**Investigate What has the Greatest Impact on the Weight of Fish**

**Dataset: FishMarket**

**Group number: 12**

Group member: Shutong Yang B00780702

Bijun Zhang B00780704

Yifan Zhang B00783528

**Investigate what has the greatest impact on the weight of fish**

**Abstract**

* This study mainly studied the different girths of fish in the fish market to understand what has the greatest impact on the weight of the fish. We studied 159 samples of 7 species of fish. Among them, Pike has the heaviest weight, about 1650kg, and Roach has the lightest weight, about 0kg.
* Species: Species name of fish
* Weight: Weight of fish in gram
* Length1: Vertical length in cm
* Length2: Diagonal length in cm
* Length3: Cross length in cm
* Height: Height in cm
* Width: Diagonal width in cm

**Introduction**

Fish are the ancestors of all modern vertebrates, including humans and all of our pets and domestic animals. Therefore, their long evolutionary history has always been an issue of interest to many biologists. Similarly, the oceans occupy most of our planet, and fish are the inhabitants there. For the research of fish bionics, it can help solve problems in the ocean, so, the information of fish is more important for the design of scientists. On the other hand, fish play an important role in economic development in both aquaculture and ornamental fish production because of their rich protein and nutritional value. Our study is not only conducive to scientists' research on biology, it also can see the economic effect of fish from this study.

Through our research direction, we found such a report, Adeboyejo O. Akintade, Clarke O. Edwin and Ekele A. Simon (2016) published a report in the International Journal of Marine Science, entitled “Length-Weight Relationship, Condition Factor and Sex-ratio of Fish Fauna in Badagry Creek, Lagos, Nigeria”. By reading this report, we are surer how to use our data to help scientists：Study the weight of fish. Use R studio software to make a prediction and judgment on the influence of different variables on the weight of fish. Our data includes Perch, Bream, Roach, Pike, Smelt, Parkki, Whitefish and their respective weights, length 1, length 2, length 3, height and width. Moreover, we set Weight to be dependent, and other variables are independent.

**Data Description**

Additional data point：We added the 160th row as additional data

1. Matrix scatter plot

Preliminary judgment, we can see the approximate relationship between the variables. There is a linear relationship between length1 and length2

1. Correlation coefficient matrix

Since |corr|=1, it proves that there is a correlation between the two variables.

**Method**

As can be seen from the drawing by RStudio, there is a significant linear relationship between them, so it is not necessary to add the denaturing independent variables, but interaction is required.

1. Model selection(setting model)
2. Get all possible models

According to Package ‘olsrr’ (2020.02.10), we get the rcode (install package “olsrr”). Then we get all possible models when y is the weight.

Choosing the best data in R^2 is based on the law that the number of independent variables rises, and R^2 must rise. Because the values of 42, 57, and 63 are too close to judge easily, we have reservations about choosing the best data in this model and can't make a final conclusion. Because the smaller the Cp, the better and we find that the values of 42, 57 and 63 are similar, so we have reservations about the data too. The model R^2 adj is mainly intended to explain how many Y's there are. Since the number of independent variables has no effect on it, its data should be as large as possible. We selected the best data from the above four models and screened it again. Finally, we found that 63 appeared most frequently, so 63 models were the best among all our data.

1. Best subset regression(olsrr)

The better models (1, 7, 22, 42, 57, and 63) that have been selected from all models are further analyzed and divided into several groups based on the number of independent variables in the model, and each the best model in the group will be selected for further screening.

The abscissa is expressed as the number of independent variables, and the ordinate is the value of each indicator corresponding to each model. By connecting the points of the four models into a straight line, it can be seen that the number of two independent variables in the model is the most appropriate.

1. Backward

We set =0.1. Assuming that all x are in the model, we compare the size of p-value and, and then one by one delete all independent variables corresponding to p-values greater than. Until the p-value of no independent variable in the model is greater than, stop backward to get the model.

1. Modified model

A good model needs to meet the Guss-Market Assumption, E(ei)=0，Var(ei)=, and the error is randomly generated. By observing our Residual vs Fitted graph, it can be seen that the small circles are uniformly distributed on the graph, which means that the error is generated randomly. Also the E (ei) line in the middle approaches zero. This proves that our model is normal and meets the assumptions.

It can be seen from the Normal QQ graph that our model basically obeys the normal distribution and does not need to be corrected.

The Scale-Location graph is used to check whether the residual variances are equal, as shown in the figure, it appears as equal variances.

From the graph of Residuals vs Leverage, it can be seen that the distribution of each small circle is relatively uniform, and the red line representing E (x) is also approaching to 0. Therefore, it can be seen from these aspects that it is a relatively good graph and does not need to be corrected.

1. VIF: Collinearity

We need to use VIF to check whether there is a problem of collinearity. We finally correct our model through several steps in Backward, which can effectively avoid the problem of multicollinearity. If VIF's inspection result finds a value greater than 10, we need to look back and check whether there is a problem on Backward. Fortunately, we didn't find anything larger than 10 in the end, so there is no multicollinearity problem in our graph, and this will be our final model.

**Result**

This part mainly explains the model, analyzing R-square, Adj-R-square, p-value, and the p-value of F-test respectively. Also explain which variable has the strongest correlation with y, which variable explains y the most, and which variable has the greatest influence on y.

According to our final model, the P-value of Species Parkki is 0.0202 less than 0.1, which means its X is significant. Meanwhile, the P-value of Species Perch is 1.69, which is less than 0.1, indicating that its X is significant. Therefore, when the P-value of Species Pike is less than 2 and conforms to the feature of less than 0.1, its X is also significant. When Species Roach's P-value is 1.74 less than 0.1, we are not surprised that its X is significant. The p-value of the final Species Smelt is 0.0933 and is less than 0.1, which means that its X is significant. However, we have also found that Species White fish is insignificant, as its P-value is 0.3375 and greater than 0.1. It can be seen from the model that our R-squared is 0.9634, which shows that R-square can explain 96.34% of y. Through the backward part, we get Length1andLength2, which are the most relevant to our Y (fish's weight), also length1 and length2 explain the most Y, which means that vertical length and diagonal length have the greatest impact on weight of the two variables.

**Conclusion**

Of the 160 samples tested, there are 35 Bream fish (21.875%), 20 Roach fish (12.5%), 6 Whitefish (3.75%), 11 Parkki fish (6.875%), 56 perky fish (35%), 17 pike fish (10.625%), and 15 Smelt fish (9.375%) were all perked. The pike fish is the heaviest, weighing about 1650kg, while Roach fish is the lightest, weighing about 0kg. The range of Length1 is between 59~7.5, pike fish is the longest and Perch fish is the shortest; Length2 is in the range of 63.4~8.4, Pike fish the longest and Perch fish is the shortest.

Regression models have shown that the P-value of all Species Bream, Parkki, Perch, Roach, and Smelt, except for Species Whitefish (P-value=0.3375), is less than 0.1, suggesting that the length (cm) of a Species is not proportional to its weight (kg). However, the regression value of most Species Pike is less than 0.1, especially the P-value of Species Pike is 1.69E-08 and is the smallest. The P value of Species Smelt is the highest of any fish less than 0.1. The regression models length1 and length2 prove that body weight increases as length increases, and for fish whose p value is less than 0.1. Those indicate the length1 and length2 have a significant effect on the weight of these fish.

In some sea areas, fishing is strictly prohibited for a certain period of time, but there will always be people who ignore this ban. So when people's supervision is lax or omission, this will give people who illegally fish have a chance. Therefore, we can use modern technology to make a bionic fish with positioning similar to fish, and put them in the sea where fishing is strictly prohibited. When they are caught and placed in the fishing boat, we can use positioning to effectively arrest illegal personnel. This helps us to supervise (Red Dot Award: Product Design, 2020). Only when our bionic fish is indistinguishable from the real fish, will it prevent illegal personnel from detecting and formulating countermeasures in advance. Our data can help Bionics Scientist who produce this area to provide information on the length and weight of the fish, making the bionic fish more realistic.

For economic development, by comparing the various information data of 7 kinds of fish and investigating which fish has the highest sales in the fish market. A classification system can be developed for precise fishing or artificial breeding of specific fish groups, which can effectively facilitate the classification and marketing of fishermen and the fish market. We can compare the weight and length of seven kinds of fish (Perch, Bream, Roach, Pike, Smelt, Parkki, and Whitefish), build an information database and simulate a virtual model, and then accurately capture the fish through the virtual model and information database. Promote the development of the sales market.

In terms of evolution, we can compare the data obtained now with the data of fish many years ago (the weight and length of the species, etc.) to see whether the two sets of data have relatively large changes, such as size and color, so as to know whether the fish has evolved. The so-called database is a data set established by comparing the basic information such as the weight and length of 7 species of fish, which can be used for comparison and analysis with previous data. This data set can be helpful for some biological researches.

**Reference**

Akintade, A. O., Edwin, C. O., & Simon, E. A. (2016). Length-weight relationship, condition factor and sex-ratio of fish fauna in Badagry Creek, Lagos, Nigeria. *International Journal of Marine Science*, *6*.

Dot, R. (2020, December 04). Bionic Positioning Fish. Retrieved December 10, 2020, from <https://www.red-dot.org/zh/project/bionic-positioning-fish-48374>

Hebbali, A., & Hebbali, M. A. (2017). Package ‘olsrr’.