Project 2: Predicting Fish Weight using Multiple Regression Analysis

Gabrielle Coronel, Kyle Zemel, Uyen Tran
Department of Statistics and Data Science, University of Central Florida
STA 4164: Statistical Methods III
Dr. Rong Zhou
December 1st, 2025

Abstract

The purpose of this study was to develop a predictive model for estimating fish weight using morphometric measurements. Using the Fish Market dataset from Kaggle, we examined the relationships between morphometric measurements (e.g., multiple length measurements, height, and width) and weight. In our initial data exploration, we identified nonlinear patterns and high correlations between the length-based measurements. These observations indicated that a linear model may not be the best model for the raw data, which led to a log-transformation on the full model. Diagnostic plots for the log-transformed full model showed improvements in linearity and homoscedasticity. Backward stepwise selection further refined the model by removing highly collinear predictors. However, the most parsimonious model was the single-predictor model using only log-transformed Width, which provided nearly identical predictive power compared to the more complex models. Overall, these results showed that multiple linear regression is not an ideal model for this dataset, and that simpler models are more appropriate when morphometric variables are inherently correlated.

Objectives

The purpose of this study is to determine the relationship of fish weight, measured in grams, with their physical attributes, measured in centimeters. Specifically, we considered vertical length, diagonal length, cross length, height, and width to find which predictor correlated most with fish weight. Because different fish characteristics influence a buyer's decision when purchasing fish, this demonstrates the importance of determining which independent variables contribute the most to a fish's weight. Additionally, by measuring the relationship between fish measurements and weight, this provides information regarding fish growth patterns and morphometric relationships, informing fisheries and conservation efforts by monitoring the health and size distribution of wild or managed fish populations. During this study, we analyzed the dataset using a multiple linear regression model to identify which predictor had the greatest impact on fish weight.

Source of Data and Environment of Study

For this project, version 2025.09.2+418 of R and RStudio were used to conduct statistical analysis on a macOS and Windows 11 operating system for the Fish Market dataset, which was obtained from Kaggle. Key libraries utilized for data manipulation, visualization, and statistical modeling included readr for data import, corrplot for visualizing correlation matrices, and car for analyzing the VIF.

The dataset contained 159 species of fish. For each fish, 6 continuous variables were measured: Weight (grams), three different length measurements in centimeters (Length1/Vertical length, Length2/Diagonal length, Length3/Cross length), Height (cm), and Width (cm). One categorical variable, Species, was also included in this dataset, which included the names: "Perch," "Bream," "Roach," "Pike," "Smelt," "Parkki," and "Whitefish." However, for this study, the Species variable was removed to create a model weight using physical traits only, without looking across multiple species.

Data Exploration

Before conducting a statistical analysis on the Fish Market dataset, data exploration was completed to gain an understanding of the initial patterns between fish physical features, namely, vertical length, diagonal length, cross length, height, and width, with fish weight. Specifically, we

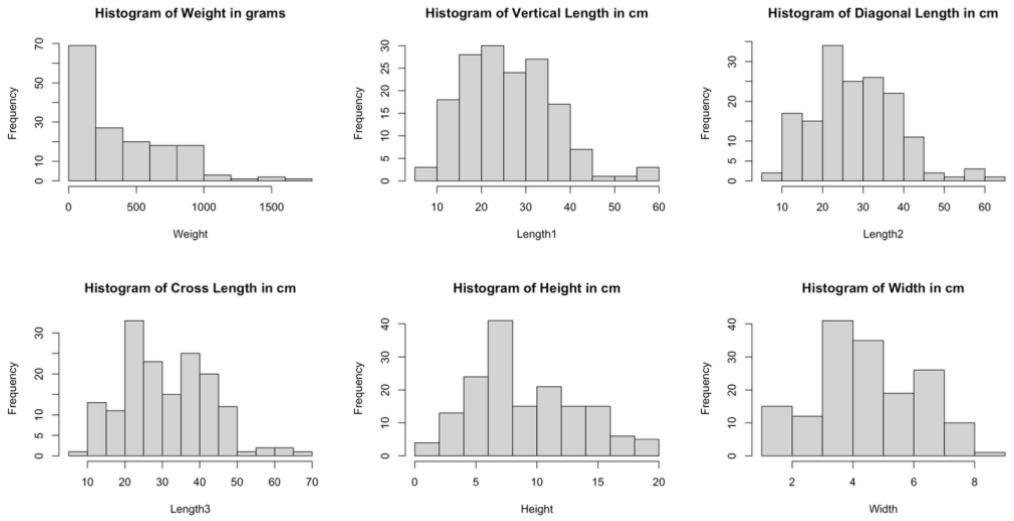
checked the summaries and descriptive statistics, using histograms, boxplots, and scatterplots, of the data. Looking at the results of the summary of each predictor, we get the following:

Summary of Length 1 (Vertical Length)					
Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
7.50	19.05	25.20	26.25	32.70	59.00
Summary of Length 2 (Diagonal Length)					
Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
8.40	21.00	27.30	28.42	35.50	63.40
Summary of Length 3 (Cross Length)					
Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
8.80	23.15	29.40	31.23	39.65	68.00
Summary of Height					
Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
1.728	5.945	29.40	8.971	12.366	18.957
Summary of Width					
Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
1.048	3.386	4.248	4.417	5.585	8.142

From the summaries of each independent variable, we observe that Weight had a wide range of 0 grams to 1650 grams, a mean of 398.3 grams, a median of 273.0 grams, and quartiles of 120 and 650 grams. Height and Width had much smaller ranges compared to Weight. Height had a range from 1.728 cm to 18.957 cm, a mean of 8.971 cm, a median of 7.786 cm, and quartiles of 5.945 cm and 12.366 cm. Width had a range from 1.048 cm to 8.142 cm, a mean of 4.417 cm, a median of 4.248 cm, and quartiles of 3.386 cm and 5.585 cm.

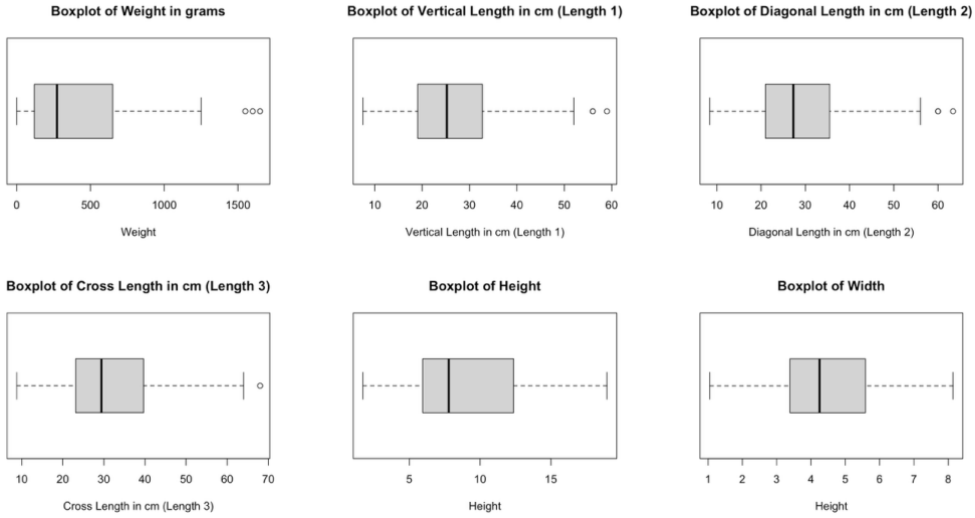
The three length-based measurements had similar distributions. Length1 ranged from 7.50 cm to 59.0 cm, with a mean of 26.25 cm, median of 25.20 cm, and quartiles of 21.0 cm and 32.70 cm. Length2 ranged from 8.40 cm to 63.40 cm, with a mean of 28.42 cm, median of 27.30 cm, and quartiles of 21.0 cm and 35.5 cm. Length3 ranged from 8.80 cm to 68.0 cm, with a mean of 31.23 cm, median of 29.40 cm, and quartiles of 23.15 cm and 39.65 cm. The similarity in these statistics across the three length variables indicated that these predictors may be highly correlated, which was later confirmed through the usage of a correlation matrix and variance inflation factor (VIF) values.

Further data exploration was completed on the Fish Market dataset using histograms, which are displayed below:



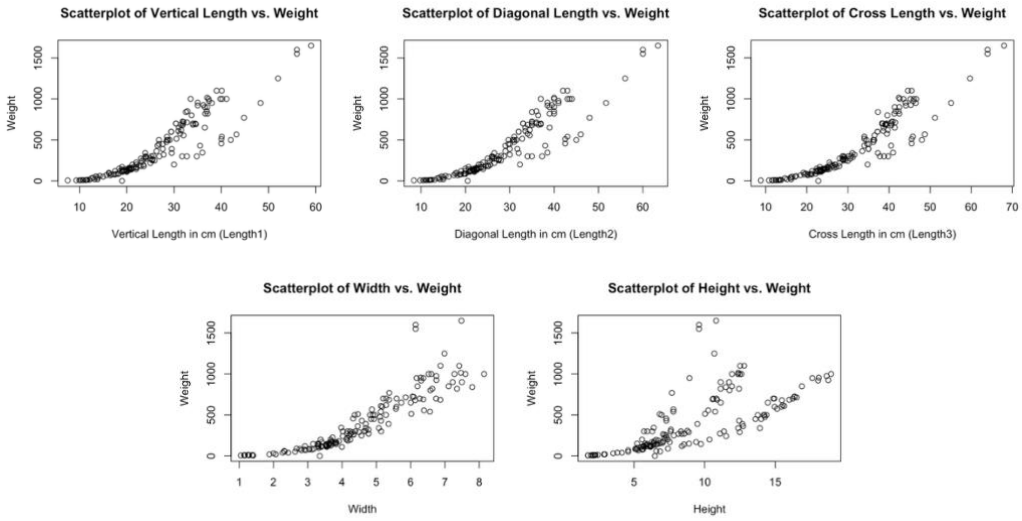
Based on these histograms, we can determine that weight is not normally distributed, but rather, it is right-skewed. However, our predictor variables have a relatively normal distribution with a slight right skew.

To find possible outliers in the dataset, boxplots were created, which gave us the following output:



The boxplots provided more insight into the variables' distribution and the presence of outliers in this dataset. Weight shows a right-skewed distribution with several high outliers. Height also displays a right skew, while Width displays a more symmetric distribution. The three length measurements show very similar distributions, each with a relatively symmetric spread and a small number of high outliers.

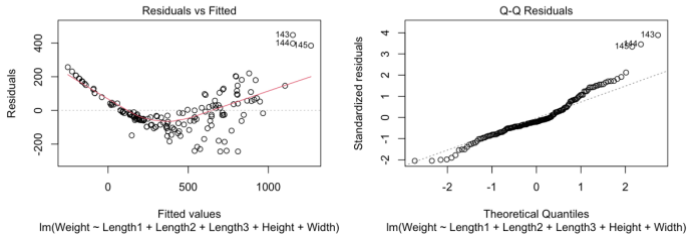
Finally, we graphed scatterplots of the independent variables against weight, which gave us the following:



The scatterplots comparing Weight to each measurement show clear positive relationships for all the predictors, indicating that larger fish tend to be heavier. However, the relationships appear curvilinear, exponential, rather than strictly linear, as each plot shows curvature, particularly for Weight, and suggests that a standard linear model may violate linearity assumptions. The scatterplots demonstrate a positive association between fish size measurements and weight, but suggest that a transformation or nonlinear models may be needed.

After performing our descriptive statistics, we created our first linear model and its diagnostic plots:

$$Weight = -4995.87 + 62.355(Length1) - 6.527(Length2) - 29.026(Length3) + 28.297(Height) + 22.473(Width).$$



Based on the diagnostic plots, multiple assumptions are violated. Specifically, linearity, homoscedasticity, and normality are all violated. After analyzing the Residuals vs. Fitted plot, we notice that the points follow a U-shaped curve, rather than a straight, horizontal line, which demonstrates violations of linearity and homoscedasticity. Additionally, after examining the Q-Q Residuals plot, we can see that the tails do not fall along the diagonal line, which shows that normality is violated.

Furthermore, we performed a summary of the linear model, which provided the following output

Variables	Estimate	Pr(>|t|)

Length1	62.355	0.12302
Length2	-6.527	0.87601
Length3	-29.026	0.09643
Height	28.297	0.00146
Width	22.473	0.27169

Based on the summarized output, we observe that most of the variables' p-values are not significant at the $\alpha = 0.05$ significance level, except for Height, since it has a p-value less than 0.05. Additionally, focusing on the Estimate column, we can spot negative coefficients for Length2 and Length3, or, namely, vertical length and cross length. However, biologically speaking, there should not be a negative relationship between length and weight, which raises some concerns for using a linear model for this dataset.

To check for collinearity in the dataset, a correlation matrix was created, which demonstrated the following results:

	Weight	**Length1**	**Length2**	**Length3**	**Height**	**Width**
Weight	1.000	0.916	0.919	0.923	0.724	0.887
Length1	0.916	1.000	1.000	0.992	0.625	0.867
Length2	0.919	1.000	1.000	0.994	0.640	0.874
Length3	0.923	0.992	0.994	1.000	0.703	0.879
Height	0.724	0.625	0.640	0.703	1.000	0.793
Width	0.887	0.867	0.874	0.879	0.793	1.000

Since the R values in the correlation matrix are all very high, close to 1, and positive, this means that the linear model shows strong, positive multicollinearity. Further inspecting the correlation matrix, we also find that the variables Length1 and Length2 have a perfect correlation, which raises additional concerns and questions. To continue checking for multicollinearity, the VIF values were analyzed, and is demonstrated by the following output:

Length1	**Length2**	**Length3**	**Height**	**Width**
1681.49649	2084.25783	422.46825	14.57009	12.27536

Because all the VIF values for the predictors are greater than 10, this provides further evidence that the linear model is multicollinear.

Data Preparation

Before conducting our study, data preparation was needed. Specifically, we checked for missing or N/A values, as these would hinder our results when answering our research question. After looking at our dataset, we found that there were no missing or N/A values. Additionally, we found that there were 159 observations, and concluded there was an appropriate amount since we have 6 total variables.

Originally, this dataset included a species column, which represented the species of each fish. However, we decided to remove the column to focus on analyzing how fish characteristics affected fish weight without considering different species of fish. It's important to note, though, that if the column were kept in our study, size dummy variables would have been implemented for the seven different species of fish.

Modeling

To understand which predictor affected fish weight the most, univariate models were created for the linear model. The following chart displays the adjusted R squares of each independent variable.

Predictors	**Adjusted R Squares**
Length 1 (Vertical Length)	0.8375
Length 2 (Diagonal Length)	0.8429
Length 3 (Cross Length)	0.8511
Width	0.7845
Height	0.5216

Based on the following output, the univariate model $Weight = \hat{\beta}_0 + \hat{\beta}_1(Length3) + \mathcal{E}$ has the highest adjusted R squared value. However, it's important to consider that the univariate models $Weight = \hat{\beta}_0 + \hat{\beta}_1(Length1) + \mathcal{E}$ and $Weight = \hat{\beta}_0 + \hat{\beta}_1(Length2) + \mathcal{E}$ also have similar adjusted R^2 values as the univariate model with Length3. Because the adjusted R squared values are almost the same, this further suggests multicollinearity in the linear model.

Since the linear model has no interaction terms, assessing confounding is appropriate. When testing for confounding against each predictor variable, we used $\hat{\beta}_{Length1|Length2|Length3|Height|Width} = 22.473$, which was calculated using R from the summary of the full model. Univariate linear models were created, and the following shows the Beta Parameters for each variable from the reduced models and calculations to determine confounding:

Variables	Beta Parameter	Calculation
Length1	32.792	$100\% \left\lvert \frac{32.792 - 22.473}{22.473} \right\rvert = 45.91732301\% > 20\%$
Length2	30.686	$100\% \left\lvert \frac{30.686 - 22.473}{22.473} \right\rvert = 36.54607752\% > 20\%$
Length3	28.4602	$100\% \left\lvert \frac{28.4602 - 22.473}{22.473} \right\rvert = 27.08672629\% > 20\%$
Height	60.496	$100\% \left\lvert \frac{60.496 - 22.473}{22.473} \right\rvert = 169.1941441\% > 20\%$
Width	188.249	$100\% \left\lvert \frac{188.249 - 22.473}{22.473} \right\rvert = 737.6674231\% > 20\%$

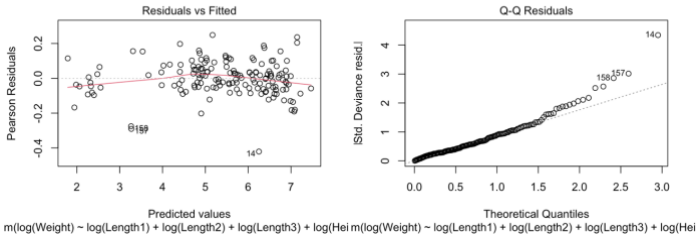
Based on the following calculations, we can see that all our variables in the models are confounders since they're all greater than 20%. All the predictors greatly affect the linear model, and there's an

appreciable change in $\hat{\beta}_{Length1|Length2|Length3|Height|Width}$ when each independent variable is added to the model. Specifically, Width affects the linear model the most as it has a percentage of about 738% which is significantly greater than 20%.

Because the Residuals vs. Fitted plot and Q-Q plot of the full linear model demonstrated multiple violated assumptions, specifically, homoscedasticity, linearity, and normality, and there was prominent multicollinearity, a log transformation was performed to test if it would improve and be a better fit. The following equation displays our new model, a summary of its output, and diagnostic plots:

$$\log(Weight) = -1.94224 + 0.40077(\log(Length1)) + 1.59553(\log(Length2)) - 0.51316\big(Log(Length3)\big) + 0.68039\big(Log(Height)\big) + 0.84464(Log(Width))$$

	Estimate	Pr(>|t|)
(Intercept)	-1.94224	<2e-16
log(Length1)	0.40077	0.5671
log(Length2)	1.59553	0.0387
Log(Length3)	-0.51316	0.1777
Log(Height)	0.68039	<2e-16
Log(Width)	0.84464	<2e-16



Based on the transformed log summary of the new model, this demonstrates an improved fit, as most of the variable coefficients are positive, excluding length3 and cross length. Furthermore, most of the p-values are also significant since they're less than $\alpha = 0.05$. However, Length1 and Length3 are still not substantial since their p-values are greater than 0.05.

Additionally, focusing on the transformed log diagnostic plots, it also displays a better fit of the data. Specifically, looking at the Residuals vs. Fitted Plot, there's no longer a U-shaped curve of points, but instead, a more horizontal spread of the data around the residual = 0 line. This shows that the assumption for linearity and homoscedasticity is met. Furthermore, looking at the Q-Q plot, we also notice that the points demonstrate a better fit of the data since the bottom tail is now following the diagonal line. However, the top tail could still be improved, as it's not on the line, and the assumption for normality could be further improved.

Additionally, since the linear model demonstrated concerns of multicollinearity, a correlation matrix and VIF values were made for the full transformed log model, which resulted in the following output:

	Log(Weight)	**Log(Length1)**	**Log(Length2)**	**Log(Length3)**	**Log(Height)**	**Log(Width)**
Log(Weight)	1.000	0.960	0.966	0.972	0.920	0.980
Log(Length1)	0.960	1.000	0.999	0.995	0.800	0.923
Log(Length2)	0.966	0.999	1.000	0.996	0.811	0.930

Log(Length3)	0.972	0.995	0.996	1.000	0.843	0.931
Log(Height)	0.920	0.800	0.811	0.843	1.000	0.901
Log(Width)	0.980	0.923	0.930	0.931	0.901	1.000

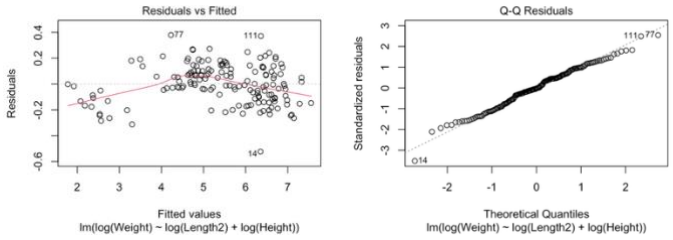
Log(Length1)	**Log(Length2)**	**Log(Length3)**	**Log(Height)**	**Log(Width)**
1308.32791	2084.25783	388.49753	18.01198	23.51414

To continue finding a better-fitted model, backwards stepwise selection was conducted on the full log-transformed model, which removed predictors that did not significantly contribute to the model, and resulted in the following new model:

$$\log(Weight) = -2.0104 + 1.4977\ (\log(Length2)) + 0.6121\ (\log(Height)) + 0.9025\ (\log(Width))$$

Since multicollinearity was an apparent issue for the linear model, we also checked the VIF values for the reduced log transformation after backward stepwise selection was performed, along with making diagnostic plots for the new model. The following chart displays the VIF results and diagnostic plots:

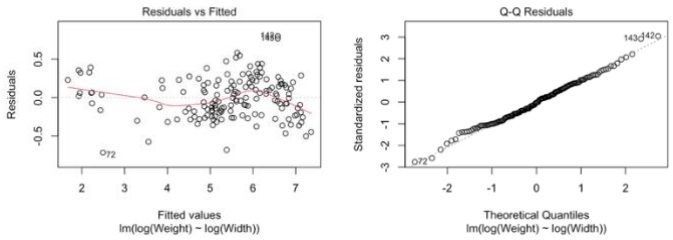
log(length2)	**log(Height)**	**log(Width)**
7.666024	5.507957	14.023230



Looking at the Residuals vs. Fitted plot, because the points are randomly scattered horizontally, this does not violate the assumptions of homoscedasticity and linearity. Additionally, focusing on the Q-Q Residuals plot, we can see that both the bottom and top tails better fit the diagonal line, further demonstrating a better fit of the model.

Additionally, looking at the VIF results, we find that Width has the highest VIF value and is still greater than 10, which continues to indicate serious collinearity. To remediate this, Width was taken out of the model, and collinearity was checked again using VIF. The following chart displays the new results with the new model and its diagnostic plots:

Log(Length2)	**Log(Height)**
2.918266	2.918266



Based on the chart, we see that the VIF values improved. However, in the search for a parsimonious model, we also considered R^2 values for the reduced log model with length2 and height, and the univariate transformed log model of Width. The following chart displays the output of the R^2 results:

Model	**Adjusted R Square**
Log(Weight) = log(Length2) + log(Height) + E	0.9872
Log(Weight) = log(Width) + E	0.9601

Looking at the following chart, we notice that both models have a similar R-squared value. Since the R-squared values are almost the same, with about a 0.02 difference, we consider the model $\log(Weight) = 1.5433 + 2.7724(\log(Width))$ as the best-fitted model because it's parsimonious.

Final Best Model

The final selected model uses log(Width) as the predictor since it achieves nearly the same R^2= value as the model with log(Length2) and log(Height). When choosing the best model, we consider a high R value and parsimony, which leaves us with the model log(Width) as a predictor.

$$\log(Weight) = 1.5433 + 2.7724(\log(Width))$$

This model satisfies most regression assumptions, has no significant collinearity, and explains 96% variability in the log of the fish weight according to its high R^2 value. The diagnostic plots for this final model showed improved normality of residuals and homoscedasticity compared to the original untransformed model.

Discussion

This model shows that a 1% increase in fish width is associated with a 2.77% increase in fish weight in this dataset. Because this model uses only one predictor, it avoids any issues in multicollinearity, while acting as the most parsimonious model of the data.

The final model showed that its single predictor, log(Width), explained 96% (R^2=0.96) of the variation in log(Weight). Although this was lower than the 98.7% explained (R^2=0.96) by the model containing log(Length2) and log(Height), this difference in R2 values was minor and was outweighed by the improvement in parsimony. Therefore, the final model had both strong predictive power and avoided multicollinearity.

Conclusion

Based on our statistical analysis of the Fish Market dataset, we found that the linear model was not a good fit for the data, as multiple assumptions were violated and multicollinearity existed. The strong correlation between the variables made it difficult to predict which predictor alone affected Weight the most in the linear model. However, after modeling refinement, we determined that the best model was the single predictor model with log(Width), since it explained most of the variation in log(Weight), avoided multicollinearity, and mostly satisfied the regression assumptions. These findings show that simpler models may outperform other models with predictors are highly correlated and measure the same underlying characteristics. For fish weight and biological growth modeling in general, future directions we could take would be to explore polynomial regression and other nonlinear models to capture better potential nonlinear relationships between morphometric measurements and weight.

References

Vipul L Rathod, Kaggle, https://www.kaggle.com/datasets/vipullrathod/fish-market, 2023.

Appendix

```
# Project 2
# By: Gabrielle Coronel, Kyle Zemel, Uyen Tran

library(readr)
library(corrplot)
library(car)

# Adding the fish dataset
fish <- read_csv("/Users/gabbiecoronel/Downloads/Fish.csv")
attach(fish)
summary(fish)

# Cleaning the fish dataset, removing species
cleanedFish <- fish[,-1]
summary(cleanedFish)

# Number of observations
numRow <- nrow(cleanedFish)
print(numRow)

# Multiple linear regression model for the fish market
model <- lm(Weight ~ Length1 + Length2 + Length3 + Height + Width, data = cleanedFish)
summary(model)

# Summary of fish Weight, Vertical Length in cm (Length1), Diagonal Length2 in cm, Cross Length
(Length3) in cm, and Height
summary(Weight)
summary(Length1)
summary(Length2)
summary(Length3)
summary(Height)
summary(Width)

# Histograms of our non-categorical variables: Weight, Vertical Length in cm (Length1), Diagonal Length2
in cm, Cross Length (Length3) in cm, and Height
hist(Weight, main = "Histogram of Weight in grams")
hist(Length1, main = "Histogram of Vertical Length in cm")
hist(Length2, main = "Histogram of Diagonal Length in cm")
hist(Length3, main = "Histogram of Cross Length in cm")
hist(Height, main = "Histogram of Height in cm")
hist(Width, main = "Histogram of Width in cm")

# Boxplots of our non-categorical variables: Weight, Vertical Length in cm (Length1), Diagonal Length2 in
cm, Cross Length (Length3) in cm, and Height
```

```
boxplot(Weight, horizontal = T, xlab = "Weight", main = "Boxplot of Weight in grams")
boxplot(Length1, horizontal = T, xlab = "Vertical Length in cm (Length 1)", main = "Boxplot of Vertical Length in cm (Length 1)")
boxplot(Length2, horizontal = T, xlab = "Diagonal Length in cm (Length 2)", main = "Boxplot of Diagonal Length in cm (Length 2)")
boxplot(Length3, horizontal = T, xlab = "Cross Length in cm (Length 3)", main = "Boxplot of Cross Length in cm (Length 3)")
boxplot(Height, horizontal = T, xlab = "Height", main = "Boxplot of Height")
boxplot(Width, horizontal = T, xlab = "Height", main = "Boxplot of Width")

# Scatterplots of our non-categorical variables: Weight, Vertical Length in cm (Length1), Diagonal Length2 in cm, Cross Length (Length3) in cm, and Height

plot(
  x = Length1,
  y = Weight,
  xlab="Vertical Length in cm (Length1)",
  ylab="Weight",
  main = "Scatterplot of Vertical Length vs. Weight"
)

plot(
  x = Length2,
  y = Weight,
  xlab="Diagonal Length in cm (Length2)",
  ylab="Weight",
  main = "Scatterplot of Diagonal Length vs. Weight"
)

plot(
  x = Length3,
  y = Weight,
  xlab="Cross Length in cm (Length3)",
  ylab="Weight",
  main = "Scatterplot of Cross Length vs. Weight"
)

plot(
  x = Height,
  y = Weight,
  xlab="Height",
  ylab="Weight",
  main = "Scatterplot of Height vs. Weight"
)

plot(
  x = Width,
  y = Weight,
```

```
  xlab="Width",
  ylab="Weight",
  main = "Scatterplot of Width vs. Weight"
)

# Residuals of the multiple linear regression model
par(mfrow = c(2, 2))
plot(model)
par(mfrow = c(1,  1))
plot(model)

# Correlation Matrix of the multiple linear regression model
numeric_data <- cleanedFish[, sapply(cleanedFish, is.numeric)]
correlation_matrix <- cor(numeric_data)
round(correlation_matrix, 3)

fish_filtered <- subset(fish, Weight > 0)

# Transformed log full model
full_model <- glm(log(Weight) ~ log(Length1) + log(Length2) + log(Length3) + log(Height) + log(Width),
data = fish_filtered)
summary(full_model)

# Correlation Matrix of the full log transformed model
all_log_vars <- data.frame(
  log_Weight = log(fish_filtered$Weight),
  log_Length1 = log(fish_filtered$Length1),
  log_Length2 = log(fish_filtered$Length2),
  log_Length3 = log(fish_filtered$Length3),
  log_Height = log(fish_filtered$Height),
  log_Width = log(fish_filtered$Width)
)
full_log_corr_matrix <- cor(all_log_vars)
round(full_log_corr_matrix, 3)

# Backwards Stepwise Selection Model for transformed log model
print("--- Starting Backward Stepwise Selection ---")
step_backward_model <- step(full_model, direction = "backward")
print("--- Final Model Selected by Backward Elimination ---")
summary(step_backward_model)

# New transformed log model based on the results of the Backwards Stepwise Selection Model
new_model <- glm(log(Weight) ~ log(Length2) + log(Height) + log(Width), data = fish_filtered)

# Residuals of the New Model
par(mfrow = c(1, 1))
plot(new_model)
```

```
# Getting the summary of the transformed log new model
new_model_log <- lm(log(Weight) ~ log(Length1) + log(Length2) + log(Length3) + log(Height) +
log(Width), data = fish_filtered)
summary(new_model_log)

# Transformed log of the predictors based on the new model
predictors_log <- data.frame(
  log_Weight = log(fish_filtered$Weight),
  log_Length2 = log(fish_filtered$Length2),
  log_Height = log(fish_filtered$Height),
  log_Width = log(fish_filtered$Width)
)

# Transformed new model log correlation matrix
new_correlation_matrix <- cor(predictors_log)
round(new_correlation_matrix, 3)

# Checking for multicollinearity using the VIF values of the mutliple linear regression model
vif_values <- vif(model)
print(vif_values)

# Checking for multicollinearity using the VIF values of the new transformed log model
vif_log <- vif(new_model_log)
print(vif_log)

vif_values_new <- vif(new_model)
print(vif_values_new)

# Based on the VIF values of the new model, creating a new model with only Length2 and Height as the
predictors
model_length2_height <- lm(log(Weight) ~ log(Length2) + log(Height), data = fish_filtered)

# Getting the summary and VIF values of the model with only Length2 and Height as the predictors
summary(model_length2_height)
vif_new <- vif(model_length2_height)
print(vif_new)

# Residuals of the model with only Length2 and Height
par(mfrow = c(2, 2))
plot(model_length2_height)
par(mfrow = c(1, 1))
plot(model_length2_height)

# simple linear regression for the log of Width
model_width <- lm(log(Weight) ~ log(Width), data = fish_filtered)

# Getting the summary and the residuals for the model with Width
summary(model_width)
```

```
par(mfrow = c(2, 2))
plot(model_width)
par(mfrow = c(1, 1))
plot(model_width)

# Univariate Models
# Length 1
lm_length1 <- lm(formula = Weight~Length1, data = cleanedFish)
summary(lm_length1)

# Length 2
lm_length2 <- lm(formula = Weight~Length2, data = cleanedFish)
summary(lm_length2)

# Length 3
lm_length3 <- lm(formula = Weight~Length3, data = cleanedFish)
summary(lm_length3)

# Width
lm_Width <- lm(formula = Weight~Width, data = cleanedFish)
summary(lm_Width)

# Height
lm_Height <- lm(formula = Weight~Height, data = cleanedFish)
summary(lm_Height)

# Testing for Confounding
# Reduced Models
glm_length1 <- glm(formula = Weight ~ Length1, data = cleanedFish)
summary(glm_length1)

glm_length2 <- glm(formula = Weight ~ Length2, data = cleanedFish)
summary(glm_length2)

glm_length3 <- glm(formula = Weight ~ Length3, data = cleanedFish)
summary(glm_length3)

glm_height <- glm(formula = Weight ~ Height, data = cleanedFish)
summary(glm_height)

glm_width <- glm(formula = Weight ~ Width, data = cleanedFish)
summary(glm_width)

# Full Model
glm_full <- glm(formula = Weight ~ Length1 + Length2 + Length3 + Height + Width, data = cleanedFish)
summary(glm_full)
```