Checking for Multicollinearity

Multicollinearity is defined as linear or near linear dependence between predictors. This can impact the usefulness of the linear regression model because it reduces or undermines the statistical significance of the predictor.

```r
library(faraway)

data <- read.csv("cleaned_data_2.csv", fileEncoding="UTF-8-BOM")
data1 <- read.csv("cleaned_data_scaled_only_2.csv", fileEncoding="UTF-8-BOM")
```

From the descriptive analysis (pairs plot), we saw that the predictors Length1, Length2, and Length3 were all highly correlated. Height and Width also looked highly correlated with with he other predictors as well, but formed groups of linear relationships. We can quantify the multicollinearty present in the model using the VIF (Variance Inflation Factor). But first, lets view a summary of our initial model.

```r
model <- lm(Weight ~ . - 1, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ . - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62240 -0.15065 -0.00678  0.11975  1.34893
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## Species_Parkki    -0.051269   0.113694  -0.451 0.652690
## Species_Perch     -0.060271   0.066537  -0.906 0.366490
## Species_Pike      -0.452263   0.190922  -2.369 0.019127 *
## Species_Roach      0.061044   0.084247   0.725 0.469846
## Species_Smelt      0.670119   0.108476   6.178 5.91e-09 ***
## Species_Whitefish  0.060055   0.131332   0.457 0.648141
## Length1            3.026638   0.931865   3.248 0.001436 **
## Length2           -0.181253   1.377454  -0.132 0.895490
## Length3           -2.107852   0.738504  -2.854 0.004929 **
## Height             0.481185   0.121979   3.945 0.000123 ***
## Width              0.004175   0.116539   0.036 0.971471
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3036 on 149 degrees of freedom
## Multiple R-squared:  0.9136, Adjusted R-squared:  0.9073
## F-statistic: 143.3 on 11 and 149 DF,  p-value: < 2.2e-16
```

Notice that the regression coefficients for Length1 is negative. This goes against our intuition, since it seems reasonable to think that the greater the length of the fish, the more it is going to weigh (regardless of what is actually being measured). Also note that these predictors do not have great significance. These issues both strongly indicate that multicollinearity is present.

```
vif(model)
```

```
## Warning in vif.lm(model): No intercept term detected. Results may surprise.
```

```
##    Species_Parkki     Species_Perch      Species_Pike     Species_Roach
##          1.542883          2.738220          6.724004          1.540307
##     Species_Smelt Species_Whitefish           Length1           Length2
##          1.787578          1.122944       1498.196392       3273.534619
##           Length3            Height             Width
##        940.955150         25.670627         23.431865
```

The VIF for the predictors shows that Length1, Length2, and Length3 are a major concern. They are almost completely linearly dependent. This makes sense given that these measurements are most likely measuring the length of the fish in three, almost identical ways. For this reason, we are justified in only keeping one of the lengths.

```
data$Length2 = NULL
data$Length3 = NULL
model <- lm(Weight ~ . - 1, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ . - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61102 -0.18494 -0.03696  0.07822  2.37901
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## Species_Parkki     0.17835    0.10755   1.658   0.0993 .
## Species_Perch      0.01546    0.05602   0.276   0.7830
## Species_Pike      -0.78434    0.18033  -4.349 2.50e-05 ***
## Species_Roach     -0.01042    0.07825  -0.133   0.8942
## Species_Smelt      0.78443    0.10548   7.437 7.21e-12 ***
## Species_Whitefish -0.02601    0.13825  -0.188   0.8510
## Length1            0.96916    0.12333   7.858 6.79e-13 ***
## Height             0.05500    0.05263   1.045   0.2976
## Width              0.16137    0.11791   1.369   0.1732
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.324 on 151 degrees of freedom
## Multiple R-squared:  0.9003, Adjusted R-squared:  0.8944
## F-statistic: 151.6 on 9 and 151 DF,  p-value: < 2.2e-16
```

```
vif(model)
```

```
## Warning in vif.lm(model): No intercept term detected. Results may surprise.
##    Species_Parkki     Species_Perch      Species_Pike     Species_Roach 
##          1.212463          1.704570          5.267820          1.166884 
##     Species_Smelt Species_Whitefish           Length1            Height 
##          1.484288          1.092661         23.044487          4.195884 
##             Width 
##         21.063370
```

We can see that removing these predictors has fixed the major issues multicollinearity was causing. The estimated regression coefficient for Length1 is now positive and has a significant linear relationship with the response. Note that Length1 and Width still have high VIF (Large VIF). This is likely due to some interaction with the indicator variables. Since Length1 and Width appear to be key variables, they can be left in for now. Hopefully after the model selection process is complete, the VIF's are more reasonable.

```
data1$Length2 <- NULL
data1$Length3 <- NULL

write.csv(data, 'cleaned_data_3.csv', row.names = FALSE)
write.csv(data1, 'cleaned_data_scaled_only_3.csv', row.names = FALSE)
```