# Multiple Linear Regression Testing On Fish Weight

Damien Ha, Kowoon Jeong, Josh Xu, Jane Zou

2023-03-17

### Introduction

**Research Question** In this study, we constructed predictive model to study how various measurements of length (such as width and height) affect the weight of fish.

**Background / Source of Data** Fish weight is a critical factor in fisheries science and management due to its value in explaining the growth and living condition of fish populations. We pulled data from Kaggle that recorded seven common different fish species in fish market sales. We then built a multiple linear regression model to predict the weight of fish based on up to five explanatory variables of fish length. Using the standard relative weight for each fish, we can evaluate the health of fisheries, such as whether they are properly managed and whether the number of competing predators is stable.

**Methodology / Paper Overview** We will first conduct a preliminary analysis to observe any relationships among the variables. Then, we will fit a multiple least squares linear regression model using all the variables and observe whether any assumptions are violated. If so, we will transform the data and refit the model. If necessary, we will then consider model selection criteria to remove redundant variables from our model. After deciding on the best model, we will interpret the result and provide analysis, including the real life application of fish weight on fisheries' health.

### Data Description

Our data set of size [159,7] contains the following variables:

- Species: Categorical variable representing the species of the fish (bream, parkki, perch, pike, roach, smelt, or whitefish)
- Weight: Response variable, weight in grams
- Length1: Vertical length (cm)
- Length2: Diagonal length (cm)
- Length3: Cross length (cm)
- Height: (cm)
- Width: (cm)

Our data set had an anomalous measurement of a roach with 0 weight. In lieu of imputing the center value of the weight column for roaches, we removed the observation from the data.

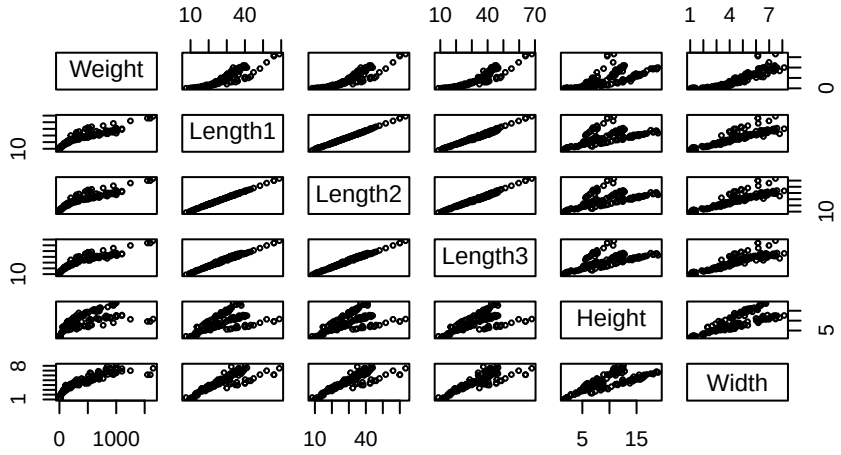
### Summary Statistics

```
##       Weight Length1 Length2 Length3 Height Width
## min     5.90    7.50    8.40    8.80   1.73  1.05
## mean  400.85   26.29   28.47   31.28   8.99  4.42
## max  1650.00   59.00   63.40   68.00  18.96  8.14
## sd    357.70   10.01   10.73   11.63   4.30  1.69
```

**Correlation Coefficient Matrix**

```
##            Weight   Length1   Length2   Length3    Height     Width
## Weight  1.0000000 0.9157195 0.9186031 0.9230903 0.7238573 0.8866536
## Length1 0.9157195 1.0000000 0.9995162 0.9920042 0.6244087 0.8666843
## Length2 0.9186031 0.9995162 1.0000000 0.9940830 0.6395032 0.8732011
## Length3 0.9230903 0.9920042 0.9940830 1.0000000 0.7026548 0.8781887
## Height  0.7238573 0.6244087 0.6395032 0.7026548 1.0000000 0.7924005
## Width   0.8866536 0.8666843 0.8732011 0.8781887 0.7924005 1.0000000
```

**Scatter Plot Matrix**




Weight has a moderately linear relationship with each variable, although its relationships with Height and Width appear closer to logarithmic. We also see strong correlation among the five predictor variables. This makes sense since we were given five different measurements for fish length when fish are three-dimensional creatures.

**Building the Model**

**Model 1: Untransformed Full Multiple Linear Regression** We start by building a multiple least squares regression model including all five predictor variables with no transformations. This was our resulting model:

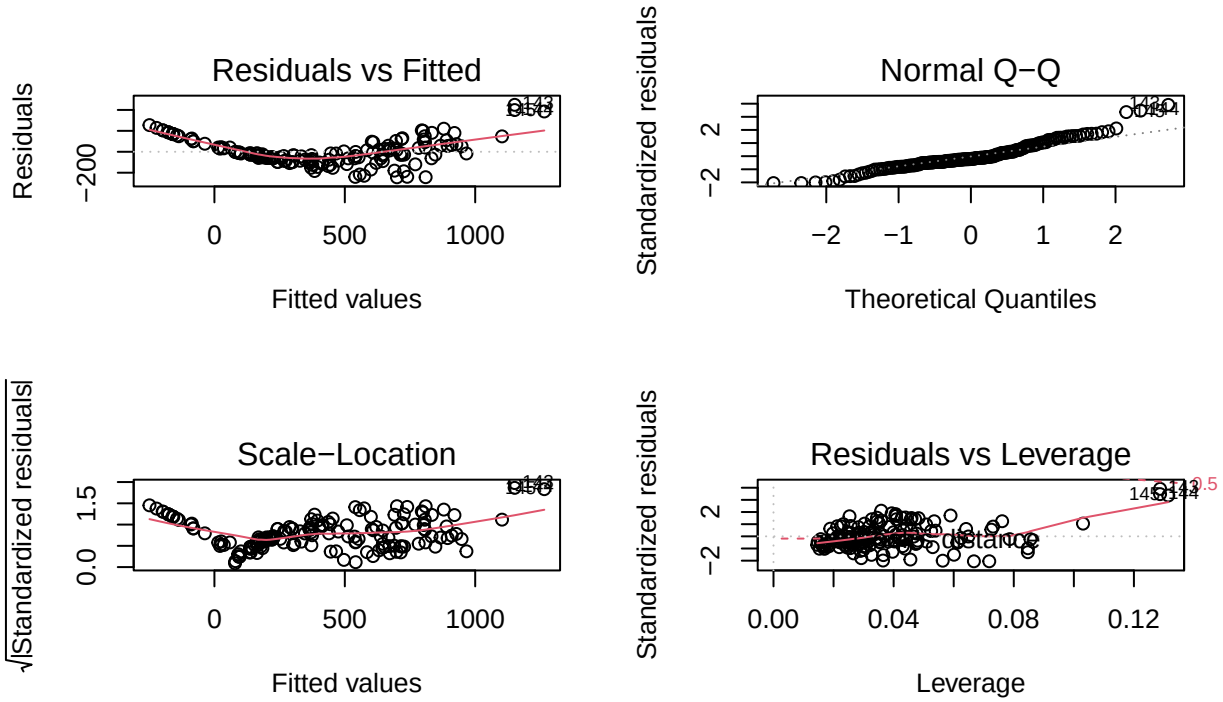
`Weight(hat) = -496.802 + 63.969Length1 - 9.109Length2 - 28.119Length3 + 27.926Height + 23.412Width`

$R^2 = 0.8855$

**Model 1 Evaluation** While our intuition would tell us that increasing the length measurements of a fish should increase its weight, we see that the coefficients of several measurements are negative. This suggests that our predictor variables are multicollinear and that some variables should be removed from the model. We can verify this by checking the variance inflation factors (VIF):

```
##    Length1    Length2    Length3     Height      Width
## 1677.71215 2082.43649  421.83683   14.55657   12.26187
```

We also check the four diagnostic plots to test our model assumptions (namely, linearity of the relationship, normality of the errors, and homoscedasticity of the errors):




The relationship between Weight and the predictors is nonlinear as shown in the Residuals v. Fitted Values plot. The Normal Q-Q plot and the Scale-Location plot show that the assumptions of normality and homoscedasticity of the errors may or may not be met. The errors appear to be normally distributed and homoscedastic except for a cluster of influential points, apparent in the Residuals vs. Leverage plot.

By checking Cook's distances, we can identify 10 influential points.

```
##          30          73         131         134         135         138         139
## 0.03072093 0.02761568 0.02626755 0.02897098 0.05020851 0.03991254 0.05418739
##         143         144         145
## 0.37589650 0.29678855 0.28458306
```

Given that the relationship between the predictors and Weight is nonlinear, we will transform the variables.

**Model 2: Power Transformation Applied to Model 1** We will apply a power transformation to simultaneously transform all variables at once. This is the result:

```
##   Y1   Y2   Y3   Y4   Y5   Y6
## 0.07 0.00 0.17 0.00 0.00 0.33
```

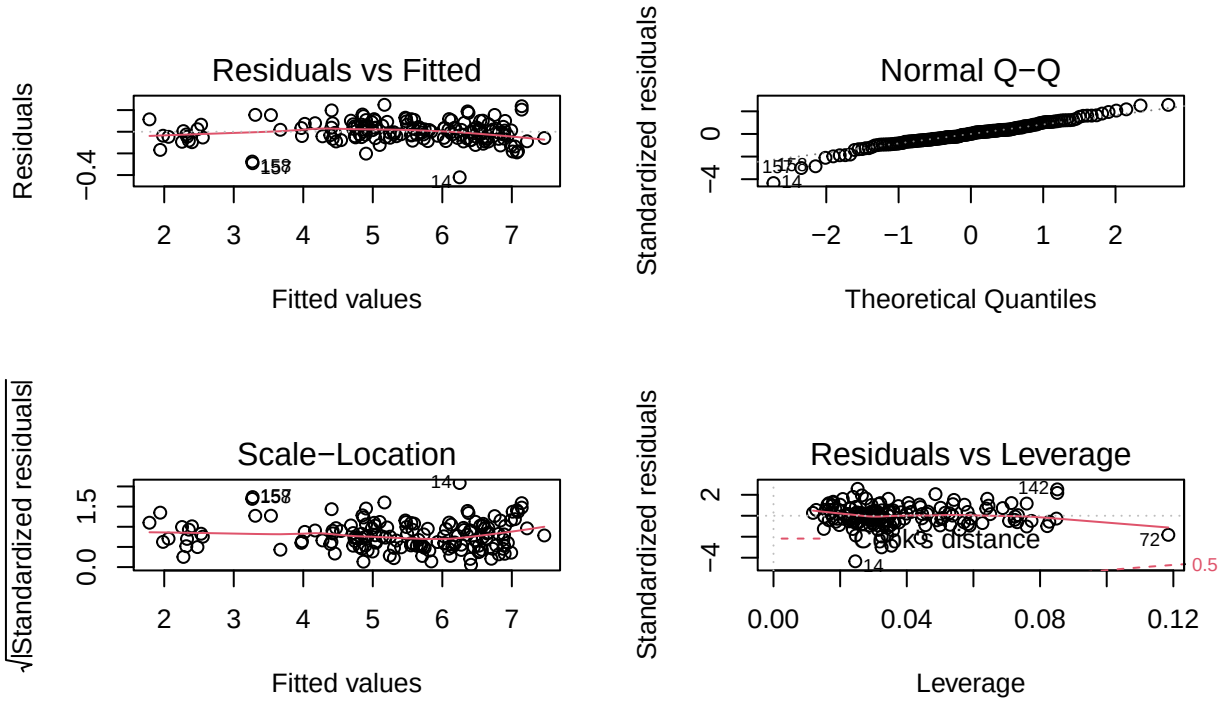
All variables should be transformed. To make our transformed model easier to interpret, we will round off every power to 0, i.e. apply a log transformation to every variable. This is our resulting model:

```
log(Weight) = -1.942 + 0.401log(Length1) + 1.596log(Length2) - 0.513log(Length3) +
0.68log(Height) + 0.845log(Width)
```

Multiple $R^2 = 0.9947$; Adjusted $R^2 = 0.9945$

**Model 2 Evaluation** We notice that the coefficient of log(Length3) is negative, despite our intuition suggesting that increasing Length measurements should increase Weight. Multicollinearity is still present among the predictors, as shown by the VIF's.

```
##   tLength1   tLength2   tLength3    tHeight     tWidth
## 1308.32791 1571.44764  388.49753   18.01198   23.51414
```




The response variable and predictors are moderately linear as shown in the scatter plot. The error terms are better normally distributed as the Normal Q-Q plot is more strictly straight and 45 degrees. The error terms have a more constant variance shown by the scale-location plot as they are relatively straight but clustered into under 1.2 and over 1.3. For further elaboration about the influential points, see the Appendix. We thus should reduce the model by selecting variables.

**Model 3: Reduced Model via Variable Selection** To choose which variables can be removed, we will apply the selection techniques of the foward stepwise Akaike information criterion (AIC), the forward stepwise Bayes information criterion (BIC), the backward stepwise AIC, and the backward stepwise BIC.

Models with these variables were selected for each method; note that logs have been taken for each variable:
- Forward stepwise AIC: Length2, Length3, Height, Width
- Forward stepwise BIC: Length2, Length3, Height, Width
- Backward stepwise AIC: Length2, Height, Width
- Backward stepwise BIC: Length2, Height, Width

We are thus left with two potential reduced models. Fortunately, the larger model is simply the other model with one extra predictor added, so we can run a partial F-test to see if adding the extra predictor (Length3) is significant or not.

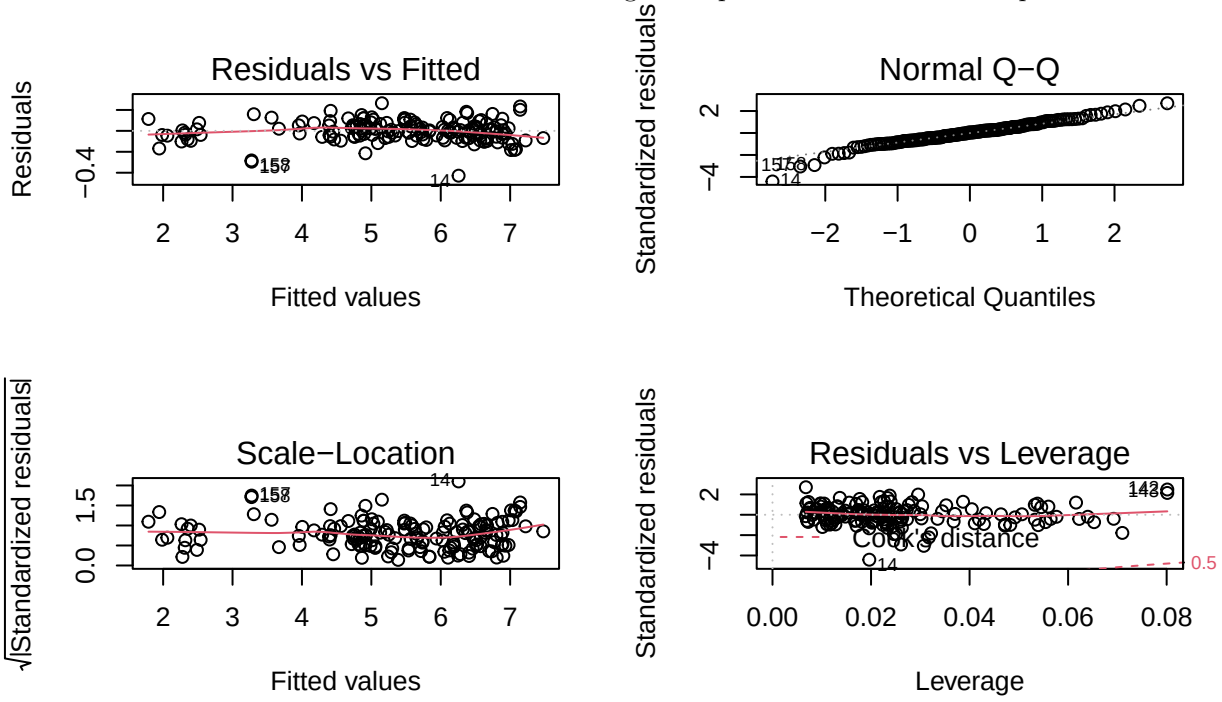
```
## Analysis of Variance Table
##
## Model 1: ty ~ tLength2 + tHeight + tWidth
## Model 2: ty ~ tLength2 + tLength3 + tHeight + tWidth
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    154 1.4802
## 2    153 1.4645  1  0.015644 1.6344  0.203
```

Since $p > 0.05$, we fail to reject the null hypothesis and conclude that there is no sufficient evidence in favor of the full model against the reduced model without Length3. Therefore this is our new model:

`log(Weight) = -2.01 + 1.498log(Length2) + 0.612log(Height) + 0.902log(Width)`

Multiple $R^2 = 0.9946$; Adjusted $R^2 = 0.9945$

**Model 3 Evaluation** We will check the four diagnostic plots to see if our assumptions are satisfied:




The relationship between Weight and the predictors is normal, and the errors are homoscedastic. We can see a small cluster of outliers in the Normal Q-Q plot, but we are satisfied that the errors are normally distributed. All of our assumptions are satisfied.

Since multicollinearity was a major issue in our two previous models, we will check the VIF's of Model 3:

```
##  tLength2   tHeight    tWidth
##  7.666024  5.507957 14.023230
```

The VIF's are still high, but they are lower than they were for models 1 and 2. It's natural to expect fish that are long in one direction to be long in every other direction, so this problem is unavoidable. Since fish are three-dimensional, it makes sense that our final model would include three length measurements.

**Model 3 Interpretation** Since 1) all the assumptions of multiple least squares regression are satisfied; and 2) we cannot simplify the model any further without losing accuracy, our final model to predict fish weight is model 3.

`log(Weight) = -2.01 + 1.498log(Length2) + 0.612log(Height) + 0.902log(Width)`

- If Height and Width are held constant, a 1% increase in Length2 results in a 1.498% increase in Weight
- If Length2 and Width are held constant, a 1% increase in Height results in a 0.612% increase in Weight
- If Length2 and Height are held constant, a 1% increase in Weight results in a 0.902% increase in Weight

Note that the y-intercept term of -2.01 is not meaningful in this context, since fish, being three-dimensional, cannot possibly take values of 0 for Length2, Height, or Width.

### Discussion

The use of linear regression models to predict the relationship between fish dimension and weight is a common approach in fisheries research. In a study by Vaseghi et al. (2020), the author investigated the relationship between length and weight of the commercially important fish species, Yellowfin tuna, using linear regression models. The result showed a strong correlation between length and weight. Similarly, a study by Polar and Ozekinci (2018) examined the relationship between length and weight of European anchovy and found

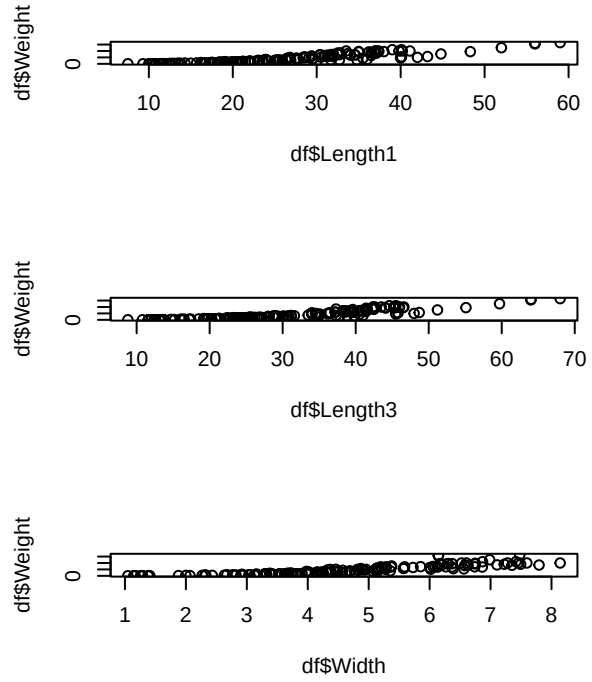
significant positive correlation between two variables. These studies provide evidence for the use of linear regression models to predict fish weight based on length measurements, supporting the findings of our present study. Our findings allow for the development of predictive models that can inform fisheries management decisions, such as setting catch limits and determining sustainable harvest levels.

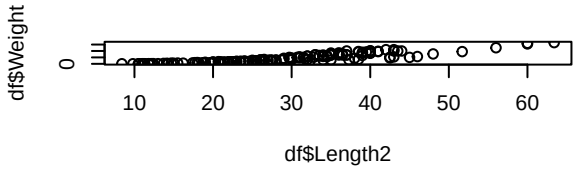
**Potential Future Improvements** In our model, we ignored the Species variable, instead choosing to use the same model to predict the weight of fish in every species. We were limited by a lack of sufficient data on some species (e.g. there were only six observations of whitefish). However, we recognize that each species is unique and may need to have their weights modeled with different predictors. We recommend that more research be conducted in this area, including, but not limited to, fitting individual models for each species once sufficient data is collected.

### Appendix

```
par(mfrow = c(3,2))
plot(df$Length1, df$Weight)
plot(df$Length2, df$Weight)
plot(df$Length3, df$Weight)
plot(df$Height, df$Weight)
plot(df$Width, df$Weight)
```

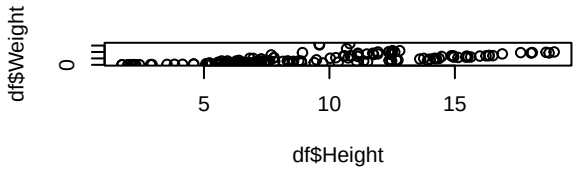
### Individual Plots of Predictors vs. Response







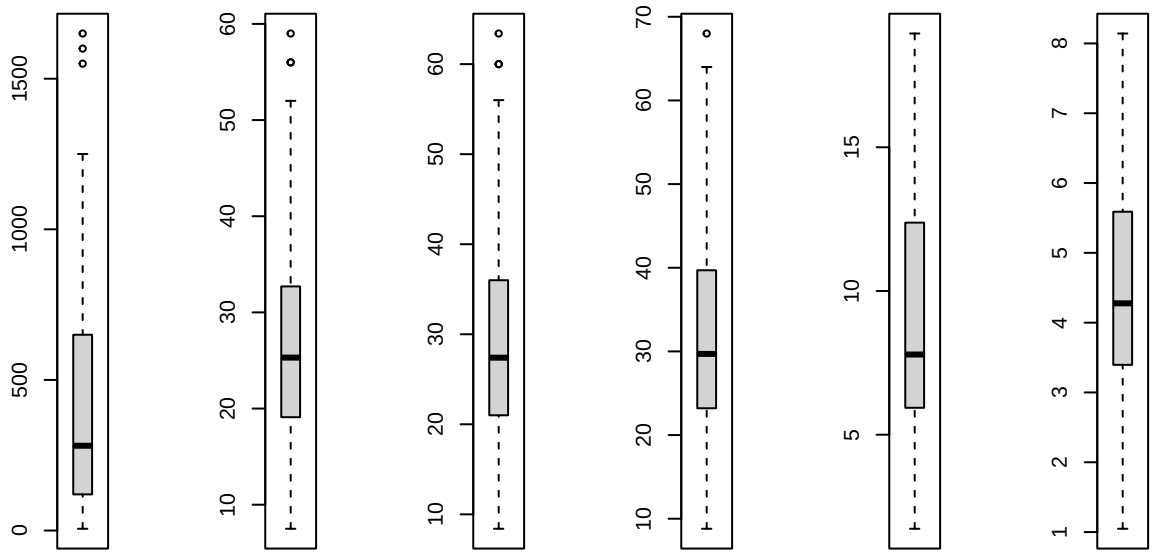






```
par(mfrow = c(1,6))
boxplot(df$Weight)
boxplot(df$Length1)
boxplot(df$Length2)
boxplot(df$Length3)
boxplot(df$Height)
boxplot(df$Width)
```

**Boxplots of Variables (EDA)**




```
summary(om1)
```

**Summary Output of Transformed Full Model**

```
##
## Call:
## lm(formula = ty ~ tLength2 + tLength3 + tHeight + tWidth)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.41982 -0.05574  0.00096  0.05450  0.24741
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.00498    0.12161 -16.487  < 2e-16 ***
## tLength2     1.97672    0.37843   5.223 5.66e-07 ***
## tLength3    -0.47655    0.37276  -1.278    0.203
## tHeight      0.67083    0.05627  11.923  < 2e-16 ***
## tWidth       0.83836    0.08003  10.476  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09784 on 153 degrees of freedom
## Multiple R-squared:  0.9947, Adjusted R-squared:  0.9946
## F-statistic:  7173 on 4 and 153 DF,  p-value: < 2.2e-16
```

```
summary(om2)
```

**Summary Output of Transformed Reduced Model**

```
##
## Call:
## lm(formula = ty ~ tLength2 + tHeight + tWidth)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.42835 -0.05739  0.00406  0.05504  0.26420
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.01044    0.12178  -16.51   <2e-16 ***
## tLength2     1.49774    0.05343   28.03   <2e-16 ***
## tHeight      0.61206    0.03251   18.83   <2e-16 ***
## tWidth       0.90246    0.06251   14.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09804 on 154 degrees of freedom
## Multiple R-squared:  0.9946, Adjusted R-squared:  0.9945
## F-statistic:  9525 on 3 and 154 DF,  p-value: < 2.2e-16
```

```
outliers_2 <- which(rstandard(model_2) > 2 | rstandard(model_2) < -2)
levpoints_2 <- which(hatvalues(model_2) > 12/158)
bad_levpoints_2 <- intersect(outliers, levpoints)
inflpoints_2 <- which(cooks.distance(model_2) > 4/156)
```

### Model 2 Influential and Leverage Points

- The leverage points are those above 12/158, which are 72, 129, 142, 143, 150, 151, 152, 153, 156
- The outliers are points 14, 77, 90, 96, 142, 143, 157, 158.
- The bad leverage points are those that are outliers and leverage points, which are points 142, 143, 144.
- The influential points are those with Cook's distance greater than 4/156, which are points 14, 36, 72, 77, 90, 96, 142, 143, 157, 158.

```
om1 <- lm(ty ~ tWidth + tLength3 + tHeight + tLength2)
p <- 4
n <- nrow(df)
Rad1 <- summary(om1)$adj.r.squared
AIC1 <- extractAIC(om1)[2]
AICc1 <- extractAIC(om1)[2] + (2 * (p + 2) * (p + 3) / (n - p - 1))
BIC1 <- extractAIC(om1, k = log(n))[2]
c(Rad1, AIC1, AICc1, BIC1)
```

### Transformed Full Model Accuracy Metrics

```
## [1]    0.9945574 -729.6085051 -729.0594855 -714.2955300
```

```
om2 <- lm(ty ~ tWidth + tHeight + tLength2)
p <- 3
n <- nrow(df)
Rad2 <- summary(om2)$adj.r.squared
AIC2 <- extractAIC(om2)[2]
AICc2 <- extractAIC(om2)[2] + (2 * (p + 2) * (p + 3) / (n - p - 1))
BIC2 <- extractAIC(om2, k = log(n))[2]
c(Rad2, AIC2, AICc2, BIC2)
```

**Transformed Reduced Model Accuracy Metrics**

```
## [1]    0.994535 -729.929654 -729.540044 -717.679274
```

**Sources** Vaseghi, S., Motallebi Moghanjoghi, A. A., & Mirshamsi, O. (2020). Length–weight relationships of yellowfin tuna (Thunnus albacares) in the Persian Gulf and Oman Sea. Journal of Applied Ichthyology, 36(2), 169-173. doi: 10.1111/jai.13963

Polat, N., & Özekinci, U. (2018). Length-weight relationships of the European anchovy Engraulis encrasicolus in the Black Sea. Turkish Journal of Fisheries and Aquatic Sciences, 18(7), 959-965. doi: 10.4194/1303-2712-v18_7_02

Alabama Cooperative Extension System. (2020, October 5). Relative Weight: An Easy-to-Measure Index of Fish Condition [Blog post]. Retrieved from https://www.aces.edu/blog/topics/fish-water/relative-weight-an-easy-to-measure-index-of-fish-condition/

Michigan Department of Natural Resources. (2013). Standard Methods for the Identification and Analysis of Aquatic Invertebrates in the Great Lakes Basin (Chapter 17: Macroinvertebrate Community Index). Retrieved from https://www2.dnr.state.mi.us/PUBLICATIONS/PDFS/ifr/manual/SMII%20Chapter17.pdf