

# 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 weight 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

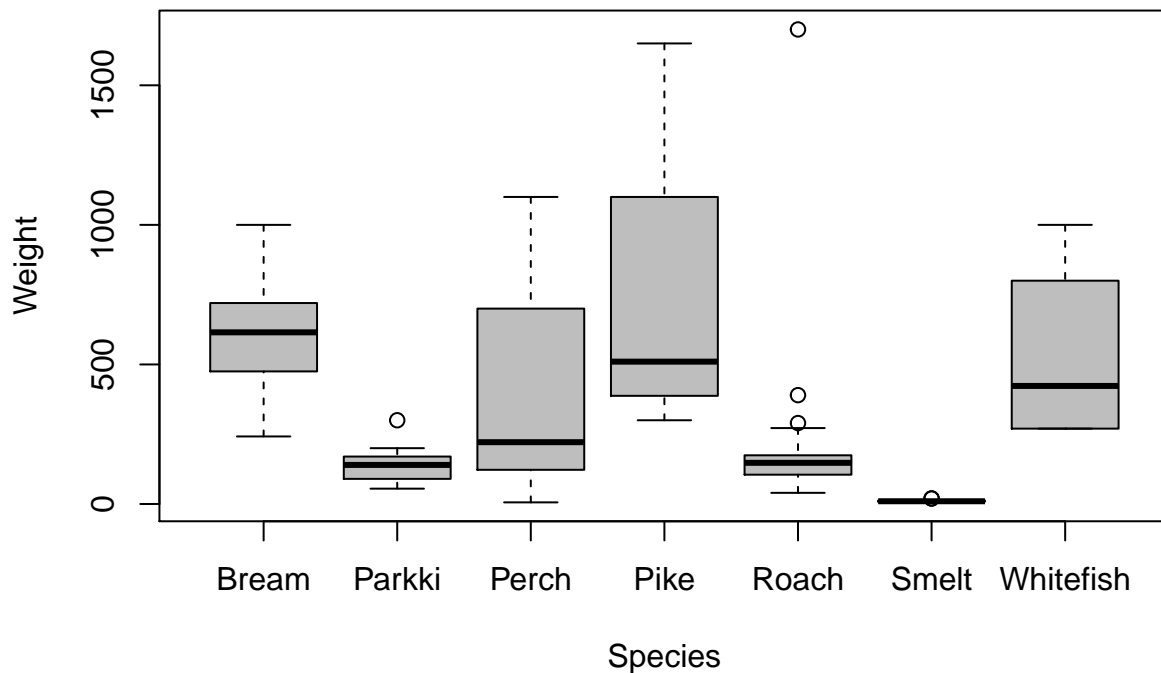
# Read the data set
fishfull = read.csv("Fish.csv",header=T, fileEncoding = 'UTF-8-BOM')
row.cnt = nrow(fishfull)
# Split the data into training and testing sets
fishtest = fishfull[(row.cnt-9):row.cnt,]
fish = fishfull[1:(row.cnt-10),]
```

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?

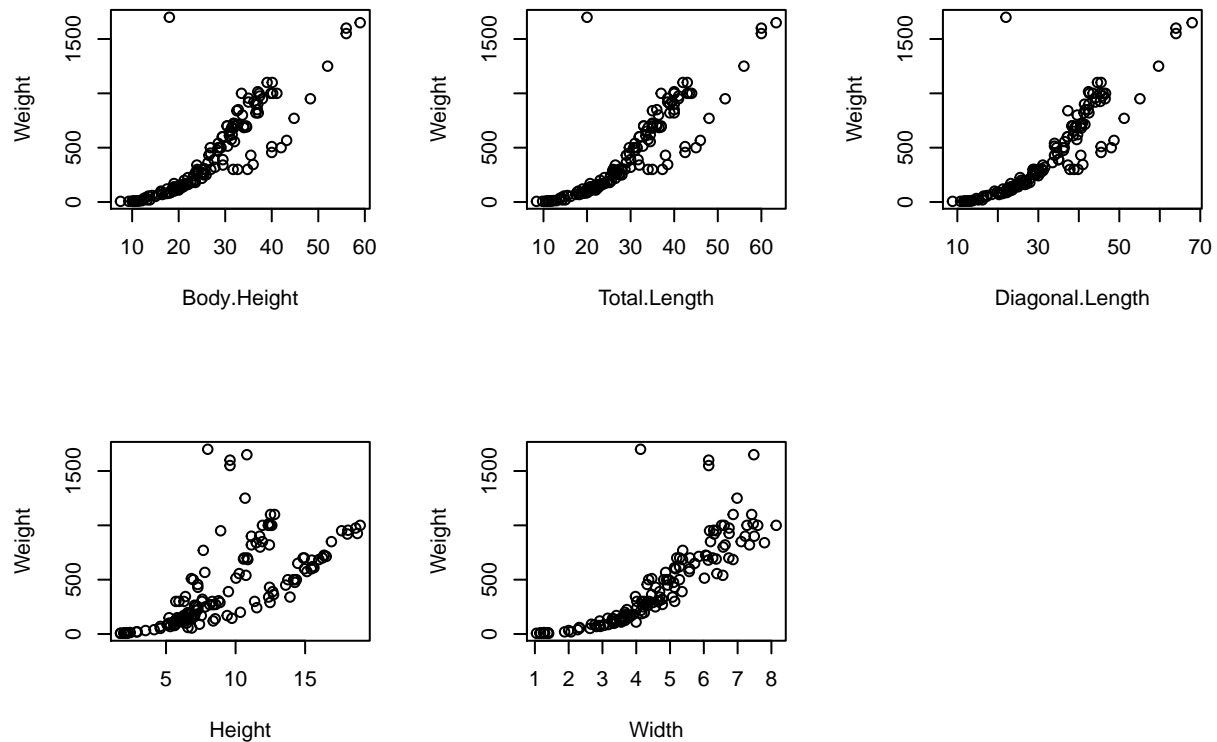
```
# Create boxplot
boxplot(Weight~Species, data=fish, col='grey')
```



From the plot, we see the medians vary across all species of fish. Specifically, Parkki, Roach, and Smelt have lower weights, while Bream, Whitefish, and Pike have higher weights. Hence, there appears to be a relationship between the predictor and the response.

(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?

```
# Create scatterplots
par(mfrow=c(2,3))
for (i in c(3:7)){
  col_name = names(fish[i])
  plot(fish[,i], fish$Weight, xlab= col_name, ylab = "Weight")
}
```



**General trend:** Graphically, it appears that all five predictors have a positive relationship with the Weight of the fish, but it seems that this relationship is non-linear since each plot suggests a curved relationship between the response and the quantitative predictor. Also, some of the plots show heteroscedasticity with variability of the weights increasing as the values of the predictors increase.

**Outliers:** We also may have a possible outlier with a large weight of over 1500 grams in the upper left portions of the graphs that will need further investigation.

(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.

```
# Correlation matrix
cor(fish[-2])
```

```
##           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
```

We see very strong positive correlation between the predicting variables and response variable.

In addition, we see strong correlation among the predicting variables suggesting that multicollinearity might be a problem in a model that uses all predictors. The least strong linear relationship is between Body.Height and Height.

**(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, numerous of these factors will be able to help determine the weight of the fish. We can attempt a multiple linear regression model first, but we are likely going to want to attempt a Box-Cox transformation to reduce the heteroskedasticity and/or incorporate non-linear associations in the linear model by including transformed versions of the predictors in the model.

## Question 2: Fitting the Multiple Linear Regression Model [9 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.**

```
# fit full model
model1<-lm(Weight~., data=fish)
# Display summary
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  $\alpha$  level of 0.01?

Because the p value associated to the F-statistic ( $<2.2e-16$ ) is less than the  $\alpha$  level, we reject the null hypothesis that all slope coefficients are equal to zero, and state that the model is statistically significant at an  $\alpha$  level of 0.01.

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

$\hat{\beta}_{Body.Height} = -176.87$ , which means for every centimeter increase in Body.Height, the expected weight of the fish DECREASES by 176.87 grams holding all other variables constant. This interpretation seems surprising suggesting that multicollinearity might be a problem in this model.

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

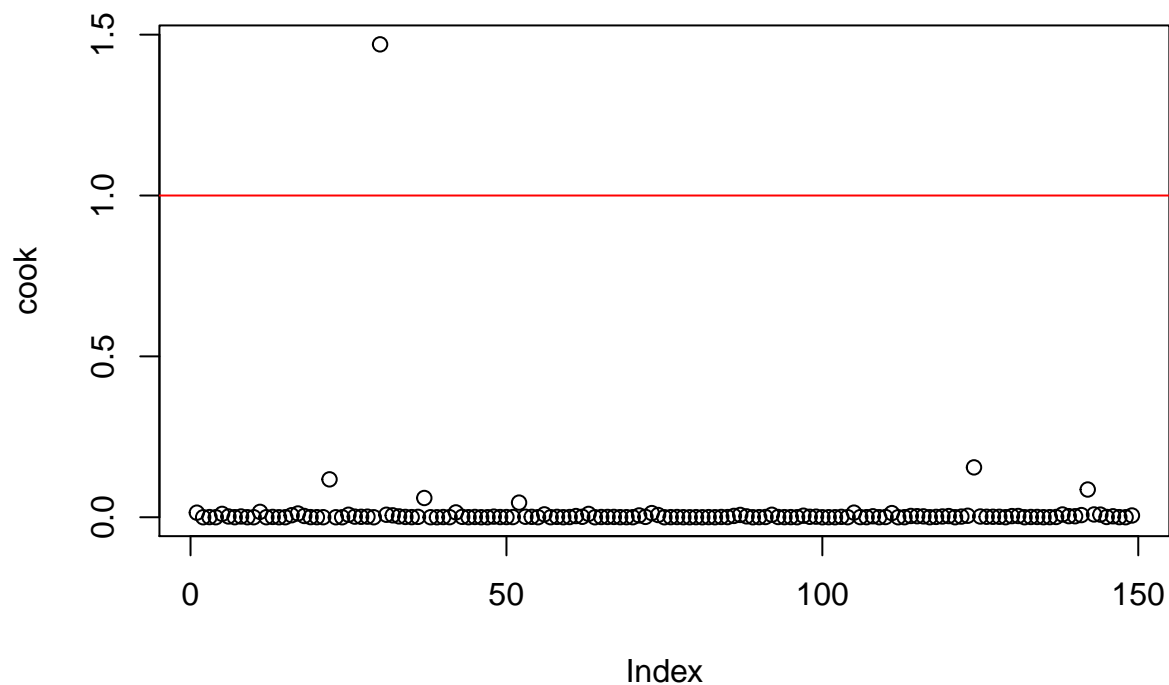
$\hat{\beta}_{SpeciesParkki} = 79.34$ , which means that if the fish is categorized as Parkki, then the expected weight of the fish would be 79.34 grams greater than the baseline species Bream holding all other variables constant. Given that the associated p-value equals 0.55, then it's paussible to state that this difference in weight is not statistically significant.

### 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.

```
# Calculating Cook's distances
cook=cooks.distance(model1)

# Plotting Cook's distances
plot(cook)
abline(h=1, col="red")
```



```
#Identify outliers
cat("Observation", which(cook>1), "has a cook's distance that is greater than 1")
```

```
## Observation 30 has a cook's distance that is greater than 1
```

(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.

```
# Remove outlier
fish2<-fish[-30,]

# Fit model2
model2<-lm(Weight~., data=fish2)

# Display summary
summary(model2)
```

```
##
## Call:
## lm(formula = Weight ~ ., data = fish2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -211.10  -50.18  -14.44   34.04  433.68
##
```

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

(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 Threshold
cat("VIF Threshold:", max(10, 1/(1-summary(model2)$r.squared))), "\n")
```

```
## VIF Threshold: 16.25583
```

```
# Calculate VIF
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
```

Using a VIF threshold of 16.26, we have high multicollinearity in the model for all predictors.

## Question 4: Checking Model Assumptions [9 points]

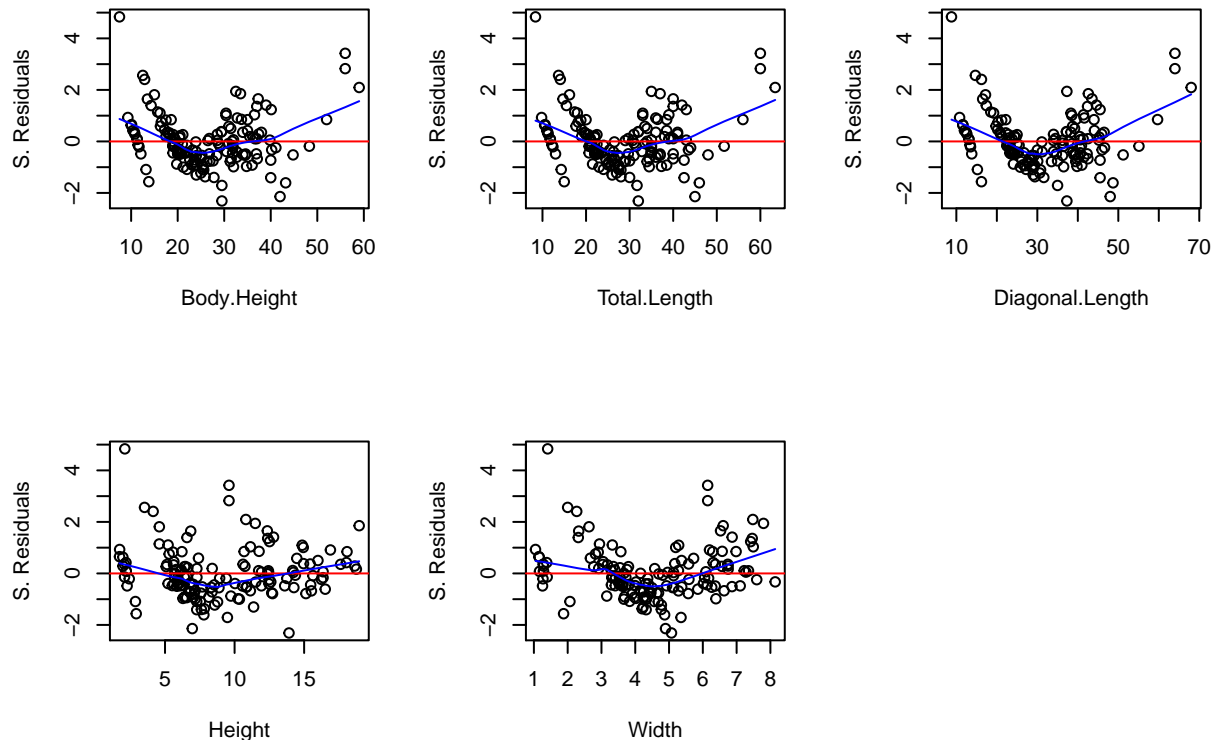
Please use the cleaned data set, which have the outlier(s) removed, and model2 for answering the following questions.

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

```

# Get standardized residuals
resids =rstandard(model2)
par(mfrow=c(2,3))
for (i in c(3:7)){
  col_name = names(fish2[i])
  plot(fish2[,i], resids, xlab= col_name, ylab = "S. Residuals")
  abline(h=0, col="red")
  lines(lowess(fish2[,i], resids), col='blue')
}

```



From the scatterplots for Body.Height, Total.Length, Diagonal.Length, and Width we can see that the residuals exhibit a clear U-shape, which provides a strong indication of non-linearity. Specifically,

Body.Height: A possible concern between 20-30cm, where the majority of the residuals are below the zero line.

Total.Length: A possible concern between 20-30cm, where the majority of the residuals are below the zero line.

Diagonal.Length: A possible concern between 27-35cm, where the majority of the residuals are below the zero line.

Width: A possible concern between 3.9-5cm, where the majority of the residuals are below the zero line.

Overall, the linearity assumption does not seem to hold.

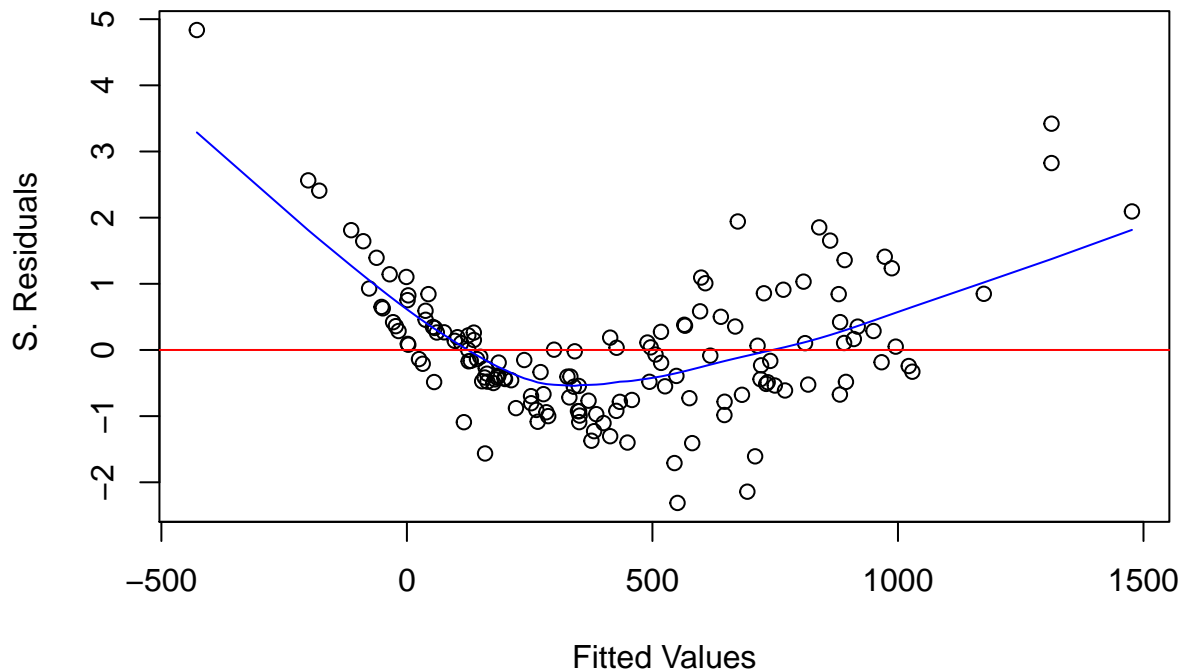
*Note: The blue line is a smooth fit to the residuals, intended to make it easier to identify a trend.*

(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 of std. residuals versus fitted values
plot(model2$fitted.values, resid, xlab="Fitted Values", ylab=" S. Residuals")
lines(lowess(model2$fitted.values, resid), col='blue')
abline(h=0, col="red")
```



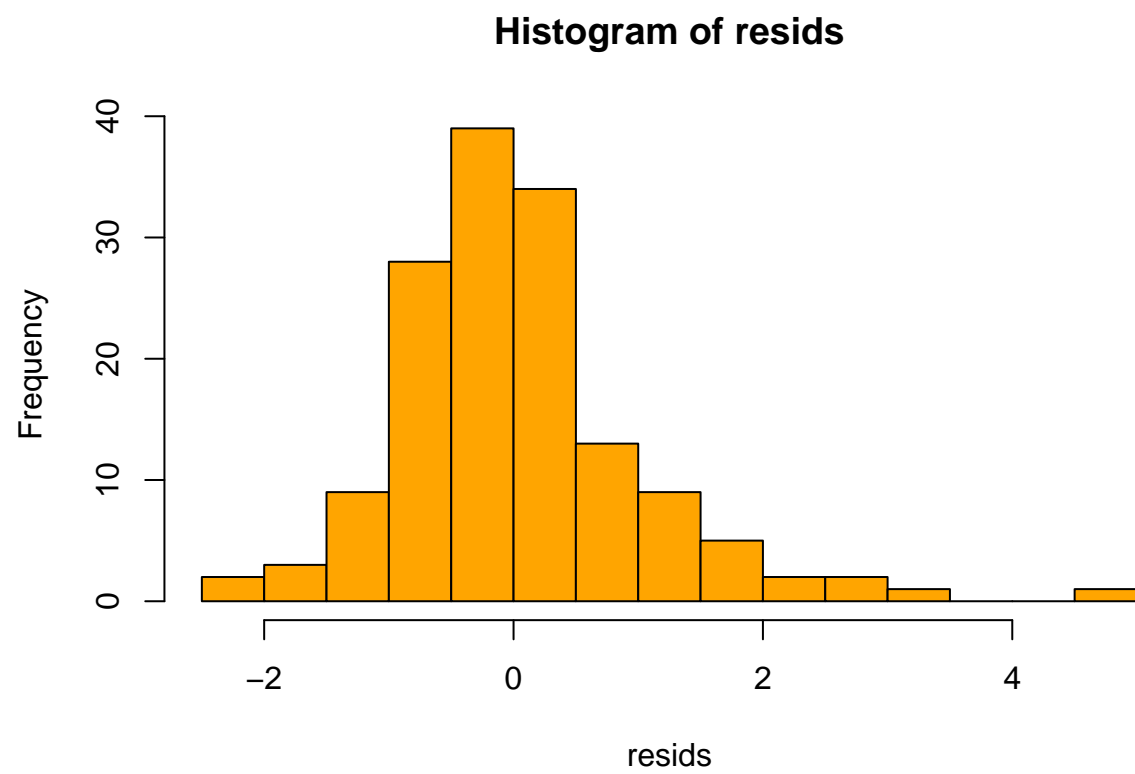
From the plot of the standardized residuals vs. fitted values, we see a U-shaped pattern which confirms that the linearity assumption does not hold.

Also the plot shows that the spread of the residuals is not roughly equal per fitted value, with the variance increasing as the fitted values increase. Constant variance does not appear to hold.

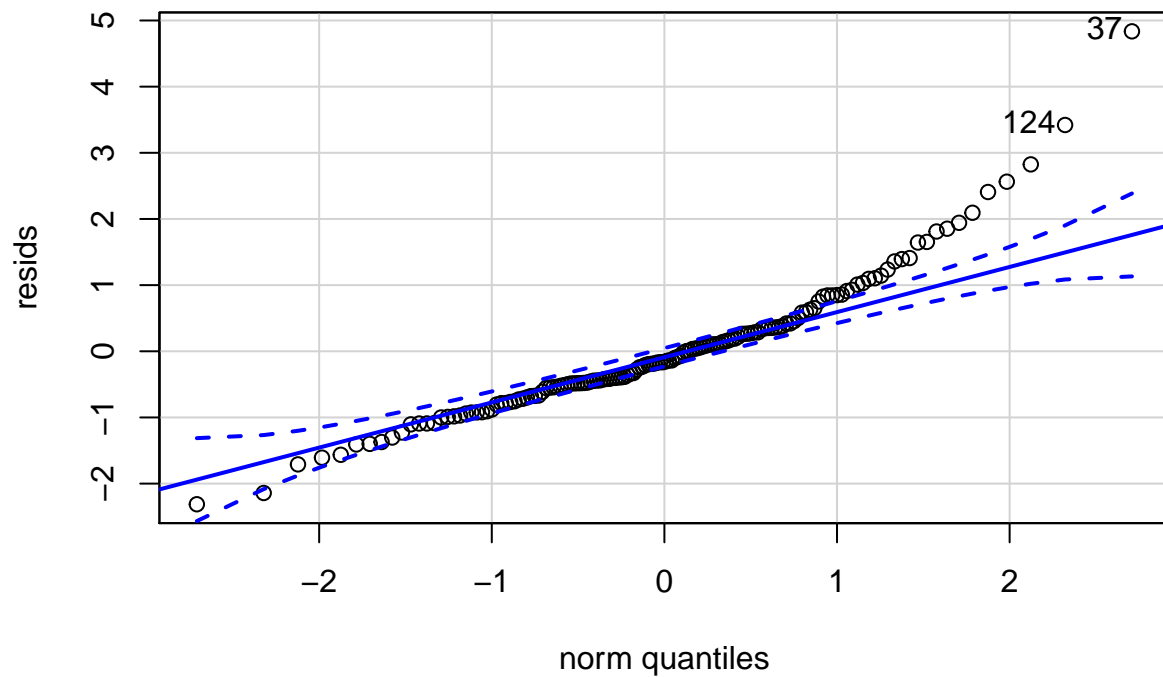
There does not appear to be any clear clustering in the residuals. This suggests that the errors might be uncorrelated.

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

```
# Plots for normality
hist(resid, col="orange", nclass=15)
```



```
qqPlot(resids)
```



```
## 37 124
## 36 123
```

From the histogram, we see a normal peak but with a heavy right tail. The heavy right tail can be confirmed with the QQ Plot. The normality assumption does not appear to hold. Thus a transformation on the response variable may better our model.

## 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`.

```
# Fit model3
model3<-lm(Weight~Species+Total.Length, data=fish2)

# Display summary
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  $\alpha$  level of 0.01?

```
# Conduct Partial F-test
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
```

$H_0 : \alpha_{Body.Height} = \alpha_{Diagonal.Length} = \alpha_{Height} = \alpha_{Width} = 0$

$H_A$ : At least one coefficient  $\neq 0$

Using a  $\alpha$  level of 0.01, since the pvalue  $> \alpha_{0.01}$  we fail to reject the null hypothesis that the  $\alpha_i = 0$   
Thus the additional variables do not provide additional explanatory power to the model.

## 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 Threshold
cat("VIF Threshold:", max(10, 1/(1-summary(model3)$r.squared))), "\n")
```

```
## VIF Threshold: 15.45466
```

```
# Calculate VIF
vif(model3)
```

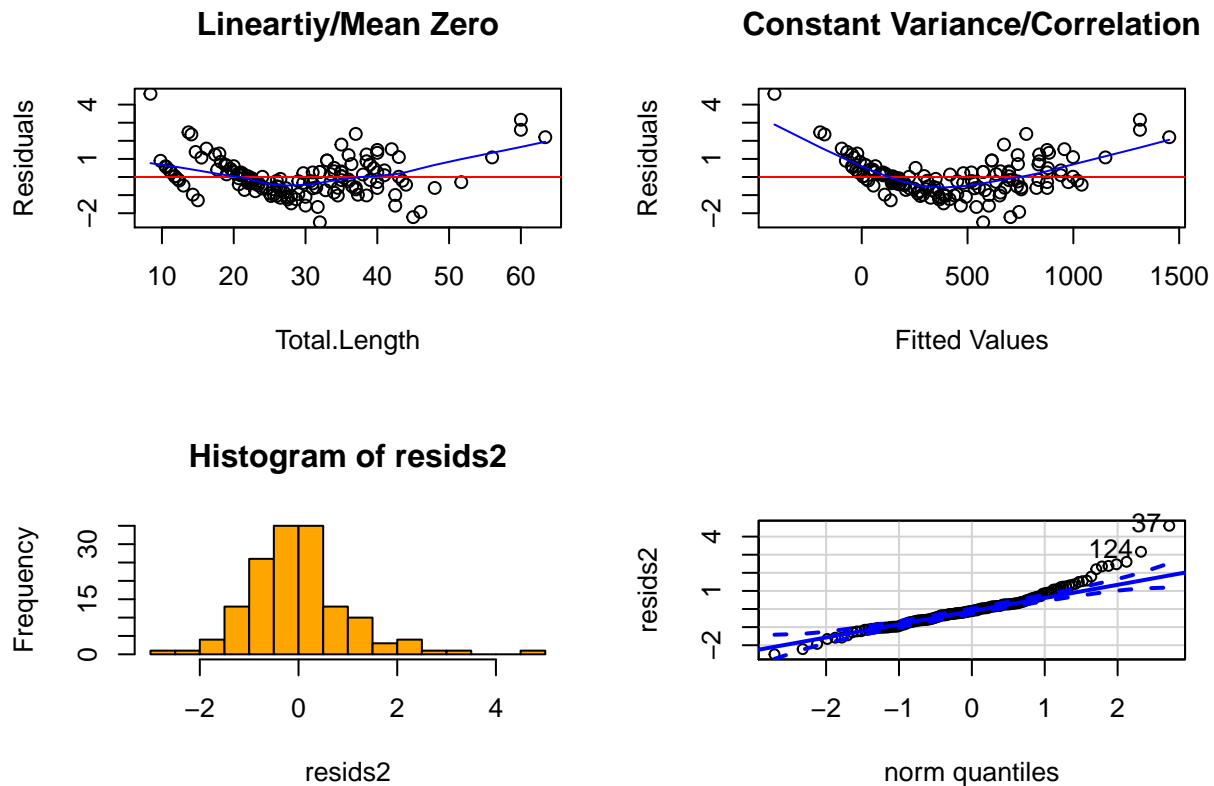
```
##              GVIF Df GVIF^(1/(2*Df))
## Species      2.654472 6      1.084755
## Total.Length 2.654472 1      1.629255
```

Since the predictors all have  $VIF < 15.45$ , the remaining multicollinearity is negligible in the model and should not affect our inferences.

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

```
# Get standardized residuals
resids2 =rstandard(model3)

# Residual plots
par(mfrow=c(2,2))
plot(fish2$Total.Length, resids2 , main="Linearity/Mean Zero",
     xlab="Total.Length", ylab="Residuals")
abline(h=0, col="red")
lines(lowess(fish2$Total.Length, resids2), col='blue')
plot(model3$fitted.values, resids2 , main="Constant Variance/Correlation",
     xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
lines(lowess(model3$fitted.values, resids2), col='blue')
hist(resids2 , col="orange", nclass=15)
qqPlot(resids2)
```



```
## 37 124
## 36 123
```

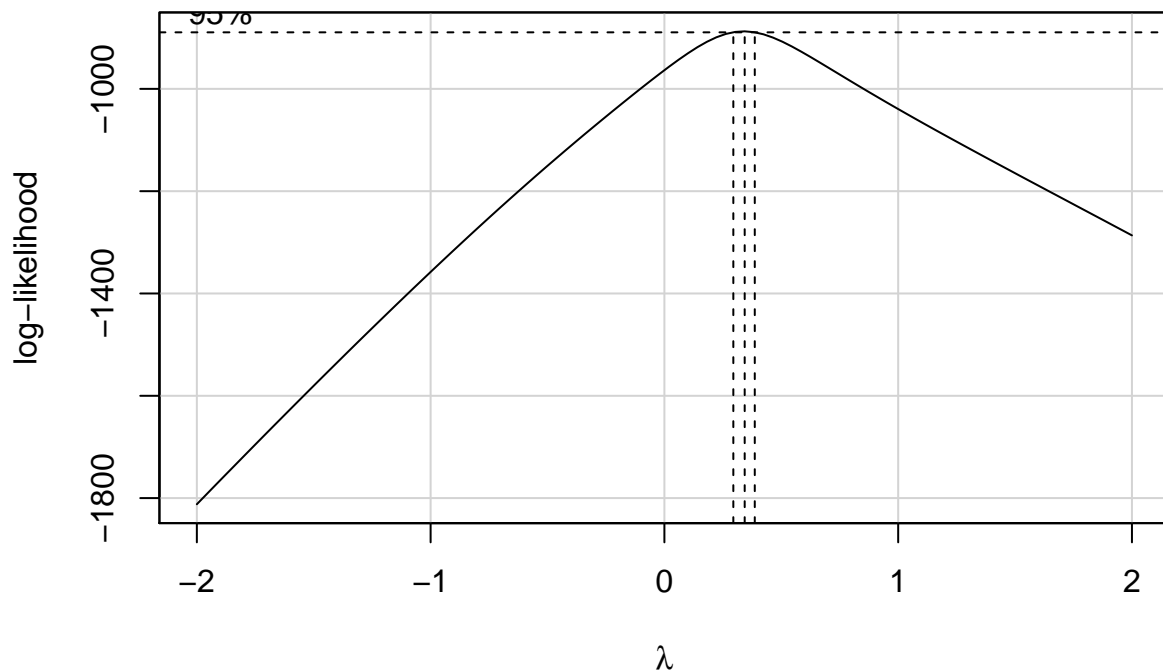
We have similar linearity/mean zero and constant variance violations as model2 with a u-shape in the residuals and increasing variance as fitted values increase.

The histogram and QQ plot both show a normal peak with a heavy right tail. Thus Normality does not hold either. A transformation of the response may help center up the residuals.

## 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

```
# Box-Cox transformation
bc<-boxCox(model3)
```



```
# Extract optimal lambda
opt.lambda<-bc$x[which.max(bc$y)]

# Rounded optimal lambda
cat("Optimal Lambda = ", round(opt.lambda/0.5)*0.5, end="\n")
```

```
## Optimal Lambda = 0.5
```

A lambda value = 0.5 equates to a square root transformation of the response variable.

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

```
model4<-lm(Weight^(1/2)~ Species + Total.Length, data=fish2)
summary(model4)
```

```
##
## Call:
## lm(formula = Weight^(1/2) ~ Species + Total.Length, data = fish2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.0111 -0.7687 -0.0579  0.6797  4.6383
##
## Coefficients:
```

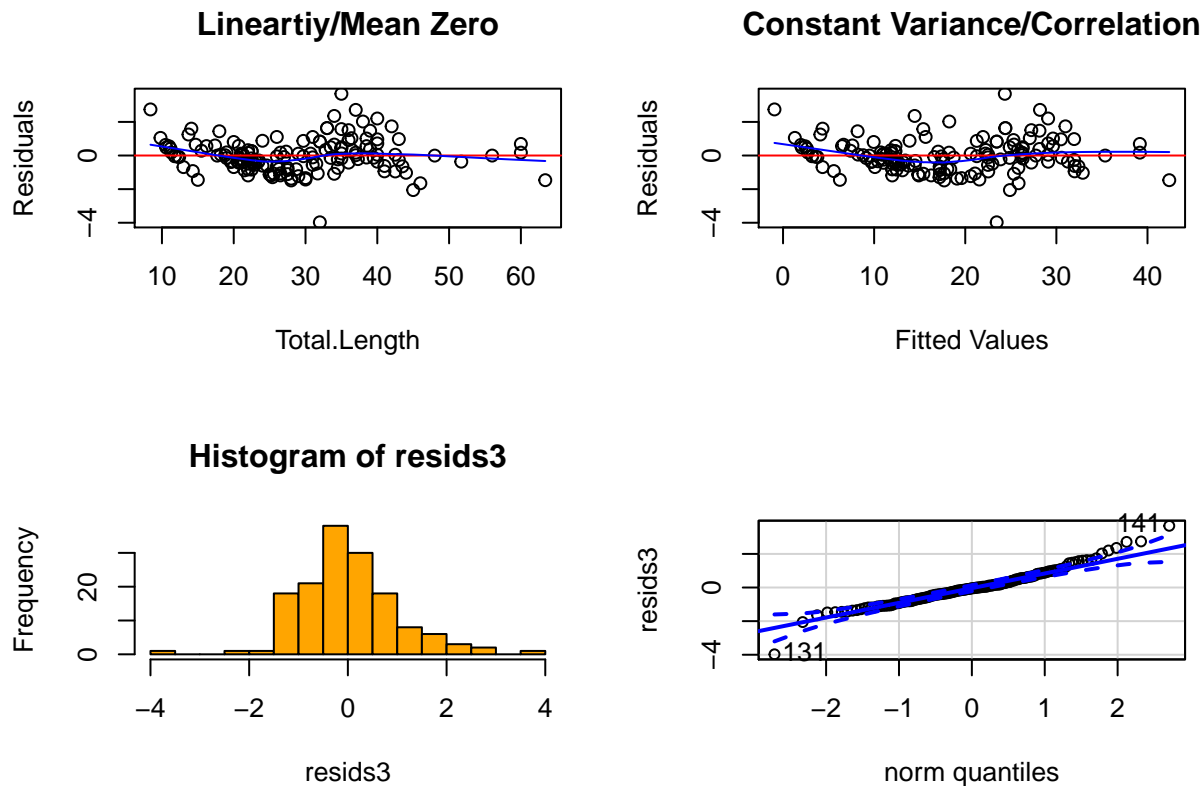
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -6.96654    0.57278 -12.163  < 2e-16 ***
## SpeciesParkki -0.36404    0.52476  -0.694   0.4890
## SpeciesPerch  -1.95734    0.29342  -6.671 5.46e-10 ***
## SpeciesPike   -10.90490    0.45001 -24.233 < 2e-16 ***
## SpeciesRoach  -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?

```
# Get standardized residuals
resids3 =rstandard(model4)

# Residual plots
par(mfrow=c(2,2))
plot(fish2$Total.Length, resids3, main="Linearity/Mean Zero",
     xlab="Total.Length", ylab="Residuals")
abline(h=0, col="red")
lines(lowess(fish2$Total.Length, resids3), col='blue')
plot(model4$fitted.values, resids3, main="Constant Variance/Correlation",
     xlab="Fitted Values", ylab="Residuals")
abline(h=0, col="red")
lines(lowess(model4$fitted.values, resids3), col='blue')
hist(resids3, col="orange", nclass=15)
qqPlot(resids3)
```





```
## 131 141
## 130 140
```

In the Linearity/Mean Zero plot, the residuals are now scattered evenly across the zero line. Linearity assumption holds.

In the Constant Variance/Correlation plot, the spread of the residuals show no pattern or increasing/decreasing trend. Constant Variance and assumption holds. Also, there does not appear to be any clear clustering in the residuals. This suggests that the errors are uncorrelated.

Lastly, the QQ plot shows that the majority of the residuals lines up considerably well with the normal line except for a slight tail and a few possible outliers. Overall, based on these plots, the normality assumption seems to hold.

Overall, the transformation was extremely successful in fixing everything from the weak linearity, to the constant variance, and the heavy left tail of the Histogram.

## 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.

**Rsquared:**

model2 = 0.9385

model3 = 0.9353

```
model4 = 0.9817
```

We expect the R-squared value of model3 to increase as we add more variables. Despite the results, we would not want to use R-squared to compare models with a different number of variables.

### Adjusted Rsquared:

```
model2 = 0.9335
```

```
model3 = 0.9321
```

```
model4 = 0.9808
```

Since the Adjusted R-squared adjusts the R-squared for the number of predicting variables in the model, we would say model2 explains more of the variance in the model than model3. Yet, this difference is extremely minimal. The transformed model4 has a much higher adjusted r-squared than the other two models. Thus, we prefer model4 over the other two models.

## 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.

```
# Calculate estimates
pred3<-predict(model3, fishtest)
# Calculate MSPE
mse.model3<-mean((pred3-fishtest$Weight)^2)

# Calculate estimates in terms of the original data
pred4<-predict(model4, fishtest)^2
# Calculate MSPE
mse.model4<-mean((pred4-fishtest$Weight)^2)

cat("The MSPE of model3 is", mse.model3, "\n")
```

```
## The MSPE of model3 is 9392.25
```

```
cat("The MSPE of model4 is", mse.model4, "\n")
```

```
## The MSPE of model4 is 2442.998
```

Since the MSPE of model4 is smaller than model3, based on this prediction accuracy measurement, model4 is preferred for predicting the weight of fish.

(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.

```
# Create new data point
new.point<-data.frame(Species="Perch", Total.Length=32)
# Calculate prediction interval
predict(model4, new.point, interval="prediction", level=0.9)^2
```

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

Model4 predicts the fish to weigh 461.94 grams with a 90% prediction interval of 374.45 to 558.61. We can be 90% confident that the weight of a fish with these specific characteristics will be between 374.45 to 558.61 grams.