HW3_Part 2 - MLR

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)
```

```
## Loading required package: carData
```

```
library(ggplot2)
library("PerformanceAnalytics")
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

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

```
##
## Attaching package: 'PerformanceAnalytics'
## The following object is masked from 'package:graphics':
## 
##     legend
```

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(MASS)
# Read the data set
fishfull = read.csv("Fish.csv",header=T, fileEncoding = 'UTF-8-BOM')
head(fishfull)
```

```
##   Weight Species Body.Height Total.Length Diagonal.Length  Height  Width
## 1    300    Pike        34.8         37.3            39.8  6.2884 4.0198
## 2    242   Bream        23.2         25.4            30.0 11.5200 4.0200
## 3    500   Bream        29.1         31.5            36.4 13.7592 4.3680
## 4    600   Bream        29.4         32.0            37.2 15.4380 5.5800
## 5    345    Pike        36.0         38.5            41.0  6.3960 3.9770
## 6   1000   Perch        40.2         43.5            46.0 12.6040 8.1420
```

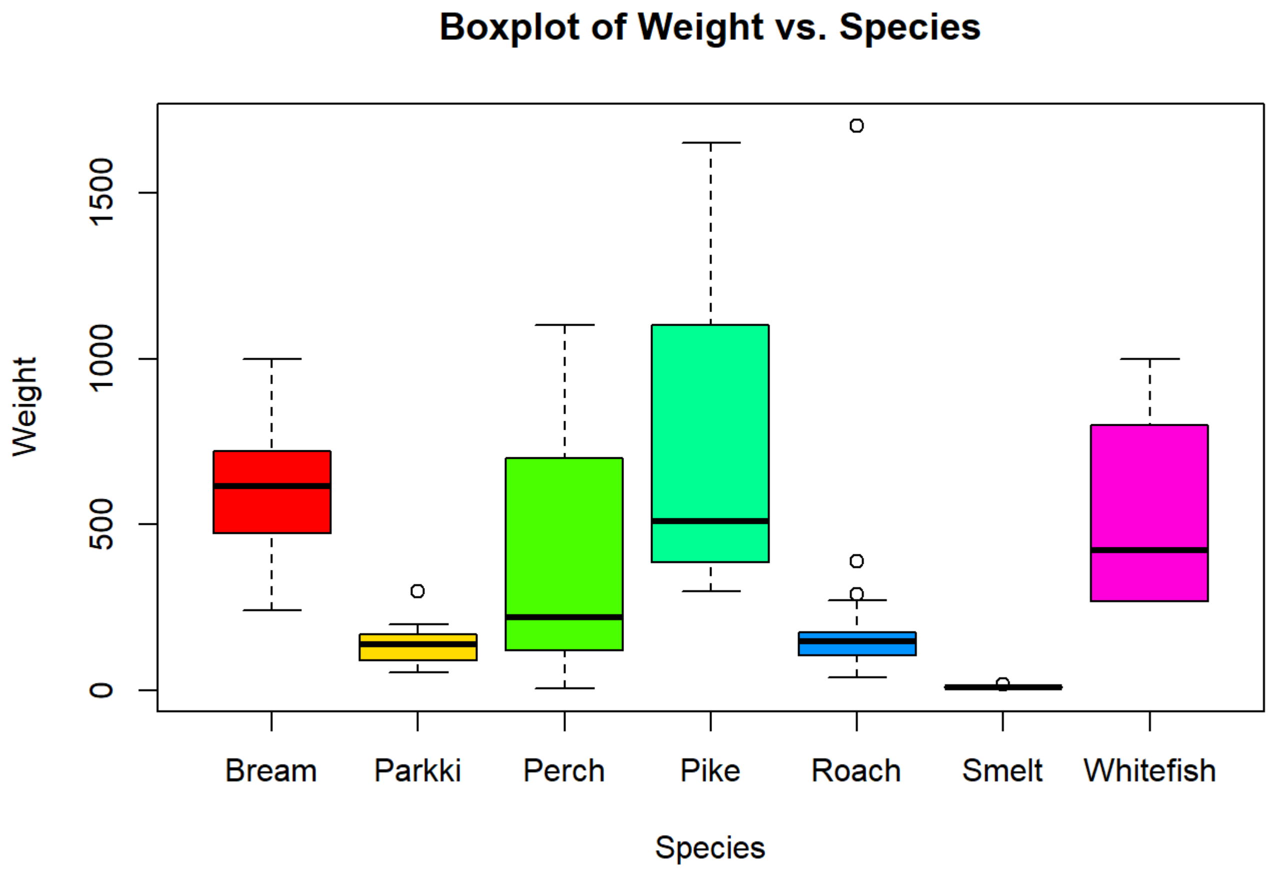
```
row.cnt = nrow(fishfull)
# Split the data into training and testing sets
fishtest = fishfull[(row.cnt-9):row.cnt,]
fish = fishfull[1:(row.cnt-10),]
```

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?

```
fish$Species = as.factor((fish$Species))
boxplot(Weight ~ Species, data = fish, main = "Boxplot of Weight vs. Species", col= rainbow(7))
```



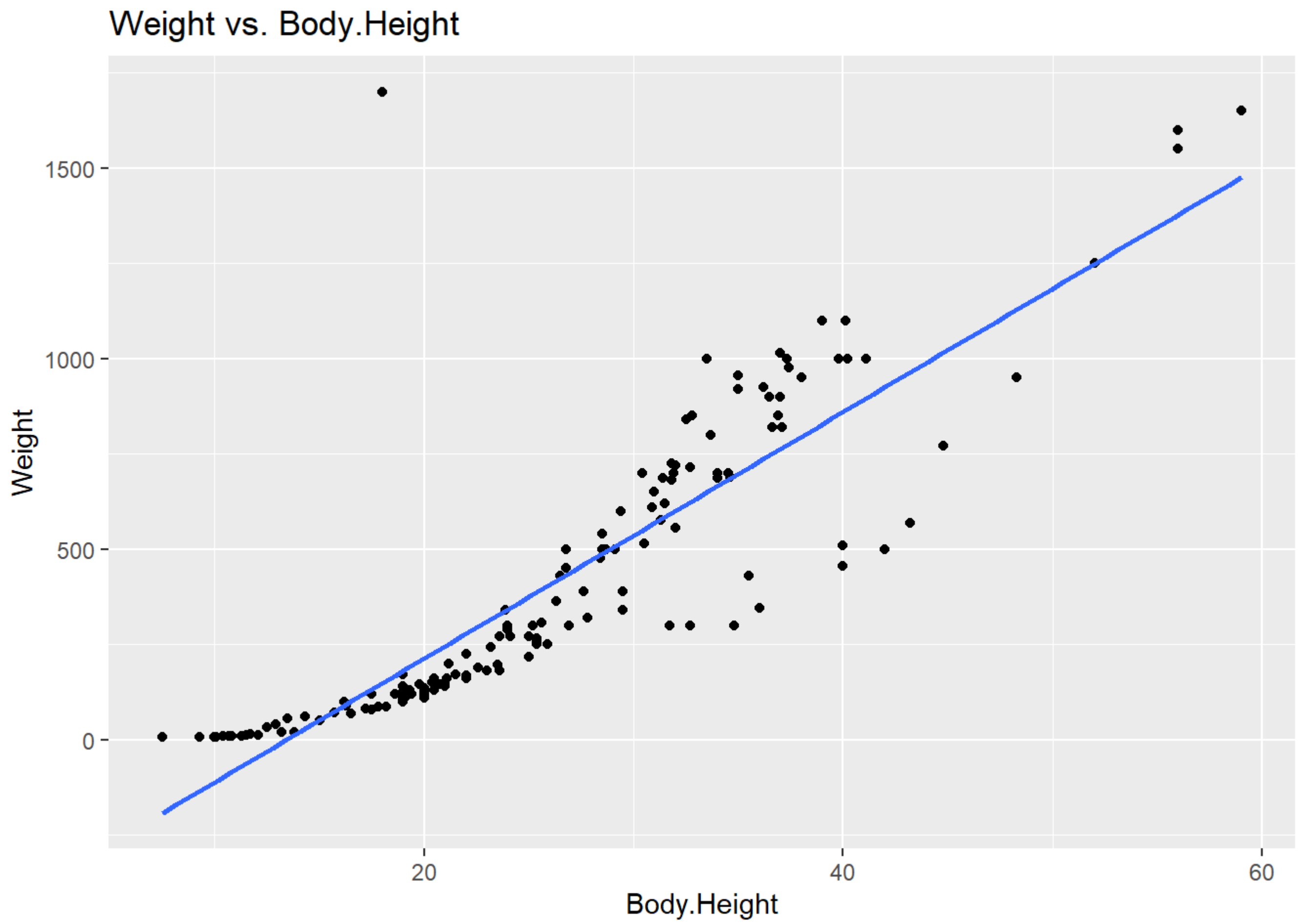
Answer

Based on the Boxplot above, it appears to be a relationship between the predicting variable $Species$ and the response variable $Weight$ due to the fact that the weight mean seems to be different across the different species. Also observe the variability within each of the species is different.

(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?**

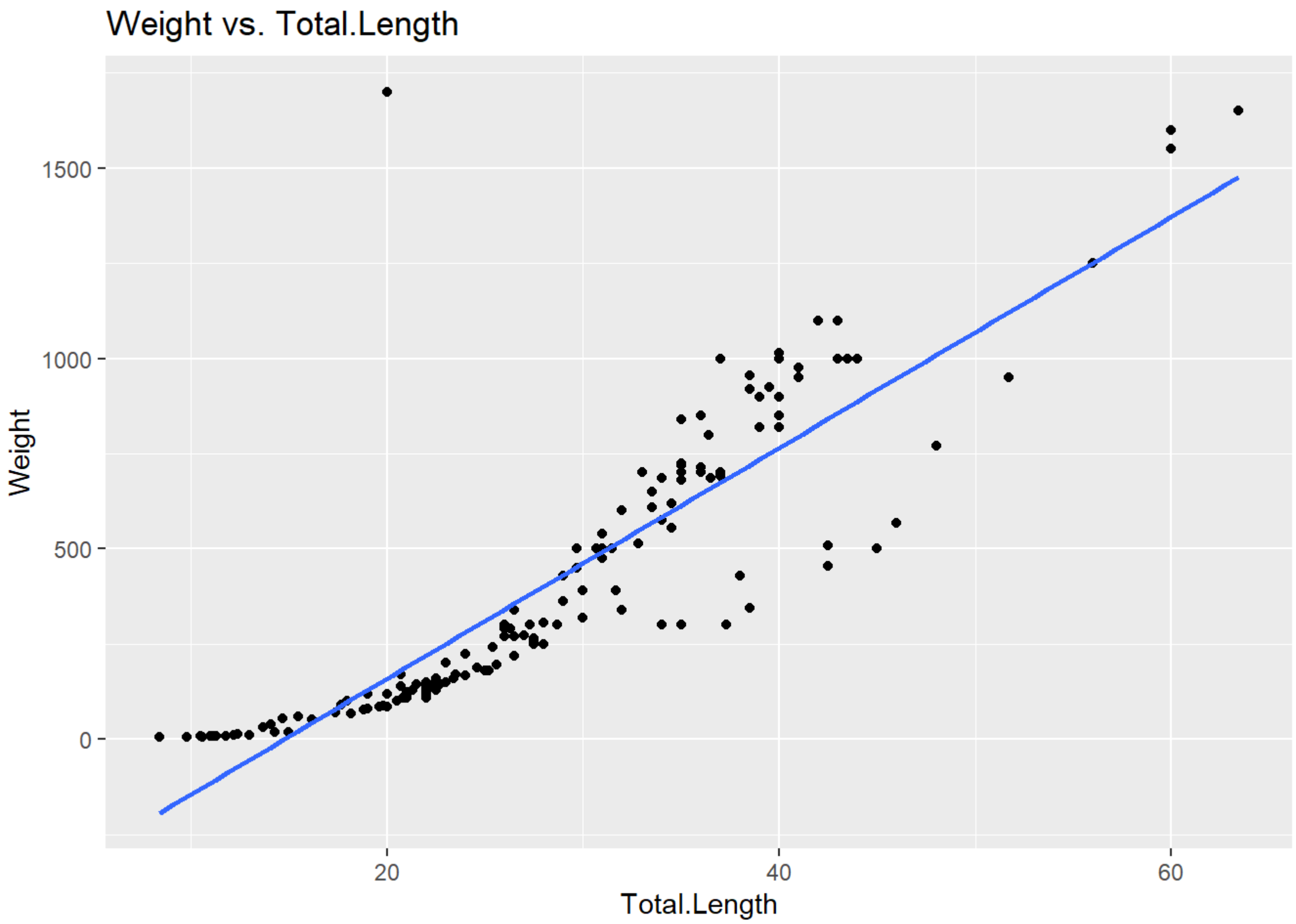
```
ggplot(fish, aes(x=Body.Height, y=Weight)) + geom_point() + ggtitle("Weight vs. Body.Height") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

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



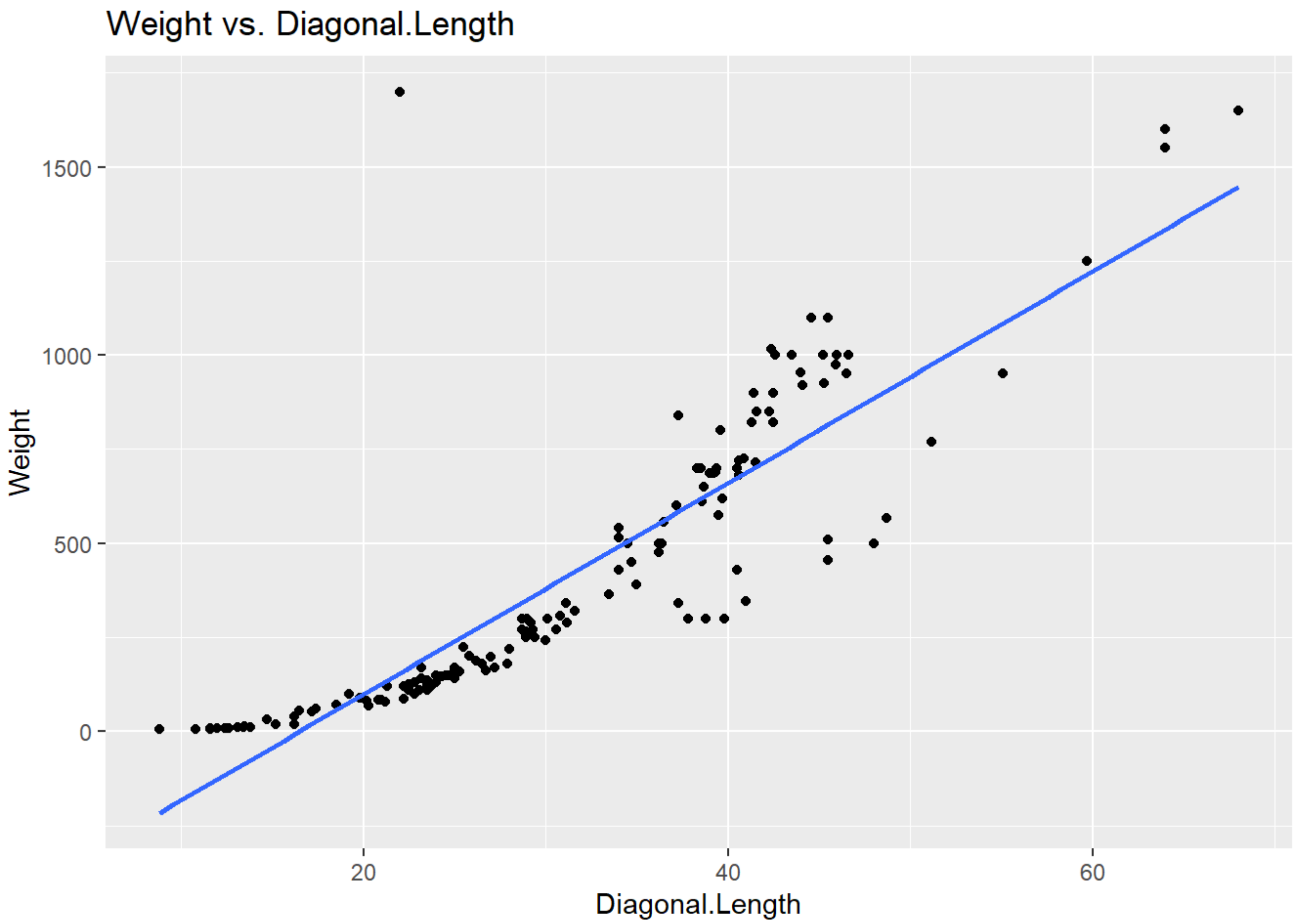
```
ggplot(fish, aes(x=Total.Length, y=Weight)) + geom_point() + ggtitle("Weight vs. Total.Length")
 +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

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



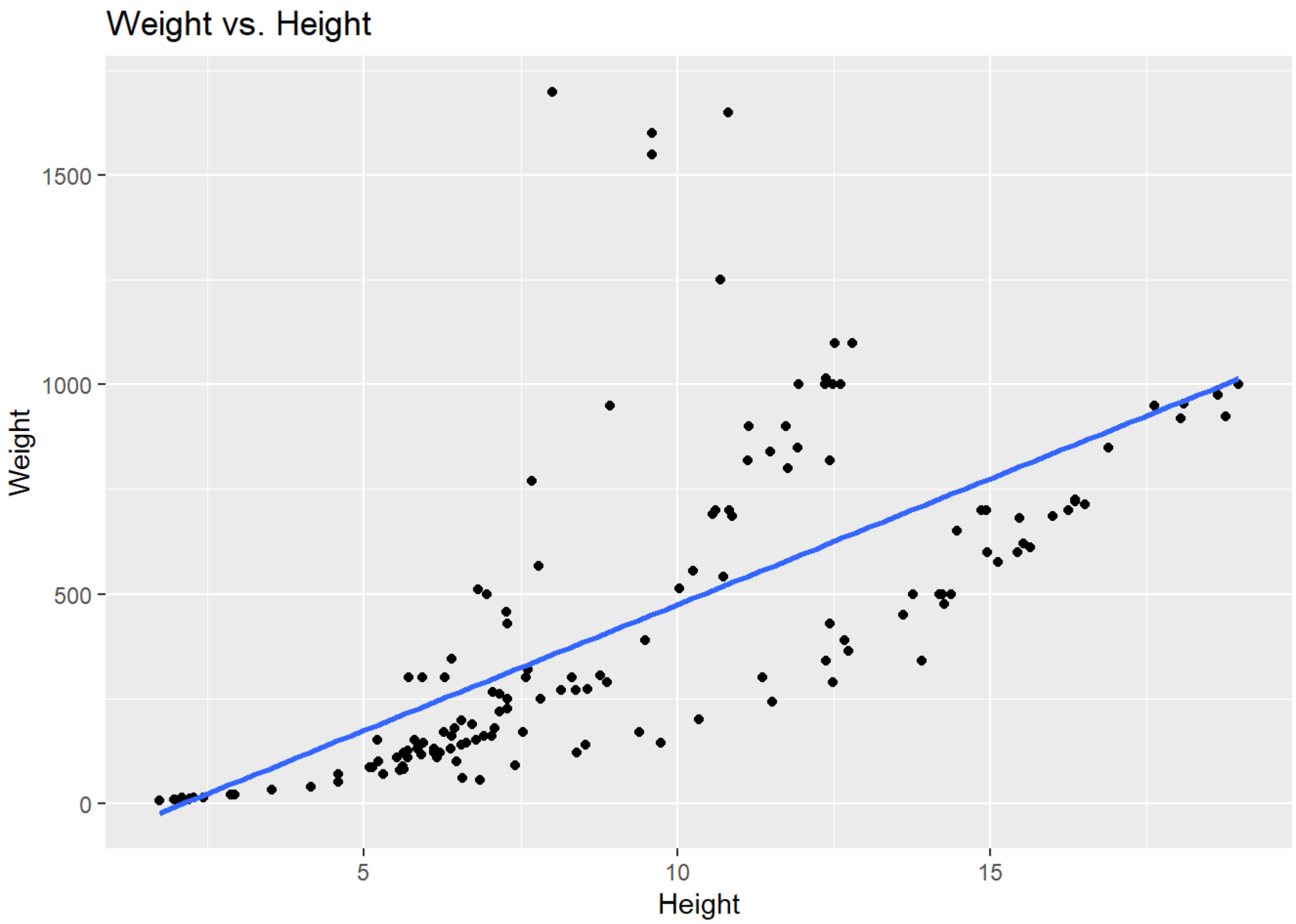
```
ggplot(fish, aes(x=Diagonal.Length, y=Weight)) + geom_point() + ggtitle("Weight vs. Diagonal.Len
gth") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

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



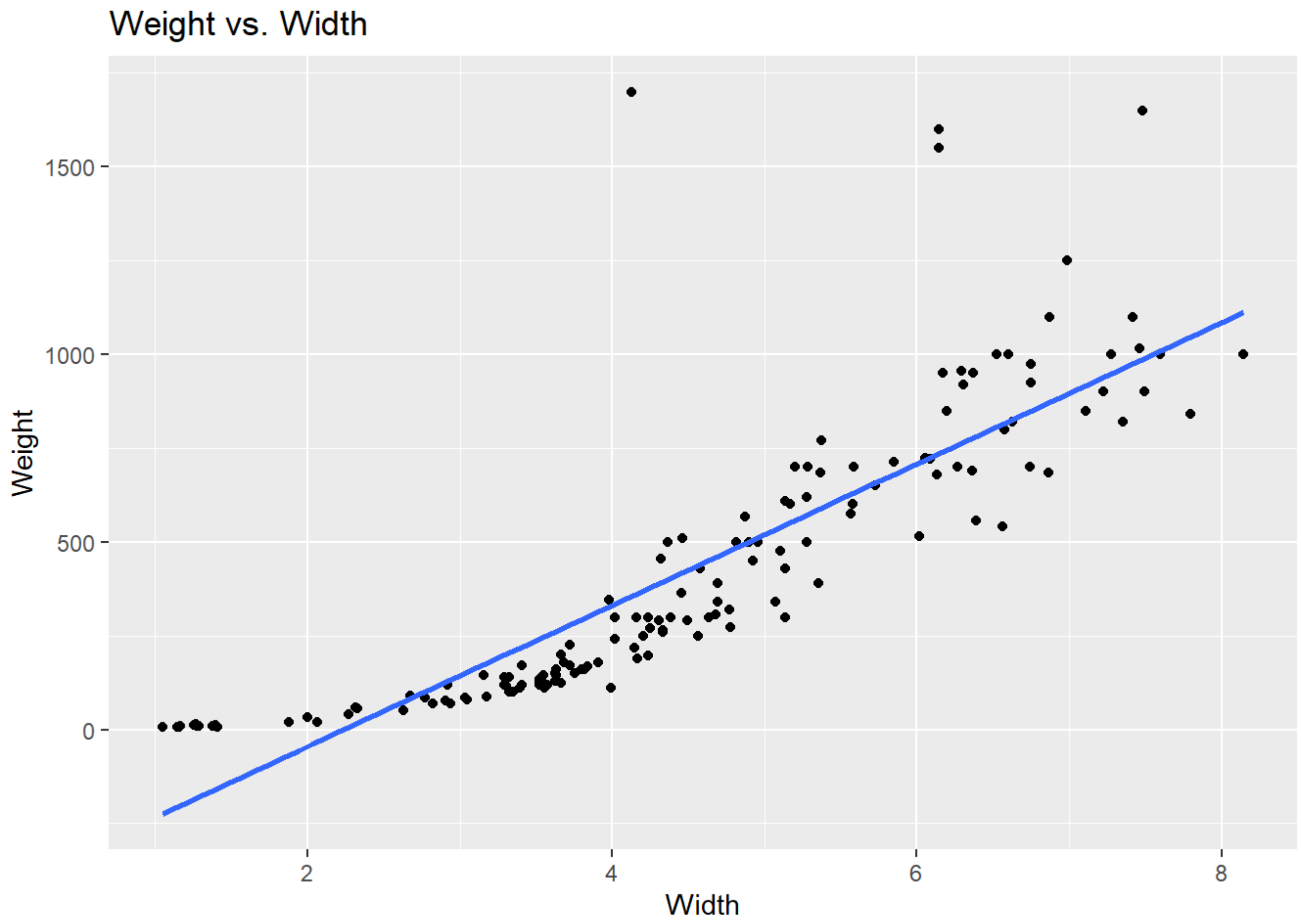
```
ggplot(fish, aes(x=Height, y=Weight)) + geom_point() + ggtitle("Weight vs. Height") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

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



```
ggplot(fish, aes(x=Width, y=Weight)) + geom_point() + ggtitle("Weight vs. Width") +
  scale_colour_hue(l=50) + # Use a slightly darker palette than normal
  geom_smooth(method="lm",   # Add linear regression line
              se=FALSE,    # Don't add shaded confidence region
              fullrange=TRUE) # Extend regression lines
```

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

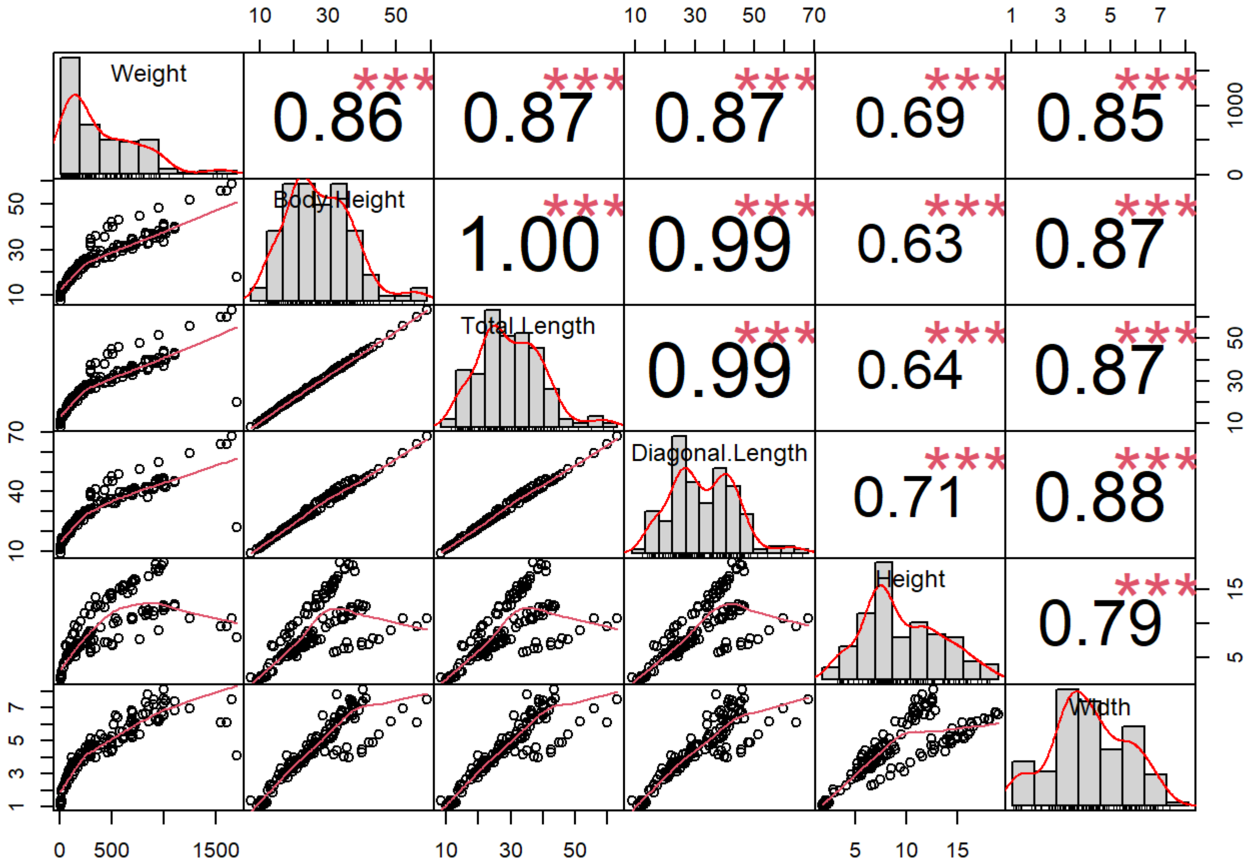


Answer

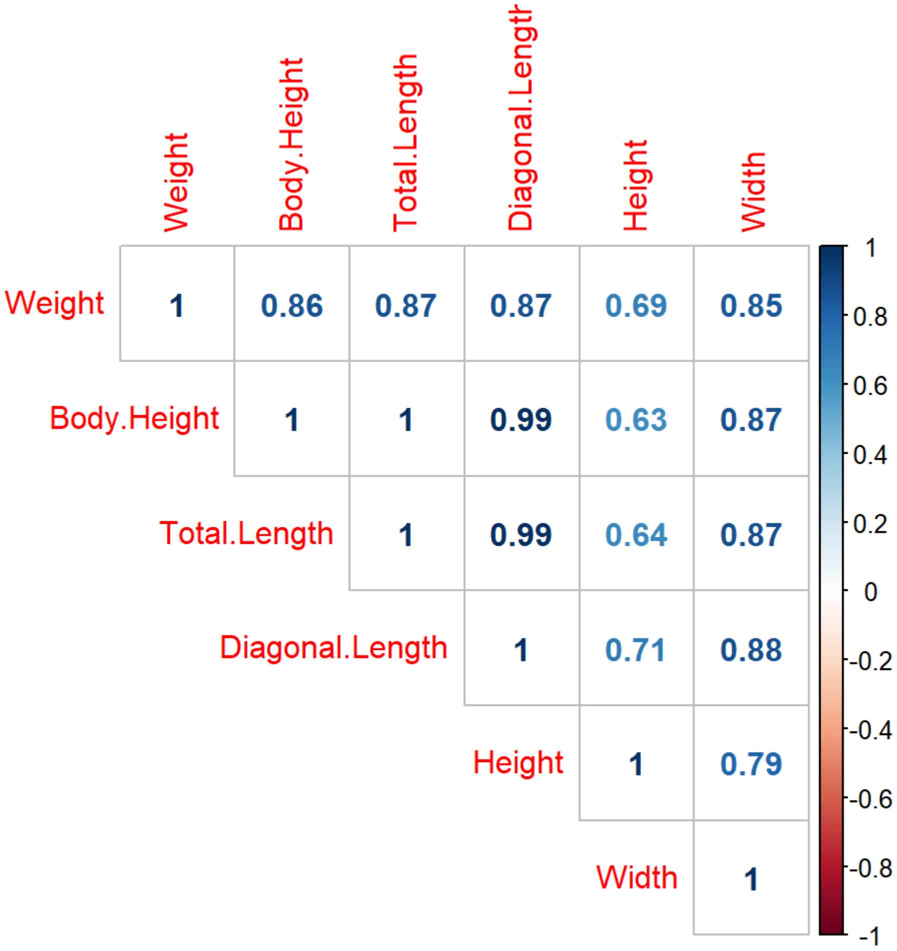
We can see there is a direct relationship between all the predictors and the response $Weight$, meaning that as the predictor increases, there is an increase in Weight. While still somewhat strong direct relationship, it appears that $Height$ has the weakest relationship with $Weight$ compared to the other predictors. Looking at the plot of $Body.Height$ we can see there might be one outlier. The $Total.Length$ and $Diagonal.Length$ plots also indicate of a potential outlier. The plots of $Height$ and $Width$ indicate of few (between ~4 to ~5) potential outliers.

(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.

```
# Exclude categorical variable
chart.Correlation(fish[,-2], histogram=TRUE, pch=25)
```



```
# Another Viz.
corrplot(cor(fish[,-2]),
  method = "number",
  type = "upper" # show only upper side
)
```



	Weight	Body.Height	Total.Length	Diagonal.Length	Height	Width
Weight	1	0.86	0.87	0.87	0.69	0.85
Body.Height		1	1	0.99	0.63	0.87
Total.Length			1	0.99	0.64	0.87
Diagonal.Length				1	0.71	0.88
Height					1	0.79
Width						1

Answer

Observe above the correlation coefficients between the response $Weight$ and each of the quantitative predictor variables. The correlation coefficient between $Weight$ and $Body.Height$ is 0.86, the correlation coefficient between $Weight$ and $Total.Length$ and $Weight$ and $Diagonal.Length$ is 0.87, the correlation coefficient between $Weight$ and $Heigh$ is 0.69 and 0.85 between $Weight$ and $Width$. Hence, we conclude there is a strong direct relationship between each predictor and the response.

Multicollinearity generally occurs when there are high correlations between two or more predictor variables. In other words, one predictor variable can be used to predict the other. Multicollinearity can cause many problems in the model and its interpretation. Based on the correlation coefficient between each pair of predictor variables, we can see that all of the predictor variables are significantly correlated (indicated by the 3 red stars). For example, the correlation coefficient between $Total.Length$ and $Diagonal.Length$ is 0.99 and $Body.Height$ and $Diagonal.Legth$ is also 0.99. This indicates multicollinearity and a potential problem if we were to model a multiple linear regression with the original 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?

Answer

As indicated above, there seems to be multicollinearity, suggesting that it is not appropriate to model a multiple linear regression between $Weight$ and the predictor variables. If we are to regress $Weight$ on the above predictor variables, we risk that, the estimated coefficients βs will be unstable,

the standard error of the estimated coefficients βs will be artificially large,

the overall F-statistic will be significant, but individual t-statistic of the estimated coefficients will not, and

the prediction will not be accurate.

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
```

Answer

See above model1 output.

(b) Is the overall regression significant at an α level of 0.01?

Answer

Based on the overall F-statistic’s p-value of 2.2e-16, which is smaller than 0.01, we conclude the overall regression is significant.

(c) What is the coefficient estimate for *Body.Height*? Interpret this coefficient.

Answer

The coefficient estimate for $\beta_{Body.Height}$ is -176.87, meaning that as Body.Height increases by 1 cm, fish’s weight decreases by 176.87 grams, ***given the other predictors and holding them fixed.***

(d) What is the coefficient estimate for the *Species* category Parkki? Interpret this coefficient.

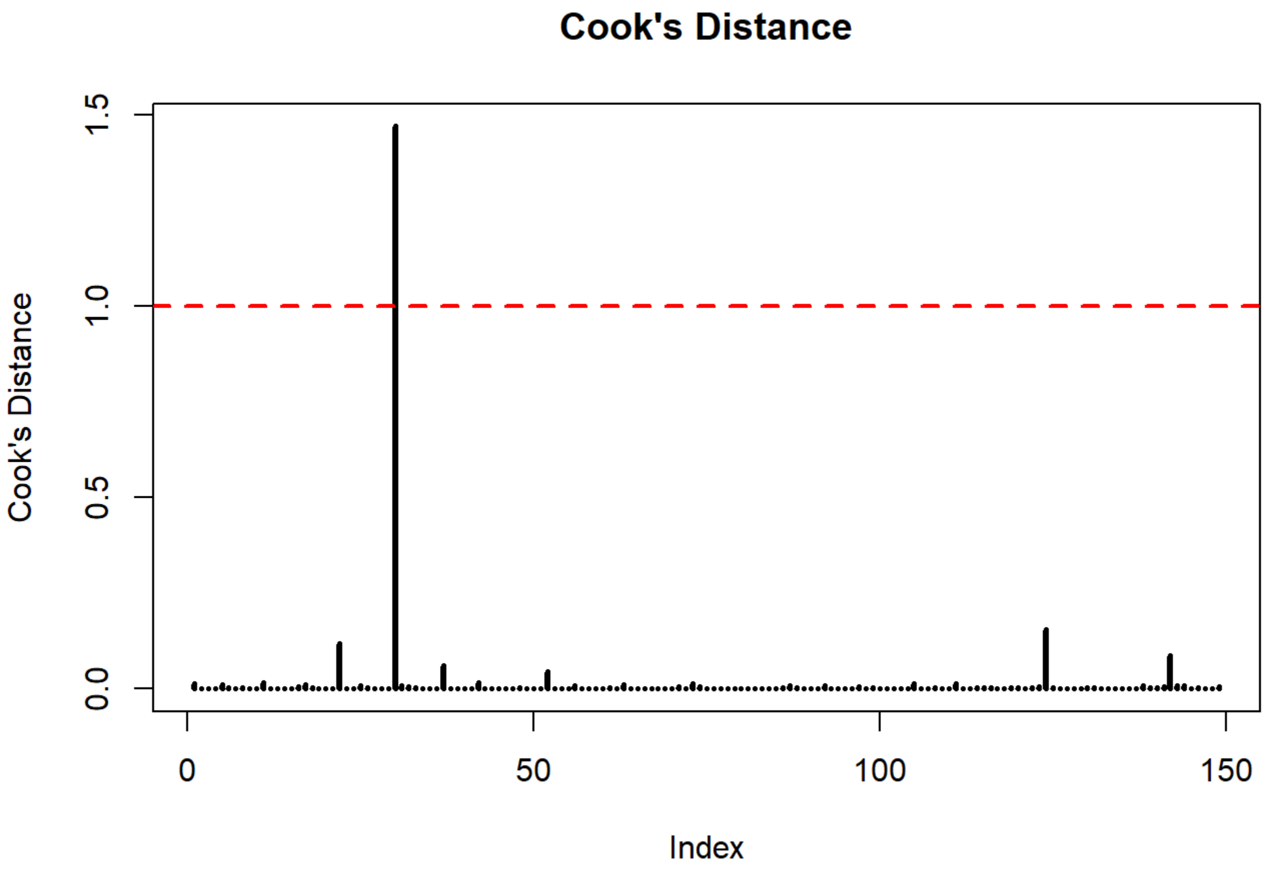
Answer

The coefficient estimate for $\beta_{SpeciesParkki}$ is 79.34. The $Species$ base case is Bream and is incorporated in the intercept estimation of -813.90, interpreted as the average fish weight among Bream, given the other predictors. The coefficient estimate for $\beta_{SpeciesParkki}$ represents the average ***difference*** between the base case Bream and Parkki. In other words, the average fish weight among Parkki is -813.90 + 79.34 = -734.56, ***given all other predictors and holding them fixed.***

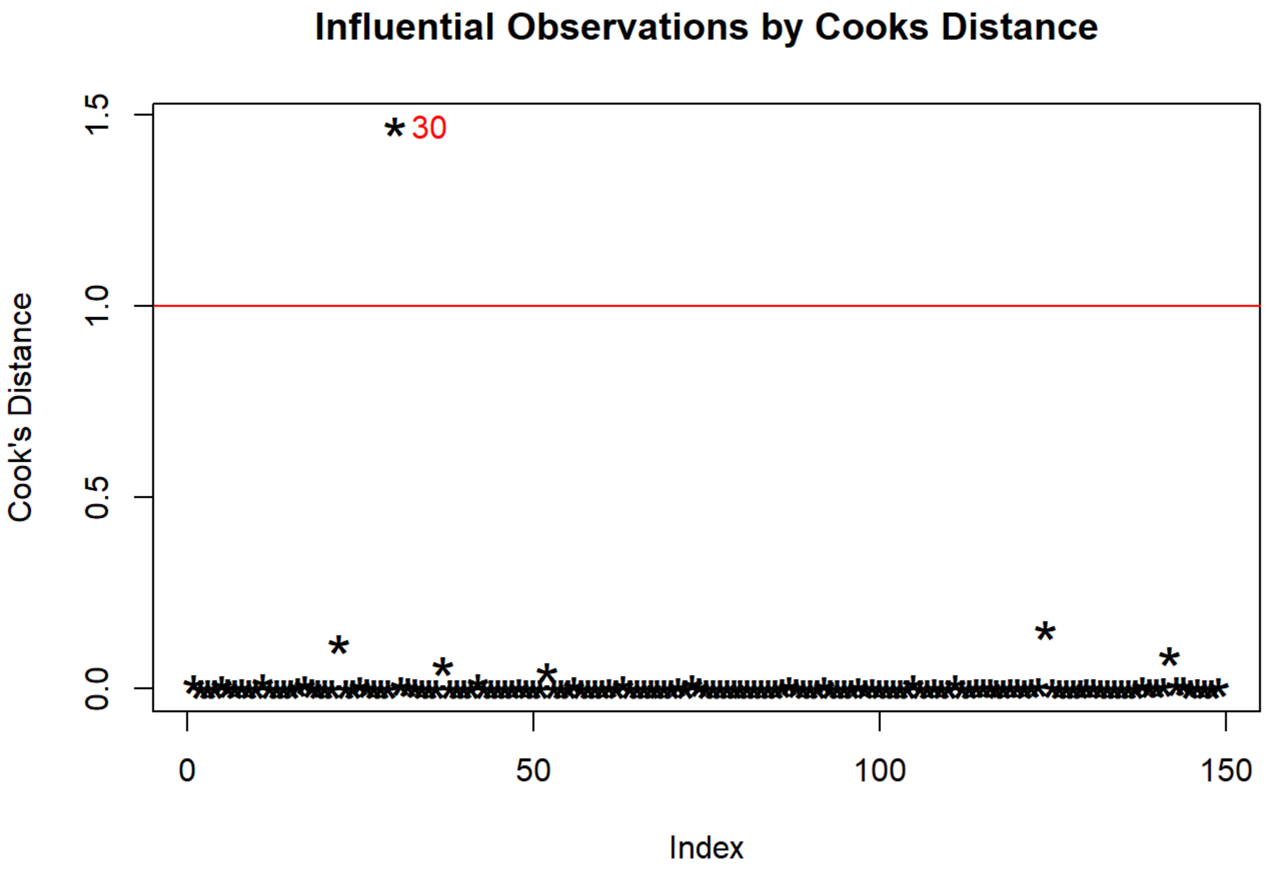
Question 3: Checking for Outliers and Multicollinearity [9 points]

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

```
# Cook’s Distance
cook = cooks.distance(model1)
plot(cook,
     type="h",
     lwd=3,
     ylab = "Cook's Distance",
     main="Cook's Distance")

abline(1, 0,
       col="red",
       lty=2, lwd=2)
```



```
plot(cook, pch="*", cex=2, ylab= "Cook's Distance", main="Influential Observations by Cooks Dist
ance")  # plot cook's distance
abline(h = 1, col="red")  # add cutoff line
text(x=1:length(cook)+5, y=cook, labels=ifelse(cook > 1, names(cook),""), col="red")  # add labe
ls
```



Answer

Observe in the plot above that based on Cook’s Distance and a threshold of 1, observation indexed 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[-30,]
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
```

Answer

See above model2 output.

(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?

```
round(vif(model2),3)
```

```
##                     GVIF Df GVIF^(1/(2*Df))
## Species         1545.550  6           1.844
## Body.Height     2371.154  1          48.694
## Total.Length    4540.477  1          67.383
## Diagonal.Length 2126.650  1          46.116
## Height            56.214  1           7.498
## Width             29.017  1           5.387
```

```
r_2 = 0.9385
thresh = 1/(1-r_2)
thresh
```

```
## [1] 16.26016
```

Answer

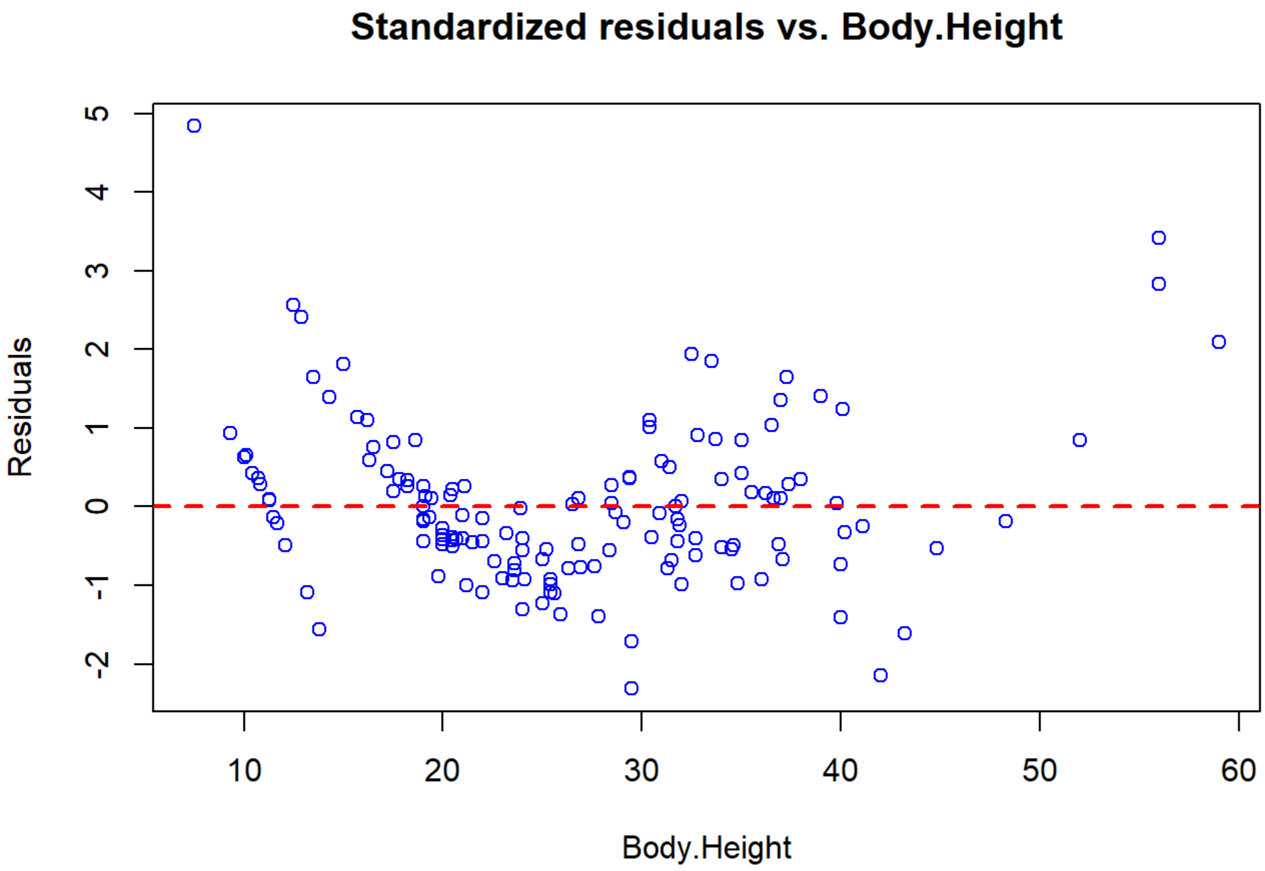
Using a VIF threshold of MAX(10, 1/(1-R^2_{model}), all predictors’ VIF is greater than the threshold, indicating a problem of multicollinearity.

Question 4: Checking Model Assumptions [9 points]

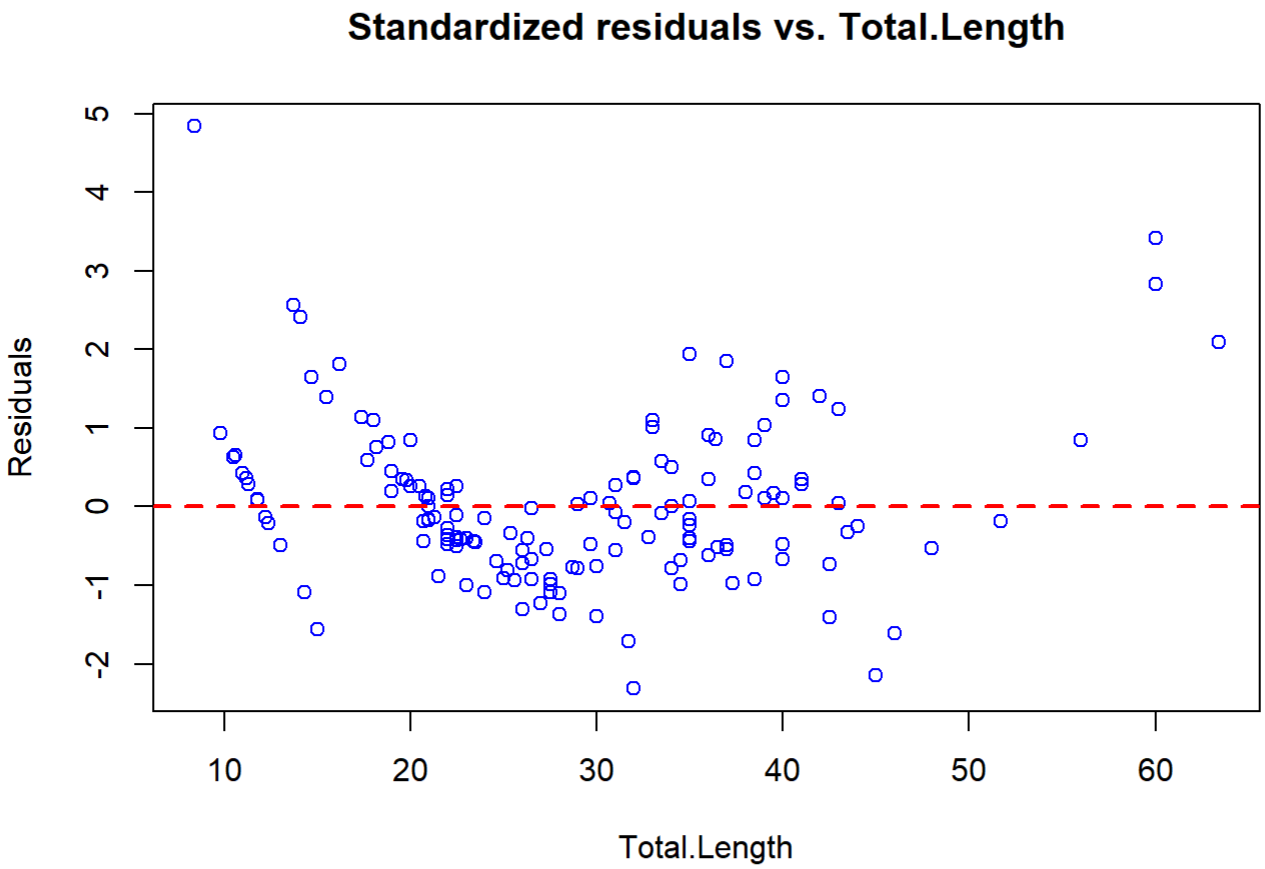
Please use the cleaned data set, which have the outlier(s) removed, and model2 for answering the following questions.

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

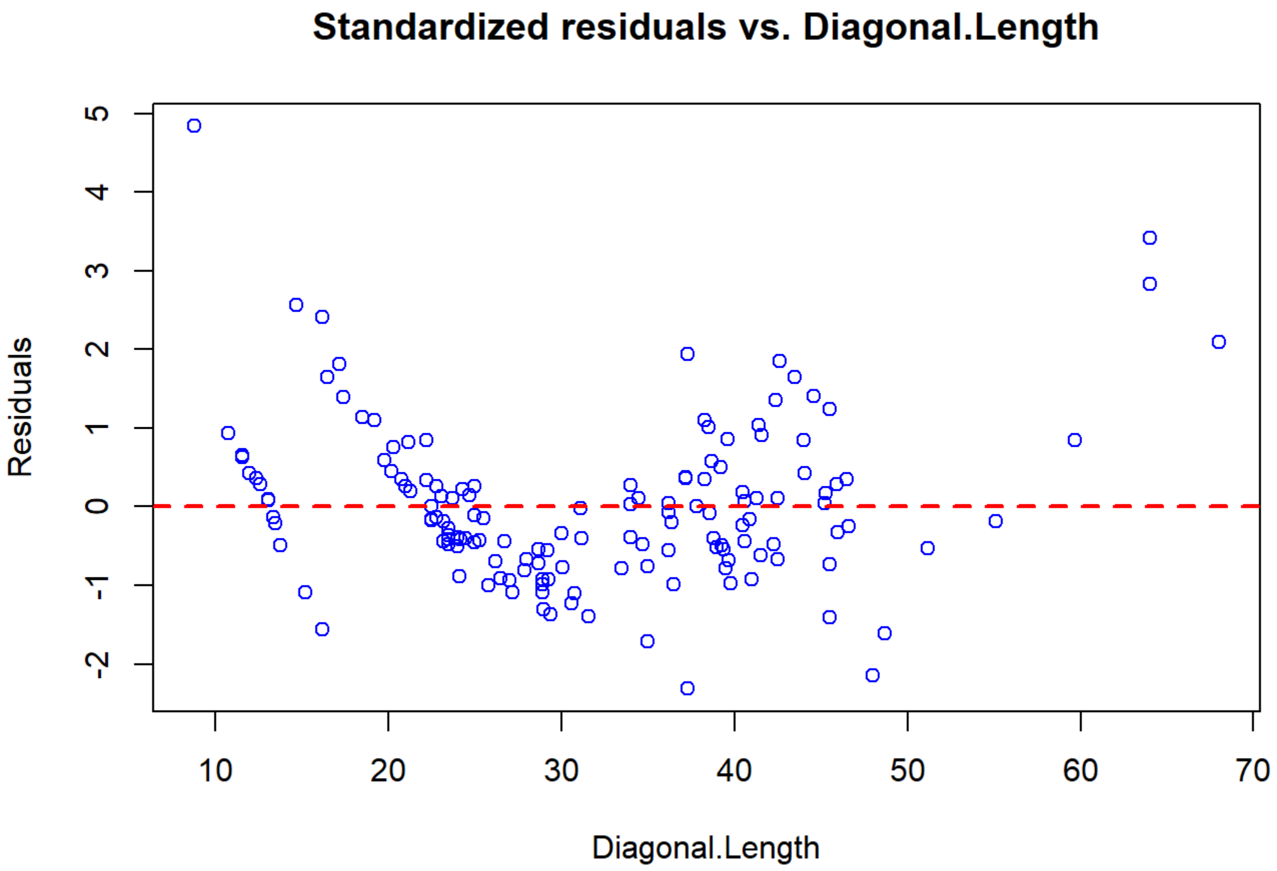
```
# Standardized residuals
resids = rstandard(model2)


# Plot the standardized residuals against
# Body.Height
plot(fish2$Body.Height, resids,
     xlab="Body.Height",
     ylab="Residuals",
     main="Standardized residuals vs. Body.Height",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



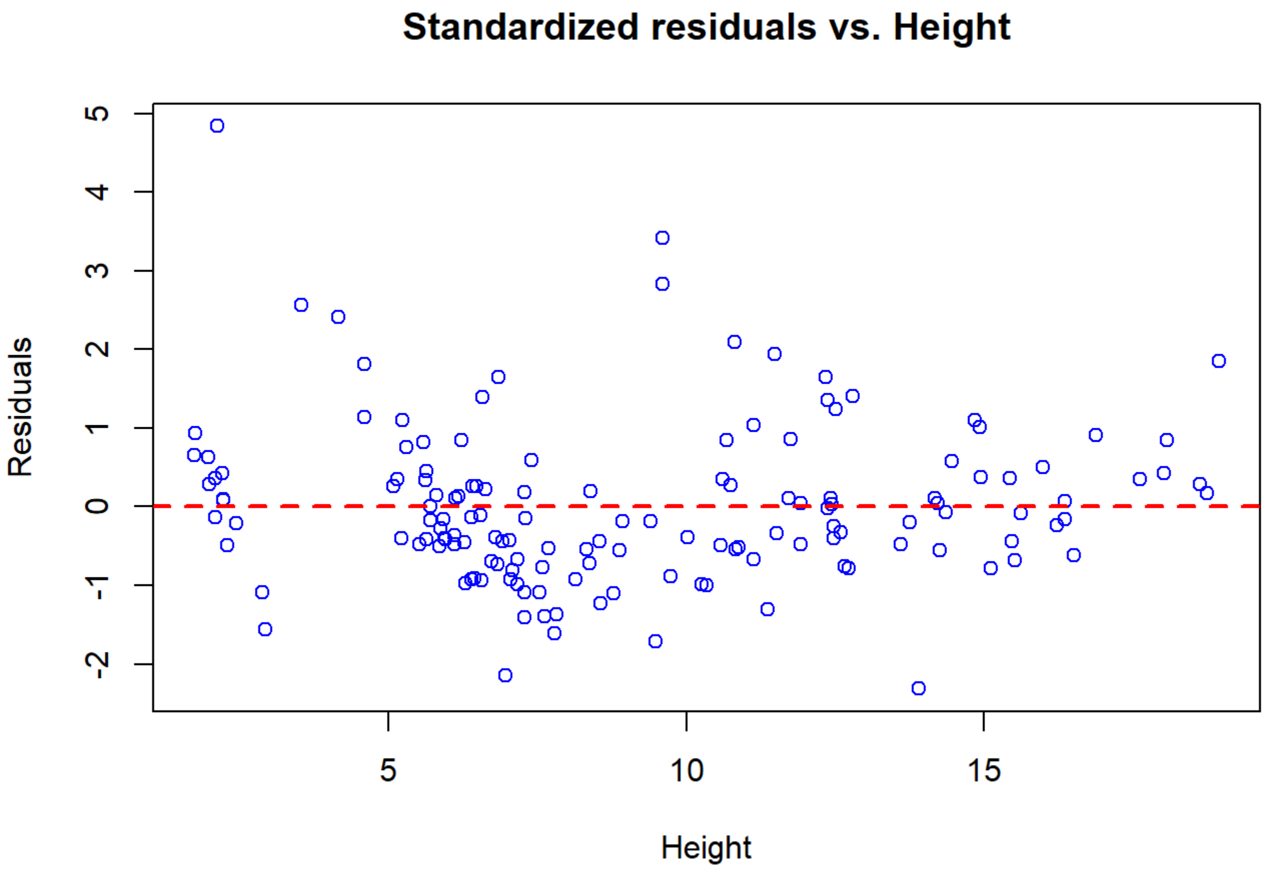
```
# Total.Length
plot(fish2$Total.Length, resids,
     xlab="Total.Length",
     ylab="Residuals",
     main="Standardized residuals vs. Total.Length",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



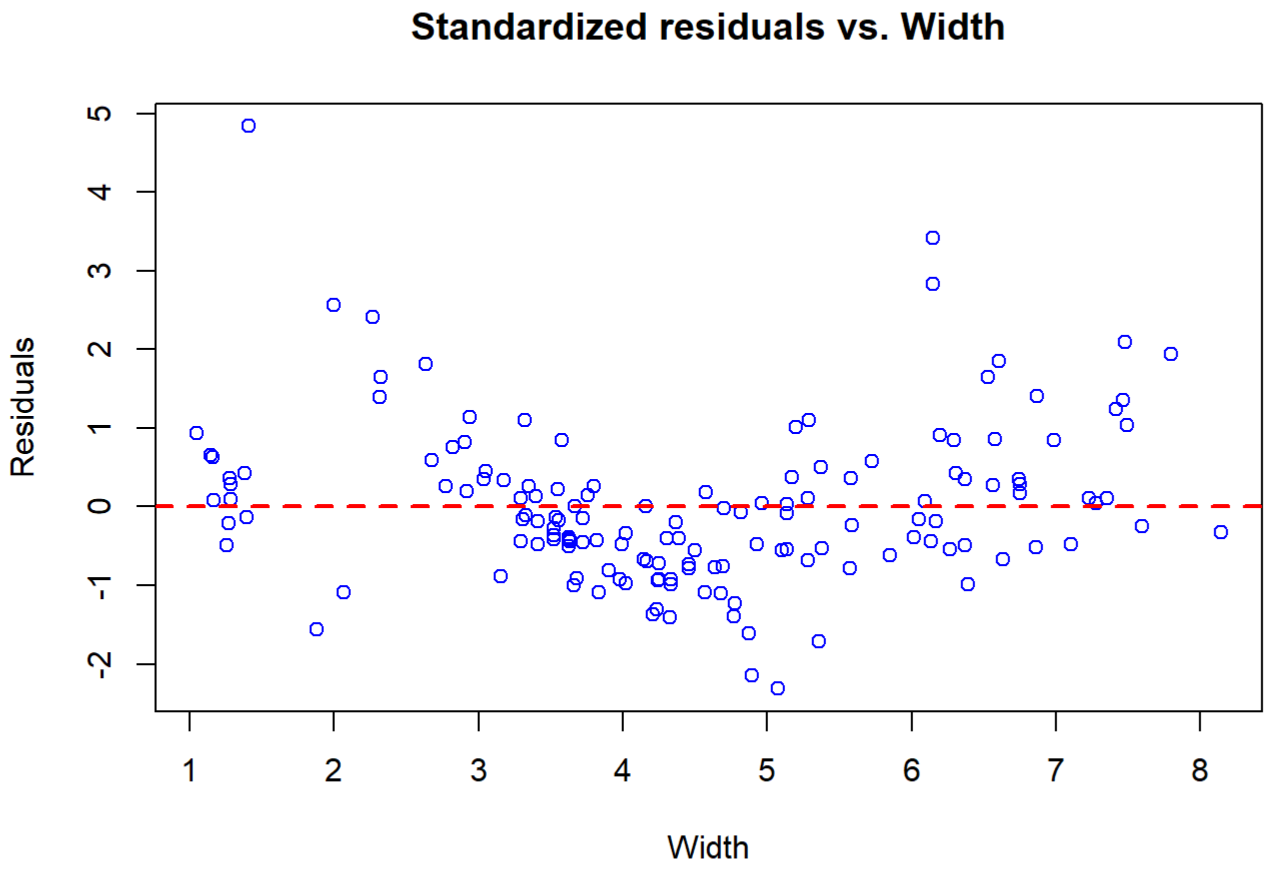
```
# Diagonal.Length
plot(fish2$Diagonal.Length, resids,
     xlab="Diagonal.Length",
     ylab="Residuals",
     main="Standardized residuals vs. Diagonal.Length",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



```
# Height
plot(fish2$Height, resids,
     xlab="Height",
     ylab="Residuals",
     main="Standardized residuals vs. Height",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



```
# Width
plot(fish2$Width, resids,
     xlab="Width",
     ylab="Residuals",
     main="Standardized residuals vs. Width",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```

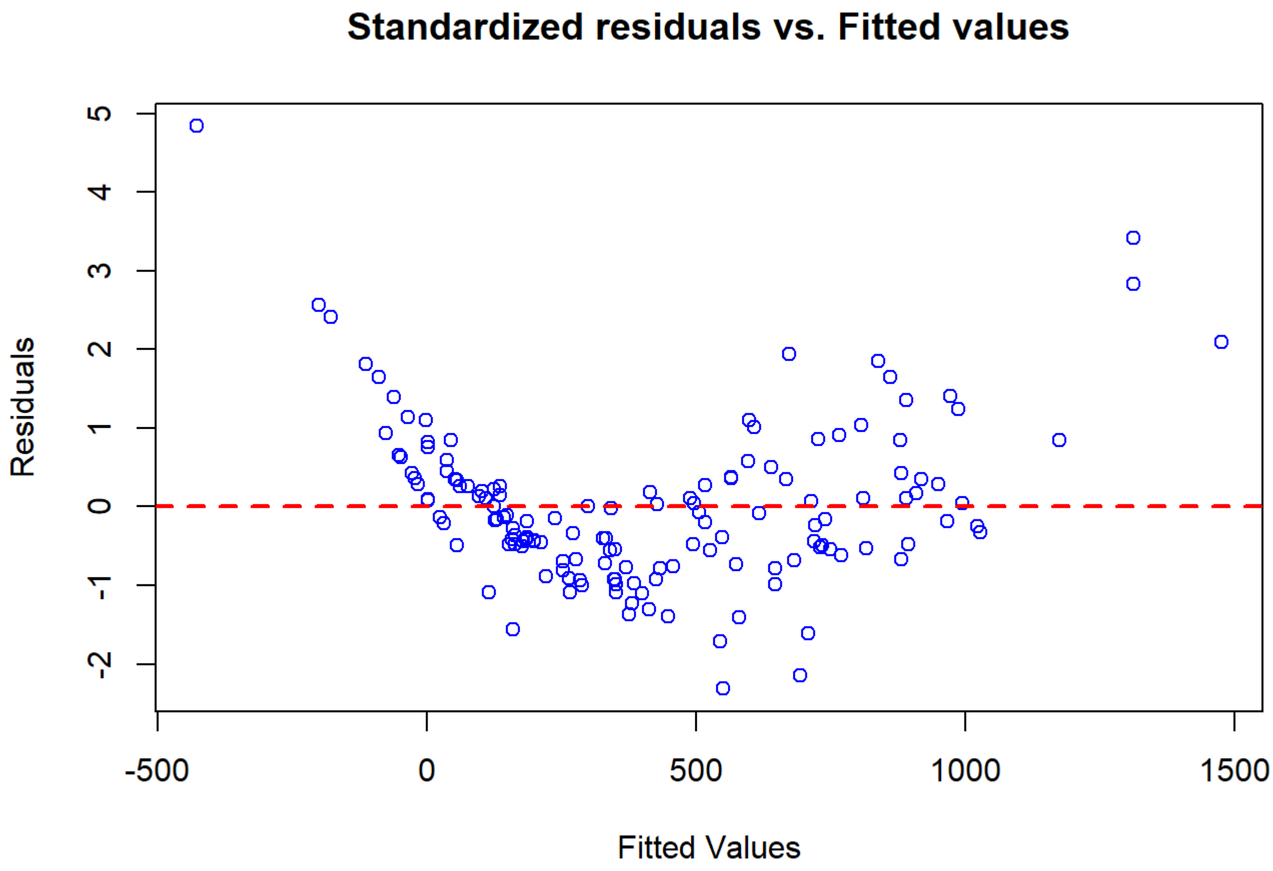


Answer

Based on the plots above, since there is a random pattern around the 0 mean line, we conclude the linearity assumption holds for all predicting variables.

(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?

```
fits = model2$fitted

# Plot the standardized residuals against
# fitted values
plot(fits, resids,
     xlab="Fitted Values",
     ylab="Residuals",
     main="Standardized residuals vs. Fitted values",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```

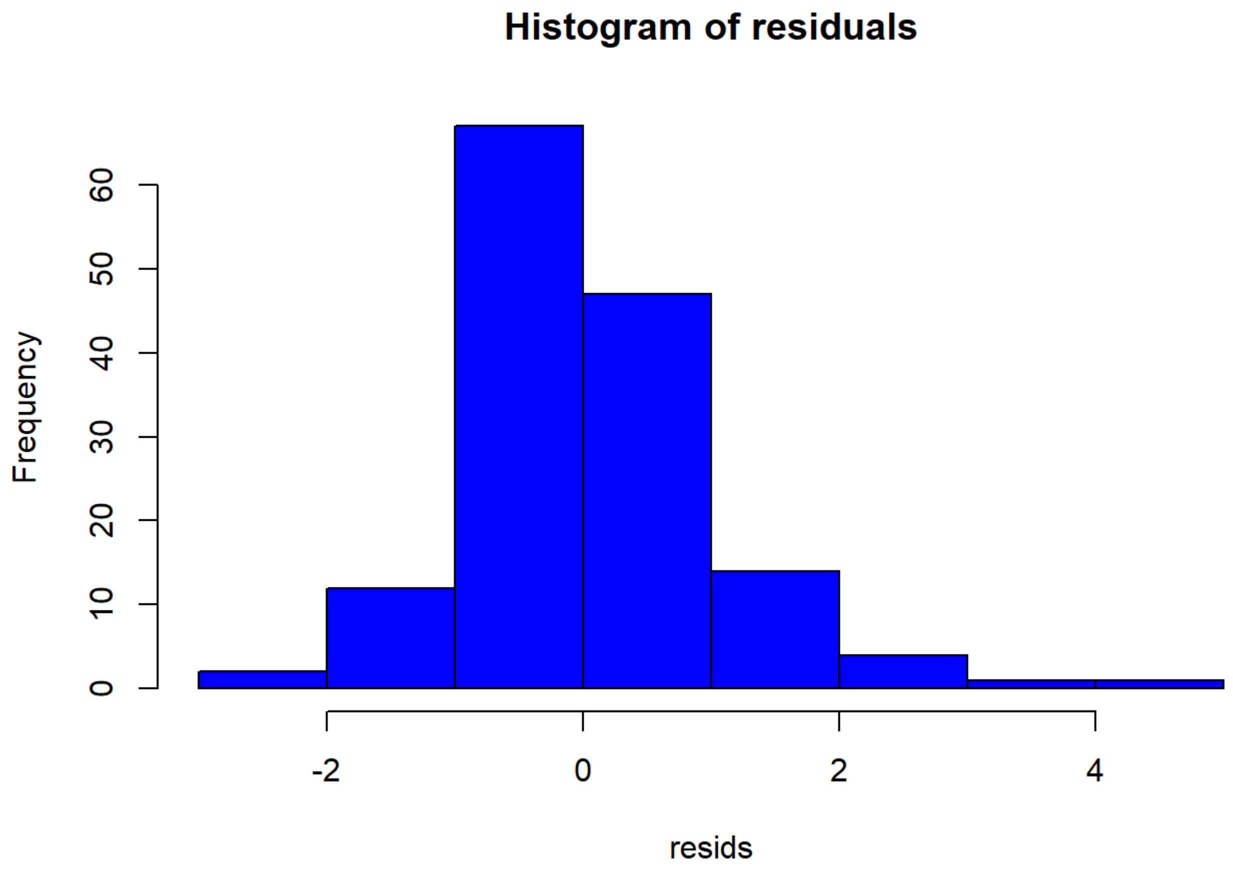


Answer

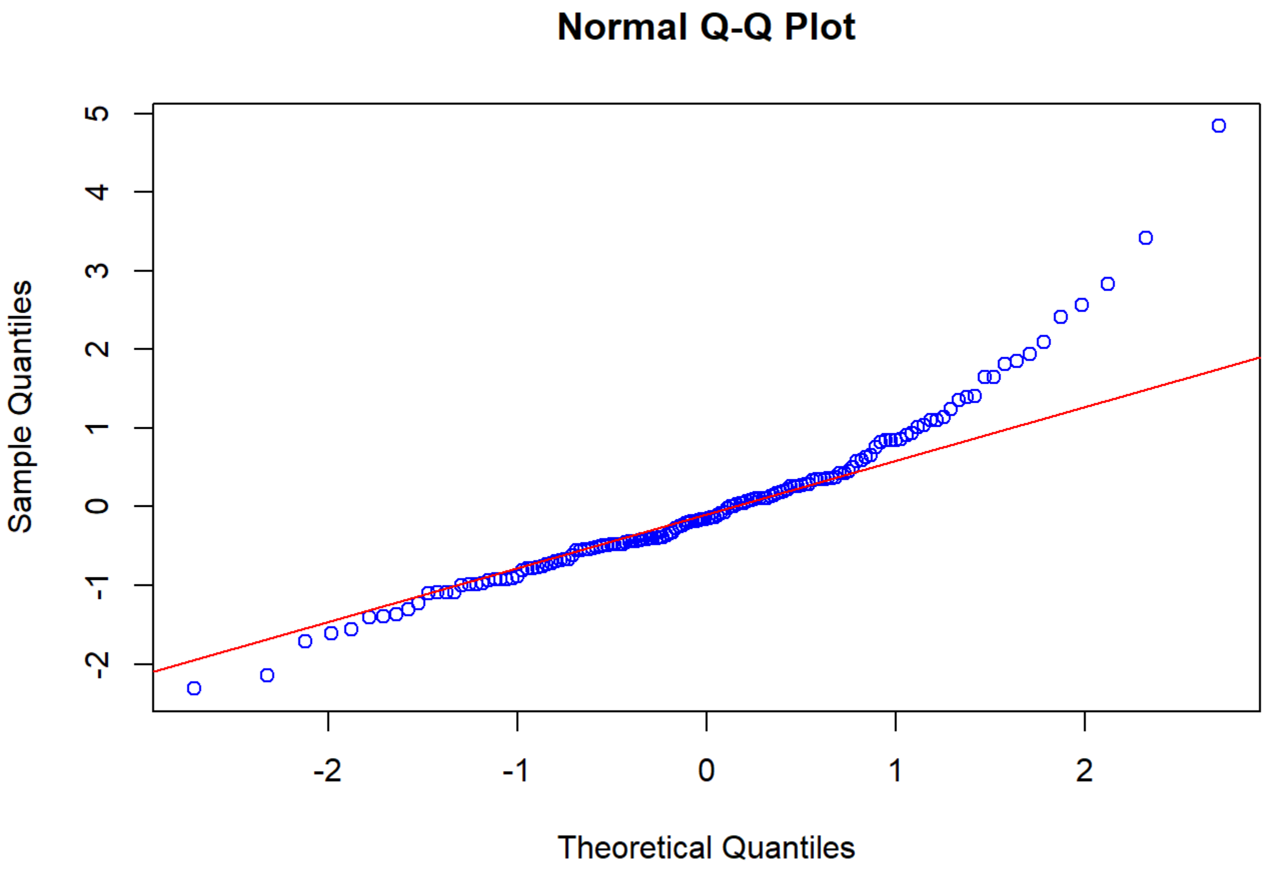
The constant variance assumption does not hold. As seen in the plot above, the variance increases as the fitted values increase. Since there is no grouping of the residuals, we can conclude the errors appear to be uncorrelated.

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

```
# Plot histogram of std residuals
hist(resids,
     col="blue",
     main="Histogram of residuals")
```



```
# qq plot of std residuals
qqnorm(resids,
       col="blue")
qqline(resids,
       col="red")
```



Answer

The Q-Q plot indicated as heavy-tailed. Histogram should have an approximately symmetric distribution with no gaps, which is not presented in our hist plot. Hence, both the Q-Q plot and histogram suggest the normality assumptions does not hold.

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
```

Answer

See above model3 output.

(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
```

Answer

Here, the null hypothesis is that the Body.Height, Diagonal.Length, Height and Width coefficients are all 0, i.e. $\beta_{Body.Height}$ = $\beta_{Diagonal.Length}$ = β_{Height} = β_{Width} = 0 and the alternative hypothesis is that at least one these coefficient is not 0. In other words, if we reject the null hypothesis, it means we can conclude that at least one of these coefficient has predictive power.

Observe above the F-statistic is 1.7626 and p-value is 0.14. Because the p-value is greater than 0.01, we ***cannot*** reject the null hypothesis that the $Body.Height$, $Diagonal.Length$, $Height$ and $Width$ coefficients are all 0, given the variables $Species$ and $Total.Length$ been taken under consideration. The conclusion of this test is that at $\alpha - level$ = 0.01, $Body.Height$, $Diagonal.Length$, $Height$ and $Width$ do not contribute significant information to the fish weight, given the $Species$ and $Total.Length$ variables.

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

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

```
round(vif(model3),3)
```

```
##               GVIF Df GVIF^(1/(2*Df))
## Species      2.654  6           1.085
## Total.Length 2.654  1           1.629
```

```
r_2 = 0.9353
thresh = 1/(1-r_2)
thresh
```

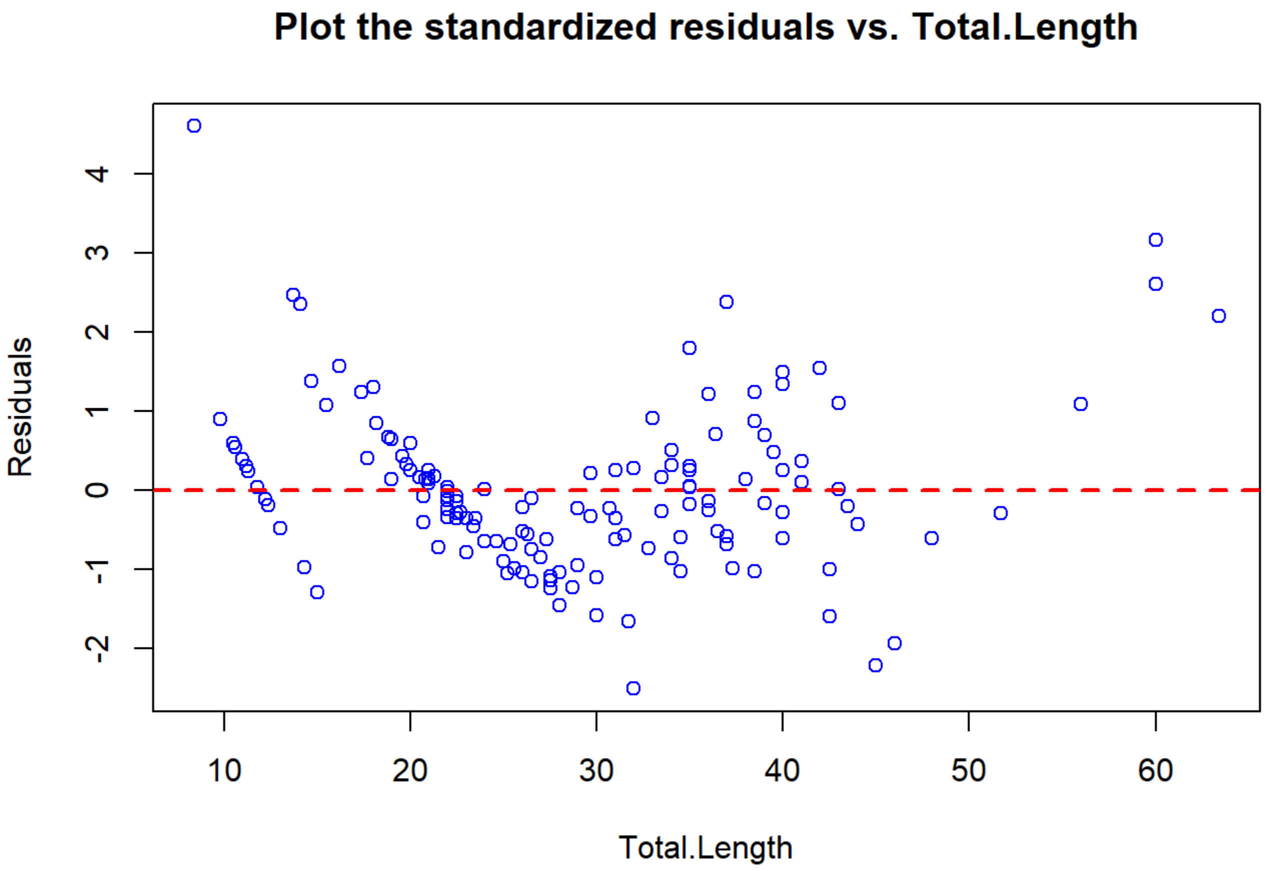
```
## [1] 15.45595
```

Answer

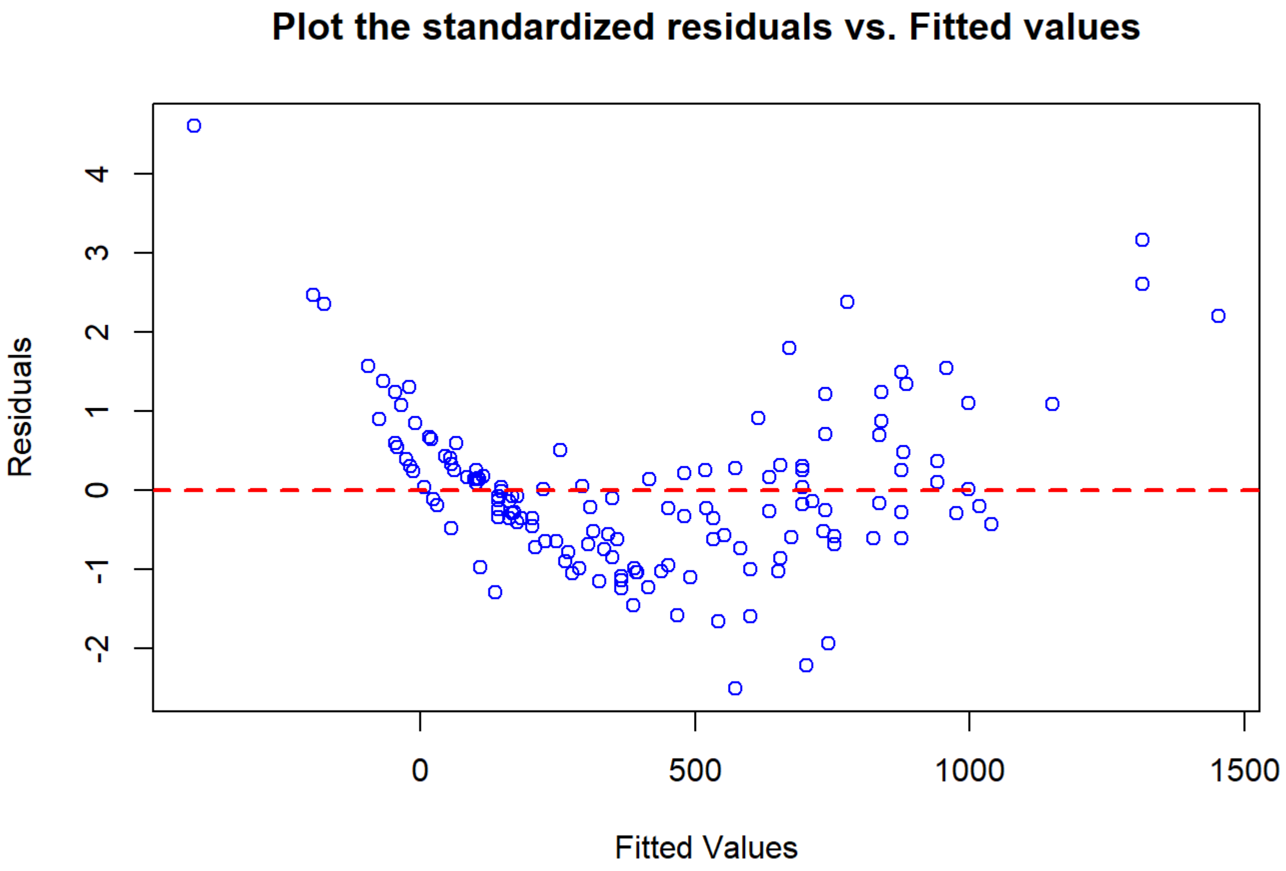
See above that the VIF of both $Species$ and $Total.Length$ is smaller than MAX(10, 1/(1-R^2_{model})), indicating there is no multicollinearity.

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

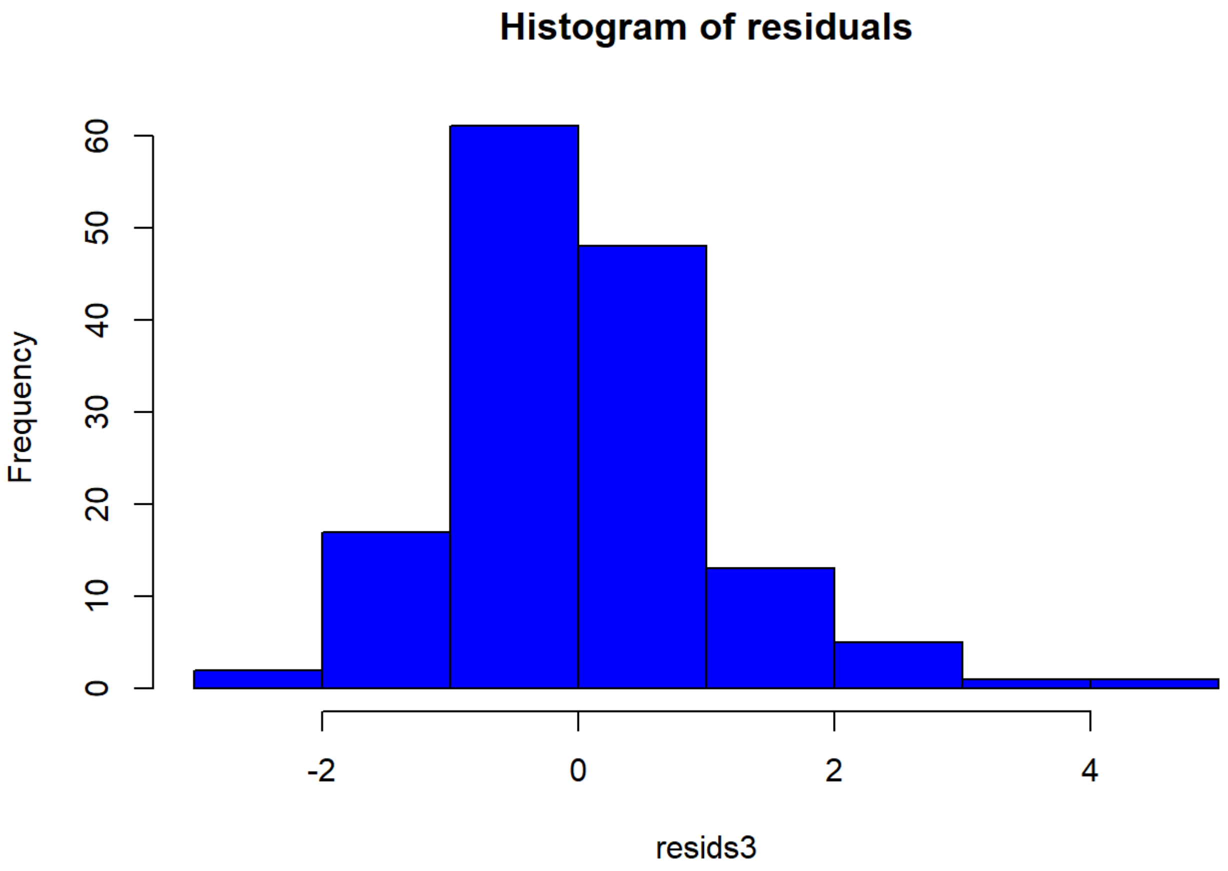
```
# Standardized residuals
resids3 = rstandard(model3)
# Fitted values
fits3 = model3$fitted

# Plot the standardized residuals against
# Total.Length
plot(fish2$Total.Length, resids3,
     main="Plot the standardized residuals vs. Total.Length",
     xlab="Total.Length",
     ylab="Residuals",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



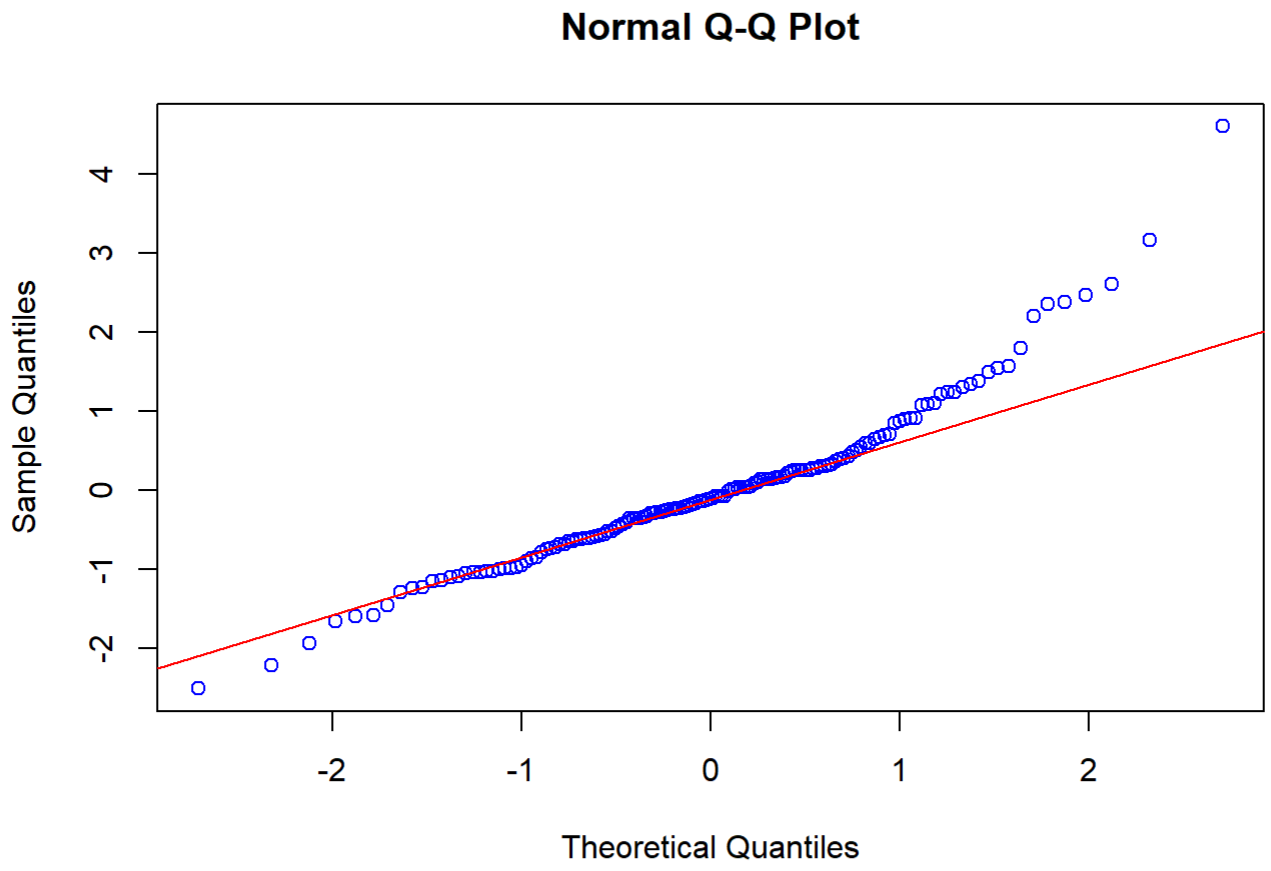
```
# Plot the standardized residuals against
# fitted values
plot(fits3, resids3,
     xlab="Fitted Values",
     ylab="Residuals",
     main="Plot the standardized residuals vs. Fitted values",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



```
# Plot histogram of std residuals
hist(resids3,
     col="blue",
     main="Histogram of residuals")
```



```
# qq plot of std residuals
qqnorm(resids3,
       col="blue")
qqline(resids3,
       col="red")
```



Answer

Based on the standardized residuals vs. Total.Length plot above, since there is a random pattern around the 0 mean line, we conclude the linearity assumption holds for the Total.Length predicting variable.

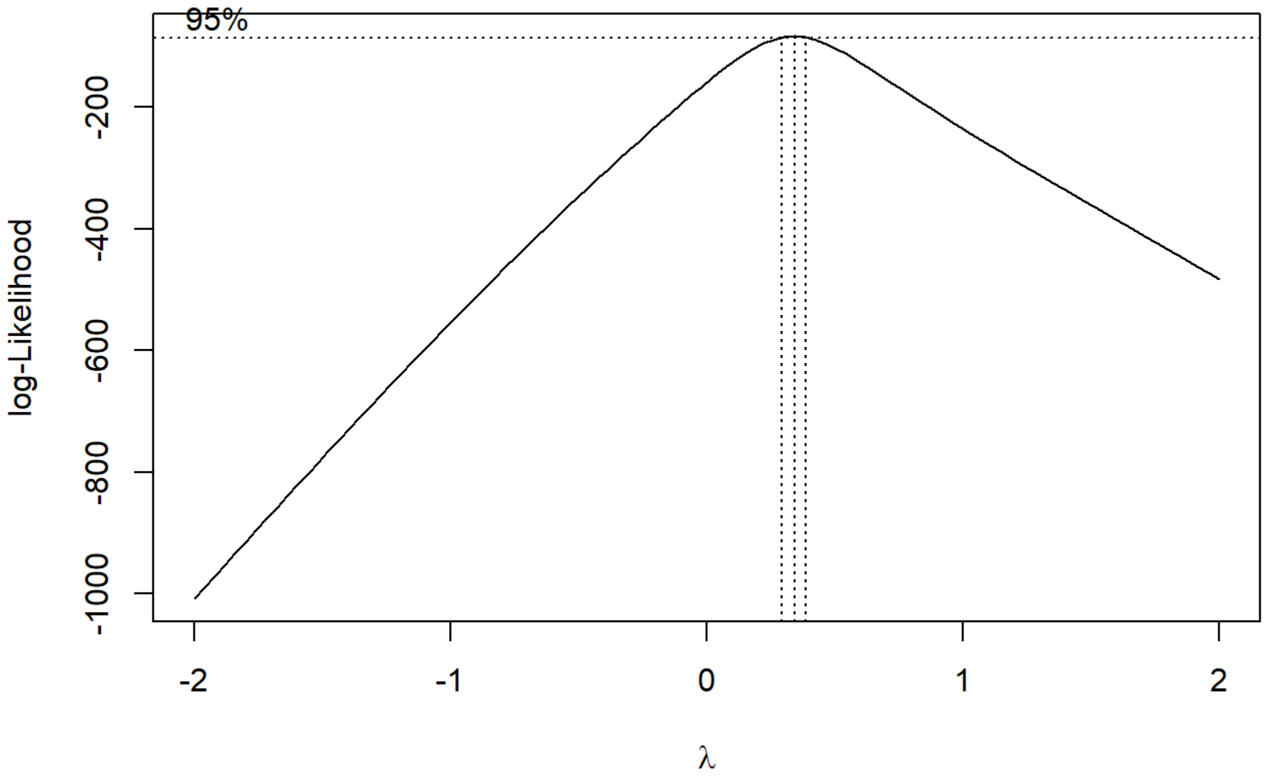
The constant variance assumption does not hold. As seen in the standardized residuals vs. Fitted values plot above, the variance increases as the fitted values increase. Since there is no grouping of the residuals, we can conclude the errors appear to be uncorrelated.

The Q-Q plot indicated as heavy-tailed. Histogram should have an approximately symmetric distribution with no gaps, which is not presented in our hist plot. Hence, both the Q-Q plot and histogram suggest the normality assumptions does not hold.

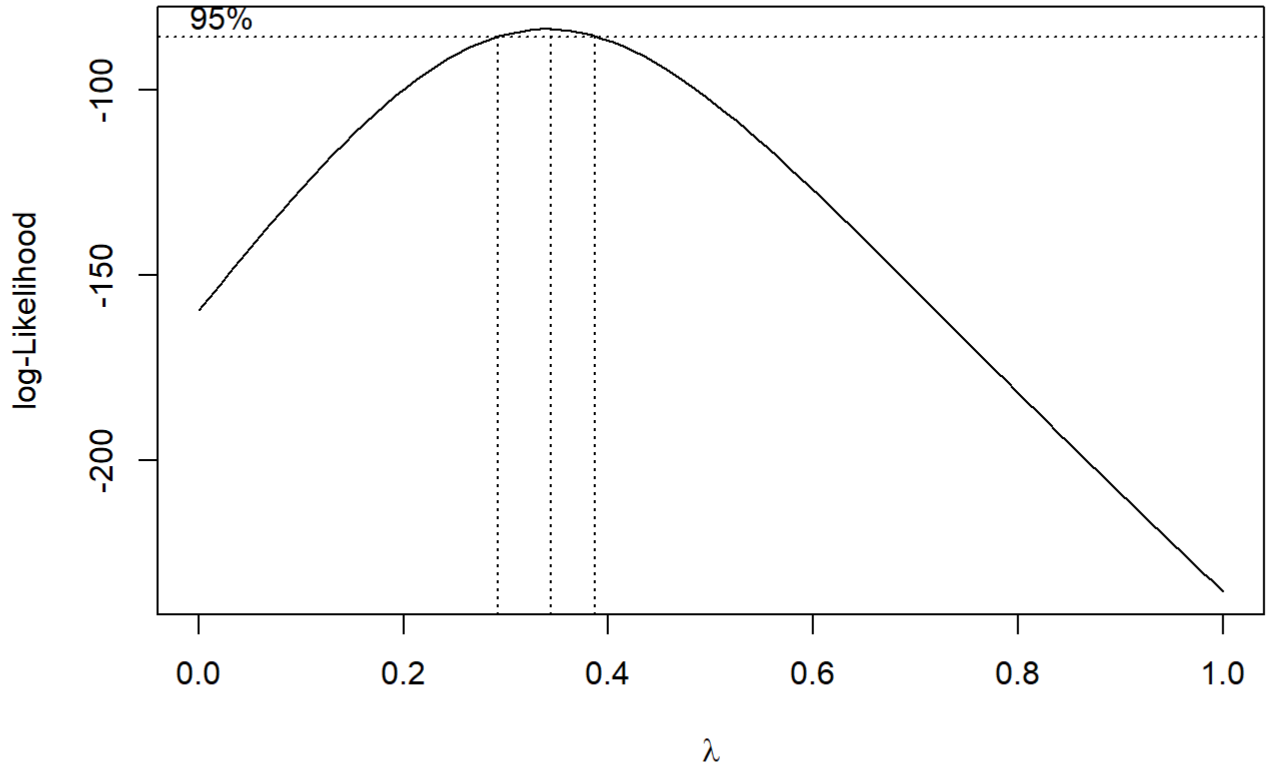
Question 7: Transformation [12 pts]

(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

```
boxcox(model3)
```



```
bc = boxcox(model3, plotit = TRUE, lambda = seq(0, 1, by = 0.1))
```



```
lambda = bc$x[which.max(bc$y)]
print(paste("Optimal lambda: ", lambda))
```

```
## [1] "Optimal lambda:  0.343434343434343"
```

```
print(paste("Optimal lambda rounded to the nearest half integer: ", round(2*lambda)/2))
```

```
## [1] "Optimal lambda rounded to the nearest half integer:  0.5"
```

Answer

See above the optimal λ value is 0.3434 and the optimal λ rounded to the nearest half integer is 0.5, which means to use the $\sqrt{Y}$ 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)
```

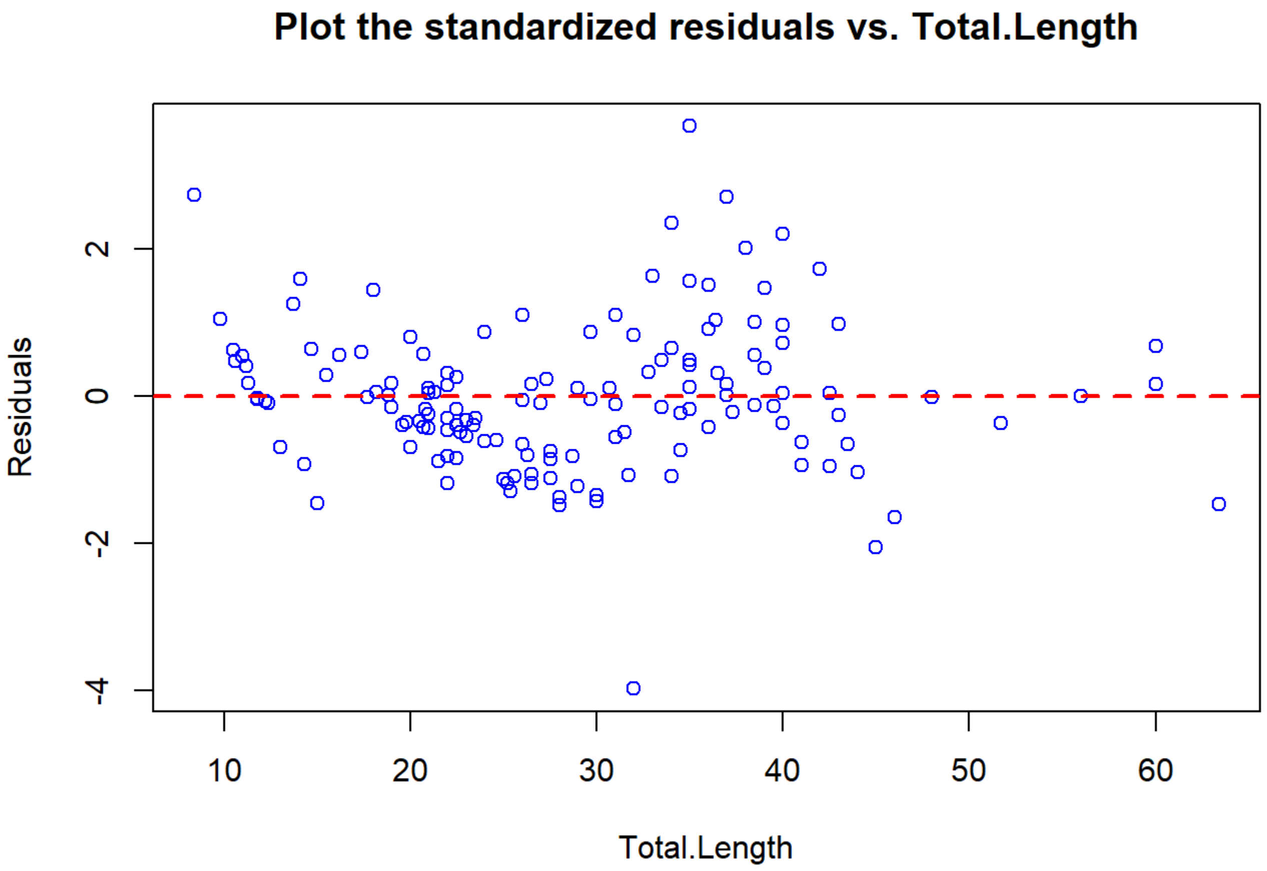
```
##
## Call:
## lm(formula = sqrt(Weight) ~ Species + Total.Length, data = fish2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0111 -0.7687 -0.0579  0.6797  4.6383
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -6.96654    0.57278 -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
```

Answer

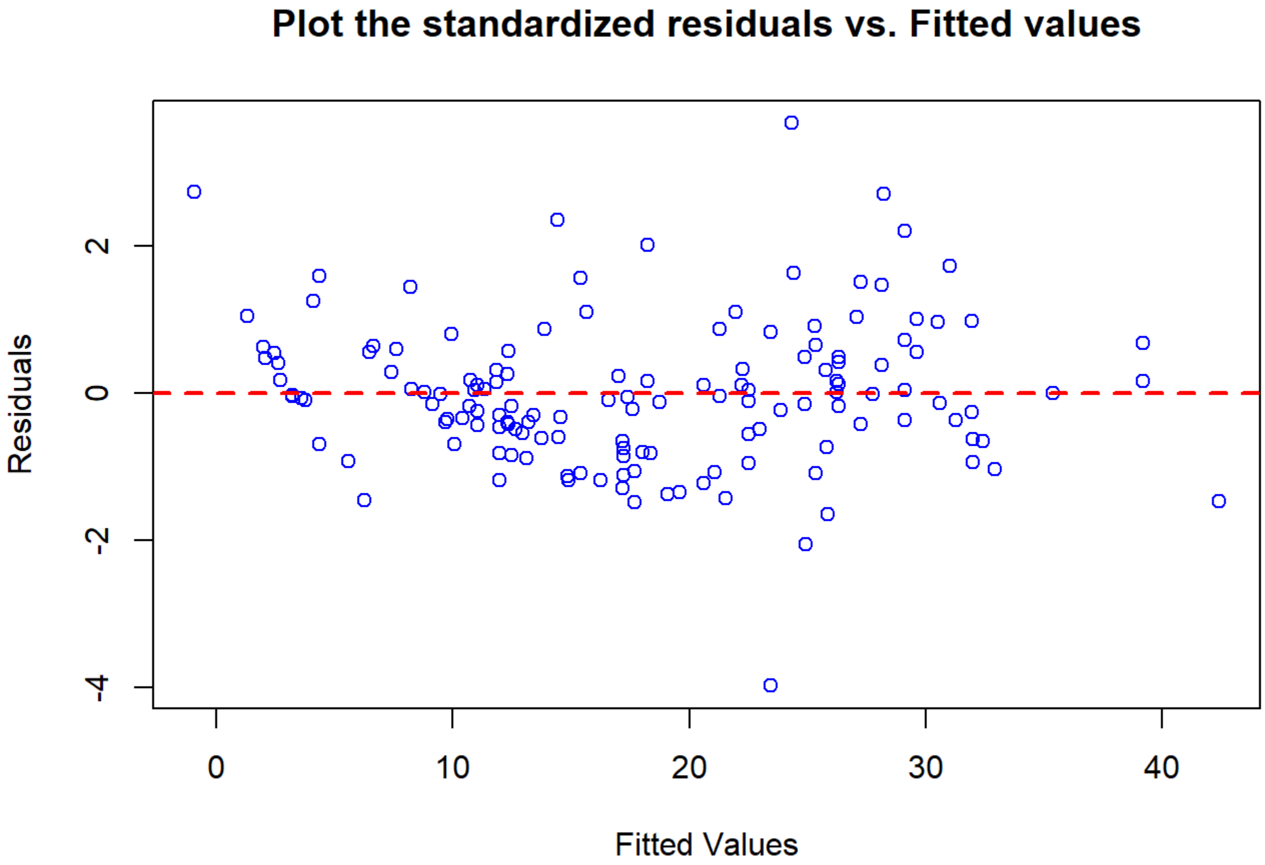
See above model4 output.

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

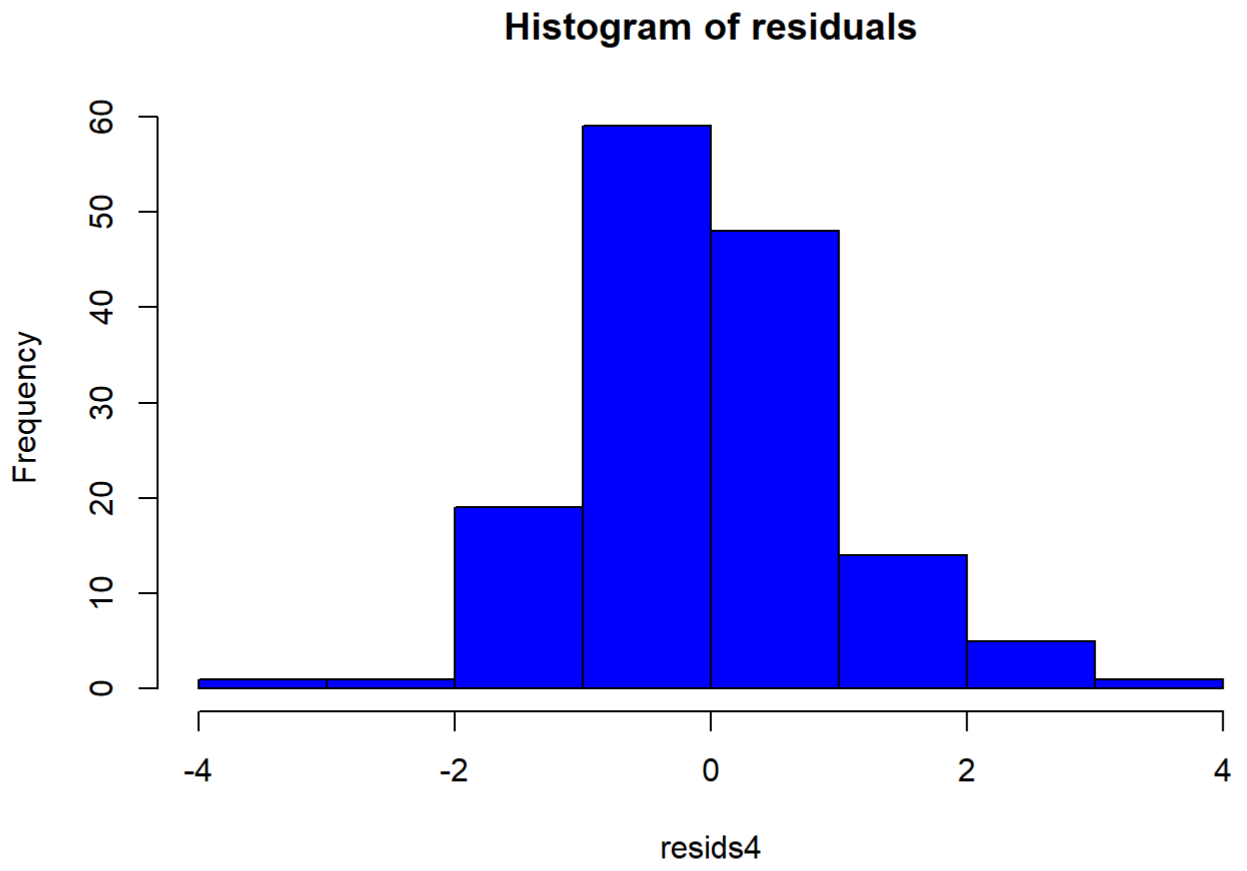
```
# Standardized residuals
resids4 = rstandard(model4)
# Fitted values
fits4 = model4$fitted

# Plot the standardized residuals against
# Total.Length
plot(fish2$Total.Length, resids4,
     main="Plot the standardized residuals vs. Total.Length",
     xlab="Total.Length",
     ylab="Residuals",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



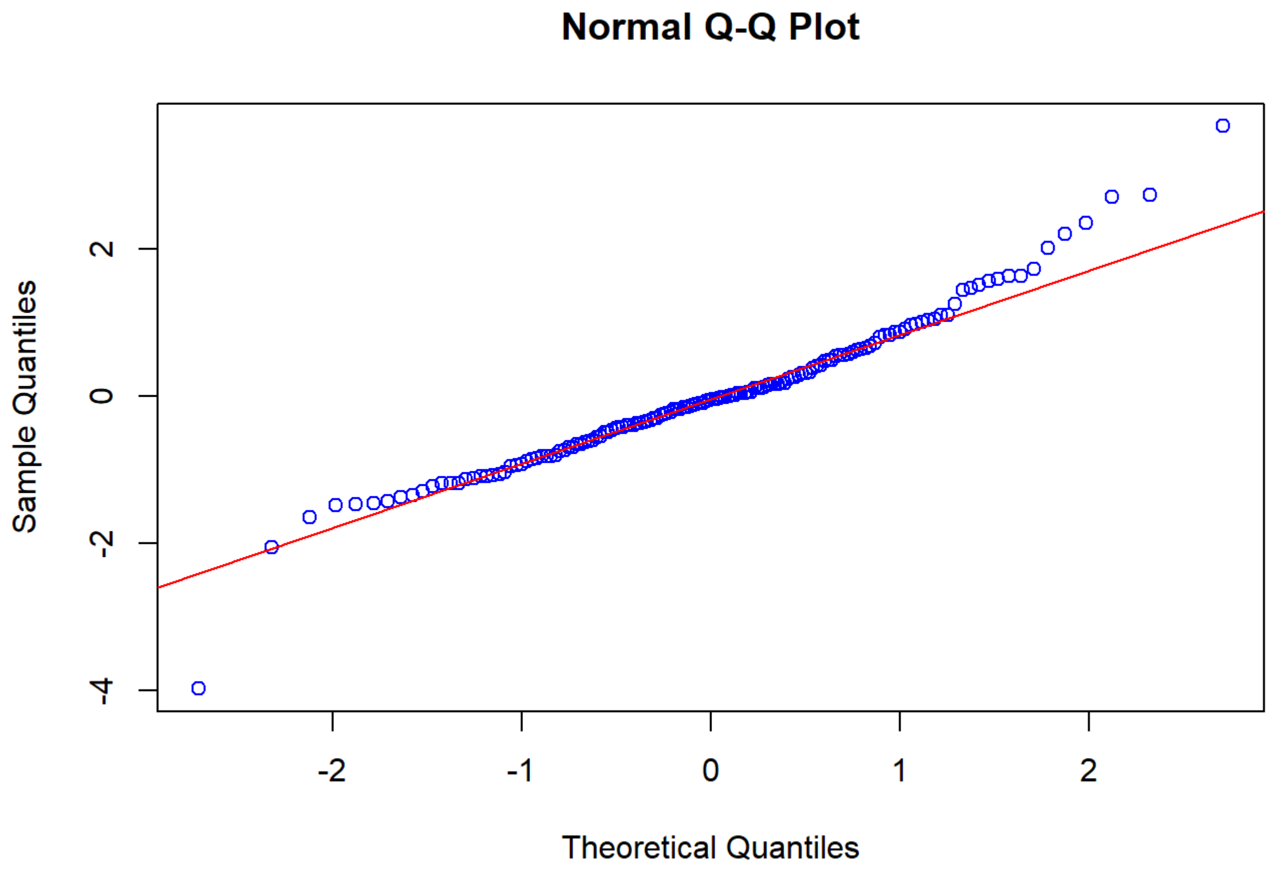
```
# Plot the standardized residuals against
# fitted values
plot(fits4, resids4,
     xlab="Fitted Values",
     ylab="Residuals",
     main="Plot the standardized residuals vs. Fitted values",
     col="blue")
abline(0, 0,
       col="red",
       lty=2, lwd=2)
```



```
# Plot histogram of std residuals
hist(resids4,
     col="blue",
     main="Histogram of residuals")
```



```
# qq plot of std residuals
qqnorm(resids4,
       col="blue")
qqline(resids4,
       col="red")
```



Answer

Based on the standardized residuals vs. Total.Length plot above, since there is a random pattern around the 0 mean line, we conclude the linearity assumption holds for the Total.Length predicting variable.

As seen in the standardized residuals vs. Fitted values plot above, the variance is the same across the fitted values, meaning the constant variance assumption holds. Since there is no grouping of the residuals, we can conclude the errors appear to be uncorrelated.

The Q-Q plot indicated as tailed, but not so heavy-tailed. Histogram also shows an improvement from the model3's histogram. Hence, we can conclude that the normality assumption holds.

Overall, the transformation seems to be successful.

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(paste("model2 R^2 is: ", summary(model2)$r.squared))
```

```
## [1] "model2 R^2 is:  0.938483625001571"
```

```
print(paste("model2 Adjusted R^2 is: ", summary(model2)$adj.r.squared))
```

```
## [1] "model2 Adjusted R^2 is:  0.933508035847286"
```

```
cat(sep="\n\n")
```

```
print(paste("model3 R^2 is: ", summary(model3)$r.squared))
```

```
## [1] "model3 R^2 is:  0.935294615139321"
```

```
print(paste("model3 Adjusted R^2 is: ", summary(model3)$adj.r.squared))
```

```
## [1] "model3 Adjusted R^2 is:  0.932059345896287"
```

```
cat(sep="\n\n")
```

```
print(paste("model4 R^2 is: ", summary(model4)$r.squared))
```

```
## [1] "model4 R^2 is:  0.981713834446085"
```

```
print(paste("model4 Adjusted R^2 is: ", summary(model4)$adj.r.squared))
```

```
## [1] "model4 Adjusted R^2 is:  0.980799526168389"
```

Answer

The coefficient of determination, or R^2, is the proportion of total variability in Y that can be explained by the linear regression model, or the amount of variance accounted for in the relationship between two (or more) variables i.e. our response variable Weight and the predictors. Simply put, as R^2 increases, the Y values would be closer to the regression line (in SLR)/ plane (in MLR). R^2 is calculated as:

R^2 = SSR / SST = 1 - (SSE - SST), where:

SST = sum of squares total = total deviation

SSE = sum of squared errors = unexplained deviation

SSR = sum of squares for regression = explained deviation

Observe above that the R^2 of model2 equals 0.938, or 93.8%. This means that 93.8% of the variation of Weight can be explained by the predictors included in model2. Similarly, R^2 of model3 is 0.935 and of model4 is 0.981. Model4's R^2 is the highest. Note that as we add predicting variables, R^2 can only increase. Hence, when we want to compare models with different number of predicting variables, we should use the Adjusted R^2, because it adjusts, or penalizes, for additional predictors included in the model. Observe above that model2's Adjusted R^2 is 0.933, model3's is 0.932, and model4's is 0.980.

Based on the Adjusted R^2 (and R^2), model4 seems to perform best, compared to model2 and model3.

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.

```
pred3 = predict(model3, fishtest, interval = 'prediction')
pred4 = predict(model4, fishtest, interval = 'prediction')
pred3 = pred3[,1]
pred4 = pred4[,1]

# Mean Squared Prediction Error (MSPE)
MSPE3 = mean((pred3-fishtest$Weight)^2)
MSPE4 = mean((pred4-fishtest$Weight)^2)

print(paste("MSPE based on model3 prediction is: ", MSPE3))
```

```
## [1] "MSPE based on model3 prediction is:  9392.24969170129"
```

```
cat(sep="\n\n")
```

```
print(paste("MSPE based on model4 prediction is: ", MSPE4))
```

```
## [1] "MSPE based on model4 prediction is:  98076.6180622229"
```

Answer

Observe the MSPE, computed as the mean of the square differences between predicted and observed, for model3 9392 and model4 98076. Based on the MSPE above, it appears that model3 performs better than model4.

(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.

```
new_fish = data.frame(Species="Perch", Body.Height=28, Total.Length=32)

pred_new = predict(model4, new_fish, interval = 'prediction', level=0.90)
pred_new
```

```
##        fit     lwr      upr
## 1 21.49286 19.3508 23.63491
```

Answer

The average weight of a Perch fish with a Body.Height of 28 cm, and a Total.Length of 32 cm, predicted using model4 at a 90% prediction interval is 21.49 grams, with a lower bound of 19.35 grams and an upper bound of 23.63 grams.

Thank you!