STAT2402: Analysis of Observations
Assignment 1 Report

R. Nazim Khan*

September 22, 2023

Executive Summary

This report is on modelling the number of rings on the shell of abalone on other physical measurements. A total of 4177 data records were available, including the Sex of the abalone, length of shell, diameter of shell, shell height, weight of abalone, weight of meat without the shell, gut weight (after bleeding), shell weight after drying, and the number of rings. The model for the log $Rings$ was the best. A model with fewer interaction terms was preferred over one with several interactions between continuous variables. Although the latter was better from a statistical point of view, the difference between the fits of the two models was considered small. The selected model showed that the log of the number of rings had a negative interaction between lengths and diameters, and was lower for higher shucked weight and viscera weight. Infants with greater height had higher log of the number of rings. Shell weight and whole weight had positive correlations with log of the number of rings. Finally, as log of the number of rings increased with length of the abalone.

1 Introduction

Abalone are a type of snail, with a single flat shell [1]. It attaches itself to rocky surfaces by its foot, which it also uses to crawl around for food sources. Abalone is highly regarded as a delicacy, and a rich and nutritious food source, and the commercial value is higher for the older abalone [3, 4]. The age of abalone is related to the number of rings in the inner shell. Usually one ring is formed each year citemehta. One way to acertain the number of rings is to cut through the shell, polish and stain it and count the number of rings under a microscope. This is time consuming, complex and expensive. Adding 1.5 to the number of rings provides a good approximation to the age of the abalone.

This makes modelling to estimate the number of rings an attractive option. Several different types of models have been reported in the literature. These are all based on the publically available data set by Warwick et al. [7]. Mehta [4] provides a review of the problem of predicting the age of abalone. These include decision trees, neural networks, and various regression models. Hossain and Chowdhury [3] Used ordinary least squares (OLS) regression and and ordered probit model. Their best OLS regression model was

$$Rings = 4.05 + 0.49 \ln \mathrm{wholeweight} + 45.12 height - 38.14 height^2 + 0.98 Female + 0.82 Male.$$

Their model suffered from non-homogeneous variance, which manifested as right skewness in the residuals. Their ordered probit model did not inlcude Sex and was unreliable.

[*] School of Mathematics and Statistics (M019), The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia. E-mail: nazim.khan@uwa.edu.au

Guo et al [2] used a linear model to estimate the number of rings. Their model had large residuals and showed evidence of heteroscedacity. Misman et al. [5] used artificial neural networks, but do not mention the effect of the variables on the number of rings.

In this article we estimate the number of rings using a linear model. This report is organised as follows. In the next section we describe the statistical methodology used, followed by the Results section. This is followed by a discussion of the findings, which are compared with the literature discussed in the Introduction.

2 Methodology

We will first explore the data by numerical and graphical summaries. Following this a linear statistical model will be fitted to the data with number of rings as response. Interaction terms will also be included, and any appropriate transformation of data will be investigated. The model will be reduced to significant terms only. A simple model that is easier to interpret is preferred to a more complex model that may be better from a statistical point of view. The final model will be interpreted to explain the dependence of the number of rings on the variables in the data.

Note that the continuous variables in the data were divided by 200 for modelling artificial neural networks. The data could be scaled back. However, this does not affect our modelling, so we chose to work with the scaled data as given.

All statistical analysis will be conducted in the `R` statistical environment [6]. Statistical significance will be taken at $\alpha = 0.05$ (5%).

3 Results

The description and summary of the variables in the data set are given in Table 1. Note the variation in the standard deviations of the variables. Plot of the variables against the number or rings also showed

Variable	Description	Summary
Sex		Female (F) = 1307, Male (M) = 1528, Infant (I) = 1342
Length (mm)	Longest shell measurement	min: 0.075, mean: 0.524, median: 0.545, max: 0.815, sd: 0.120
Diameter (mm)	perpindicular to length	min: 0.055, mean: 0.408, median: 0.425, max: 0.65, sd: 0.099
Height (mm)	With meat in shell	min: 0.010, mean: 0.140, median: 0.140, max: 1.13, sd: 0.417
Wholewt (g)	Whole abalone	min: 0.002, mean: 0.829, median: 0.800, max: 2.83, sd: 0.490
Shuckedwt (g)	Weight of meat	min: 0.001, mean: 0.359, median: 0.336, max: 1.49, sd: 0.222
Viscerawt (g)	Gut weight (after bleeding)	min: 0.0005, mean: 0.181, median: 0.171, max: 0.76, sd: 0.110
Shellwt (g)	After being dried	min: 0.002, mean: 0.239, median: 0.234, max: 1.01, sd: 0.139
Rings (count)	+1.5 gives age in years	min: 1, mean: 9.93, median: 9, max: 29, sd: 3.22

Table 1: Summary of data.

evidence of non-homogeneous variance. The plots also depicted some non-linearity, resembling a square root or log function.

A linear model was fitted to `Rings` including all second order interaction terms. The model was initially reduced using the `stepAIC` procedure from the library `MASS`. The resulting model (M1) still contained several interaction terms between continuous variables. Given the non-linearity in the data, and based on plots of residuals against fitted values, model M2 was considered inadequate. We fitted a model for log(`Rings`) against all the variables, including all second order interaction terms. This model was then reduced using stepAIC. The resulting model as further reduced by omitting all insignificant terms (model M2, AIC -2063). This was compared with model M3 (AIC -2063), obtained from M2, that contained only main effects. Clearly M2 is a better model than M3.

We next added in some interaction terms to M3, namely `Sex:Diameter, Sex:Height, Sex:Shuckedwt`. After reduction, we obtained model M4 (AIC -1705). Again M3 is better than M4.

To further select between m3 and M4, we examined plots of fits (see Figure 1). The plots show that the two models are comparable on the basis of fits. Further, the plots of residuals against fitted values are not too different. The plots of model fits against log(Rings) differ only at the lower values. Finally, the normal probability plot for Model 4 is close to the straight line except at the ends. We consider Model 4 to be adequate for this data set.

The model equation for Model 4 is

$$\begin{aligned}\log(\texttt{Rings}) = 1.79 &- 0.612 \times \texttt{SexInfant} - 0.193 \times \texttt{SexMale} + \\ &1.15 \times \texttt{Diameter} + 0.279 \times \texttt{Height} + 0.646 \times \texttt{Wholewt} - 1.57 \times \texttt{Shuckedwt} \\ &- 0.725 \times \texttt{Viscerawt} + 0.706 \times \texttt{Shellwt} \ \ + 4.02 \times \texttt{SexI:Height} \\ &1.25 \times \texttt{SexM:Height}. \end{aligned} \tag{1}$$

4 Discussion

From the model equation (1), all the variables in the data except `Length` are significant predictors of the number of rings. The main effect of `Height` is not significant, but the interaction with `Sex` is. The effects of the variables on the log of number of rings and therefore age is as follows.

1. The rate of change of the number of rings is different by Sex. The rate for females is 1.3218073, for males is 4.613561 and for infants is 57.2770489.

2. Whole weight and Shell weight have a positive effect on the number of rings. For every unit increase in Whole weight the number of rings increases by a factor 1.942547, and for every unit increase in Shell weight the number rings increases by a factor 2.0258715.

3. Shucked weight and Viscera weight have a negative effect on the number of rings. For every unit increase in Shucked weight, the number of rings decreases by a factor 0.2080452, while for every unit increase in Viscera weight the number of rings decreases by a factor 0.4843246.

4. Diameter has a positive effect on the number of rings. For every unit increase in Diameter, the number of rings increases by a factor 3.1581929.

5. For a given Height, compared to females, infants have a lower number of rings by a factor 0.5422653, and males have a lower number of rings by a factor 0.824482.

Our model is closest to that of Hossain and Chowdhury [3]. However, our results are different from theirs. In particular, their model only included Whole weight, Height and its square, and Sex as significant variables. This is mainly because they did not include any interactions in their model. Further, instead of

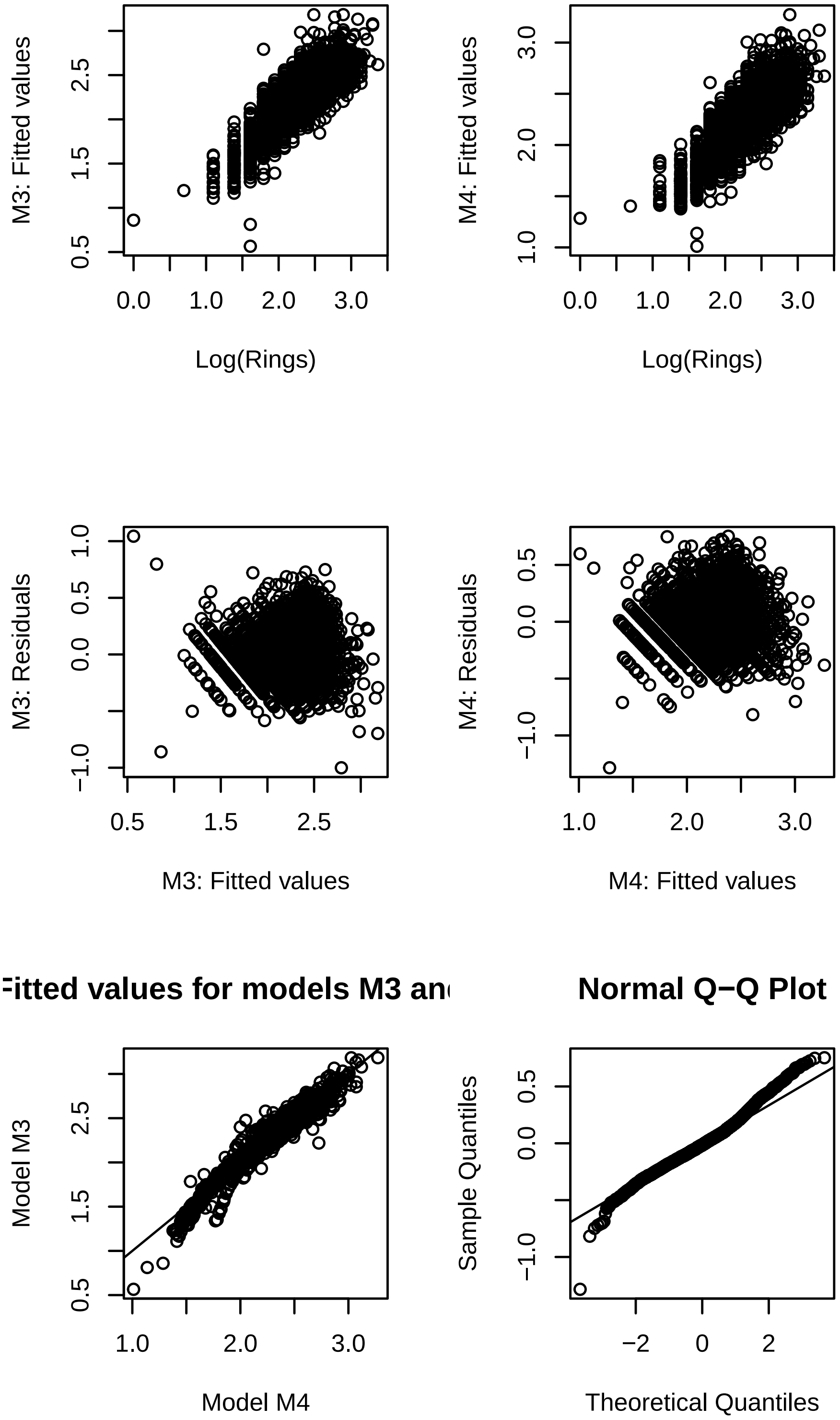

Figure 1: Comparison of Model 3 (with interactions) and Model 4 (with selected interactions).

addressing non-linearity in the data by taking the log of the response variable, they chose to include the square of Height. This does not completely account for the non-linearity with respect to other variables.

We have obtained a fairly simple model for the number of rings on an abalone based on a model using physical measurements. This model can be easily implemented for predicting the number of rings for an abalone.

References

[1] G. Anderson. Marine science: Abalone introduction. www.marinebio.net/marinescience/06future/abintro.htm, 2003. [Online; accessed 6 Sep 2023].

[2] R. Guo, Luo J., and W. Gao. A new method of measuring the age of abalone based on data visualization analysis. *Journal of Physics: Conference Series*, 1744(4):1–7, feb 2021.

[3] M. M. Hossain and M. N. M. Chowdhury. Econometric ways to estimate the age and price of abalone. *Munich Personal RePEc Archive*, 2019. Accessed 6 Sep 2023.

[4] K. Mehta. Abalone age prediction problem: A review. *JInternational Journal of Computer Applications*, 178, 2019.

[5] M. F. Misman, A. A. Samah, N. A. Aziz, H. A. majid, Z. A. Shah, H. Hashim, and M. F. Harun. Prediction of abalone age using regression-basedneural network. *2019 1st International Conference on Artificial Intelligence and Data Sciences*, 2019.

[6] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2023.

[7] J. N. Warwick, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. 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. Technical report, Sea Fisheries Division, Technical Report No. 48, 2019.