# Executive Summary– Build Non-Linear Models Part 1 and Continuation of Regression with an Abalone Dataset Kaggle Competition (Late Submission).

By Samuel Mbah Nde – DDS 8555 – Assignment 3

## Project Overview.

This project is the continuation of the work I have done in previous weeks, participating in the Kaggle competition of Regression with an Abalone Dataset. The work is divided into 2 sections. In section 1, I focus on answering the conceptual questions from Chapter 6 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023). In section 2, I highlight the changes I implemented to improve my regression model in the Abalone Dataset competition in Kaggle (Reade & Chow, 2024). The code used for this analysis can be found in my GitHub repository (Nde, 2024).

Section 1: Conceptual Questions.

Question 1: Question 1 was extracted from page 283 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.

A screenshot of a paper

AI-generated content may be incorrect.

Solution.

1. The **best subset selection** should have the smallest training **Residual Sum of Squares (RSS)** for a given number of *k* predictors.

This is because **best subset selection** evaluates all possible combinations of *k* predictors and selects the one that minimizes training RSS. Forward stepwise selection is constrained because it builds the model sequentially, only adding one predictor at a time Thus, it may not find the absolute best model for a given *k.* Likewise, backward stepwise selection starts with all *p* predictors and removes them one by one, which means it does not necessarily explore all possible models with *k* predictors. Since best subset selection searches across all possible models, it has the flexibility to find the model that achieves the lowest training RSS.

1. None of the methods guarantee the smallest test RSS for a fixed *k*, but forward and backward stepwise selection often perform better in terms of test error compared to best subset selection.
2. These are the responses with the reasons.

i) **True**. The sequential nature in which forward stepwise process selects predictors means predictors in the *k-variable* model must be included in the *(k+1)-variable* model. This ensures a nested structure where additional predictors are added iteratively.

ii) **True**: This is because the backward stepwise predictor selection process works in a sequential manner similar to forward stepwise process.

iii) **False:** Predictor selection forward and backward stepwise happen in sequential manner (with the forward process continually adding predictors while the backward process removes them) but there is no guarantee that both processes will “converge” on some *k* variables.

iv) **False:** There is no requirement for the forward stepwise selection model of size *k* to be contained within the *k+1* variable model chosen by backward selection.

v) **False:** Unlike the forward and backward stepwise predictor selection processes that are sequential and can be nested, the best subset is not. The predictors in each model of *m* predictors is chosen independently.

Question 2: Question 2 was extracted from page 283 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.

A screenshot of a math test

AI-generated content may be incorrect.

Solution.

This problem is solved in the notebook attached to this homework submission files and linked in my GitHub repository (Nde, 2024). The following paragraph summarizes the work that was done.

In my analysis of the **Regression with an Abalone Dataset**, I applied **Lasso Regression** and **Principal Component Regression (PCR)** to perform feature selection before fitting a **Linear Regression model**. The expectation of using Lasso regression was that it would help identify the most relevant predictors by shrinking some coefficients to 0 but this was not the case (Erick, George, & Michael, 2021). Failure to shrink any coefficients to 0 indicates that all the variables in the model were useful in predicting the rings (Effrosynidis & Arampatzis, 2021). The lasso regression model was the first model that resulted in invalid predictions which suggests that the process might have resulted in loss of some information.

PCR transformed the original features into orthogonal principal components to capture the most variance in the data. I used cross validation to determine the optimal number of features and as shown in Figure 1 below, the optimal number of components turned out to be 6. These 6 components explained more than 97% of the variability in the predictors.

The Lasso-based and PCR-based models **underperformed compared to the Polynomial Linear Regression model** I had implemented in the previous week. The polynomial regression approach was able to capture more complex relationships in the dataset, leading to better predictive accuracy, suggesting that **nonlinear relationships are crucial for modeling abalone age** effectively.

Other techniques that may improve the model could include applying lasso regression on the polynomial features generated from the dataset or applying principal components on the polynomial features model.

A graph with blue lines

AI-generated content may be incorrect.

Figure 1: *Cross validation Error on test set by number of Principal components.*

# References

Effrosynidis, D., & Arampatzis, A. (2021). An evaluation of feature selection methods for environmental data. *Ecological Informatics, 61*, 101-224. doi:10.1016/j.ecoinf.2021.101224

Erick, O. O., George, O. O., & Michael, W. K. (2021). Feature Selection for Classification using Principal Component Analysis and Information Gain. *Expert Systems with Applications, 174*, 114765. doi:10.1016/j.eswa.2021.114765

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python.* Springer Nature.

Nde, S. M. (2024, 02). *DDS-8555*. Retrieved from Github: https://github.com/ndesamuelmbah/DDS-8555.git

Reade, W., & Chow, A. (2024). *Regression with an Abalone Dataset*. Retrieved from Kaggle: https://kaggle.com/competitions/playground-series-s4e4