# Executive Summary– Regression with an Abalone Dataset Kaggle Competition (Late Submission).

By Samuel Mbah Nde – DDS 8555 – Assignment 2

## Project Overview

This document is an executive summary of the work I did as part of Assignment 2. I divided this work into two parts both of which are tightly connected with the python jupyter notebook file attached to this document in the submission. The first part focuses on providing responses to the conceptual questions from ISLP while the second part summarizes the project contribution to Regression with an Abalone Dataset Kaggle competition.

**Part 1 – Responses to Conceptual Questions.**

Question 1): The first question in the screenshot below is extracted from page 128 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is shown in the screenshot below.

A screenshot of a paper

Description automatically generated

1. The prediction function is.

If x1 and x2 are constant and x3 is binary with 0 for high school and 1 for college, only the terms in x3 will be useful predictor. The other terms can be viewed as a constant k. So, the new equation will become.

With **x1 being non-negative**, the values of *f* will be larger for x3 = 1 than x3 =0 hence the second assertion is correct.

1. For a college graduate with IQ of 110 and a GPA of 4.0, the predicted value will be.

Thus, the predicted salary for the college graduate is $ 137,000.

1. False. The small coefficient of the interaction between GPA and IQ, means the combine effect of GPA and IQ on Salary is close to the sum of their individual effects. The significance of this coefficient is what will give us evidence of interaction between the two. In other words, evidence of interaction can only be gotten from a statistical test.

**Question 2. This question is extracted from page 130 of ISLP.**

A white paper with black text

Description automatically generated

**Part 2 – Participation in Regression in Regression in an Abalone Dataset Kaggle Competition.**

The objective of participating in this competition is to highlight my ability to implement regression models, evaluate it and interpret any findings from such an analysis.

1. **Introduction to the dataset:**

The Abalone dataset is a well-known dataset used in predictive analytics and machine learning. It consists of various physical measurements of abalones, a type of marine mollusk, and aims to predict the age of an abalone based on its attributes such as Weight, Sex, Weight, Height, and Diameter. Since abalones do not have easily countable growth rings like trees, estimating their age is crucial for ecological research and commercial fisheries.

1. **Exploratory Data Analysis.**

I downloaded the data from Kaggle.com and explored it, computing measures of central tendency and spread on the dataset and visualizing the relationship between variables in the dataset before creating the model. The following table shows a summary of some of the descriptive statistics I computed from the dataset. This table highlights that all the columns in the dataset have outliers.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Column Name | Unique Values | Mean | Std | Kurtosis | Skewness | Outliers Percent |
| Diameter | 126 | 0.402 | 0.098 | 3.001 | -0.695 | 1.744 |
| Height | 90 | 0.135 | 0.038 | 16.453 | 0.31 | 0.384 |
| Length | 157 | 0.517 | 0.118 | 3.133 | -0.732 | 1.611 |
| ShellWeight | 1,129 | 0.226 | 0.13 | 3.096 | 0.479 | 0.74 |
| ShuckedWeight | 1,799 | 0.341 | 0.204 | 3.284 | 0.592 | 1.067 |
| VisceraWeight | 979 | 0.169 | 0.101 | 2.796 | 0.477 | 0.486 |
| WholeWeight | 3,175 | 0.789 | 0.458 | 2.815 | 0.429 | 0.633 |

Table 1: Subset of summary statistics from dataset computed on training set.

I also plotted a pair plot of the variables in the training set colored by sex as shown in table 1 below.

A graph of a number of colored lines

Description automatically generated with medium confidence

Figure 1: Pair plot of numeric variables in dataset, colored by Sex

*Note that there is some polynomial degree of interaction between some variables in the dataset and that Male and Female Sex values tend to be in the same range while infant values tend to be small in separated from the other sex values.*

1. **Fitting Models on the dataset.**

I divided my dataset into training and validation sets using a train-test split ratio of 0.15 with the sample stratified on the Sex column to ensure that I get appropriate representation of each Sex in the training and validation sets which helped me minimize sampling bias while training the model (Ndung'u, 2022). I also excluded the feature id because it is just a serial sample identifier and does not capture any specific feature of the dataset.

Since Sex is a category, I used pandas to get the dummy variables from Sex and apply the same transformation on the test set during evaluation. To prevent the introduction of multi-collinearity in the model, I only used two columns from the dummies to create my model (Yip & Tsang, 2007).

Then I implemented two regression models, one, an ordinary least squares linear regression and the summary of that model is shown in figure 2 below.

A screenshot of a computer

Description automatically generated

Figure 2: Summary of OLS Linear Regression model on the training Set.   
*Note: All but one coefficients of the model are highly statistically significant.*

For a second model, guided by the clear polynomial relationships between the predictors as shown in Figure 1 above, I implemented a polynomial transformation of the inputs and fitted a second model which outperformed the first in mean squared error, R-squared and adjusted R-squared, mean absolute error, and mean log absolute error (Ostertagová, 2012). Figure 3 below is the residual plots for both models.

A screenshot of a graph

Description automatically generated

Figure 3: Comparison of residual plots for both models showing smaller spread in residuals for polynomial regression model.

1. **Preparation of data and submission to competition.**

I exported my predictions on the test set and submitted them to the competition. The screenshot below shows my submission (as a late entry). My submission score was 0.15482 earning me the 2049th position.

A screenshot of a computer

Description automatically generated

Figure 4: Late Entry submission result – Regression with an Abalone Dataset Kaggle competition.

1. **Looking Forward.**

While this exercise achieved its primary goal, there is room for improvement. Key enhancements that could be done include:

* 1. **Feature Selection with LASSO Regression**: LASSO can shrink insignificant coefficients, aiding feature selection and model validation as discussed in (Tibshirani, 1996).
  2. **KNN for Regression**: Given the dataset's wide-spread and presence of outliers, KNN—being locally sensitive—could provide more robust predictions.
  3. **Ensemble Modeling**: Combining multiple models can reduce errors and improve predictions, especially in datasets with many outliers.
  4. **Refining Dummy Variable Strategy**: Instead of encoding only Male and Infant categories, representing Male and Female together might yield different insights (Yip & Tsang, 2007).

Implementing these refinements could lead to a more accurate and resilient predictive model.

# Bibliography

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning with Applications in Python.* Springer. doi:10.1007/978-3-031-38747-0

Ndung'u, R. (2022). Data Preparation For Machine Learning Modelling. *International Journal of Computer Applications Technology and Research, 11*, 231-235. doi:10.7753/IJCATR1106.1008}

Ostertagová, E. (2012). Modelling using Polynomial Regression. *Procedia Engineering, 48*, 500-506. doi:10.1016/j.proeng.2012.09.545

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 1*, 267-288. Retrieved from http://www.jstor.org/stable/2346178

Yip, P., & Tsang, E. (2007). Interpreting Dummy Variables and Their Interaction Effects in Strategy Research. *Strategic Organization - STRATEG ORGAN, 5*, 13-30. doi:10.1177/1476127006073512