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 ablone, 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.

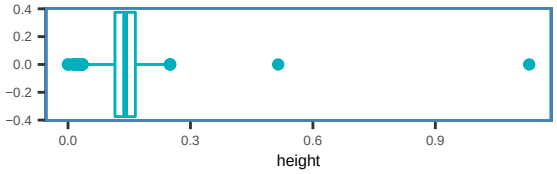
2. Data Set

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. 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. 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

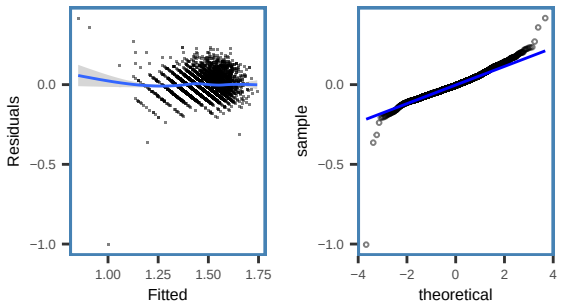
3.1 Transformations. Prior to selecting an appropriate model, with reference to appendix 1, clearly the independent variables do not possess a linear relationship with number of rings. The variables clearly rise rapidly and reach a plateau, thus it was found best to perform log transformations of the 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.

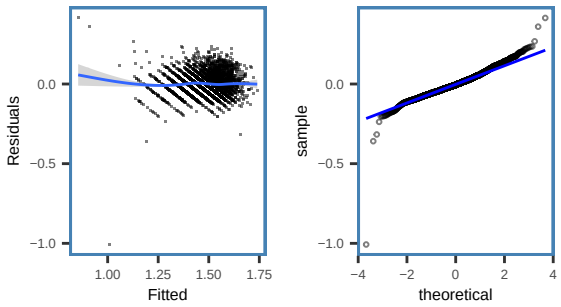
	Forward Model		Backward Model	
Predictors	Estimates	p	Estimates	p
(Intercept)	1.43	**<0.001**	1.45	**<0.001**
log shell	0.11	**<0.001**	0.11	**<0.001**
log shucked	-0.19	**<0.001**	-0.19	**<0.001**
log whole	0.19	**<0.001**	0.20	**<0.001**
sex infant	-0.02	**<0.001**	-0.01	**<0.001**
log viscera	-0.03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



Backward Residual Verse Fitted/QQ Plot.

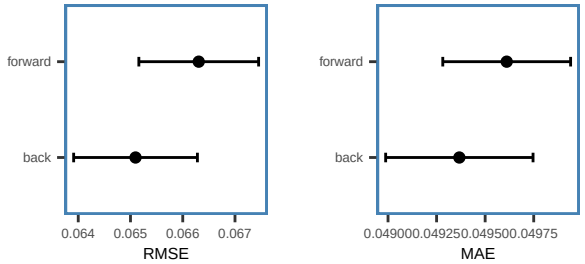


Assumption checks for both forward and backward models;

- **Linearity**: The residual vs fitted values plot indicates no obvious curvature for both thus the linearity assumption is satisfied.
- **Independence**: Referencing to appendix 3 the data was collected over 5 different regions in the Tasman Sea, hence there will be independence between the observations from the differing locations that the data was pulled from.
- **Homoskedasticity**: The residuals do not appear to be fanning out or changing over the range of the fitted values for both, thus the constant error variance assumption is met.
- **Normality**: The normality assumption is at least approximately satisfied. In the QQ plot, the points are reasonably close to the diagonal line, however the sample size is large enough to rely upon the central limit theorem ensuring the inferences are approximately valid.

4. Results

Provided both the Forward and Backward approach resulted in the same adjusted R^2, the RMSE and MAE were computed to justify which model was the better approach, shown in the graph below.



Thus, it is clearly seen that the Backward model is the better approach as it has a lower Residual Mean Square Error and Mean Absolute Error, however referring to appendix 4 there appears to be significant multicollinearity within the dataset which may reduce the precision of the estimate coefficients, lessening the statistical power. This was expected as living organisms tend to grow with age, hence the weight and height will all grow at a rapid rate til it eventually slows down at a certain point in the life.

Additionally, the p values for the model all are statistically significant except the sex factor where p-value of sex_f not providing enough evidence to reject the null hypothesis that coefficient is equal to zero, interpreting as to whether the abalone is adult or infant.

Furthermore, with strong multicollinearity within the data set and all the variables are statistically significant the optimal solution was to remove some of the highly correlated independent variables of weight. It was found that the two more significant weight variables were shucked weight and whole weight, based on the standardized regression coefficients, shown below.

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 vere two were removed which led us to our final model as follows;

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

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

5. Discussion/Conclusion

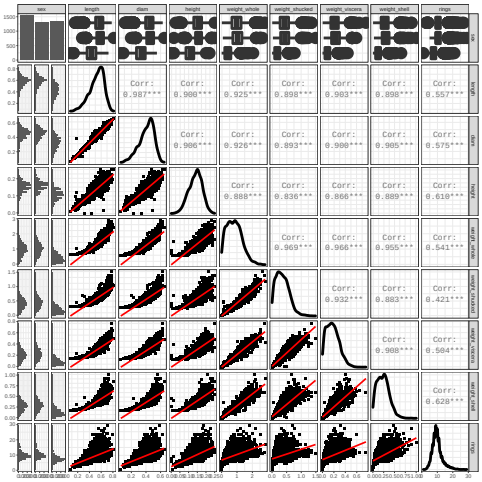
Conducting this analysis was fruitful, where the results demonstrated that we can indeed construct a model that will approximate an Abalone's age without needing any arduous ring-counting. Such a tool is very useful for our introductory scenario, when monitoring large marine ecosystems, in which, research time is better spent collecting and analysing observations rather than counting rings.

5.1 Limitations.

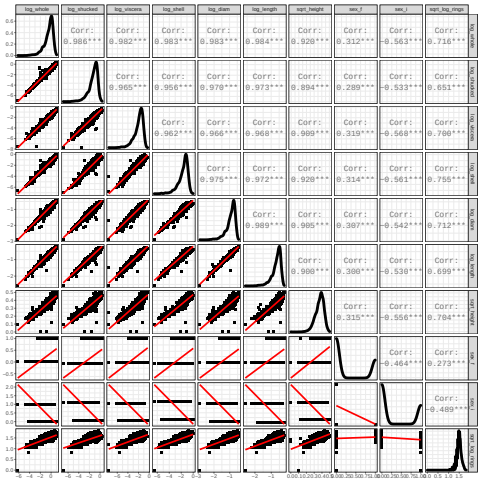
- Our data only pertains to *Haliotis Rubra*, and so we cannot claim that our model accounts for species, or will even perform generally among Haliotes. Any conservational or environmental inferences will be limited as such.
- The provided variables were not necessarily equally useful in the model. There is high co-linearity among the weight variables, and this reduces the weight and usefulness of each. In considering future research, it would be more profitable to forego one of these measurements in favour of another that would add more breadth to our profile of the abalone - i.e. depth found or total volume.
- The data was only collected around the Tasman Sea rendering the model to potentially become unreliable in differing locations as they are found all over the globe, where factors from the surrounding environment migth influence the age.

6. Appendix

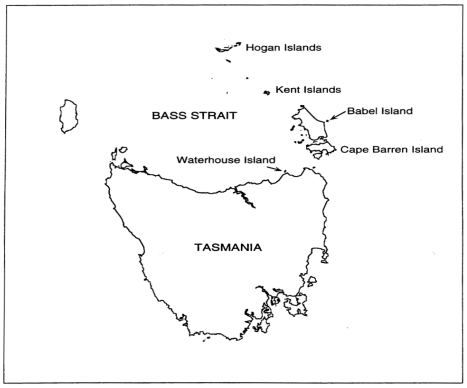
Appendix 1: Correlation matrix of initial dataset



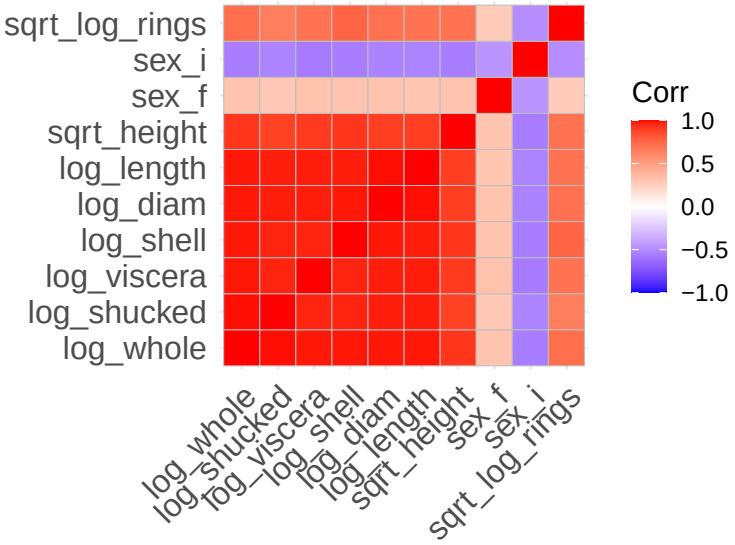
Appendix 2: Correlation matrix of transformed variables



Appendix 3: Locations of where data was collected (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). *T*he 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.