HW2 Peer Assessment

Background

The fishing industry uses numerous measurements to describe a specific fish. Our goal is to predict the weight of a fish based on a number of these measurements and determine if any of these measurements are insignificant in determining the weigh of a product. See below for the description of these measurments.

Data Description

The data consists of the following variables:

1. **Weight**: weight of fish in g (numerical)
2. **Species**: species name of fish (categorical)
3. **Body.Height**: height of body of fish in cm (numerical)
4. **Total.Length**: length of fish from mouth to tail in cm (numerical)
5. **Diagonal.Length**: length of diagonal of main body of fish in cm (numerical)
6. **Height**: height of head of fish in cm (numerical)
7. **Width**: width of head of fish in cm (numerical)

Read the data

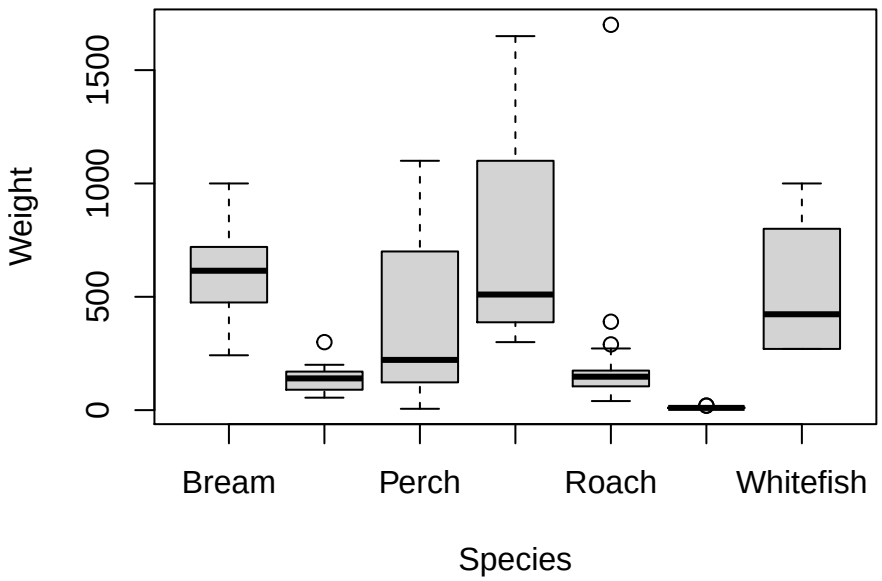
```
# Import library you may need
library(car)
# Read the data set
fishfull = read.csv("Fish.csv", header = T, fileEncoding = "UTF-8-BOM")
row.cnt = nrow(fishfull)
# Split the data into training and testing sets
fishtest = fishfull[(row.cnt - 9):row.cnt, ]
fish = fishfull[1:(row.cnt - 10), ]
```

Please use fish as your data set for the following questions unless otherwise stated.

Question 1: Exploratory Data Analysis [10 points]

(a) Create a box plot comparing the response variable, *Weight*, across the multiple *species*. Based on this box plot, does there appear to be a relationship between the predictor and the response?

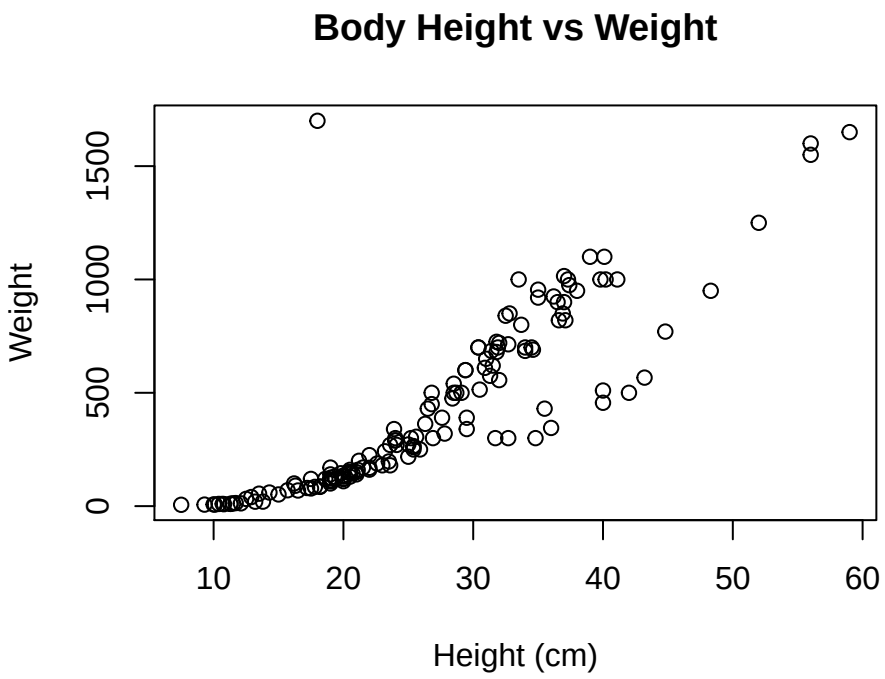
```
boxplot(Weight ~ Species, data = fish, xlab = "Species", ylab = "Weight")
```



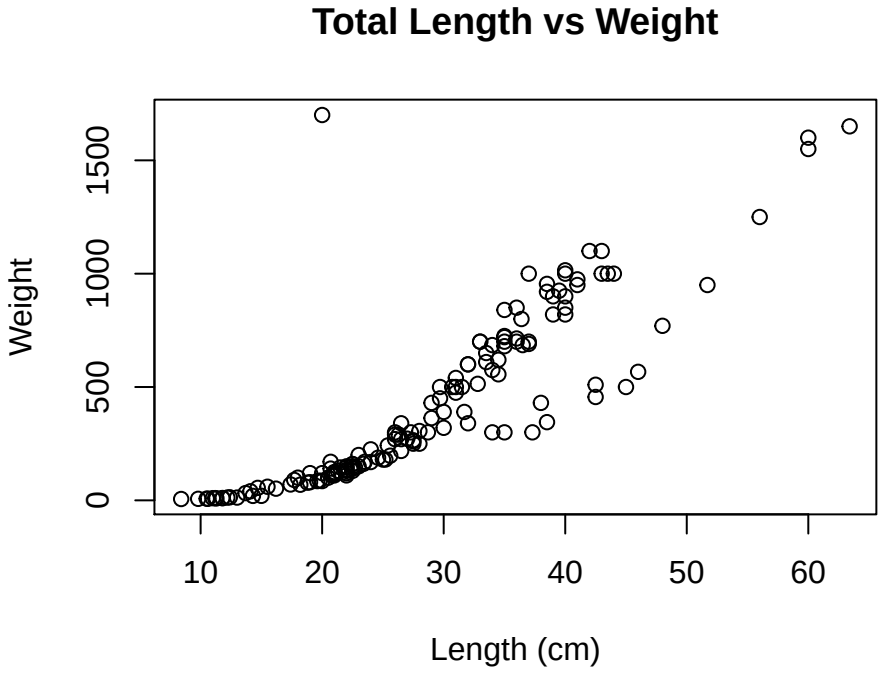
Based on this box plot, there is a relationship between species and weight. Most species have a distinct range and/or average weight - Parkki and Roach species are the most similar but hopefully we can use the other predictors to distinguish between the two.

(b) Create plots of the response, *Weight*, against each quantitative predictor, namely Body.Height**,** Total.Length**,** Diagonal.Length**,** Height**, and** Width**. Describe the general trend of each plot. Are there any potential outliers?**

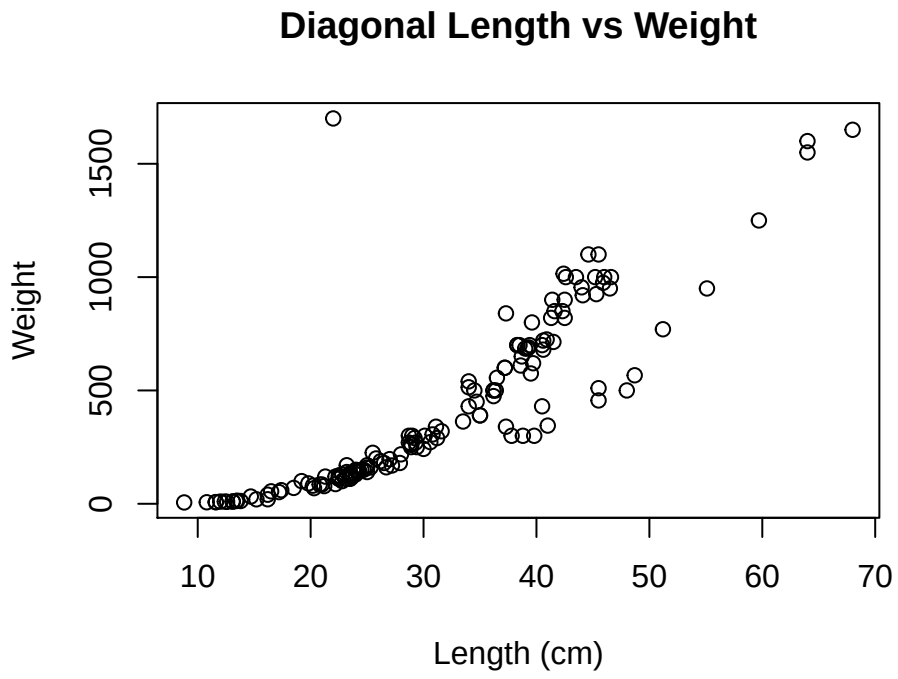
```
plot(x = fish$Body.Height, y = fish$Weight, main = "Body Height vs Weight", xlab = "Height (cm)",
    ylab = "Weight")
```



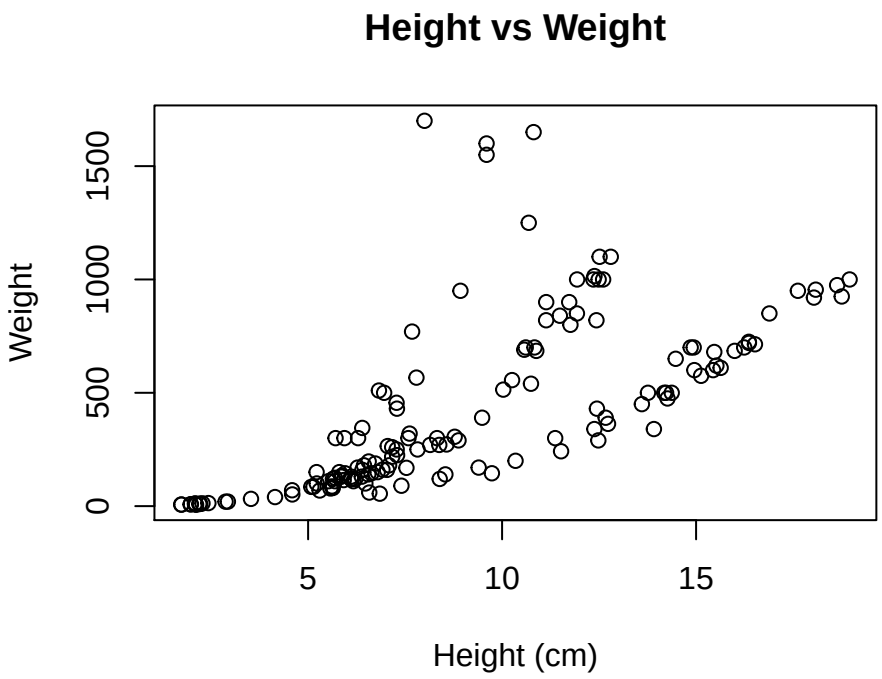

```
plot(x = fish$Total.Length, y = fish$Weight, main = "Total Length vs Weight", xlab = "Length (cm)",
    ylab = "Weight")
```



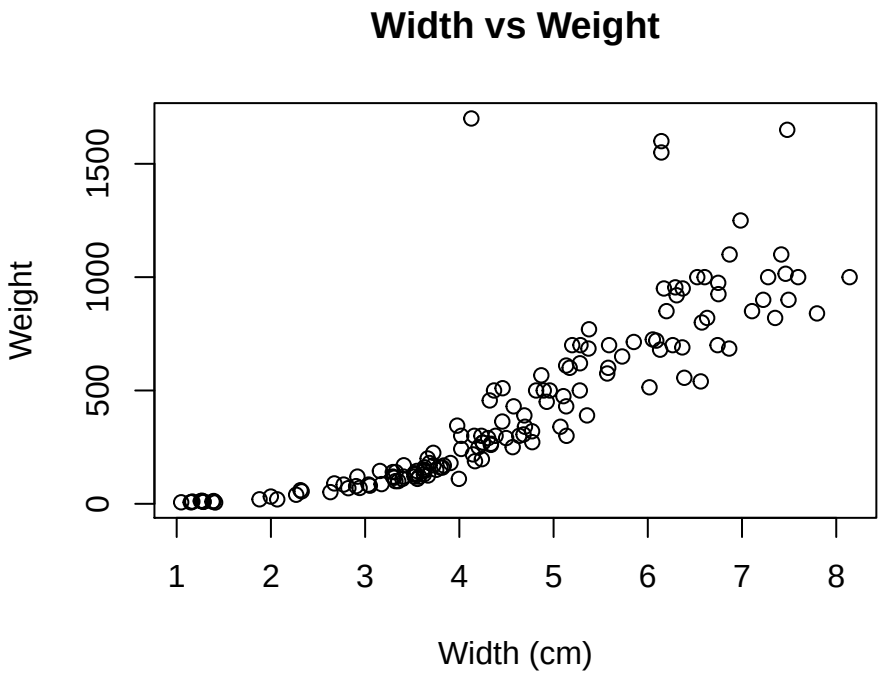
```r
plot(x = fish$Diagonal.Length, y = fish$Weight, main = "Diagonal Length vs Weight",
    xlab = "Length (cm)", ylab = "Weight")
```




```r
plot(x = fish$Height, y = fish$Weight, main = "Height vs Weight", xlab = "Height (cm)",
    ylab = "Weight")
```




```
plot(x = fish$Width, y = fish$Weight, main = "Width vs Weight", xlab = "Width (cm)",
    ylab = "Weight")
```



The general trend between predictors and Weight appears to be exponential. It looks like there is one outlier

in the upper left part of the graphs.

(c) Display the correlations between each of the variables. Interpret the correlations in the context of the relationships of the predictors to the response and in the context of multicollinearity.

```
cor(x = fish[, -1:-2])
```

```
##                 Body.Height Total.Length Diagonal.Length    Height     Width
## Body.Height       1.0000000    0.9995134       0.9919502 0.6268604 0.8661882
## Total.Length      0.9995134    1.0000000       0.9940896 0.6422261 0.8728030
## Diagonal.Length   0.9919502    0.9940896       1.0000000 0.7052116 0.8770361
## Height            0.6268604    0.6422261       0.7052116 1.0000000 0.7908491
## Width             0.8661882    0.8728030       0.8770361 0.7908491 1.0000000
```

Each predictor has a near perfect or strong positive correlation with other predictors. This indicates that multicollinearity exists between the predictors.

(d) Based on this exploratory analysis, is it reasonable to assume a multiple linear regression model for the relationship between *Weight* and the predictor variables?

Based on the correlation analysis, a multiple linear regression model seems reasonable. However, the plots in part b suggest that a transformation may be required to satisfy the linearity assumption of the model.

Question 2: Fitting the Multiple Linear Regression Model [11 points]

Create the full model without transforming the response variable or predicting variables using the fish data set. Do not use fishtest

(a) Build a multiple linear regression model, called model1, using the response and all predictors. Display the summary table of the model.

```
model1 <- lm(Weight ~ ., data = fish)
summary(model1)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -211.37  -70.59  -23.50   42.42 1335.87
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -813.90     218.34  -3.728 0.000282 ***
## SpeciesParkki       79.34     132.71   0.598 0.550918
## SpeciesPerch        10.41     206.26   0.050 0.959837
## SpeciesPike         16.76     233.06   0.072 0.942775
## SpeciesRoach       194.03     156.84   1.237 0.218173
## SpeciesSmelt       455.78     204.92   2.224 0.027775 *
```

```
## SpeciesWhitefish    28.31     164.91   0.172 0.863967
## Body.Height       -176.87      61.36  -2.882 0.004583 **
## Total.Length       266.70      77.75   3.430 0.000797 ***
## Diagonal.Length    -72.49      49.48  -1.465 0.145267
## Height              38.27      22.09   1.732 0.085448 .
## Width               29.63      40.54   0.731 0.466080
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 156.1 on 137 degrees of freedom
## Multiple R-squared:  0.8419, Adjusted R-squared:  0.8292
## F-statistic:  66.3 on 11 and 137 DF,  p-value: < 2.2e-16
```

(b) Is the overall regression significant at an α level of 0.01?
Yes; the p-value for the F-statistic is less than 0.01 and indicates that at least one of the predictors has predictive power (ie - the regression coefficient is different from 0).

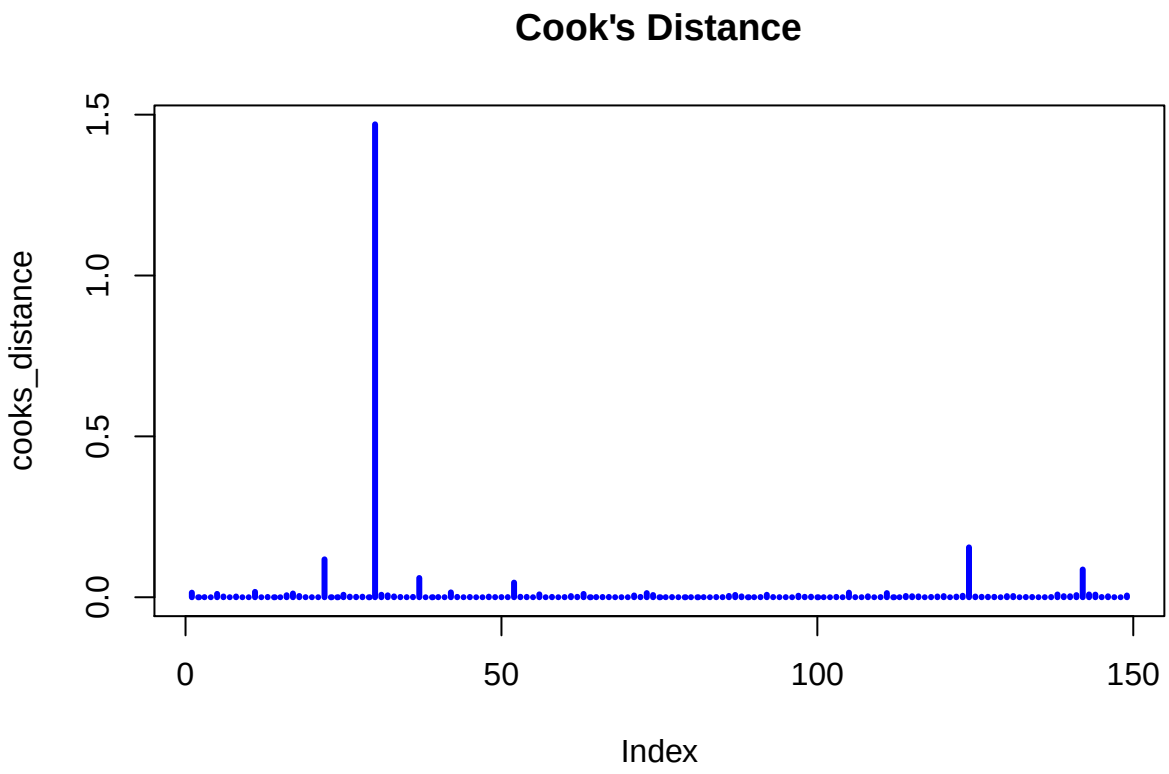
(c) What is the coefficient estimate for *Body.Height*? Interpret this coefficient.
-176.87. All other things equal, 1 cm increase in height will decrease weight by -177 grams.

(d) What is the coefficient estimate for the *Species* category Parkki? Interpret this coefficient.
79.34. All other things equal, having Parkki as the species will increase weight by 79 grams.

Question 3: Checking for Outliers and Multicollinearity [9 points]

(a) Create a plot for the Cook's Distances. Using a threshold Cook's Distance of 1, identify the row numbers of any outliers.

```
cooks_distance <- cooks.distance(model1)
plot(cooks_distance, type = "h", lwd = 3, col = "blue", main = "Cook's Distance")
```




```
influential <- as.numeric(names(cooks_distance)[(cooks_distance > 1)])
influential
```

```
## [1] 30
```

Row 30 is an outlier.

(b) Remove the outlier(s) from the data set and create a new model, called model2, using all predictors with *Weight* as the response. Display the summary of this model.

```
fish2 <- fish[-influential, ]
model2 <- lm(Weight ~ ., data = fish2)
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -211.10  -50.18  -14.44   34.04  433.68
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -969.766    131.601  -7.369 1.51e-11 ***
```

```
## SpeciesParkki     195.500     80.105   2.441 0.015951 *
## SpeciesPerch      174.241    124.404   1.401 0.163608
## SpeciesPike      -175.936    140.605  -1.251 0.212983
## SpeciesRoach      141.867     94.319   1.504 0.134871
## SpeciesSmelt      489.714    123.174   3.976 0.000113 ***
## SpeciesWhitefish  122.277     99.293   1.231 0.220270
## Body.Height       -76.321     37.437  -2.039 0.043422 *
## Total.Length       74.822     48.319   1.549 0.123825
## Diagonal.Length    34.349     30.518   1.126 0.262350
## Height             10.000     13.398   0.746 0.456692
## Width              -8.339     24.483  -0.341 0.733924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.84 on 136 degrees of freedom
## Multiple R-squared:  0.9385, Adjusted R-squared:  0.9335
## F-statistic: 188.6 on 11 and 136 DF,  p-value: < 2.2e-16
```

(c) Display the VIF of each predictor for model2. Using a VIF threshold of max(10, 1/(1-R^2) what conclusions can you draw?

```
library(car)
threshold <- max(10, 1/(1 - summary(model2)$r.squared))
vif(model2)
```

```
##                       GVIF Df GVIF^(1/(2*Df))
## Species          1545.55017  6        1.843983
## Body.Height      2371.15420  1       48.694499
## Total.Length     4540.47698  1       67.383062
## Diagonal.Length  2126.64985  1       46.115614
## Height             56.21375  1        7.497583
## Width              29.01683  1        5.386727
```

```
threshold
```

```
## [1] 16.25583
```

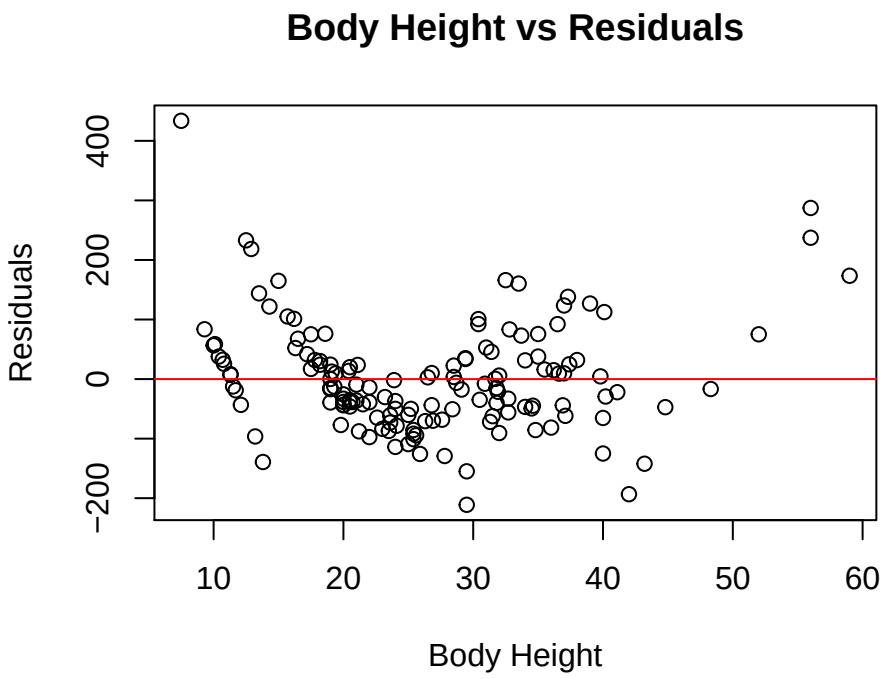
All of the predictors have a VIF greater than max(10, 1/(1-R^2) which indicates that multicollinearity exists among the predictors.

Question 4: Checking Model Assumptions [9 points]

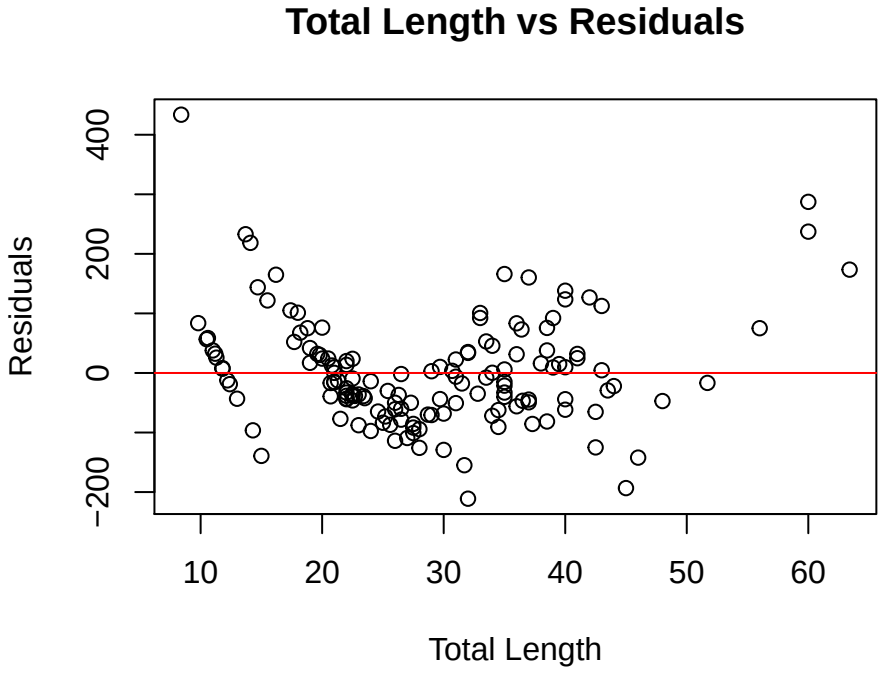
Please use the cleaned data set, which have the outlier(s) removed, and model2 for answering the following questions.

(a) Create scatterplots of the standardized residuals of model2 versus each quantitative predictor. Does the linearity assumption appear to hold for all predictors?

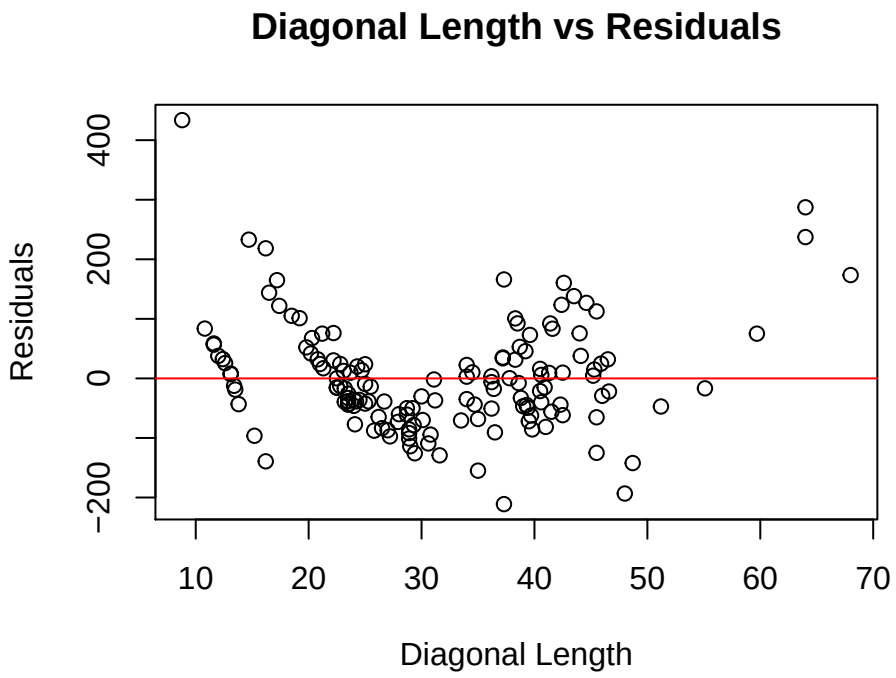
```
plot(fish2$Body.Height, residuals(model2), xlab = "Body Height", ylab = "Residuals",
    main = "Body Height vs Residuals")
abline(0, 0, col = "red")
```



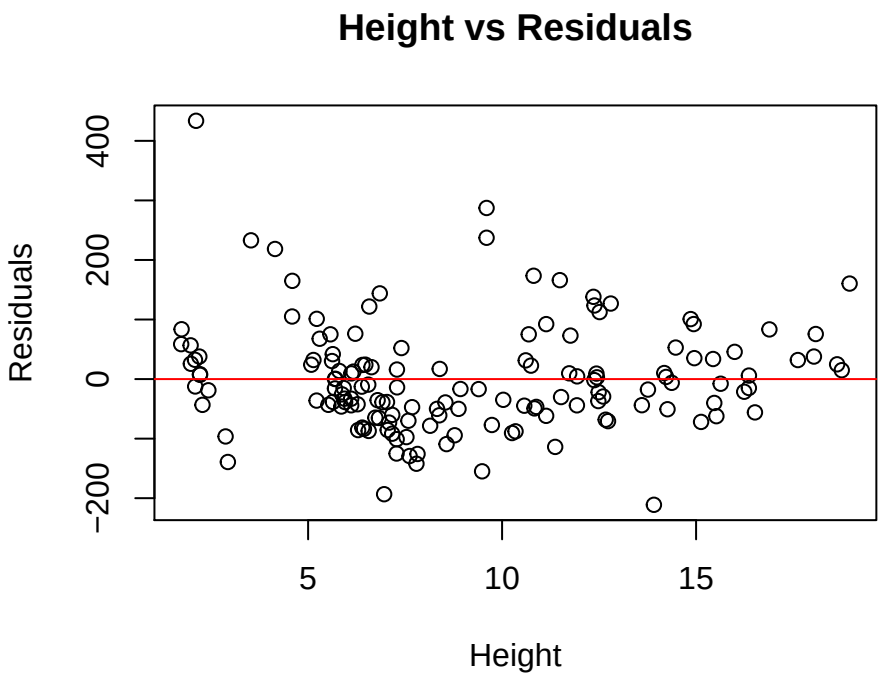

```
plot(fish2$Total.Length, residuals(model2), xlab = "Total Length", ylab = "Residuals",
    main = "Total Length vs Residuals")
abline(0, 0, col = "red")
```



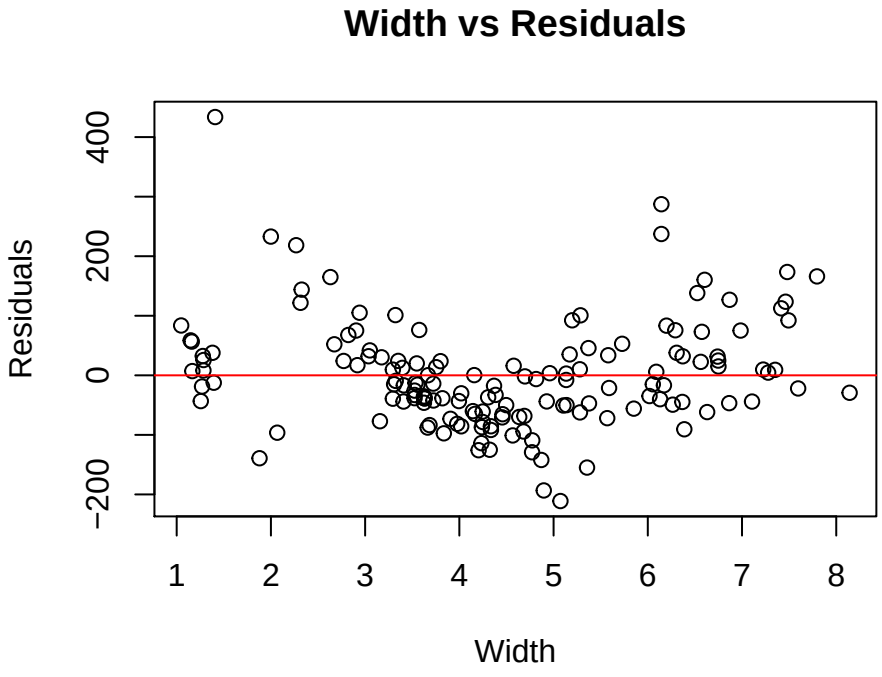
```r
plot(fish2$Diagonal.Length, residuals(model2), xlab = "Diagonal Length", ylab = "Residuals",
    main = "Diagonal Length vs Residuals")
abline(0, 0, col = "red")
```




```r
plot(fish2$Height, residuals(model2), xlab = "Height", ylab = "Residuals", main = "Height vs Residuals")
abline(0, 0, col = "red")
```



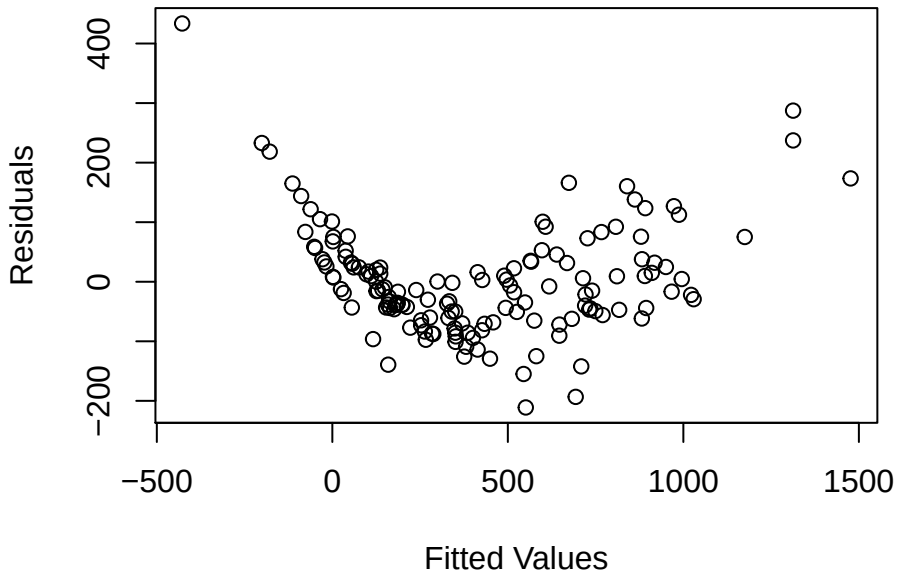

```
plot(fish2$Width, residuals(model2), xlab = "Width", ylab = "Residuals", main = "Width vs Residuals")
abline(0, 0, col = "red")
```



The linearity assumption does not hold as all the residuals are not randomly scattered around 0. There is a slight parabolic curve to each of the graphs.

(b) Create a scatter plot of the standardized residuals of model2 versus the fitted values of model2. Does the constant variance assumption appear to hold? Do the errors appear uncorrelated?

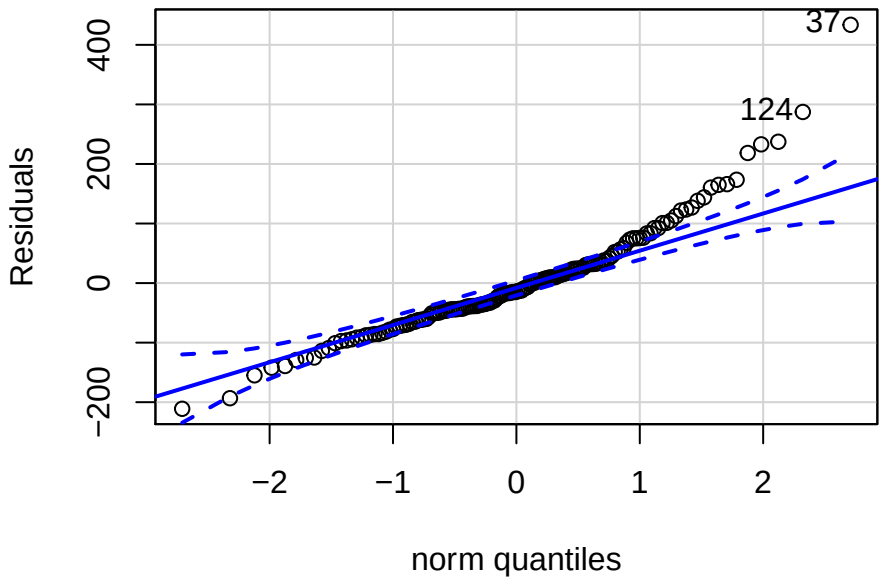
```
plot(model2$fitted.values, model2$residuals, xlab = "Fitted Values", ylab = "Residuals")
```



Constant variance does not hold as there is a parabolic curve to the graph.

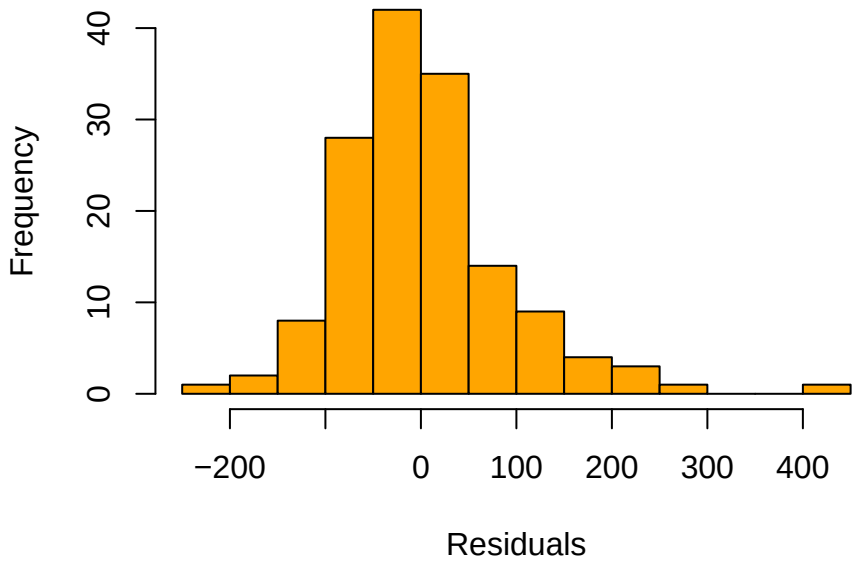
(c) Create a histogram and normal QQ plot for the standardized residuals. What conclusions can you draw from these plots?

```
qqPlot(model2$residuals, ylab = "Residuals", main = "")
```




```
##  37 124
##  36 123
```

```
hist(model2$residuals, xlab = "Residuals", main = "", nclass = 10, col = "orange")
```



Curvature at the ends of

the QQ plot suggest that the normality assumption is violated.

Question 5 Partial F Test [6 points]

(a) Build a third multiple linear regression model using the cleaned data set without the outlier(s), called model3, using only *Species* and *Total.Length* as predicting variables and *Weight* as the response. Display the summary table of the model3.

```
model3 <- lm(Weight ~ Species + Total.Length, data = fish2)
summary(model3)
```

```
##
## Call:
## lm(formula = Weight ~ Species + Total.Length, data = fish2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -233.83  -56.59  -10.13   34.58  418.30
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -730.977     42.449 -17.220  < 2e-16 ***
## SpeciesParkki       63.129     38.889   1.623    0.107
## SpeciesPerch       -23.941     21.745  -1.101    0.273
## SpeciesPike        -400.964     33.350 -12.023  < 2e-16 ***
## SpeciesRoach        -19.876     30.111  -0.660    0.510
## SpeciesSmelt        256.408     39.858   6.433 1.85e-09 ***
## SpeciesWhitefish    -14.971     42.063  -0.356    0.722
## Total.Length         40.775      1.181  34.527  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.86 on 140 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9321
## F-statistic: 289.1 on 7 and 140 DF,  p-value: < 2.2e-16
```

(b) Conduct a partial F-test comparing model3 with model2. What can you conclude using an α level of 0.01?

```
anova(model3, model2)
```

```
## Analysis of Variance Table
##
## Model 1: Weight ~ Species + Total.Length
## Model 2: Weight ~ Species + Body.Height + Total.Length + Diagonal.Length +
##     Height + Width
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    140 1259746
## 2    136 1197659  4     62087 1.7626   0.14
```

The p-value of 0.14 is greater than 0.01 so we fail to reject the null hypothesis and can conclude the additional predictors (Body.Height, Diagonal.Height, Diagonal.Length, Height and Width) add no explanatory power to the model.

Question 6: Reduced Model Residual Analysis and Multicollinearity Test [10 points]

(a) Conduct a multicollinearity test on model3. Comment on the multicollinearity in model3.

```
library("car")
vif(model3)
```

```
##                   GVIF Df GVIF^(1/(2*Df))
## Species      2.654472  6        1.084755
## Total.Length 2.654472  1        1.629255
```

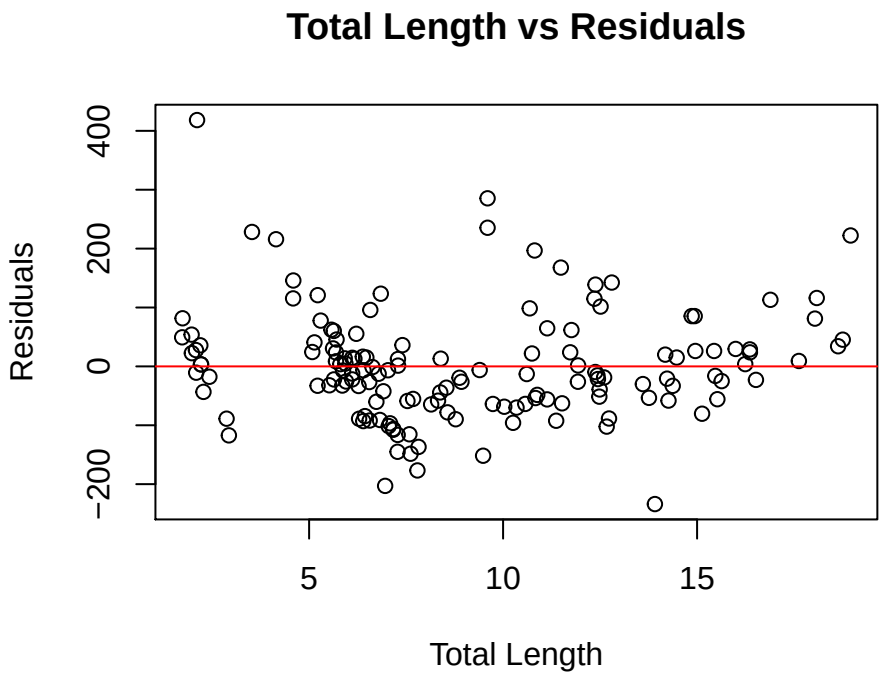
```
threshold <- max(10, 1/(1 - summary(model3)$r.squared))
threshold
```

```
## [1] 15.45466
```

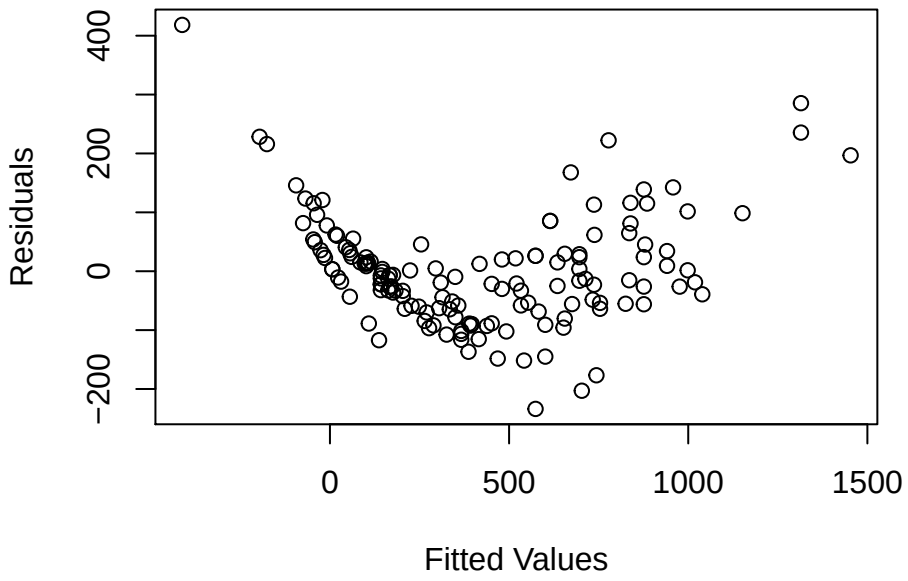
VIF for both Species and Total.Length is less than max(10, 1/(1-R^2) which indicates that multicollinearity does not exist among the predictors.

(b) Conduct residual analysis for model3 (similar to Q4). Comment on each assumption and whether they hold.

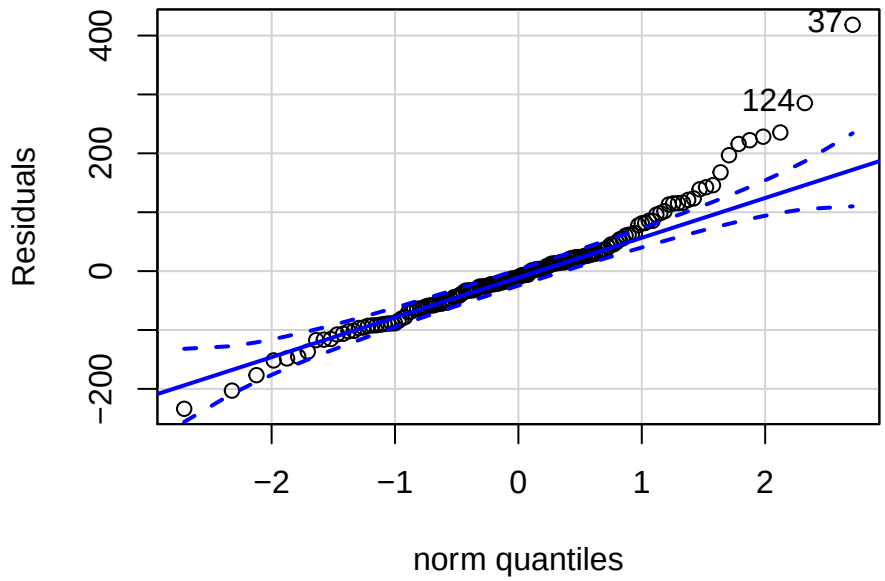
```
# Linearity
plot(fish2$Height, residuals(model3), xlab = "Total Length", ylab = "Residuals",
    main = "Total Length vs Residuals")
abline(0, 0, col = "red")
```



```r
# Constant Variance
plot(model3$fitted.values, model3$residuals, xlab = "Fitted Values", ylab = "Residuals")
```

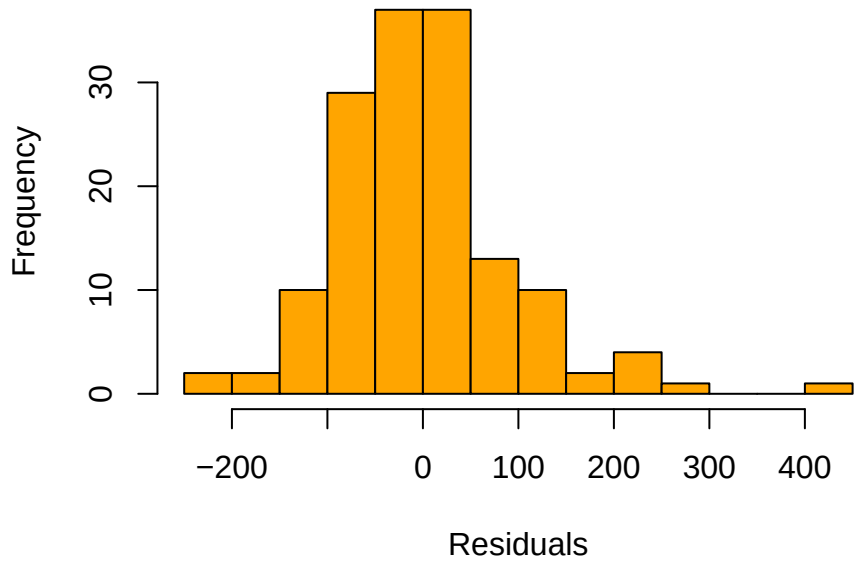



```r
# Normality Assumption
qqPlot(model3$residuals, ylab = "Residuals", main = "")
```




```
##  37 124
##  36 123
```

```
hist(model3$residuals, xlab = "Residuals", main = "", nclass = 10, col = "orange")
```



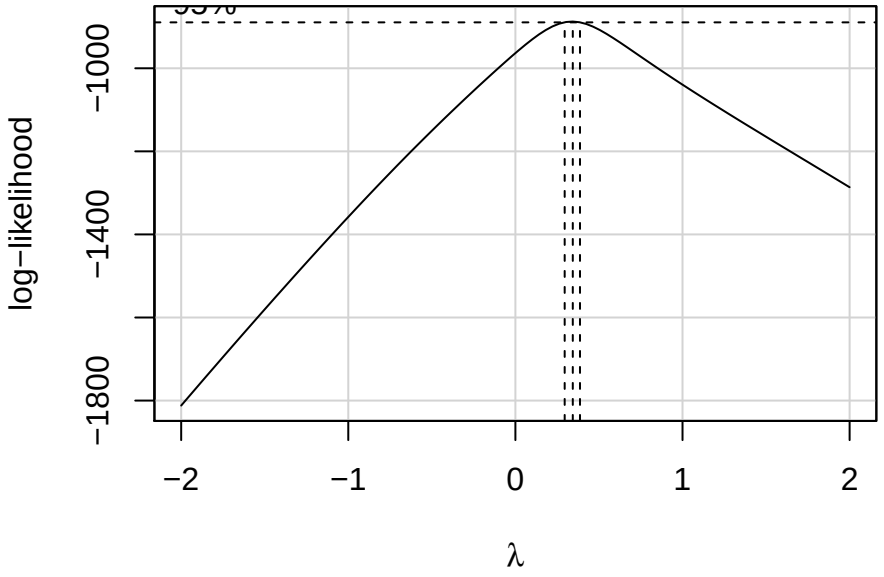
Linearity Assumption

holds better for model 3 than it does for model 2. Constant Variance is still violated in model 3 as the parabolic curve still exists. Normality assumption is also still violated using model 3 as the curvature in the right tail of the graph still exists.

Question 7: Transformation [12 pts]

(a) Use model3 to find the optimal lambda, rounded to the nearest 0.5, for a Box-Cox transformation on model3. What transformation, if any, should be applied according to the lambda value? Please ensure you use model3

```
library("car")
box_cox <- boxCox(model3)
```




```
lambda <- box_cox$x[which.max(box_cox$y)]
lambda <- round(lambda/0.5) * 0.5
lambda
```

```
## [1] 0.5
```

Lambda of 0.5 suggests an square root transformation.

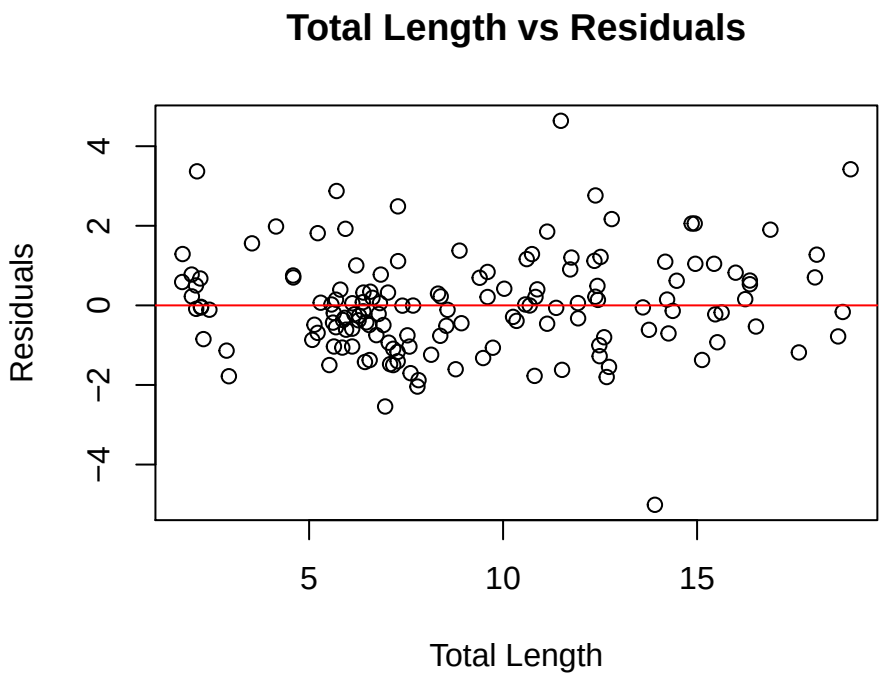
(b) Based on the results in (a), create model4 with the appropriate transformation. Display the summary.

```
model4 <- lm(sqrt(Weight) ~ Species + Total.Length, data = fish2)
summary(model4)
```

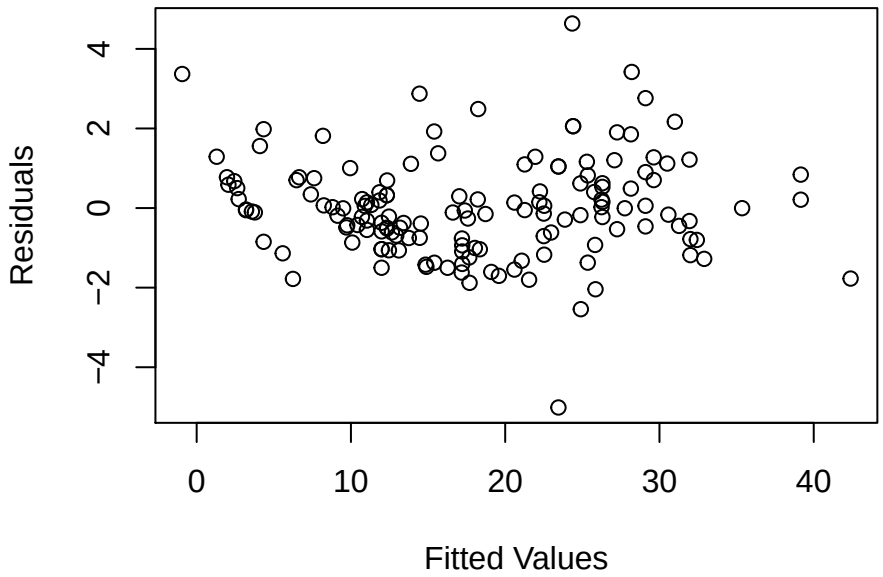
```
##
## Call:
## lm(formula = sqrt(Weight) ~ Species + Total.Length, data = fish2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0111 -0.7687 -0.0579  0.6797  4.6383
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.96654    0.57278 -12.163  < 2e-16 ***
## SpeciesParkki     -0.36404    0.52476  -0.694   0.4890
## SpeciesPerch      -1.95734    0.29342  -6.671 5.46e-10 ***
## SpeciesPike      -10.90490    0.45001 -24.233  < 2e-16 ***
## SpeciesRoach      -2.09340    0.40630  -5.152 8.58e-07 ***
## SpeciesSmelt      -1.04994    0.53782  -1.952   0.0529 .
## SpeciesWhitefish  -0.55048    0.56758  -0.970   0.3338
## Total.Length       0.95052    0.01594  59.649  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.28 on 140 degrees of freedom
## Multiple R-squared:  0.9817, Adjusted R-squared:  0.9808
## F-statistic:  1074 on 7 and 140 DF,  p-value: < 2.2e-16
```

(c) Perform Residual Analysis on model4. Comment on each assumption. Was the transformation successful/unsuccessful?

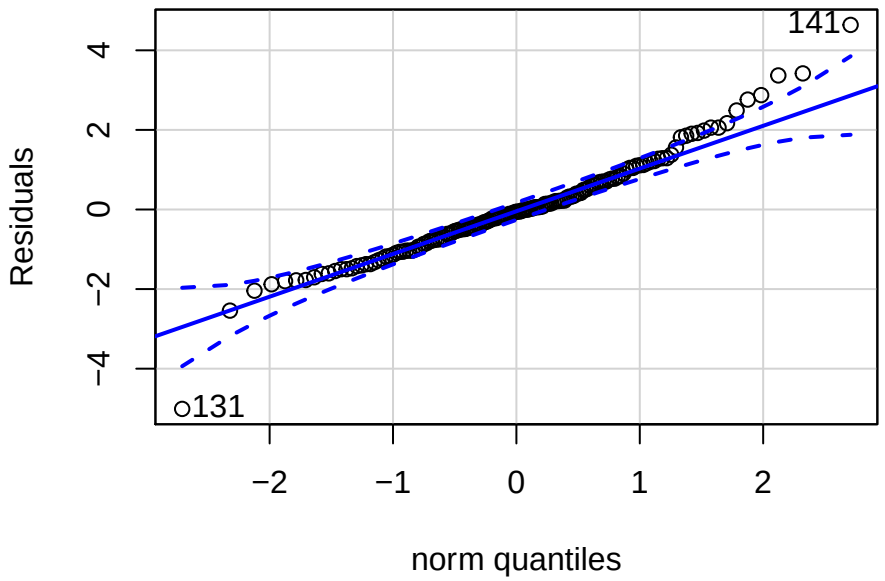
```r
# Linearity
plot(fish2$Height, residuals(model4), xlab = "Total Length", ylab = "Residuals",
    main = "Total Length vs Residuals")
abline(0, 0, col = "red")
```




```r
# Constant Variance
plot(model4$fitted.values, model4$residuals, xlab = "Fitted Values", ylab = "Residuals")
```

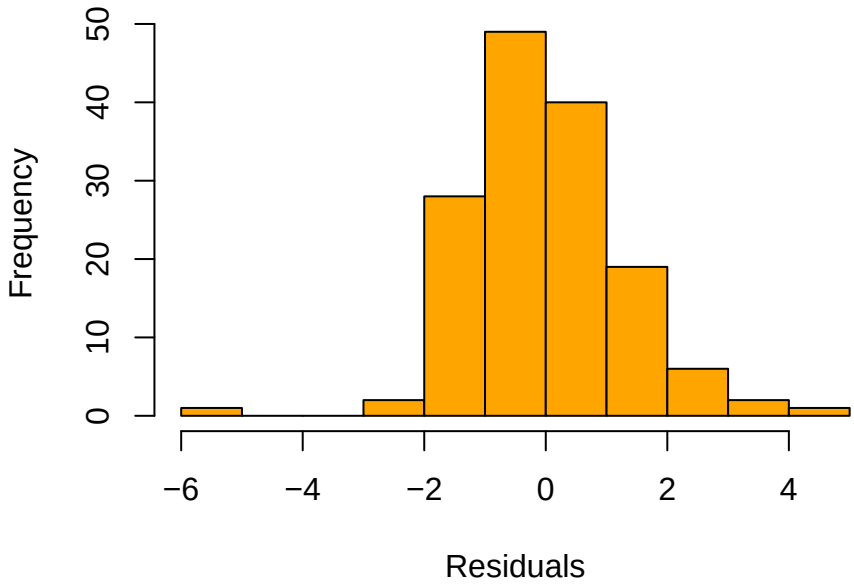


```r
# Normality Assumption
qqPlot(model4$residuals, ylab = "Residuals", main = "")
```




```
## 131 141
## 130 140
```

```r
hist(model4$residuals, xlab = "Residuals", main = "", nclass = 10, col = "orange")
```



The transformation was successful as all 3 assumptions are not violated: * Linearity - residuals are scattered around 0 * Constant Variance - the parabolic curve is gone and data looks to be randomly scattered * Normality - right tail of the QQ plot has less curvature than that of model 3 and is within the outlined boundaries.

Question 8: Model Comparison [3pts]

(a) Using each model summary, compare and discuss the R-squared and Adjusted R-squared of model2, model3, and model4.

```
print("Model 2")
```

```
## [1] "Model 2"
```

```
print(paste0("r-squared: ", summary(model2)$r.squared))
```

```
## [1] "r-squared: 0.938483625001571"
```

```
print(paste0("adj r-squared: ", summary(model2)$adj.r.squared))
```

```
## [1] "adj r-squared: 0.933508035847286"
```

```
print("Model 3")
```

```
## [1] "Model 3"
```

```r
print(paste0("r-squared: ", summary(model3)$r.squared))
```

```
## [1] "r-squared: 0.935294615139321"
```

```r
print(paste0("adj r-squared: ", summary(model3)$adj.r.squared))
```

```
## [1] "adj r-squared: 0.932059345896287"
```

```r
print("Model 4")
```

```
## [1] "Model 4"
```

```r
print(paste0("r-squared: ", summary(model4)$r.squared))
```

```
## [1] "r-squared: 0.981713834446085"
```

```r
print(paste0("adj r-squared: ", summary(model4)$adj.r.squared))
```

```
## [1] "adj r-squared: 0.980799526168389"
```

Reducing the number of predictors from Model 2 to 3 decreased the r-squared and adjusted r-squared. Transforming model 3 into model 4 using the square root transformation increased the r-squared and adjusted r-squared significantly.

Question 9: Estimation and Prediction [10 points]

(a) Estimate Weight for the last 10 rows of data (fishtest) using both model3 and model4. Compare and discuss the mean squared prediction error (MSPE) of both models.

```r
predict3 <- predict(model3, fishtest[, -1], interval = "prediction")
predict4 <- predict(model4, fishtest[, -1], interval = "prediction")

mean((predict3[, 1] - fishtest[, 1])^2)
```

```
## [1] 9392.25
```

```r
mean((predict4[, 1]^2 - fishtest[, 1])^2)
```

```
## [1] 2442.998
```

The mean squared prediction error for model 4 is significantly less than that of model 3. This is likely the case due the smaller scale of model 4 where the predicted variable is transformed using the square root function.

(b) Suppose you have found a Perch fish with a Body.Height of 28 cm, and a Total.Length of 32 cm. Using model4, predict the weight on this fish with a 90% prediction interval. Provide an interpretation of the prediction interval.

```r
# model4 <- lm(sqrt(Weight) ~ Species + Total.Length, data = fish2)
df <- data.frame("Perch", 32)
colnames(df) <- c("Species", "Total.Length")

new_predict4 <- predict(model4, df, interval = "prediction", level = 0.9)
new_predict4^2  ## Squaring the resuls to 'untransform' the results
```

```
##        fit      lwr      upr
## 1 461.9429 374.4536 558.6091
```

The 90% prediction interval of the weight of a Perch fish with a total length of 32 cm (body height is not used in model 4) is between 374 and 559 grams. The approximate estimate is 462 grams.