

An Abalone-Age Investigation

470408957 | 480423142 | 490209370 | 490384806 | 490443251 |

This version was compiled on November 19, 2020

In this project we investigated two predictive models for the predicting the number of rings in an abalone which is used to find the age by adding 1.5 to the number. In order for our models to satisfy the linearity assumptions we had to perform transformations on the data set, where we then could conduct multiple linear regression through backward and forward selection using Akaike Information Criterion (AIC). The resulting model selection was further manipulated as there was strong observed collinearity with some independant variables which led to our final model.

1. Introduction

The original article for this data set was utilised to explore the age of an abalone. However, the classical method for determining age is to cut the shell, stain it, and count the number of rings under a microscope. This is a very tedious process, which provides reasoning to encompass easier ways in calculating the number of rings without manually counting the rings. We therefore aim to construct a regression model that can predict the number of rings an abalone has (calculating the age), using only an abalones physical attributes which are easily and quickly measured. To do so, linear relationships need to be considered between the variables and the value they aim to predict.

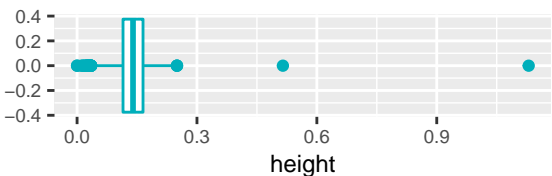
2. Data Set

This biological study looks at abalones which are a common type of sea snail, where the majority of abalone species are found in cold waters, such as off the coasts of New Zealand, South Africa, Australia, and Tasmania. The data set is provided by the University of California Irvine Machine Learning Repository which explores variables describing physical characteristics for abalones. The original data was collected by Marine Resources Division in Taroona, Tasmania, containing 4177 observations with 9 different variables which has no missing values.

2.1 Variables.

- Sex; factor; male, female or infant
- Length (mm); continuous; longest shell measurement
- Diameter (mm); continuous; perpendicular to length
- Height (mm); continuous; with meat in shell
- Whole Weight (g); continuous; whole abalone
- Shucked Weight (g); continuous; weight of meat
- Viscera Weight (g); continuous; gut weight (after bleeding)
- Shell Weight (g); continuous; after being dried
- Rings; integer; number of rings. +1.5 gives age in years

2.2 Outliers. In exploring the different variables there appeared to be two values identified as clear outliers in the height variable. Thus, concluding they are likely to be errors induced when entering the data and will be filtered out.



3. Analysis

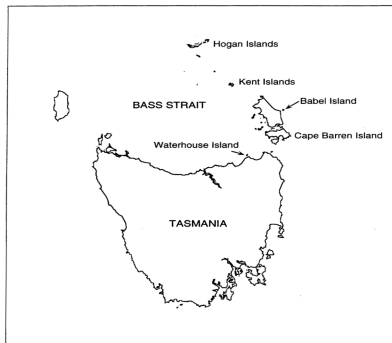
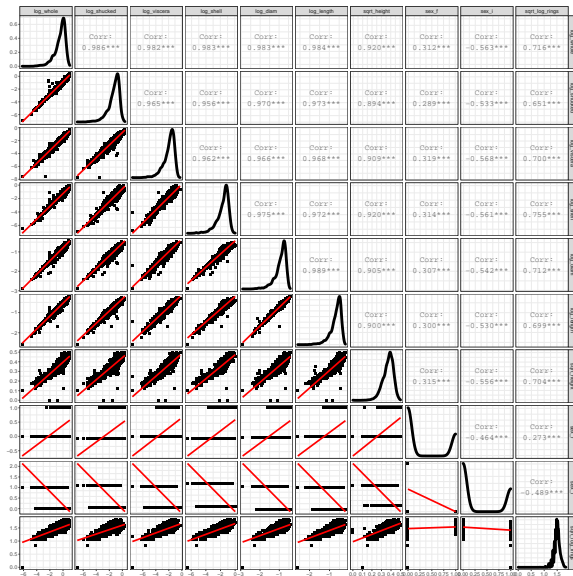
3.1 Transformations. Prior to selecting an appropriate model, with reference to appendix 1, clearly the independent variables do not possess a linear relationship with number of rings. The variables clearly rise rapidly and reach a plateau, thus it was found best to perform log transformations of the variable rings, length, diameter, weight shucked, weight viscera, weight shell and a square root transformation was applied to height and log of rings. With reference to appendix 2, the data set now adopts a linear relationship with the predictive variable, allowing for a linear regressive model to work appropriately.

3.2 Model Selection. Provided the linearity assumptions with respect to the dependent variable are satisfied, a model search was justified by the Akaike information criterion through a backward and forward variable selection. After conducting the relevant search it was found that the forward model did not include log of diameter and log of length while the backward approach included all variables, shown in the table below.

Predictors	Forward Model		Backward Model	
	Estimates	p	Estimates	p
(Intercept)	1.43	<0.001	1.45	<0.001
log shell	0.11	<0.001	0.11	<0.001
log shucked	-0.19	<0.001	-0.19	<0.001
log whole	0.19	<0.001	0.20	<0.001
sex infant	-0.02	<0.001	-0.01	<0.001
log viscera	-0.03	<0.001	-0.02	<0.001
sqrt height	0.13	0.007	0.12	0.012
log diameter			0.07	0.005
log length			-0.08	0.005
Observations	4175		1.45	
R ² / R ² adjusted	0.647 / 0.647		0.648 / 0.647	
AIC	-10882.310		-10887.886	

3.3 Assumption Checking.

Forward Residual Verse Fitted/ QQ Plot.



Locations of where data was collected.

References

- Warwick J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn *The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and Islands of Bass Strait* URL https://www.researchgate.net/profile/Warwick_Nash2/publication/287546509_The_Population_Biology_of_Abalone_Haliotis_species_in_Tasmania_I_Blacklip_Abalone_H_rubra_from_the_North_Coast_and_Islands_of_Bass_Strait/links/5d949460458515202b7bf592/The-Population-Biology-of-Abalone-Haliotis-species-in-Tasmania-I-Blacklip-Abalone-H-rubra-from-the-North-Coast-and-Islands-of-Bass-Strait.pdf.
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science. URL <http://archive.ics.uci.edu/ml>
- Allaire J, R Foundation, Wickham H, Journal of Statistical Software, Xie Y, Vaidyanathan R, Association for Computing Machinery, Boettiger C, Elsevier, Broman K, Mueller K, Quast B, Pruim R, Marwick B, Wickham C, Keyes O, Yu M (2017). *rticles: Article Formats for R Markdown*. R package version 0.4.1, URL <https://CRAN.R-project.org/package=rticles>.
- MacFarlane J (2017). *Pandoc: A Universal Document Converter*. Version 1.19.2.1, URL <http://pandoc.org>.
- Xie Y (2017). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.17, URL <https://yihui.name/knitr/>.