1. Introduction

Background and Motivation

Fish play an essential role in feeding the world's growing population by providing nutrients to people's diets. Fish and fish products account for almost 20% of global animal-based protein consumption, and they are also critical sources of micronutrients necessary for human health, including iron, zinc, vitamins A and B12, and essential fatty acids (Environmental Defense Fund, 2018). Consequently, understanding the ecology of fish species is important to protect this valuable resource. The relationship between weight and length of a fish varies by species. Looking at this relationship can provide more insight to the underlying ecology of different fish species. As a result, the purpose of this study is to find the best fitting linear regression model to predict the weight of various fish species.

Data source and descriptions

The dataset was originally collected in 1917 and measured 159 fishes caught from Lake Laengelmavesi near Tampere in Finland.

There are 159 observations, 6 explanatory variables and 1 response variable. The response variable is Weight and the explanatory variables are Species, Length1, Length2, Length3, Height, Width. Species is a categorical variable with 7 levels; Length1, Length2, Length3, Height, Width are all numeric variables.

The detailed information of each variables from this dataset are listed below:

Variable	Description	Unit
Species	name of the 7 species : 'Bream' 'Parkki' 'Perch' 'Pike' 'Roach' 'Smelt' 'Whitefish'	/
Weight	Weight of the fish	cm
Length1	Length from the nose to the beginning of the tail	cm
Length2	Length from the nose to the notch of the tail	cm
Length3	Length from the nose to the end of the tail	cm
Height	Maximal height as percentage% of length from the nose to the end of the tail	%
Width	Maximal width as percentage% of length from the nose to the end of the tail	%

Data pre-processing

Firstly, the data was preprocessed to remove row 41, which had a weight of 0 despite having other measurements which is physically impossible. Afterwards, the data was split into training and testing sets with an 80/20 split, respectively.

In the original data set, there are 3 variables describing the length of the fish. They are the lengths from the nose to the beginning, the notch and the end of the tail. Collinearity diagnostics are used to check if these three variables share inherent collinearity.

Condition indices are larger than 30 suggests collinearity, which means only one variable should be kept. The following form shows variance inflation factors(VIF) and condition indices of length1, length2 and length3.

Variable	VIF	Condition Indices
Length1	2360.4	62.0
Length2	4307.9	342.8
Length3	2076.8	585.2

These numbers indicate that Length1, Length2 and Length3 are highly correlated variables, so two of these can be removed from the model. In our case, according to measurement used by the fishing industry, the length of a fish is measured from the tip of the snout to the posterior end of the last vertebra or to the posterior end of the mid-lateral portion of the hypural plate. Thus, length1 is kept in our model.

2. Analysis

Data visualization

After preprocessing steps, the relationship between each of the explanatory variables and the response variable, weight, was visualized.

Figure1: the scatterplot of fish Length1 and weight

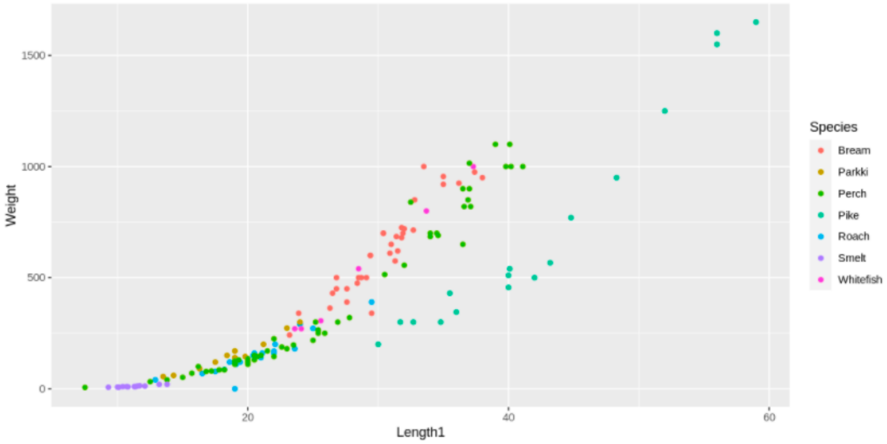


Figure2: the scatterplot of fish Length2 and weight

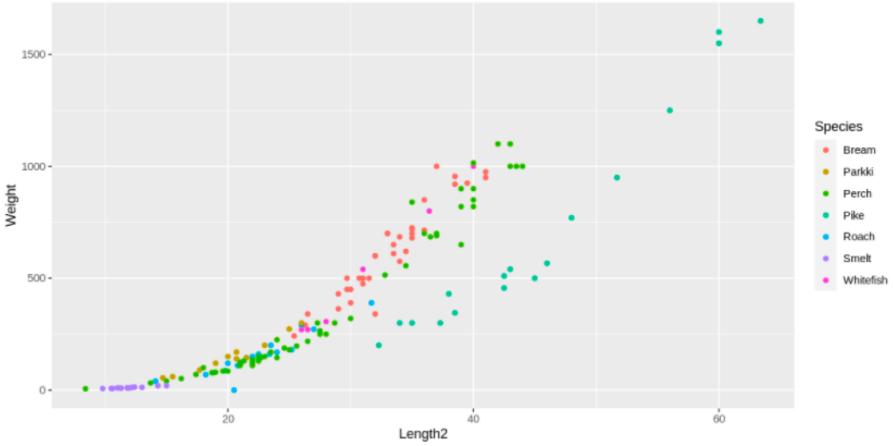


Figure3: the scatterplot of fish Length3 and weight

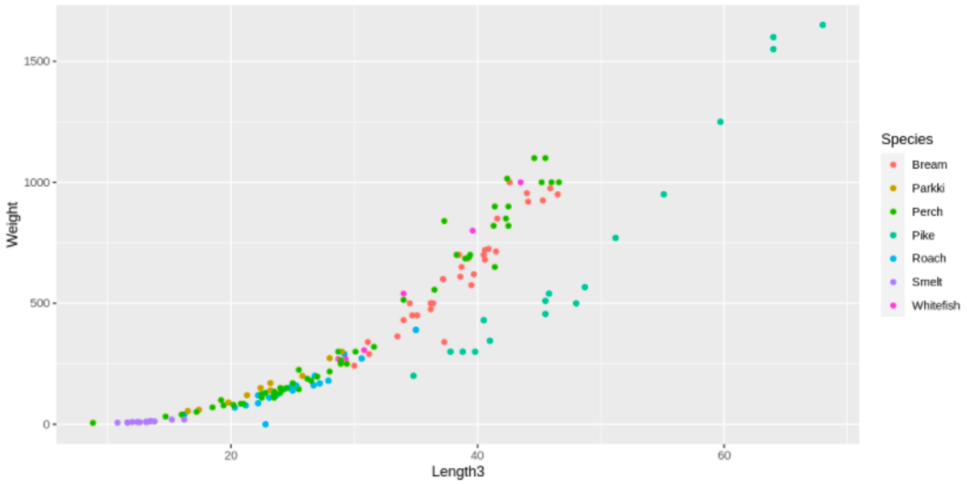


Figure4: the scatterplot of fish Height and weight

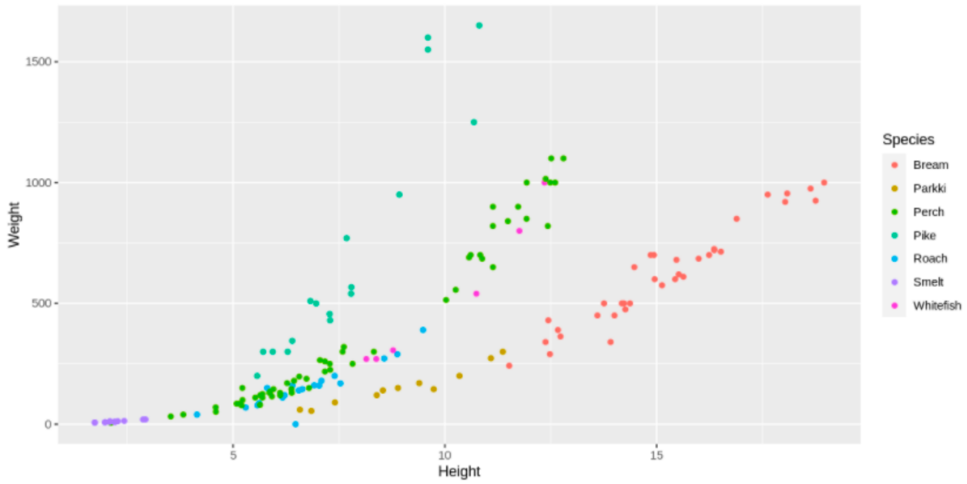
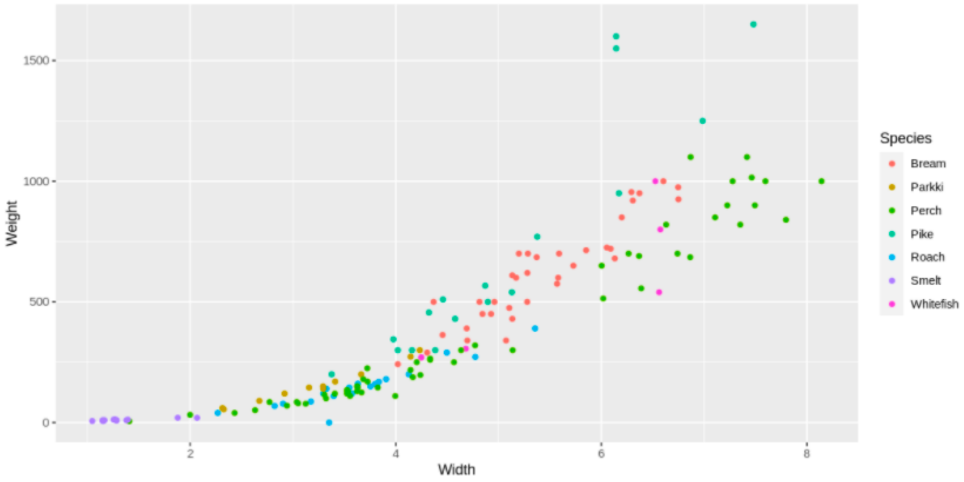


Figure5: the scatterplot of fish Width and weight

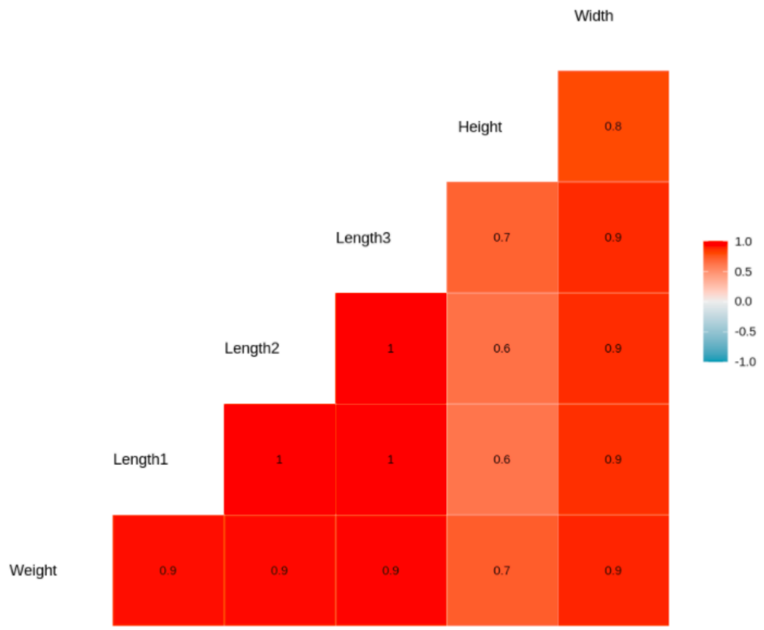


From Figure 1 to 5, there is a clear positive relationship between each explanatory variable with fish weight. However, the patterns of the plots indicate that some transformations are needed when fitting the model.

As indicated from Figure 1, 2, and 3, although different fish generally have different weight and length, most of them are distributed along a curved line, and this is very helpful for building a model around it. There is an exception though, for poke fish, the length is longer than other fish at the same weight. However, dummy variables can be added for different species to adjust for this. We can see from the plot for length 1, length 2, and length3 that the data points are on the right side of the other data points for the rest of the species. All lengths for other species seem to be distributed along the same curve, and this is true for all length 1, length 2, and length 3.

Figure 4 indicates that for the same Weight, Bream fish have the largest average Width than any other species. It is also clear that smek fish have the smallest average Width among all other species.
Plot 5 indicates that for the same Height, Perch fish have the largest Width than any other species. It is also clear that still smek fish have a average smallest With among all other species

Figure6: correlation between each variable



To be more clear, the correlation Figure (Figure 6) is generated to see if there is any correlation from each of the variables.

3.Discussion

<u>Baseline Model 0</u>

First a simple model with no transformations was fit to the training data for baseline comparison. The explanatory variables Species, Width, Height, and Length1 are included in the model.

The model is:

$$Y_{Weight} = \beta_0 + \beta_{Length1}x_{Length1} + \beta_{Parkki}x_{Parkki} + \beta_{Perch}x_{Perch} + \beta_{Pike}x_{Pike} + \beta_{Rpach}x_{Roach} + \beta_{Smelt}x_{Smelt} + \beta_{Whitefish}x_{Whitefish} + \beta_{Height}x_{Height} + \beta_{Width}x_{Width}$$

Fitting the model to the training data in R:

$$Y_{Weight} = -766.99 + 40.56x_{Length1} + 69.55x_{Parkki} + 7.41x_{Perch} - 316.52x_{Pike} + 12.85x_{Roach} + 296.38x_{Smelt} + 1.68x_{Whitefish} + 0.10x_{Height} + 10.74x_{Width}$$

The fitted model has an R^2 value of 0.9394, which is very high as it is close to 1. However, despite the high R^2 value, this model has a few issues with it. Firstly, a Shapiro-Wilk Normality Test was performed on the residuals in R to test the following hypotheses:

H_0: The residuals follow a normal distribution
H_a: The residuals do not follow a normal distribution

This gives a test statistic of W = 0.95954, and a p-value of 0.0007389. Thus, the null hypothesis can be rejected at significance level 0.05 in favour of the alternative hypothesis that the residuals are not normally distributed. This violates an important assumption of the model that the errors are normally distributed, which implies the model does not accurately explain the data.

Secondly, a residual versus fitted values plot was generated. Figure 7 shows a pattern, implying that the variance of error is not constant.

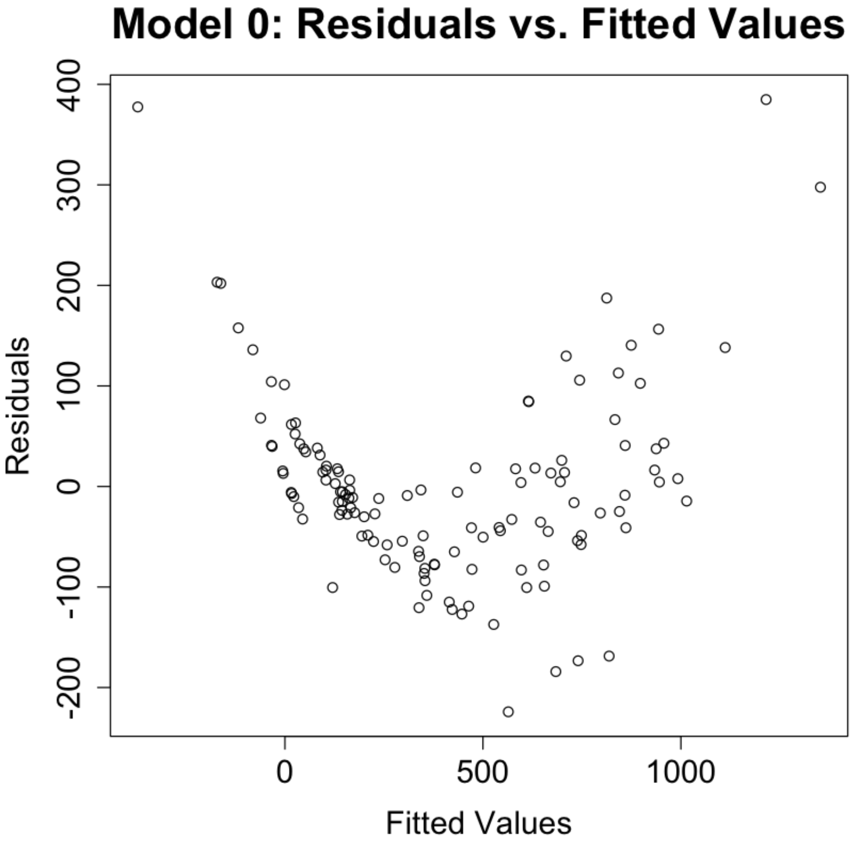


Figure 7

Lastly, the following hypothesis test was conducted for each coefficient estimated in the model.

$$H_0 : \beta_i = 0$$
$$H_a : \beta_i \neq 0 \ ,$$

where *i = (Width, Length1, Species).*

The null hypothesis is rejected for the explanatory variables Length1, SpeciesPike and SpeciesSmelt, as they have a p-value less than the significance level alpha = 0.05, in favour of the alternative hypothesis that the slope coefficients do not equal 0. Conversely, since the p-values for the other explanatory variables are greater than the significance level, alpha = 0.05, the null hypothesis is not rejected for these cases. This means that the only significant coefficients are Length1, SpeciesPike, and SpeciesSmelt. Hence, the model can be significantly reduced.

The baseline model was also used to predict the response variable of the test set. The normalized root mean squared error was calculated to compare prediction error. The RMSE was normalized in order to be able to compare it to future models that include transformations. The normalized RMSE value for the baseline model is 0.1118.

Next, model 1 was fitted to the training data.

In the fishing industry, the weight of the fish is estimated by the following formula.

$$\mathrm{W} = \frac{\text{length} \times \text{girth} \times \text{girth}}{\alpha}$$

W is the estimated weight of the fish and is a constant related to the species of the fish. This method has been proposed a long time ago. Herbert Spencer in his Principles of Biology of 1864–1867 (here cited from the 1966 reprint of the 1898 edition) restated the first part of Galileo's law as follows: ‘In similarly-shaped bodies the masses, and therefore the weights, vary as the cubes of the dimensions.’ This subsequently became known as the ‘cube law.’ Accordingly, a fish which doubles its length increases by eight times in weight.

The first model is roughly based on the aforesaid empirical formula. The variable length1 is raised to the power 3 to be an estimate of the factor. Using this model, the constant will be estimated for all kinds of fish as following:

$$\mathrm{W} = \beta_{Bream}L^3 + \beta_{Parki}L^3x_{Parki} + \beta_{Perch}L^3x_{Perch} + \beta_{Pike}L^3x_{Pike} + \beta_{Roach}L^3x_{Roach} + \beta_{Smelt}L^3x_{Smelt} + \beta_{Whitefish}L^3x_{Whitefish}$$

Fitting the model to the training data in R:

$$Y_{Weight} = 6.66 \times 10^{-2} + 2.14 \times 10^{-2}L^3x_{Length1}{}^3 + 2.18 \times 10^{-2}L^3x_{Parkki} + 1.66 \times 10^{-2}L^3x_{Perch} + 8.42 \times 10^{-3}L^3x_{Pike} + 1.63 \times 10^{-2}L^3x_{Roach} + 7.68 \times 10^{-3}L^3x_{Smelt} + 2.10 \times 10^{-2}L^3x_{Whitefish}$$

This model gives an R^2 value of 0.9759, which can be interpreted as 97.59% of the variation in weight can be explained by the explanatory variables in the model. Cleary, this model fits the data better than model 0 since it gives a higher R^2 value.

Next, the following hypothesis test was conducted for each coefficient estimated in the model.

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0 \ ,$$

where *i* = (Species).

The above hypothesis test to see whether the slope coefficients are statistically different from 0 finds that all coefficients are significant at alpha =0.001, except for the interaction between the smelt species and the transformed Length1. Thus, the null hypothesis is rejected for every estimated coefficient except for smelt species, in favour of the alternative hypothesis.

In addition, this model was used to predict the response variable values of the test set. The normalized root mean squared error was calculated to compare prediction error. The RMSE value is 0.0630 which is lower than the baseline model, meaning that it is more accurate at predicting response values.

One major problem with this model is that it breaks an important assumption that the residuals are normally distributed. A Shapiro-Wilk Normality Test was performed on the residuals in R to test the following hypotheses:

$$H_0\text{: The residuals follow a normal distribution}$$
$$H_a\text{: The residuals do not follow a normal distribution}$$

This gives a test statistic of W=0.86022, and p-value = 1.216e-09. Thus, the null hypothesis can be rejected at significance level 0.05 in favour of the alternative hypothesis that the residuals are not normally distributed. This implies the results here may be misleading, which is problematic.

Another problem with this model is that it also breaks the assumption of constant variance. The plot of fitted values versus residuals, figure 8, shows a funnel pattern, implying non constant variance.

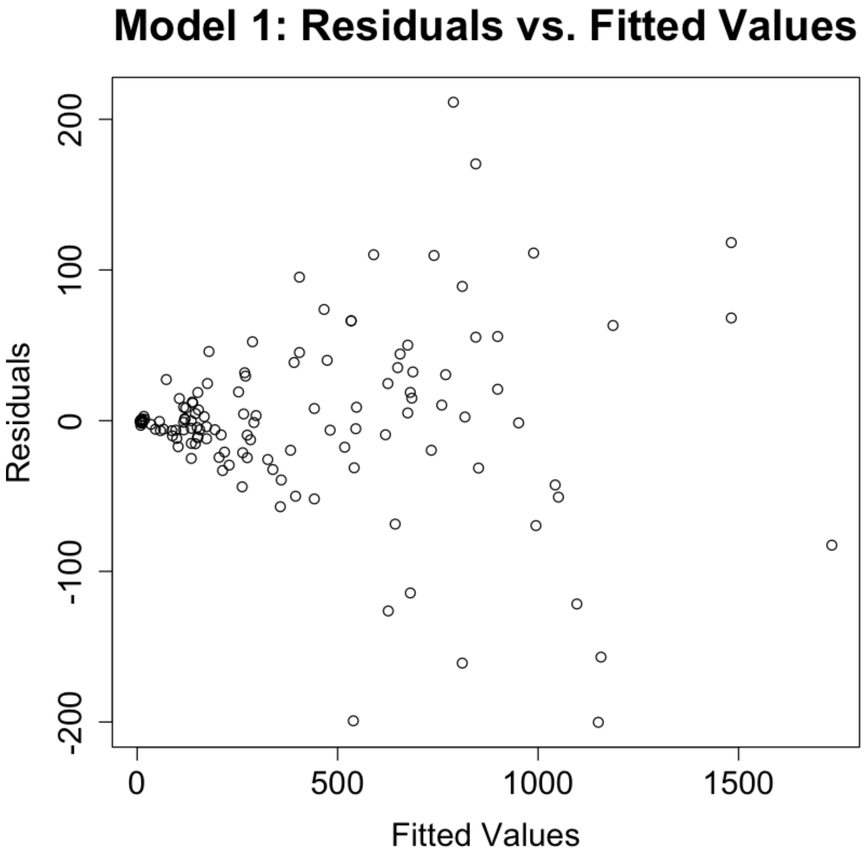


Figure 8

Model2:

Lastly, model 2 was fitted to the training data.

The aforementioned cube-law has proven to be incorrect by multiple researches. Instead of raising the length to the predetermined power 3, scientists suggest that the power should be estimated. Fulton(1904) has found that the weight of the fish increases slightly faster than the cubic term of length. In a 1928 article, Keys specifically pointed out the cube-law is an incorrect formulation of the weight–length relation and proposed the following weight-length relationship,

$$\mathrm{W} = \alpha \mathrm{L}^{\beta},$$

where W is the estimated weight, L is the length of the fish, and are parameters.

Clark(1928) also found this relationship in its logarithmic form, which has been deemed as the modern WLR.

$$\log W = \log \alpha + \beta \log L$$

Model 2 is an advanced interpretation of this relationship. Considering species play an important role in estimating the weight of fish, dummy variables regarding species are added to the logarithmic estimation.

The model 2 is defined as the following:

$$\mathrm{W} = \alpha L^{\beta}\, e^{\beta_{Bream} + \beta_{Parki} x_{Parki} + \beta_{Perch} x_{Perch} + \beta_{Pike} x_{Pike} + \beta_{Roach} x_{Roach} + \beta_{Smelt} x_{Smelt} + \beta_{Whitefish} x_{Whitefish}}$$

Fitting the model to the training data in R:

$$Y_{Weight} = 0.098{x_{Length1}}^{2.34}e^{-0.025x_{Parkki}-0.32x_{Perch}-0.73x_{Pike}-0.29x_{Roach}-1.07x_{Smelt}-1.14x_{Whitefish}+0.14x_{Weight}}$$

This model has significant benefits over model 1 and model 0. Firstly, the R^2 value is 0.9955, which is higher than the previous models.

Next, the following hypothesis test was conducted for each coefficient estimated in the model.

$$H_0: \beta_i = 0$$
$$H_a: \beta_i \neq 0 \ ,$$

where i = (Species).

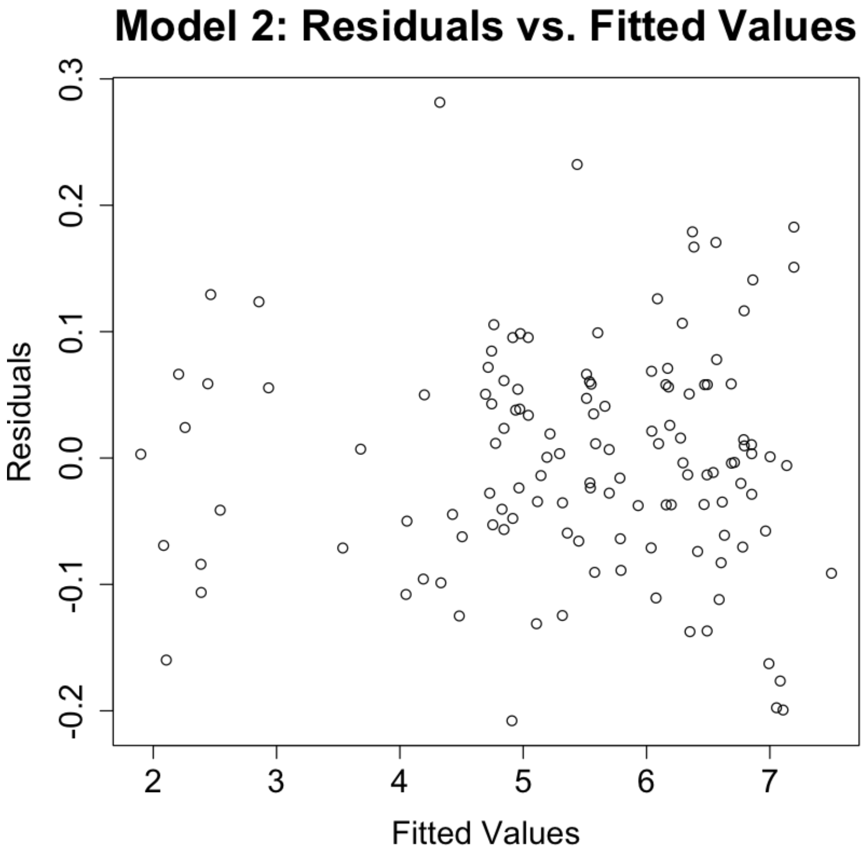
This finds that all the slope coefficients are statistically significant except for Parkki Species, meaning they are important to the model. Hence, the null hypothesis is rejected for all coefficients except the Parkki Species, in favour of the alternative hypothesis.

Moreover, a Shapiro-Wilk Normality Test was performed on the residuals in R to test the following hypotheses:

$$H_0\text{: The residuals follow a normal distribution}$$
$$H_a\text{: The residuals do not follow a normal distribution}$$

This gives a test statistic of W =0.9924, and a p-value of 0.7209. Hence, the null hypothesis is not rejected since the p-value was greater than the significance level alpha = 0.05, and it can be concluded that the residuals follow a normal distribution. This is a key assumption of the linear regression model, so this is a major benefit over the previous models, which both violated this assumption.

Next, a plot of the fitted values versus residuals was made, figure 9. Figure 9 shows no pattern, implying that there is constant variance. Since constant variance is an assumption of the model, this is another reason why this model fits the data better than the previous models.



Lastly, the model was used to predict the response values of the test set. The normalized RMSE was again calculated to assess prediction error. The RMSE value is 0.0309, which is lower than the previous models, meaning there is less prediction error.

Overall, it is clear that model 2 is the best model based on a higher R^2, lower prediction error, and the fact that it meets the normality and constant variance assumptions of linear regression models.

4.Conclusion

In this project, we tried to fit the data to different models, including two linear models and a log model. The best model we found to predict a fish's weight is

$$W = \alpha L^{\beta} e^{\beta_{Bream}+\beta_{Parki}x_{Parki}+\beta_{Perch}x_{Perch}+\beta_{Pike}x_{Pike}+\beta_{Roach}x_{Roach}+\beta_{Smelt}x_{Smelt}+\beta_{Whitefish}x_{Whitefish}}$$

This model indicates that the fish's weight has a dependence on fish's length and kind of species. Alphas and the Beta in the model are constants, and we fit the data to the model to find the optimal values for them.

In the original dataset, only physical features are included in the model. However, in order to build a comprehensive and scientific model, it is meaningful for future research to consider adding external factors like fish's growth rate in week, fish feed in kilograms per week into the model to predict the weight of fish.

Reference

Brofeldt, Pekka: Bidrag till kaennedom on fiskbestondet i vaera sjoear. Laengelmaevesi T.H.Jaervi: Finlands Fiskeriet Band 4, Meddelanden utgivna av fiskerifoereningen i Finland. Helsingfors 1917

Environmental defense fund,"Why should you care about fisheries? They can help feed the world", 2018,

Keys, A. B., 1928: The weight-length relationship in fishes. Proceedings of the National Academy of Science, Vol. XIV, no. 12, Washington, DC, pp. 922–925.

Clark, F. N., 1928: The weight–length relationship of the California sardine (Sardina caerulea) at San Pedro. Division of Fish and Game, Fish Bull. No. 12. 59 pp.

Fulton, T. W., 1904: The rate of growth of fishes. Twenty-second Annual Report, Part III. Fisheries Board of Scotland, Edinburgh, pp. 141–241.

Spencer, H., 1864-1867: The Principles of Biology, 2 volumes. Williams & Norgate, London. 678, pp.https://blogs.edf.org/edfish/2018/05/15/why-should-you-care-about-fisheries-they-can-help-feed-the-world/

Glossary Search for Standard length, 1991
https://www.fishbase.se/Glossary/Glossary.php?q=standard+length&sc=is