Fish Market Analysis

Yufei Yin

Abstract

In this project, I will apply what I learned from **Linear Models In Applied Statistics** to a real-life example. I will focus on predict the Weight for different kind of fish. This project is going to focus on using statistical analysis and visualization techniques. Moreover, using Multiple Linear Regression and a series of tests and plots to build different models, and decide which is the optimal choice to complete our task.

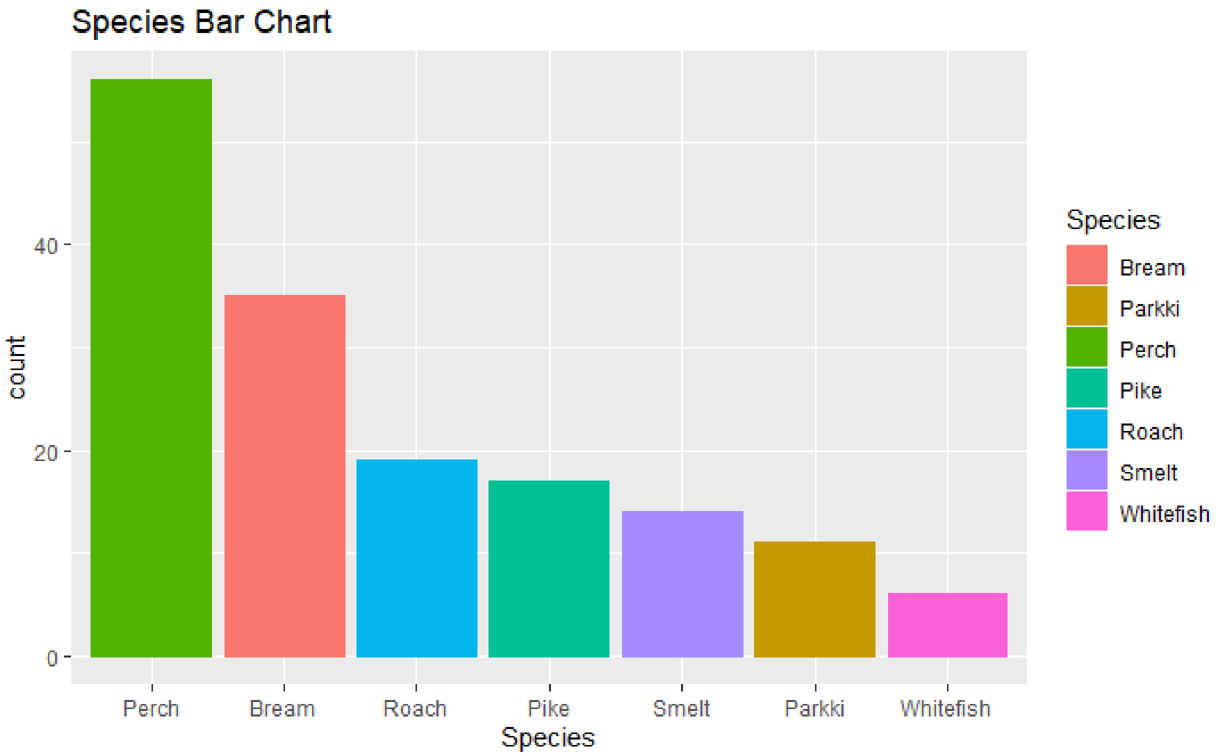
Introduction

In class, our datasets are usually clean and focus on only one problem for academic purpose; However, datasets can be really messy in real life. Statisticians spend 80% of their time on data wrangling, and only 20% for exploration and modeling. In order to have a better performance in the future, I am going to start a **Linear Regression Analysis** on a real-life example. The goal of this project is to predict the Weight for different kind of fish. The dataset fish market is from Kaggle.

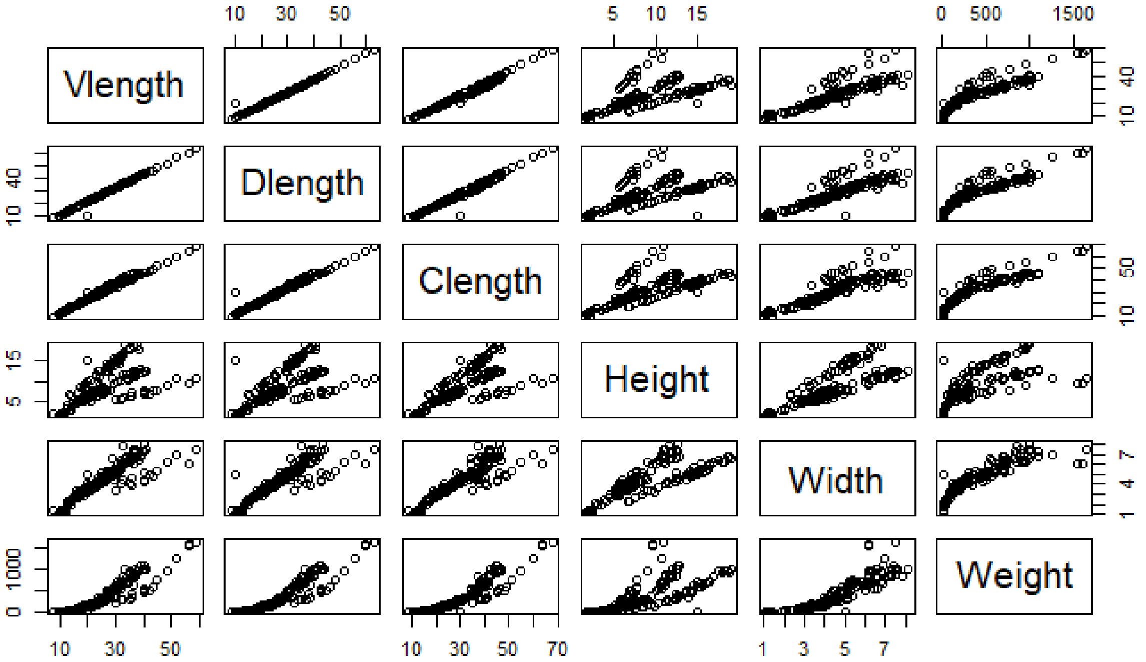
First, I will understanding the data and all its variables. Second, I am going to apply the method I learned from class to have a better understand of the data and the relationship between variables. Third, introduce the method we are using to build our model. Next get the result from our analysis using statistical analysis and visualization techniques to explain the result. Finally choose our model and validate our model.

Data Description

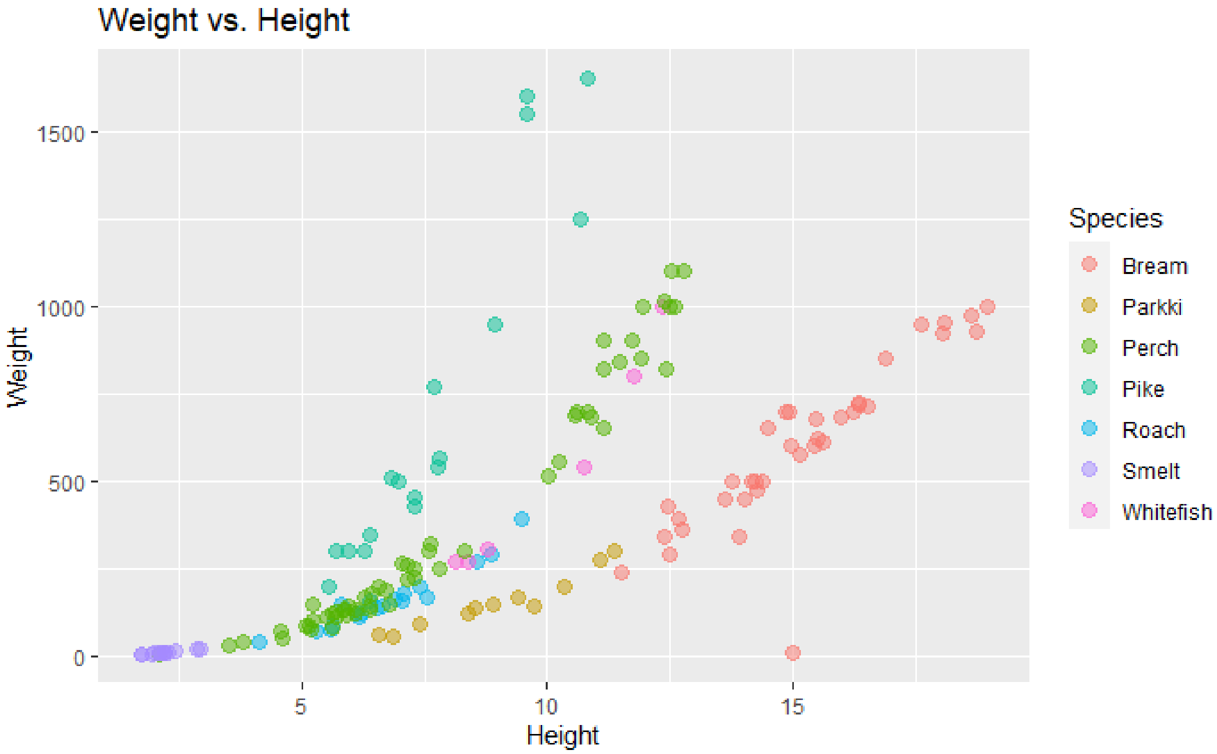
In our dataset, There are 7 variables in total. 1 qualitative variable Species and 6 quantitative variables (Vertical Length, Diagonal Length, Cross Length, Height, Width, Weight). Species, Vertical Length, Diagonal Length, Cross Length, Height and Width are predictor variables, and Weight is a respond variable. After taking a closer look of our dataset, we are lucky that our dataset is almost clean. We do not have any missing values, but one of our observation have Weight equal to 0 which does not make any sense so we delete this point.We add one extra point to our dataset in order to study the effect of outlier. After finish data cleaning, we can start analyze our data.



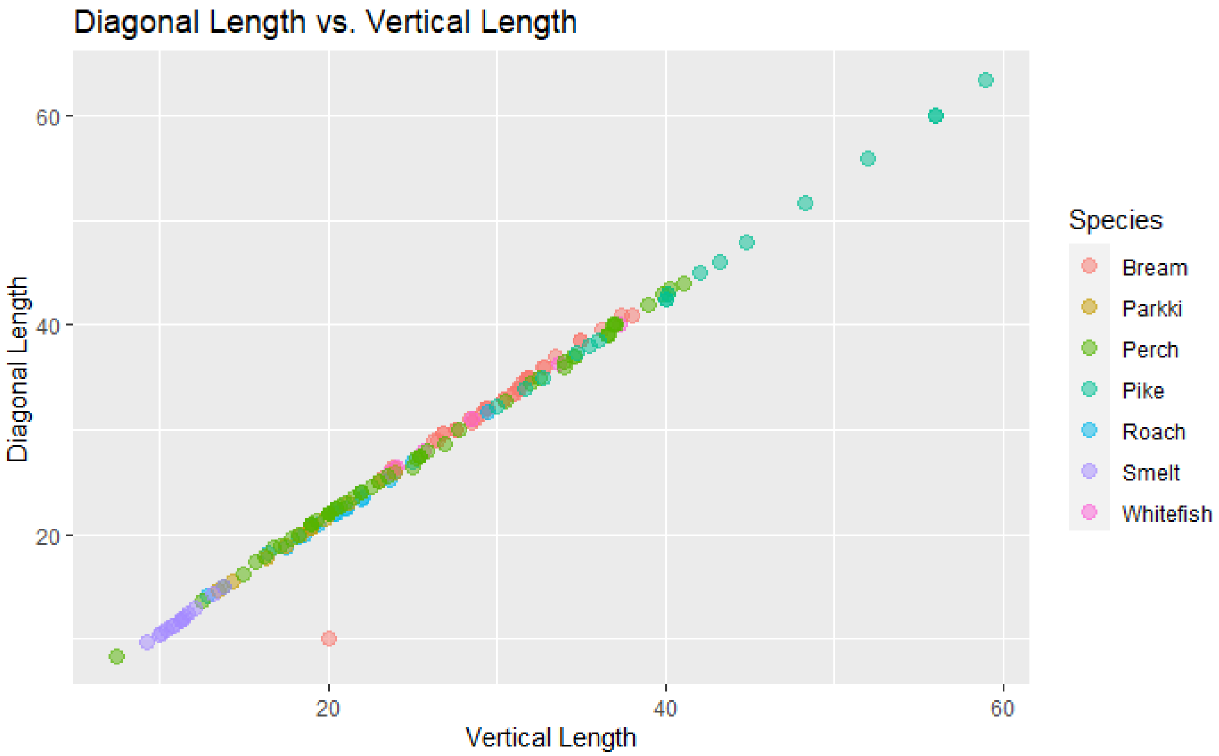
The bar chart shows the total amount of each Species. We can see the most Species is Perch and the least Species is Whitefish.



According to the pair plot, we noticed that the relationship between Weight and other variables seems linear; Moreover, There is a severe problem of multicollinearity especially between Vertical Length, Diagonal Length, and Cross Length. It is reasonable, since it is obvious that the fish have longer Vertical Length will have longer Diagonal Length and longer Cross Length. We will use ggplot() package in R to have a better visualization.



From the Weight versus Height plot, we can see that the relation between Weight and Height are different for each Species, and it may not be a linear relationship. We will take a closer look when we fit our model.



The relationship between Diagonal Length and Vertical Length is strictly linear.

The plot of other combination of variables can be found in Appendix.

Method

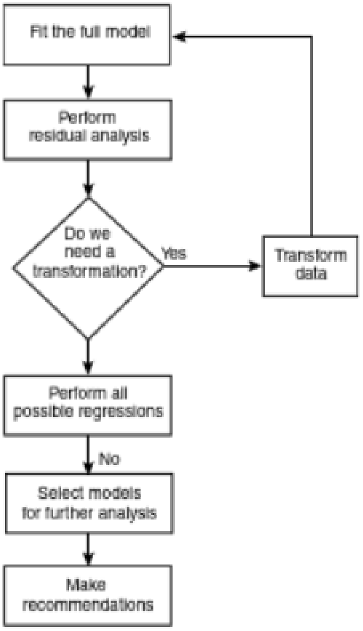

Fit the full model
Perform residual analysis
Do we need a transformation?
Yes
Transform data
Perform all possible regressions
No
Select models for further analysis
Make recommendations

Figure 10.11 Flowchart of the model-building process.

The figure shows the process of building our model. First, We fit the full model includes all the indicator variables (7 Species in total, so we will have 6 indicator variables), 5 quantitative variables and interaction terms. After fit the full model, we will perform residual analysis to check if we violate our assumptions. Then, it is always recommend to do a transformation even if our residual analysis is satisfied because we will have more options and we can choose the model that have the best performance.

Transformation

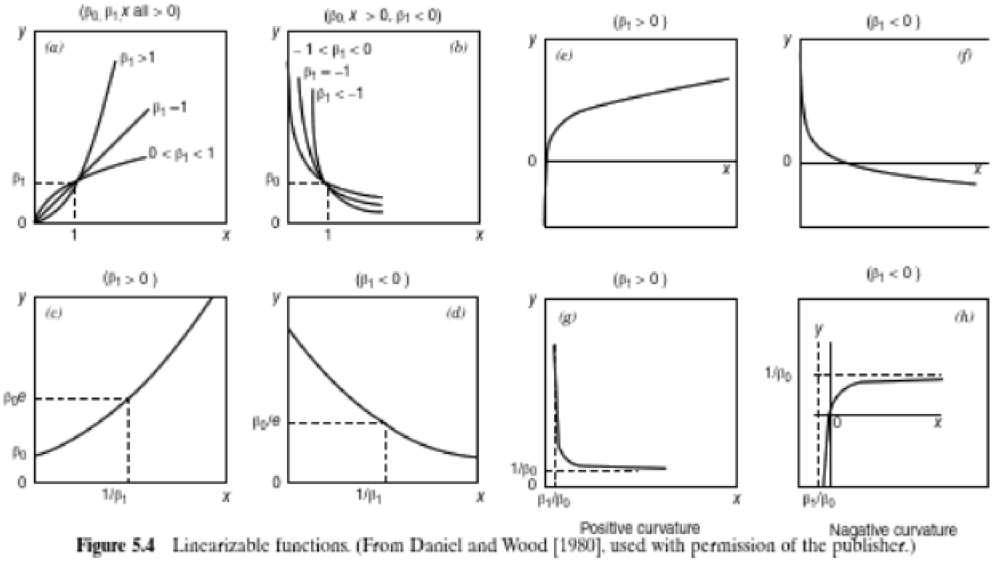


Figure 5.4 Linearizable functions. (From Daniel and Wood [1980], used with permission of the publisher.)

Figure 1: transformation-figures

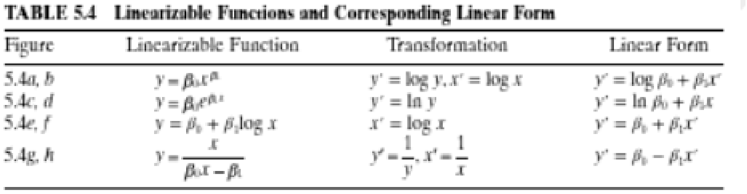


TABLE 5.4 Linearizable Functions and Corresponding Linear Form

Figure	Linearizable Function	Transformation	Linear Form
5.4*a*, *b*	$y = \beta_0 x^{\beta_1}$	$y' = \log y, x' = \log x$	$y' = \log \beta_0 + \beta_1 x'$
5.4*c*, *d*	$y = \beta_0 e^{\beta_1 x}$	$y' = \ln y$	$y' = \ln \beta_0 + \beta_1 x$
5.4*e*, *f*	$y = \beta_0 + \beta_1 \log x$	$x' = \log x$	$y' = \beta_0 + \beta_1 x'$
5.4g, *h*	$y = \dfrac{x}{\beta_0 x - \beta_1}$	$y' = \dfrac{1}{y}, x' = \dfrac{1}{x}$	$y' = \beta_0 - \beta_1 x'$

Figure 2: transformation-table

The Figures and table above shows how we perform transformations. In our case, our plot is similar to Figure 5.4a, so maybe a loglog transformation is applied. Again, we can try as much as possible until we get the model we satisfied.

High leverage points

Compare the diagonal of hat matrix with 2p/n (p is the number of estimated coefficients, n is the number of observations) if hat matrix is large than 2p/n then the observations can be considered as leverage points.

High leverage only indicates the location in x space. Observations with large leverage and large studentized residuals are potential influential points.

Influential points, Outliers

The points have cook distance large than 1 are consider to be influential. We will refit the model without outliers to see the difference of regression coefficients and decide whether we delete the outliers. However, outliers sometimes can be special cases which need further investigation.

After finishing transformation, we need to fit all possible regressions and perform model selection.

- Criteria for Evaluating Subset Regression Models
    - T-test, P-value
    - Coefficient of Multiple Determination
    - Adjusted R^2
    - Residual Mean Square
    - Mallow's C_p Statistic
    - The Akaike Information Criterion and Bayesian Analogues (AIC and BIC)

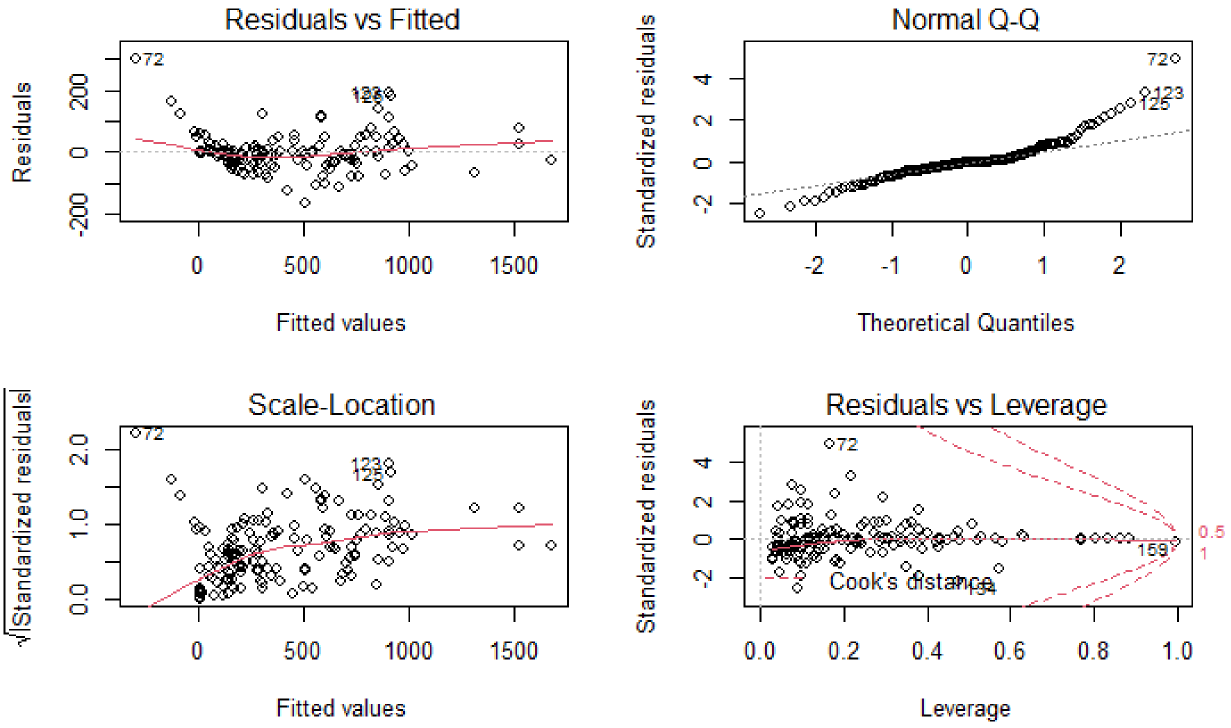
We will use step() function to specify stepwise regression algorithm to perform model selection and the default criteria is AIC. regsubsets() function is recommended if you want to use other criteria.

We need perform some further analysis after model selection. For example, To see whether we solve the multicollinearity problem.

- Detecting multicollinearity
    - pairwise correlations
    - variance inflation factors (VIF)
    - using the eigenvalues
        - condition number
        - condition index
- Methods for dealing with multicollinearity
    - collecting additional data
    - model respecification
    - ridge regression

Finally after finishing all the analysis above we will find the best model, then we need to perform data splitting also called cross validation. We split our dataset into two part (train data and test data) usually 80% to train data 20% to test data. Next we fit the model using train data, and predict Weight based on test data to see the performance of our model.

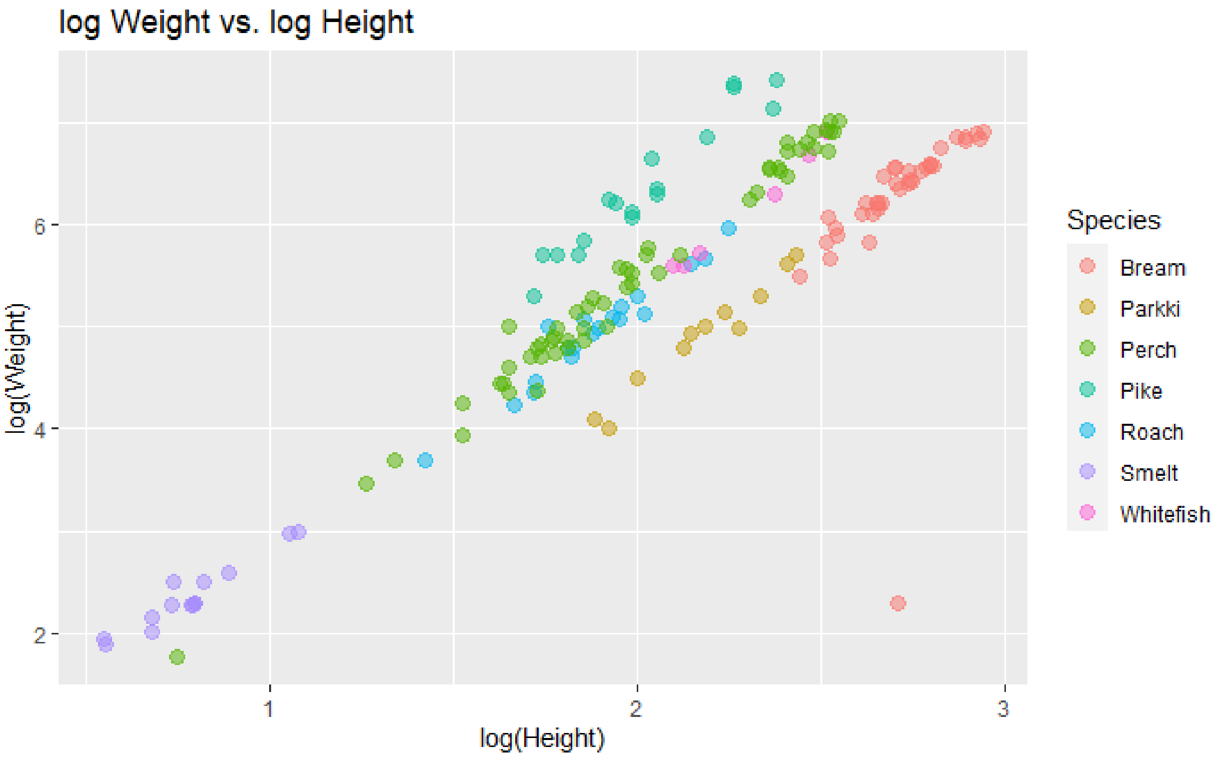
Result



The plot above is the residual plot after we fit the full model. We would like to see all the points from Residuals vs. Fitted plot and Scale Location Plot distributed around zero, with the same variance throughout the whole space, the points on Normal Q-Q plot should fall on the straight dotted line, and the residuals vs. Leverage plot we do not want to see any points have high leverage and high standardized residuals, we also do not want to see any points appear outside of Cooks Distance line marked by the dotted red line. From the plot, the full model do not perform very well, So we need to perform transformations.

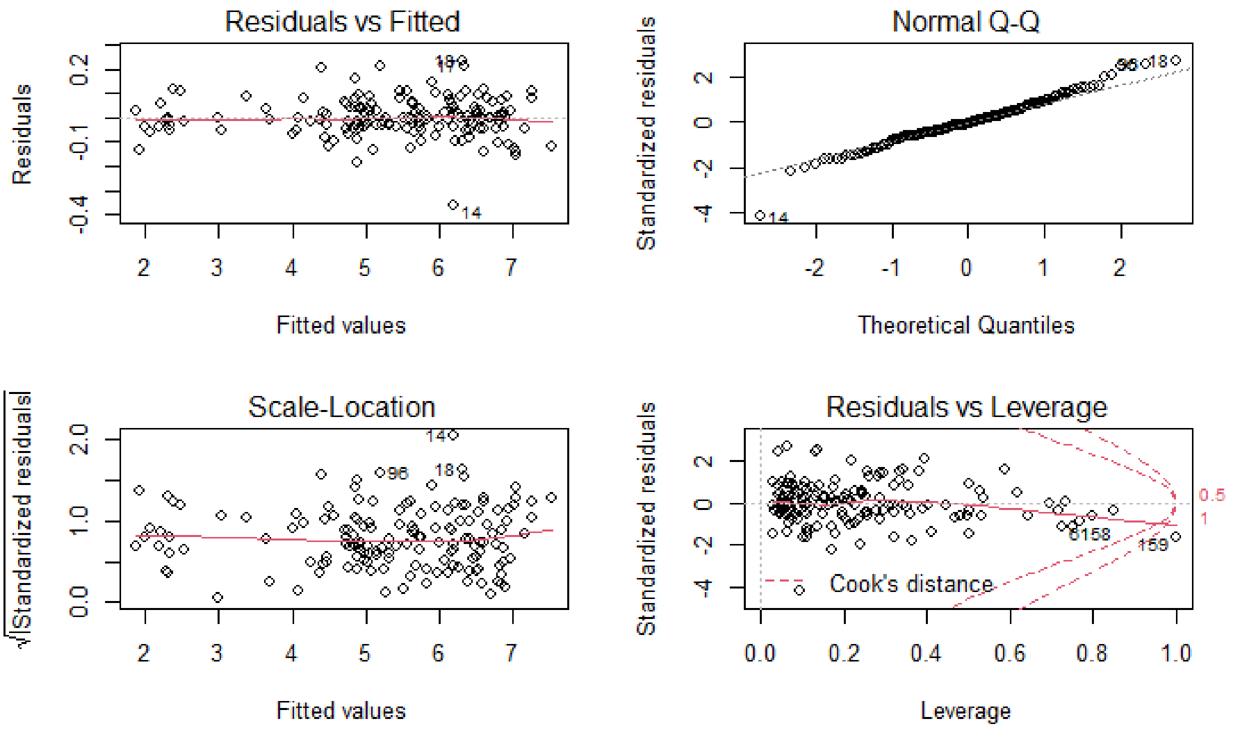
Transformation

Loglog transformation perform really well compares to other transformations.



From the plot we could see that it is more linear after transformation. The left lower-tail used to be curve shape, but it is almost straight after transformation.

Next we will look at the residual plot again after transformation.



From the plot, our assumption is met. However, there are still some problems we need to further investigate.

High leverage points

We have some high leverage points, but high leverage only indicates remote points. We do not have points have both high leverage and high studentized residuals. Observation 14 have studentized residuals larger than 3, but there is not a significant difference between coefficients after we remove it, so we will keep observation 14.

Influential points

Observation 159 have cook distance larger than 1 which is outside the red dotted line from redisual plot. This point is added on purpose to help us understand the effect of outliers. Without observation 159 we noticed a huge change in regression coefficients, So we should better delete this point.

If we have lots of outliers, robust regression is recommended.

Model Selection

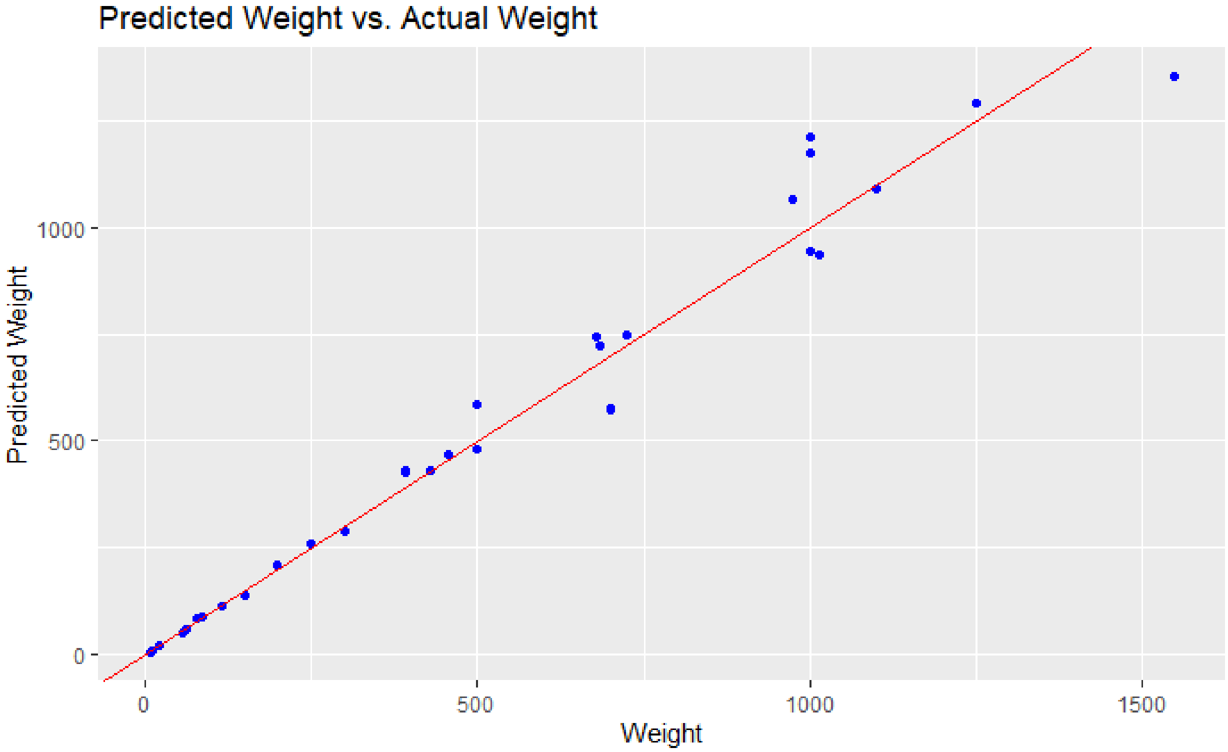
We use step() function to perform model selection, the result removed all the interaction terms indicates the slope will not change for different Species. the result kept all the indicator variable indicates the intercept is different for each Species. In addition, the result kept Cross Length, Height, and Width.

Multicollinearity

The correlation between Vertical Length, Diagonal Length, and Cross Length are almost 1 which means there is a severe problem of multicollinearity. It is not easy to solve by adding just one extra data point. The method we applied here is model respecification. After model selection we kept Cross Length, Height and Width. The VIF are 8.97, 11.94, and 7.57 respectively. The condition number is also larger than 1000. Since VIF for height is larger than 10, I will remove this variable to check if we can solve the problem. After we removed variable Height, VIF are 6.66 which is acceptable.

Then our final model includes Species, Cross Length, and Width.

Cross Validation



We expect to see most of the point falls on the red line, the plot indicates our model did a decent job to predict Weight. Repeat data splitting multiple times to further validate our model, R^2 is frequently used to check the performance of our model. I repeated data splitting 5 times, and the R^2 are 0.991, 0.968, 0.961, 0.986, 0.967 respectively. The R^2 closer to 1 means our model did well on predict weight.

Our Final Model:

Species Bream : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) - 3.28334

Species Parkki : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) + 0.17591

Species Perch : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) - 0.04493

Species Pike : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) - 0.44947

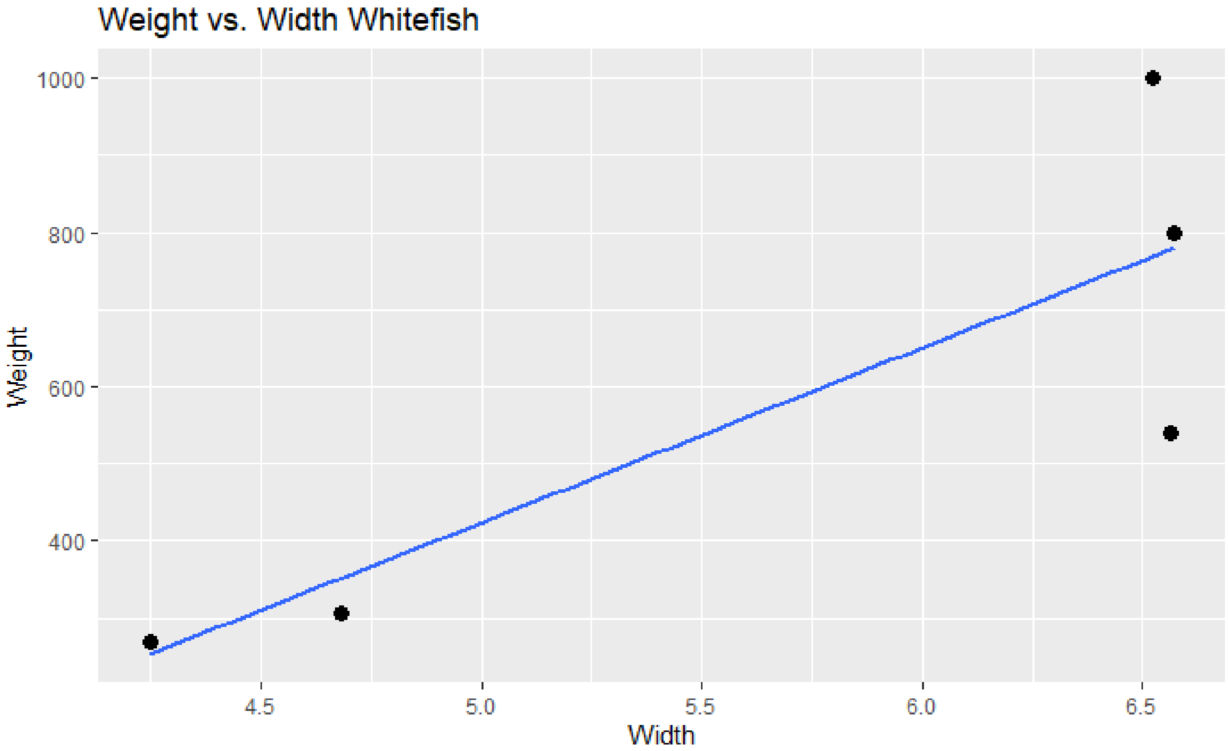
Species Roach : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) - 0.11548

Species Smelt : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) - 0.43618

Species Whitefish : log(Weight) = 2.28842×log(Cross Length) + 0.78238×log(Width) + 0.04059

Limitation

One thing i noticed is that our dataset is relatively small. The total number of some species are quite small. For example, The number of Parkki, Pike, Roach, Smelt and Whitefish are 11, 17, 19, 14, 6 which all less than 20.



Especially we only have 6 observations which is Whitefish, maybe a polynomial regression is more suitable for Whitefish. We need more samples to check our model and increase the accuracy of our prediction.

Conclusion

In this study, We predicted weight for different kind of fish. We find out using loglog transformation leads to a better model by far. However, our sample size for some of Species are quite small. More data points are needed for further analysis.

Appendix

Here is the link to my GitHub Repositories : **https://github.com/henryyinyufei/fish-market**

Here is all the images: image

Here is all the Rmarkdown files: Rmarkdown