Proposal to Study the Age and Sex of Abalones

Brook Assefa, Shrikant Mishra, Tom Lever

11/13/22

Introduction to Study and Data Set

According to Candy Abalone, blacklip abalone is a marine univalve mollusc containing a large muscular foot that is highly sought after for eating. Blacklip Abalone is found along Australia's southern coast. There is a tighly monitored and controlled quota on Australia wild abalone to ensure its ecological sustainability.

There is value is studying the population biology of blacklip abalones, including the relationship of age and sex of with length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight. The age of a blacklip abalone is the sum of the number of rings of the abalone and 1.5. A data set with these data are available at https://archive.ics.uci.edu/ml/datasets/Abalone. Please see attached CSV file entitled `Data_Set--Abalone_Marine_Snails--With_Column_Names.csv`.

This data was coincidentally used for a weekly assignment in our course Programming for Data Science, but no ideas were carried over from that course. That assignment dealt with learning how to use common dplyr functions, such as select and filter, and did not delve into any kind of statistical analysis of the data.

Per the UCI website listed above as well as the associated names file that comes with the data, here are the descriptions of the albalone physical characteristics: The sex of a blacklip abalone refers to either Male, Female, or Infant. The length of a blacklip abalone is the length of the longest shell. The diameter of a blacklip abalone is the measurement of the length that is perpendicular to the length variable. The height of a blacklip abalone includes the meat and the shell. The whole weight of a blacklip abalone is the weight of the entire abalone, including the meat, the shell. The shucked weight of a blacklip abalone is the weight of the meat without the shell. The viscera weight of a blacklip abalone is the gut weight after bleeding. The shell weight of a blacklip abalone is the weight of the shell (after drying the meat). The number of rings of a blacklip abalone is the difference between the age of the abalone and 1.5.

Questions of Interest

We propose to conduct data analysis of the above data set towards answering the following two questions:

1. How is the age of a blacklip abalone related to and/or predicted from the sex, length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone?
2. How is the sex of male and female blacklip abalones related to and/or predicted from the age and number of rings, length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone?

Addressing how the age of a blacklip abalone is related to and/or predicted from the length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone may be valuable in determining ways to promote the longevity of blacklip abalones and their species. Ways to promote the longevity of other abalone species may be determined. Determining the success of any population-boosting or remediation program may be enhanced.

Addressing how the sex of male and female blacklip abalones is related to and/or predicted from the length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of abalones may be

valuable in determining ways to preserve a balance of male and female abalones. Ways to promote a balance for blacklip abalone or other abalone species may be determined. Determining the success of any balance-improving or remediation program may be enhanced.

Models

We propose to conduct data analysis by studying a multiple linear regression model and a multiple logistic regression model. The response of the multiple linear regression model will be age in years; the response of the multiple logistic regression model will be adult sex (i.e., male or female).

According to Keith G. Calkins (https://www.andrews.edu/~calkins/math/edrm611/edrm05.htm), "Correlation coefficients whose magnitude[s] are between 0.9 and 1.0 indicate variables which can be considered very highly correlated. Correlation coefficients whose magnitude[s] are between 0.7 and 0.9 indicate variables which can be considered highly correlated. Correlation coefficients whose magnitude[s] are between 0.5 and 0.7 indicate variables which can be considered moderately correlated. Correlation coefficients whose magnitude[s] are between 0.3 and 0.5 indicate variables which have... low correlation[s]. Correlation coefficients whose magnitude[s] are less than 0.3 have little if any (linear) correlation."

The age and number of rings of a blacklip abalone has a moderate correlation with all variables other than sex and shucked weight of the abalone, for which correlation is low. These six predictors may be the most significant predictors; the correlation of these six predictors seems to be a good indication that these predictors could be used in a multiple linear regression model to answer our first question. That being said, each variable is correlated with every other variable. For any multiple linear regression model, the model may suffer from significant multicollinearity.

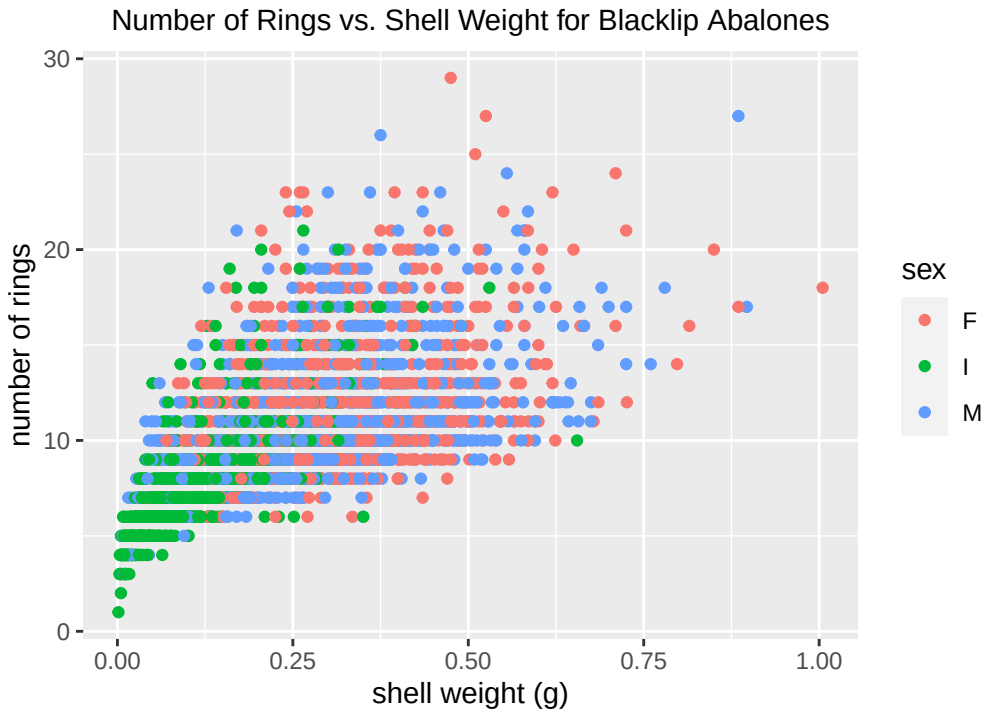
Both bidirectional and backward selection using R and an Akaike Information Criterion (AIC) suggest conducting a multiple linear regression of the age of a blacklip abalone versus the diameter, height, sex, shell weight, shucked weight, viscera weight, and whole weight of the abalone. Bidirectional selection adds predictors to an intercept-only multiple linear regression model in the following order: shell weight, shucked weight, diameter, whole weight, sex, viscera weight, and height. Shell weight may be added to the model without introducing multicollinearity. Shell weight may be the most important predictor, followed by shucked weight, diameter, whole weight, sex, viscera weight, and height, according to bidirectional selection. That being said, each variable is correlated with every other variable. For any multiple linear regression model, the model may suffer from significant multicollinearity.

The sex of a blacklip abalone has low correlation with all other variables. Forward, bidirectional, and backward selection using R and an AIC suggest conducting a multiple logistic regression of the adult sex (i.e., male or female) of a blacklip abalone versus the diameter, shucked weight, viscera weight, and height of an abalone. Diameter may be added to the linear model without introducing multicollinearity. Diameter may the most important predictor, followed by shucked weight, viscera weight, and height. That being said, each variable is correlated with every other variable. For any multiple linear regression model, the model may suffer from significant multicollinearity.

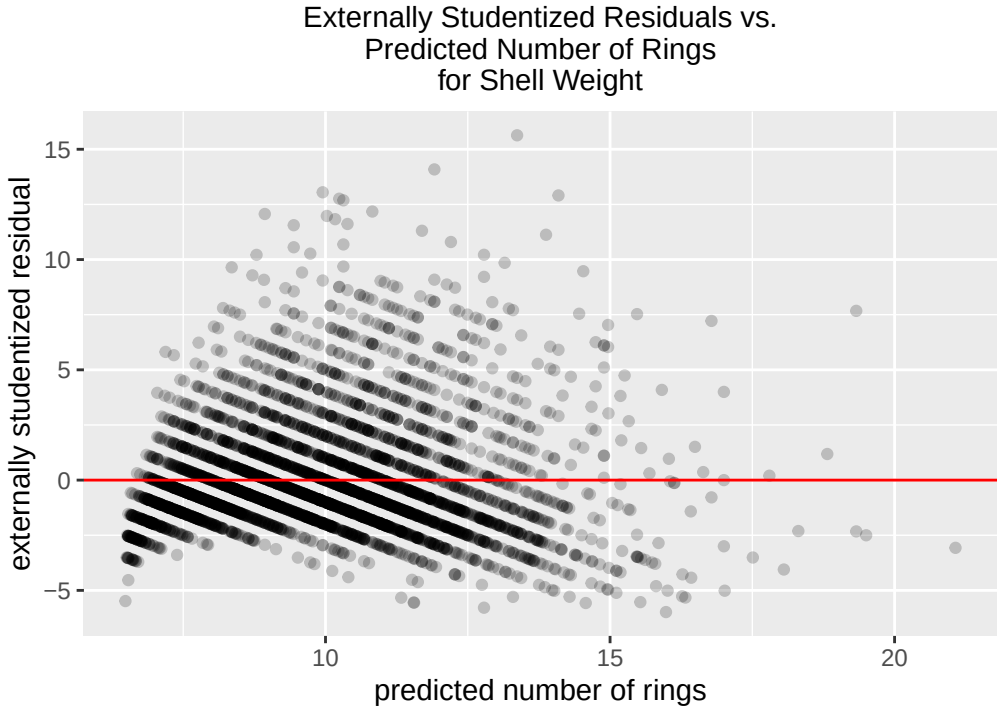
Data Visualizations

Data Visualizations for Multiple Linear Regressions of Number of Rings

A graph of number of rings versus shell weight in grams and sex for blacklip abalones is depicted below. Reasonably, the number of rings and shell weight for infant blacklip abalones are smallest. The number of rings and variance in number of rings grows at a decreasing rate as the shell weight grows. The number of rings and shell weights of male and female adult abalones seem similar. It seems that number of rings has a logarithmic or fractional power relationship with shell weight. Additionally, despite the low correlation between number of rings and sex, the number of rings seems to be affected by sex values, especially between infants and adults.

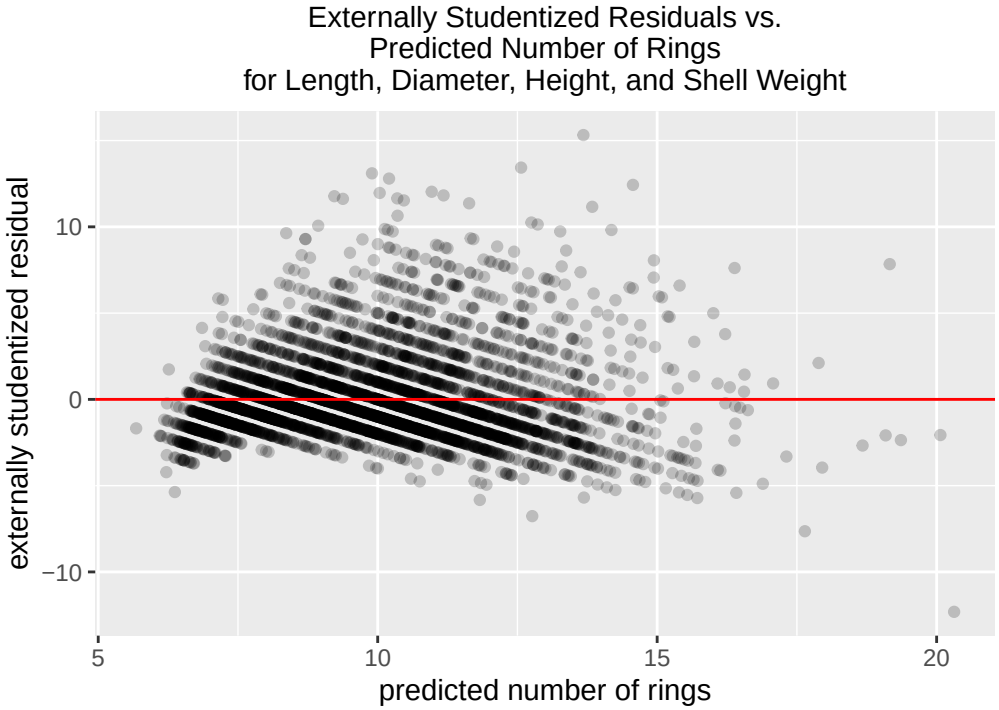


A graph of externally studentized residuals versus predicted number of rings for a multiple linear model of number of rings versus shell weight of a blacklip abalone is presented below. This graph seems to be a rotated version of the above graph. Both graphs suggest that transformations of shell weights and/or number of rings may help us linearize number of rings versus shell weight data before applying a simple linear regression model.

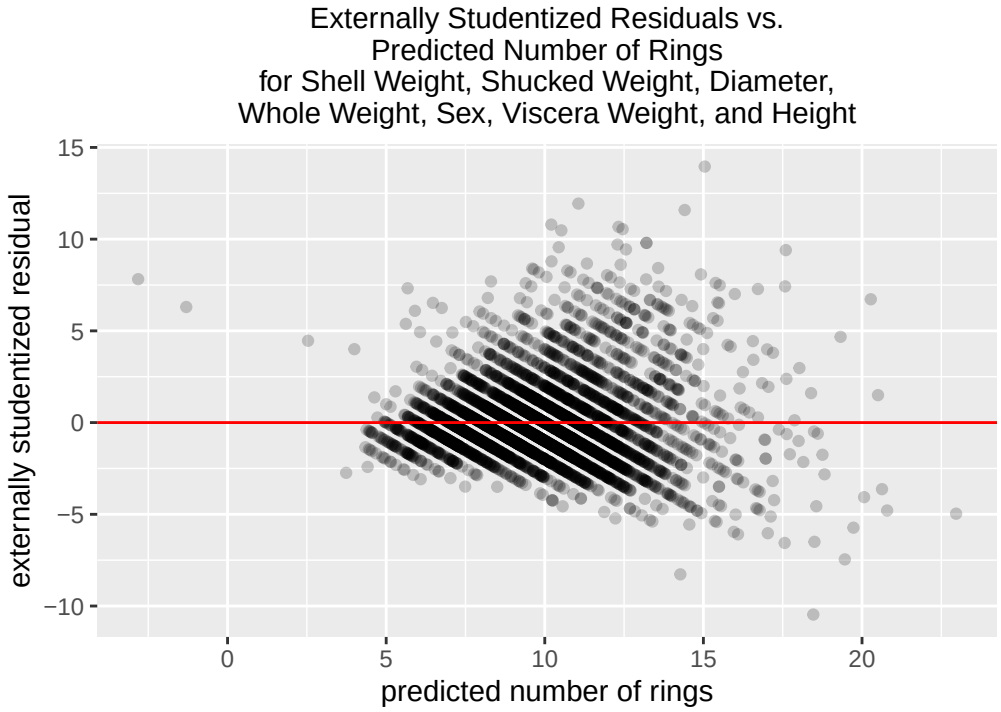


A graph of externally studentized residuals versus predicted number of rings for a multiple linear model of

number of rings versus length, diameter, height, and shell weight of a blacklip abalone is presented below. This graph resembles the above graph of externalized studentized residuals versus predicted number of rings and is better fit into a horizontal band, suggesting that this linear model may be more appropriate for determining how the age of a blacklip abalone is related to and/or predicted from the length, diameter, height, whole weight, shucked weight, viscera weight, and/or shell weight of the abalone than a simple linear model of number of rings versus shell weight.

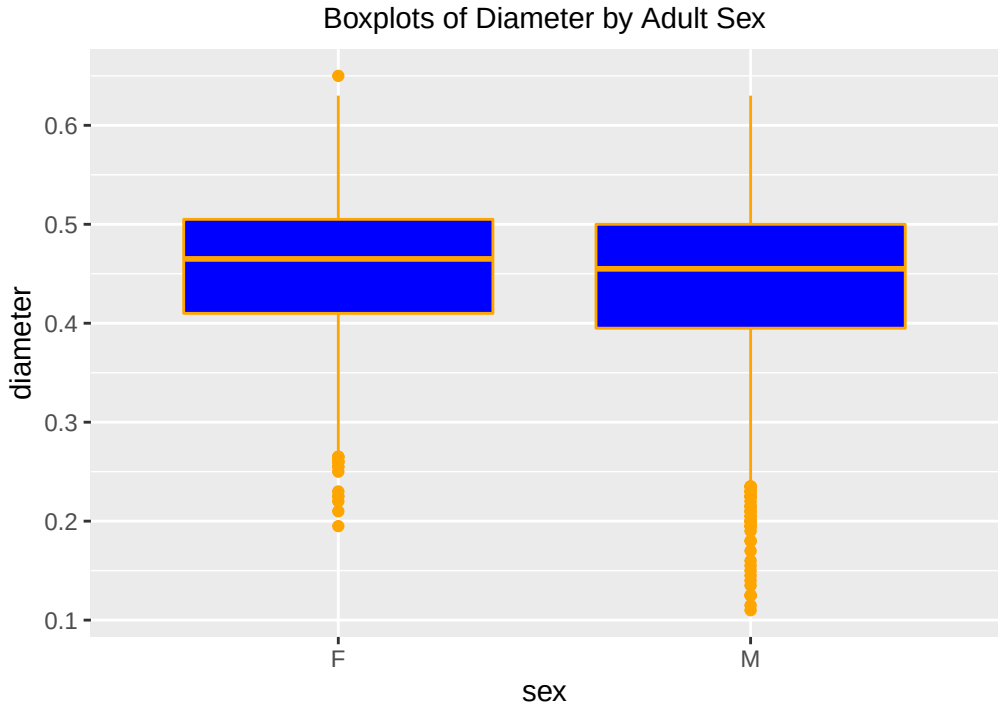


A graph of externally studentized residuals versus predicted number of rings for a multiple linear model of rings versus shell weight, shucked weight, diameter, whole weight, sex, viscera weight, and height of a blacklip abalone is presented below. This graph exhibits a right-opening funnel shape that indicates that the variance of the residuals grows with predicted number rings and transformations of number of rings and/or shell weight, shucked weight, diameter, whole weight, sex, viscera weight, and height may help us linearize number of rings versus predictors data before applying a multiple linear regression model.

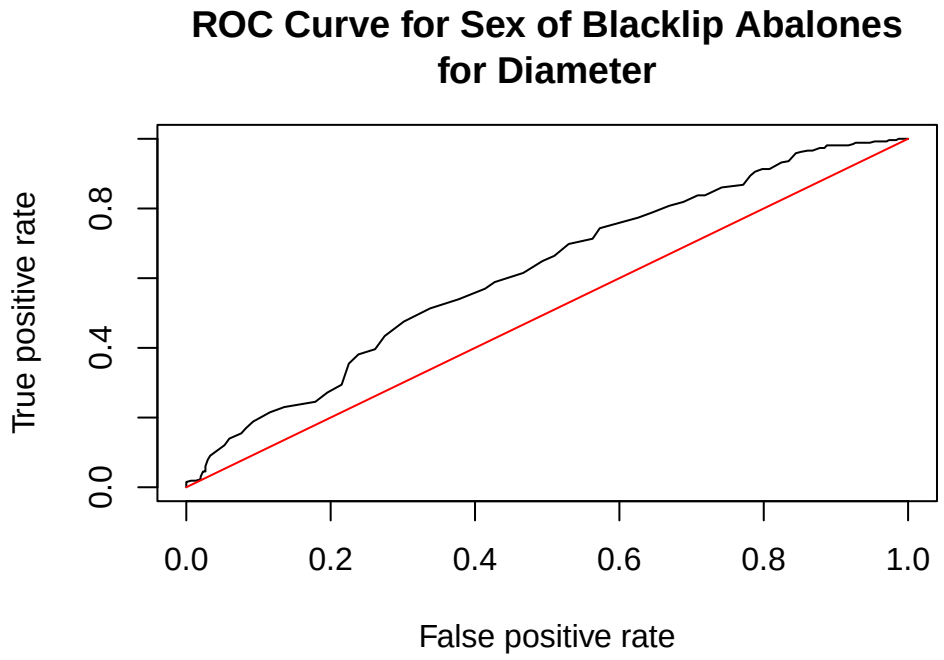


Data Visualizations for Multiple Logistic Regressions of Sex

Boxplots of diameter in millimeters by adult sex of blacklip abalones are presented below. Male blacklip abalones may have more outliers with lower diameters; female blacklip abalones may have more outliers with higher diameters. Males have a slightly lower minimum, first quartile, median, third quartile, and maximum diameter and have a slightly greater interquartile range of diameters. Male blacklip abalones may be smaller than female blacklip abalones. A binary classifier may be able to detect this trend.



The Receiver Operating Characteristic (ROC) curve for a logistic regression model of sex of blacklip abalones versus diameter is presented in black below along with the ROC curve for a random model. Ideally, this ROC curve would climb up the left side of the graph until it reached the top of the graph and continue along the top of the graph to the right. The ROC curve for the logistic regression model is a little better than the ROC curve for the random model. Since the ROC curve appears to be above the diagonal red line, we can assume that our logistic regression model will perform a little better than random guessing.



The Receiver Operating Characteristic (ROC) curve for a logistic regression model of sex of blacklip abalones versus diameter, shucked weight, viscera weight, and height is presented below along with the ROC curve for a random model. Ideally, this ROC curve would climb up the left side of the graph until it reached the top of the graph and continue along the top of the graph to the right. The ROC curve for the logistic regression model is similar to the above ROC curve.