HW3 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)
```

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

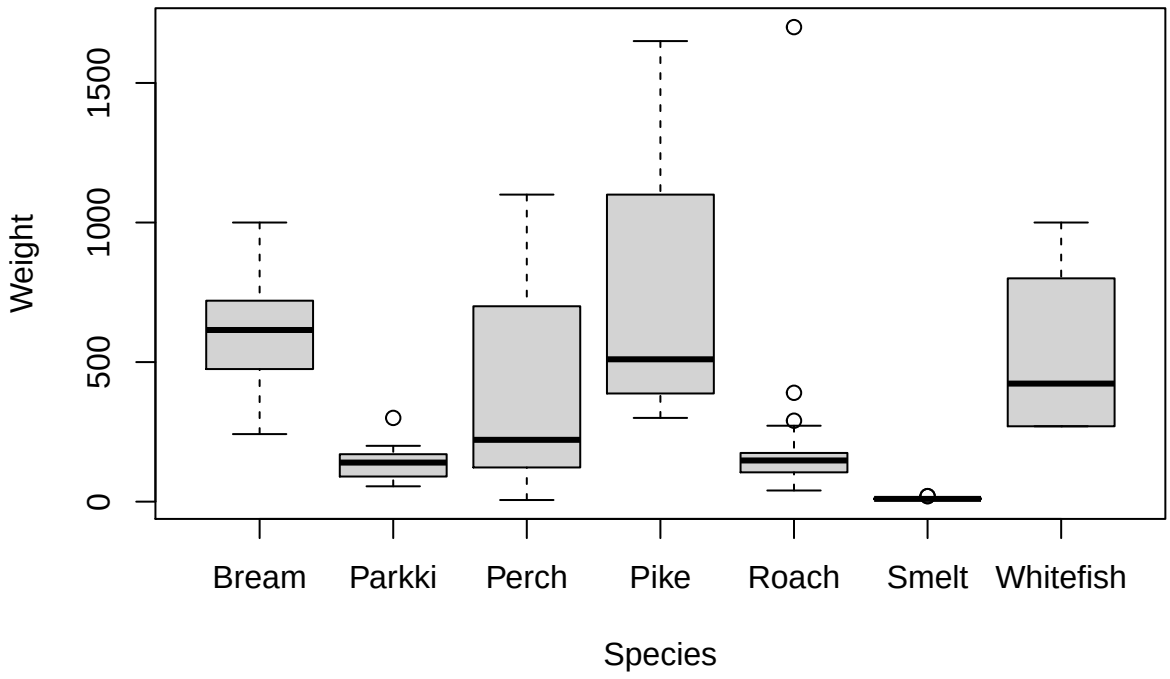
```
# Read the data set
fishfull = read.csv("C:/Users/mjpearl/Desktop/omsa/ISYE-6414-OAN/hw3/Fish.csv",header=T, fileEncoding =
row.cnt = nrow(fishfull)
# Split the data into training and testing sets
fishtest = fishfull[(row.cnt-9):row.cnt,]
fish = fishfull[1:(row.cnt-10),]
```

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?

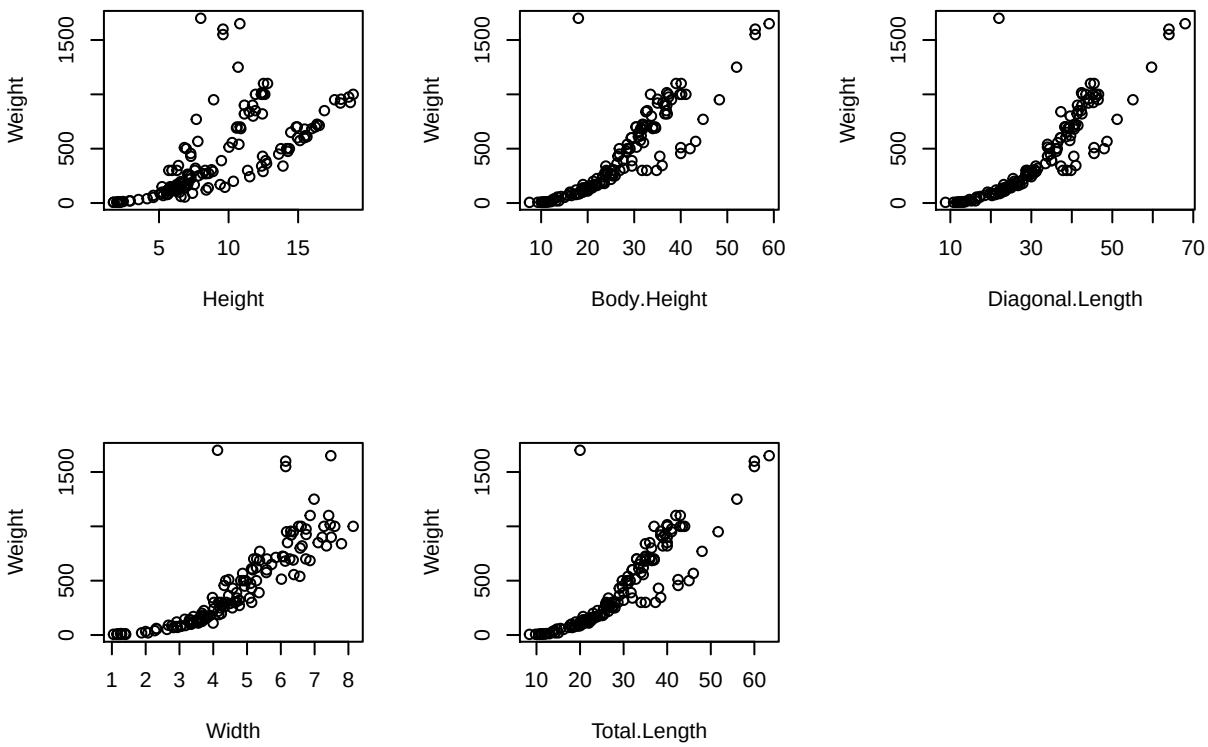
```
boxplot(Weight~Species,data=fish)
```



Based on our results, we can see that the weight tends to differ significantly based on the Species type, and therefore can conclude there does appear to be a relationship with weight across the multiple species.

(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?**

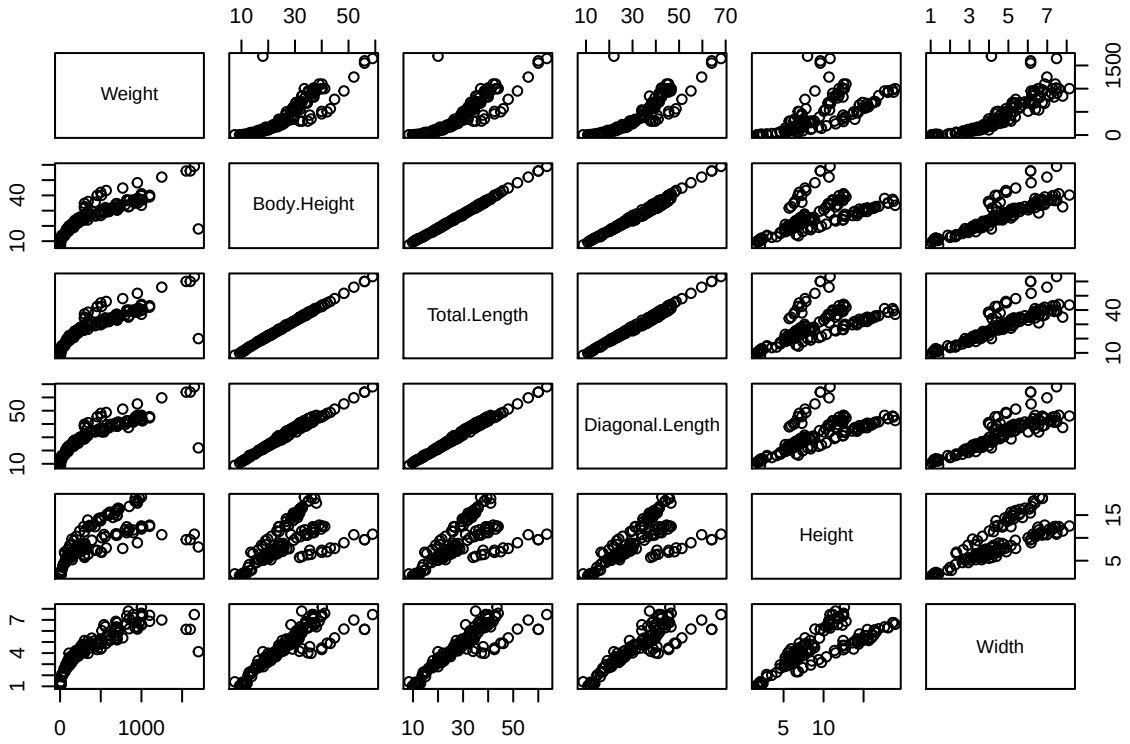
```
par(mfrow=c(2,3))
plot(Weight~Height, data=fish)
plot(Weight~ Body.Height, data=fish)
plot(Weight~Diagonal.Length, data=fish)
plot(Weight~Width, data=fish)
plot(Weight~ Total.Length, data=fish)
```



When observing the first plot on the top left, we can see a general linear trend between the height and the weight. When looking at the observations on the top left, one could argue that these 4 observations could be classified as outliers, however, they're not drastically far off from the 3rd quartile of observations, so one would need to do a regression with and without these observations to really assess the impact.

We can see from the last 4 plots that there exhibits a nice linear relationship between the response and this predictor variable. However there does seem to be an influential point that seems to be far from the means of both the x and y and is situated in the top left portion of each plot.

```
nums <- unlist(lapply(fish, is.numeric))
plot(fish[ , nums])
```



(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(fishfull[ , nums])
```

```
##                    Weight Body.Height Total.Length Diagonal.Length    Height
## Weight          1.0000000   0.8599272    0.8636929       0.8678228 0.6907124
## Body.Height     0.8599272   1.0000000    0.9995199       0.9920566 0.6247620
## Total.Length    0.8636929   0.9995199    1.0000000       0.9941380 0.6398512
## Diagonal.Length 0.8678228   0.9920566    0.9941380       1.0000000 0.7027728
## Height          0.6907124   0.6247620    0.6398512       0.7027728 1.0000000
## Width           0.8485527   0.8661069    0.8727830       0.8776244 0.7929377
##                     Width
## Weight          0.8485527
## Body.Height     0.8661069
## Total.Length    0.8727830
## Diagonal.Length 0.8776244
## Height          0.7929377
## Width           1.0000000
```

Based on our correlation chart, we can see that we're exhibiting high correlation between the response variable and all attributes, except for the Height attribute, however 0.69% is still a modest correlation score. We can also see we're exhibiting multicollinearity because several variables are showing high correlation with other predictive variables (i.e. Body Height and Total Length = 0.99% correlation).

(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? Yes, based on this initial assessment a multiple linear regression model appears to be reasonable for the relationship between Weight and all the predictor variables.Even though we're experiencing multicollinearity, it doesn't mean that we don't have predictive variables that provide high explanatory power for the response variable weight. When we address the outliers and multicollinearity with feature selection criteria, we should reasonably have a model with high explanatory power.

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.

```
Species <- as.factor(fish$Species)
model1 <- lm(Weight~Height+Body.Height+Total.Length+Diagonal.Length+Height+Width+Species ,data=fish)
summary(model1)
```

```
##
## Call:
## lm(formula = Weight ~ Height + Body.Height + Total.Length + Diagonal.Length +
##     Height + Width + Species, 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 ***
## Height              38.27      22.09   1.732 0.085448 .
## 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
## Width               29.63      40.54   0.731 0.466080
## 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
## ---
## 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?

Since $\alpha = 0.01$ exceeds the observed significance level, p = <2.2e-16, we reject the null hypothesis. The data provide strong evidence that at least one of the slope coefficients is nonzero. The overall model appears to be statistically useful in predicting weight

(c) What is the coefficient estimate for *Body.Height*? Interpret this coefficient. The result means that holding all other predictors constant, a 1 unit increase in Body.Height, we will see a -176.87 reduction in Weight. This coefficient does appear to be statistically significant as the p-value 0.004583 < alpha-level. However, we tend to see drastic changes in the results of the coefficients with these values, which may be alluding to potential multicollinearity.

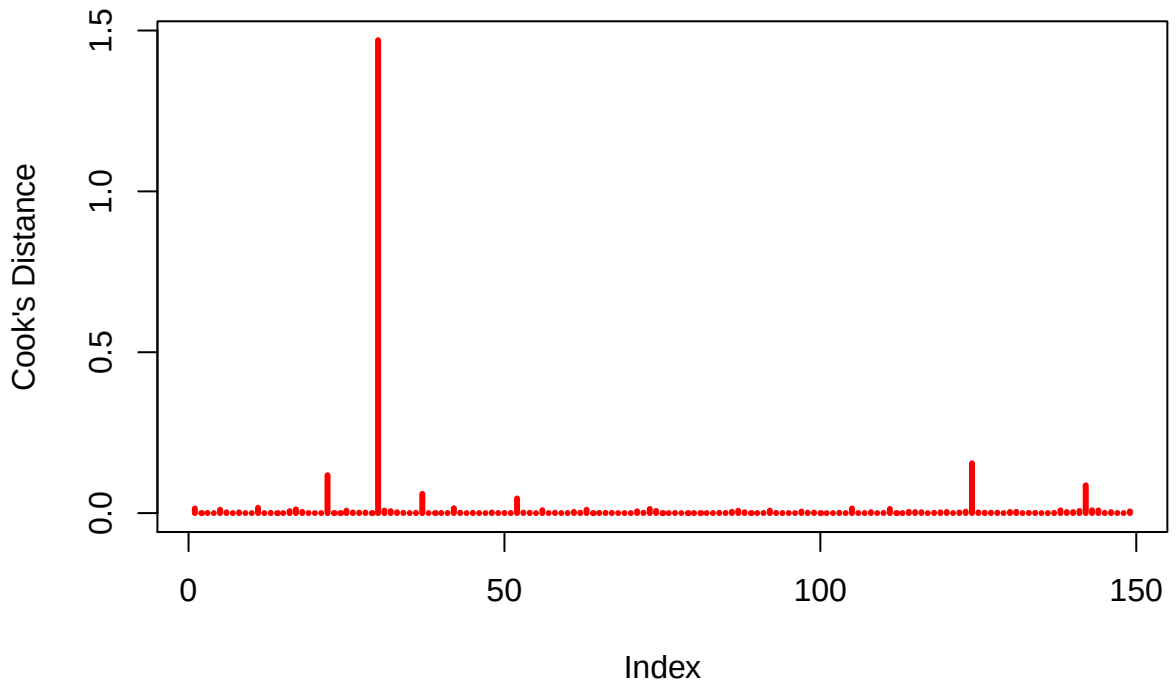
(d) What is the coefficient estimate for the *Species* category Parkki? Interpret this coefficient. Based on our approach for using as.factor for dummy variables, the Parkki variable is the baseline dummy variable for the rest of the species types.

The result means that holding all other predictors constant, having a species value of Parkki, is associated with a 79.34 increase in the value for Weight. This coefficient does not appear to be statistically significant as the p-value 0.550918 > alpha-level

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



```
cook[cook>1]
```

```
##       30
## 1.469853
```

Based on the results, we can see that there's one outlier acting as a heavy influence point, as it's Cook's distance is > 1 for a value of 1.423786.

(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.

```
removeRows <- function(rowNum, data) {
      newData <- data[-rowNum, , drop = FALSE]
      rownames(newData) <- NULL
      newData
}

fish2 = removeRows(30,fish)
model2 <- lm(Weight~Height+Body.Height+Total.Length+Diagonal.Length+Height+Width+Species ,data=fish2)
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ Height + Body.Height + Total.Length + Diagonal.Length +
##     Height + Width + Species, 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 ***
## Height              10.000     13.398   0.746 0.456692
## 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
## Width               -8.339     24.483  -0.341 0.733924
## 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
## ---
## 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
```

We can see from the result that removing this outlier had a drastic increase on the model's performance as the R2 coefficient of determinant went up from 0.79 to 0.9335.

(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?

```
vif(model2)
```

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

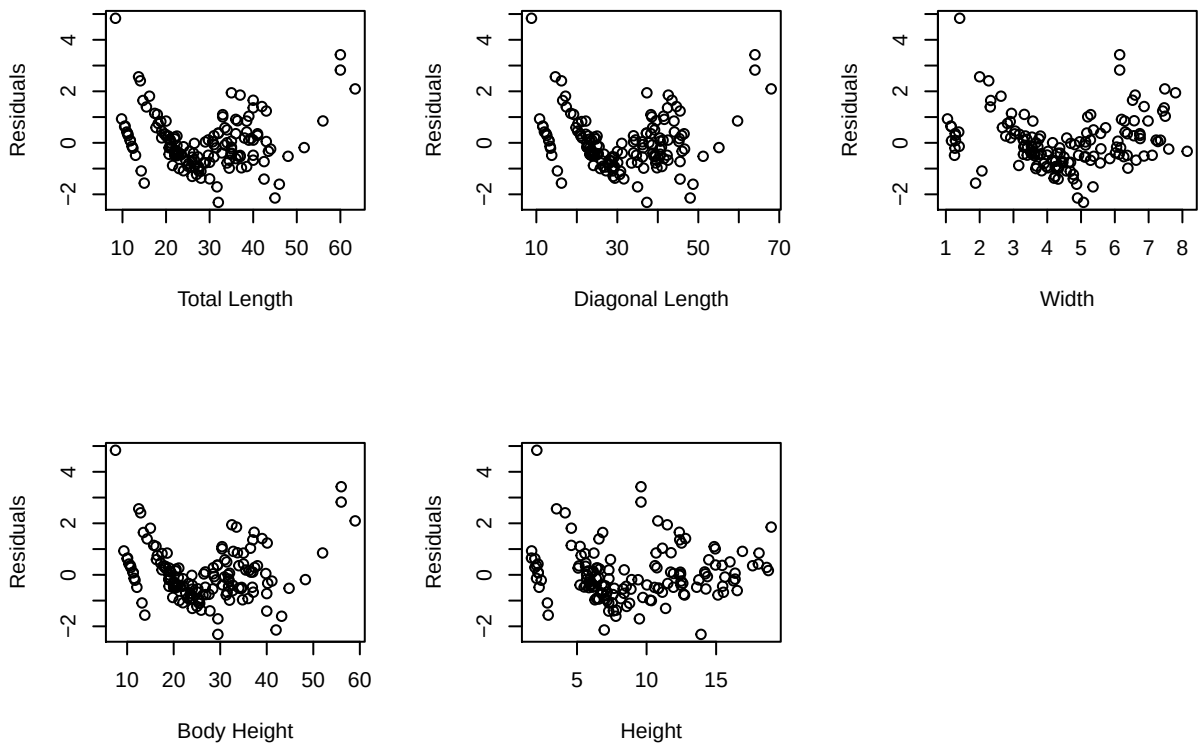
Based on our result, we can see that all GVIF values are > 10 and 1/(1-0.9335)=15. Therefore, we can conclude that we are exhibiting multicollinearity in our model for all predicting variables.

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?

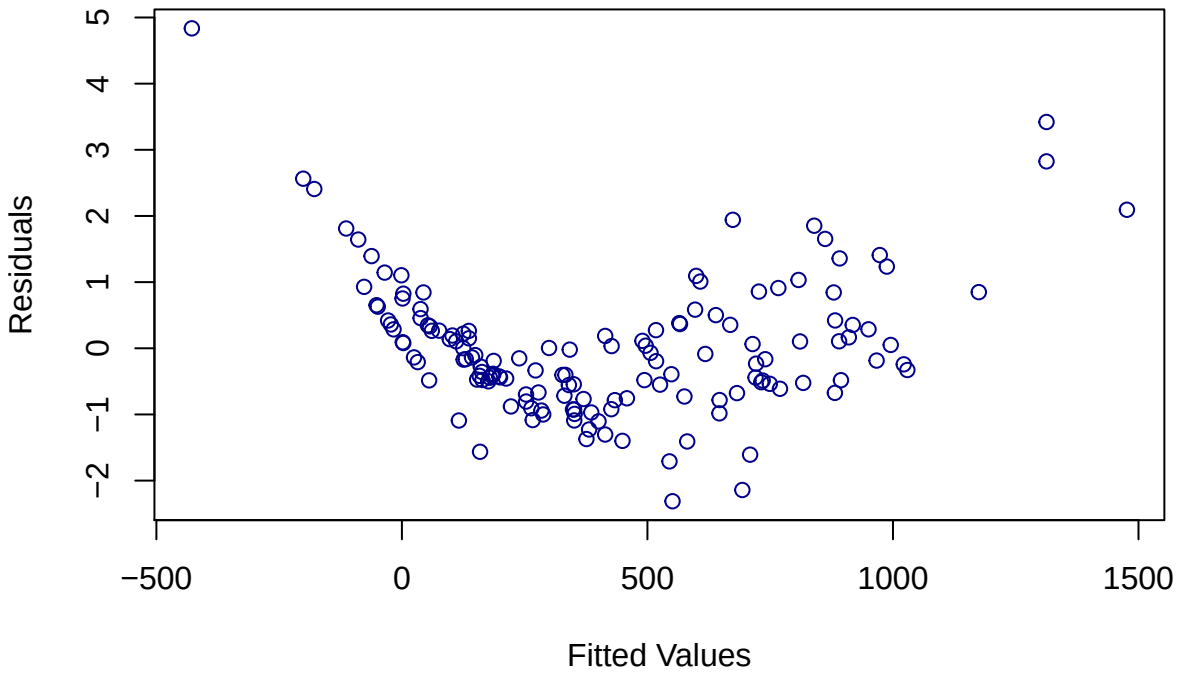
```
library(MASS)
fits = model2$fitted.values
resids = stdres(model2)
par(mfrow=c(2,3))
plot(fish2$Total.Length,resids,xlab='Total Length',ylab='Residuals')
plot(fish2$Diagonal.Length,resids,xlab='Diagonal Length',ylab='Residuals')
plot(fish2$Width,resids,xlab='Width',ylab='Residuals')
plot(fish2$Body.Height,resids,xlab='Body Height',ylab='Residuals')
plot(fish2$Height,resids,xlab='Height',ylab='Residuals')
```



We can see from our results that our predictive variables are scattered evenly across the 0 line when compared against the standardized residuals, and therefore our linearity assumption holds.

(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?

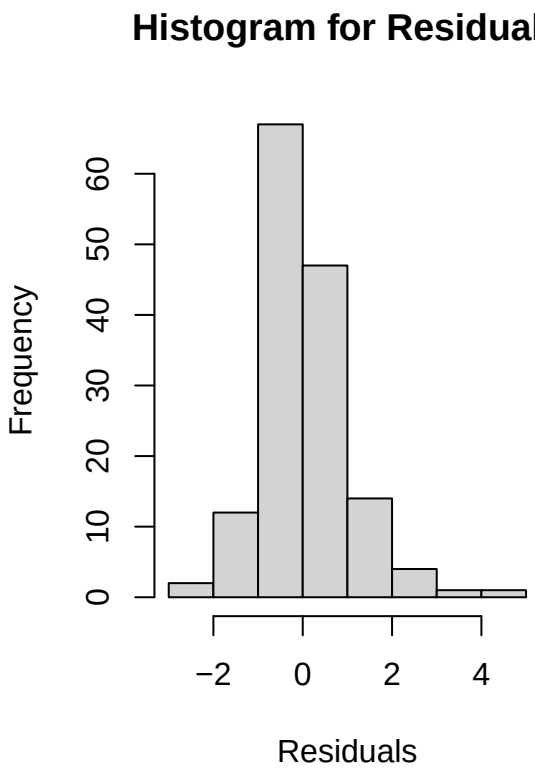
```
plot(fits,resids,xlab='Fitted Values',ylab='Residuals',col='darkblue')
```

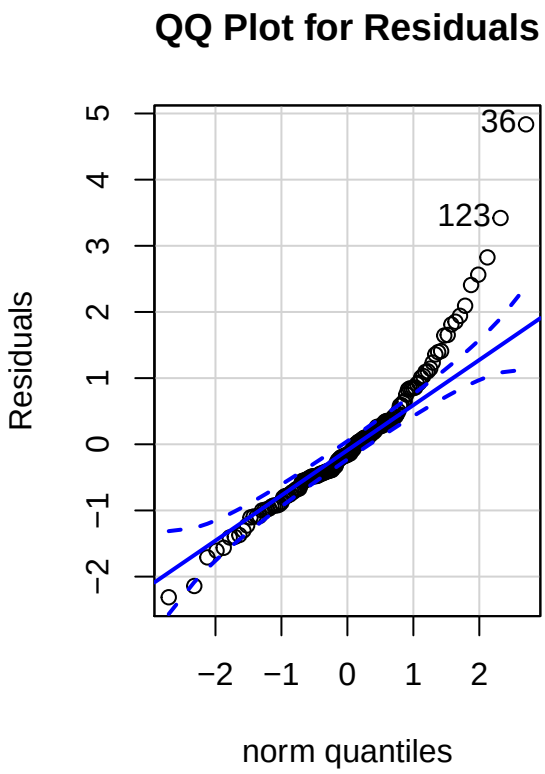


Based on the result of our plot, we can see that the variance seems to be spread out evenly along the 0 line and don't seem to form any cone shapes as our values increase. There does seem to be a few outlier values > 3, but the general range of values seems to -2 to 2, which is consistent across most of the plot. When we look at values from 200 to 750 on the x-axis, we could argue that a cone-shape is starting to emerge, however this seems to flatten out > 750 and a constant variance emerges again. Therefore, overall I will say the constant variance assumption holds.

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

```
par(mfrow=c(1,2))
hist(resids,xlab='Residuals',main="Histogram for Residuals")
qqPlot(resids,ylab='Residuals',main="QQ Plot for Residuals")
```






```
## [1]  36 123
```

From our results we can see that the normality assumption does not hold as our histogram is show a slight skewness, which is reaffirmed in our qqplot as there seems to be a significant tale forming at the end of the plot.

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

H0: a = 0 Meaning that the additional variable does not provide any additional explanatory power HA: a != 0 Meaning that the additional variable does provide additional explanatory power

Based on our results, we see a p-value for the partial F-test of 0.14 which is > alpha level 0.01. This means that we reject the null hypothesis, meaning that 1 or more variables that we've excluded as part of model3 do provide explanatory power against the response variable Weight.

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

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

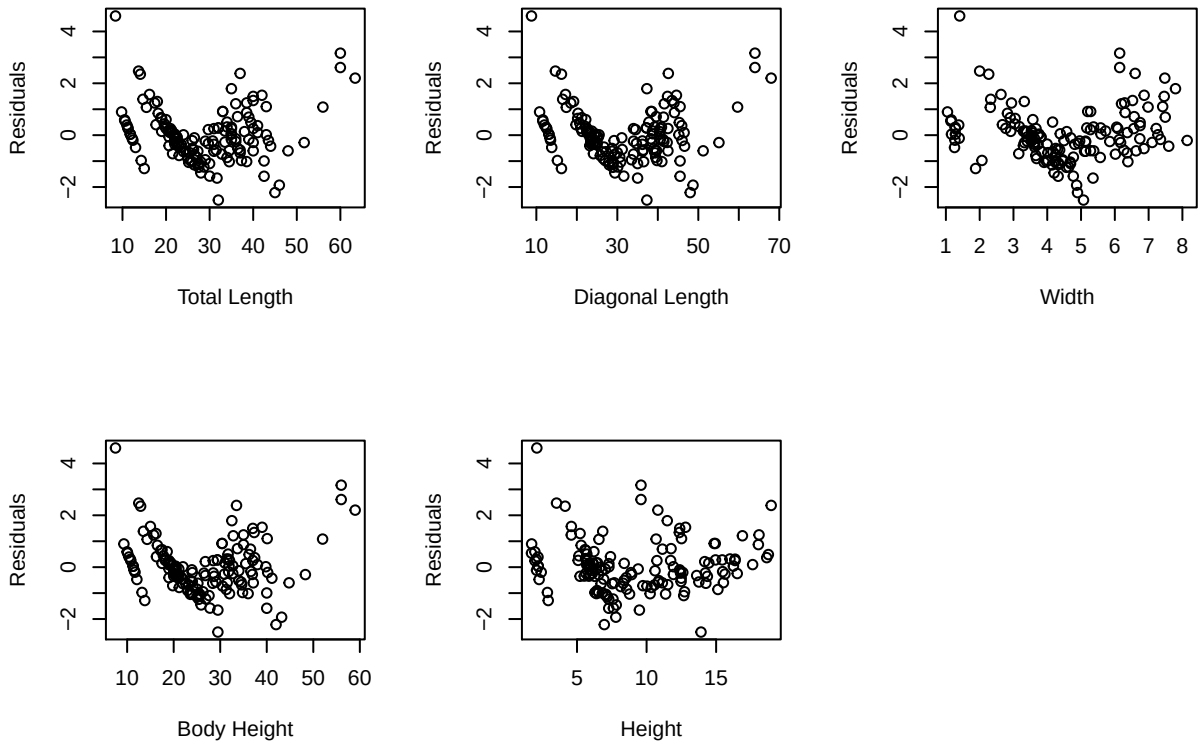
```
vif(model3)
```

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

From our result, we can see that the VIF < 10, and therefore we are not exhibiting any multicollinearity in our model3.

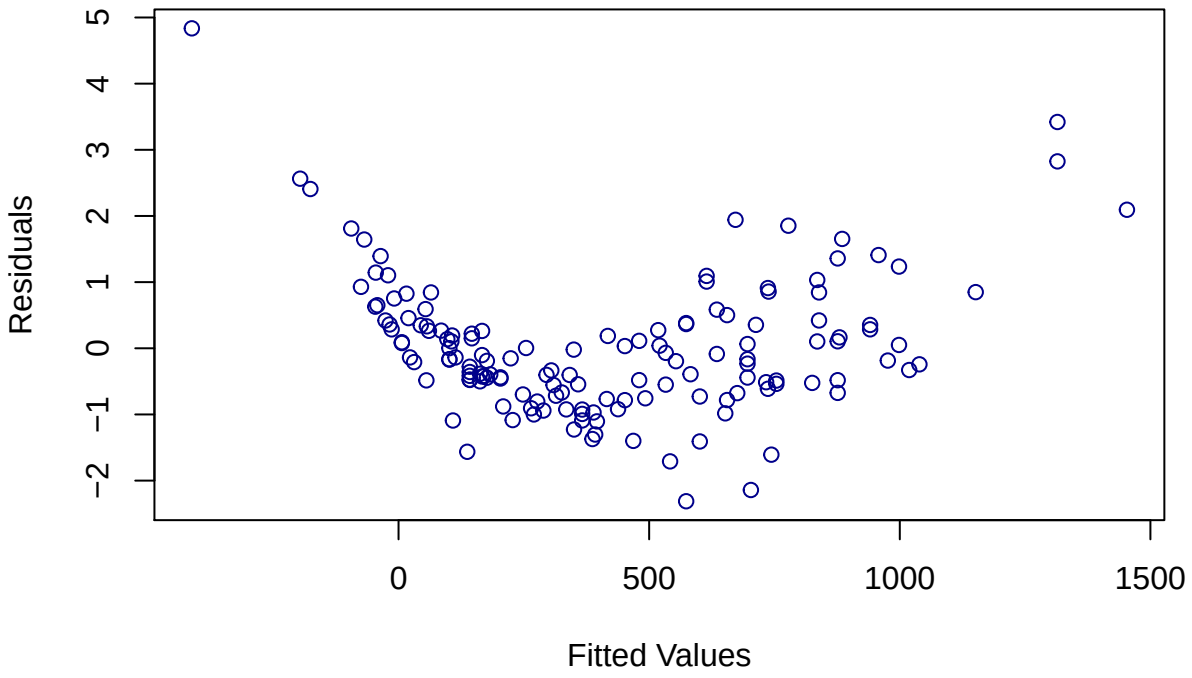
(b) Conduct residual analysis for model3 (similar to Q4). Comment on each assumption and whether they hold.

```
fits3 = model3$fitted.values
resids3 = stdres(model3)
par(mfrow=c(2,3))
plot(fish2$Total.Length,resids3,xlab='Total Length',ylab='Residuals')
plot(fish2$Diagonal.Length,resids3,xlab='Diagonal Length',ylab='Residuals')
plot(fish2$Width,resids3,xlab='Width',ylab='Residuals')
plot(fish2$Body.Height,resids3,xlab='Body Height',ylab='Residuals')
plot(fish2$Height,resids3,xlab='Height',ylab='Residuals')
```



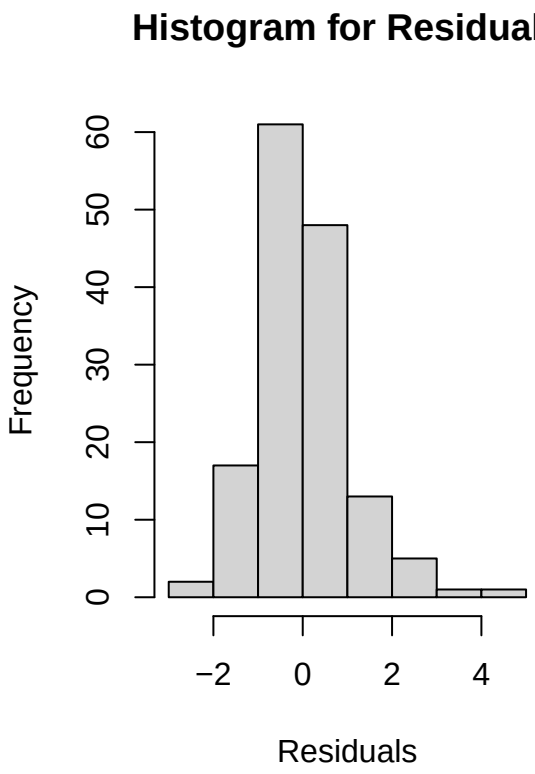
Similar to the previous result, we can see that the observations are spread across the 0 line, and therefore our linearity assumptions holds.

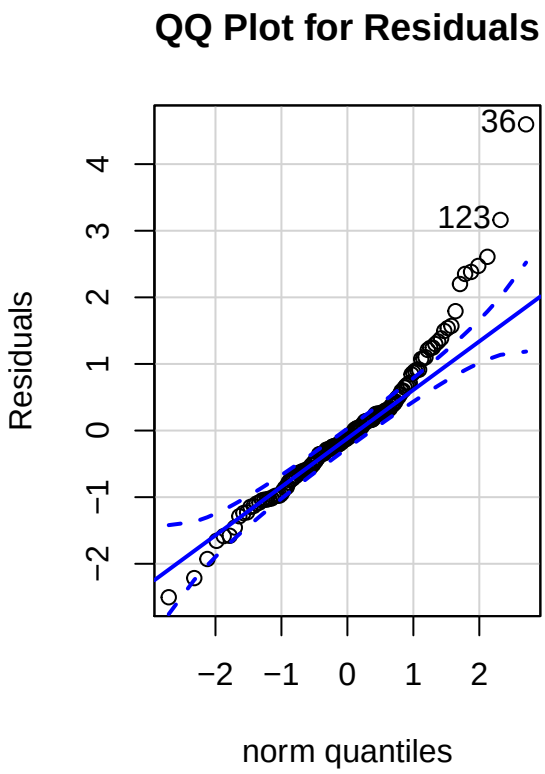
```
plot(fits3,resids,xlab='Fitted Values',ylab='Residuals',col='darkblue')
```



Based on the result of our plot, we can see that the variance seems to be spread out evenly along the 0 line and don't seem to form any cone shapes as our values increase. There does seem to be a few outlier values > 3, but the general range of values seems to -2 to 2, which is consistent across most of the plot. When we look at values from 200 to 750 on the x-axis, we could argue that a cone-shape is starting to emerge, however this seems to flatten out > 750 and a constant variance emerges again. Therefore, overall I will say the constant variance assumption holds.

```
library(car)
par(mfrow=c(1,2))
hist(resids3,xlab='Residuals',main="Histogram for Residuals")
qqPlot(resids3,ylab='Residuals',main="QQ Plot for Residuals")
```





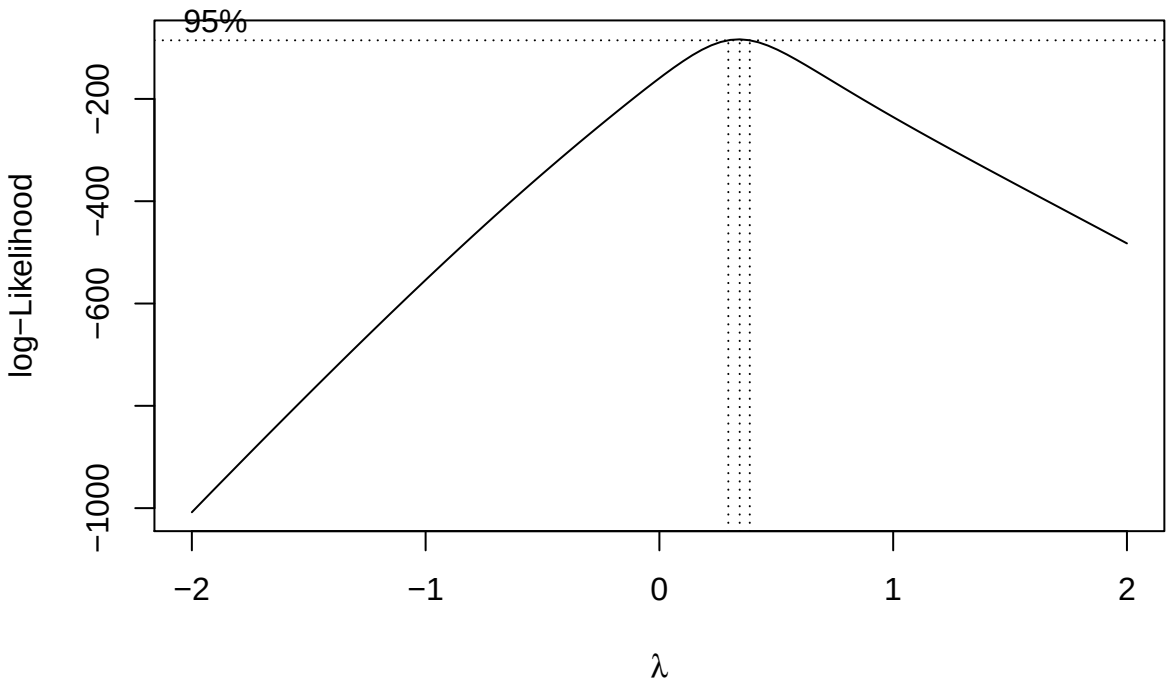

```
## [1]  36 123
```

Similar to the previous result, from our results we can see that the normality assumption does not hold as our histogram is show a slight skewness, which is reaffirmed in our qqplot as there seems to be a significant tale forming at the end of the plot.

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

```
bc <- boxcox(model3,data=fish2)
```




```
bc$x[which.max(bc$y)]
```

```
## [1] 0.3434343
```

When we look for the highest lambda value on this chart and round to the nearest 0.5, we can see that our lambda value = 0.5. Therefore, the suggest transformation would be to take the square root of the response variable.

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

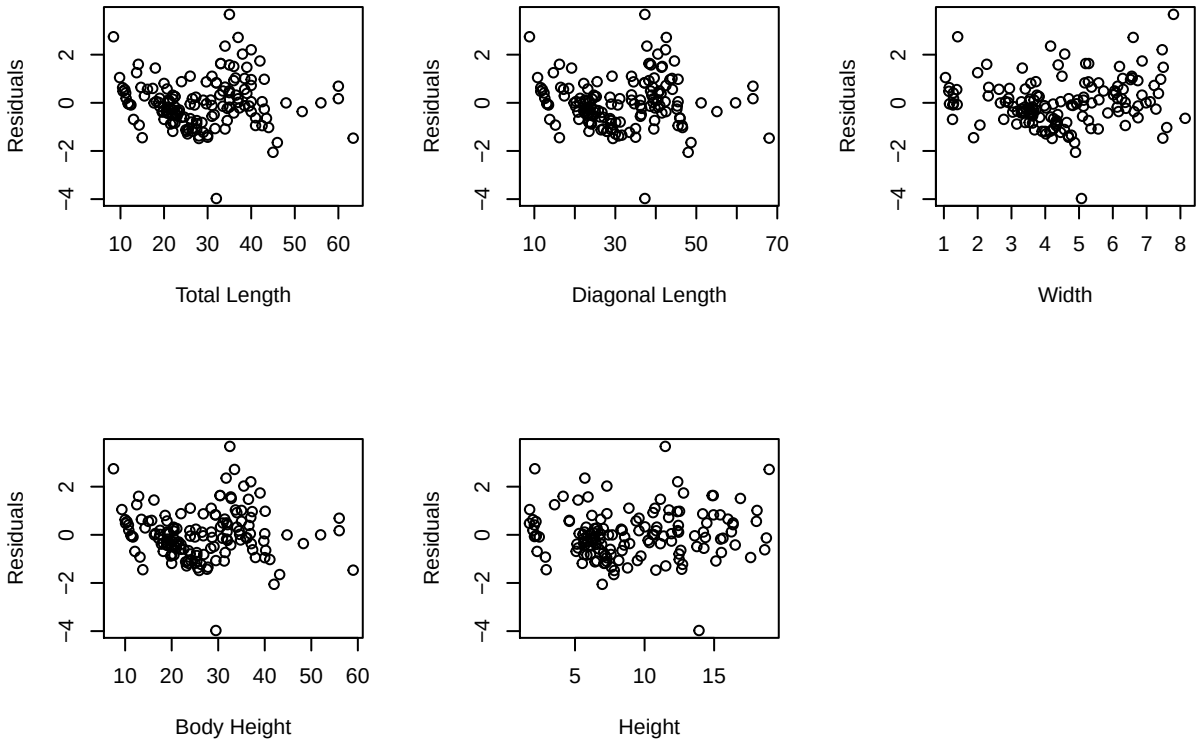
```
model4 <- lm(sqrt(Weight)~Species+Total.Length,data=fish2)
summary(model4)
```

```
##
## 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
```

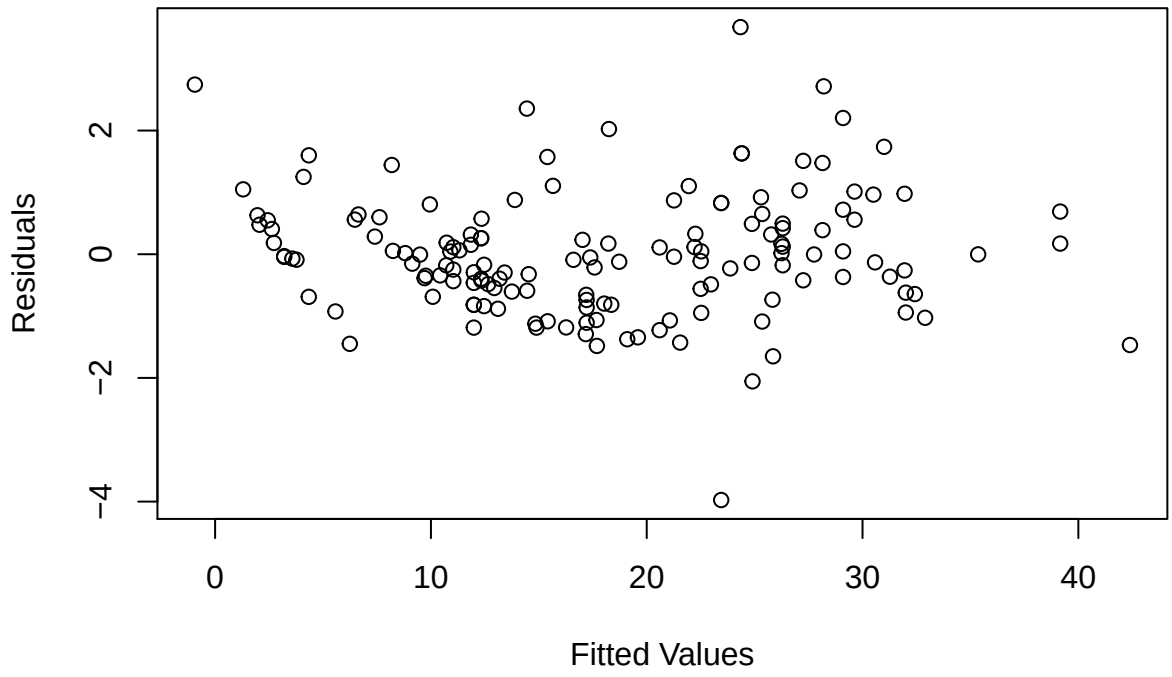
(c) Perform Residual Analysis on model4. Comment on each assumption. Was the transformation successful/unsuccessful?

```
fits4 = model4$fitted.values
resids4 = stdres(model4)
par(mfrow=c(2,3))
plot(fish2$Total.Length,resids4,xlab='Total Length',ylab='Residuals')
plot(fish2$Diagonal.Length,resids4,xlab='Diagonal Length',ylab='Residuals')
plot(fish2$Width,resids4,xlab='Width',ylab='Residuals')
plot(fish2$Body.Height,resids4,xlab='Body Height',ylab='Residuals')
plot(fish2$Height,resids4,xlab='Height',ylab='Residuals')
```



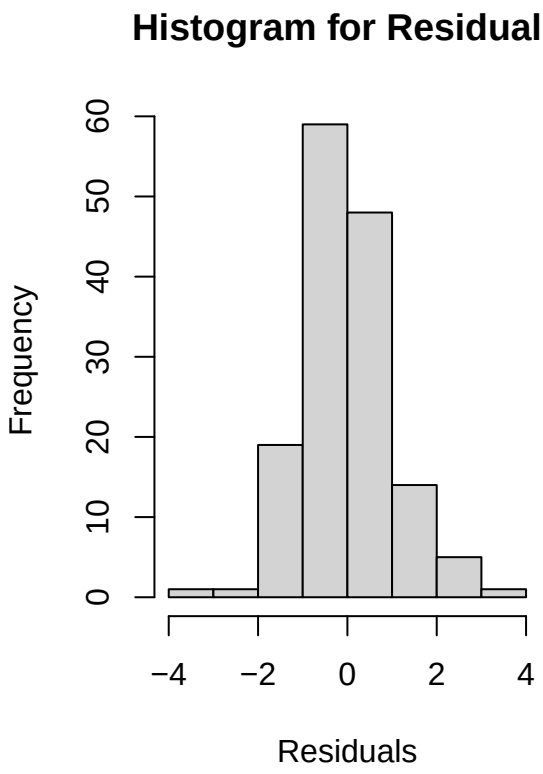
Based on our results we can see that linearity assumption holds as the results are spreadout evenly across 0.

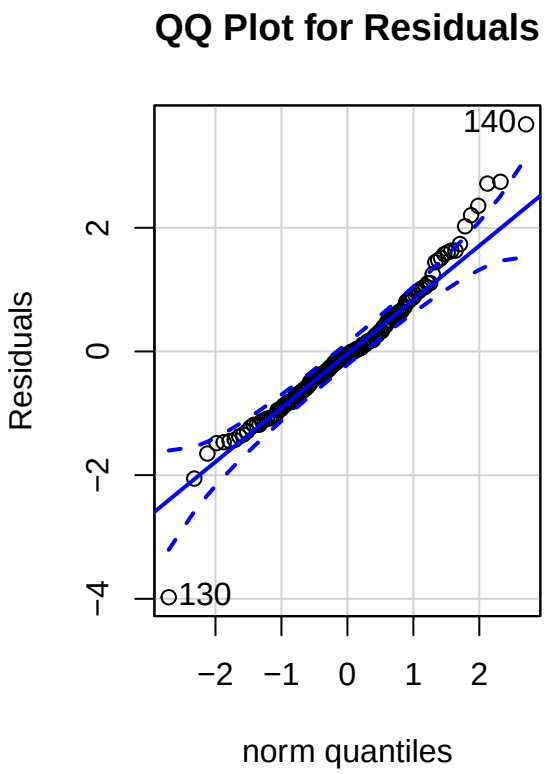
```
plot(fits4,resids4,xlab='Fitted Values',ylab='Residuals')
```



From our result we can see that we have an improvement when assessing constant variance assumption, as the observations with a residual value > 3 seems to be fewer compared to the previous plot, therefore showing an even strong confinement to the (-2,2) range.

```
par(mfrow=c(1,2))
hist(resids4,xlab='Residuals',main="Histogram for Residuals")
qqPlot(resids4,ylab='Residuals',main="QQ Plot for Residuals")
```






```
## [1] 130 140
```

Comparing our results from before, we can see that normality assumption now holds for model4, as the skewness as been significantly reduced in the histogram and we don't see a significant number of observations outside the tail of our QQplot compared to before.

We can say that model4 holds all assumptions.

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.

```
par(mfrow=c(1,3))
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ Height + Body.Height + Total.Length + Diagonal.Length +
##     Height + Width + Species, 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 ***
## Height              10.000     13.398   0.746 0.456692
## 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
## Width               -8.339     24.483  -0.341 0.733924
## 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
## ---
## 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
```

```
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
```

```
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
```

Based on our results, we can see that Model4 has the best performance, with an Adjusted R2 of 0.9808. Following model2 with 0.9335 and the last model3 with 0.9321. However, notice that there's only a 0.014 difference between model2 and model3's performance.

We can see that when removing the correlated variables, there doesn't seem to be a significant difference in the model performance. Meaning that a majority of the variance in Weight can be explained by the Species and Total Length variable. The boxcox transformation also seemed to improve the distribution of our data in order for the normality assumption to hold, which lead to a significant uptake in the model performance.

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.

```
pred_model3 <- predict(model3, fishtest)
pred_model4 <- predict(model4, fishtest)
mean((fishtest$Weight - pred_model3) ^ 2)
```

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

```
mean((fishtest$Weight - pred_model4^2) ^ 2)
```

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

Based on our results, because we applied a sqrt to the response variable to model4, we need to make sure we reverse this operation by squaring the prediction results. This will then provide a result at the same level of units as the Weight variable.

When we compare the two results, we can see that Model4 performs significantly better with a smaller MSPE score of 2442.9 compared to model3 of 9392.25.

(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.

```
# Code to predict point and prediction interval...
fishtest2 <- fishtest[1,]
fishtest2['Species']='Perch'
fishtest2['Body.Height']=28
fishtest2['Total.Length']=32

predict(model4, fishtest2, interval="prediction",level = 0.90)^2
```

```
##          fit      lwr      upr
## 150 461.9429 374.4536 558.6091
```

Based on our results, when we square the result of our fitted value, we retrieve a fitted value of 461.94 for the weight with a lower bound of 374 and an upperbound of 558 for the prediction interval at 0.9 level.