Abalone Age Prediction: A Linear Regression Approach

Saif Ali Athyaab - 23810756

1 Executive Summary

An abalone is a marine animal belonging to the biological family of shellfish. Its presence in a select few regions around the world coupled with intensive localised farming and marketing efforts has garnered a niche consumer base. A common problem faced by the producers is concerned with an accurate estimation of the age of an abalone. The aim of this analysis is to mathematically model the estimation of its age using other physical measurements of the abalone. A dataset collected by researchers (Nash et. al) in 1994 [1] is used to aid in the analysis of this report. There are 4177 observations recorded in the dataset exposited by 8 variables, 3 of which explain the visual attributes such as length, shell diameter and height. Traditionally, the ageing process involved cutting each abalone, staining it in a chemical solution, and then closely examining for the number of rings. Arriving at the age of an abalone involved adding a constant value of 1.5 to the number of rings. The cutting process introduced the remaining 4 variables to the dataset – the whole weight, weight after shucking out the meat from the shell, gut weight post-bleeding and shell weight after drying. Major findings include the presence of some invalid data points and the selected model with accuracy of 62%.

1.1 Introduction

It is of utmost importance to predict the age of an abalone accurately. Reasons include:

- Limited quantity and seasonal nature of its growth.
- Notoriously juvenile and irregular growth behaviour even among the same species [6].
- International market standards [3]
- Prevention of over-fishing/farming, while also maintaining distribution of sizes in the fishing environment [6]
- Timing of the catch to ensure optimal maintenance costs [3]

Various methods in the past have been employed to determine the age of an abalone including, econometrics [6], neural networks [4,5], stable oxygen isotopes [2], logistical model [3]. While these are all equally intensive and useful methods, there is a need to model the other variables in a simple manner. Siddeek and Johnson [8] performed a growth parameter estimate using length-frequency data for Omani abalones. Non-Linear Least Square Fitting method for fitting the growth curve to modal length at-age data was demonstrated. Another similar method (Tarbath and Officer, 2003) involving growth curves was fitted by non-linear regression of age-length couplets using the von Bertalanffy growth function (VBGF) [9]. The dataset is specified below highlighting the type of each variable along with a set of possible values or range.

Table 1 - Dataset Description

Attribute	Type	Values
Sex	Categorical	Male, Female, Infant
Length	Numerical	0.07 - 0.81 (mm)
Diameter	Numerical	0.05 - 0.65 (mm)
Height	Numerical	0.00 - 1.13 (mm)
Whole Weight	Numerical	0.002 - 2.82 (g)
Shucked Weight	Numerical	0.001 - 1.48 (g)
Visceral Weight	Numerical	0.0005 - 0.76 (g)

Attribute	Type	Values
Shell Weight	Numerical	0.0015 - 1.005 (g)
Rings	Numerical	1 - 29

2 Methodology

Linear Regression is the main statistical method employed to predict the number of rings. The other variables of the dataset are used as predictor variables to generate the response variable - Rings. R is the tool used to make all calculations including descriptive statistics, model fitting, model diagnostics and examination of normality plots and histograms. The analysis started off with examining a simple linear model factoring in all the variables.

$$Rings = \beta_0 Sex + \beta_1 Length + \beta_2 Height + \beta_3 Diameter + \beta_4 Wholewt + \beta_5 Shuckedwt + \beta_6 Viscerawt + \beta_7 Shellwt + \epsilon$$

The data is assumed to have a normal distribution $\epsilon \sim N(0, \sigma^2)$

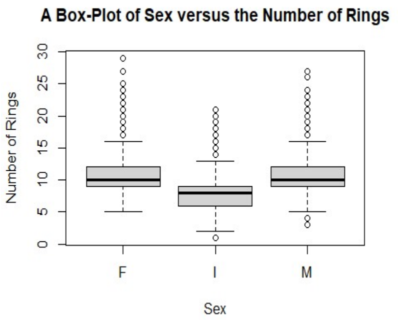
Next, StepAIC was used to iteratively reduce the full model into one consisting of mostly statistically significant terms. Following this, two-way interaction models were explored before finally arriving at a simpler one-variable interaction model. The metrics used to measure the suitability of the model include Standard residual error, R-squared, Adjusted R-squared and the AIC comparison.

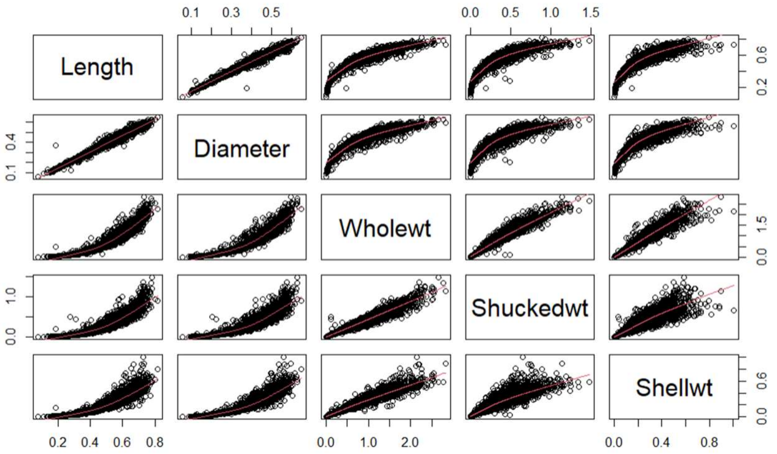
3 Results

The analysis can be divided into 2 main parts. Part 1 discusses initial exploratory data analysis followed by Part 2 which is the actual model fitting.

3.1 Part 1

During the exploratory phase, the main focus was to get a thorough understanding of the data itself. The first observation was that regarding the structure of the dataset, which consisted one categorical variable with 3 factors - M, F and I. The remaining ones were numerical variables including the response variable - Rings. Secondly, descriptive statistics were explored to identify any missing or invalid values. A simple check for NA values was also performed. It was discovered that the minimum height was 0, which should be a mistake. The relevant observations (just 2) were identified and removed from the dataset. A boxplot was plotted to identify any outliers present. It can be observed that around 10 are present in each Sex. Following this, a pair-wise correlation plot was plotted between some of the variables. During this iterative process, it was observed the Diameter variable represents a curved and not a straight-line correlation.





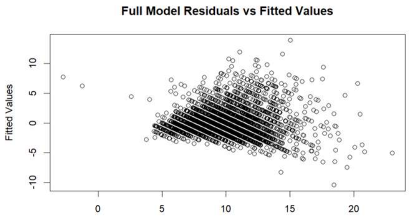
3.2 Part 2

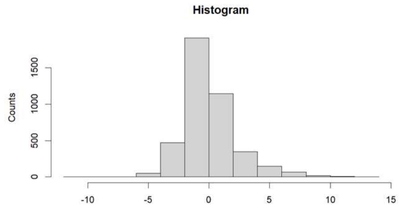
3.2.1 Standard Linear Regression

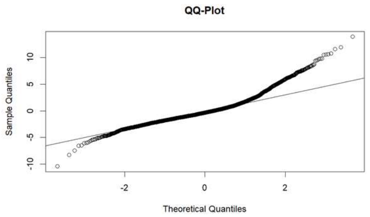
As briefly touched upon in the preceeding section, model fitting commenced by modelling all variables. The fitted model equation in this case is:

$$\log(\widehat{Rings}) = 1.34 - 0.09\widehat{Infant} + 0.008\widehat{Male} + 0.53\widehat{Length} + 1.41\widehat{Diameter} + 1.21\widehat{Height} + 0.60\widehat{Wholewt} - 1.65\widehat{Shuckedwt} - 0.83\widehat{Viscerawt} + 0.61\widehat{Shellwt}$$

To check our assumptions regarding normality, a residuals plot, histogram and QQ plot were plotted.


Full Model Residuals vs Fitted Values


Histogram


QQ-Plot

- The residuals plot seemed to be fanning towards higher ring sizes without log transformation.
- The histogram, although normally distributed, is slightly skewed towards the right.
- This is highlighted in the QQ plot which has its head straying away from the fitted line.

3.2.2 Interaction Terms

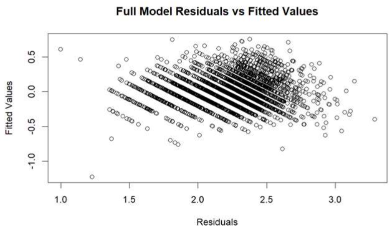
A two-way interaction between all variables was modelled resulting in 54 combinations predictors. StepAIC was again employed to intelligently reduce the predictors. A 24% reduction in the number of predictors was observed. An enviable 98% accuracy was achieved at the cost of model complexity.

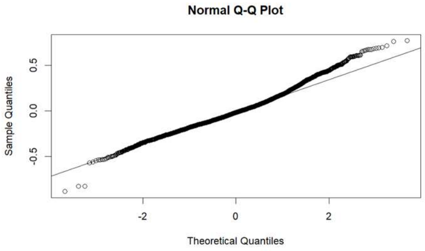
3.2.3 Interaction Model with only the Sex variable

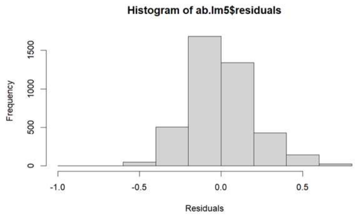
Drawing from the results of the interactions models, a simpler model involving only the Sex variable for the interaction with the numerical variables was finalized. After a couple of iterations through StepAIC, the final model is:

$$\log(\widehat{Rings}) = 1.99 - 0.89\widehat{Infant} - 0.37\widehat{Male} - 0.45\widehat{Length} + 1.04\widehat{Diameter} + 0.35\widehat{Height} + 0.67\widehat{Wholewt} - 1.57\widehat{Shuckedwt} - 0.45\widehat{Viscerawt} + 0.71\widehat{Shellwt} + 0.95\widehat{InfantLength} + 0.48\widehat{MaleLength} + 3.37\widehat{InfantHeight} + 1.12\widehat{MaleHeight} - 0.83\widehat{InfantViscerawt} + 0.34\widehat{MaleViscerawt}$$

Below, it can be seen that fanning is now relatively non-existent, while the histogram of residual counts and the normal qqplot seem improved. The model still feels slightly right skewed.


Full Model Residuals vs Fitted Values


Normal Q-Q Plot


Histogram of ab.lm5$residuals

3.2.4 Model Metrics

4 main metrics are used to compare the different models, the first 3 being obtained from the model summaries and the last 2 from AIC. The way to interpret each variable is as follows:

- Lower Standard Error and AIC indicate a better model.
- Higher Multiple R-squared indicates better model and higher Adjusted R-squared indicates utilizing lesser variables.

Table 2 - Model Comparison

Model Type	Standard Error	Multiple R-squared	Adjusted R-squared	Goodness of Fit (AIC)	Number of Covariates
Full Model	0.2026	0.599	0.5981	-1472.800	11
Full Model with log (Diameter)	0.197	0.6205	0.6198	-1704.968	10
Reduced Model with Sex Interaction Terms	0.1964	0.6236	0.6222	-1725.344	17
Reduced all-variable 2-way Interactions	0.03416	0.9886	0.9886	-16302.995	43

4 Discussion

During the analysis, more than 10 models have been studied including standard full models, log transformation of the response variable, transformations for the Diameter variable, and 2-way interactions. Finally a model with only the Sex variable interaction terms is selected. The decision to choose this particular one is because it performs marginally better than the baseline. Although the 2-way interaction model had a whopping 98% accuracy, it is disregarded due to complex covariates of the order 43. This is not suitable for easy dissemination. There are around 10 outliers for the Rings variable which have not been omitted for data integrity purposes. Future scope of improvements includes iteratively modelling each variable along-with suitable transformations to improve accuracy. A close examination of some variables may indicate high correlation and hence the ability to skip them.

5 References

1. Nash, Warwick J., Tracy L. Sellers, Simon R. Talbot, Andrew J. Cawthorn, and Wes B. Ford. “The population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait.” Sea Fisheries Division, Technical Report 48 (1994): p411.

2. Gurney, L.J., C. Mundy, and M.C. Porteus. “Determining Age and Growth of Abalone Using Stable Oxygen Isotopes: a Tool for Fisheries Management.” Fisheries Research 72, no. 2 (2005): 353–60. https://doi.org/10.1016/j.fishres.2004.12.001.

3. Susanto, Marliadi, Mamika Ujianita Romdhini, Siti Raudhatul Kamali, and Laya Zurfani. “Logistic Model of Abalon’s Length Growth in Sekotong, West Lombok.” In AIP Conference Proceedings, Vol. 2199. Melville: American Institute of Physics, 2019. https://doi.org/10.1063/1.5141285.

4. Sahin, Egemen, Can Jozef Saul, Eran Ozsarfati, and Alper Yilmaz. “Abalone Life Phase Classification with Deep Learning.” In 2018 5th International Conference on Soft Computing & Machine Intelligence (ISCMI), 163–67. IEEE, 2018. https://doi.org/10.1109/ISCMI.2018.8703232.

5. Misman, Muhammad Faiz, Azurah A Samah, Nur Azni Ab Aziz, Hairudin Abdul Majid, Zuraini Ali Shah, Haslina Hashim, and Muhamad Farhin Harun. “Prediction of Abalone Age Using Regression-Based Neural Network.” In 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS), 23–28. IEEE, 2019. https://doi.org/10.1109/AiDAS47888.2019.8970983.

6. Hossain, Md Mobarak, and Md Niaz Murshed Chowdhury. “Econometric Ways to Estimate the Age and Price of Abalone.” IDEAS Working Paper Series from RePEc, 2019.

7. Prince, JD, TL Sellers, WB Ford, and SR Talbot. “A Method for Ageing the Abalone Haliotis Rubra (Mollusca : Gastropoda).” Marine and Freshwater Research 39, no. 2 (1988): 167–. https://doi.org/10.1071/MF9880167.

8. Siddeek, M. S. M., and D. W. Johnson. “Growth parameter estimates for Omani abalone (Haliotis mariae, Wood 1828) using length-frequency data.” Fisheries research 31, no. 3 (1997): 169-188. https://doi.org/10.1016/S0165-7836(97)00036-2 .

9. Tarbath, David, and Rickard A. Officer. Size limits and yield for blacklip abalone in northern Tasmania. No. 17. Tasmanian Aquaculture and Fisheries Institute, University of Tasmania, 2003.