

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)

data1 <- read.csv("cleaned_data_2.csv", fileEncoding="UTF-8-BOM")
data <- read.csv("cleaned_data_scaled_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 the other predictors as well, but formed groups of linear relationships. We can quantify the multicollinearity present in the model using the VIF (Variance Inflation Factor). But first, let's view a summary of our initial model.

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

```
##
## Call:
## lm(formula = Weight ~ . - 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.61452 -0.16026  0.00386  0.11097  0.99209
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## Species_Parkki   -0.041129    0.103127  -0.399   0.69060
## Species_Perch    -0.006328    0.062118  -0.102   0.91899
## Species_Pike     -0.565196    0.179374  -3.151   0.00197 **
## Species_Roach     0.054568    0.077854   0.701   0.48445
## Species_Smelt     0.613500    0.097630   6.284 3.45e-09 ***
## Species_Whitefish 0.106220    0.120320   0.883   0.37876
## Length1          3.484141    0.670441   5.197 6.57e-07 ***
## Length2         -0.986473    1.196934  -0.824   0.41116
## Length3         -1.648915    0.688777  -2.394   0.01791 *
## Height           0.453491    0.107867   4.204 4.50e-05 ***
## Width           -0.095669    0.105431  -0.907   0.36566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2805 on 149 degrees of freedom
## Multiple R-squared:  0.9263, Adjusted R-squared:  0.9208
## F-statistic: 170.1 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
## 1.486632 2.794973 6.950804 1.540472
## Species_Smelt Species_Whitefish Length1 Length2
## 1.695738 1.103812 908.203782 2894.694859
## Length3 Height Width
## 958.560486 23.509296 22.459362
```

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
## -0.65161 -0.17949 -0.03852  0.10274  2.45330
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## Species_Parkki   0.17810    0.10766   1.654   0.1002
## Species_Perch    0.09727    0.05547   1.754   0.0815 .
## Species_Pike    -1.05407    0.17669  -5.966 1.66e-08 ***
## Species_Roach     0.04869    0.07811   0.623   0.5340
## Species_Smelt     0.76853    0.10558   7.279 1.72e-11 ***
## Species_Whitefish 0.05582    0.13874   0.402   0.6880
## Length1          1.20600    0.12182   9.900 < 2e-16 ***
## Height           0.12014    0.05354   2.244   0.0263 *
## Width            -0.10868    0.11825  -0.919   0.3595
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3243 on 151 degrees of freedom
## Multiple R-squared:  0.9001, Adjusted R-squared:  0.8942
## F-statistic: 151.2 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
##	1.212207	1.667420	5.045785	1.159955
##	Species_Smelt	Species_Whitefish	Length1	Height
##	1.483688	1.097952	22.433565	4.333418
##	Width			
##	21.136621			

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(data1, 'cleaned_data_3.csv', row.names = FALSE)
write.csv(data, 'cleaned_data_scaled_3.csv', row.names = FALSE)
```