Modelling the Age of Abalone

Phua, Benjamin
451032759

Abstract

The Blacklip Abalone, scientific name Haliotis rubra, is common to several parts of Australia.[1] They are fished recreationally and also farmed. A data set provided by University of California Irvine, originally from the Department of Primary Industry and Fisheries, Tasmania, consists of several physical characteristics, including the number of rings of the Blacklip Abalone shell. The number of rings acts as proxy for the age of the Abalone. We set out to determine and compare both the best and most practical model of age of the abalone using multiple regression. This could then be used to develop and enforce recreational fishing regulations and also to maintain helathy farmed populations. Beginning with a full model, we employ step backward model selection using the Akaike information criterion (AIC). This is compared with a step forward AIC and also a practical model using easy to measure explantory variables. The practical model first removes highly correlated variables using a test for multicollinearity. We then remove difficult to measure explantory variables and test the performance of the resultant model. Our analysis finds that a practical model performs well compared to the stepwise model and can form the basis for developing best harvesting practices both recreationally and for the Blacklip Abalone Farming industry.

1 Introduction

Abalone meat is used as a food delicacy, its shells are a foundation for certain types of jewellery and it functions ecologically to stablize kelp forrests and algae in its rocky reef habitat. Blacklip Abalone reach sexual maturity after three to six years and spawning occurs between Spring and Autumn.[2] Given the importance of a healthy Abalone population, both in the wild and when farmed, it is crucial to develop, apply and regulate best harvesting practices to maintain healthy Abalone populations. With this in mind, we question how well an unrestricted model using all explanatory variables predicts the age of the Blacklip Abalone. An attempt is then made to build a practical model that can be deployed by farmers and regulators alike that provides a similar level of predictive power but doesn't necessitate killing the Abalone or require the use of specialized equipment.

2 Data Set

The data is provided by the Machine Leanring Repository at University of California Irvine. It was orginally obtained from the Department of Primary Industry and Fisheries, Tasmania. The data summary is shown in Table 1.

Variable	Type	Dimension	Description
Sex	nominal		M, F and I (infant)
Length	continuous	mm	Longest shell measurement
Diameter	continuous	mm	perpendicular to length
Height	continuous	mm	with meat in shell
Whole_weight	continuous	g	whole abalone
Shucked_weight	continuous	g	weight of meat
Viscera_weight	continuous	g	gut weight (after bleeding)
Shell_weight	continuous	g	after being dried
Rings	integer		+ 1.5 gives age in years

Table 1: Description of Data Set

After assigning new variable names to each column, we checked the assumption for multicollinearity. Linear regression assumes that there is little or no multicollinearity in the data; however, from the ggpairs output, most of the variables were relatively highly correlated with one another. The variables with few of the highest correlations were between length, height, and diameter, as well as whole weight and the other weight predictors. We will use this information later to build a practical model.

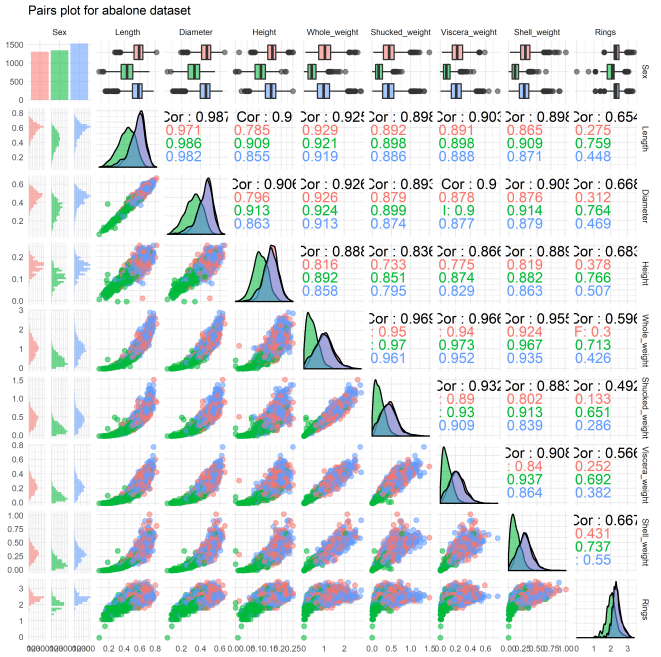

Figure 1: GGPairs showing high colinearity

3 Analysis

The first step is to determine an appropriate model for multiple regression of the data. We use the AIC model selection method in the backwards direction to achieve this. We begin with the full model and remove the least informative variable after every iteration. Crucially, we undergo a log transformation of the dependent variable

Rings to ensure normality and homoscedasticity assumptions are held for regression.

(Intercept)	1.30
SexI	-0.09
SexM	0.01
Length	0.46
Diameter	1.21
Height	2.63
Whole_weight	0.60
Shucked_weight	-1.61
Viscera_weight	-0.90
Shell_weight	0.49

Table 2: Stepback AIC coefficients

The model determined that the following independent variables `Sex`, `Diameter`, `Height`, `Shucked_weight` and `Viscera_weight` all have a relationship with the `log(Rings)`. Their respective p-values are smaller than the critical value and thus we reject the null hypothesis of no relationship with the `log(Rings)` variable. Interestingly enough, the degrees of freedom for sex is 2, meaning infants have a more significant effect on the number of rings than males with respect to females. Looking at the qqnorm plot, most of the points are close to the 45 degree line. Hence the normality assumption holds. Furthermore, the residual points are symmetrically distributed above and below the zero line and their spread is fairly constant. Hence the linearity and homoscedasticity assumptions hold as well. Our r-square value is 0.60 which is good as 60% of the `log(Rings)` variables can be explained by our constructed model.

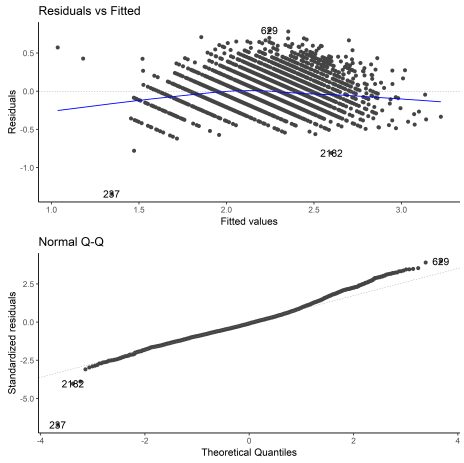

Figure 2: Variance and Normality of Log-Rings

4 Results

However, the variables chosen by the model, including `Sex`, `Shucked_weight` and `Viscera_weight` can only be determined after cracking open the abalone. This defeats the purpose of the model, as we want to predict as accurate as possible the age of the abalone without opening it to ensure sustainable and efficient farming practices. Thus, a more practical model is chosen with two parameters which are easy to measure without damaging the abalone: weight and diameter.

`log(Rings) = 1.26 + -0.10 x_Diameter + 2.61 x_Weight` has a lower r^2 value 0.45 (lower than 0.57). However, this is still a very good result as only two variables are used compared to 8 determined by the AIC and yet almost half of the log(Rings) variables can be explained by this model. In addition, from plotting the residual and qqplot of the practical model, the normality, linearity and homoscedasticity assumptions all hold as well.

Performing 10-fold cross validation on both the full model and the practical model (with only whole weight and diameter), we observe a RMSE of 1.60 and 1.70, respectively. This gives further validation to our practical model, as we only observed a reduction of out-of-sample performance by less than 10%.

5 Discussion & Conclusion

Our analysis is limited by the need to preserve the life of abalone such that other indicators of age such as shucked weight and viscera weight cannot be easily measured. As abalone production is shifting to aquaculture, where conditions are less volatile as compared to abalone in the wild, and where farmers could control the enviroment and nutriets abalone are exposed to, we should expect to be able to better predict the age of abalone. For abalone living in the wild, data on weather patterns, water PH levels, location, and food availability could be collected, and could provide a better indication of the age of abalone.

Although the full model gives the highest r^2 value, the presence of colinearity in our data set allows us to generate a simplified model that would also be most practical to abalone farmers and regulators. As whole weight and diameter can easily be measured without harming the life of abalone, our practical model is a fairly good predictor of the age of abalone. The interpretation of the r^2 value given by the model indicates that we can predict the number of rings in abalone to a precision of plus/minus 1.7 rings, which for commercial or regulatory purposes, should be deemed sufficient.

References

[1] Abalone. (2018). Dpipwe.tas.gov.au. Retrieved 27 October 2018, from `https://dpipwe.tas.gov.au/sea-fishing-aquaculture/community-resources/fish-facts/abalone-blacklip`

[2] Wild Fisheries Research Program. (2018). Dpi.nsw.gov.au. Retrieved 27 October 2018, from `https://www.dpi.nsw.gov.au/__data/assets/pdf_file/0009/375858/BlacklipAbalone.pdf`