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)
```

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

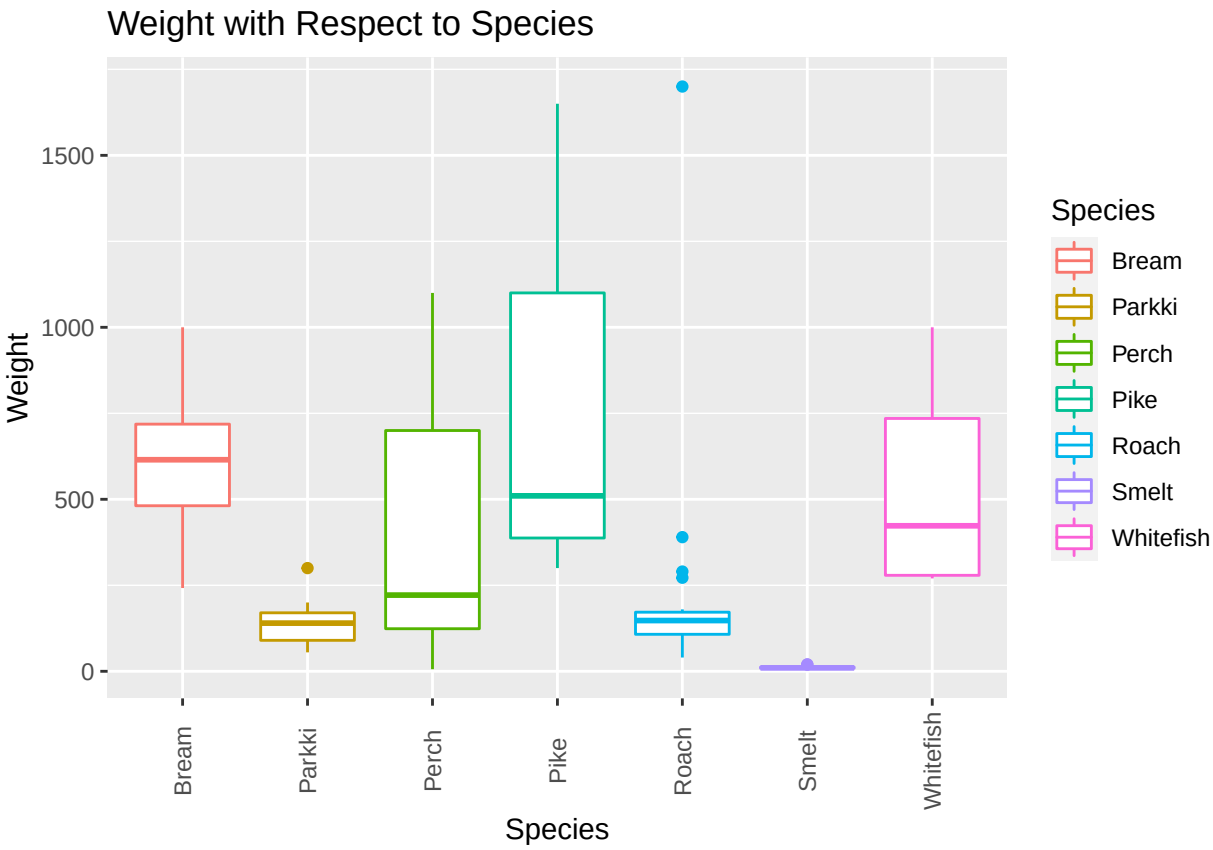
```
# Read the data set
fishfull = read.csv("Fish.csv",header=T, fileEncoding = 'UTF-8-BOM')
row.cnt = nrow(fishfull)
# Split the data into training and testing sets
fishtest = fishfull[(row.cnt-9):row.cnt,]
fish = fishfull[1:(row.cnt-10),]
```

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?

```
library(ggplot2)
ggplot(fish, aes(x=Species, y=Weight, color=Species),
       xlab = "Species",
       ylab = "Weight") +
  geom_boxplot() +
  theme(axis.text.x  = element_text(angle=90, vjust=0.5))+
  ggtitle("Weight with Respect to Species")
```



Answer

- Bream - Almost normal with a slight skew towards lower weight. with not a lot of variations
- Parkki - Almost normal with a slight skew towards lower weight. with not a lot of variations. There might be some potential outliers
- Perch - the data is not normally distributed.. there is a long tail towards higher weights.
- Pike - It is similar to Perch but with a longer tail towards higher weight
- Roach - Alomst normal distribution with potentialy a lot of outliers
- Smelt - The weight is not distributed over a long range. It might also have potential outliers
- Whitefish - Similar to Perch and Pike.. with a tail towards higher weight

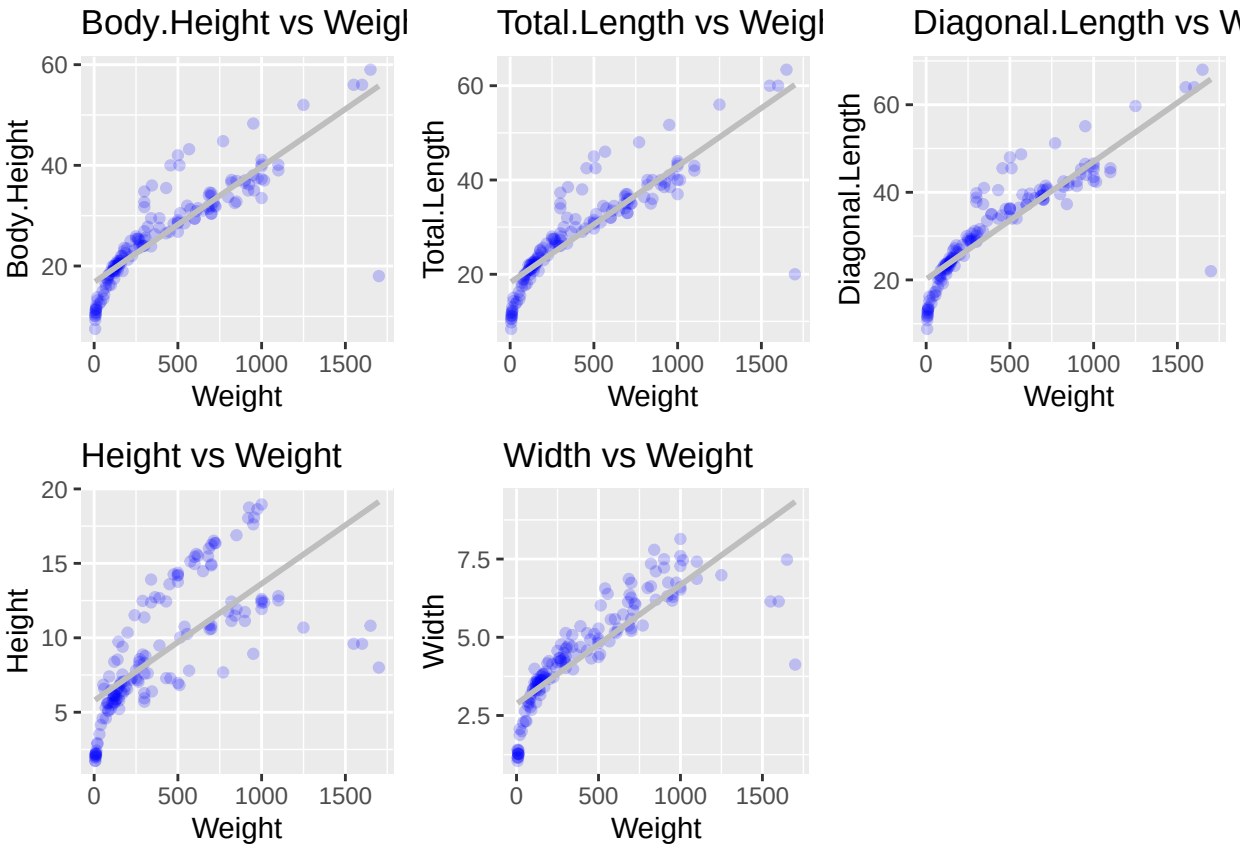
We might need some transformation for Perch, Pike and Whitefish.

(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?**

```
library("grid")
library("ggplot2")
library("gridExtra")
par(mfrow=c(2,3))
plot(fish$Weight, fish$Body.Height)
plot(fish$Weight, fish$Total.Length)
plot(fish$Weight, fish$Diagonal.Length)
plot(fish$Weight, fish$Height)
plot(fish$Weight, fish$Width)
plot1 = ggplot(data=fish, aes(x=Weight, y=Body.Height)) + geom_point(alpha=I(0.2),color='blue') +
  xlab('Weight') + ylab('Body.Height') + ggtitle('Body.Height vs Weight') +
  geom_smooth(method= "lm",color='gray', se=FALSE)
```

```
plot2 = ggplot(data=fish, aes(x=Weight, y=Total.Length)) + geom_point(alpha=I(0.2),color='blue') +
  xlab('Weight') + ylab('Total.Length') + ggtitle('Total.Length vs Weight') +
  geom_smooth(method= "lm",color='gray', se=FALSE)
plot3 = ggplot(data=fish, aes(x=Weight, y=Diagonal.Length)) + geom_point(alpha=I(0.2),color='blue') +
  xlab('Weight') + ylab('Diagonal.Length') + ggtitle('Diagonal.Length vs Weight') +
  geom_smooth(method= "lm",color='gray', se=FALSE)
plot4 = ggplot(data=fish, aes(x=Weight, y=Height)) + geom_point(alpha=I(0.2),color='blue') +
  xlab('Weight') + ylab('Height') + ggtitle('Height vs Weight') +
  geom_smooth(method= "lm",color='gray', se=FALSE)
plot5 = ggplot(data=fish, aes(x=Weight, y=Width)) + geom_point(alpha=I(0.2),color='blue') +
  xlab('Weight') + ylab('Width') + ggtitle('Width vs Weight') +
  geom_smooth(method= "lm",color='gray', se=FALSE)
grid.arrange(plot1, plot2, plot3, plot4, plot5, ncol=3)
```

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



Answer

Linearity Assumption holds for all variables.. there is strong linear relationship with all the predictors. Though the data for the plots show a curve towards the lower part of the Weight, so we might need some transformation.

(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.

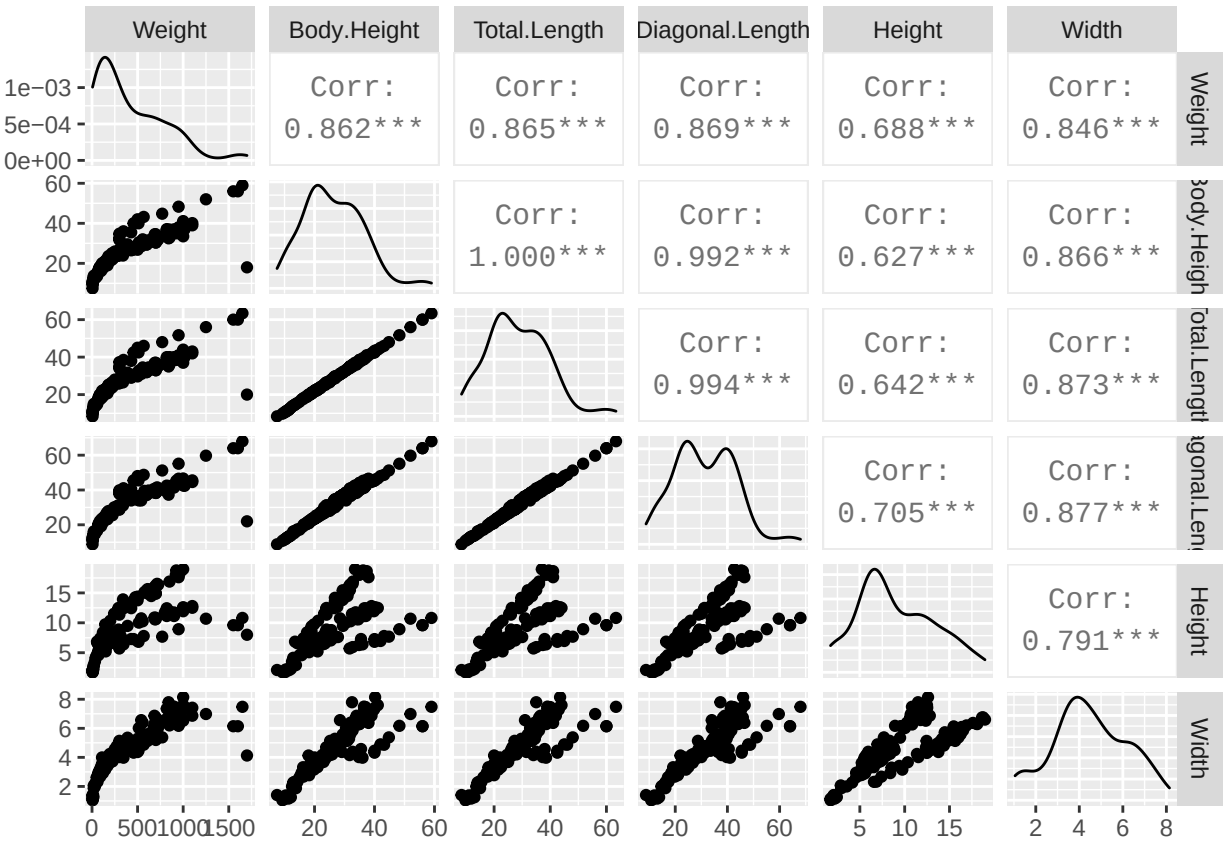
```
#library(corrplot)
#res <- cor(fish[ ,!(colnames(fish) == "Species")])
#corrplot(res )

cor(fish[ ,!(colnames(fish) == "Species")])
```

```
##                    Weight Body.Height Total.Length Diagonal.Length    Height
## Weight          1.0000000   0.8616894    0.8654773       0.8688250 0.6879801
## Body.Height     0.8616894   1.0000000    0.9995134       0.9919502 0.6268604
## Total.Length    0.8654773   0.9995134    1.0000000       0.9940896 0.6422261
## Diagonal.Length 0.8688250   0.9919502    0.9940896       1.0000000 0.7052116
## Height          0.6879801   0.6268604    0.6422261       0.7052116 1.0000000
## Width           0.8456717   0.8661882    0.8728030       0.8770361 0.7908491
##                     Width
## Weight          0.8456717
## Body.Height     0.8661882
## Total.Length    0.8728030
## Diagonal.Length 0.8770361
## Height          0.7908491
## Width           1.0000000
```

```
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(fish[ ,!(colnames(fish) == "Species")])
```



Answer

- There is a moderate to high correlation between Weight and all the other predictors rnging from (0.688 - 0.869)
- Correlation between Total.Length and Body.Height is 1
- Body.Height has a
  - very high correlated to Diagonal.Length and Width
  - High correlation with Width and Weight
  - Low to moderate correlation to Height
- Total.Length
  - is highly correlated to Diagonal.Length and Width
  - high correlation to Weight and Width
  - Low to Moderate correlation to Height
- Diagonal.Length
  - has Very high correlation with Body.Height and Totla.Length
  - High correlation to Width and Weight
  - Moderate correlatin to Height
- Finally Height is also highly correlate to Width

(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 : Based on the analysis i would still recommend a multiple linear regrssion model for the relationship between *Weight* an other predicting Variables.

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
```

```
#anova(model1)
```

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

Answer : Final regresson p-value: $<$ 2.2e-16. this means the that the overall regresson is significant

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

Answer: - estimated coefficient for *Body.Height* is -176.87, This means that there is an inverse relationship to weight *provided all other variables are fixed.* If all the other predictors as same, we observe that for every unit of change in the Body.Height, there will be -176.87 times change in the Weight

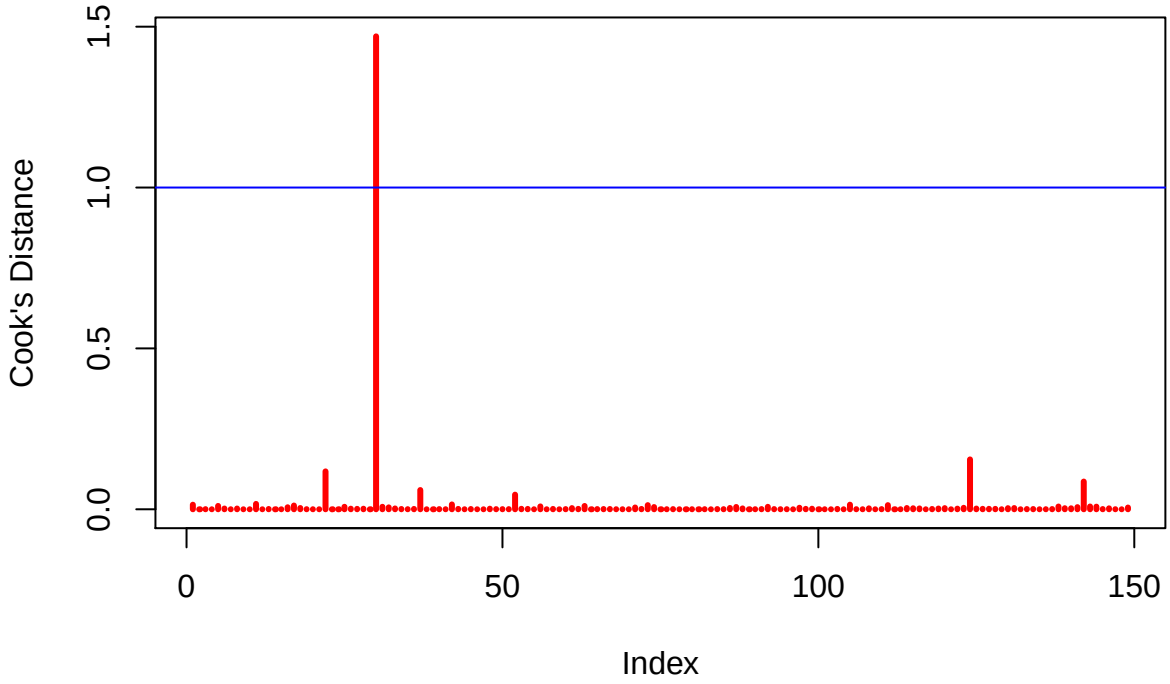
(d) What is the coefficient estimate for the *Species* category Parkki? Interpret this coefficient.

Answer : the estimated coefficient for *Species* category Parkki is 79.34. this means that if the species is Parakki, then the intercept of the regression line will be -813.90 + 79.34 = -734.56. So the intercept for the regression line changes to -734.56.

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.

```
library(car)
cook = cooks.distance(model1)
plot(cook,type="h",lwd=3,col="red", ylab= "Cook's Distance")
abline(1,0,col="blue")
```




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

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

Answer : Row number for the outlier - 30

(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 ~ Species + Total.Length + Body.Height + Diagonal.Length + Height + Width, data = fis
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ Species + Total.Length + Body.Height +
##     Diagonal.Length + Height + Width, 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
## Total.Length       74.822     48.319   1.549 0.123825
## Body.Height       -76.321     37.437  -2.039 0.043422 *
## 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?

```
model2.r2 = summary(model2)$r.squared
model2.r2
```

```
## [1] 0.9384836
```

```
model2.threshold = max(10, 1/(1-model2.r2))
model2.threshold
```

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

```
model2.vif_values = car::vif(model2)
model2.vif_values
```

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

Answer : $max(10, 1/(1-R^2) = 16.256$. Based on this value... i think there is multicollinearity between all the predictors , with height and width having less compared to others

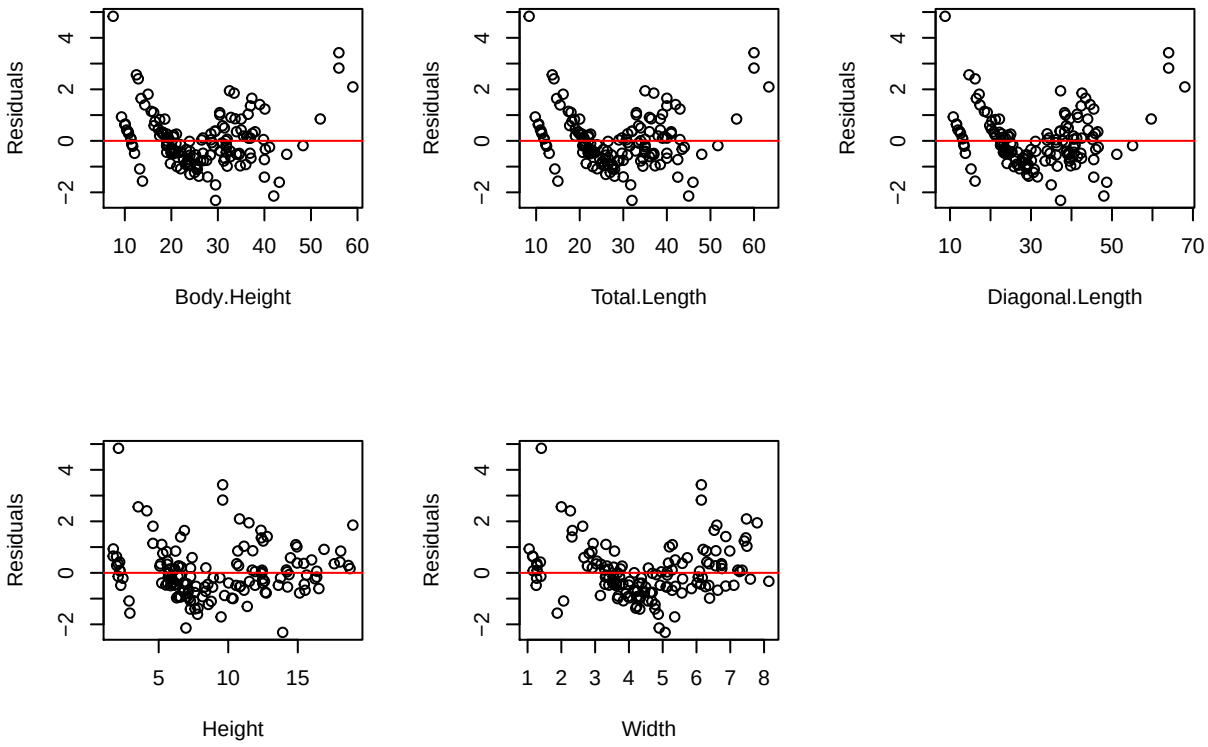
Question 4: Checking Model Assumptions [9 points]

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)
model2.resids= stdres(model2)
par(mfrow=c(2,3))
plot(fish2$Body.Height,model2.resids,xlab="Body.Height",ylab="Residuals")
```

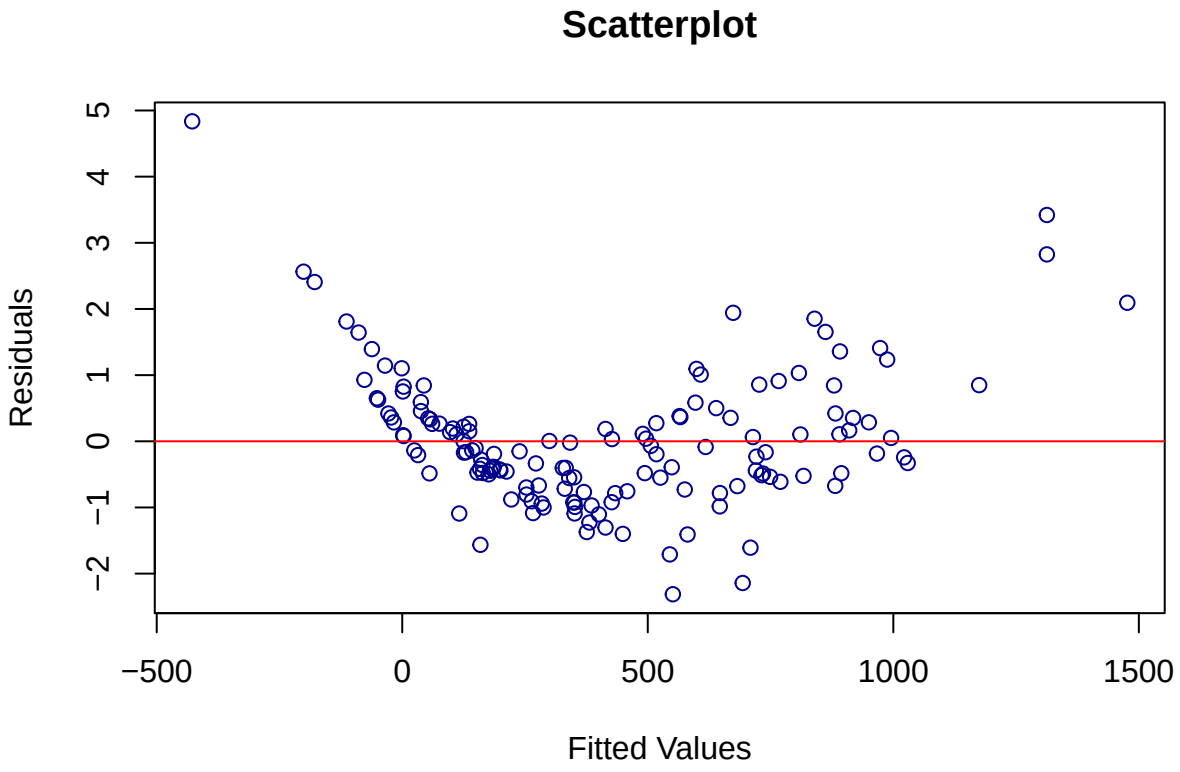
```
abline(0,0,col="red")
plot(fish2$Total.Length,model2.resids,xlab="Total.Length",ylab="Residuals")
abline(0,0,col="red")
plot(fish2$Diagonal.Length,model2.resids,xlab="Diagonal.Length",ylab="Residuals")
abline(0,0,col="red")
plot(fish2$Height,model2.resids,xlab="Height",ylab="Residuals")
abline(0,0,col="red")
plot(fish2$Width,model2.resids,xlab="Width",ylab="Residuals")
abline(0,0,col="red")
```



Answer : Body.Height, Total.Length and Diagonal.Length do not show a random pattern around the 0 line. On the other hand Height and Weight show random distribution.. though the data points are more concentrated around the lower values

(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?

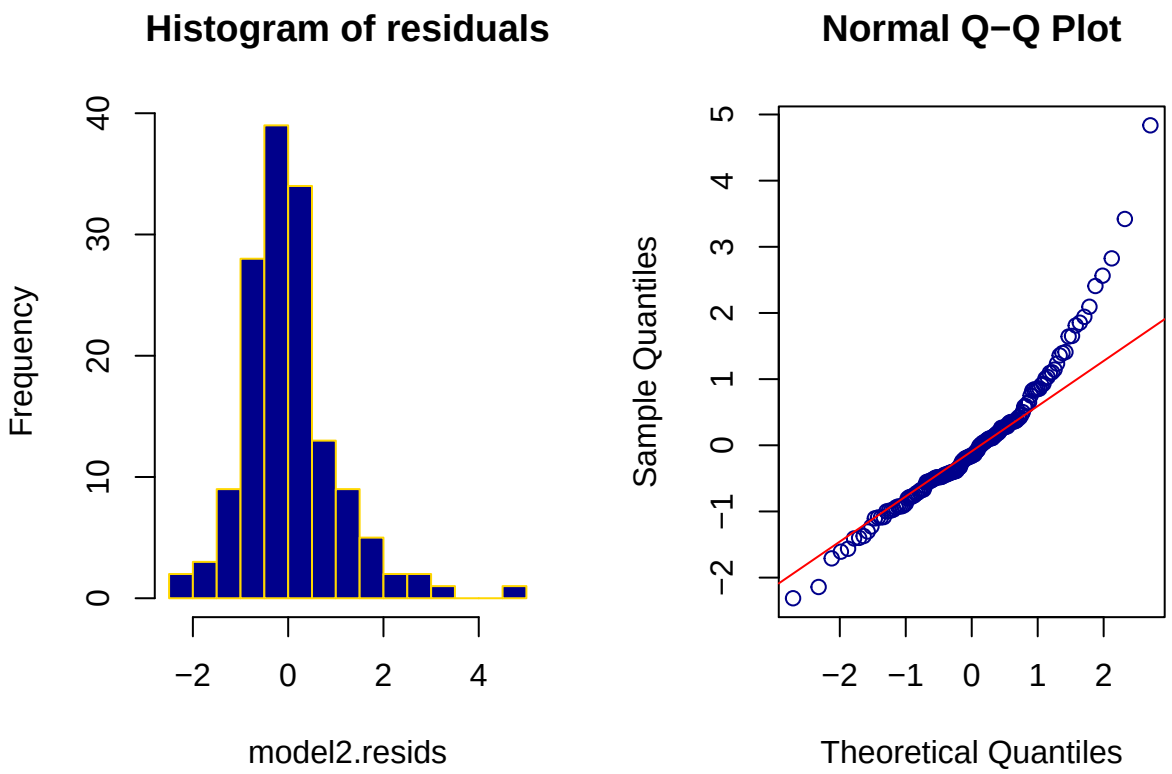
```
model2.resids = rstandard(model2)
model2.fits = model2$fitted
plot(model2.fits, model2.resids, xlab="Fitted Values", ylab="Residuals", main="Scatterplot",
     col="darkblue")
abline(0,0,col="red")
```



Answer : the constant variance does not seem to hold true as the variance is not consistent.. the variance is not distributed around the 0 line. The variance is higher for < 0 and > 1000.

(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(model2.resids,nclass=20, col="darkblue", border="gold",main="Histogram of residuals")
qqnorm(model2.resids,col="darkblue")
qqline(model2.resids,col="red")
```



Answer : The residual do not seem to be normally distributed. They have a long right tail. . . and looks like also have an outlier. QQ plot shos that the residuals have a right tailed distribution. There might be a need to transform the predictors and/or the response.

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 + Total.Length + Body.Height + 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 : the F Value is 1.7626 and the p-value is 0.14. At α level of 0.01 we fail to reject the null hypothesis that the coefficients of all the extra predictors in Model 1 are all 0. There is a significant probability that all the extra coefficients(Body.Height, Diagonal.Length, Height & Width) in model 1 are = 0.

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

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

```
model3.r2 = summary(model3)$r.squared
model3.r2
```

```
## [1] 0.9352946
```

```
model3.threshold = max(10, 1/(1-model3.r2))
model3.threshold
```

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

```
model3.vif_values = car::vif(model3)
model3.vif_values
```

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

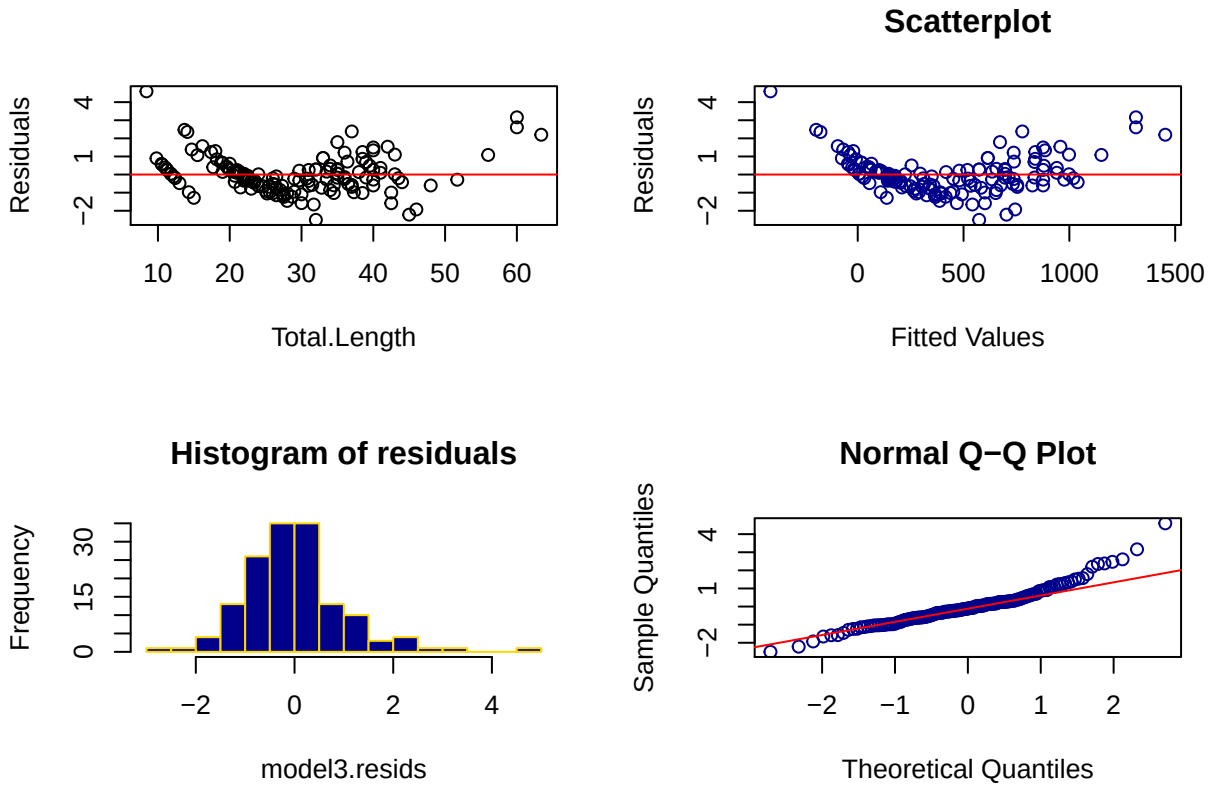
```
#barplot(model3.vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue")
#abline(v = model3.threshold, lwd = 3, lty = 2)
```

Answer : *max(10, 1/(1-R^2)* = 15.45466. Based on this value... Since the vif value for all coefficients is less than 15.455, there is no significant correlation between the predictors.

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

```
library(MASS)
model3.resids= stdres(model3)
```

```r
model3.fits = model3$fitted

par(mfrow=c(2,2))
plot(fish2$Total.Length,model3.resids,xlab="Total.Length",ylab="Residuals")
abline(0,0,col="red")

plot(model3.fits, model3.resids, xlab="Fitted Values", ylab="Residuals", main="Scatterplot",col="darkblu
abline(0,0,col="red")

hist(model3.resids,nclass=20, col="darkblue", border="gold",main="Histogram of residuals")
qqnorm(model3.resids,col="darkblue")
qqline(model3.resids,col="red")
```

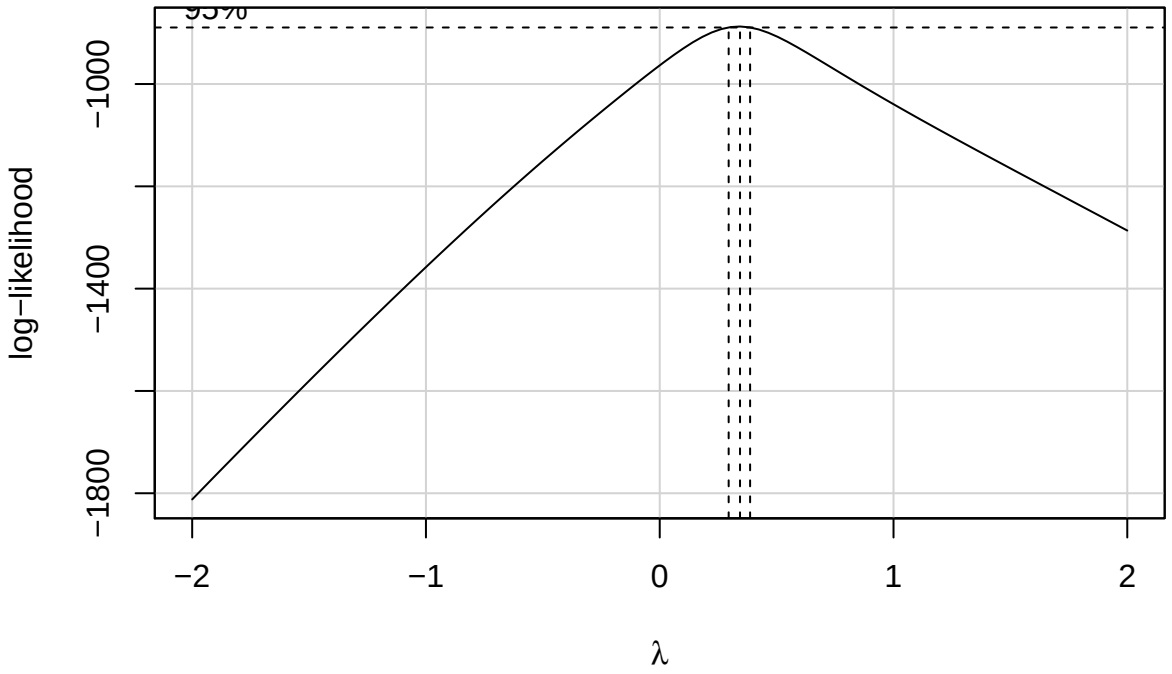


Answer : The residual plots against Total.Length does not show a random distribution around 0. Does not look like the linearity assumption holds. The Residual vs fitted value does not show a random distribution around 0. Constant Variance assumptions does not seem to hold. Residuals do not strictly follow the normality plot. The plot looks right tailed. Looks like we would beed a some data transformation.

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

```
b = boxCox(model3)
```




```
lambda <- b$x # lambda values
lik <- b$y # log likelihood values for SSE
bc <- cbind(lambda, lik) # combine lambda and lik
sorted_bc <- bc[order(-lik),] # values are sorted to identify the lambda value for the maximum log like
head(sorted_bc, n = 10)
```

```
##          lambda       lik
##  [1,] 0.3434343 -887.5466
##  [2,] 0.3030303 -888.6628
##  [3,] 0.3838384 -889.1869
##  [4,] 0.2626263 -892.6394
##  [5,] 0.4242424 -893.3920
##  [6,] 0.2222222 -899.1985
##  [7,] 0.4646465 -899.7915
##  [8,] 0.5050505 -907.8777
##  [9,] 0.1818182 -907.9427
## [10,] 0.5454545 -917.1797
```

Answer : Base on the boxcox transformation ... looks like λ value of 0.5 should be used

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

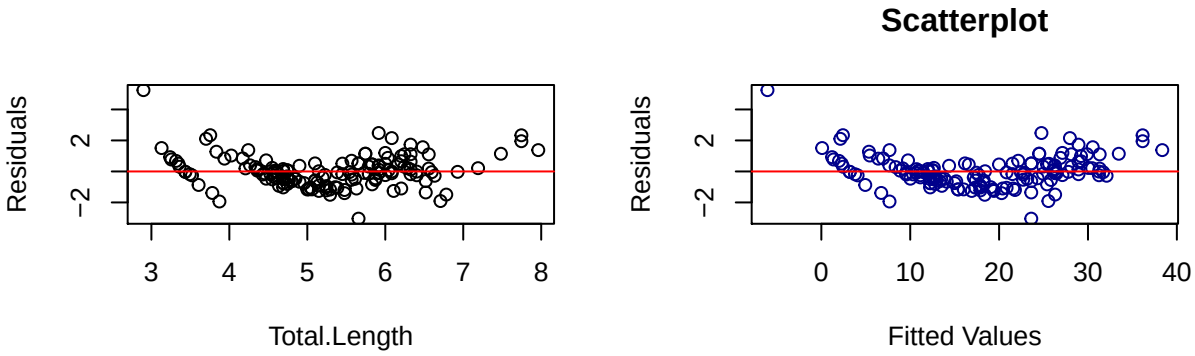
```
model4 = lm(sqrt(Weight) ~ Species +  sqrt(Total.Length), data = fish2)
summary(model4)
```

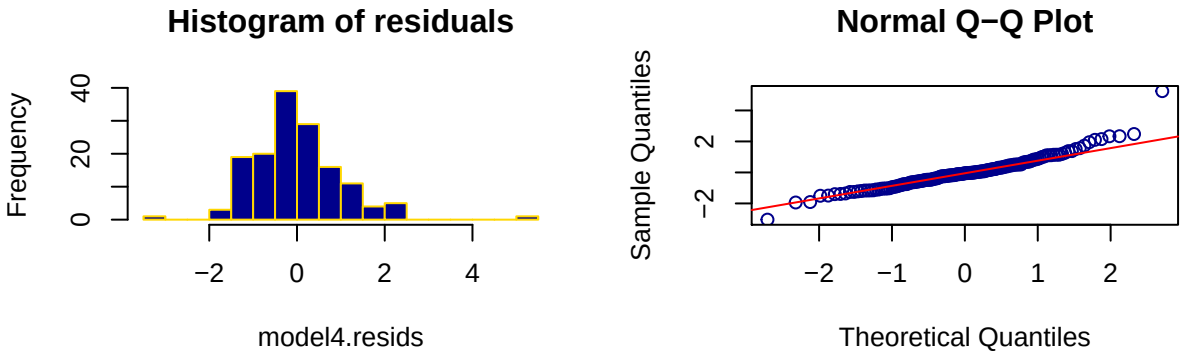
```
##
## Call:
## lm(formula = sqrt(Weight) ~ Species + sqrt(Total.Length), data = fish2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -5.1790 -0.9774 -0.0938  0.8348  8.4907
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -34.1650     1.3821 -24.719  < 2e-16 ***
## SpeciesParkki        0.3555     0.7164   0.496 0.620476
## SpeciesPerch        -1.5017     0.3991  -3.762 0.000247 ***
## SpeciesPike         -8.8115     0.5875 -14.997  < 2e-16 ***
## SpeciesRoach        -1.7616     0.5529  -3.186 0.001778 **
## SpeciesSmelt         2.2761     0.7809   2.915 0.004144 **
## SpeciesWhitefish    -0.5554     0.7653  -0.726 0.469211
## sqrt(Total.Length)  10.2147     0.2347  43.520  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.726 on 140 degrees of freedom
## Multiple R-squared:  0.9668, Adjusted R-squared:  0.9651
## F-statistic: 581.6 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?

```
model4.resids= stdres(model4)
model4.fits = model4$fitted

par(mfrow=c(2,2))
plot(sqrt(fish2$Total.Length),model4.resids,xlab="Total.Length",ylab="Residuals")
abline(0,0,col="red")

plot(model4.fits, model4.resids, xlab="Fitted Values", ylab="Residuals", main="Scatterplot",col="darkblu
abline(0,0,col="red")

hist(model4.resids,nclass=20, col="darkblue", border="gold",main="Histogram of residuals")
qqnorm(model4.resids,col="darkblue")
qqline(model4.resids,col="red")
```





Answer : Looks like the transformation has improved the linearity, normality and constant variance .. the Residuals look to have tails on both the sides and some outliers.

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.

Answer :

Model	R^2	$Adj - R^2$
Model2	0.9384836	0.933508
Model3	0.9352946	0.9320593
Model4	0.9667547	0.9650924

Model 2 & Model 3 have cery similar R^2 values but Model 4 has a higher R^2 as well as $Adj - R^2$. Model 4 does a better job of explaining the variance in the model.

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.

```
fishtest
```

```
##     Weight Species Body.Height Total.Length Diagonal.Length  Height  Width
## 150  650.0   Perch        36.5         39.0            41.4 11.1366 6.0030
## 151  450.0   Bream        27.6         30.0            35.1 14.0049 4.8438
```

```
## 152  273.0  Parkki        23.0         25.0            28.0 11.0880 4.1440
## 153   78.0   Perch        16.8         18.7            19.4  5.1992 3.1234
## 154  145.0   Perch        22.0         24.0            25.5  6.3750 3.8250
## 155   40.0   Perch        13.8         15.0            16.0  3.8240 2.4320
## 156  200.0    Pike        30.0         32.3            34.8  5.5680 3.3756
## 157  540.0    Pike        40.1         43.0            45.8  7.7860 5.1296
## 158  150.0  Parkki        18.4         20.0            22.4  8.8928 3.2928
## 159    9.8   Smelt        11.4         12.0            13.2  2.2044 1.1484
```

```
model3.pred = predict(model3, fishtest, interval = 'prediction')
model4.pred = predict(model4, fishtest, interval = 'prediction')
model3.pred
```

```
##             fit        lwr       upr
## 150  835.320989  644.30713 1026.3348
## 151  492.283575  301.85813  682.7090
## 152  351.535704  153.48729  549.5841
## 153    7.581312 -183.04331  198.2059
## 154  223.690686   34.10391  413.2775
## 155 -143.287496 -335.10709   48.5321
## 156  185.102637  -11.41631  381.6216
## 157  621.398920  427.53262  815.2652
## 158  147.658936  -50.02673  345.3446
## 159   14.734838 -179.88560  209.3553
```

```
'^'(model4.pred, 2)
```

```
##           fit          lwr       upr
## 150 790.96465 6.075195e+02 998.57745
## 151 474.50870 3.355834e+02 637.43647
## 152 298.04755 1.865415e+02 435.56061
## 153  72.33930 2.534478e+01 143.42800
## 154 206.63837 1.193886e+02 317.67117
## 155  15.16854 1.536397e-01  54.72142
## 156 227.31047 1.326127e+02 347.37056
## 157 576.27791 4.194367e+02 757.97936
## 158 140.94633 6.848162e+01 239.28383
## 159  12.22118 2.029875e-03  49.51677
```

```
model3.pred1 <-model3.pred[,1]
model4.pred1 <-'^'(model4.pred[,1],2)


model3.mspe = mean((model3.pred1 - fishtest$Weight)^2)
model4.mspe = mean((model4.pred1 - fishtest$Weight)^2)

model3.mspe
```

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

```
model4.mspe
```

```
## [1] 2769.68
```

Answer : MSPE for Model3 is 9392.2496917 and for Mode l4 = 2769.6799582. Model 4 is able to predict the weight closer to the actual weight compare to Model 3.

(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.

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

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

```
##        fit      lwr      upr
## 1 489.1344 369.6697 625.2985
```

Answer : the lower and upper bound for the new prediction with 90 % confidence Interval is 369.67 and 625.30 respectively.