Predicting the Age of Abalones: An Analysis of Transformations in Regression Modelling

Angela Jacinto
23778435

Executive Summary

An Abalone is a type of single-shelled marine mollusc considered a delicacy in various parts of the world. Commercial fisheries in Western Australia are major seafood exporters due to the continuous increase in demand caused by poor fisheries management and overfishing in many countries.

The study's focus was to model the number of rings on the shell of abalone, an indicator of age, using other physical measurements from the dataset 'abalone.txt'. The research involved examining 4,177 data records, encompassing variables such as the abalone's sex, shell measurements (length, diameter, height), and various weight metrics (total, meat, gut, shell). Initial observations suggest that there is a positive correlation between the physical measurements and the abalone's age. The analysis further revealed that the log-transformed number of rings was best modelled by a relatively simple approach, favoring fewer interaction terms among continuous variables.

Introduction

Abalones are marine snails with a unique shell structure that is considered a very popular delicacy in various parts of the world. Given this, they play a vital role in the marine ecosystem and have a significant economic value. The age of an abalone represented by the number of rings on its shell is not just a physical attribute but also an indicator of its value. Economically, its value is directly correlated to the seafood's age. Traditionally, the age of an abalone can be determined by counting the number of rings on their shells under a microscope. This method involves cutting and staining the shell, which can be very tedious, damaging to the species and costly for fisheries. In addition, due to staining limitations, an approximation of the abalone's age is determined by adding 1.5 to the ring count. Given these constraints, the traditional method is not effective in providing a precise and accurate estimation of the abalone's age. Given other physical measurements, which are easier to obtain from the abalone, it is possible to predict the age without the use of intrusive methods.

Various studies have been conducted and researchers have been using different methods in order to determine the age of an abalone suggesting correlations between size, weight, and age of marine molluscs. For example, Mohammed, G. R. et al. used a Just Neural Network tool [64] to build a multilayer Artificial Neural Network where factors that were considered to influence predicting the age of an abalone were classified as input variables. These inputs were then fed into a neural network resulting in the output variable as the predicted age of the abalone based on the input variables. The use of a feed forward back propagation algorithm for training the model was essential in obtaining high accuracy in predictions. The study proved that an ANN model for predicting the age of an abalone from physical measurements was successful with an accuracy rate of 92.22 %.

Our dataset 'abalone.txt', comprises several useful information. Containing records of a sample size of 4,177 abalones will provide a comprehensive set of physical measurements to analyse. The dataset contains nine variables namely, Sex - a categorical variable where an Abalone could be categorised as Male (M), Female (F), or Infant (I). Length - The longest shell measurement in millimetres. Diameter - The shell's diameter is in millimetres. Height - The measurement of the abalone's shell height in millimetres. Whole weight - The total weight of the abalone in grams. Shucked weight - The weight of the abalone's meat in grams. Viscera weight - The gut-weight of the shell post-bleeding in grams. Shell weight - The weight of the shell post-drying in grams. Rings - The number of rings on the shell is +1.5 which gives the age of the abalone in years.
The primary aim of this study is to identify the most appropriate and suitable model that can accurately predict the age of the abalone while exploring the possibilities of a non-invasive method using the given physical measurements by analysing the relationship between the number of rings and other variables.

Methodology

We began the analysis with initial data pre-processing where the categorical variable "Sex" was transformed into a factor for R to recognize it as a categorical variable while avoiding the treatment of the variable as continuous as it can lead to incorrect and misleading results.
A multiple regression analysis was then conducted to model the relationship between the response variable "Rings" and the independent variables. Model diagnostics were done to assess the assumptions - linearity, normally distributed residuals, homoscedasticity, and independence where errors are uncorrelated. Based on the diagnostic plots and statistical significance, the model was refined by removing insignificant variables and introducing interactions between specific predictors to better capture the patterns in the data. To better address the model assumptions, two model transformations were explored – a log transformation and a square root transformation on the "Rings" variable. Regression analyses and

diagnostic checks were performed on both transformed models to assess the appropriateness of the transformation. Finally, the original and log-transformed models were compared and analysed according to several metrics such as R^2 values, which explained how much of the variance was explained by each model. The residual plots were also compared to visualize the distribution, pattern and spread of the data. Lastly, histograms and QQ plots were used to assess the normality of the errors. The findings from these diagnostics were then used for proper model selection.

Results

Figure 1

	Length	Diameter	Height	Wholewt	Shuckedwt	Viscerawt	Shellwt	Rings
Min	0.0750	0.550	0.0000	0.0020	0.0010	0.0005	0.0015	1
Max	0.8150	0.6500	1.1300	2.8255	1.4880	0.7600	1.0050	29
Mean	0.5234	0.4079	0.1395	0.8287	0.3594	0.1806	0.2388	9.9337
Median	0.5450	0.4250	0.1400	0.7995	0.3360	0.1710	0.2340	9

Figure 2

Sex	Number of Observations
F	1307
I	1342
M	1528

We first examined the data to discern the nature of the values and variables that will be dealt with. The tables illustrate a somewhat even distribution between the levels of the categorical variable "sex". Interestingly, there is an irregularity with the "height" variable as it contains a minimum value of 0.

Full Model

$$\text{Rings} = \beta0 + \beta1 \times \text{Sex} + \beta2 \times \text{Length} + \beta3 \times \text{Diameter} + \beta4 \times \text{Height} + \beta5 \times \text{Whole} + \beta6 \times \text{Shuckedwt} + \beta7 \times \text{Viscerawt} + \beta8 \times \text{Shellwt} + \beta9 \times \text{Sex: Length} + \beta10 \times \text{Sex: Diameter} + \beta11 \times \text{Sex: Height} + \beta12 \times \text{Sex: Wholewt} + \beta13 \times \text{Sex: Shuckedwt} + \beta14 \times \text{Sex: Viscerawt} + \beta15 \times \text{Sex: Shellwt} + \beta16 \times \text{Length: Diameter} + \beta17 \times \text{Length: Height} + \beta18 \times \text{Length: Wholewt} + \beta19 \times \text{Length: Shuckedwt} + \beta20 \times \text{Length: Viscerawt} + \beta21 \times \text{Length: Shellwt} + \beta22 \times \text{Diameter: Height} + \beta23 \times \text{Diameter: Wholewt} + \beta24 \times \text{Diameter: Shuckedwt} + \beta25 \times \text{Diameter: Viscerawt} + \beta26 \times \text{Diameter: Shellwt} + \beta27 \times \text{Height: Wholewt} + \beta28 \times \text{Height: Shuckedwt} + \beta29 \times \text{Height: Viscerawt} + \beta30 \times \text{Height: Shellwt} + \beta31 \times \text{Wholewt: Shuckedwt} + \beta32 \times \text{Wholewt: Viscerawt} + \beta33 \times \text{Wholewt: Shellwt} + \beta34 \times \text{Shuckedwt: Viscerawt} + \beta35 \times \text{Shuckedwt: Shellwt}$$

We constructed a linear regression model where we included the main effects of each variable, their quadratic terms and all two-way interactions to create a full model that will capture all possible effects and patterns present within the data.

Refined Model I

$$\text{Rings} = \beta0 + \beta1 \times \text{Sex} + \beta2 \times \text{Length} + \beta3 \times \text{Diameter} + \beta4 \times \text{Height} + \beta5 \times \text{Whole} + \beta6 \times \text{Shuckedwt} + \beta7 \times \text{Viscerawt} + \beta8 \times \text{Shellwt} + \beta9 \times \text{Sex: Length} + \beta10 \times \text{Sex: Height} + \beta11 \times \text{Sex: Wholewt} + \beta12 \times \text{Sex: Shuckedwt} + \beta13 \times \text{Sex: Viscerawt} + \beta14 \times \text{Length: Diameter} + \beta15 \times \text{Length: Wholewt} + \beta16 \times \text{Length: Shuckedwt} + \beta17 \times \text{Diameter: Wholewt} + \beta18 \times \text{Height: Wholewt} + \beta19 \times \text{Height: Shuckedwt} + \beta20 \times \text{Wholewt: Shuckedwt} + \beta21 \times \text{Wholewt: Viscerawt} + \beta22 \times \text{Shuckedwt: Viscerawt} + \beta23 \times \text{Shuckedwt: Shellwt} + \beta24 \times \text{Viscerawt: Shellwt}$$

We then applied a model selection technique using backward elimination with the Akaike Information Criterion (AIC) as a metric to decide which variable to retain in the model. The function "stepAIC" from the "MASS" package was used to perform stepwise model selection based on the AIC where we got an AIC of 6221.61. The summary of this model was then obtained in order to get the refined model that resulted from the elimination process. The results of the process give us Refined Model I.

Refined Model II

$$\text{Rings} = \beta0 + \beta1 \times \text{Sex} + \beta2 \times \text{Length} + \beta3 \times \text{Diameter} + \beta4 \times \text{Height} + \beta5 \times \text{Whole} + \beta6 \times \text{Shuckedwt} + \beta7 \times \text{Viscerawt} + \beta8 \times \text{Shellwt} + \beta9 \times \text{Sex: Length} + \beta10 \times \text{Sex: Shuckedwt} + \beta11 \times \text{Length: Diameter} + \beta12 \times \text{Length: Wholewt} + \beta13 \times \text{Length: Shuckedwt} + \beta14 \times \text{Diameter: Wholewt} + \beta15 \times \text{Height: Wholewt} + \beta16 \times \text{Height: Shuckedwt} + \beta17 \times \text{Wholewt: Shuckedwt} + \beta18 \times \text{Wholewt: Viscerawt} + \beta19 \times \text{Shuckedwt: Viscerawt} + \beta20 \times \text{Shuckedwt: Shellwt}$$

We further refined the model by removing variables that are insignificant based on its p-values. *"Sex:Wholewt, Sex:Viscerawt, Sex:Height, Viscerawt:Shellwt"* were dropped from the model as they were not significant, arriving at Refined Model II.

Figure 3

Min	-10.6921
Median	-0.2737
Max	14.6972
Multiple R^2	0.5778

As shown in the statistical summary from the refined model in Figure 3, the model captures a significant proportion of the data which explains approximately 57.78 % of the variance in the age of the abalone. However, the residuals ranging from -10.6921 to 14.6972 indicate a large number of potential outliers.

Log-transformed Model

$$\log(\text{Rings})\hat{} = \beta 0 + \beta 1 \cdot \text{Sex} + \beta 2 \cdot \log(\text{Length}) + \beta 3 \cdot \log(\text{Diameter}) + \beta 4 \cdot \text{Height} + \beta 5 \cdot \log(\text{Wholewt}) + \beta 6 \cdot \log(\text{Shuckedwt}) + \beta 7 \cdot \log(\text{Viscerawt}) + \beta 8 \cdot \log(\text{Shellwt}) + \beta 9 \cdot \text{Sex: Height} + \beta 10 \cdot \text{Sex} \cdot \log(\text{Wholewt}) + \beta 11 \cdot \text{Sex} \cdot \log(\text{Shuckedwt}) + \beta 12 \cdot \log(\text{Length}) \cdot \log(\text{Diameter}) + \beta 13 \cdot \text{Height} \cdot \log(\text{Wholewt}) + \beta 14 \cdot \text{Height} \cdot \log(\text{Shuckedwt}) + \epsilon$$

To improve the accuracy and performance of the model, we used a log transformation approach to optimize the regression model. Where the following interaction terms were dropped:
"log(Shuckedwt):log(Shellwt), log(Viscerawt):log(Shellwt), log(Shuckedwt):log(Viscerawt), log(Wholewt):log(Viscerawt), log(Wholewt):log(Shuckedwt), log(Length):log(Shuckedwt), log(Length):log(Wholewt), Sex:log(Viscerawt), Sex:log(Length), log(Diameter):log(Wholewt)"
The resulting model was expressed as the Log-transformed Model above.

Figure 4

Min	-1.0057
Median	-0.0133
Max	0.7946
Multiple R^2	0.663

As seen in Figure 4, the log-transformed model had an R^2 of 0.663 suggesting that approximately 66.3 % of the variability in the log-transformed age of the abalone can be explained by the predictors. We also found that the residuals ranged between -1.00568 and 0.79459 - narrower than that of the refined model.

Figure 5

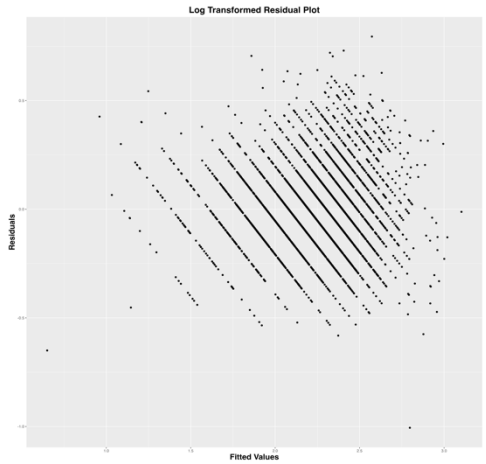


Figure 6

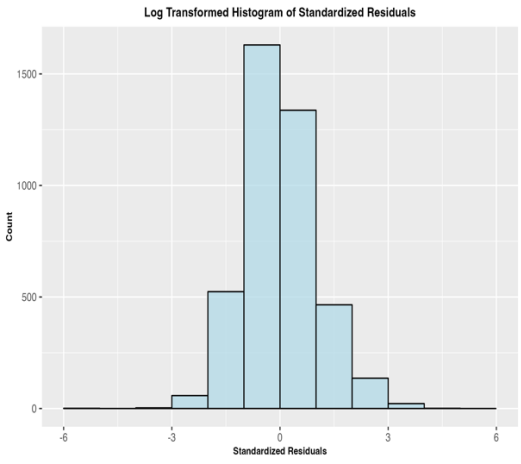
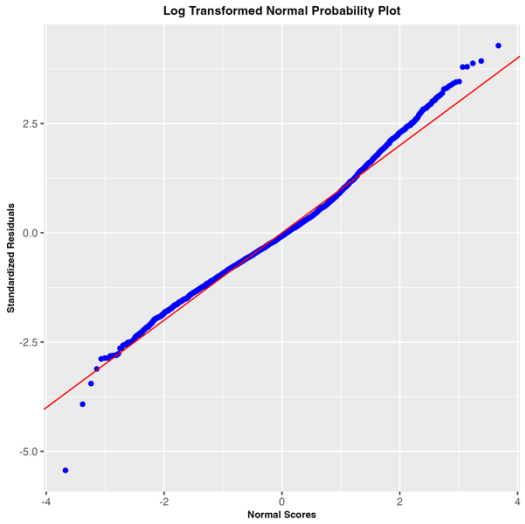


Figure 7



To further select between the refined model and the log-transformed model, we examined its model fit. The following plots provide a comprehensive understanding of the log-transformed model's performance. The residual plot illustrates that the residuals are randomly scattered around the zero line without any clear pattern while the spread of the residuals appears to be consistent across the range of fitted values. There are a few points that are far from the zero line, which suggests potential outliers. However, most of the residuals are close to the zero line, which tells us that the predictions of the model are relatively accurate. Overall, we can say that the transformed model seems to be a good fit as its assumptions are met.

The histogram for the residuals of the log-transformed model shows that most of the residuals are centered around zero and demonstrate a slight bell shape, which indicates that it follows a good approximation of normal distribution. However, there is a bit of a right-skewed distribution, suggesting a few positive outliers.

The QQ plot shows that most of the data points are close to the straight line, which tells us that the residuals are roughly normally distributed. There are some deviations from the straight line at the upper and lower tail, which might indicate some non-normal behavior.

Given the detailed analysis of the residual plot, histogram, and QQ plot for the log-transformed model, we conclude that this model is most appropriate for determining the age of abalone.

The model equation for the log-transformed model is

*log(rings) = 1.89923 + (-0.05700 * SexI) + (0.08643 * SexM) + (-0.45491 * log(Length)) + (0.01183 * log(Diameter)) + (-1.32842 * Height) + (0.39378 * log(Wholewt)) + (-0.42734 * log(Shuckedwt)) + (-0.08529 * log(Viscerawt)) + (0.24858 * log(Shellwt)) + (2.45125 * SexI:Height) + (-0.43285 * SexM:Height) + (-0.41271 * SexI:log(Wholewt)) + (0.04418 * SexM:log(Wholewt)) + (0.39767 * SexI:log(Shuckedwt)) + (0.01356 * SexM:log(Shuckedwt)) + (-0.22498 * log(Length):log(Diameter)) + (3.12359 * Height:log(Wholewt)) + (-2.63997 * Height:log(Shuckedwt))*

Discussion

The log-transformed model for predicting the age of abalones based on their physical measurements demonstrates notable fitting characteristics. The randomness and consistently even spread of the residuals around the zero line in the residual plot adheres to the model's assumptions of linearity, independence, and homoscedasticity. Despite the presence of a few potential outliers, most of the residuals that lie close to zero indicate that the model's predictions are significantly accurate. While the residual plot suggests a normal distribution, the histogram demonstrates a slight skew to the right and the QQ plot demonstrates minor deviations at the tails, which indicate a few potential outliers that can potentially impact the model's predictions. The model can be enhanced by addressing these outliers, or by using other forms of transformations to find patterns not captured by the current model.

In summary, the log-transformed regression model has addressed some of the limitations in the original model, as indicated by the residual plot, histogram and QQ plot. With an R^2value of 0.663, the model explains 66.3 % of the variability in the log-transformed age of the abalone, which indicates a notable fit. Given this, it is reasonable to say that the model is effective in predicting the age of abalones based on their physical measurements as the model's diagnostics provide strong evidence of its reliability.

The effects of the variables on the log of number of rings on the abalone's age are as follows.

1. Being male is associated with a slight increase in the number of rings in which every male abalone, the log number of rings increases by 0.08643.

2. Whole weight and shell weight have a positive effect on the number of rings. For every unit increase in the logarithm of the whole weight, there is an approximate increase of 0.39378 in the logarithm of the number of rings. This indicates that heavier abalones tend to have more rings. While for every unit increase in the logarithm of the shell weight, the logarithm of the number of rings increases by approximately 0.24858, suggesting that abalones with heavier shells are likely to have a higher number of rings.

3. An abalone's length, height, shucked weight, and viscera weight have a negative effect on the number of rings. For every unit increase in the logarithm of the length, the logarithm of the number of rings decreases by approximately 0.45491, indicating that abalones with longer shells tend to have fewer rings. For every unit increase in the height of the abalone, the logarithm of the number of rings decreases by approximately 1.32842, suggesting that taller abalones are likely to have fewer rings. For every unit increase in the shucked weight and viscera weight of abalone, the logarithm of the number of rings decreases by 0.42734 and 0.08529, respectively. This suggests that abalones with higher shucked and viscera weights are likely to have fewer rings.

References

Department of Fisheries, Western Australia. (n.d.). Abalone. Retrieved from https://www.fish.wa.gov.au/Documents/recreational_fishing/fact_sheets/fact_sheet_abalone.pdf

Mohammed, G. R., Abu Shbikah, J. R., & Al-Zamili, M. M. (2021). Age of Abalone Prediction from Physical Measurements Using ANN. *International Journal of Academic Engineering Research (IJAER)*, 5(4), 1-7. Retrieved from http://dstore.alazhar.edu.ps/xmlui/bitstream/handle/123456789/2710/IJAER210401.pdf?sequence=1&isAllowed=y