An Abalone-Age Investigation

| 470408957 | 480423142 | 490209370 | 490384806 | 490443251 |

Data2002 Group Project | November 2020

In this report, we investigate whether the age of *Haliotis Rubra (Blacklip Abalone)* can be estimated from external physical attributes. We constructed and evaluated two multiple linear regression models using the Akaike Information Criterion (AIC). After refinement of the selected model, we found that given two weights, three dimensions, and the sexual maturity of an abalone, we could explain 62.8% of the the variance in our target variable. Provided these measurements, predictions could in turn be untransformed to generate age estimates for abalone.

1. Introduction

Marine biologists and conservationists often study the age and growth patterns of a species in order to understand its demographics in and across various ecosystems. As a sought after commodity within the fishing industry, this is especially true of Abalone. However, the classical method for determining an abalone's age is arduous and time inefficient; counting the rings in a specially prepared shell under a microscope (Dheeru Dua and Casey Graff (2017)). We therefore aim to find a technique for estimating an abalone's age using only physical attributes which are easily and quickly measured. We will construct a multiple regression model in order to predict the number of rings an abalone has, and evaluate whether this model can effectively predict observed values and would therefore have any utility when applied to new observations.

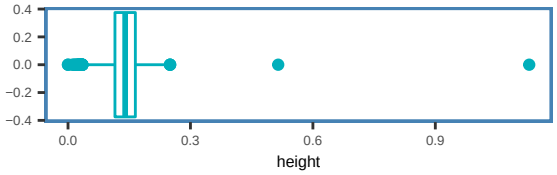
2. Data Set

This data pertains to *Haliotis Rubra*, an Australian species of abalone found predominantly in cold waters, such as off the coast of Tasmania. The relevant data were originally collected by *the Marine Resources Division in Taroona, Tasmania* to explore neural network techniques for estimating the age of abalone. The data were made available by the University of California Irvine Machine Learning Repository(Dheeru Dua and Casey Graff (2017)). The dataset contains 4177 observations upon 9 different variables, and it contains no missing values. Each variable describes some physical property - a weight, dimension, sex, ring count - of the observed abalone.

2.1 Variables.

Name	Type	Description
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. Initial data exploration reveals two clear anomalies in the height variable. These two observations are so far from the range of all other 4175 observed values that they are considered to be erroneous, and are discarded from the dataset.



3. Analysis

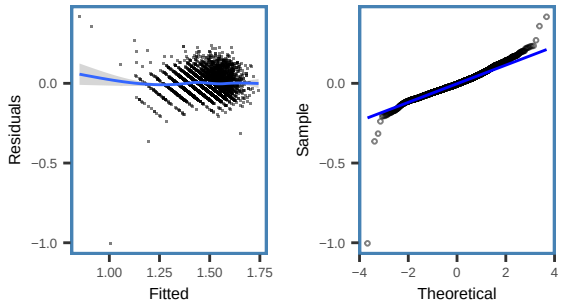
3.1 Transformations. Prior to selecting an appropriate model, we must acknowledge that the observed variables do not demonstrate a linear relationship with the observed number of rings *(Appendix 1)*, and we cannot consider sexual maturity in its current state as a ternary factor. Due to the nature of the observed curves, the ideal transformations for the predictor variables were as follows: Log transformations for length, diameter, and all the weights, and a square root transformation to height. Transforming the predicted variable (rings) to the square root of its logarithm proved ideal. Each predictor variable now adopts a linear relationship with the predictive variable *(Appendix 2)*, allowing for a linear regressive model to work appropriately. Additionally, the sexual maturity factor was encoded using a contrast matrix.

3.2 Model Selection. Having conducted our transformations, models could now be constructed. Two models were constructed; one using forward stepping variable selection, and the other using backward stepping variable selection. Both of these models were evaluated considering their R^2 and AIC values. The produced models were remarkably similar in regard to the above criteria, and the only notable difference between them is the omission of diameter and length from the forward model. The produced models are shown in the table below.

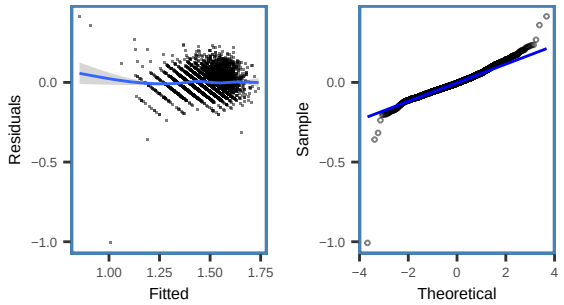
	Forward Model		Backward Model	
Predictors	Estimates	p	Estimates	p
(Intercept)	1.43	**<0.001**	1.45	**<0.001**
log shell	0.11	**<0.001**	0.11	**<0.001**
log shucked	-0.19	**<0.001**	-0.19	**<0.001**
log whole	0.19	**<0.001**	0.20	**<0.001**
sex infant	-0.02	**<0.001**	-0.01	**<0.001**
log viscera	-0.03	**<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 vs Fitted/QQ Plot.



Backward Residual vs Fitted/QQ Plot.

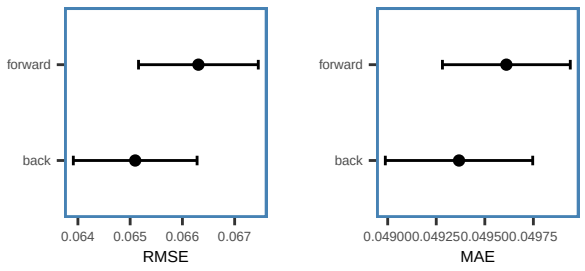


We must state and justify our assumptions - for both models - to validate any inferences made in our results.

- **Linearity**: The residual plot displays no obvious curvature for either model, thus the linearity assumption is satisfied.
- **Independence**: The data were collected across 5 different regions in the Tasman Sea *(Appendix 3)*, with no systematic or intentional collection grouping. Granted these facts, there is no reason to believe there is any dependence between observations. Hence, independence can reasonably be assumed.
- **Homoskedasticity**: For both models, the residuals do not appear to be fanning out or changing over the range of fitted values. Thus the constant error variance assumption is met.
- **Normality**: The normality assumption is at least approximately satisfied. For the QQ plot of each model, the points are reasonably close to the diagonal line. Regardless, the sample size is large enough to rely upon the central limit theorem.

4. Results

Since the models constructed using the forward and backward approach share the same adjusted R^2, the Residual Mean Square Error (RMSE) and Mean Absolute Error (MAE) were computed for each in order to determine and justify the better model. Graphs are shown below.



It is evident that the backward model is the better model, as it has a lower RMSE and MAE. It is worth noting that the p-values for all the original variables are statistically significant, excepting one of the sexual maturity factors produced from the dummy coding contrast matrix. The p-value for `sex_f` was consistent with the null hypothesis that the **gender** of the abalone is immaterial, while `sex_i` was significant, indicating that the sexual **maturity** of the abalone was meaningful.

It must be conceded that there is apparent multicollinearity within the dataset *(Appendix 4)*; which may reduce the precision of the estimate coefficients and lessen the statistical power of the model. This multicollinearity is to be expected. Living organisms tend to grow physically as they age, with a rate that decreases over time. Naturally, our measured values display that trend.

It is a challenge to address multicollinearity within a data set where all the variables are significant. The optimal solution is to omit collinear variables which are already well represented by similar measurements. In our specific case, this was a number of the weight variables. The two more significant weight variables - according to the standardized regression coefficients - were shucked weight and whole weight;

log whole	log shucked	log viscera	log shell
1.4790451	−1.5000279	−0.1885621	0.8269786
log diam	**log length**	**sqrt height**	**sex infant**
0.19408468	−0.19821088	0.06175294	−0.06326012

Thus the viscera weight and shell weight were ignored in our final model;

$$\widehat{\sqrt{log(rings)}} = 1.330 + 0.297log(whole) - 0.243log(shucked) + 0.153log(diameter) - 0.079log(length) + 0.205\sqrt{height} - 0.013Sex_{infant}$$

Our model can predict the square root of the log of the number of rings with 62.8% explainable variance when using all the provided variables, making for a respectable regressive model.

5. Discussion and Conclusion

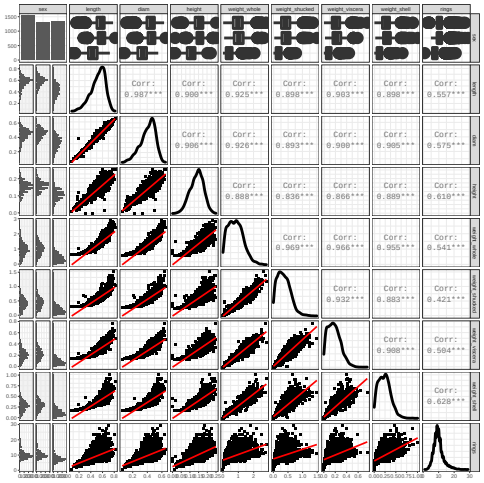
We have constructed a model that will approximate an abalone's age from easily measured attributes - a useful tool when monitoring large marine ecosystems, where research time is far better spent collecting and analysing observations than counting rings.

5.1 Limitations.

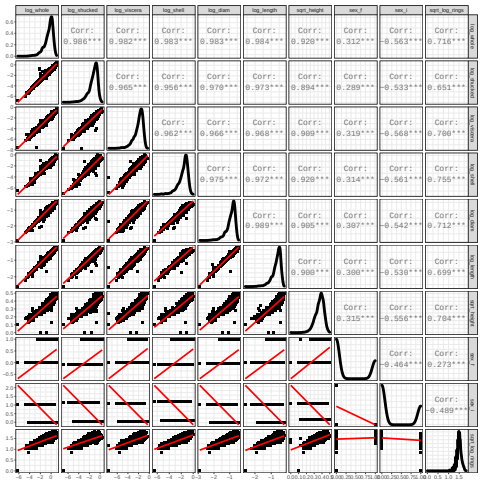
- Our data only pertains to *Haliotis Rubra*. The model does not account for species, and cannot claim to perform generally among Haliotes. Any conservational or environmental inferences are thus limited.
- As noted above, there is high collinearity among the weight variables, and together this reduces the usefulness of each. It would perhaps be more profitable to forego one of these measurements in favour of another that would add more breadth to our profile of the abalone.
- The data were only collected from waters surrounding Tasmania *(Appendix 3)*. Although Blacklip Abalone **is** prevalent in these waters, they are found in coastal waters reaching all the way from lower NSW to lower WA. This restricts the usefulness of the model, since it can only be used with confidence for Abalone in Tasmanian waters - a portion of a much larger population.

6. Appendix

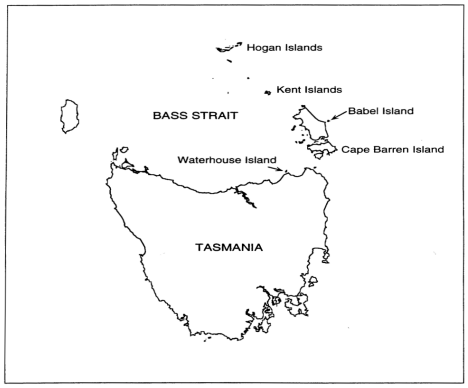
Appendix 1: Correlation matrix of initial dataset



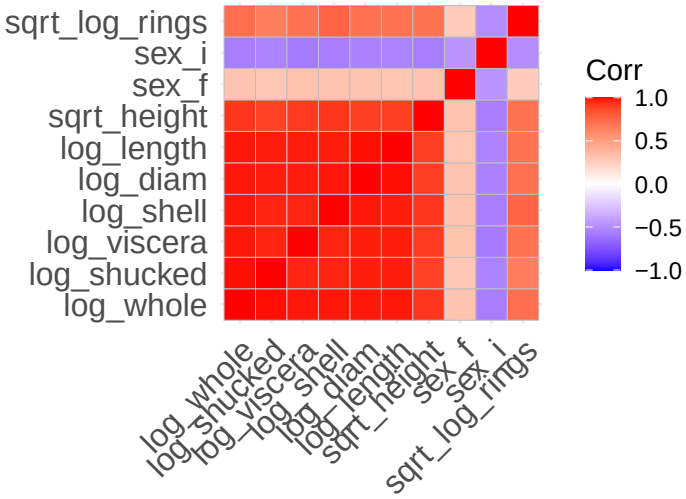
Appendix 2: Correlation matrix of transformed variables



Appendix 3: Data Collection Sites (Warwick *et al.* (1994))



Appendix 4: Correlation Matrix



References

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/.

Karl W. Broman (2015). R/qtlcharts: interactive graphics for quantitative trait locus mapping. *G*enetics, 199:359–361, URL http://www.genetics.org/content/genetics/199/2/359.full.pdf.

Dheeru Dua and Casey Graff (2017). UCI machine learning repository, URL https://archive.ics.uci.edu/ml/datasets/abalone.

Dirk Eddelbuettel and James Joseph Balamuta (August 2017). Extending *R* with *C++*: A brief introduction to *Rcpp*. *P*eerJ Preprints, 5:e3188v1, URL https://doi.org/10.7287/peerj.preprints.3188v1.

Max Kuhn (2020). *c*aret: Classification and Regression Training. R package version 6.0-86, URL https://CRAN.R-project.org/package=caret.

Daniel Lüdecke (2020). *sj*Plot: Data Visualization for Statistics in Social Science. R package version 2.8.6, URL https://CRAN.R-project.org/package=sjPlot.

Warwick Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn, and W.B. Ford (1994). 7he population biology of abalone (haliotis species) in tasmania. i. blacklip abalone (h. rubra) from the north coast and islands of bass strait. *S*ea Fisheries Division, Technical Report No, 48, URL https://www.researchgate.net/publication/287546509_7he_Population_Biology_of_Abalone_Haliotis_species_in_Tasmania_I_Blacklip_Abalone_H_rubra_from_the_North_Coast_and_Islands_of_Bass_Strait

Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Jason Crowley(2020). *G*Gally: Extension to 'ggplot2'. R package version 2.0.0, URL https://CRAN.R-project.org/package=GGally.

Taiyun Wei and Viliam Simko (2017). *R* package "corrplot": Visualization of a Correlation Matrix. (Version 0.84), URL https://github.com/taiyun/corrplot.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani (2019). Welcome to the tidyverse. *J*ournal of Open Source Software, 4(43):1686.