# Executive Summary– Build Non-Linear Models Part 2 and Continuation of Regression with an Abalone Dataset Kaggle Competition (Late Submission).

By Samuel Mbah Nde – DDS 8555 – Assignment 4

## 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 two sections. In section 1, I focus on answering the conceptual questions from Chapter 7 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 https://github.com/ndesamuelmbah/DDS-8555/tree/main/AssignmentFiles/Week4 (Nde, 2024).

Section 1: Conceptual Questions.

Question 1: This is question 3 extracted from page 326 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.

A math equations on a white background

AI-generated content may be incorrect.

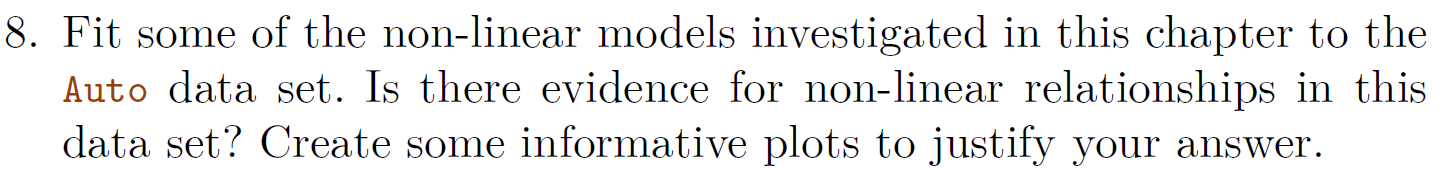
Solution.  
**Figure 1**  
*Sketch of curve of the estimated linear regression model equation.*

A graph with a line and a dotted line

AI-generated content may be incorrect.

Note. *The important point such as intercept, transition point and the inflexion point are annotated in the graph.*

Question 2: This is question 8 was extracted from page 327 of ISLP (James, Witten, Hastie, Tibshirani, & Taylor, 2023) and is displayed in the screenshot below.



Solution.  
I solved this problem in the notebook attached to this homework submission files and linked in my GitHub repository (Nde, 2024). The following paragraphs summarizes the work that I did.

I performed some exploratory data analysis of the auto dataset and plotted a pair plot of the variables in it. From the pair plot shown in figure 1 below, I observed a polynomial relationship between mpg and horsepower and was curious to better understand that relationship.   
**Figure 2**  
 *Pairplot of variables in Auto dataset.*A screenshot of a graph

AI-generated content may be incorrect.  
Note. *The hue is a custom column which I computed by concatenating the origin of the vehicle with the number of cylinders it has – A combination which helps us understand the relationship between other features and miles per gallon. Also note the curve-like (polynomial) relationship between horsepower and mpg.*

To determine the optimal degree of polynomial to use for modeling the relationship between horsepower and mpg, I performed a series Anova test for polynomials of degrees up to 6 to get the optimal polynomial that will model this relationship (Holzer & Precht, 1992). Each test compares each polynomial of degree *n* with the polynomial of degree *n-1* testing the null hypothesis that the polynomial of degree *n-1* is sufficient to explain the data against the alternative hypothesis that a more complex model of degree *n* is required. Table 1 below summarizes the finding from the Anova tests.

**Table 1**

*Result table of Anova tests to help determine the optimal degree of polynomial for relationship between horsepower and mpg.*

A table with numbers and letters

AI-generated content may be incorrect.

Note: *When we compare models[0] (linear) with model[1] (quadratic), the p-value is essentially 0 meaning that the linear relationship is not sufficient to capture the relationship.* *The comparison between models[1] (quadratic) and models[2] (cubic) provides a p value of 0.36 which tells us that the quadratic is a good fit for the data. Similar comparison between models[2] (cubic) and models[3] (quartic) has a p-value of 0.23 which also means a cubic is a good fit but there is no point in fitting any polynomial beyond the cubic.*

Since both the **quadratic** and **cubic** polynomial models perform well, I would opt for the **quadratic** model as it offers a good balance between **accuracy and complexity, reducing the risk of overfitting while still capturing key patterns in the data**. Figure 3 below shows a 2-degree polynomial using horsepower to predict mpg.

**Figure 3**  
*Polynomial of degree 2 (quadratic) in blue with a 95% confidence interval plotted in red dashes.*

A graph of a graph with a red line and blue line

AI-generated content may be incorrect.

*Note. Most of the point lie close to the model line in blue supporting that a quadratic is a good fit for the relationship between horsepower and mpg in the auto dataset.*

Section 2: Regression on an Abalone dataset – Continued.

This week, I improved my previous model by performing polynomial regression on the dataset. To determine the optimal value of degree of polynomial to use for the regression, I captured metrics for polynomials up to degree 4 with focus on the adjusted R-squared score and the custom metric log root mean squared error (Karch, 2019). Figure 4 below shows a summary of the results.

**Figure 4**

*Summary results using polynomials of degrees 1 to 4 to predict Rings from the Abalone dataset.*

A screenshot of a computer

AI-generated content may be incorrect.

*Note. We see a decline in the custom score Root Mean Squared Log Erro for models from degree 1 to degree 4 accompanied with an increase in adjusted R-squared. After this point, the model starts showing a decline in both scores indicating that we have started overfitting. Thus, a cubic would be a better fit for the data.*

The predictions from my cubit model produced the best score yet for my dataset. Figure 5 below shows my submissions to the competition.

Figure 5  
*Kaggle Submissions showing cubic polynomial as my best model yet.*A screenshot of a computer

AI-generated content may be incorrect.

# References

Holzer, C., & Precht, M. (1992). Multiple comparison procedures for normally distributed ANOVA models in SAS, SPSS, BMDP, and MINITAB. *Computational Statistics & Data Analysis, 13*(3), 351-358. doi:10.1016/0167-9473(92)90141-2

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

Karch, J. (2019, 09). Improving on Adjusted R-Squared. doi:10.31234/osf.io/v8dz5

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