DATA2002 - Group 13 Report

**Abstract**

We aimed to see if it was possible to predict the age of abalone using various physical traits of the abalone. To do this we used multiple linear regression (MLR) on a data set of 4177 samples of abalone measurements in an attempt to create such a model. We were able to create a model with a multiple correlation coefficient of approximately 0.74 indicating a good level of prediction to predict the age of abalone.

**Introduction**

The process of determining the age of an abalone is normally a complicated process. It involves taking the shell, staining it and then counting the number of layers (rings) using a microscope in order to learn the age. In this project we aim to use MLR analysis on the provided dataset in order to determine if we can create an accurate model for predicting abalone age using various physical traits of the abalone.

**Data set**

The abalone data set consists of 4177 samples of physical attributes provided by the University of California, Irvine. The data originally came from another study, “Technical report / Division of Sea Fisheries, Tasmania”. The original data set contained had the missing values removed and the continuous variables had been scaled for use with an artificial neural network.

Each sample contains 9 measured attributes:

- Sex - M (male), F (female), I (infant)
- Length - longest shell measurement (mm)
- Diameter - perpendicular to the length (mm)
- Height - with meat in shell (mm)
- Whole weight - weight of the whole abalone (g)
- Shucked weight - weight of the abalone meat (g)
- Viscera weight - weight of the internal organs after bleeding (g)
- Shell weight - weight of the shell after being dried (g)
- Rings - number of layers in the shell, adding 1.5 gives the age in years (integer)

**Analysis**

We created a correlation heat map in order to gain a basic idea about the correlations amongst all variables and we can see that each pair of variables has a positive correlation, different weight variables have relatively high correlation with each other, length, height and diameter have high correlation with each other and rings variables has low correlation with the others. (Figure 1)

We then used GGally to create specific graphs for variable analysis. (Figure 2) is categorised by sex into 3 groups, male (blue), female (red) and infant (green). Infants have an overall lower quantity and males have the greatest quantity. (Figure 3) reveals that length has a very strong linear relationship with diameter and height indicating that it may need to be dropped. Shucked weight also has near perfect collinearity with all the weight variables. From this exploratory analysis, we need to drop some variables.

There are 5 assumptions: that all predictor variables have an approximate linear relationship with the response variable (linearity), which was assessed to be satisfied using the scatter plots with a loess line in figure 2. No perfect collinearity, which although there was high levels of collinearity between most of the variables, was only an issue for length which was removed from the final model. All the data is independent and identically distributed. The errors initially increased in variance as the value of the fitted value increased, but after using a log transformation on the response variable the errors assumed constant variance (homoscedasticity) (Figure 4). The dataset also satisfied the Central Limit Theorem (n = 4177) and the errors followed a normal distribution as seen in Figure 5. Thus, all assumptions were met.

**Modelling and Selection**

The dataset had all missing values already removed, from the exploratory analysis we noted the height observations contained some 0 values and large outliers which where cleaned from the set. We generated 4 models, a simple linear model (FIgure 6), MLR containing all predictors in the dataset (Figure 7), a backwards stepwise MLR (Figure 8) and an exhaustive search which considered all combinations of variables (Figure 9). Notably the stepwise and exhaustive search models dropped the same variable, which was length, but kept every other variable in the dataset. We performed cross validation to assess the accuracy of the model. Since we have a baseline of the simpler models, we can assess whether our model selection found any better options. We used the Mean absolute error to choose the model, as it is an appropriate indication of accuracy. We chose our model based off the lowest MAE score of 1.55, which used the following terms.

log(Rings) = Sex + Diameter + Height + Shucked Weight + Whole Weight + Shell Weight + Viscera Weight

If we look at some of the results of exhaustive search to further justify this model, we can see Mallows complexity parameter (Figure 10) and BIC (Figure 11) decreasing and then a very slight increase with the inclusion of the 8th variable which is length. Our adjusted R squared should penalise overfitting.

**Results**

We have obtained a multiple correlation coefficient of approximately 0.74 which indicates a good level of prediction. The proportion of variance, or $R^2$, was found to be roughly 0.5429 and had a minimal difference compared to the adjusted $R^2$ value which was 0.542. These are strong indications which suggest that the independent variables have directly affected the dependent variable. Furthermore, the p-values obtained for each of these independent variables were found to be well below the $p<0.5$ threshold

except for the ‘Sex M’ and ‘Sex F’ variables which did not show significant changes in the prediction of the age of abalone, and were therefore removed as a variable during the analysis of our data. ‘Sex I’ was also found to have a minimal unstandardized coefficient but had a p-value of $3.27e^{-15}$ which we then decided to use. Aside from the gender of the abalone however, all 7 other variables were calculated to have an unstandardized coefficient of larger than 7.

The final equation we have come up with to predict the age of abalone based of physical measurements is:
3.605 – (0.804 x sex I) + (0.051 x sex M) + (8.481 x Diameter) + (21.157 x Height) + (8.875 x Whole weight) – (19.531 x Shuckled weight) – (11.136 x Viscera weight) + (7.899 x Shell weight)

**Discussion and conclusion**

Abalone are considered a cultural delicacy. As such demand for farming the animal has resulted in conservational issues where overfishing has caused the population to decrease. Therefore the most useful MLR model for a marine biologist or abalone farmer would be one that does not require the animal to be dissected. This can be achieved by constructing a model that drops the variables of shucked weight, viscera weight and shell weight:

$$log(Rings) = Sex + Diameter + Height + Length + WholeWeight$$

The resulting model is less accurate than our full model with an R squared of 0.5 compared to 0.6, and has a higher MAE, 0.17 compared to 0.15 but is sufficiently accurate considering its purpose.

If we wanted to ensure conservation of the animal, we could set a critical value, where an animal would be rejected if the model calculated it was below a certain age. Alternatively a logistic regression could be more useful to give a binary answer as to whether the animal is old enough to be farmed.

**Appendix**

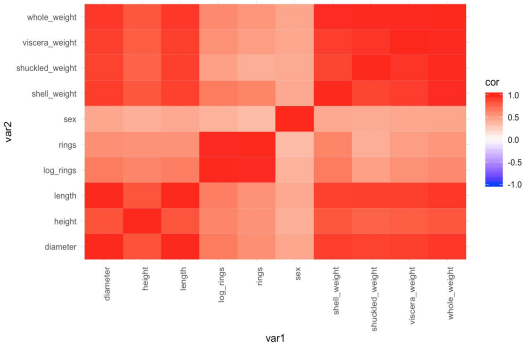


**Figure 1 - Correlation Heat Map**

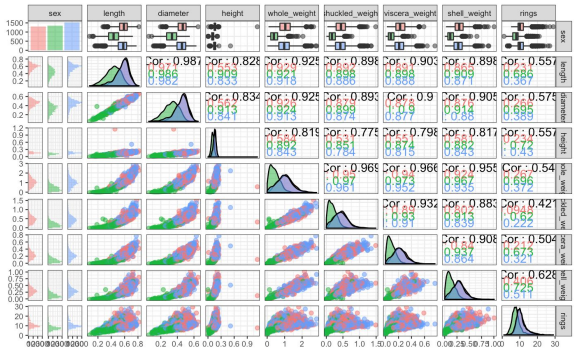

**Figure 2 - Exploratory Analysis Graph 1**

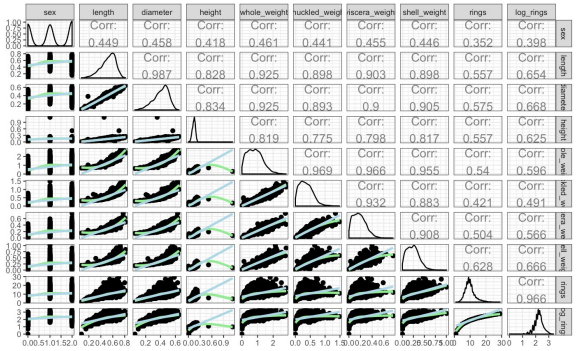

**Figure 3 - Exploratory Analysis Graph 2**

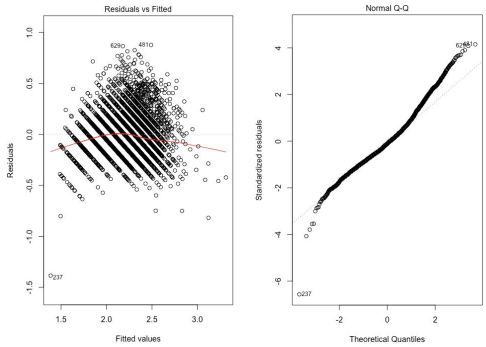


**Figure 4 & 5 - Normality/Homoscedasticity Check**

| RMSE | Rsquared | MAE |
|---|---|---|
| 2.55251 | 0.3752872 | 1.849458 |

**Figure 6 - Simple Linear Regression**

| RMSE | Rsquared | MAE |
|---|---|---|
| 2.187342 | 0.5421348 | 1.57983 |

**Figure 7 - MLR including all variables**

| RMSE | Rsquared | MAE |
|---|---|---|
| 2.181897 | 0.541999 | 1.576802 |

**Figure 8 - MLR generated using backwards stepwise**

| RMSE | Rsquared | MAE |
|---|---|---|
| 0.2011523 | 0.6040806 | 0.1550811 |

**Figure 9 - MLR generated by exhaustive search**

| RMSE | Rsquared | MAE |
|---|---|---|
| 0.2239764 | 0.5087743 | 0.1721334 |

**Figure 10 - MLR using variables that can be obtained without dissecting abelone**

GitHub Repository:
https://github.sydney.edu.au/hlin4599/DATA2002-Assignment
Dataset source:
https://archive.ics.uci.edu/ml/datasets/abalone