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("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 [8 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?

```
weight.species <- boxplot(Weight ~ Species, data = fish)
```
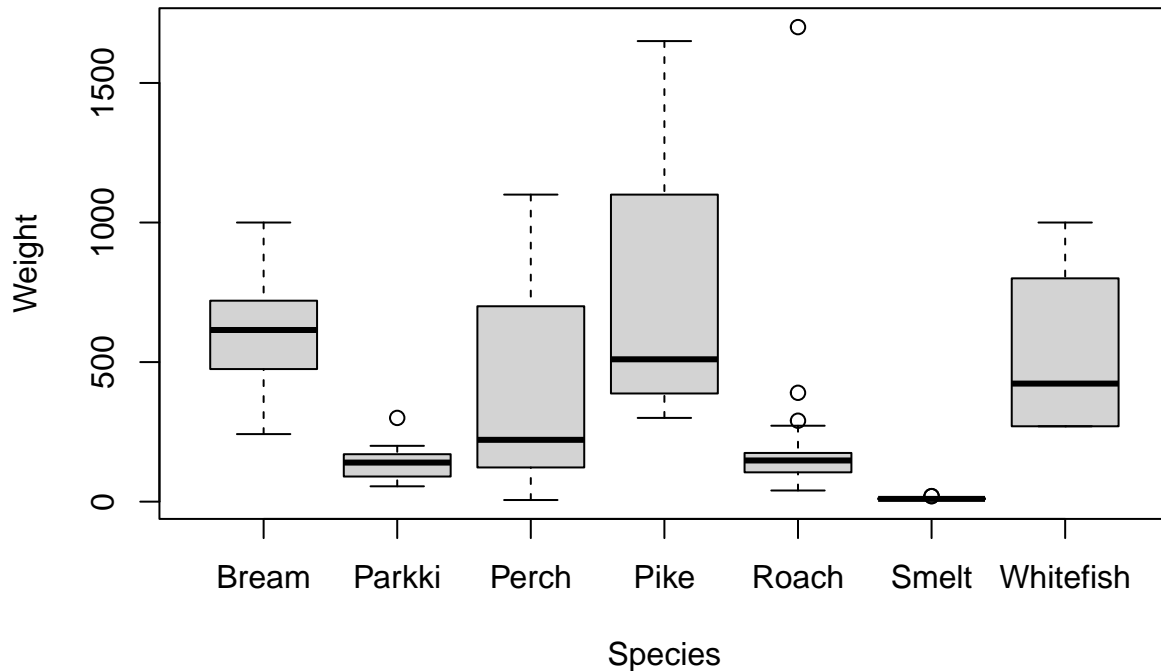

Answer to 1.a: Based on the results from the boxplot there seems to be a relationship between fish species and weight as the means of each fish are different indicating variability between species.

(b) Create scatterplots 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?

```
Weight = fish$Weight
Body.Height = fish$Body.Height
Total.Length = fish$Total.Length
Diagonal.Length = fish$Diagonal.Length
Height = fish$Height
Width = fish$Width

plot(Body.Height, Weight, main ="Width vs Weight")
```
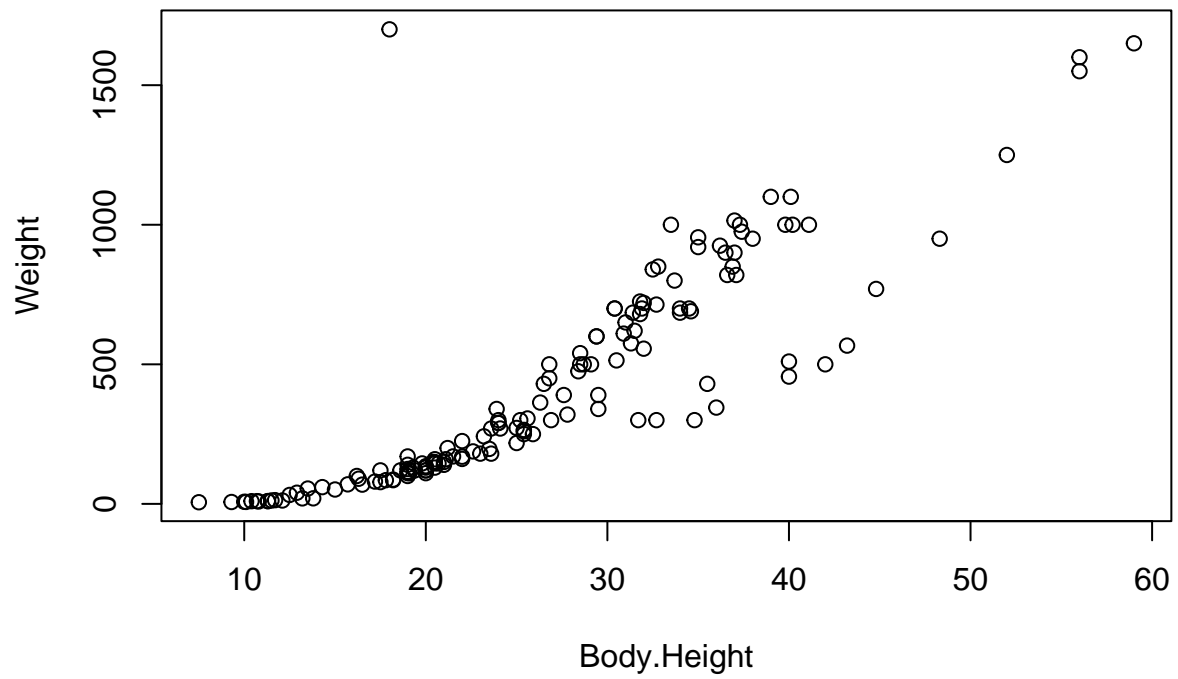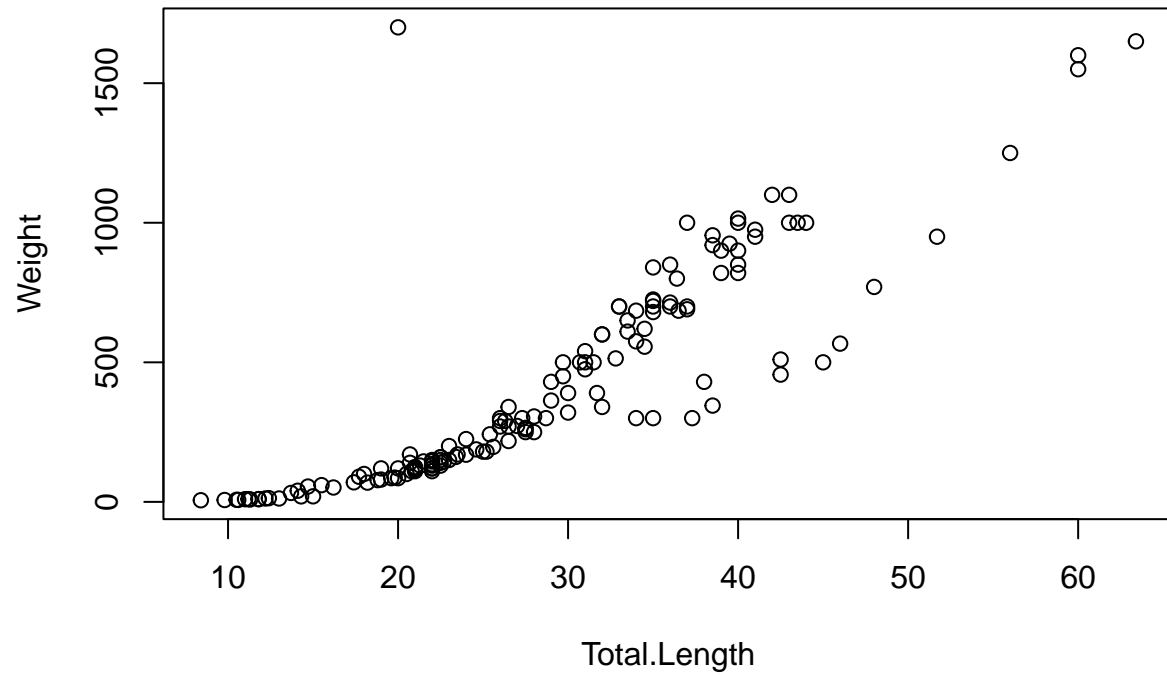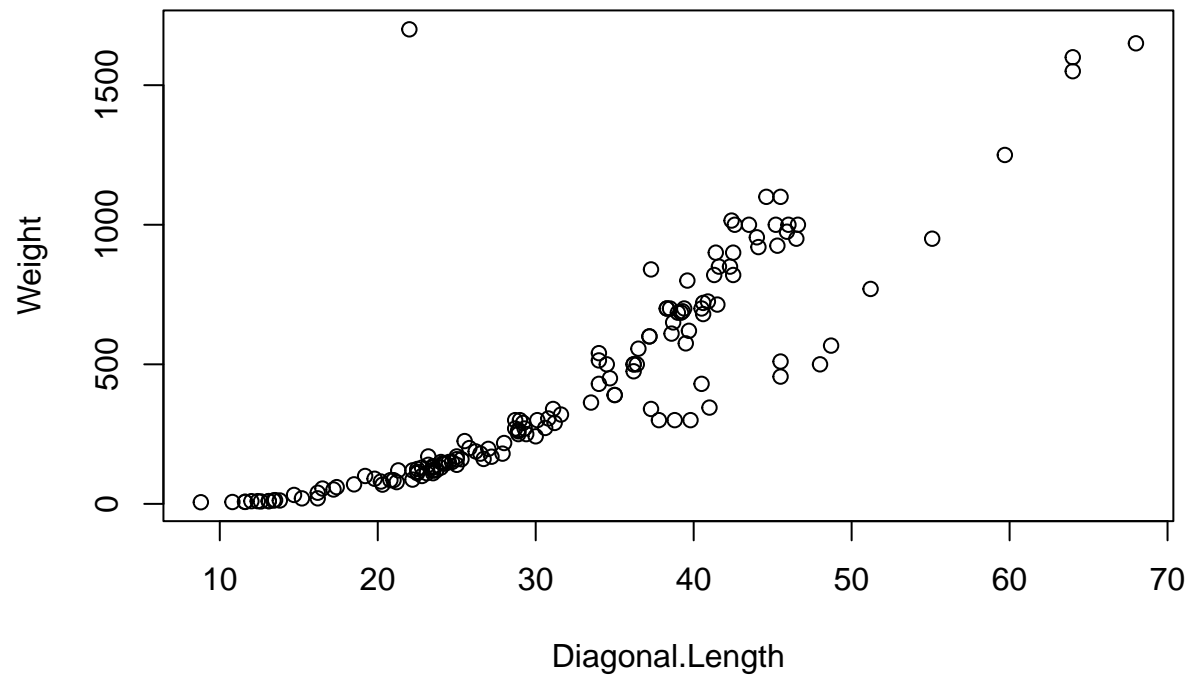
Width vs Weight


```
plot(Total.Length, Weight, main ="Total.Length vs Weight")
```
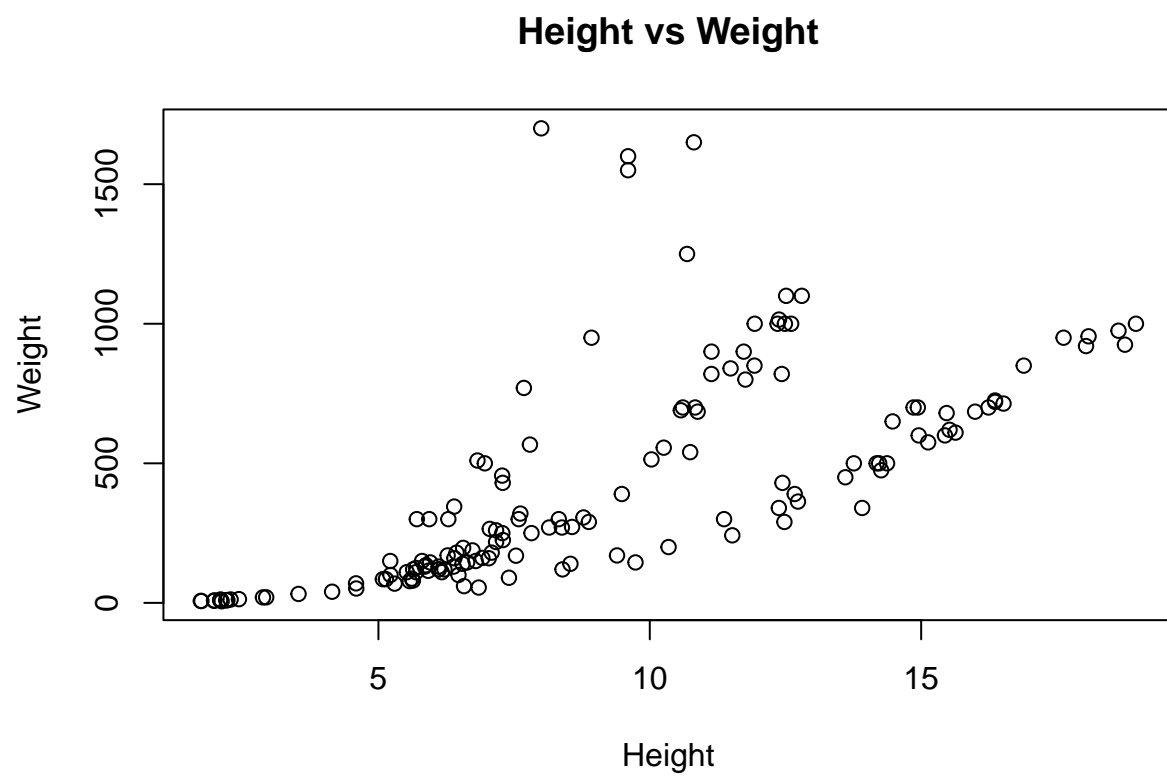
Total.Length vs Weight


```
plot(Diagonal.Length, Weight, main ="Diagonal.Length vs Weight")
```
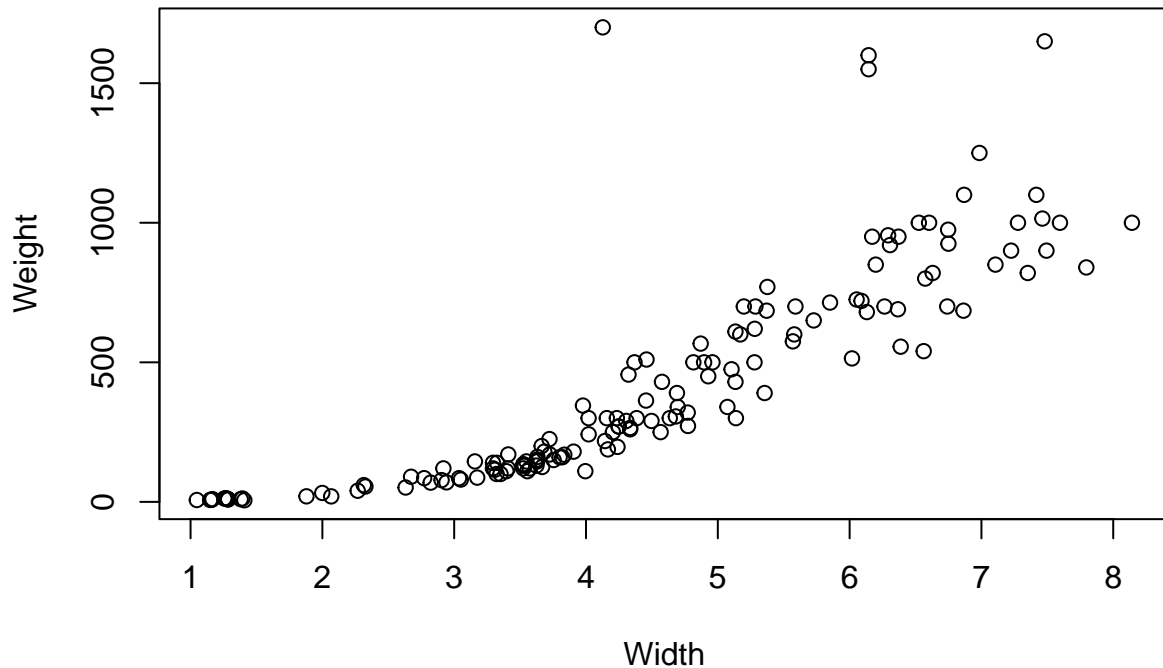
Diagonal.Length vs Weight


```r
plot(Height, Weight, main ="Height vs Weight")
```

Height vs Weight


```
plot(Width, Weight, main ="Width vs Weight")
```

Width vs Weight

Answer to 1.b: There is a positive linear correlation with each predictor and weight. Each plot has at least 1 outlier at around 1500 grams.

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

```
cor(fish[,3:7])
```

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

Answer to 1.c: The predicting variables show high levels of correlation with the highest being 0.999 in Total.Length/Body.Height and lowest being 0.626 in Height/Body.Height. There is potential for multicollinearity, we can further analyze by finding the VIF value.

(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 to 1.d: From the exploratory analysis, we can see that there is a relationship between Weight and the predictor variables. So it is reasonable to use a multiple linear regression model, but we need to take into account the possibility of multicollinearity.

Question 2: Fitting the Multiple Linear Regression Model [8 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? Explain.

Answer to 2.b:

At $\alpha = 0.01$ the overall regression is significant and we can reject the null hypothesis as p-value: $< 2.2e\text{-}16$.

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

Answer to 2.c:

The coefficient estimate for Body.Height is -176.87. For every unit of increase in Body.Height there is a -176.87 gram decrease in weight as long as all other factors stay constant.

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

Answer to 2.d:

The coefficient estimate for Species Parkki is 79.34. If the species is Parkki then it is 79.34 grams heavier than Bream Species as long as all other factors stay constant.

Question 3: Checking for Outliers and Multicollinearity [6 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.d = cooks.distance(model1)
size = nrow(fish)
plot(cook.d, type= 'h', lwd=1, col='blue')
abline(h=1)
text(x=1:length(cook.d)+1, y=cook.d, labels=ifelse(cook.d>4/size, names(cook.d),""), col="red")
```
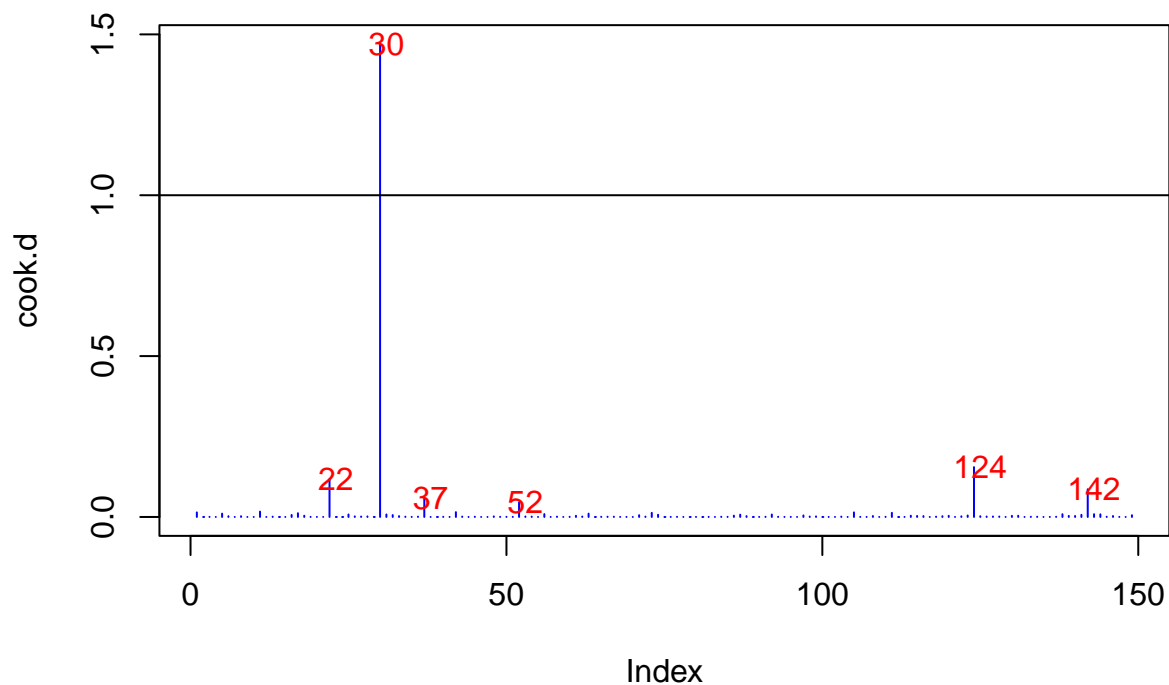


```
outliers <- which(cook.d >= 1)
```

Answer to 3.a:
Using a threshold of 1, row 30 contains an outlier with a cook's distance of approximately 1.5.

(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.

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

```
##
## Call:
```

9

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

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

```
library(car)
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
```
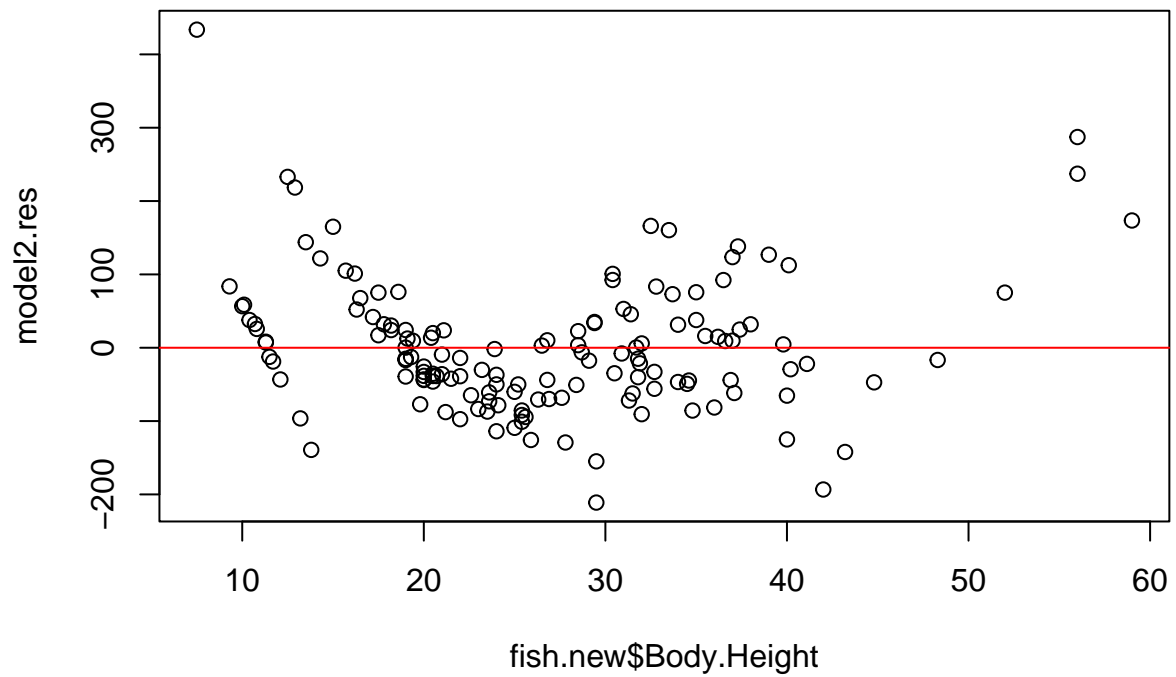
Answer to 3.c:
Based on the results from vif(model) we can see that all the predictors have a VIF > max(10, 16.26). This indicates the presence of multicollinearity among the predictor variables and the model needs to be adjusted.

Question 4: Checking Model Assumptions [6 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?
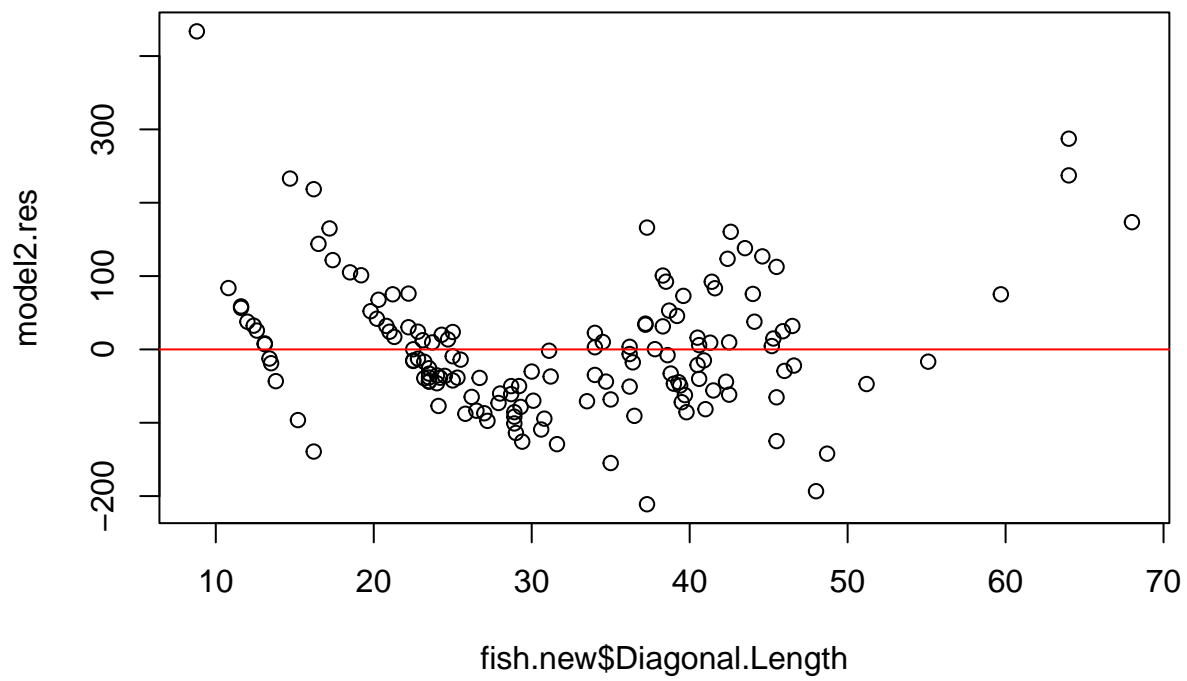
```
model2.res <- resid(model2)
plot(fish.new$Body.Height, model2.res)
abline(0,0, col="red")
```
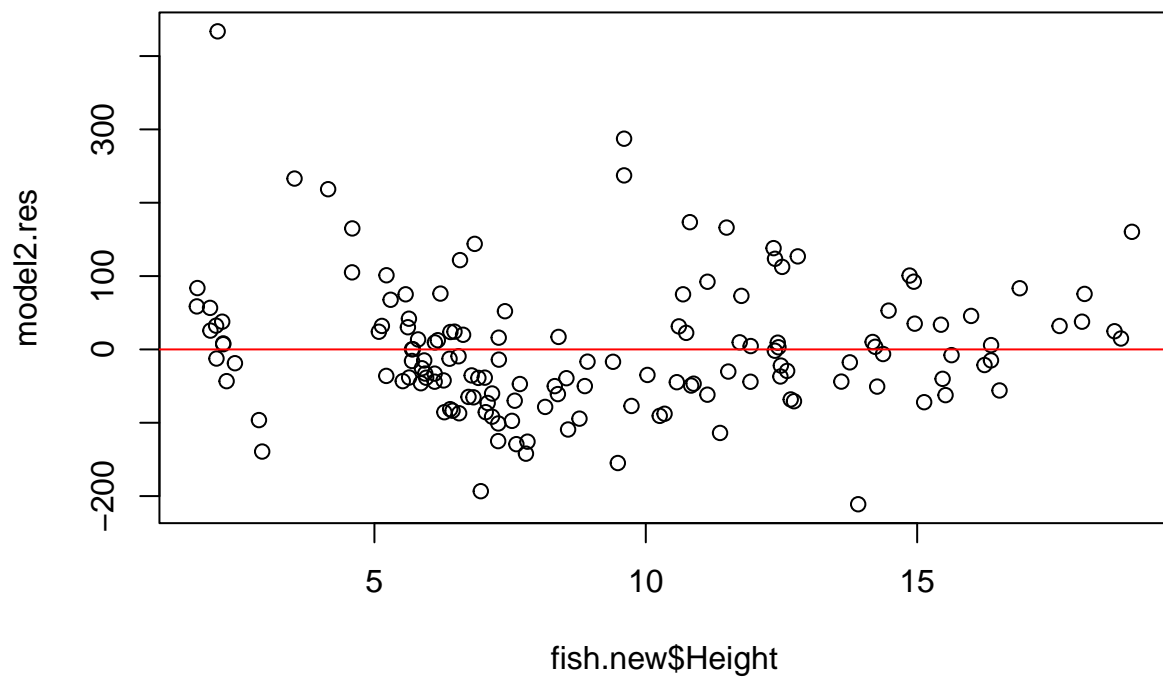


```
plot(fish.new$Total.Length, model2.res)
abline(0,0, col="red")
```
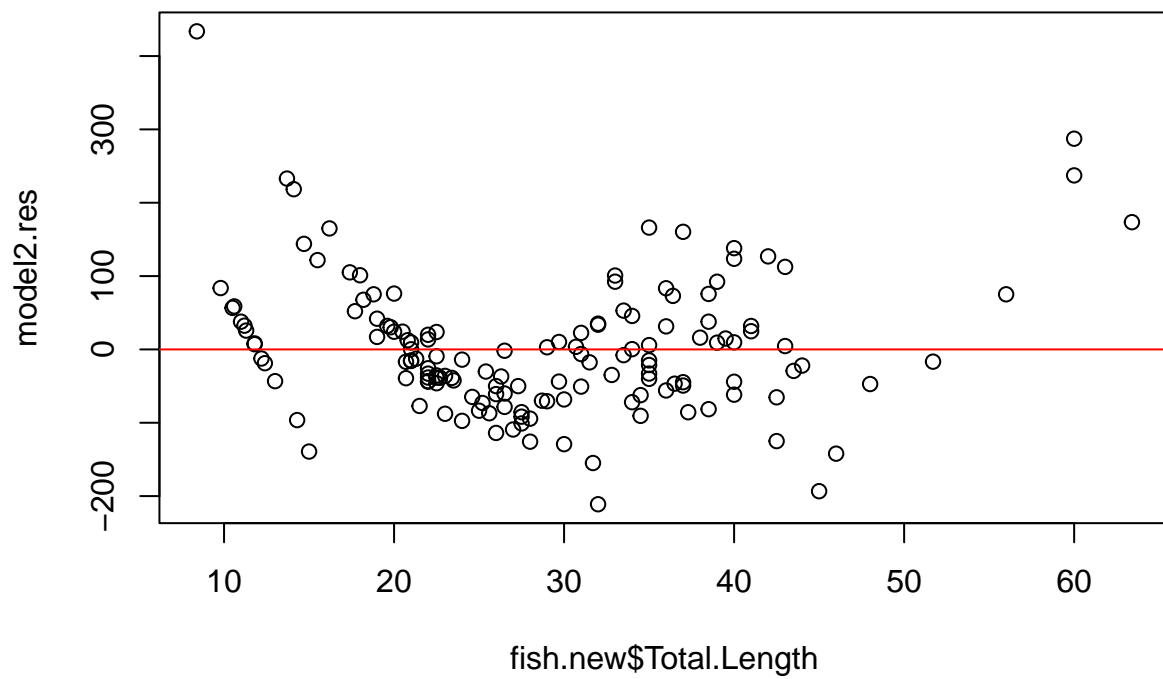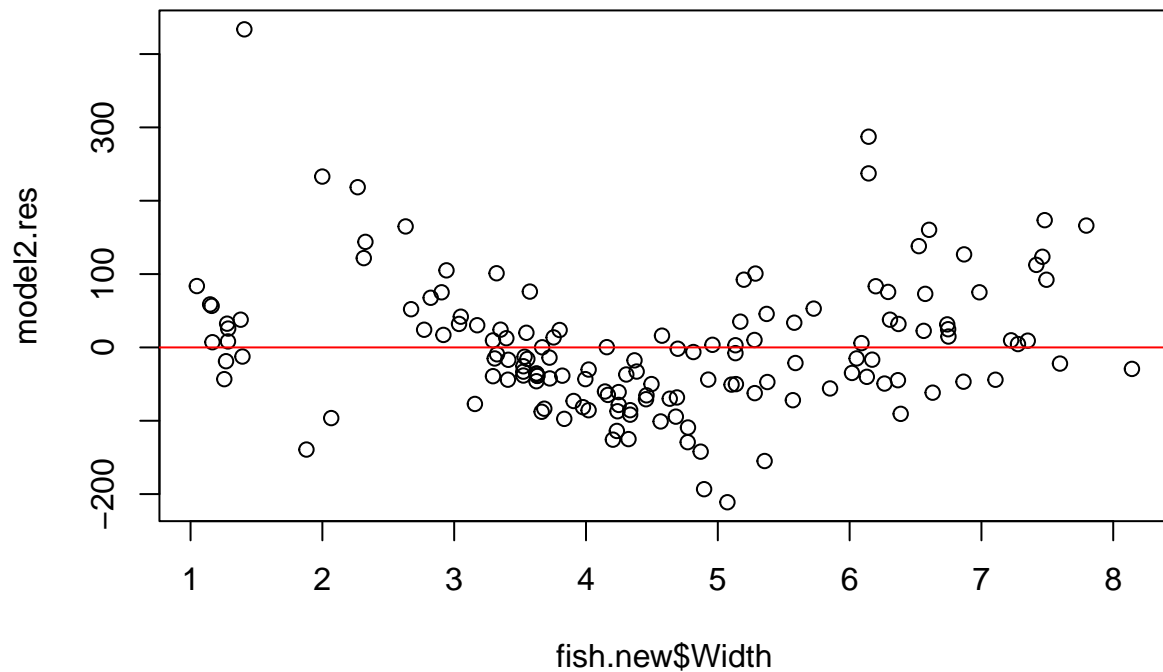
```
plot(fish.new$Diagonal.Length, model2.res)
abline(0,0, col="red")
```

```
plot(fish.new$Height, model2.res)
abline(0,0, col="red")
```

```
plot(fish.new$Width, model2.res)
abline(0,0, col="red")
```
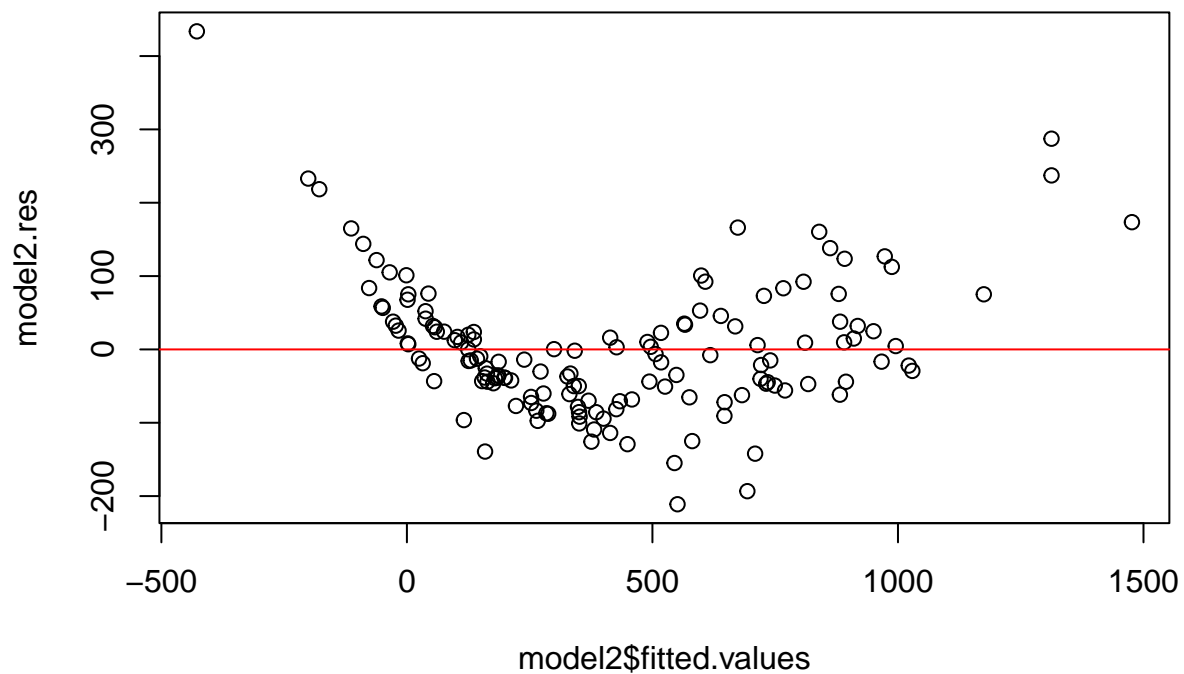
Answer to 4.a:

The linearity assumption does not hold as there is little to no linear relationship between the residuals and quantative 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?

```
plot(model2$fitted.values, model2.res)
abline(0,0, col="red")
```
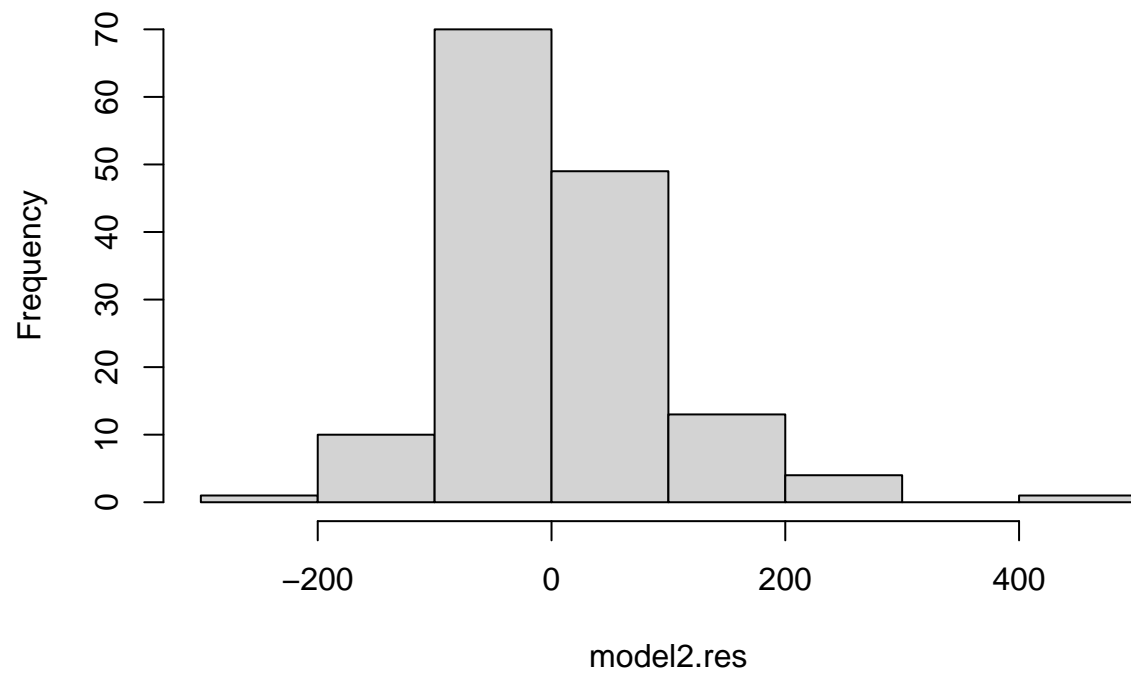
Answer to 4.b:

The constant variance assumption does not hold as the plots are not randomly scattered and there is clustering around (0,0).

(c) Create a histogram and normal QQ plot for the standardized residuals. What conclusions can you draw from these plots?
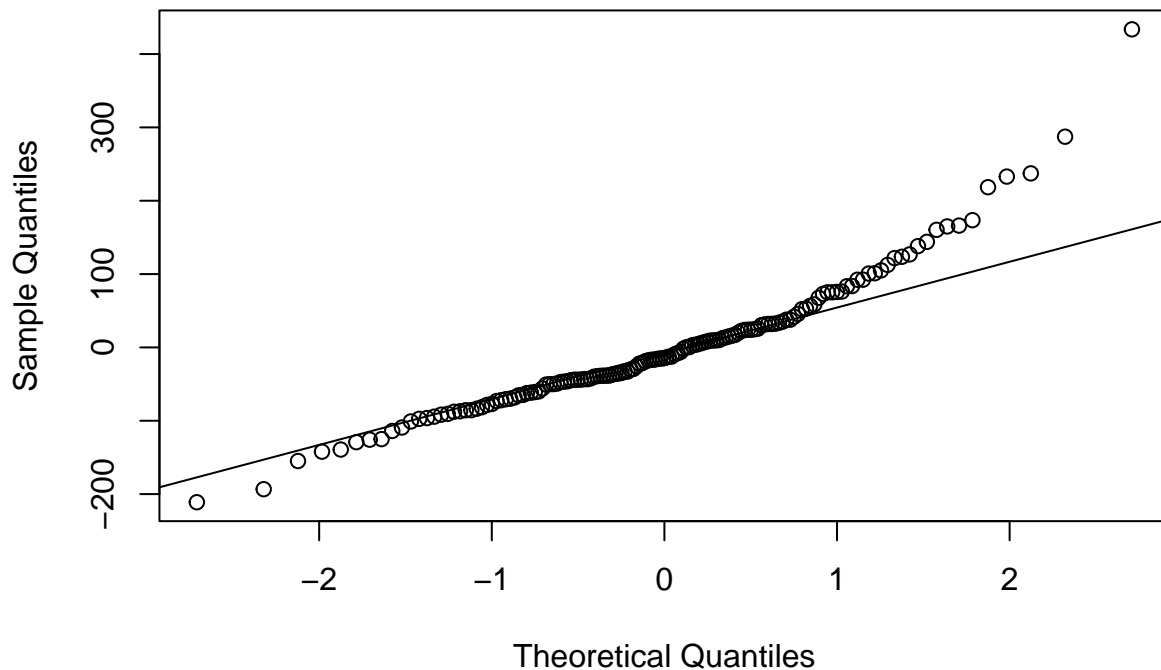
```
hist(model2.res)
```

Histogram of model2.res


```
qqnorm(model2.res)
qqline(model2.res)
```

Normal Q–Q Plot

Answer to 4.c:
We can see that the QQ plot has a right skew and the histogram indicates a slight right skew with most of the data centered around 0. This indicates that the normality assumption 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= fish.new)
summary(model3)
```

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

18

```
## SpeciesPerch       -23.941     21.745  -1.101    0.273
## SpeciesPike        -400.964    33.350 -12.023   < 2e-16 ***
## SpeciesRoach        -19.876    30.111  -0.660    0.510
## SpeciesSmelt        256.408    39.858   6.433 1.85e-09 ***
## SpeciesWhitefish   -14.971     42.063  -0.356    0.722
## Total.Length        40.775      1.181  34.527   < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 94.86 on 140 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9321
## F-statistic: 289.1 on 7 and 140 DF,  p-value: < 2.2e-16
```

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

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

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

Answer to 5.a:
After conducting a partial F-test by comparing model3 and model2 we have a p-value of 0.14 and F-stat of 1.7626. Because p-value > alpha value, we fail to reject the null hypothesis.Therefore, the additional variables from model2 does not improve model3.

Question 6: Reduced Model Residual Analysis and Multicollinearity Test [7 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
```
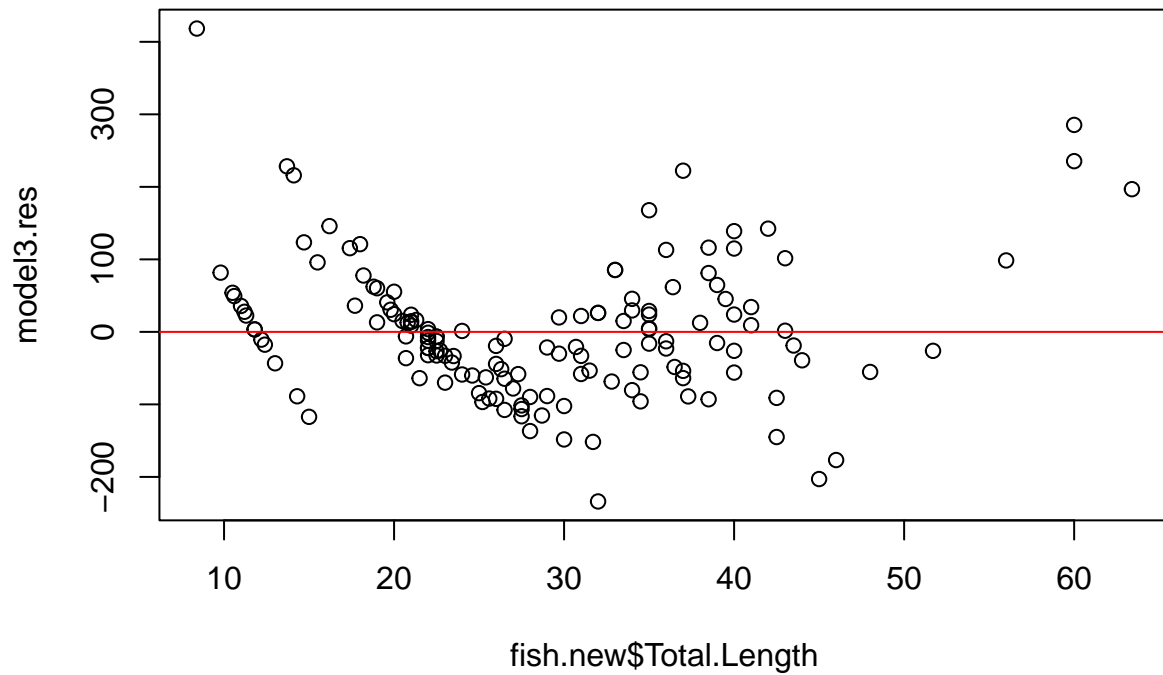
Answer to 6.a:
The VIF for the variables in model3 are less than the max(10, 15.456), but also less than 3 meaning low correlation, which means that there is no indication of multicollinearity for model3.
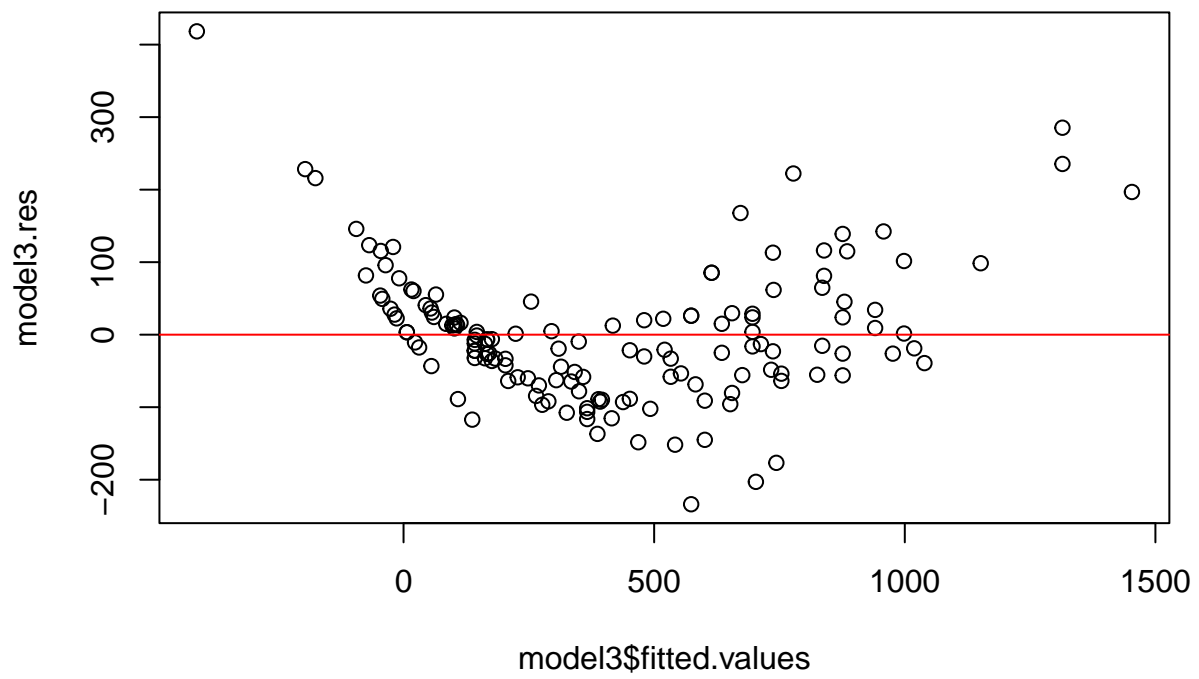
(b) Conduct residual analysis for model3 (similar to Q4). Comment on each assumption and whether they hold.

```
model3.res <- resid(model3)

plot(fish.new$Total.Length, model3.res)
abline(0,0, col="red")
```
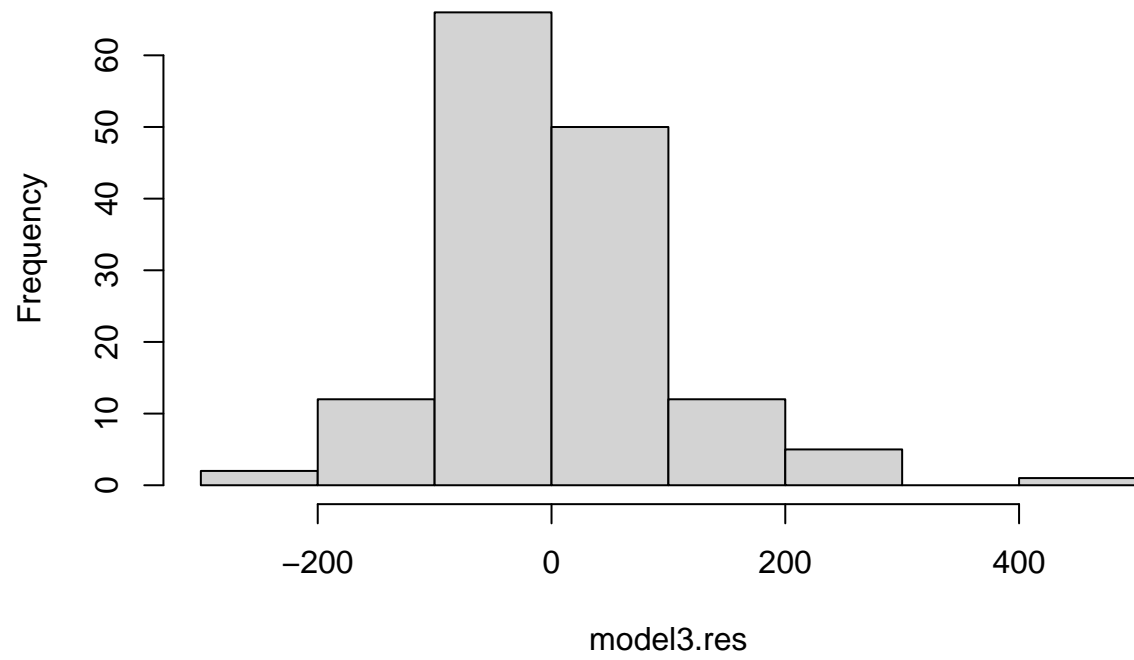


```
plot(model3$fitted.values, model3.res)
abline(0,0, col="red")
```
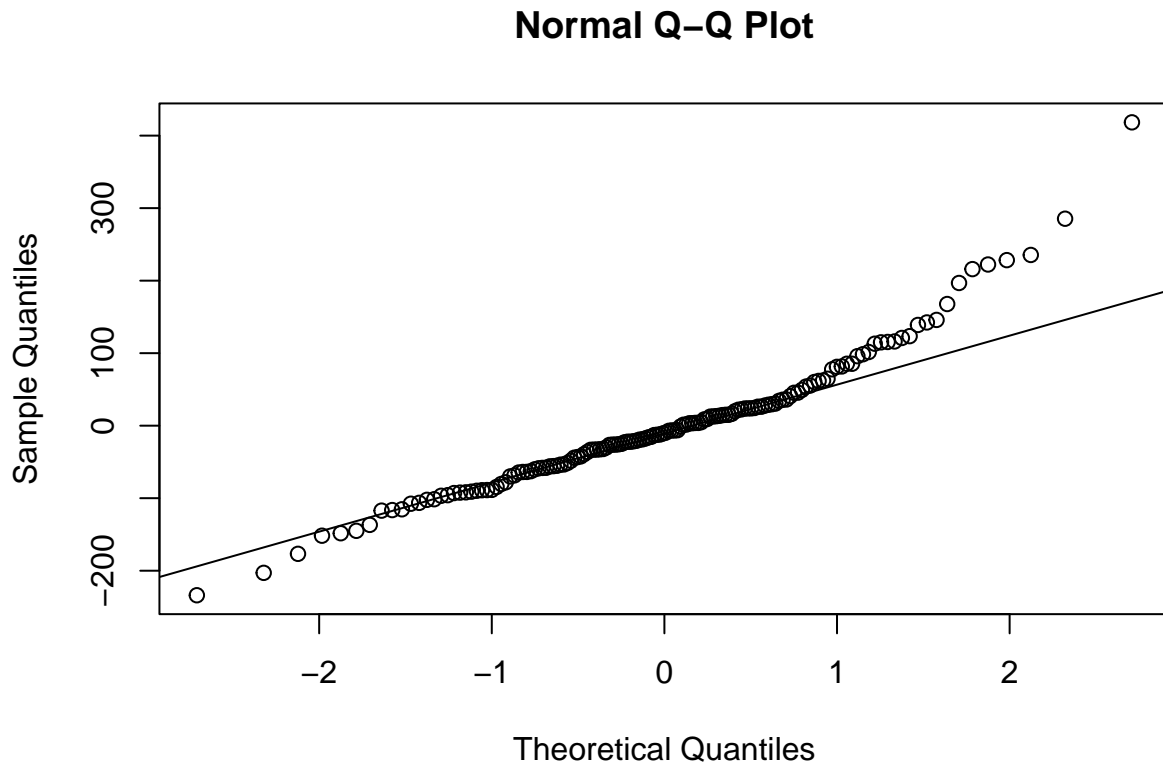
```
hist(model3.res)
```

Histogram of model3.res


```
qqnorm(model3.res)
qqline(model3.res)
```

Normal Q–Q Plot

Answer to 6.b:
After conducting residual analysis for model3 we can still see the same issues as in model2. The scatter plot for Residual v Total.Length is still non linear as the points are not randomly scattered. The residual v fitted values shows non random and clustering. Lastly, from the qqplot and histogram we can also see the graph showing heavy tails and skewing. These all indicated that the linearity, independence, normalilty and constant variance assumptions do not hold.

Question 7: Transformation [9 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

```
model3.bc <- boxCox(model3)
```
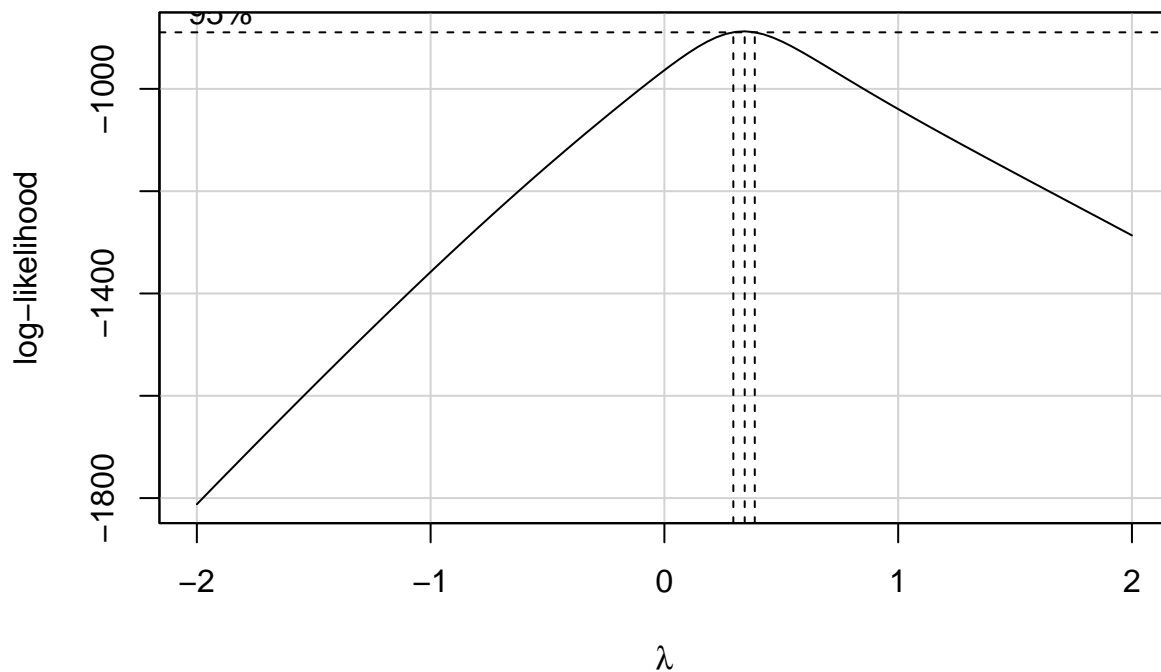
```
opt.lambda = model3.bc$x[which.max(model3.bc$y)]
cat("Optimal lambda:", round(opt.lambda/0.5)*0.5, end="\n")
```

```
## Optimal lambda: 0.5
```

Answer to 7.a:

lambda value is rounded to 0.5, so we apply a square root transformation.

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

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

```
##
## Call:
## lm(formula = sqrt(Weight) ~ Species + Total.Length, data = fish.new)
##
## 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
```
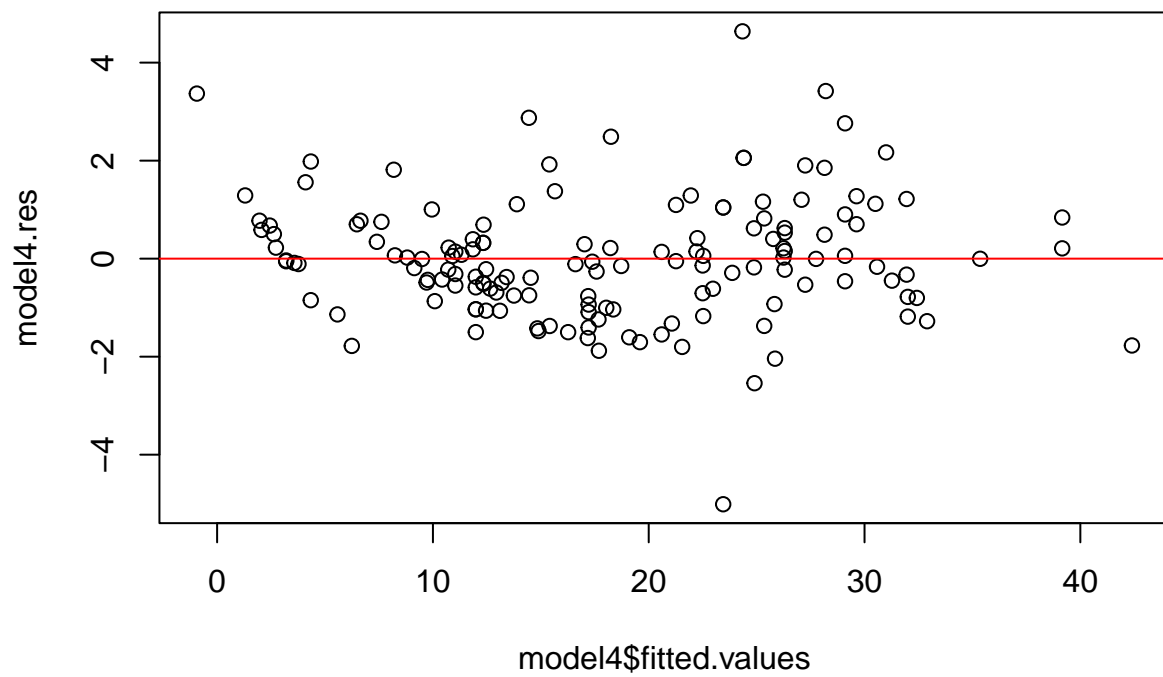
(c) Perform Residual Analysis on model4. Comment on each assumption. Was the transformation successful/unsuccessful?

```
model4.res <- resid(model4)

plot(fish.new$Total.Length, model4.res)
abline(0,0, col="red")
```
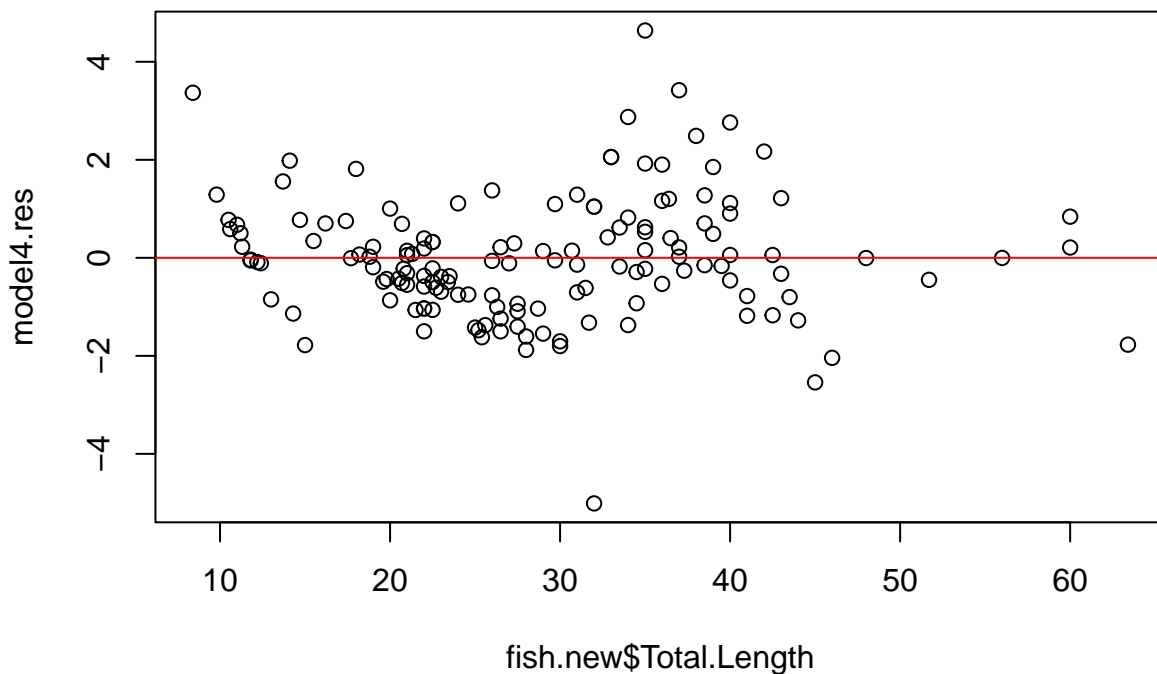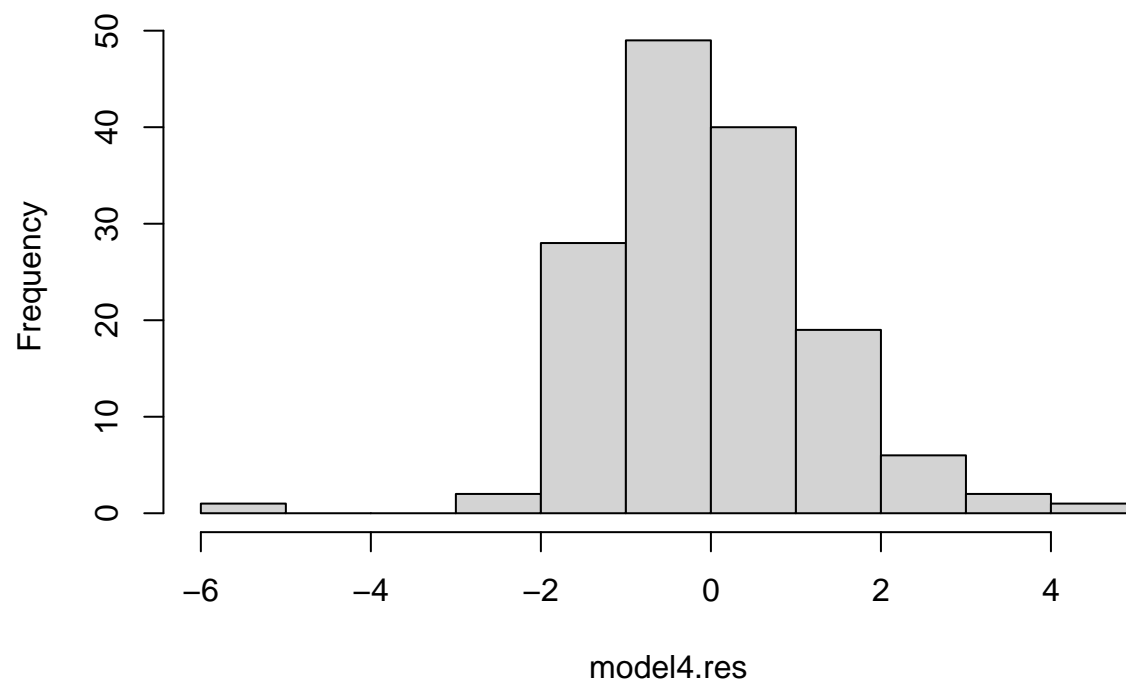


```
plot(model4$fitted.values, model4.res)
abline(0,0, col="red")
```
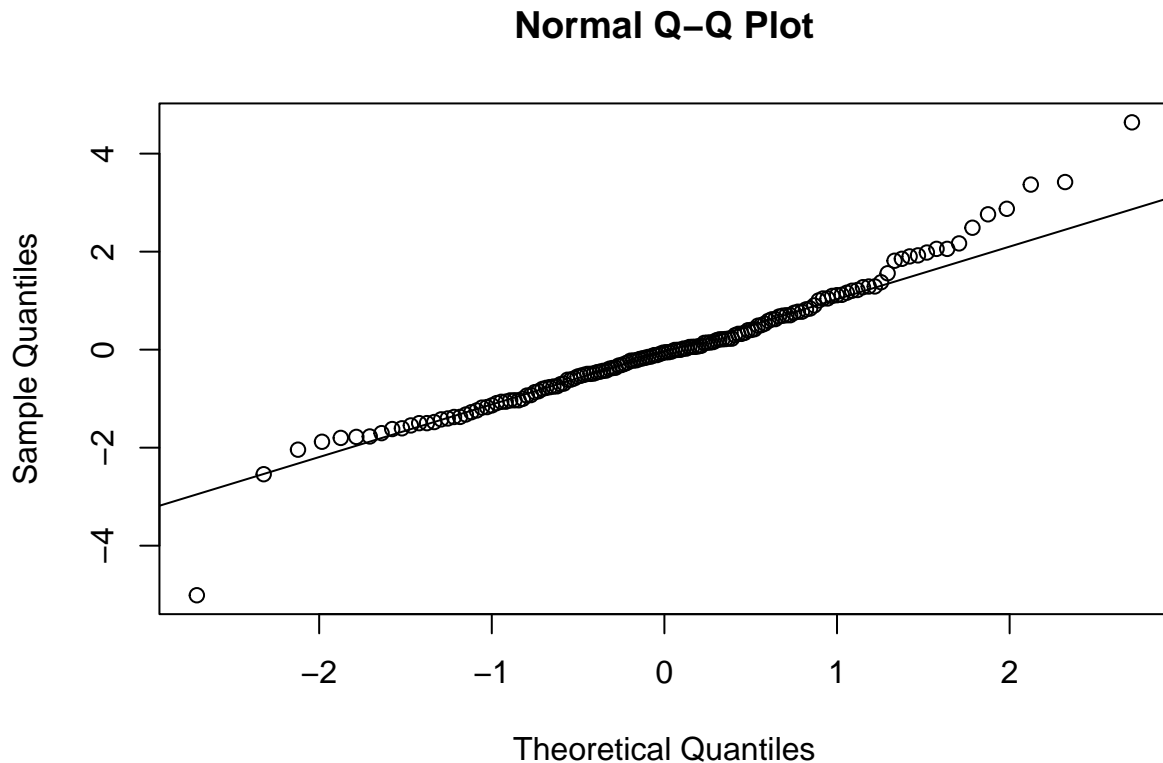
```
hist(model4.res)
```

Histogram of model4.res


```
qqnorm(model4.res)
qqline(model4.res)
```

Normal Q–Q Plot

Answer to 7.c:
For the linearity assumption we see an improvement in linearity indicated by the more even distribution across the line. For the constant variance assumption we see an improvement as the points are more randomly spread out across the plot and less clustering. For the qqplot and histogram we also see an improvement as the distribution is more uniform and a reduction in the heavy tail. The transformation was successful as there was improvement in all of the assumptions.

Question 8: Model Comparison [2 pts]

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

```
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish.new)
##
## 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
```

summary(model3)

```
##
## Call:
## lm(formula = Weight ~ Species + Total.Length, data = fish.new)
##
## 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 = fish.new)
##
## 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 to 8.a:
Model2: $R^2 = 0.9385$, Adjusted $R^2 = 0.9335$ Model3: $R^2 = 0.9353$, Adjusted $R^2 = 0.9321$ Model4: $R^2 = 0.9817$, Adjusted $R^2 = 0.9808$

R^2 explains the variability of the model explained by the independent variables. The adjusted r^2 explains the percentage of variation by the independent variables.

In this case model4 has the highest r^2 and adjusted r^2 indicating that it is the most preferred model to use. The higher r^2 value can be explained by the transformation which improved its overall fit.

Model2 and model3 are close; however, model2 has a slightly higher r^2 value due to having more variables.

Question 9: Prediction [8 points]

(a) Predict 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.pred <- predict(model3, fishtest)
model3.mse <- mean((model3.pred-fishtest$Weight)^2)

model4.pred <- predict(model4, fishtest)
model4.mse <- mean((model4.pred-fishtest$Weight)^2)
```

Answer to 9.a:
Model4 MSE is 98076.62 and Model3 MSE is 9392.25. Because the MSPE of Model 4 is higher it indicates that Model 3 is preferred as lower values indicate better prediction ability.

(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.

```
perch <- data.frame(Species='Perch', Body.Height=28, Total.Length = 32)
predict(model4, perch, interval = 'prediction', level=0.90)^2
```

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

Answer to 9.b:

Using model4 need to square the prediction as we had done a square root transformation. The result of 90% confidence that the average weight of a perch fish with those measurements is 461.9429g with a lower bound of 374.4536g and upper bound of 558.6091g.