STAT844: STATISTICAL LEARNING, ADVANCED REGRESSION PROJECT PROPOSAL

BY HONGYU CHEN[1,a]

[1] *Faculty of Mathematics, University of Waterloo,* [a] *h542chen@uwaterloo.ca*

1. Introduction. Abalone is a type of mollusk. The shell of abalones is convex, rounded to oval, and may be highly arched or flattened. People love to eat them in many places like Latin America, France, New Zealand, Southeast Asia, China, Vietnam, Japan, and Korea because they're tasty and full of good nutrients. They're also farmed for their shiny shells, which are used for decoration. (1) This makes abalone a highly sought after commodity and economically significant. The price of an abalone is positively correlated to its age. (2) Figuring out how old an abalone is can be quite tricky. As it grows, rings appear in its inner shell, with one new ring each year. To count these rings, the shell needs to be cut open. Then, after polishing and staining, a lab technician checks a small piece of the shell under a microscope to count the rings. (3) The **research question** of this project is to explore the possibility of predicting the age of abalone using physical measurements that are easier to obtain, so we can streamline the age-determination process. Predicting the age of abalone from physical measurements offers a faster alternative way to the time-consuming process of manually counting rings through a microscope. By analyzing easily obtainable physical attributes such as shell size and weight, predictive models can estimate the age of abalone more efficiently. These models can establish correlations between these measurable characteristics and the number of rings which simplifies the age-determination process.

2. Dataset Description. The dataset used in this project is sourced from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/1/abalone) (3) and contains measurements of the physical attributes of abalone.

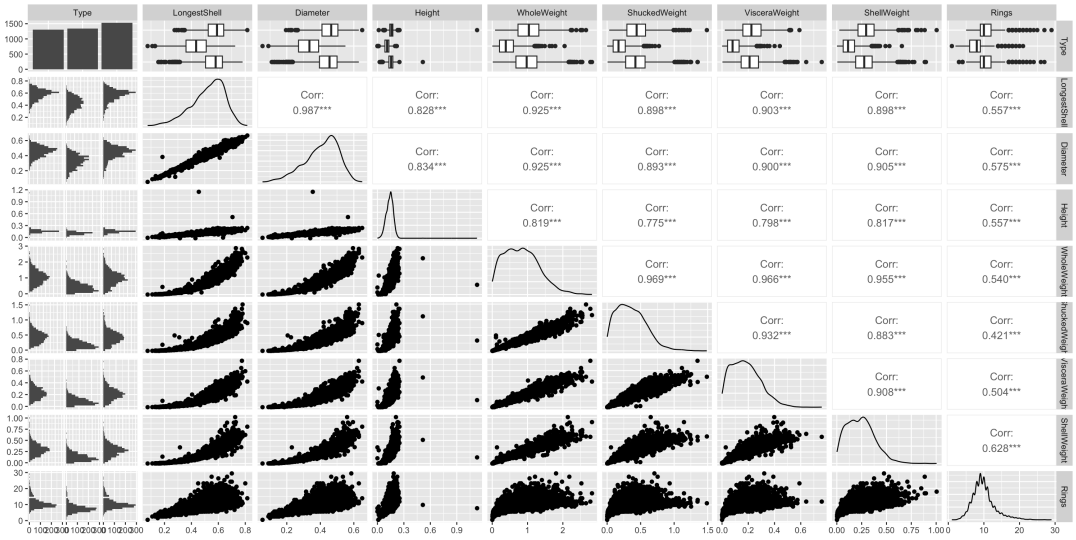

FIG 1. *Relational matrix*

It includes attributes such as sex, length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight. The response variable of interest is the number of rings in the shell. From the summary, we observe that most of the variables have positive correlations with the number of rings which indicates older abalone have larger shell sizes and higher weights. These attributes are likely to be useful predictors in our future model prediction for estimating the age of abalone, as they provide valuable insights into the physical characteristics associated with age. But we also need to be aware that the whole weight is mostly the sum of shell and viscera weight and other features. So while building a model we shouldn't select them together. We will do a further feature selection.

3. Analysis Plan.

3.1. *Exploratory Data Analysis.* Includes a comprehensive examination of data summary and statistics, covering both categorical and numeric data. Uni-variate analysis is conducted to explore the distribution of each attribute. Followed by bivariate analysis, which involves exploring relationships between pairs of attributes using different plots, correlation analysis, and visualizations such as ggpairs. Additionally, multivariate analysis techniques like Principal Component Analysis (PCA) or Factor Analysis of Mixed Data (FAMD) are employed to investigate interactions between multiple attributes. Finally, the discussion on data quality by examining the overall summary. Doing data cleaning or the addition of new variables if necessary to enhance the analysis.

3.2. *Model Development.* My objective is to employ a diverse range of regression techniques to predict and assess the performance of my predictions, encompassing both models within and beyond the course curriculum. In this project, I aim to train the following models which will select 3 models from the course's scope and explore others outside the scope.
Models covered in the course, including those introduced later. Presently, my preferences lean towards K-nearest Neighbors and Additive Models. Other options to consider are Lasso Regression, Ridge Regression, Spline Regression, and Polynomial Regression, alongside potentially other models discussed in forthcoming lectures. I prefer exploring a model that was not implemented in our assignments.
Beyond the course: (A brief explanation of each of these models will be included in the paper)

- Elastic Net: Offers regularization benefits by combining L1 and L2 penalties.
- Random Forest: Provides robustness to overfitting and handles large datasets effectively.
- Regression Trees: Offers interpretability and handles non-linear relationships well.
- Support Vector Machines: Excels in high-dimensional spaces and provides flexibility with different kernel functions.
- Neural Networks: Offers powerful learning capabilities for complex patterns and large datasets.

Interaction and polynomial transformation will be used to enhance the model's performance.

3.3. *Evaluation.*

- Use Root Mean Square Error (RMSE) to quantify the average difference between predicted and observed values.
- Construct ANOVA-F tests to assess the significance of differences between model fits.
- Implement Stepwise Regression, AIC, and BIC for variable selection to add or remove predictors based on their statistical significance.
- Calculate the Variance Inflation Factor to assess multicollinearity and detect potential issues with predictor variables.
- Construct plots to visualize model performance and facilitate interpretation of the results.

4. REFERENCES.

REFERENCES

[1] Abalone: https://en.wikipedia.org/wiki/Abalone

[2] Hossain, M, & Chowdhury, N (2019) Econometric Ways to Estimate the Age and Price of Abalone. Department of Economics, University of Nevada

[3] UCI Machine Learning Repository, Abalone dataset:https://archive.ics.uci.edu/ml/datasets/Abalone