Extra credit

Chimdi Chikezie

5/31/2023

```
library(Sleuth3)
library(ggplot2)
library(Rmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: plyr
```

```
library(graphics)
library(leaps)
```

Problem: Fish Market Data Analysis A dataset was collected on 7 different common fish species in fish market sales. The goal of this dataset was to find the relationship between the weight (Weight, response variable) of fish (in grams) and the height (Height, explanatory variable) of the fish in cm, the diagonal width (Width, explanatory variable) of the fish in cm, the vertical length (Length1, explanatory variable) of the fish in cm, the diagonal length (Length2, explanatory variable) of the fish in cm, the cross length (Length3, explanatory variable) in cm, and the type of the fish (Species, categorical variable with 7 categories). We perform linear regression analysis with this dataset with the goal of explaining the relationship between the response and the explanatory variables. The dataset is attached to the assignment.

```
load("Fish_Data.RData")
head(Data)
```

```
##   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
```

```
#View(Data)
```

(a) Fit the linear regression model of mean weight against all the explanatory variables, but do not include any interaction term in the model. Report the summary() output of the linear model fit. Plot the residual plot for fitted vs. residual values based on the fitted regression model. What conclusions can you draw from the plot specifically regarding the linear model assumptions? (look for evidence of violations of the model assumptions, identify unusual observations)

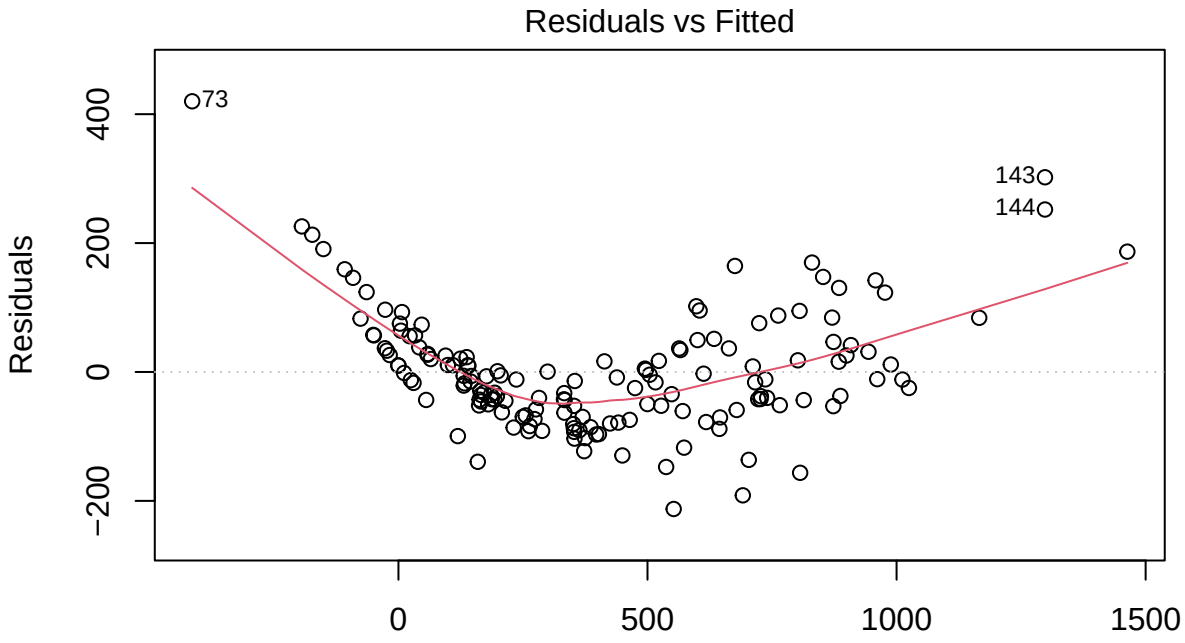
```
lin_mod1 <- lm(Weight ~ Species + Length1 + Length2 + Length3 + Height + Width, data = Data)
summary(lin_mod1)
```

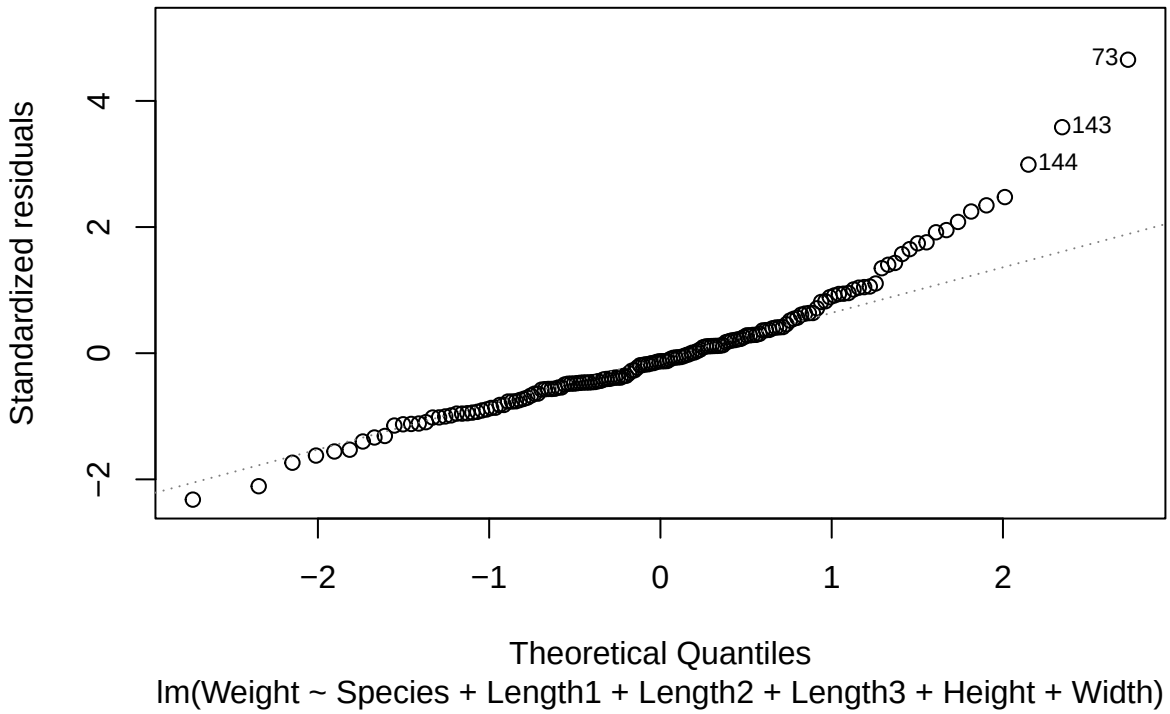
```
##
## Call:
## lm(formula = Weight ~ Species + Length1 + Length2 + Length3 +
##     Height + Width, data = Data)
```

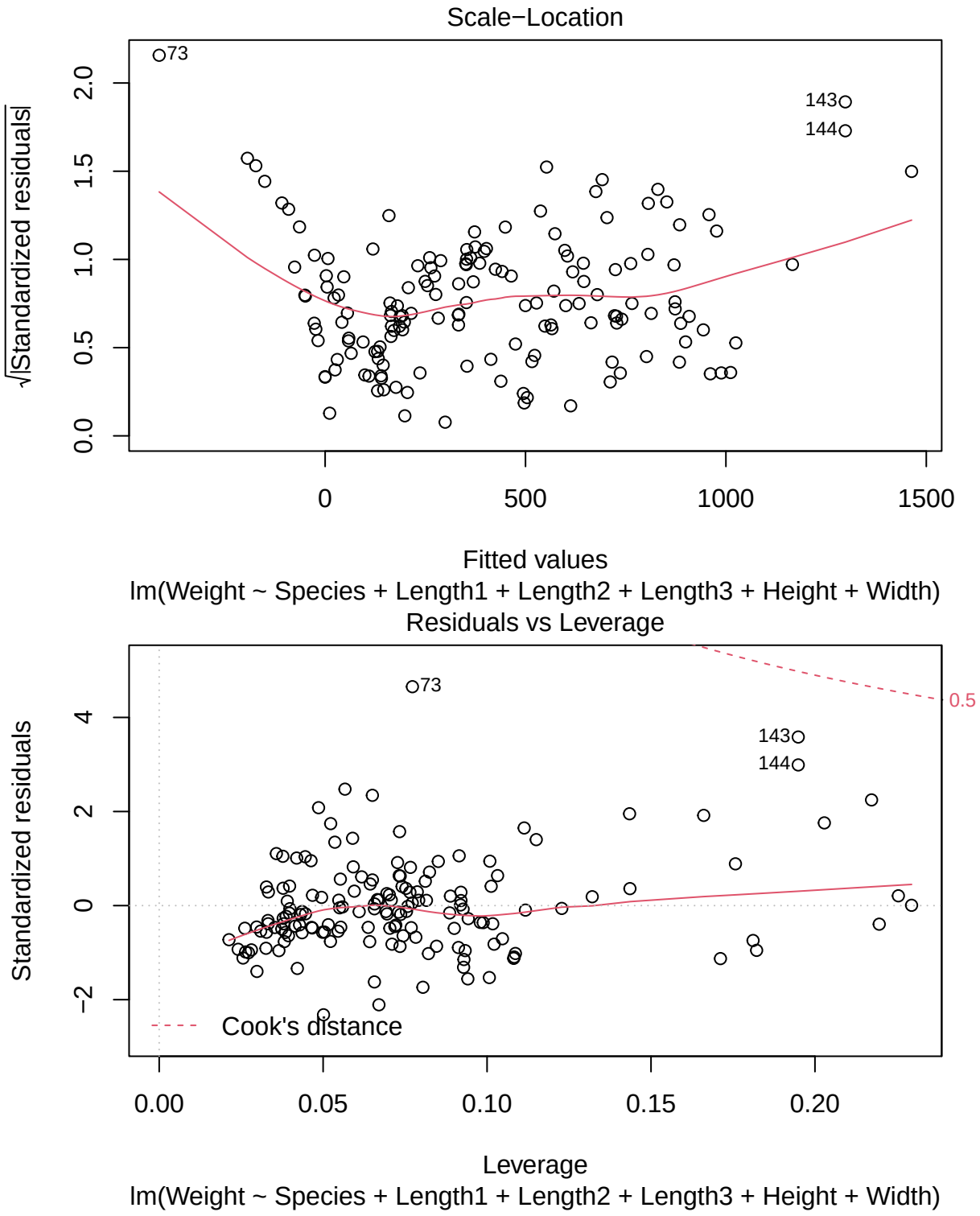
```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -212.56  -52.52  -11.70   36.55  419.97
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -912.7110   127.4597  -7.161 3.64e-11 ***
## SpeciesParkki     160.9212    75.9591   2.119 0.035824 *
## 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
```

From the lm output, the pvalues: 3.64e-11 (Intercept β_0), 0.035824 (SpeciesParkki β_1), 0.000311 (SpeciesSmelt β_5) and 0.029552 (Length1 β_7) are less than $\alpha = 0.05$ which means that these parameters are statistically significant. Therefore, we conclude that the means of SpeciesParkki and SpeciesSmelt are different from the mean of other species and that they these sopecies as well as Length1 have an effect on the mean weight of fishes.

```
plot(lin_mod1)
```


Residuals vs Fitted
Residuals
Fitted values
lm(Weight ~ Species + Length1 + Length2 + Length3 + Height + Width)
73
143
144
400
200
0
−200
0
500
1000
1500


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Weight ~ Species + Length1 + Length2 + Length3 + Height + Width)
73
143
144
4
2
0
−2
−2
−1
0
1
2



From the plot, there is a violation of linearity as there is a u shape made by the observations. There is a violation of equal variance as the spread of the residuals increases as the fitted values increase.

(b) Transform the response values using the cube-root transformation. You can use the following code to do the transformation Data$Weight_new <- (Data$Weight)^(1/3)$ Fit the linear regression model of mean transformed weight against all the explanatory variables, but do not include any interaction term in the model. Report the summary() output of the linear model fit. Plot the residual plot for fitted

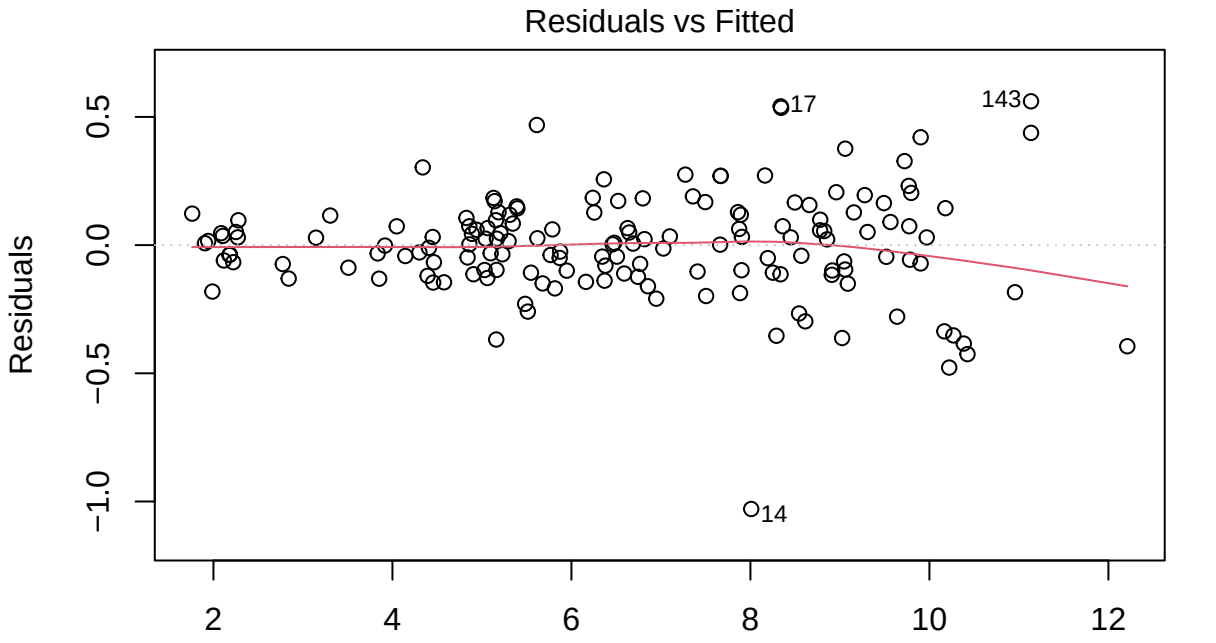
vs. residual values based on the fitted regression model in part (d). What conclusions can you draw from this plot and comparing this plot with the residual plot in part (b)?

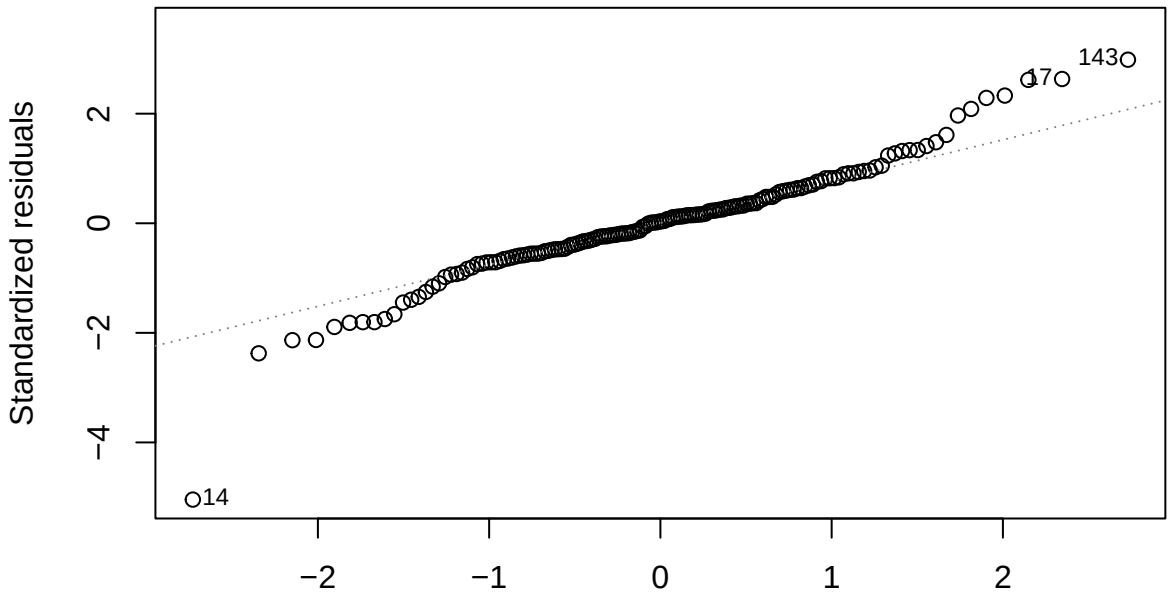
```
Data$Weight_new <- (Data$Weight)^(1/3)
lin_mod2 <- lm(Weight_new ~ Species + Length1 + Length2 + Length3 + Height + Width, data = Data)
summary(lin_mod2)
```

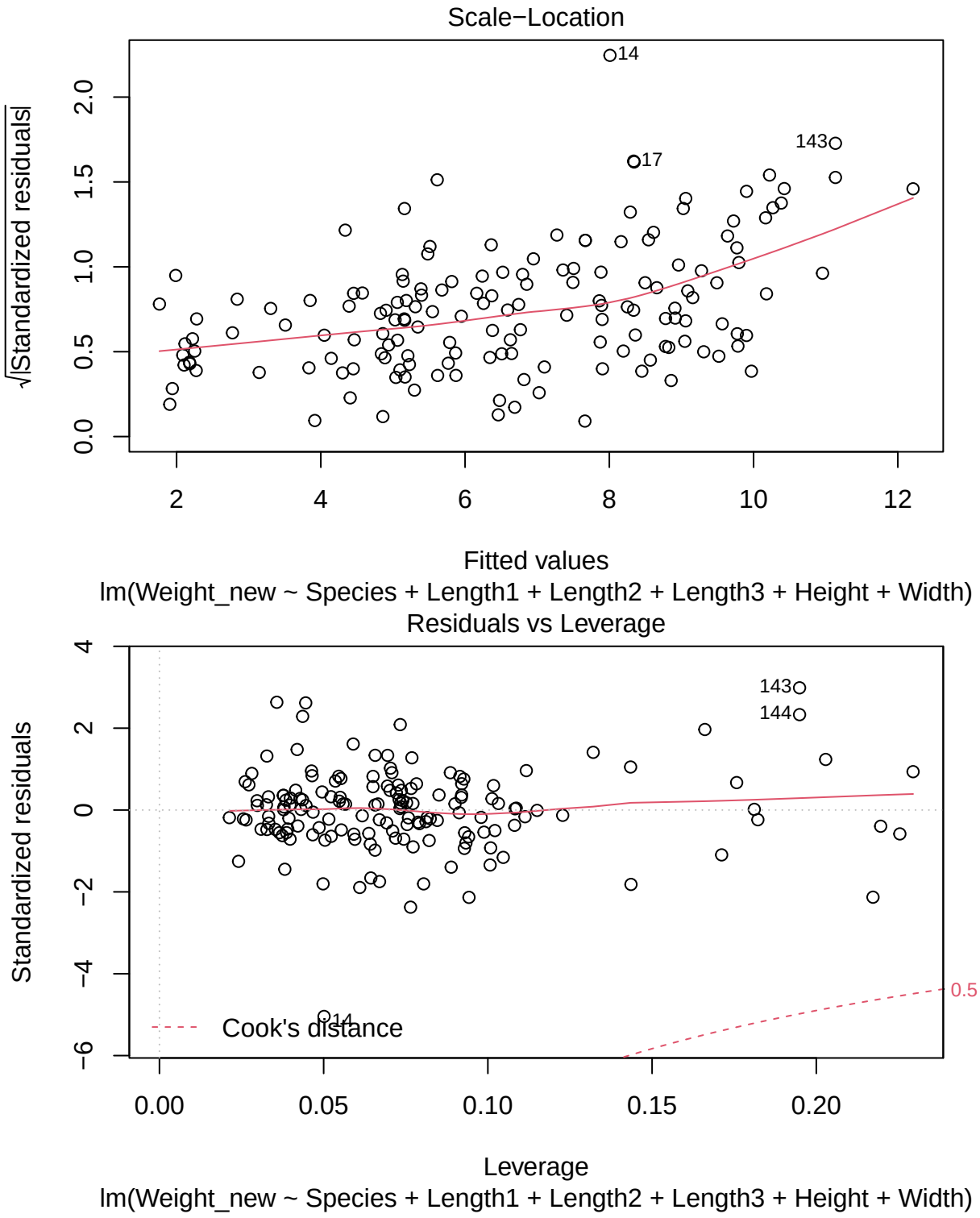
```
##
## Call:
## lm(formula = Weight_new ~ Species + Length1 + Length2 + Length3 +
##     Height + Width, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02924 -0.10228  0.00658  0.10425  0.56092
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -0.30004    0.28406  -1.056  0.29259
## SpeciesParkki     0.33183    0.16928   1.960  0.05188 .
## SpeciesPerch      0.45419    0.26879   1.690  0.09321 .
## SpeciesPike      -0.16471    0.30196  -0.545  0.58625
## SpeciesRoach      0.26178    0.20384   1.284  0.20109
## SpeciesSmelt      0.06640    0.26675   0.249  0.80377
## SpeciesWhitefish  0.62872    0.21580   2.913  0.00414 **
## Length1           0.16128    0.08097   1.992  0.04825 *
## Length2          -0.12969    0.10216  -1.269  0.20629
## Length3           0.10999    0.06571   1.674  0.09629 .
## Height            0.14456    0.02918   4.954 1.99e-06 ***
## Width             0.31290    0.05337   5.863 2.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2094 on 146 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9925
## F-statistic:  1887 on 11 and 146 DF,  p-value: < 2.2e-16
```

From the lm output, the pvalues: 3.64e-11 (Intercept β_0), 0.035824 (SpeciesParkki β_1), 0.000311 (SpeciesSmelt β_5) and 0.029552 (Length1 β_7) are less than $\alpha = 0.05$ which means that these parameters are statistically significant. Therefore, we conclude that the means of SpeciesParkki and SpeciesSmelt are different from the mean of other species and that they these sopecies as well as Length1 have an effect on the mean weight of fishes.

```
plot(lin_mod2)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
−1.0
17
143
14
2
4
6
8
10
12
Fitted values
lm(Weight_new ~ Species + Length1 + Length2 + Length3 + Height + Width)


Normal Q−Q
Standardized residuals
2
0
−2
−4
143
17
14
−2
−1
0
1
2
Theoretical Quantiles
lm(Weight_new ~ Species + Length1 + Length2 + Length3 + Height + Width)



From the residual plot, there is no violation of linearity as the observations are randomly dispersed around the horizontal line. There does not seem to be a violation of equal variance as there is no increase spread of the residuals as the fitted values increase.

c) Based on the transformed weight in part (b), test the hypothesis that the mean transformed weight of Smelt fish is same as the mean transformed weight of White Fish, keeping the values of all the other explanatory variables same.

```
new_species <- relevel(Data$Species, ref="Whitefish")
lin_mod3 <- lm(Weight_new ~ new_species + Length1 + Length2 + Length3 + Height + Width, data = Data)
summary(lin_mod3)
```

```
##
## Call:
## lm(formula = Weight_new ~ new_species + Length1 + Length2 + Length3 +
##     Height + Width, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02924 -0.10228  0.00658  0.10425  0.56092
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.32868    0.15111   2.175 0.031236 *
## new_speciesBream  -0.62872    0.21580  -2.913 0.004138 **
## new_speciesParkki -0.29689    0.13710  -2.166 0.031971 *
## new_speciesPerch  -0.17453    0.11829  -1.475 0.142235
## new_speciesPike   -0.79343    0.18739  -4.234 4.04e-05 ***
## new_speciesRoach  -0.36695    0.10702  -3.429 0.000788 ***
## new_speciesSmelt  -0.56232    0.14034  -4.007 9.78e-05 ***
## Length1            0.16128    0.08097   1.992 0.048253 *
## Length2           -0.12969    0.10216  -1.269 0.206290
## Length3            0.10999    0.06571   1.674 0.096288 .
## Height             0.14456    0.02918   4.954 1.99e-06 ***
## Width              0.31290    0.05337   5.863 2.91e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2094 on 146 degrees of freedom
## Multiple R-squared:  0.993,  Adjusted R-squared:  0.9925
## F-statistic:  1887 on 11 and 146 DF,  p-value: < 2.2e-16
```

H_0 : mean transformed weight of Smelt fish = mean transformed weight of White Fish H_A : mean transformed weight of Smelt fish $\neq$ mean transformed weight of White Fish

Since the p-value (9.78e-05) associated with β_6 (the parameter representing the difference in mean transformed weight of Smelt fish and White Fish) is less than $\alpha = 0.05$. It means β_6 is significant. Therefore, we reject the null hypothesis and conclude that there hat the mean transformed weight of Smelt fish is not the same as the mean transformed weight of White Fish.

(d) Fit the linear regression model of mean transformed weight against all the explanatory variables as well as the interaction terms between Species and Height and the interaction terms between Species and Width. Compare the model in part (b) with this model using ANOVA F-test. Which model will you choose?

```
lin_mod4 <- lm(Weight_new ~ Species + Length1 + Length2 + Length3 + Height + Width + Species:Height + Sp
summary(lin_mod4)
```

```
##
## Call:
## lm(formula = Weight_new ~ Species + Length1 + Length2 + Length3 +
##     Height + Width + Species:Height + Species:Width, data = Data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -1.02112 -0.09179 -0.00935  0.09439  0.54854
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -0.004191   0.373099  -0.011 0.991054
## SpeciesParkki            0.210710   0.528881   0.398 0.690964
## SpeciesPerch             0.068577   0.376699   0.182 0.855820
## SpeciesPike             -0.504327   0.403239  -1.251 0.213228
## SpeciesRoach             0.047703   0.434993   0.110 0.912840
## SpeciesSmelt             0.075582   0.505805   0.149 0.881440
## SpeciesWhitefish        -0.745007   0.669152  -1.113 0.267547
## Length1                  0.056913   0.094757   0.601 0.549108
## Length2                 -0.089573   0.117017  -0.765 0.445341
## Length3                  0.155837   0.075370   2.068 0.040600 *
## Height                   0.194092   0.051612   3.761 0.000252 ***
## Width                    0.133877   0.137019   0.977 0.330297
## SpeciesParkki:Height    -0.211242   0.177746  -1.188 0.236758
## SpeciesPerch:Height      0.009542   0.073678   0.130 0.897152
## SpeciesPike:Height       0.145594   0.145600   1.000 0.319131
## SpeciesRoach:Height     -0.111531   0.136931  -0.815 0.416801
## SpeciesSmelt:Height     -0.083093   0.337726  -0.246 0.806030
## SpeciesWhitefish:Height  0.240045   0.187499   1.280 0.202669
## SpeciesParkki:Width      0.603764   0.449129   1.344 0.181123
## SpeciesPerch:Width       0.144580   0.161018   0.898 0.370843
## SpeciesPike:Width       -0.034571   0.233434  -0.148 0.882489
## SpeciesRoach:Width       0.301589   0.271107   1.112 0.267942
## SpeciesSmelt:Width       0.080592   0.430680   0.187 0.851844
## SpeciesWhitefish:Width  -0.122391   0.306163  -0.400 0.689972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.21 on 134 degrees of freedom
## Multiple R-squared:  0.9936, Adjusted R-squared:  0.9924
## F-statistic: 897.9 on 23 and 134 DF,  p-value: < 2.2e-16
```

```
anova(lin_mod2, lin_mod4)
```

```
## Analysis of Variance Table
##
## Model 1: Weight_new ~ Species + Length1 + Length2 + Length3 + Height +
##     Width
## Model 2: Weight_new ~ Species + Length1 + Length2 + Length3 + Height +
##     Width + Species:Height + Species:Width
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    146 6.4010
## 2    134 5.9087 12   0.49226 0.9303 0.5189
```

Based on the anova result, the p-value (0.5189) for the f-test comparison is greater than $\alpha = 0.05$ which means it is not significant. Therefore, we conclude that the interaction terms that were added do not have any effect on the mean transformed weight of fishes and we prefer the simpler model.

(e) For the model in part (b), based on the anova() command output, produce the ANOVA table with just the three rows of Regression, Residual, and Total.

```
anova(lin_mod2)
```

```
## Analysis of Variance Table
##
## Response: Weight_new
##            Df Sum Sq Mean Sq  F value    Pr(>F)
## Species     6 512.18   85.36 1947.032 < 2.2e-16 ***
## Length1     1 390.64  390.64 8909.999 < 2.2e-16 ***
## Length2     1   0.14    0.14    3.080   0.08136 .
## Length3     1   1.02    1.02   23.351 3.378e-06 ***
## Height      1   4.66    4.66  106.312 < 2.2e-16 ***
## Width       1   1.51    1.51   34.373 2.915e-08 ***
## Residuals 146   6.40    0.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
vals <- matrix(c(11, 910.14, 82.74, 146, 6.40, 0.04, 157, 916.54, 5.837), ncol = 3, byrow = TRUE)
colnames(vals) <- c("Df", "SS", "MS")
rownames(vals) <- c("Regression", "Residual", "Total")
vals <- as.table(vals)
vals
```

```
##                 Df      SS      MS
## Regression  11.000 910.140  82.740
## Residual   146.000   6.400   0.040
## Total      157.000 916.540   5.837
```

f) Based on the fitted model in part (b), find the predicted transformed weight of a Bream fish of height of 16 cm, diagonal width of 5 cm, vertical length of 25 cm, diagonal length of 30 cm, cross length of 35 cm. Also, find the 95% prediction interval of the transformed weight of this fish. If you back-transform (that is, take a cubic power), what is the 95% prediction interval of the weight of this fish?

```
test_value <- data.frame(Species = "Bream", Length1 = 25, Length2 = 30, Length3 = 35, Height = 16, Widt
pred_weight <- predict(lin_mod2, newdata = test_value)
sprintf("predicted transformed weight = %f", pred_weight)
```

```
## [1] "predicted transformed weight = 7.568339"
```

```
sprintf("Prediction interval")
```

```
## [1] "Prediction interval"
```

```
pred_int <-predict(lin_mod2, test_value, interval ="prediction", level=0.95 )
pred_int
```

```
##        fit      lwr      upr
## 1 7.568339 7.001071 8.135608
```

```
sprintf("Transformed Prediction interval")
```

```
## [1] "Transformed Prediction interval"
```

```
pred_int^(3)
```

```
##        fit      lwr      upr
## 1 433.5127 343.1575 538.4806
```