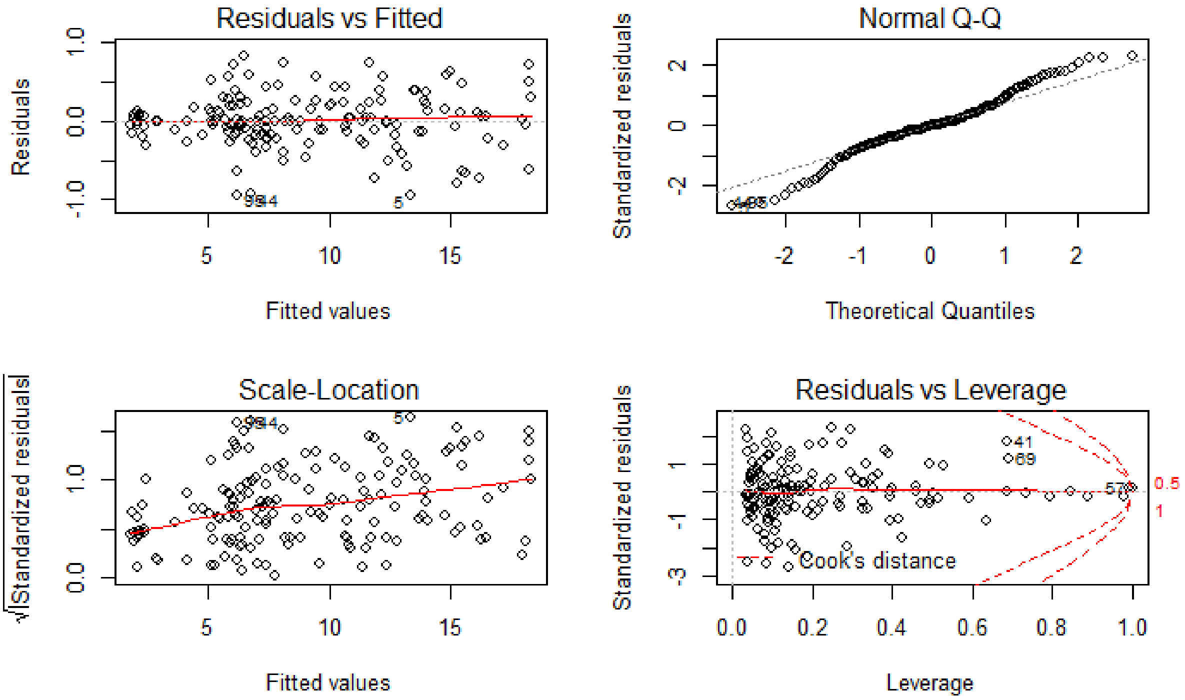
```
                  Df Sum of Sq    RSS     AIC
<none>                         17.805 -281.31
- Width:Species    6   1.53882 19.344 -280.05
+ Length2          1   0.07158 17.734 -279.95
- Weight:Species   6   1.63594 19.441 -279.24
- Length1:Species  6   2.11088 19.916 -275.38
- Length3:Species  6   2.43175 20.237 -272.82

Call:
lm(formula = Height ~ Length1 + Length3 + Weight + Width + Species + 
    Length1:Species + Length3:Species + Weight:Species + Width:Species, 
    data = fish)
```

Summary of stepwise regression of model 2

```
> summary(m2_new)

Call:
lm(formula = Height ~ Length1 + Length3 + Weight + Width + Species + 
    Length1:Species + Length3:Species + Weight:Species + Width:Species, 
    data = fish_new)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9456 -0.1680  0.0009  0.1831  0.8275 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               9.1856824  1.0156923   9.044 2.43e-15 ***
Length1                  -0.4329889  0.2545065  -1.701 0.091375 .  
Length3                   0.5679984  0.2497780   2.274 0.024670 *  
Weight                    0.0049808  0.0010642   4.680 7.34e-06 ***
Width                     0.1112336  0.2976662   0.374 0.709271    
SpeciesParkki             4.4560143  5.3525435   0.833 0.406712    
SpeciesPerch             -4.3328021  1.2222135  -3.545 0.000553 ***
SpeciesPike              -8.6960663  1.4153659  -6.144 9.91e-09 ***
SpeciesRoach             -4.1654081  2.4501147  -1.700 0.091601 .  
SpeciesSmelt             -6.4750810  5.3724391  -1.205 0.230387    
SpeciesWhitefish         -4.3019958  2.6738070  -1.609 0.110152    
Length1:SpeciesParkki    -3.1101118  3.2418035  -0.959 0.339221    
Length1:SpeciesPerch      0.4960196  0.3094512   1.603 0.111480    
Length1:SpeciesPike       1.2022204  0.3626636   3.315 0.001200 ** 
Length1:SpeciesRoach      0.0808139  0.3808095   0.212 0.832284    
Length1:SpeciesSmelt      0.5315168  1.2897011   0.412 0.680954    
Length1:SpeciesWhitefish  3.4220296  6.0142027   0.569 0.570383    
Length3:SpeciesParkki     2.7598087  2.7369509   1.008 0.315234    
Length3:SpeciesPerch     -0.4755035  0.2971774  -1.600 0.112108    
Length3:SpeciesPike      -1.3417852  0.3749942  -3.578 0.000493 ***
Length3:SpeciesRoach     -0.0810769  0.3625491  -0.224 0.823410    
Length3:SpeciesSmelt     -0.5835065  1.0923038  -0.534 0.594153    
Length3:SpeciesWhitefish -3.4448816  5.8408993  -0.590 0.556399    
Weight:SpeciesParkki      0.0101556  0.0136459   0.744 0.458138    
Weight:SpeciesPerch      -0.0036497  0.0012140  -3.006 0.003198 ** 
Weight:SpeciesPike       -0.0026670  0.0015610  -1.709 0.090022 .  
Weight:SpeciesRoach      -0.0057025  0.0028810  -1.979 0.049974 *  
Weight:SpeciesSmelt       0.0506081  0.1206719   0.419 0.675654    
Weight:SpeciesWhitefish   0.0008167  0.0089907   0.091 0.927763    
Width:SpeciesParkki      -1.8627857  2.0827500  -0.894 0.372833    
Width:SpeciesPerch        0.4509554  0.3311808   1.362 0.175755    
Width:SpeciesPike         1.1918384  0.4095306   2.910 0.004277 ** 
Width:SpeciesRoach        0.6635906  0.5922474   1.120 0.264665    
Width:SpeciesSmelt       -0.0611935  0.9345077  -0.065 0.947895    
Width:SpeciesWhitefish    0.5049614  0.4387324   1.151 0.251947    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.3774 on 125 degrees of freedom
Multiple R-squared:  0.994,	Adjusted R-squared:  0.9923 
F-statistic: 605.7 on 34 and 125 DF,  p-value: < 2.2e-16
```

Diagnostic plot of stepwise regression of model 2

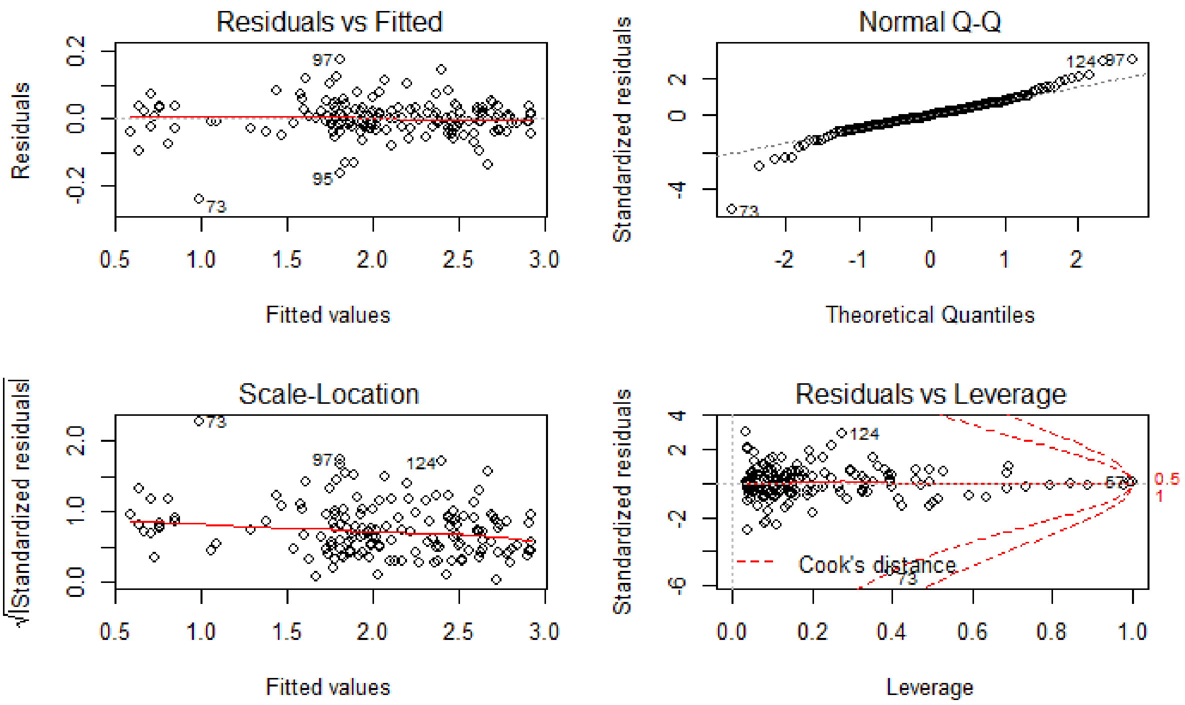


Summary of log-transformation

```
> summary(m2_new_log)

Call:
lm(formula = log(Height) ~ Length1 + Length3 + Weight + Width + 
    Species + Length1:Species + Length3:Species + Weight:Species + 
    Width:Species, data = fish_new)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.238607 -0.025787 -0.000327  0.026715  0.174085 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               2.349e+00  1.596e-01  14.721  < 2e-16 ***
Length1                  -3.355e-02  3.999e-02  -0.839  0.40301    
Length3                   4.449e-02  3.924e-02   1.134  0.25910    
Weight                    3.006e-04  1.672e-04   1.797  0.07469 .  
Width                    -1.071e-03  4.677e-02  -0.023  0.98176    
SpeciesParkki             2.105e-01  8.410e-01   0.250  0.80272    
SpeciesPerch             -5.468e-01  1.920e-01  -2.847  0.00515 ** 
SpeciesPike              -1.256e+00  2.224e-01  -5.647 1.04e-07 ***
SpeciesRoach             -6.130e-01  3.850e-01  -1.592  0.11383    
SpeciesSmelt             -1.092e+00  8.441e-01  -1.293  0.19833    
SpeciesWhitefish         -5.860e-01  4.201e-01  -1.395  0.16552    
```

```
Length1:SpeciesParkki    -3.357e-01  5.093e-01  -0.659  0.51111
Length1:SpeciesPerch     -5.380e-03  4.862e-02  -0.111  0.91207
Length1:SpeciesPike       1.556e-01  5.698e-02   2.731  0.00724 **
Length1:SpeciesRoach     -6.076e-03  5.983e-02  -0.102  0.91928
Length1:SpeciesSmelt      1.147e-01  2.026e-01   0.566  0.57247
Length1:SpeciesWhitefish  4.139e-01  9.449e-01   0.438  0.66209
Length3:SpeciesParkki     3.091e-01  4.300e-01   0.719  0.47363
Length3:SpeciesPerch      3.156e-02  4.669e-02   0.676  0.50038
Length3:SpeciesPike      -1.616e-01  5.892e-02  -2.744  0.00697 **
Length3:SpeciesRoach      2.561e-02  5.696e-02   0.450  0.65374
Length3:SpeciesSmelt     -7.186e-02  1.716e-01  -0.419  0.67615
Length3:SpeciesWhitefish -4.084e-01  9.177e-01  -0.445  0.65709
Weight:SpeciesParkki      3.332e-04  2.144e-03   0.155  0.87674
Weight:SpeciesPerch      -9.318e-04  1.908e-04  -4.885 3.10e-06 ***
Weight:SpeciesPike       -6.064e-05  2.453e-04  -0.247  0.80511
Weight:SpeciesRoach      -8.811e-04  4.527e-04  -1.947  0.05383 .
Weight:SpeciesSmelt       2.027e-02  1.896e-02   1.069  0.28708
Weight:SpeciesWhitefish   2.482e-04  1.413e-03   0.176  0.86079
Width:SpeciesParkki      -8.575e-02  3.272e-01  -0.262  0.79372
Width:SpeciesPerch        1.131e-01  5.203e-02   2.174  0.03156 *
Width:SpeciesPike         1.636e-01  6.434e-02   2.542  0.01224 *
Width:SpeciesRoach        1.237e-01  9.305e-02   1.330  0.18610
Width:SpeciesSmelt        7.723e-03  1.468e-01   0.053  0.95814
Width:SpeciesWhitefish    7.343e-02  6.893e-02   1.065  0.28882
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0593 on 125 degrees of freedom
Multiple R-squared:  0.9915,	Adjusted R-squared:  0.9892 
F-statistic: 431.1 on 34 and 125 DF,  p-value: < 2.2e-16
```

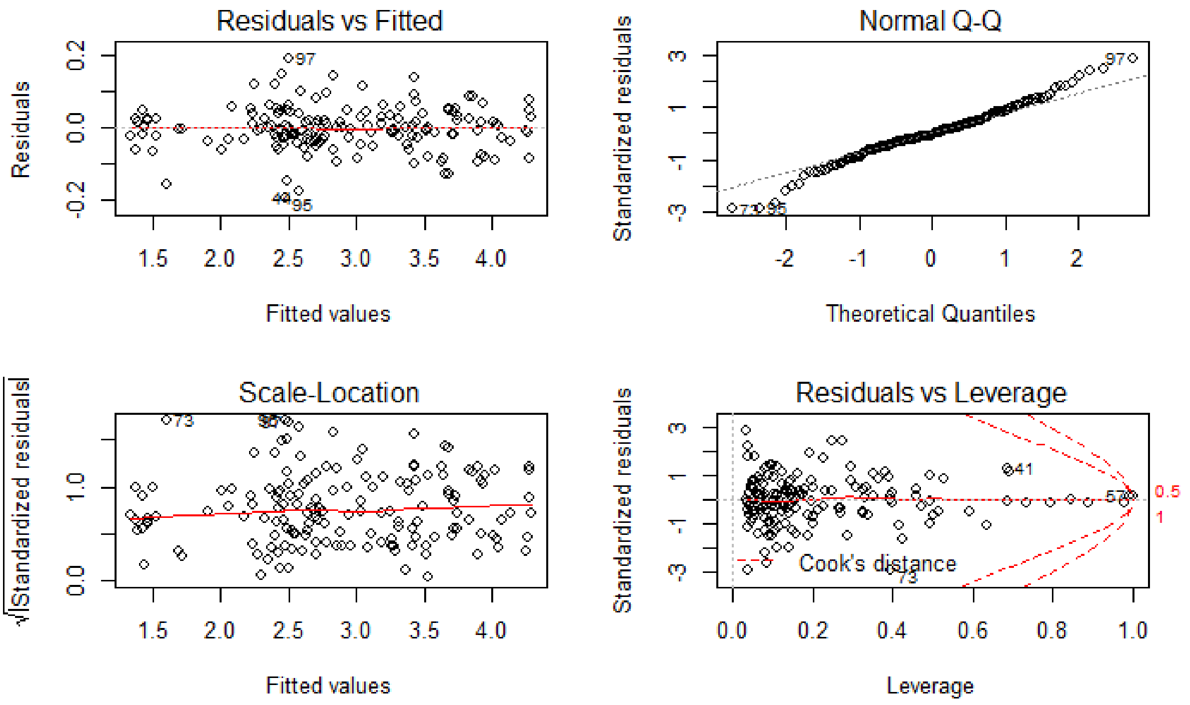
Diagnostic plot for log-transformation



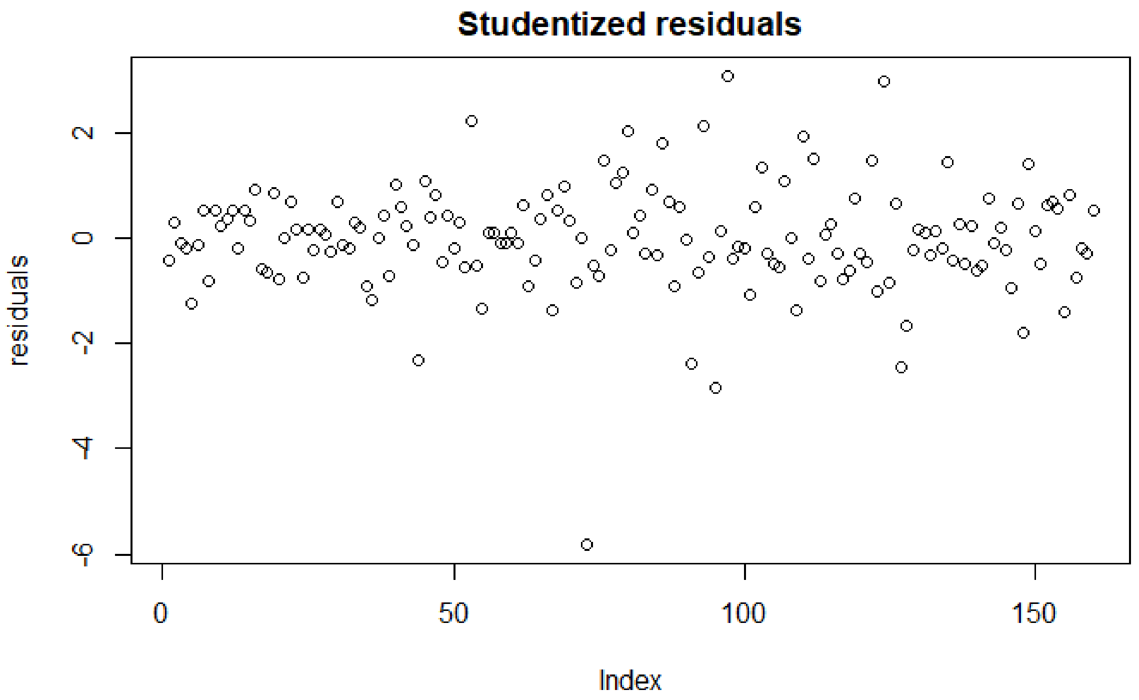
Summary of square root transformation

```
> summary(m2_new_sqrt)

Call:
lm(formula = sqrt(Height) ~ Length1 + Length3 + Weight + Width + 
    Species + Length1:Species + Length3:Species + Weight:Species + 
    Width:Species, data = fish_new)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.19293 -0.03066 -0.00114  0.03244  0.19189 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)               3.1516766  0.1827855  17.242  < 2e-16 ***
Length1                  -0.0604867  0.0458014  -1.321 0.189037    
Length3                   0.0797628  0.0449504   1.774 0.078421 .  
Weight                    0.0006116  0.0001915   3.193 0.001780 ** 
Width                     0.0059917  0.0535684   0.112 0.911120    
SpeciesParkki             0.5097860  0.9632515   0.529 0.597581    
SpeciesPerch             -0.8197133  0.2199513  -3.727 0.000292 ***
SpeciesPike              -1.6786645  0.2547113  -6.590  1.1e-09 ***
SpeciesRoach             -0.8322741  0.4409262  -1.888 0.061404 .  
SpeciesSmelt             -1.4097205  0.9668320  -1.458 0.147326    
SpeciesWhitefish         -0.8203850  0.4811822  -1.705 0.090690 .  
Length1:SpeciesParkki    -0.5136277  0.5833997  -0.880 0.380330    
Length1:SpeciesPerch      0.0494328  0.0556893   0.888 0.376432    
Length1:SpeciesPike       0.2140264  0.0652655   3.279 0.001348 ** 
Length1:SpeciesRoach      0.0002629  0.0685310   0.004 0.996946    
Length1:SpeciesSmelt      0.1077012  0.2320965   0.464 0.643429    
Length1:SpeciesWhitefish  0.5943708  1.0823247   0.549 0.583874    
Length3:SpeciesParkki     0.4642417  0.4925457   0.943 0.347737    
Length3:SpeciesPerch     -0.0302125  0.0534805  -0.565 0.573137    
Length3:SpeciesPike      -0.2303602  0.0674845  -3.414 0.000865 ***
Length3:SpeciesRoach      0.0128315  0.0652449   0.197 0.844408    
Length3:SpeciesSmelt     -0.0930750  0.1965726  -0.473 0.636690    
Length3:SpeciesWhitefish -0.5919312  1.0511367  -0.563 0.574352    
Weight:SpeciesParkki      0.0011388  0.0024557   0.464 0.643651    
Weight:SpeciesPerch      -0.0008691  0.0002185  -3.978 0.000117 ***
Weight:SpeciesPike       -0.0002366  0.0002809  -0.842 0.401335    
Weight:SpeciesRoach      -0.0010488  0.0005185  -2.023 0.045220 *  
Weight:SpeciesSmelt       0.0163800  0.0217163   0.754 0.452104    
Weight:SpeciesWhitefish   0.0002786  0.0016180   0.172 0.863558    
Width:SpeciesParkki      -0.2191226  0.3748147  -0.585 0.559860    
Width:SpeciesPerch        0.1193281  0.0595998   2.002 0.047432 *  
Width:SpeciesPike         0.2230943  0.0736997   3.027 0.003000 ** 
Width:SpeciesRoach        0.1472318  0.1065817   1.381 0.169621    
Width:SpeciesSmelt        0.0043668  0.1681754   0.026 0.979326    
Width:SpeciesWhitefish    0.0999413  0.0789549   1.266 0.207938    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06792 on 125 degrees of freedom
Multiple R-squared:  0.9935,	Adjusted R-squared:  0.9918 
F-statistic: 565.3 on 34 and 125 DF,  p-value: < 2.2e-16
```

Diagnostic plot for square root transformation



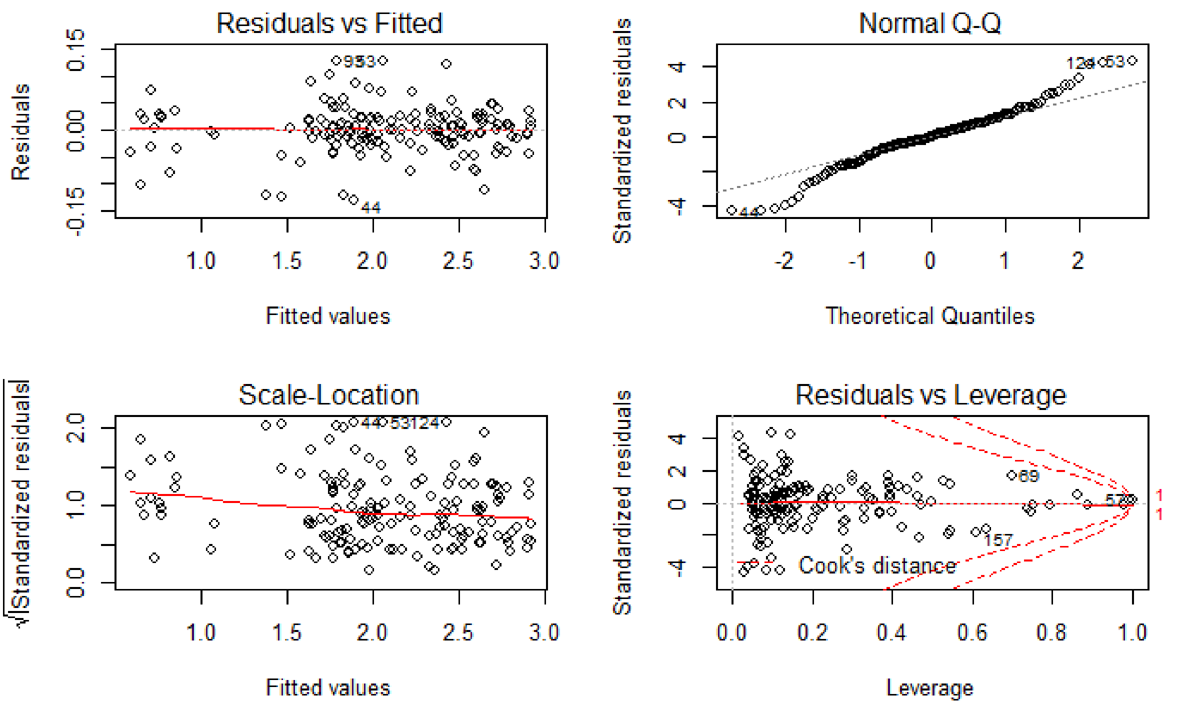
Studentized residual plot for log-transformation



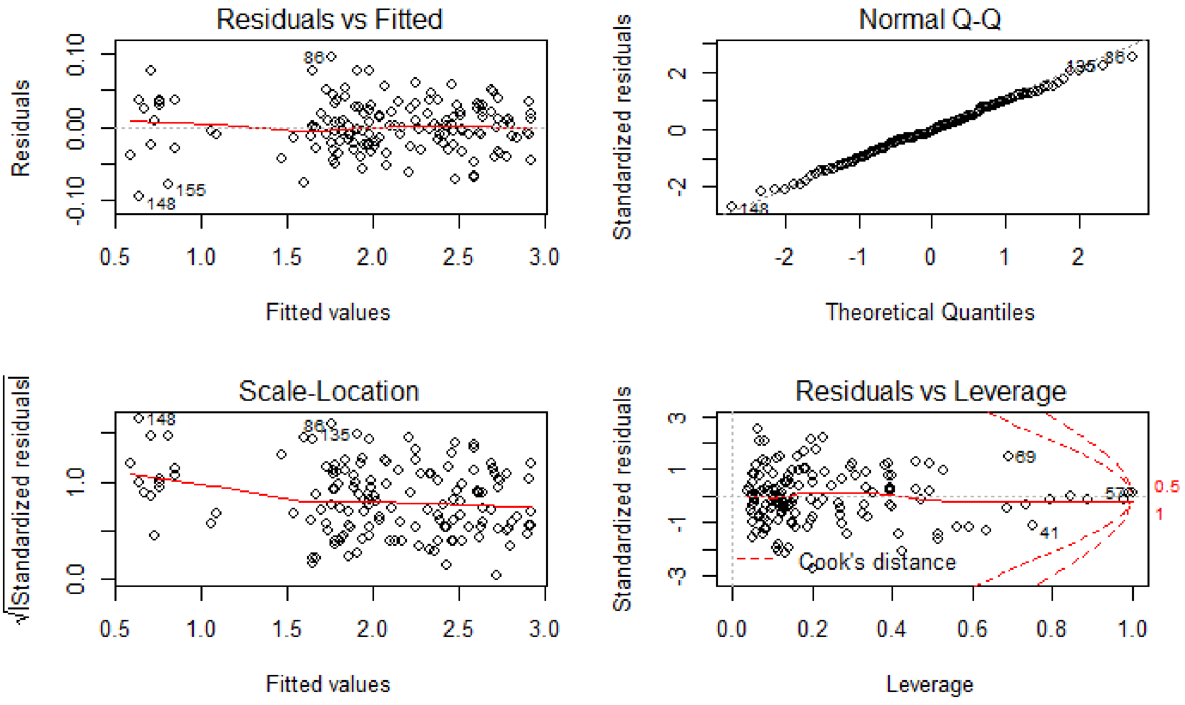
Summary of robust regression with romoved outliers and log transformation

```
> summary(rr.m2_new_log2)

Call: rlm(formula = log(Height) ~ Length1 + Length3 + Weight + Width + 
    Species + Length1:Species + Length3:Species + Weight:Species + 
    Width:Species, data = fish_infremove2, psi = psi.huber)
Residuals:
       Min         1Q     Median         3Q        Max 
-0.1315655 -0.0204992  0.0001975  0.0209630  0.1272408 

Coefficients:
                         Value   Std. Error t value
(Intercept)               2.3305  0.1184    19.6757
Length1                  -0.0309  0.0297    -1.0395
Length3                   0.0413  0.0291     1.4187
Weight                    0.0003  0.0001     2.3421
Width                     0.0058  0.0347     0.1681
SpeciesParkki             0.1169  0.6242     0.1873
SpeciesPerch             -0.7183  0.1438    -4.9971
SpeciesPike              -1.2689  0.1651    -7.6881
SpeciesRoach             -0.6247  0.2857    -2.1864
SpeciesSmelt             -0.9387  0.6265    -1.4984
SpeciesWhitefish         -0.5672  0.3118    -1.8192
Length1:SpeciesParkki    -0.3176  0.3780    -0.8401
Length1:SpeciesPerch      0.0391  0.0366     1.0666
Length1:SpeciesPike       0.1555  0.0423     3.6757
Length1:SpeciesRoach     -0.0052  0.0444    -0.1162
Length1:SpeciesSmelt      0.1227  0.1504     0.8156
Length1:SpeciesWhitefish  0.4112  0.7013     0.5864
Length3:SpeciesParkki     0.2950  0.3192     0.9244
Length3:SpeciesPerch     -0.0207  0.0354    -0.5842
Length3:SpeciesPike      -0.1634  0.0437    -3.7370
Length3:SpeciesRoach      0.0254  0.0423     0.6001
Length3:SpeciesSmelt     -0.0719  0.1274    -0.5644
Length3:SpeciesWhitefish -0.4052  0.6811    -0.5949
Weight:SpeciesParkki     -0.0002  0.0016    -0.1198
Weight:SpeciesPerch      -0.0006  0.0001    -4.2094
Weight:SpeciesPike        0.0000  0.0002    -0.0761
Weight:SpeciesRoach      -0.0011  0.0003    -3.1452
Weight:SpeciesSmelt       0.0176  0.0141     1.2483
Weight:SpeciesWhitefish   0.0003  0.0010     0.2462
Width:SpeciesParkki      -0.0277  0.2429    -0.1141
Width:SpeciesPerch        0.1099  0.0387     2.8389
Width:SpeciesPike         0.1661  0.0478     3.4778
Width:SpeciesRoach        0.1326  0.0691     1.9194
Width:SpeciesSmelt        0.0025  0.1090     0.0226
Width:SpeciesWhitefish    0.0665  0.0512     1.3003

Residual standard error: 0.03108 on 122 degrees of freedom
```

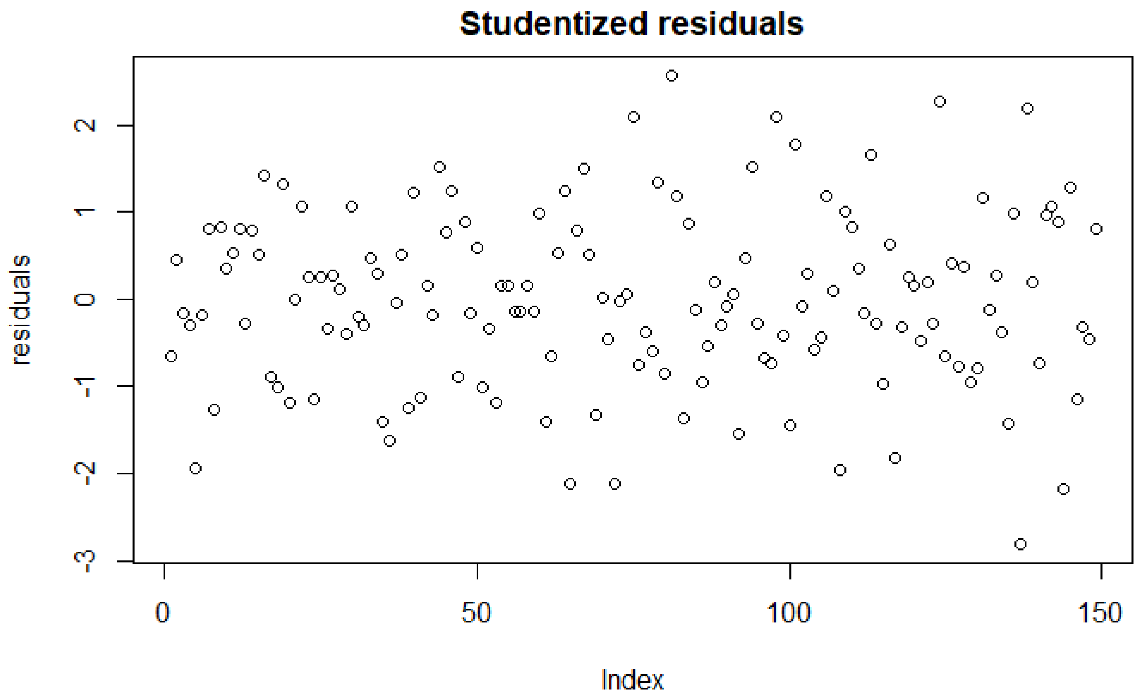
Diagnostic plot of robust regression with romoved outliers and log transformation



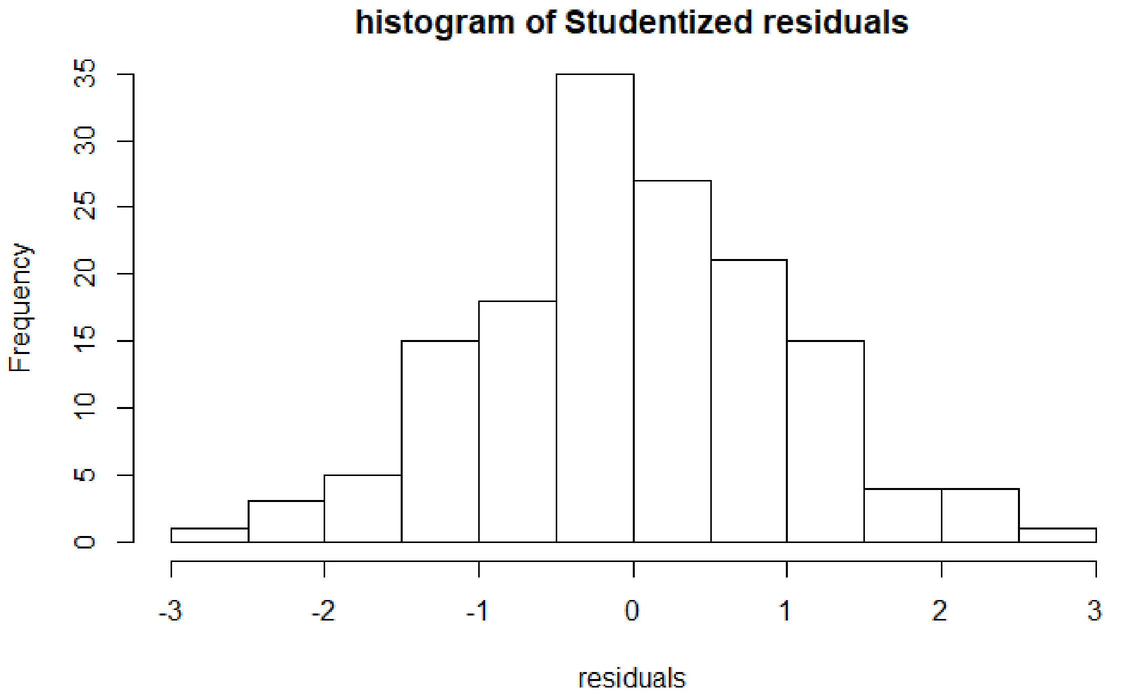
Diagnostic plot of final model



Studentized residual plot of final model



Histogram of Studentized residual in final model



ANOVA of final model

```
> anova(m2_new_log3)
Analysis of Variance Table

Response: log(Height)
                 Df  Sum Sq Mean Sq    F value    Pr(>F)    
Length1           1 21.7693 21.7693 14519.4242 < 2.2e-16 ***
Length3           1 14.6477 14.6477  9769.5444 < 2.2e-16 ***
Weight            1  0.0104  0.0104     6.9599 0.0095000 ** 
Width             1  7.4687  7.4687  4981.3598 < 2.2e-16 ***
Species           6  3.9284  0.6547   436.6806 < 2.2e-16 ***
Length1:Species   6  0.1449  0.0241    16.1063 2.377e-13 ***
Length3:Species   6  0.0191  0.0032     2.1217 0.0560683 .  
Weight:Species    6  0.0371  0.0062     4.1185 0.0008847 ***
Width:Species     6  0.0318  0.0053     3.5339 0.0030296 ** 
Residuals       114  0.1709  0.0015                         
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

variance inflation factor (VIF)

```
> vif(m2_new_log3)
                        GVIF Df GVIF^(1/(2*Df))
Length1         6.743672e+03  1        82.11986
Length3         8.769305e+03  1        93.64457
Weight          1.496073e+02  1        12.23141
Width           2.593501e+02  1        16.10435
Species         2.311278e+16  6        23.10225
Length1:Species 2.913187e+23  6        90.23272
Length3:Species 5.436983e+23  6        95.04877
Weight:Species  4.102141e+13  6        13.62754
Width:Species   4.276361e+15  6        20.07195
```