

Multivariate Linear Regression on Fish Market Data

COGS 109 Final Project Proposal

Team Members

Jaidyn Patricio - A15110718

Ata Tafazoli Yazdi - A12326075

Andrew Hernandez - A14632498

Susanna Anil - A15571255

Background Information

Analysis from this project concerns itself with the nature of fish anatomy from species commonly seen in markets. Our interest in this topic is largely related to the effects of anthropogenic climate change as it has impacted several species of fish and their biomasses. These changes, in part, are stimulated by climate change increases in ocean temperature and levels directly affect habitats for all sea life. Particularly within the last hundred years, ocean temperature has consistently risen 0.13°F every decade due to greenhouse gas emissions¹ — forcing several species to migrate to less shallow waters. Due to the dispersion of fish species, invasive species are close to extinct and species that were once heavily populated are seeing a decrease in biomass and reproduction².

Although we perform fish weight analysis on species common to markets, this topic remains relevant as observable and predictable changes in fish may provide insight on those that are being impacted. For this project, the scope of our data analysis involves fish weight prediction using various anatomical measurements as predictors. To accomplish this goal, we have chosen a fish market dataset recording these anatomical measurements. As such, a real-life application of this dataset could track those changes for common fish species caught and sold at fish markets.

Research Question

We are looking to utilize multivariate linear regression to explore how the physical characteristics of any particular fish (such as its height and width) combined with differing measurements of length (as they are recorded vertically, diagonally, or as a cross-section) influence its weight. As this is the case, our project explores whether a vertical, diagonal, or cross measurement is most useful in producing a model based on height and width to predict fish weight. From our analysis, we attempt to answer the following question:

How can we make efficient use of the existing labeled data to build a robust model for predicting fish weight?

For this dataset, we consider the three length measurements (vertical, diagonal, or cross) and both height and width as predictors to include within our model. We also take into account an interaction term between height, width, and each of the three length measurements. Thus using the specific variables included within this dataset to answer our question above, the overall scope of our analysis translates to the following:

Which length measurement (vertical, diagonal, or cross), combined with height and width, creates a model that best predicts fish weight?

Data

In order to answer our research question, we use a fish market dataset containing a record of seven species of fish that are common for market sale totaling 159 samples. This dataset is primarily composed of continuous float measurements which track a fish's weight, vertical length, diagonal

¹ <https://theconversation.com/how-is-climate-change-affecting-fishes-there-are-clues-inside-their-ears-110249>

² <https://climefish.eu/climate-change-and-impacts-on-fisheries/>

length, and cross length measurements, as well as its height and width. Measurements are recorded in centimeters and weight is recorded in grams. For each sample, physical and anatomical attributes are associated with a particular species of fish.

1. Common Fish Species by Market

- **Samples:** 159
- **Variables:** *species, weight, length1, length2, length3, height, width*
- **Labels:** weight
- **Link:** <https://www.kaggle.com/aungpyaeap/fish-market>

Although *species* — which includes 'Bream', 'Roach', 'Whitefish', 'Parkki', 'Perch', 'Pike', and 'Smelt' fish types — could be considered a viable label to perform analysis on this dataset, the relevant label for our multivariate analysis is *weight* as linear regression is useful for predicting continuous outputs rather than categorical variables. While the original dataset is labeled by species of fish, we are tracking the change in weight due to environmental factors and thus have decided to label the dataset by weight.

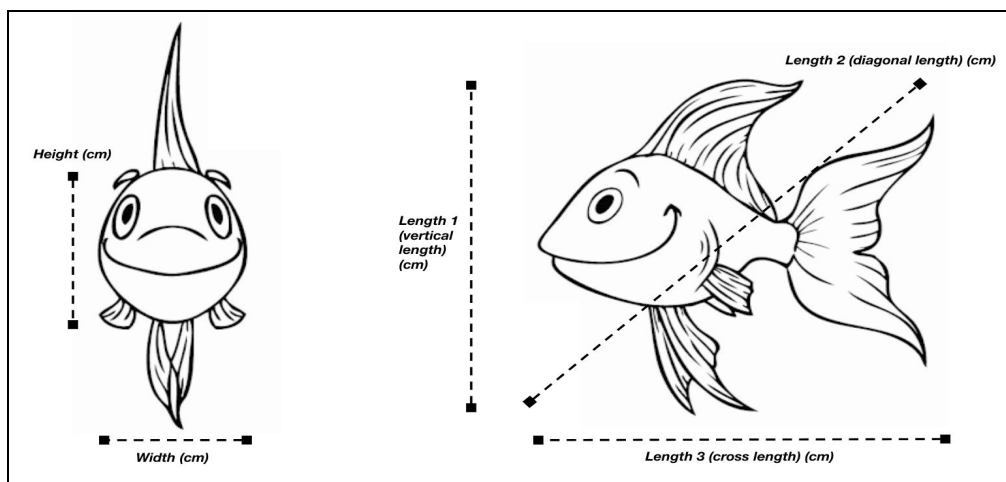


Figure 1. A diagram which depicts *height*, *weight*, *length1* (vertical), *length2* (diagonal), and *length3* (cross) measurements as recorded in the dataset. From the diagram above, we note that the difference between height and vertical length is based on the difference between the measurement of fish fins in comparison to the main fish body.

How is this analysis different from others on Kaggle?

Our analysis concerns itself with how the *vertical*, *diagonal*, and *cross* measurements affect the overall accuracy of the model when each is combined with an interaction term, *height*, and *weight* as predictors. In this case, we decide to include the features *height* and *weight* in every model because this provides us with a baseline to compare the effectiveness of the remaining variables that our research question is particularly interested in. Thus, the scope of our analysis is unique as we have centered the design of each model around our research question to specifically allow for the evaluation of which measurement is most useful in comparison to its remaining counterparts.

While our analysis and many others found on Kaggle similarly use the Common Species by Fish Market dataset to attempt to predict fish weight, the four models of our analysis were constructed by our team with the certain goal in mind to evaluate the difference between the *vertical*, *diagonal*, and *cross* measurements. This goal differentiates our analysis from others as multivariate linear regression has not yet been used to explore this particular comparison chosen by our team. As this is the case, each model and its variables included are distinctive from other analyses and, as such, the results of our research have the potential to provide new insights on the relative importance of *vertical*, *diagonal*, *cross*, and *species* measurements in the prediction of fish weight.

Methods

Data Cleaning

To address our research question, we begin our analysis by cleaning our dataset. Data cleaning is a crucial first step as it gives us the opportunity to detect any potential bias existing within the data from missing or duplicate values, as well as provides us with the ability to reformat our dataset into a structure that works well for our specific analysis. In this step, we rename ambiguous variables — *length1*, *length2*, and *length3* — to the length measurements ‘vertical’, ‘diagonal’, and ‘cross’ as a means to ensure clarity throughout our analysis. We then check for and drop missing data and duplicate sample entries to ensure that all samples have measurement data for every column and appear only once throughout the entire dataset.

Exploratory Data Visualization

Our next step concerns itself with exploratory data visualization. With this step, we generate histograms, barplots, and heatmaps to better understand the range of our data. The information gained from this process is fundamental because it gives us context which we can use to construct appropriate models. Histograms of each measurement variable allow us to understand how the data is distributed, and a barplot of the *species* variable gives us an overall picture of how the dataset is spread by species. Generating a correlation heatmap and scatterplots between all measurements is useful as we are interested in seeing if any measurements are strongly correlated with each other. To gain a better understanding of strongly correlated variables, we would have a total of 36 scatterplots as there are six measurement variables within this dataset.

Data Analysis

After we explore and visualize the data, we then construct models to address our research question. In total, our analysis consists of 4 multivariate linear regression models which allow us to explore how the length measurements (*vertical*, *diagonal*, and *cross*), paired with height and width, impact weight. With this being the case, 3 of our models each include one length measurement while our fourth model does not include any length measurement to serve as a baseline to compare our analyses. For all 4 models, we similarly include an interaction term ($w_4 * height * width * length$) to account for variables impacting each other. As such, our models are listed as follows:

Model 1: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * vertical_length) + (w_4 * height * width * vertical_length)$

Model 2: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * diagonal_length) + (w_4 * height * width * diagonal_length)$

Model 3: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * cross_length) + (w_4 * height * width * cross_length)$

Model 4: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * height * width)$

For each model w_i corresponds to a weight multiplied by each predictor variable. To perform our analysis, we split our data into an 80% training set and 20% testing set. Using our training data, we solve for each w_i by putting our data into an augmented array of the form $y = A * w$, where y represents the output of the data we train each model on, A represents each value for *height*, *width*, the relevant *length* measurement, and interaction term per model, and w represents a vectors that consists of the weights we solve for during training.

Reporting Data Analysis

Once we have solved for w_i , we report the model by replacing each w with its calculated value. We then calculate the SSE and MSE to assess our models. From this cross validation, we pay particular attention to SSE_{test} and MSE_{test} scores to evaluate each model. To compare our results, we generate a bar graph of each model's MSE_{test} score.

Summary Data Analysis Pipeline

Below we have outlined a pipeline of the steps necessary to perform our analysis. Sections include details pertaining to data cleaning, data visualization in terms of both exploratory data analysis and reporting data analysis, as well as the analysis itself (multivariate linear regression).

1. Data Cleaning

- a. Rename variables *length1*, *length2*, and *length3* to better represent the respective vertical, diagonal, and cross measurements of the dataset.
- b. Check for missing data as it has the potential to produce bias within the model. If missing data exists for any attribute within a sample, drop the observation completely.
- c. Check for duplicate sample entries, which can also lead to a biased model. If duplicate sample entries exist, keep the first observation and drop its duplicates.

2. Data Visualization

- a. **Exploratory Data Analysis** (to be performed before creating each model)
 - i. Create histograms to view the distributions of each attribute of interest within the dataset (*weight*, *length1*, *length2*, *length3*, *height*, *width*). This will allow us to gain a better understanding of the data we are working with, as well as help us to identify skewed distributions.
 - ii. Create a barplot comparing the total amount of each species included within the dataset. This graph will be useful in visualizing an overall picture of how the dataset is distributed by species.
 - iii. Create a heatmap to present correlation between each attribute. In this case, we are particularly interested in seeing if any of the three length measurements (vertical, diagonal or cross) are strongly correlated with either height or width, as a strong correlation between these variables could affect the model in light of our research question.
 - iv. Create a scatterplot matrix to visualize the pairwise relationships between each variable within the dataset. For this dataset, we will have a total of 36 subplots that give us the ability to understand and visualize how each variable is correlated in scatterplot form.

- b. **Reporting Data Analysis** (to be performed after linear regression, allowing us to visualize the results of each model)
 - i. Use a bar plot to compare the SSE/MSE for each of the 4 models. For each graph, we will use the test scores to evaluate the performance of each model.
 - ii. Compare the SSE/MSE scores of each model within a 1x4 subplot figure from the bar plots generated above.

3. Data Analysis

- a. Build 4 multivariate linear regression models to predict fish weight using height, width, and either the vertical, diagonal, or cross measurements as predictors.
 - o Model 1: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * vertical_length) + (w_4 * height * width * vertical_length)$
 - o Model 2: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * diagonal_length) + (w_4 * height * width * diagonal_length)$
 - o Model 3: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * cross_length) + (w_4 * height * width * cross_length)$
 - o Model 4: $weight = w_0 + (w_1 * height) + (w_2 * width) + (w_3 * height * width)$
- b. Split data into an 80% training set and 20% testing set.
- c. Perform linear regression and solve for each w using the training data. This will be repeated for each of the four models.
- d. For each model, calculate SSE/MSE for training and test data.

4. Results

- a. Report each model alongside its testing SSE/MSE (See Data Visualization, Reporting Data Analysis (Section 2b).
- b. Compare each model and identify the best one. For this analysis, we plan to pay close attention to the testing SSE/MSE scores for each model and any overfitting that may have occurred during model evaluation.
- c. Include an explanation behind our conclusions, as well as any last remarks to conclude our analysis.

Results & Discussion

To reiterate, our analysis is framed by the following question: *How can we make efficient use of the existing labeled data to build a robust model for predicting fish weight?* With this in mind, we particularly look to explore whether a vertical, diagonal, or cross measurement is most useful in producing a model based on height and width to predict fish weight such that our research poses the question: *Which length measurement (vertical, diagonal, or cross), combined with height and width, creates a model that best predicts fish weight?*

Before performing analysis, we suspected that our results could reflect that models including a length measurement would perform better than the model that does not include a length measurement. We considered this possibility due to how adding more parameters to a linear model allows for the model to better represent patterns found in our dataset. Yet adding more parameters also causes SSE_{train} to decrease, meaning that we must pay close attention to SSE_{test} and MSE_{test} to evaluate the results of our analysis against our hypothesis.

After performing analysis on each of the four models, we obtain the following results:

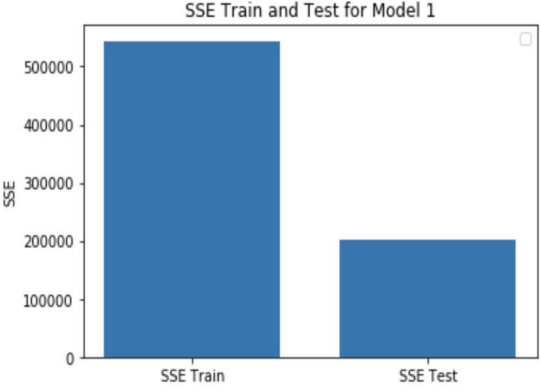
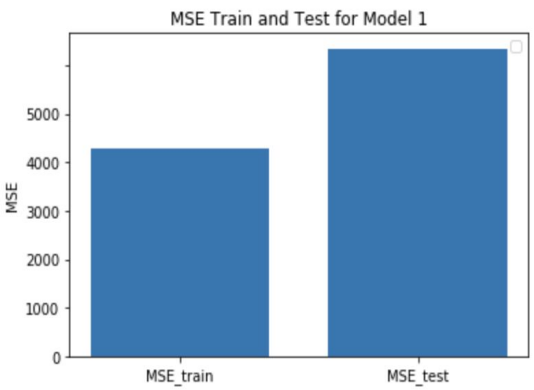
Model 1 $weight = -83.87 + (-17.94 * height) + (-1.83 * width) + (11.66 * vertical_length) + (0.24 * height * width * vertical_length)$			
			
SSE_{train}	SSE_{test}	MSE_{train}	MSE_{test}
543,936.58	202,931.36	4,282.96	6,341.60

Table 1. Results for Model 1, which evaluates how the *vertical* length measurement predicts weight.


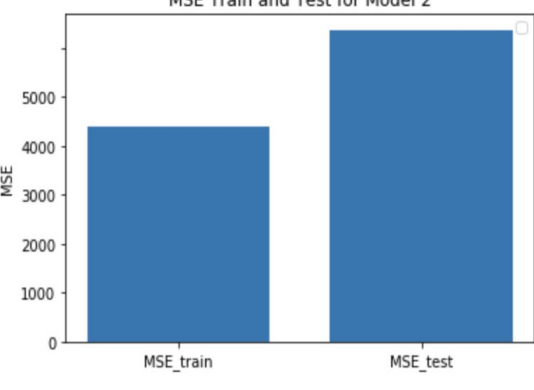
Model 2 $weight = -88.98 + (-19.68 * height) + (0.03 * width) + (11.26 * diagonal_length) + (0.22 * height * width * diagonal_length)$			
			
SSE_{train}	SSE_{test}	MSE_{train}	MSE_{test}
557,955.41	203,886.75	4,393.34	6,371.46

Table 2. Results for Model 2, which evaluates how the *diagonal* length measurement predicts weight.

Results for Model 3 and Model 4 are outlined below:

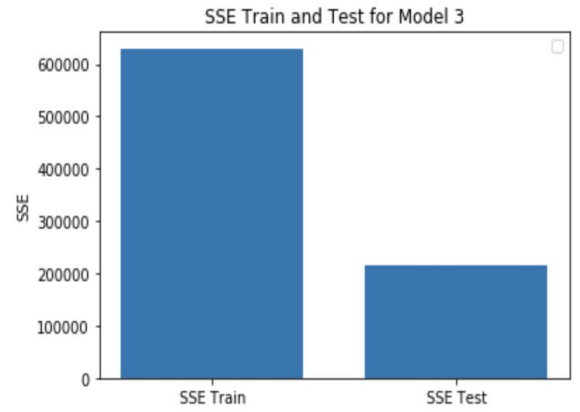

Model 3 $weight = -97.10 + (-17.94 * height) + (-30.40 * width) + (24.24 * cross_length) + (10.17 * height * width * cross_length)$			
			
SSE_{train}	SSE_{test}	MSE_{train}	MSE_{test}
630,005.39	215,009.16	4,960.67	6,719.03

Table 3. Results for Model 3, which evaluates how the *cross length* measurement predicts *weight*.

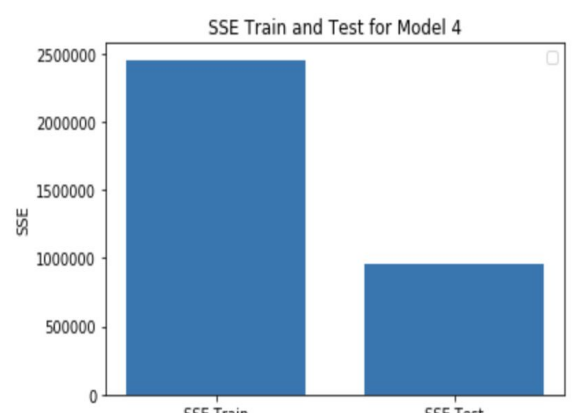
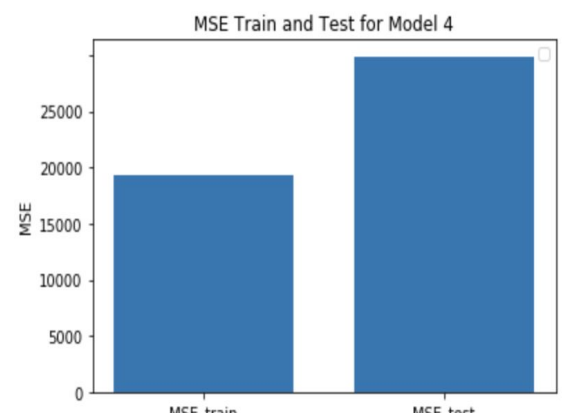
Model 4 $weight = -84.13 + (-49.37 * height) + (98.80 * width) + (10.81 * height * width)$			
			
SSE_{train}	SSE_{test}	MSE_{train}	MSE_{test}
2,452,554.61	956,738.50	19,311.45	29,898.07

Table 4. Results for Model 4, which evaluates how *height* and *width* predicts *weight* data.

Taking a closer look at each model's MSE_{train} and MSE_{test} measurements, we can see that MSE_{train} is lower than MSE_{test} across the board. Although cases where a model's testing error is much higher than its training error may indicate that the model is overfitting to the training data, cases where testing error is, in comparison, only slightly higher than its training error is not out of the ordinary. In our case, Model 2 has the least discrepancy between train and test error as its difference is equal to 1,978.12. Model 1 follows closely behind with the second-least amount of discrepancy between train and test error with a difference of 2,058.64. These discrepancies may be explained by a possible limitation of our model, being that our dataset only contains 159 samples. With a larger dataset, more training has the potential to reduce this discrepancy even further. Yet as discrepancies between training and testing error is to be expected, these models do not necessarily indicate extreme overfitting to training data and thus robustly represent generalizable patterns in the dataset to some extent.

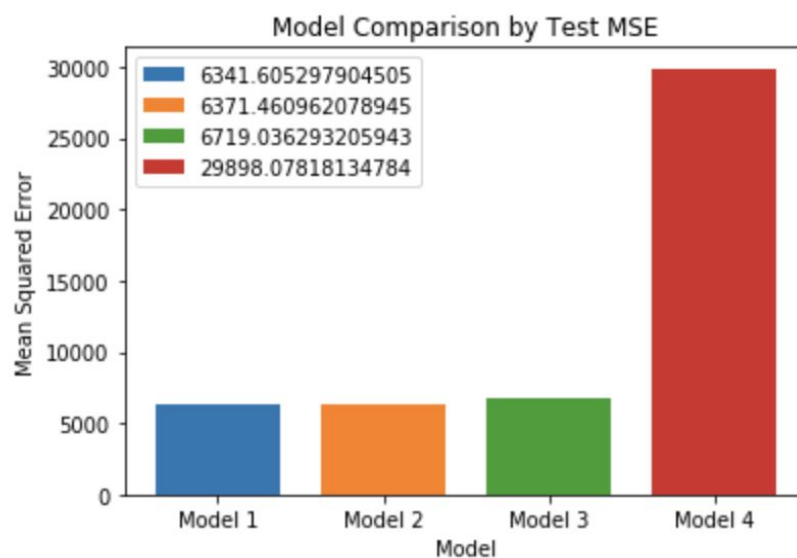


Figure 2. A bar graph comparing the MSE_{test} scores of Model 1 (*vertical*), Model 2 (*diagonal*), Model 3 (*cross*), and Model 4 (*height and width only*).

Based on the results shown above it can be seen that the use of length measures in predicting weight significantly improves prediction results, but the very small relative differences in MSE_{test} for the 3 models show that although the vertical length of the fish was the best predictor (Model 1), there was little variance between the models that used different length measurements to predict fish weight. Thus the results of our analysis confirm our hypothesis that the models which include a length measurement perform better than the model that does not include a length measurement. A possible explanation for this result could be that a fish's length measurements provide additional valuable information about its anatomy such that weight is able to be more accurately predicted.

Addressing our research question, we can conclude that the inclusion of a *length* measurement alongside *height* and *width* measurements produces a model that predicts fish *weight* more accurately than a model that does not consider *length*. This conclusion is based on the consistently lower MSE_{test} scores observed for Models 1-3. As such, the multivariate linear regression analysis

was appropriate for both this dataset and our research question as we were able to utilize the continuous length measurements recorded to predict fish weight.

To expand upon our findings in future work, we suggest that the distribution of fish species within this dataset be further explored to determine how this distribution impacts each model's performance as our understanding of the data suggests that not all species are equally represented. As this is the case, further exploration of species bias within this particular dataset could be useful to improve model design.

Code Appendix:

COGS 109 Final Project Fish Weight Prediction ([pdf](#)) of Jupyter Notebook code)