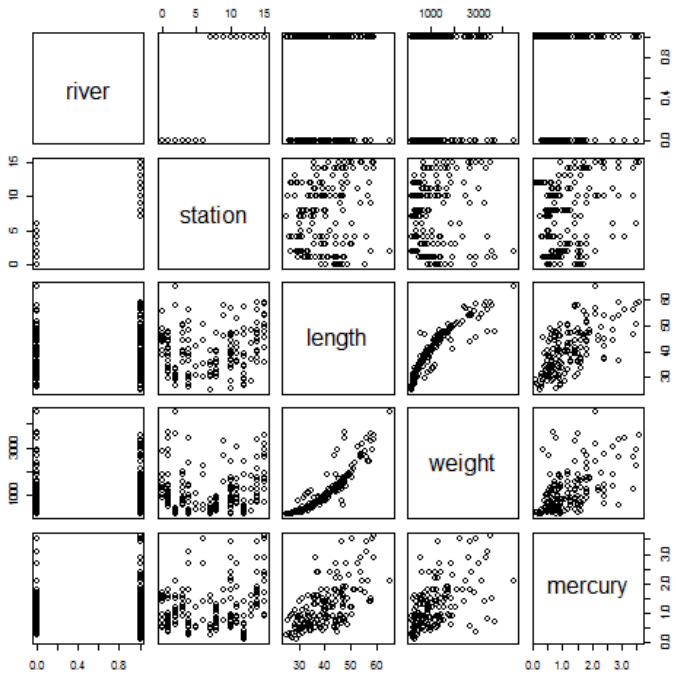
Introduction:

Something smells a little bit fishy! The North Carolina Department of Health and Human Services has monitored the levels of mercury in largemouth bass in the local waterways. They have sampled fish from two different rivers, of which we will call River0 and River1. In addition, these rivers were further divided up into 16 locations, known as “Stations”, and fish from each station were sampled and analyzed. In this study, we will model a relationship between mercury found in the fish given their station location, their length, weight, and whether or not they are deemed safe to eat.

Preliminary Analysis:

I first looked at the data to see if I could spot any collinearity, or if I should transform any of the data.

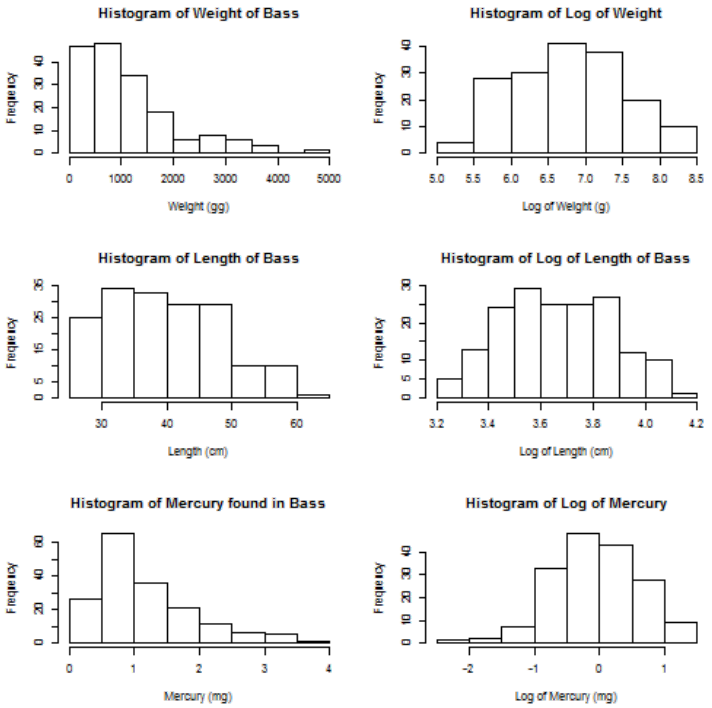
Plot 1.1A



There appears to be high correlation between length and weight, and thus I will remove their respective interaction term from all the models.

I also looked at histograms of the weight, length and mercury levels of the fish, in order to see if transformations would be necessary.

Plot 1.1B

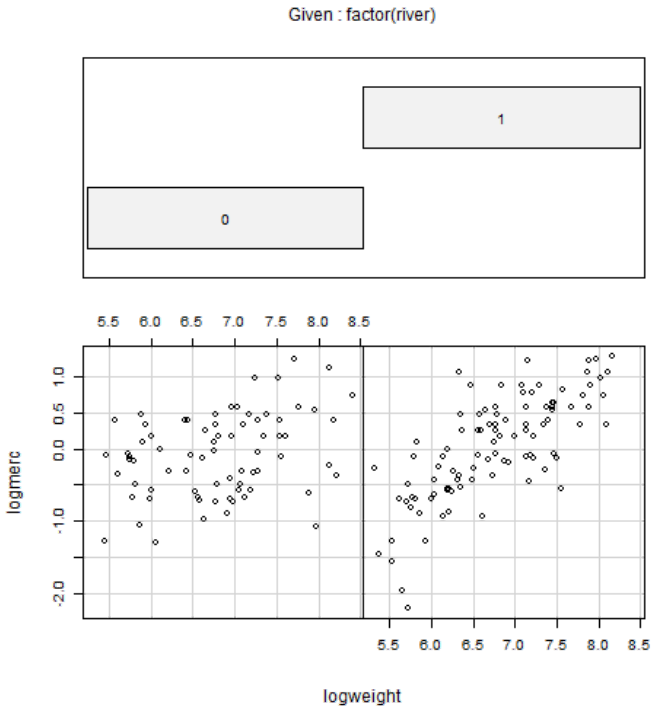


In order to minimize the increasing variance, I transformed that data with log functions, as the distributions appeared to become more normal by doing so. And since I was curious if location was correlated with mercury levels in the fish, I also introduced an indicator variable, data$safe, that described whether the fish sampled was toxic (containing more than 2 mg of mercury).

The next few plots show the relationships between location of the bass and the levels of mercury found.

Below is a graph (**Plot 1.2a**) of the mercury Vs weight (both are transformed) dependent on River0 and River1

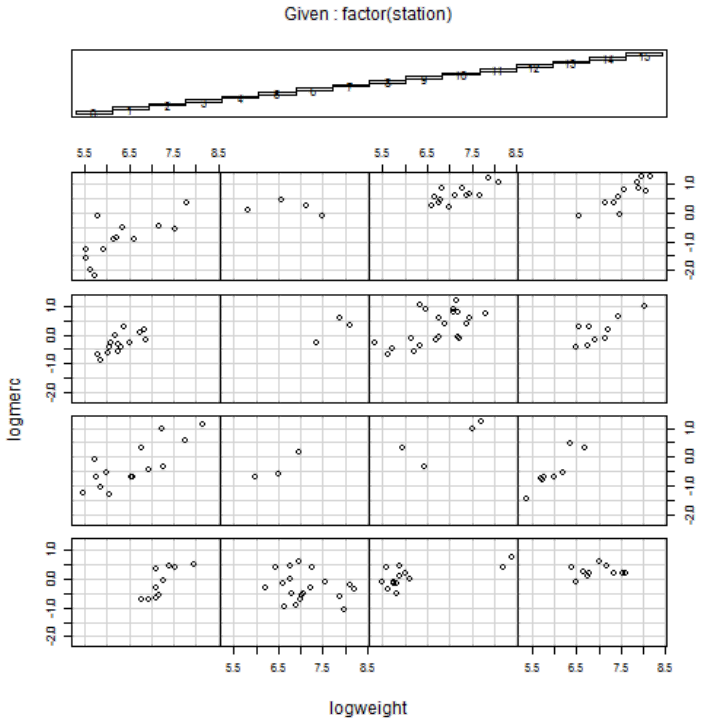
Plot 1.2a



There appear to be more fish in River 1 that have higher levels of mercury. In addition, in River0, there appears to be no relationship between mercury and weight, whereas in River1, there looks like there is a clear linear relationship between weight and mercury levels. In addition, it appears that smaller fish are less likely to have dangerous levels of mercury in them.

Below is a graph (**Plot 1.2b**) of mercury on weight (both are transformed) given the station of the fish.

Plot 1.2b



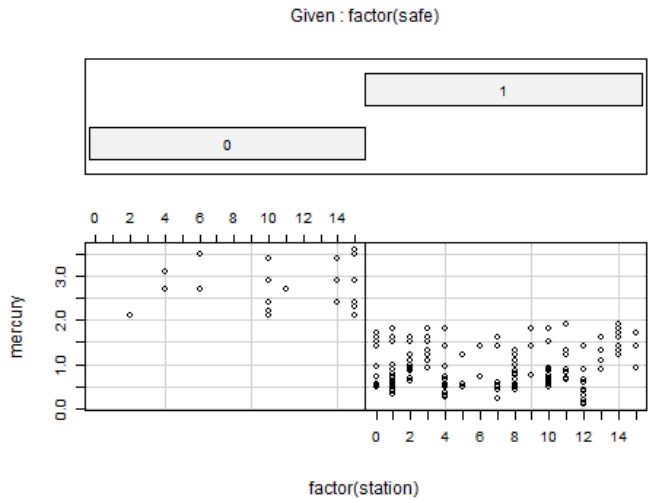
In many of the stations, there is a positive linear relationship between mercury and weight. Stations 1-3, and 13 show no clear relationship between mercury and weight. However, due to small sample sizes from these locations, removing any particular station from the data, can skew results dramatically, and therefore I have decided to keep all stations.

Table 1.1 in the appendix clearly shows the safety of certain fish in certain locations of the river. For example, only 45% of the fish in station 15 sample have less than 2mg of mercury. Also, from the third co-plot below, it can be seen that a smaller proportion of lighter fish are deemed unsafe, where as the proportion increases as the fish weight increases.

1 = Fish have less than 2mg mercury

0 = Fish have more than 2mg mercury (dangerous)

Plot 1.3



The above plot displays the relationship between stations and whether the fish are safe.The factor of 1, represents that the fish have less than 2 milligrams of mercury, and the factor of 0 represents that the fish have more than 2 milligrams of mercury. From this plot, it appears that there are high levels of toxic fish in stations 15 and 10 that would not be safe for public consumption.

Analysis:

Our first model is the full model, as I regress the log of mercury on the log of weight, the log of length, river location, river type, and the interactions among them. The subsequent models would remove certain terms (either parameters themselves, or their interactions). Again, I did not include the interaction between length and weight because the high collinearity between the two parameters. **Table 1.2** in the appendix describes the different models.

There is high correlation between the rivers, and station numbers. Thus, I removed river type from the data set. I chose river over station, because the residual sum of squares increased dramatically upon the removal of station from the data set. The residual sum of squares slightly increased when removing river.

```
##                              RSS
## With River and Station     15.68
## With Station/Remove River  15.68
## With River/Remove Station  34.84
```

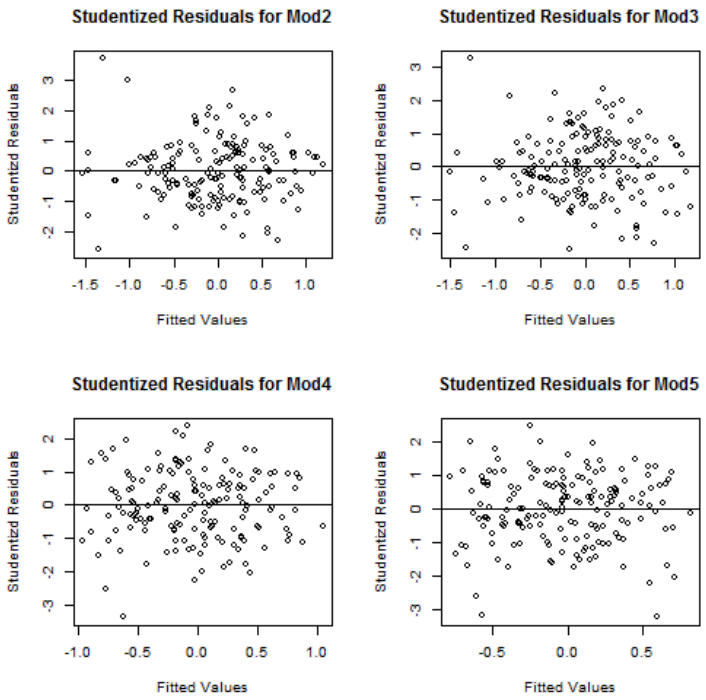
The table above displays the residual sum of squares for 3 models. As one can see, the RSS remains 15.677 when dropping river from the data set, whereas it jumps to 34.845, when dropping river. Therefore, river will be removed from the data set. The remaining models test the significance of the interaction terms.

```
##      Res.Df   RSS  Df  Sum of Sq      F    Pr(>F)
## mod0    123 15.68  NA         NA     NA        NA
## mod1    123 15.68   0 -3.020e-13     NA        NA
## mod2    138 17.84 -15 -2.161e+00  1.130 3.368e-01
## mod3    153 22.71 -15 -4.873e+00  2.549 2.468e-03
## mod4    168 40.04 -15 -1.732e+01  9.062 7.295e-14
## mod5    169 49.33  -1 -9.297e+00 72.941 4.251e-14
```

Looking at the anova table above, it appears that mod2 fits the data well better than the other models. However, not all off the coefficients in this model are significant. This can be due to the small sample sizes of each station as well as outliers and leverage points. I will analyze the outliers of all the models in an attempt to achieve better fits. However, mod2 is currently the model of choice because of its relatively high adjusted R^2, the analysis of deviance table, and the relatively low RSS. I have excluded mod1 from further analysis, because none of its coefficients were significant.

I will next observe the outliers from each of model (excluding mod0 and mod1 for reasons mentioned above), and then remove them in order to create a better fit.

Plot 1.4a



In each graph above, there are noticeable outliers. I will remove residuals greater than 2 and -2 from each model. Each updated model will be noted (i.e, “mod2update”). Then to further validate the model selection, I will look at other criteria, such as AIC, BIC, and Mallow's Cp to choose the best model.

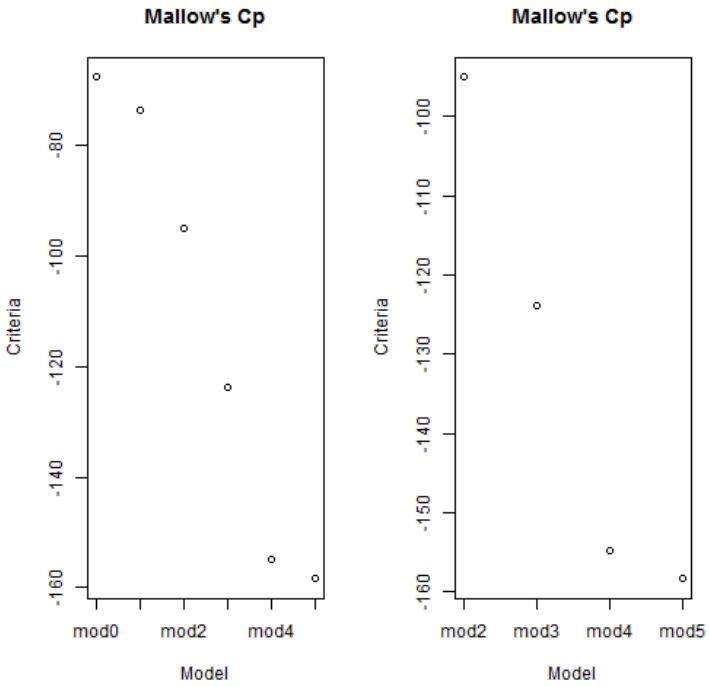
The AIC and BIC tables are provided below. Notice that mod2 and mod3 have the lowest levels for both compared to the rest of the models.

```
##            Df    AIC    BIC
## mod0       48 -312.6 -161.8
## mod1       48 -312.6 -161.8
## mod2update 33 -364.6 -260.9
## mod3update 18 -340.5 -283.9
## mod4update  3 -272.5 -263.1
## mod5update  2 -240.5 -234.2
```

Discussion:

By observing a plot of Mallow's Cp below (**Plot 1.4b**), we can see that Mod3,Mod4 and Mod5 have the lowest value.

Plot 1.4b

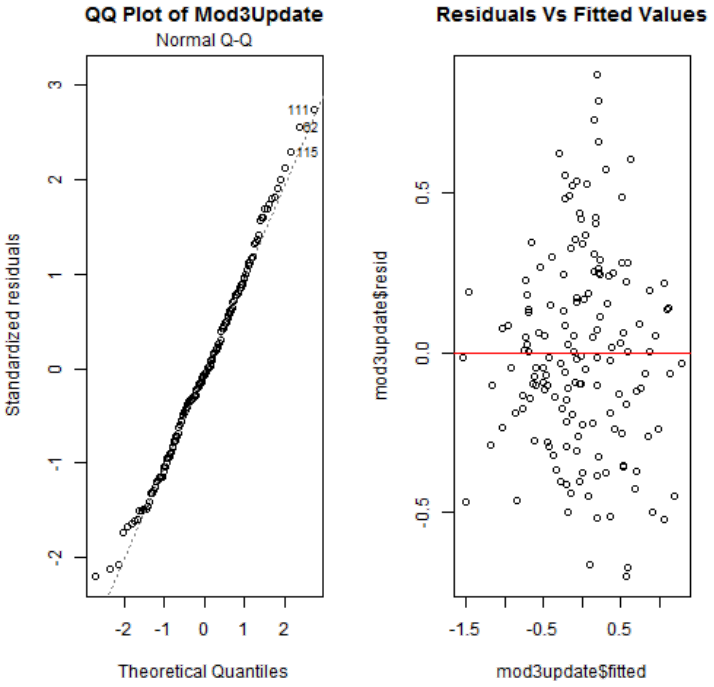


This leads us to the question, of which model we should choose? Mod2update has the highest R^2 and one of the lowest AIC and BIC values. However, many of its coefficients are NOT significant, which signifies that there is over-fitting. This can explain why it's Mallow's Cp value is not as low as the others. Mod4update and mod5update have the lowest Mallow's Cp values, however, they do not accurately describe the model. Both the models have very low adjusted R^2, and very high residual standard errors (0.4295 and 0.4797 respectively). Mod3 has one of the lowest AIC values and the lowest BIC value. In addition, it also has the second lowest residual standard error (behind mod2update), and nearly all of its coefficients are significant. Also, it has a relatively high adjusted R^2 value of 0.7496. Thus for these reasons, I believe mod3update is the best model. The table below describes these findings.

```
##             Df DfModel    AIC    BIC      Cp Adjusted R^2
## mod0        48     123 -312.6 -161.8  -67.55       0.7083
## mod1        48     123 -312.6 -161.8  -73.55       0.7083
## mod2update 33     129 -364.6 -260.9  -95.00       0.7894
## mod3update 18     142 -340.5 -283.9 -123.78       0.7432
## mod4update  3     160 -272.5 -263.1 -154.90       0.5073
## mod5update  2     163 -240.5 -234.2 -158.39       0.3997
##             Residual Standard Error
## mod0                         0.3570
## mod1                         0.3570
## mod2update                   0.2967
## mod3update                   0.3273
## mod4update                   0.4295
## mod5update                   0.4797
```

In order to verify that the mod3update is a great linear fit, I analyze its QQ plot and its residuals (**Plot 1.5**).

Plot 1.5



The QQ plot above follows a linear slope, and the residuals appear to have no patterns with the fitted values. In addition, the mean of the residuals is zero. Thus, mod3update appears to fit the data well.

Conclusion: Upon looking at our model and analysis, we found that mercury levels in the fish were best described by regressing on the log of weight, the log of length, and the location of the fish in the river (station). We also noticed through our analysis that a majority of fish that were smaller, tended to have safer amounts of mercury as compared to larger fish (**Plot 1.2a**). Fish in certain locations, such as station 10 and 15 seemed to have higher amounts of mercury than other locations (**Plot 1.2b**). However, despite all this, I believe this model will do a great job in explaining the mercury levels for largemouth bass.

APPENDIX:

Table 1.1

```
##            Percentage of Fish < 2 MG of Mercury
## Station 0                                 1.0000
## Station 1                                 1.0000
## Station 2                                 0.9231
## Station 3                                 1.0000
## Station 4                                 0.8571
## Station 5                                 1.0000
## Station 6                                 0.5000
## Station 7                                 1.0000
## Station 8                                 1.0000
## Station 9                                 1.0000
## Station 10                                0.6667
## Station 11                                0.8889
## Station 12                                1.0000
## Station 13                                1.0000
## Station 14                                0.6923
## Station 15                                0.4545
```

Table 1.2

```
##      Parameters
## mod0 Log of Weight, Log of length, river, station
## mod1 Log of Weight, Log of length, station
## mod2 Log of Weight, Log of length, station
## mod3 Log of Weight, Log of length, station, safety
## mod4 Log of Weight, Log of length
## mod5 Log of Weight
##      Interaction Terms
## mod0 logweight and river; logweight and station; loglength and river; loglength and
station
## mod1 loglength and station; logweight and safety; loglength and safety
## mod2 logweight and station
## mod3 No Interaction
## mod4 No Interaction
## mod5 No Interaction
```