Regression Modelling Of Abalone Age

T09ol_ontime_5

Introduction

This report had a singular aim; to determine how accurately one could predict the response variable, age of a given abalone, using predictors found in the data set. To achieve this, three tasks were created:

- Determining whether producing regression models per subpopulation (the groups infant, male, female found in the "Sex" predictor) would give significant advantages in accuracy & efficiency over modelling across the entire population.
- Determining the best models, destructive and non-destructive, for predicting abalone age; that is, determining the best model when considering only the predictor "WholeWeight" out of the weight type predictors, and determining the best model when considering all weight type predictors except "WholeWeight".
- Investigating the predictive performance of the best models and analysing the error involved in predictions across age.

Data Set

The data set came from an original (non-machine-learning) study [@warwick_j_nash_population_1994]. It contained variables outlined in *Table 1.* and had 4176 entries prior to cleaning, reduced to 4170 entries after cleaning.

Table 1: Variables Of Data Set

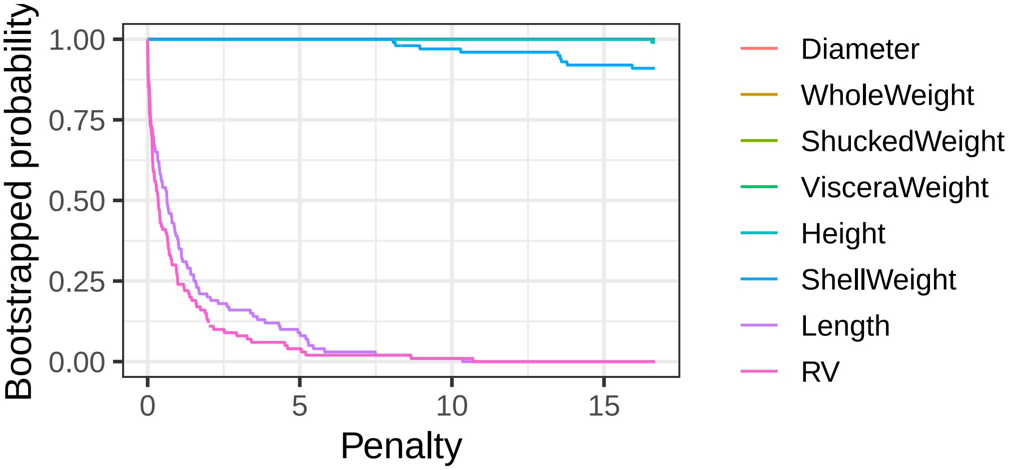
Name	Description
Sex	Infant, Male, or Female
Length	Longest shell measurement (mm)
Diameter	Perpendicular to length (mm)
Height	With meat in shell (mm)
WholeWeight	Whole abalone (g)
ShuckedWeight	Weight of meat (g)
VisceraWeight	Gut weight (after bleeding) (g)
ShellWeight	After being dried (g)
Rings	+1.5 Gives the age in years
Age	*see below

Note:

*Age was calculated from Rings in this report and not part of the original data set.

Analysis

As part of our EDA, we performed an ANOVA to compare the means of each of the independent variables in different-sex groups. As the differences in the means of each were significant across the groups, we decided to produce a model for both the full population and each group of the population selected by sex.



We performed an exhaustive search of all models to determine the best model for each group allowing for different numbers of parameters. Models using 3 or 4 predictors seemed the best compromise between the number of predictors and representation across the sexes. Using bglmnet, we determined bootstrapped selection probabilities for each of the variables, shown in the figure above, and selected the 3 to 4 most frequently occurring variables for the model of each group.

Having selected these models, we used diagnostic plots to perform assumption checking. Linearity, homoscedasticity and normality assumptions were satisfied, and high-leverage outlier points were identified and removed. When rerunning the above steps of the outlier-removed we obtained the same models. To validate the stability of the models selected, we used mplot with an adaptive fence procedure, using 80 bootstrap samples over a grid of 100 parameter values.

Results

The R^2 value for the combined population model is 0.5181, indicating that around half of the variance in the age variable was explained by our model. This is likely due to the presence of moderate amounts of noise in the data.

There is generally little difference in out-of-sample predictive ability - measured using mean absolute error and mean squared error - between models for different groups. There is little to no improvement using models per Sex versus a model ignoring Sex, although it is worth noting that the model for the infant group performs somewhat better than male and female. As the balance of sexes in the data is roughly even, this may be why the overall population model performs similar to, or better than, individual models for the male and female populations.

Discussion and conclusion

Our analysis of the data is mainly limited by the presence of the noise in the response variable. Additionally, there appears to be relatively high variance in the independent variables of diameter, height and weight within the adult population and low correlation with age, when compared with the infant population. Infant abalones are still growing, unlike adult abalones, which may explain why measures of their size are so much more strongly associated with their age.

All in all, we were able to use mplot to perform an exhaustive search of models and select a subset of independent variables which produced a stable model. The model we chose performed reasonably well at predicting values of the age of an abalone given physical measurements despite the presence of noise in the dataset.

References