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 measurements.

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(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# 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),]
```

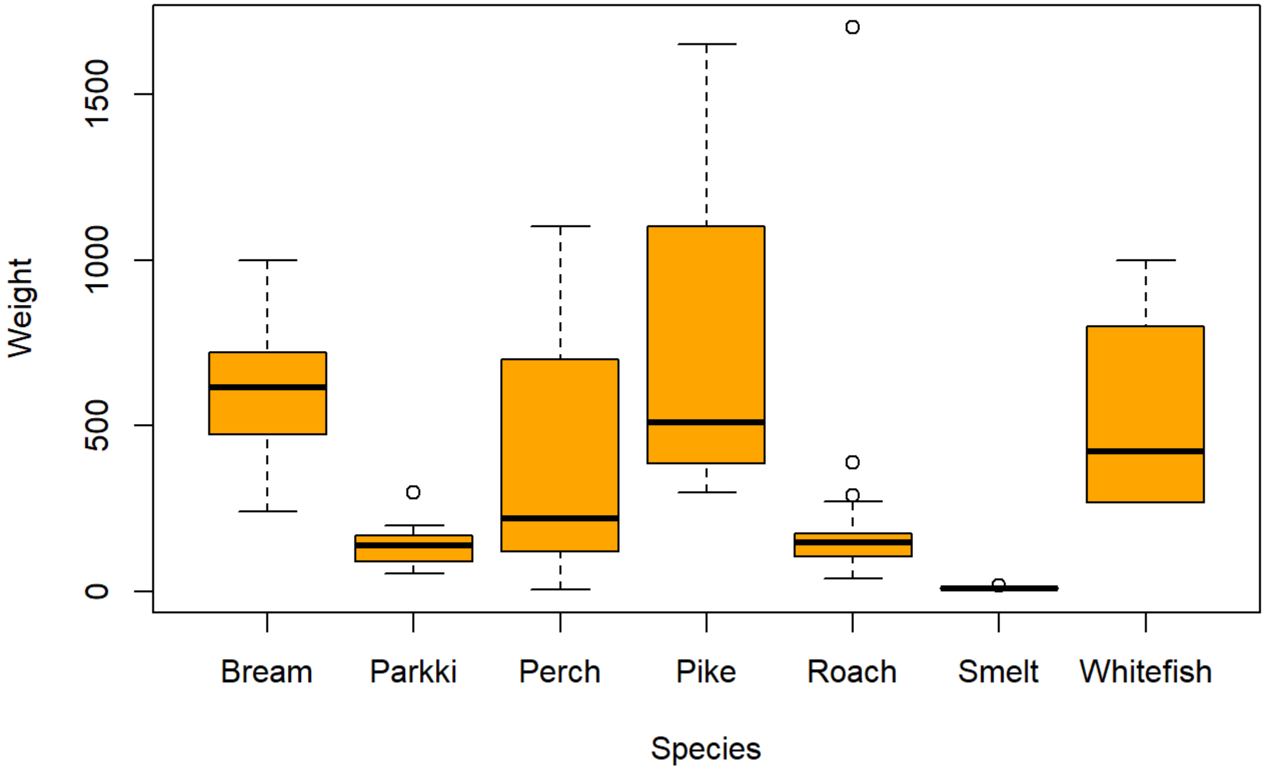
Please use fish as your data set for the following questions unless otherwise stated.

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?

From the below graph we can say that different species have different weights and hence there is some relationship between weight and species.

```
boxplot(Weight~Species, col= "orange", data = fish)
```

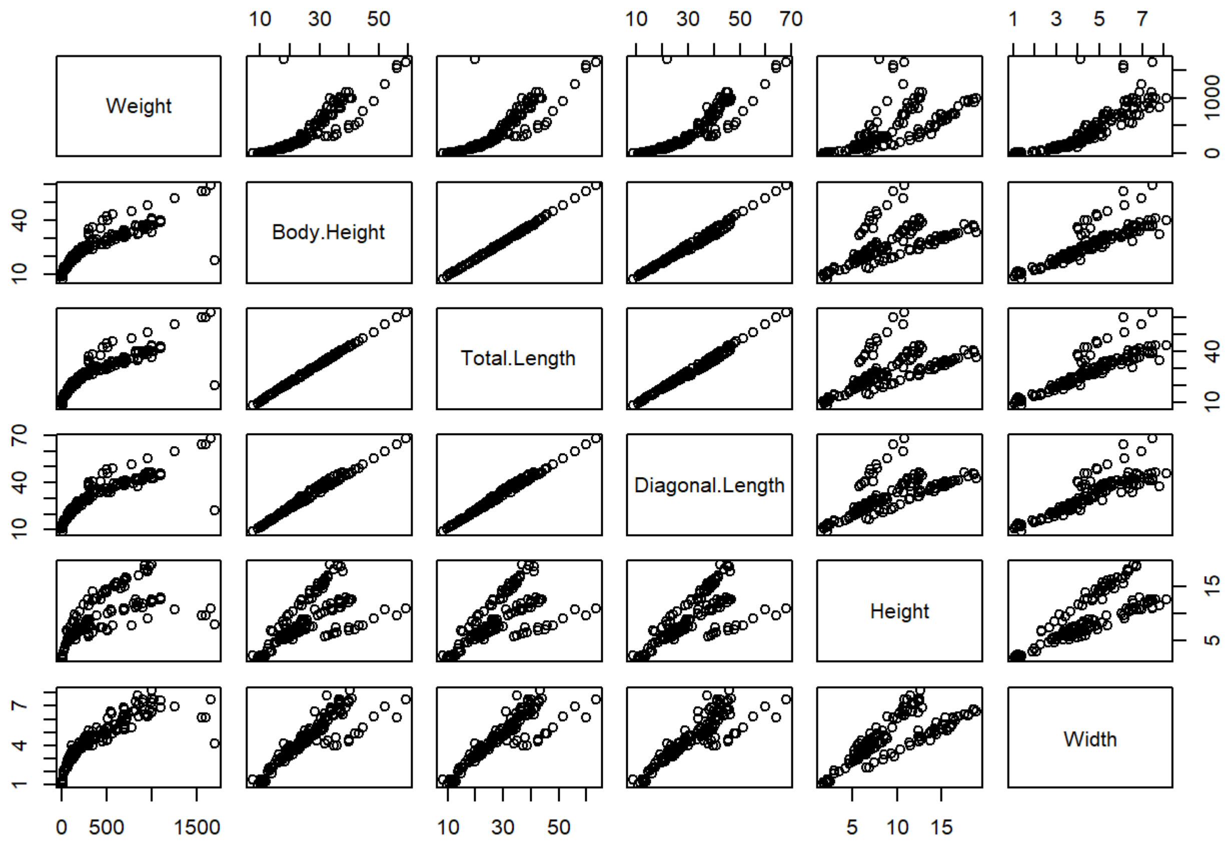


(b) Create plots of the response, *Weight*, against each quantitative predictor, namely Body.Height, Total.Length, Diagonal.Length, Height, **and** Width**. Describe the general trend of each plot. Are there any potential outliers?**

From the below plot we can see that there is positive linear relationship between weight and all other quantitative variables.

There appears to be some potential outliers.

```
new_fish <- fish[,-2]
plot(new_fish[,1:6])
```



(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(new_fish[,1:6])
```

```
##                    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
```

From the above table we can say that all the variables have very strong positive linear relationship with the response variable.
We can also see that predicting variables are strongly related to each other causing multicollinearity. FOr example Body.Height is strongly related to Total.Length and Diagonal.Length. The lowest correlation coefficient is 0.6268604 between Body.Height and Height, which also suggest the strong relationship between them.

(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?

No, based on the above analysis I would not suggest multiple linear regression model for the relationship between Weight and the predictor 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
```

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

Yes, the p-value is 2.2e-16 which is very small compare to 0.01.

(c) What is the coefficient estimate for *Body.Height*? Interpret this coefficient.
The coefficient estimate for Body.Height = -176.87.
Which means that expected additional loss in weight is 176.87 for one unit increase in Body.Height holding all other predictors fixed.

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

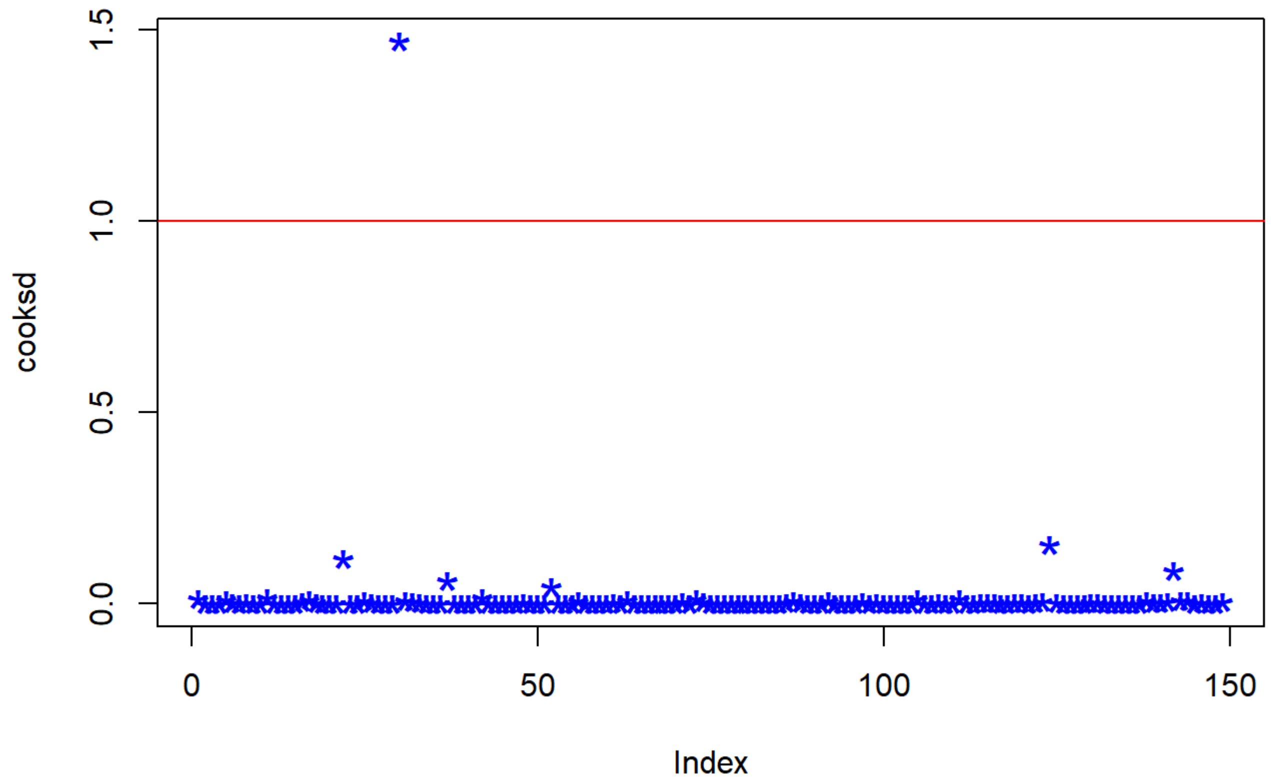
When the species type is Parkki the weight increases by 79.34 unit holding all other variables fixed.

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.

```
cooksd <- cooks.distance(model1)

plot(cooksd, pch = "*", cex= 2, col ="blue")
abline(h=1,col ="red")
```



```
influenttial <- as.numeric(names(cooksd)[cooksd>1])
fish[influenttial, ]
```

	Weight <dbl>	Species <chr>	Body.Height <dbl>	Total.Length <dbl>	Diagonal.Length <dbl>	Height <dbl>	Width <dbl>
30	1700	Roach	18	20	22	8	4.1272

1 row

Row number 30 is the 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.

```
new_datafish <- fish %>% slice(-c(30))
model2 <- lm(Weight~., data= new_datafish)
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = new_datafish)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -211.10  -50.18  -14.44   34.04  433.68
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -969.766    131.601  -7.369 1.51e-11 ***
## SpeciesParkki     195.500     80.105   2.441 0.015951 *
## SpeciesPerch      174.241    124.404   1.401 0.163608
## SpeciesPike      -175.936    140.605  -1.251 0.212983
## SpeciesRoach      141.867     94.319   1.504 0.134871
## SpeciesSmelt      489.714    123.174   3.976 0.000113 ***
## SpeciesWhitefish  122.277     99.293   1.231 0.220270
## Body.Height       -76.321     37.437  -2.039 0.043422 *
## Total.Length       74.822     48.319   1.549 0.123825
## Diagonal.Length    34.349     30.518   1.126 0.262350
## Height             10.000     13.398   0.746 0.456692
## Width              -8.339     24.483  -0.341 0.733924
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 93.84 on 136 degrees of freedom
## Multiple R-squared:  0.9385, Adjusted R-squared:  0.9335
## F-statistic: 188.6 on 11 and 136 DF,  p-value: < 2.2e-16
```

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

```
vif(model2)
```

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

From the above VIF we can say that all the predictors VIF is larger than 10 hence indicates the 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?

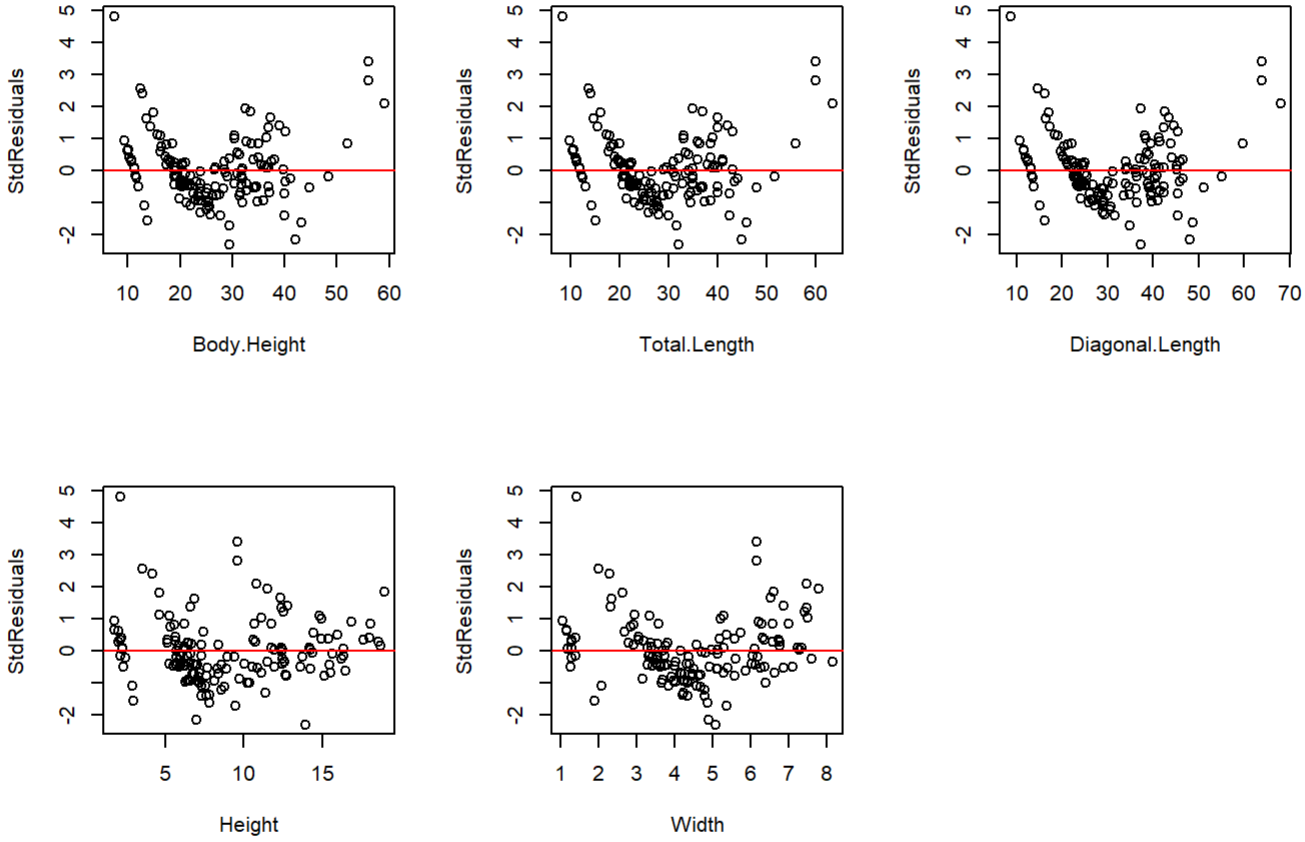
From the below plots it looks like there are some clusters, higher residuals on the edge and lower residuals in the middle.Hence, the linearity assumption doesn’t hold.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

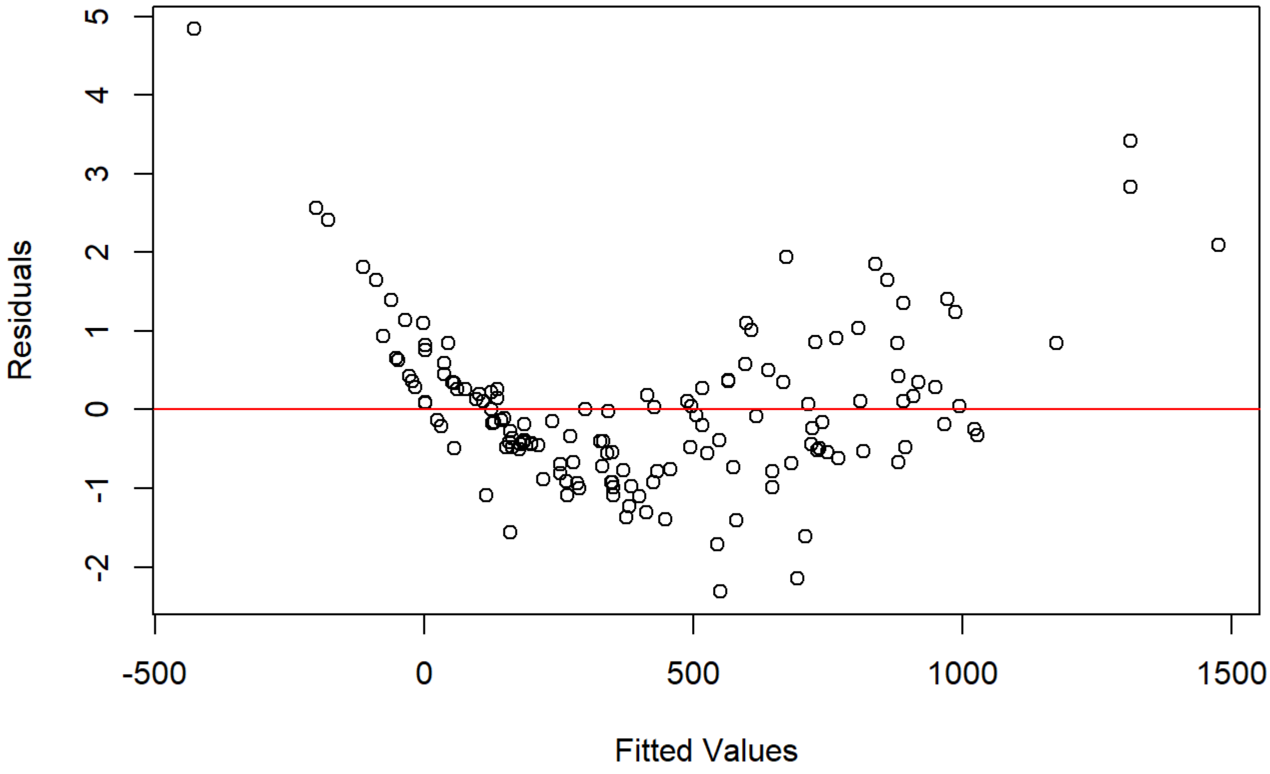
```
resids2 <- stdres(model2)
par(mfrow = c(2,3))
plot(new_datafish[,3],resids2, xlab = "Body.Height", ylab = "StdResiduals")
abline(0,0,col="red")
plot(new_datafish[,4],resids2, xlab = "Total.Length", ylab = "StdResiduals")
abline(0,0,col="red")
plot(new_datafish[,5],resids2, xlab = "Diagonal.Length", ylab = "StdResiduals")
abline(0,0,col="red")
plot(new_datafish[,6],resids2, xlab = "Height", ylab = "StdResiduals")
abline(0,0,col="red")
plot(new_datafish[,7],resids2, xlab = "Width", ylab = "StdResiduals")
abline(0,0,col="red")
```



(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?

From the below plot we can say that the constant variance assumption does not hold and errors appear to be related.

```
fits <- model2$fitted
plot(fits, resids2, xlab ="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
```



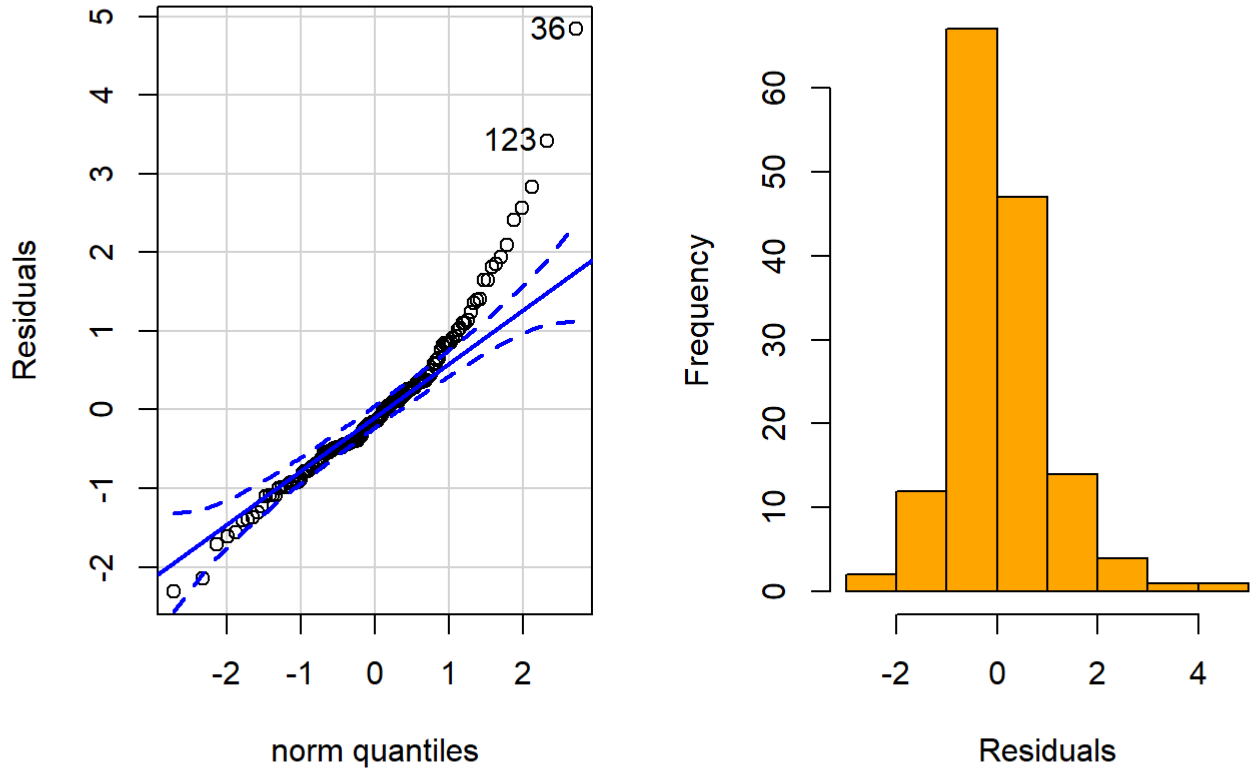
(c) Create a histogram and normal QQ plot for the standardized residuals. What conclusions can you draw from these plots?

From below plots it looks like the data is slightly skewed to the right.Hence, we can say the normality assumption doesn’t hold.

```
par(mfrow = c(1,2))
qqPlot(resids2, ylab="Residuals", main = "")
```

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

```
hist(resids2, xlab="Residuals", main = "",nclass=10,col="orange")
```



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 = new_datafish)
summary(model3)
```

```
##
## Call:
## lm(formula = Weight ~ Species + Total.Length, data = new_datafish)
##
## 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)
```

	Res.Df <dbl>	RSS <dbl>	Df <dbl>	Sum of Sq <dbl>	F <dbl>	Pr(>F) <dbl>
1	140	1259746	NA	NA	NA	NA
2	136	1197659	4	62086.66	1.762561	0.1399744

2 rows

The P-value(0.14) for the partial F-test is higher than alpha level of 0.01, hence we can not reject the null hypothesis that the regression coefficients for Body.Height, Diagonal.Length, Height and Width are zero at alpha level of 0.01

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

(a) Conduct a multicollinearity test on model3. Comment on the multicollinearity in model3.
From the below table we can say the the VIF value for both the variables are very less(2.65) then 10. Hence there is no multicollinearity.

```
vif(model3)
```

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

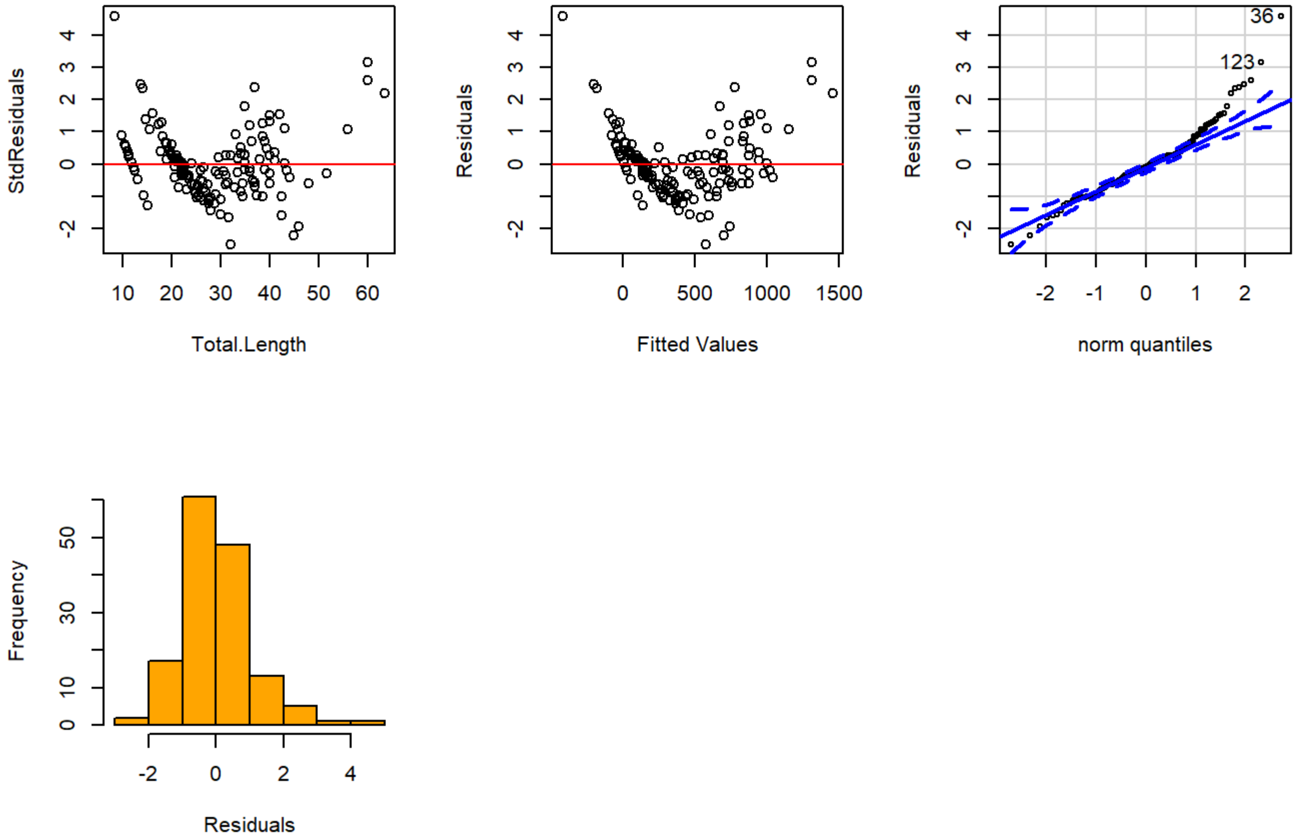
(b) Conduct residual analysis for model3 (similar to Q4). Comment on each assumption and whether they hold.

From the below Total.Length vs Std Residuals graph we can say that linearity assumption doesn’t hold. From the below Fitted Values vs Std Residuals graph we can say that constant variance doesn’t hold. From the qqplot and histogram we can say that residuals are not normally distributed and they have heavy tails hence the normality assumption doesn’t hold.

```
resids3 <- stdres(model3)
fits3 <- model3$fitted
par(mfrow = c(2,3))
plot(new_datafish[,4],resids3, xlab = "Total.Length", ylab = "StdResiduals")
abline(0,0,col="red")
plot(fits3, resids3, xlab ="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
qqPlot(resids3, ylab="Residuals", main = "")
```

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

```
hist(resids3, xlab="Residuals", main = "",nclass=10,col="orange")
```

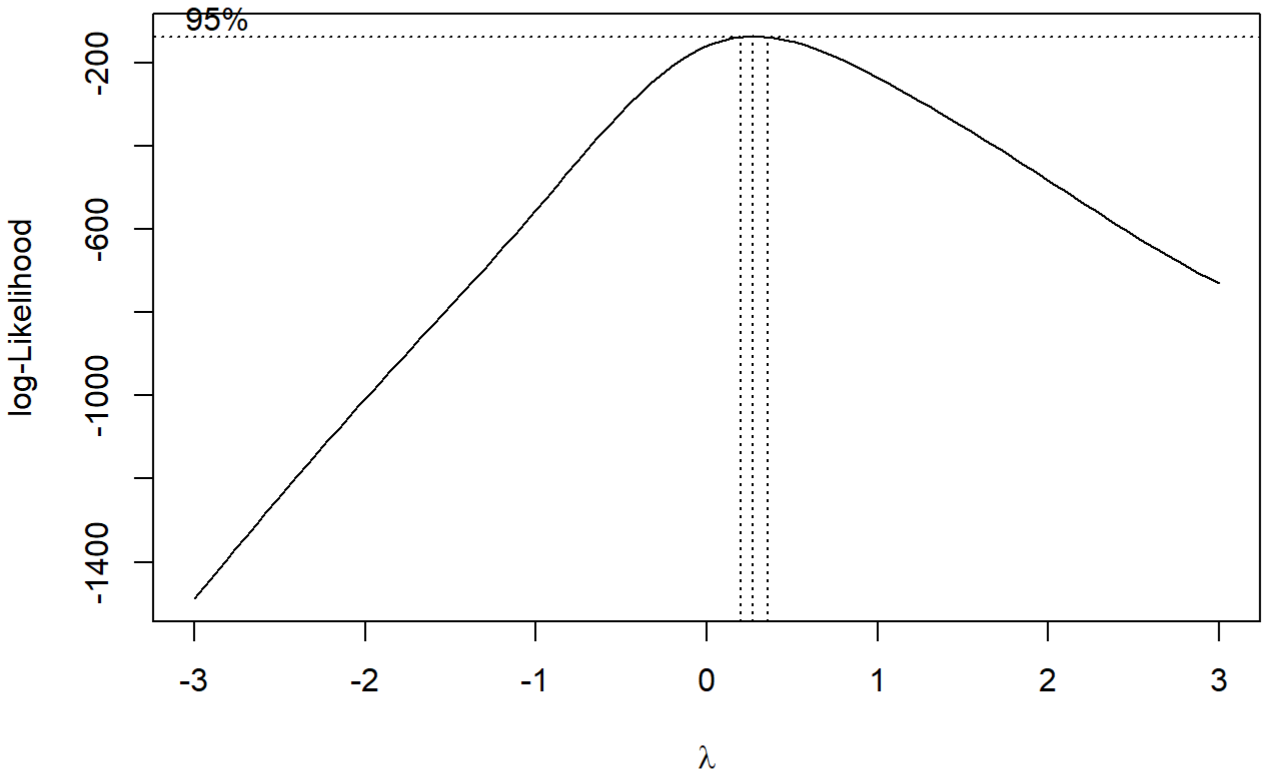


Question 7: Transformation [12 pts]

(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

Based on the below analysis best lambda value is 0.27 rounding it to 0.5, we should apply square root(y) transformation.

```
bc = boxcox(model3, lambda= seq(-3,3))
```



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

```
## [1] 0.2727273
```

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

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

```
##
## Call:
## lm(formula = sqrt(Weight) ~ Species + Total.Length, data = new_datafish)
##
## 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
```

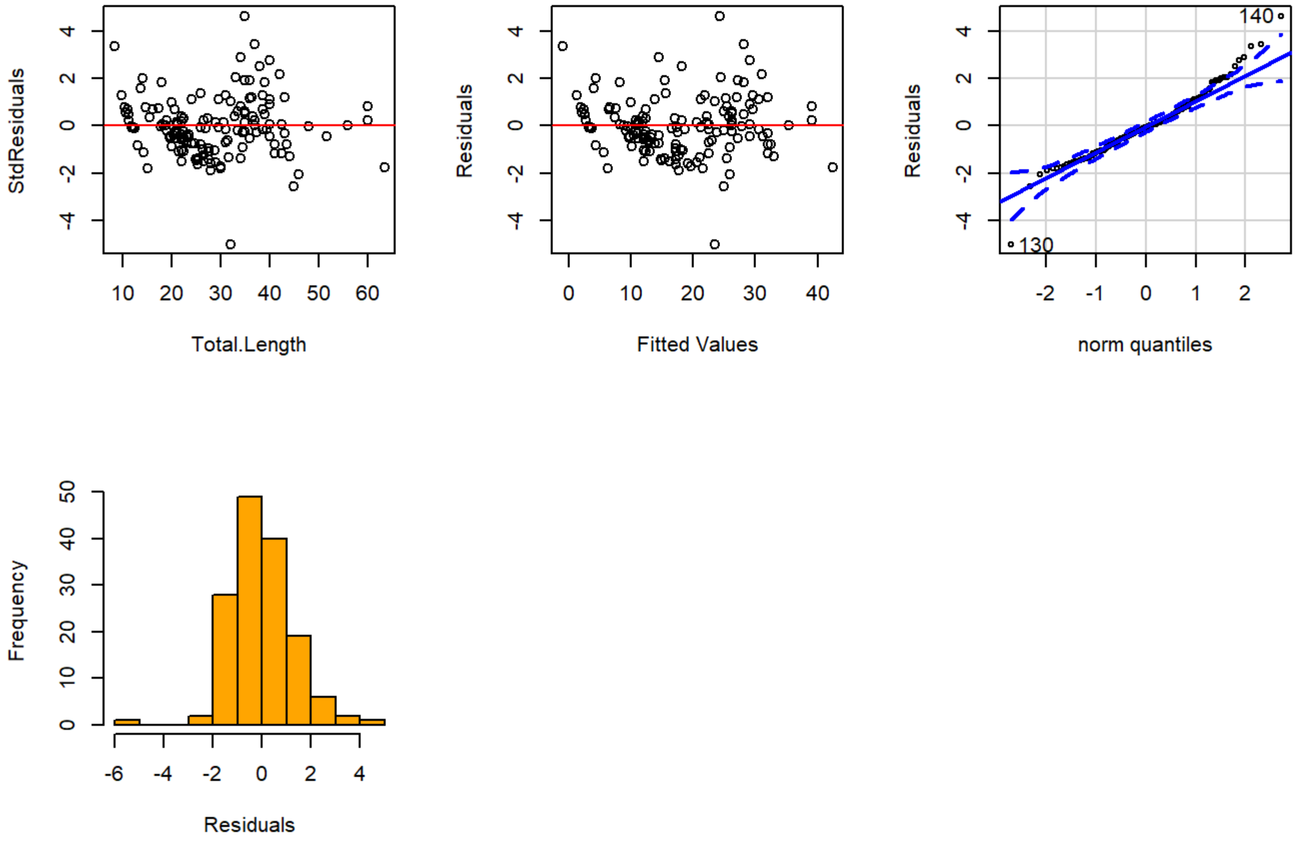
(c) Perform Residual Analysis on model4. Comment on each assumption. Was the transformation successful/unsuccessful?

From the Standard Residuals vs Total.Length scatter plot we can say that the linearity assumption does not hold.
From Fitted values vs std residuals graph we can say that the constant variance assumption holds.
From the qq plot and histogram we can say that the residuals are not normally distributed. Hence the normality assumption does not hold.

```
resids4 <- resid(model4)
fits4 <- model4$fitted
par(mfrow = c(2,3))
plot(new_datafish[,4],resids4, xlab = "Total.Length", ylab = "StdResiduals")
abline(0,0,col="red")
plot(fits4, resids4, xlab ="Fitted Values",ylab="Residuals")
abline(0,0,col="red")
qqPlot(resids4, ylab="Residuals", main = "")
```

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

```
hist(resids4, xlab="Residuals", main = "",nclass=10,col="orange")
```



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.

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

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

```
summary(model2)$adj.r.squared
```

```
## [1] 0.933508
```

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

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

```
summary(model3)$adj.r.squared
```

```
## [1] 0.9320593
```

```
summary(model4)$r.squared
```

```
## [1] 0.9817138
```

```
summary(model4)$adj.r.squared
```

```
## [1] 0.9807995
```

From the above summary we can say that when we applied the transformation on model4 the R-squared and Adjusted R-squared values improved. The model4 explains the 98% of the variance while model2 and model3 explains the 93% of variance. THe Adjusted r-squared value of all models are little less than their r-squared values as adjusted r-squred values add penalties for number of variables used in 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.

```
model3.predict <- predict(model3, fishtest, interval = "prediction")
model3.predict
```

```
##             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.predict <-predict(model4, fishtest, interval = "prediction")
model4.predict
```

```
##           fit        lwr       upr
## 150 28.146519 25.5690732 30.723964
## 151 21.549151 18.9796449 24.118657
## 152 16.432491 13.7601243 19.104857
## 153  8.850901  6.2787076 11.423095
## 154 13.888673 11.3304840 16.446862
## 155  5.333966  2.7456480  7.922284
## 156 12.830454 10.1787260 15.482183
## 157 23.001051 20.3851159 25.616986
## 158 11.679876  9.0124039 14.347347
## 159  3.389796  0.7636846  6.015907
```

```
m3predict <- model3.predict[,1]
mean((m3predict-fishtest$Weight)^2)
```

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

```
m4predict <- model4.predict[,1]
mean((m4predict-fishtest$Weight)^2)
```

```
## [1] 98076.62
```

The MSPE is very high for model4 compare to model3 because we have transformed the response variable. We need to transformed it back to original scale and compare the MSPE with model3.

```
model4T.predict <-(predict(model4, fishtest, interval = "prediction"))^2
model4T.predict
```

```
##           fit         lwr       upr
## 150 792.22652 653.7775038 943.96198
## 151 464.36590 360.2269190 581.70960
## 152 270.02675 189.3410209 364.99556
## 153  78.33845  39.4221692 130.48709
## 154 192.89524 128.3798673 270.49928
## 155  28.45119   7.5385831  62.76258
## 156 164.62056 103.6064626 239.69799
## 157 529.04835 415.5529496 656.22997
## 158 136.41949  81.2234246 205.84637
## 159  11.49071   0.5832142  36.19113
```

```
m4Tpredict <- model4T.predict[,1]
mean((m4Tpredict-fishtest$Weight)^2)
```

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

Now we can compare Model3 and Model4 MSPE. Model3 has higher MSPE while the Model4 has very less MSPE which means that Model4 is better than model3 in predicting true value.

(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)
new_predict <- predict(model4, newdata, interval = "prediction",level = .90)
new_predict
```

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

The estimated average weight of a Perch fish for above given data is 21.49. The 90% confidence interval would be 19.3508 for the lower bound and 23.63491 for the upper bound. We are 90% confident that the mean weight for Perch fish with these specific characteristics is between 19.3508 and 23.63491.