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.

```
library(faraway)
data <- read.csv("cleaned_data_scaled_only.csv", fileEncoding="UTF-8-BOM")
data1 <- read.csv("cleaned_data.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.

```
model <- lm(Weight ~ Length1 + Length2 + Length3 + Height + Width - 1, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Length2 + Length3 + Height +
##     Width - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.68273 -0.17635 -0.07472  0.16207  1.25518
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## Length1  1.78126    1.11626   1.596   0.1126
## Length2 -0.25832    1.24277  -0.208   0.8356
## Length3 -0.91950    0.55952  -1.643   0.1023
## Height   0.33591    0.10391   3.233   0.0015 **
## Width    0.10963    0.09537   1.149   0.2522
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3422 on 154 degrees of freedom
## Multiple R-squared:  0.8859, Adjusted R-squared:  0.8822
## F-statistic: 239.1 on 5 and 154 DF,  p-value: < 2.2e-16
```

Notice that the regression coefficients for Length2 and Length3 are 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 are insignificant at $\alpha = 0.05$. These issues both strongly indicate that multicollinearity is present.

```
vif(lm(Weight ~ Length1 + Length2 + Length3 + Height + Width, data = data))
```

```
##    Length1    Length2    Length3     Height      Width
## 1681.49649 2084.25783  422.46825   14.57009   12.27536
```

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.

```
model <- lm(Weight ~ Length1 + Height + Width - 1, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = Weight ~ Length1 + Height + Width - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.69766 -0.20417 -0.08122  0.20661  1.23987
##
## Coefficients:
##         Estimate Std. Error t value Pr(>|t|)
## Length1  0.63298    0.05611  11.282  < 2e-16 ***
## Height   0.16325    0.04587   3.559 0.000494 ***
## Width    0.20868    0.07185   2.905 0.004212 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3439 on 156 degrees of freedom
## Multiple R-squared:  0.8832, Adjusted R-squared:  0.881
## F-statistic: 393.2 on 3 and 156 DF,  p-value: < 2.2e-16
```

```
vif(lm(Weight ~ Length1 + Height + Width, data = data))
```

```
##  Length1   Height    Width
## 4.204450 2.810496 6.894227
```

We can see that removing these predictors has fixed the issues multicollinearity was causing. The estimated regression coefficients are now positive and they have a significant linear relationship with the response.

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

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