Abalone Age Prediction from Physical Measurements

Joseph Griffin
Lewis University
DATA550 Final Project

Abstract—**Abalones are a type of mollusk whose age can be directly determined by counting the number of rings present on the inside of the shell. This is a timely process, however, as the counting must be done under a microscope only after the shell has been cut and the rings on the inside have been stained. Thus, this paper seeks to streamline this process by avoiding the ring-counting process altogether. Utilizing physical measurements that are much easier to obtain, a model has been developed to predict the age of the animals.**

I. Introduction

The objective of this project was to develop a model to predict the age of abalones as the normal process to do so is highly time consuming. The data set on the abalones and their physical measurements was obtained from a study by Warwick J Nash et. al [2]. Both linear and polynomial regression were used as a model.

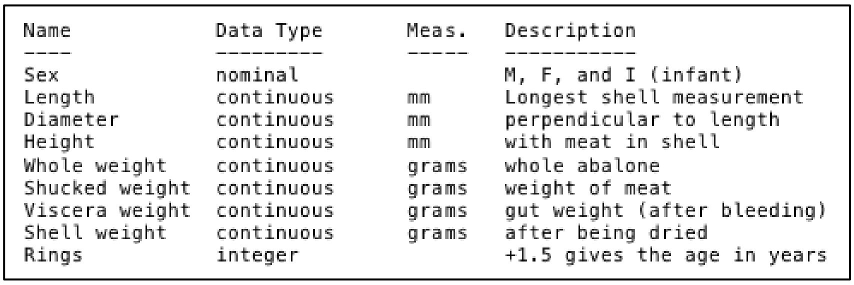
The attributes given in the data set are displayed in the table below. There was already some cleaning and preparation of the data done before it was made available to download. After these processes, there were 4177 instances remaining in the data set. All instances with missing values were removed, and ranges of continuous values were scaled using an ANN by dividing by 200. Thus, no cleaning was needed to be done. All of the given data are numerical except for the categorical variable labeled "Sex."

It is stated that the age of the abalones is positively correlated, and thus directly determined by the number of rings on the inside of their shell. The number of rings plus 1.5 years yields the age of the animal. For the sake of the project, the model only worries about the relationship between the various measurements and the number of rings. Additional details about the data set can be found at the bottom of this page in Figures 1 and 2.

II. Models / Results

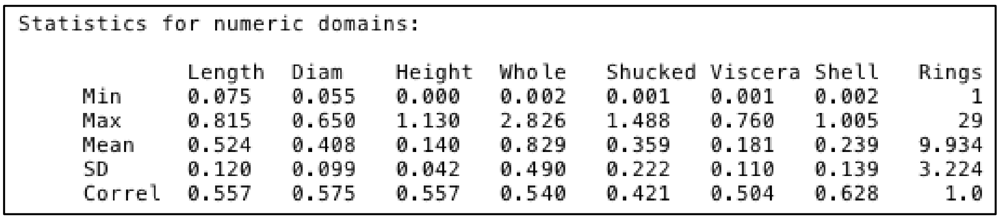
A. Linear Regression

Given the rather straight-forward nature of the data set and the goal of the project, it was clear that regression was a good first step to analyze the relationship between each of the numerical independent variables and the dependent variable. I put together a matrix of all of these scatter plots comparing the independent variables to the dependent variable (Figure 3). 250 samples are displayed in the plots. They are color coded with three different colors to refer to the last variable in the data set: 'sex.' The three categories for this variable are 'male,' 'female,' and 'infant.' 'Male' is colored red, 'female' is green, and 'infant' is blue. From these plots, some correlation became apparent. As one might expect, infant was highly correlated with younger age. However, no major difference was apparent between males and females.



```
Name            Data Type       Meas.   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      grams   whole abalone
Shucked weight  continuous      grams   weight of meat
Viscera weight  continuous      grams   gut weight (after bleeding)
Shell weight    continuous      grams   after being dried
Rings           integer                 +1.5 gives the age in years
```

Figure 1. Descriptions of all variables in the data set.



```
Statistics for numeric domains:

             Length  Diam    Height  Whole   Shucked Viscera Shell    Rings
     Min     0.075   0.055   0.000   0.002   0.001   0.001   0.002        1
     Max     0.815   0.650   1.130   2.826   1.488   0.760   1.005       29
     Mean    0.524   0.408   0.140   0.829   0.359   0.181   0.239    9.934
     SD      0.120   0.099   0.042   0.490   0.222   0.110   0.139    3.224
     Correl  0.557   0.575   0.557   0.540   0.421   0.504   0.628      1.0
```

Figure 2. Statistics for of numeric variables in the data set.

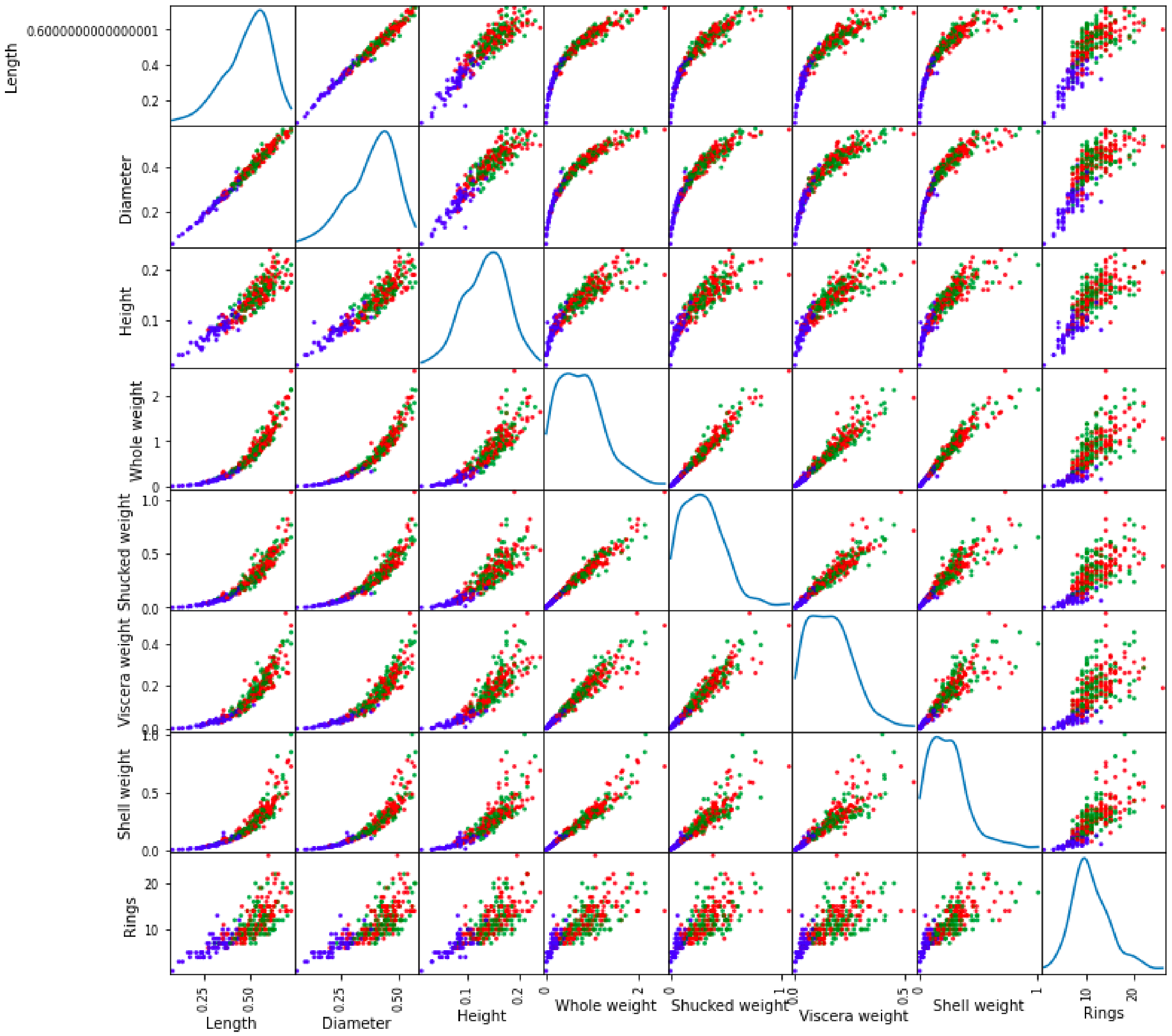


Figure 3. Matrix of linear regression plots of all variables compared to 'rings' variable. 'Infant' is blue, 'male' is red, and 'female' is green.

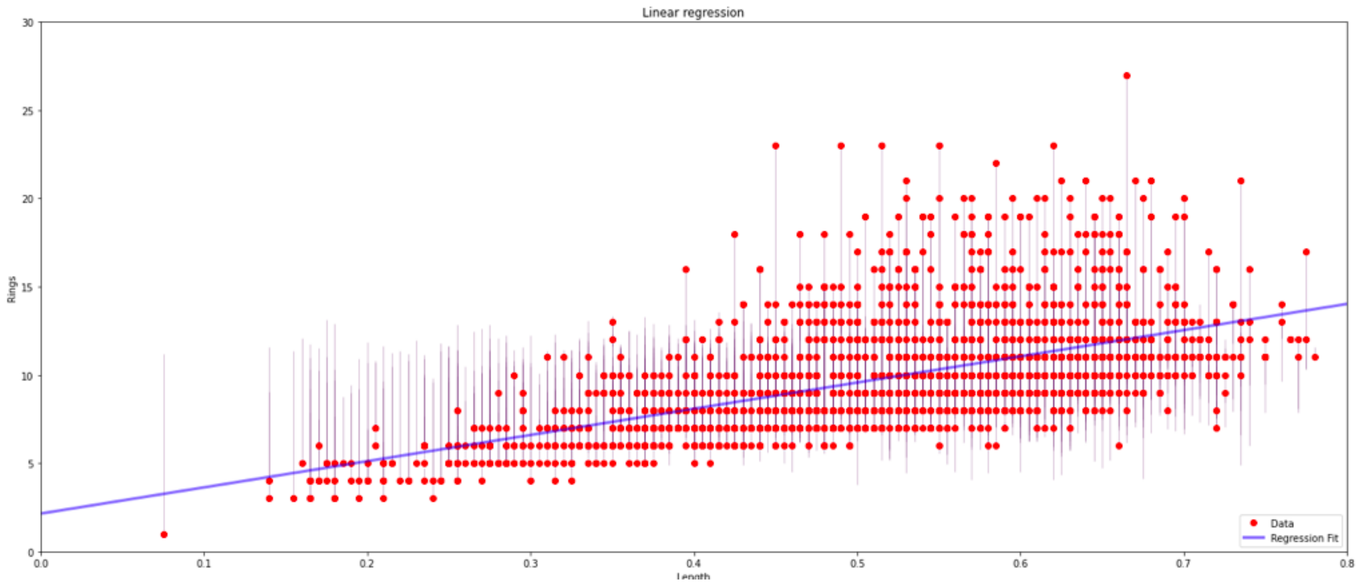


Figure 4. Linear regression plot with 'length' vs. 'rings.' Trendline is displayed in blue.

I then ran a full linear regression plot with the first variable in the data set: ‘length.’ The trendline is displayed and the variances from this trendline are shown above in Figure 4. The variance score for this plot was .303 (a score of 1 is perfect prediction).

I repeated for all the 7 independent numerical variables. Rather than displaying all of the full plots, I have put all of the variance scores into Table 1 below.

Variable	**Variance Score**
Length	.303
Diameter	.324
Height	.246
Whole Weight	.289
Shucked Weight	.172
Viscera Weight	.250
Shell Weight	.388

Table 1. Variables and their variance scores from linear regression with ‘rings’ variable.

B. Polynomial Regression

Following linear regression, polynomial regression was the next step to try to improve the model. All of the numeric variables were put into this model at first, and the variance score improved from the linear regression model immediately with a score of .514. I then continued to test different polynomial regression models experimenting with and without various variables. I was able to improve the model by removing ‘height’ and ‘shell weight’ from the polynomial regression. This resulted in a variance score of .530. Given that there are five dimensions in this model, I was not able to produce plots as I did with the linear regression plots.

III. Discussion

This model can definitely be a practical way to at least give a good estimate for the age of abalones without having to go through the long process of counting the number of rings present under a microscope. With more development, possibly from other physical measurements such as shell color, the model could certainly be improved more. One aspect that could also be improved from data that is already available is the use of the categorical variable ‘sex.’ There is a visible correlation between ‘infants’ and fewer rings. The same could be said for ‘males’ and ‘females’ having more rings. This variable was not used in the models.

This model can also be adapted based on the ease of acquiring the physical measurements. Some of the measurements are definitely easier to obtain than others, and for maximum time efficiency the model could also be run without some of the variables that are more difficult or time consuming to obtain. For example, the viscera weight refers to the gut weight after bleeding, according to the data set description file. This measurement certainly requires a bit of effort to obtain. Thus, it could be more efficient overall to skip the steps required for this measurement if the model still produces has a similar variance score.

Whether or not it would make sense to adjust the model would also depend on the application of the data. For instance, if researchers simply want to identify older abalones for another study, a less accurate model would likely suffice. However, in other situations, a highly accurate model may be more necessary.

Another method to improve the model would be to look at the characteristics that appear to have a non-linear relationship with rings. This can be seen in Figure 3. All of the weight-related variables seem to fit this profile. Further scaling or preparation of the data before running regression could possibly improve the model.

Lastly, perhaps the most intriguing result of this work was the fact that the polynomial regression model ran better without ‘shell weight.’ This variable produced the highest variance score (.388) in the initial linear regression. This could be due to the fact that there is high correlation between all four of the variables that deal with weight: 'whole weight', 'shucked weight', 'viscera weight', and 'shell weight.’ This high level of correlation can be clearly seen in the plot matrix in Figure 3.

IV. References

Much of the code used in this project came from course material provided by Dr. Mahmood Al-Khassaweneh [1]. He is the professor of the DATA550 course at Lewis University. The data used for this project originated from a study by Warwick J Nash et. al [2].

[1] Mahmood Al-Khassaweneh, DATA550 Course Material. Lewis University, 2021.

[2] Warwick J Nash, Tracy L Sellers, Simon R Talbot, Andrew J Cawthorn and Wes B Ford (1994) "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (_H. rubra_) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-328