Studying the Age and Sex of Abalones

Group 9: Brook Assefa, Shrikant Mishra, and Tom Lever

11/25/22

1. Executive Summary

According to Candy Abalone, a blacklip abalone is a marine univalve mollusc containing a large muscular foot that is highly sought after as seafood for eating. A blacklip abalone (*Halitois rubra*), is found in regions of southern coastal Australia, such as the island of Tasmania. According to the Department of Primary Industries in New South Wales, the species is currently recovering from historically low levels due to overfishing and mortality from a certain parasite. To help populations of blacklip abalones recover, specific bans on harvesting blacklip abalones with lengths less than a certain threshold have been implemented.

There is value in studying the population biology of blacklip abalones, including the relationship of number of rings / age, or sex, of blacklip abalones with the other, and with other physical attributes, including length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight. Understanding the relationship of age of blacklip abalones with the physical attributes of the abalones would allow marine biologists and environmental scientists to determine optimal methods for determining abalone age. Additionally, this study could help ensure healthy, balanced, sustainable, and diverse populations of abalones in terms of age and/or maturity. Scientists may identify trends in physical attributes for different ages over time, which may be useful in identifying individual and population health for blacklip abalones. Understanding the relationship of sex with other physical attributes could allow scientists to explore alternative ways of determining sex, as well as help ensure rebounds of abalone populations to healthy levels.

We are interested in using observations of blacklip abalones, such as those in the Blacklip Abalone Data Set, to answer the following two important questions. See the linked website or below descriptions of variables for further information.

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

In "A First Question of Interest: Data Visualizations", we note that, as expected, infant blacklip abalones generally have the lowest numbers of rings, ages, and shell weights. Adult male and female blacklip abalones have comparable numbers of rings and ages, though females may be slightly larger and older on average. In "A First Question of Interest: Model Building", we develop and consider the value of a linear model, of a version of number of rings / age for a blacklip abalone, and versions of other physical attributes of the abalone. After much consideration into different attributes and versions of those attributes, we recommend a linear model involving a version of number of rings, a version of shell weight, and two variables that together indicate sex. Evaluating this model, we find that the model is not ideal in predicting age, given that the model is only able to account for about 55 percent of the variation in rings. We may wish to consider nonlinear models of the relationship between number of rings / age of a blacklip abalone and other physical attributes of the abalone, and other methods of predicting age.

In "A Second Question of Interest: Data Visualizations", we note that, while it is easy to distinguish between infant and adult blacklip abalones, it is not easy to distinguish between adult male and adult female blacklip abalones. In "A Second Question of Interest: Model Building", we consider male and female blacklip abalones

only, and consider the relationship between the sex of a blacklip abalone and other physical attributes of the abalone, and a method to predict sex, by developing a binary classification / logistic regression model. After considering a logistic regression model with all physical attributes other than sex, and considering logistic regression models with only some of those physical attributes, a working model with the four physical attributes *length*, *height*, *shucked weight*, and *rings* was preferred. In evaluating this model, we determined that this model performed only slightly better than a model that guesses randomly, performed poorly in terms of key metrics, and was an ineffective predictor of sex. That being said, it did perform slightly better than a model that guesses randomly, suggests a relationship between the sex of a blacklip abalone and other physical attributes of the abalone, and moves us toward understand this relationship and the ideal methods for predicting sex.

2. Data Description

a. Source Description

Our data set with observations, corresponding to $n = 4,177$ blacklip abalones, of number of rings, sex, length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight is available at https://archive.ics.uci.edu/ml/datasets/Abalone.

b. Description of Variables

Per the UCI website listed above as well as the associated names file that comes with the data, here are the descriptions of the abalone physical characteristics:

- The sex of a blacklip abalone refers to either Male, Female, or Infant.
- The length (mm) of a blacklip abalone is the length of the longest horizontal line segment l along the abalone's shell.
- The diameter (mm) of a blacklip abalone is the measurement of the length of the horizontal line segment perpendicular to l.
- The height (mm) of a blacklip abalone is the measurement of the length of the longest vertical line segment h along the abalone's soft tissue and shell.
- The whole weight (g) of a blacklip abalone is the weight of the entire abalone, including the abalone's soft tissue and shell.
- The shucked weight (g) of a blacklip abalone is the weight of the abalone's soft tissue without the shell.
- The viscera weight (g) of a blacklip abalone is the weight of the abalone's soft tissue after bleeding.
- The shell weight (g) of a blacklip abalone is the weight of the shell without the soft tissue.
- According to Hao Chen of the University of California Davis, the number of rings of a blacklip abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. The age of a blacklip abalone is determined by the number of rings in the shell of the abalone. The age of a blacklip abalone in years is the sum of the number of rings of the abalone and 1.5.

3. A First Question of Interest

a. Introduction

i. Question Statement

We conduct data analysis of the above data set towards answering the following question:

"How is the number of rings / 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?"

ii. Value in Exploring Question

Addressing how the number of rings / 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. Determining the success of any population-boosting or remediation program may be enhanced.

b. Data Visualizations

i. Data Wrangling

No data wrangling was needed in order to produce the below data visualizations.

ii - iii. Presentation and Interpretation of Data Visualizations

A graph of number of rings versus shell weight in grams and sex for blacklip abalones is depicted below.

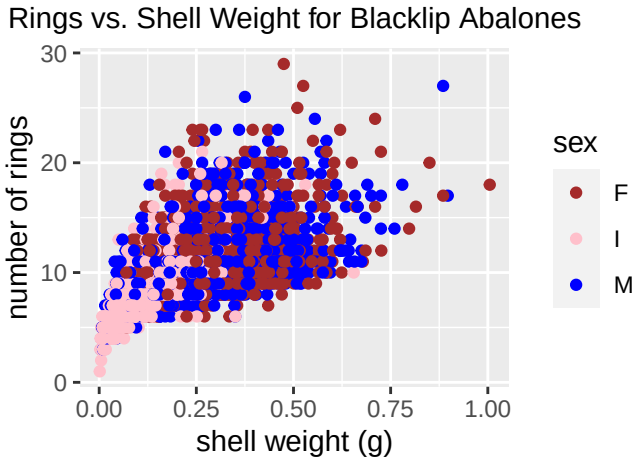


Figure 1: Scatterplot of number of rings versus shell weight and sex

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 as depicted in a correlation matrix in section 3.*c.ii*, the number of rings seems to be affected by sex values, especially between infants and adults.

Boxplots of the numbers of rings of blacklip abalones by sex is depicted below. Adult male and female blacklip abalones have similar distributions of number of rings. The adult female blacklip abalone with the greatest number of rings has a greater number of rings than the adult male blacklip abalone with the greatest number of rings. The adult male blacklip abalone with the least number of rings has a lower number of rings than the adult female blacklip abalone with the least number of rings. Infant blacklip abalones have the least numbers of rings.

Violin plots of number of rings / age by sex are similarly top-skewed. The most common number of rings for female and male adult blacklip abalones is about 10. The most common number of rings for infant abalones is about 7. The female adult blacklip abalone with the greatest number of rings / age has a greater number of rings than the male adult blacklip abalone with the greatest number of rings. The male adult blacklip abalone with the least number of rings / age has a lower number of rings than the female adult blacklip abalone with the least number of rings. Infant blacklip abalones have the least numbers of rings.

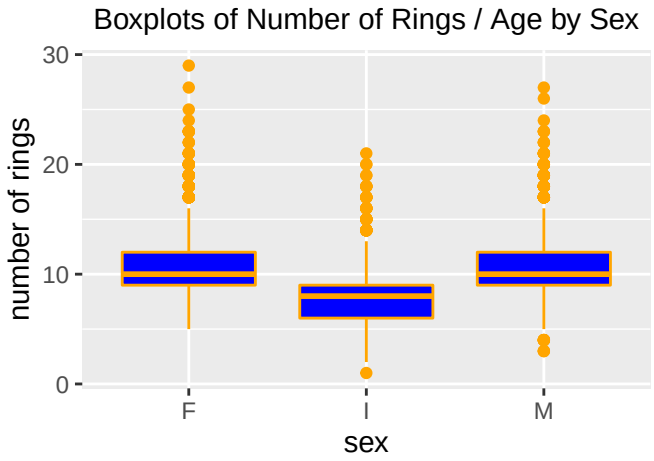


Figure 2: Boxplots comparing numbers of rings for infant, male, and female blacklip abalones

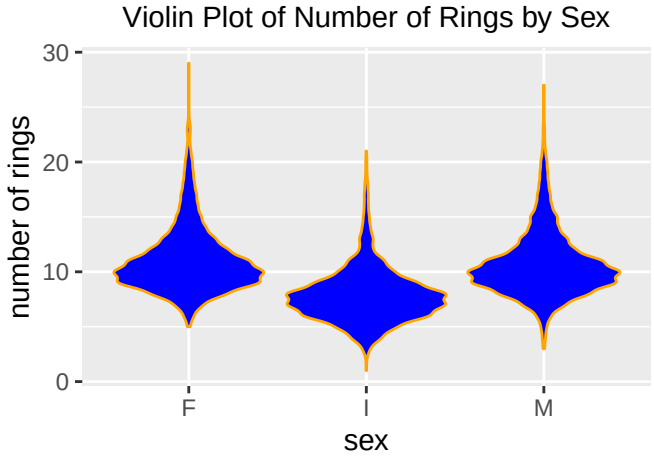


Figure 3: Violin plots comparing numbers of rings for infant, male, and female blacklip abalones

A histogram and probability density distribution of number of rings is depicted below. The divergence between the histogram and probability density distribution and the oscillation in the probability density distribution is due to half of the histogram bars having a height of 0, due to no observations having number of rings corresponding to those bins. The mean number of rings of a blacklip abalone is approximately 10 with a standard deviation of approximately 3. The histogram and probability density distribution are approximately normal and slightly right-skewed.

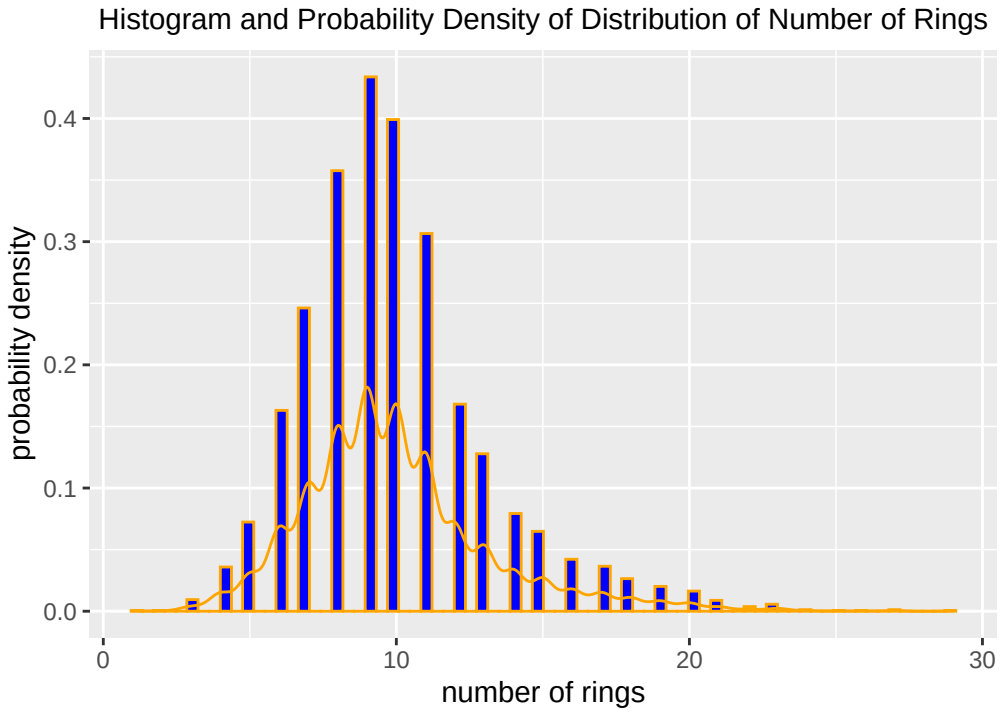


Figure 4: Histogram and probability density distribution of number of rings

c. Model Building

i. Choice of Initial Linear Regression Model

We study the age of abalones by constructing a multiple linear regression model. The response of the multiple linear regression model will be number of rings / age. The $p = 9$ columns and the first three rows of our data set are presented below.

```
##   sex length diameter height whole_weight shucked_weight viscera_weight
## 1   M  0.455    0.365  0.095       0.5140         0.2245         0.1010
## 2   M  0.350    0.265  0.090       0.2255         0.0995         0.0485
## 3   F  0.530    0.420  0.135       0.6770         0.2565         0.1415
##   shell_weight rings
## 1         0.15    15
## 2         0.07     7
## 3         0.21     9
```

We randomize the order of the rows of our data set so that AutoCorrelation Function (ACF) values for nonzero lags for a multiple linear regression model are insignificant, and so that an assumption that the residuals of a multiple linear regression model are uncorrelated is met.

We consider our data set to be a version of our data set with two indicator variables, *male* and $female$, substituted for the trinary predictor *sex*. The indicator variable *male* equals 1 when *sex* is M; *male* equals 0

when sex is not M. The indicator variable $female$ equals 1 when sex is F; $female$ equals 0 when sex is not F. An abalone being an infant is encoded by both $male$ and $female$ being equal to 0. The $p = 10$ columns and the first three rows of our data set are presented below.

```
##      length diameter height whole_weight shucked_weight viscera_weight
## 1422   0.72    0.575  0.170       1.9335          0.913          0.389
## 1017   0.63    0.485  0.155       1.2780          0.637          0.275
## 2177   0.57    0.450  0.170       1.0980          0.414          0.187
##      shell_weight rings male female
## 1422        0.510    13    0      1
## 1017        0.310     8    1      0
## 2177        0.405    20    0      1
```

For the remainder of our study of the number of rings / age of abalones, we consider $rings$ to be response and the other $k = 9$ parameters in our data set to be predictors.

After examining the below correlation matrix in section 3.*c.ii*, because all predictors other than $male$ and $female$ are at least moderately correlated, we consider an initial and working linear regression model with one predictor. Specifically, we choose the simple linear regression model with lowest Akaike Information Criterion (AIC):

$$rings = \beta_0 + \beta_1(shell\ weight) + \epsilon$$

ii. Improvements to Initial Model

A plot of residuals versus predicted values for our initial linear model is presented below.

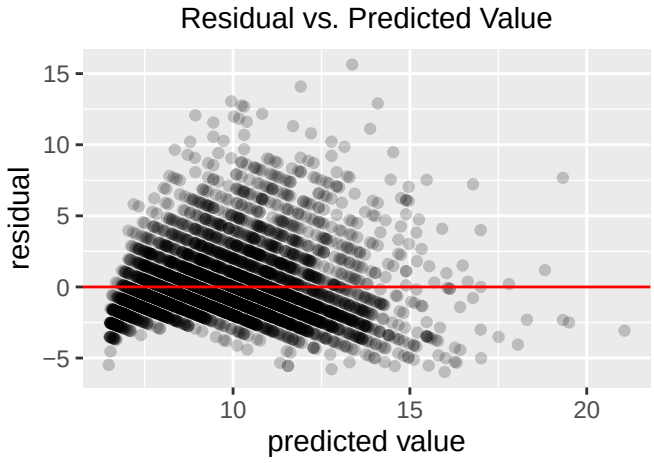


Figure 5: Plot of residuals versus predicted values for our initial simple linear regression model

The plot in Figure 5 exhibits:

1. a logarithmically shaped band that indicates that an assumption that the relationship between number of rings / age and shell weight is linear is not met,
2. a positive bias that indicates that an assumption that the mean residual is 0 is not met, and
3. a right-open funnel shape indicating that the variance of residuals increases with predicted value and that an assumption that the variance of residuals is constant is not met.

Below, we describe the process we used to determine a recommended multiple linear regression model.

We consider $k = 9$ simple linear regression models, each with response $rings$ and one of the $k = 9$ predictors. For each simple linear regression model, we meet simple linear regression assumptions that 1) the relationship between response and predictor is linear, 3) the variance of residuals is constant, and 2) the mean residual is

0. To meet these assumptions, we examine residual plots and apply the natural logarithm to *rings* if the predictor is not one of *male* and *female*, and to the predictor if the predictor is one of *height*, *whole weight*, *shucked weight*, *viscera weight*, and *shell weight*.

Specifically, we consider the plot of residuals versus predicted values for each simple linear regression model. The residual plot for each simple linear regression model with predictor other than *male* and *female* exhibits a right-opening, positively biased funnel shape suggesting that the residuals of the appropriate simple linear regression model have variance that increases with *rings* and mean residual greater than 0. For example, see Figure 5 above. The residual plots for the simple linear regression models with predictors *height*, *whole weight*, *shucked weight*, *viscera weight*, and *shell weight* additionally offer an impression of a logarithmically shaped band, suggesting that the relationship between *rings* and predictor may be logarithmic. See Figure 5 above.

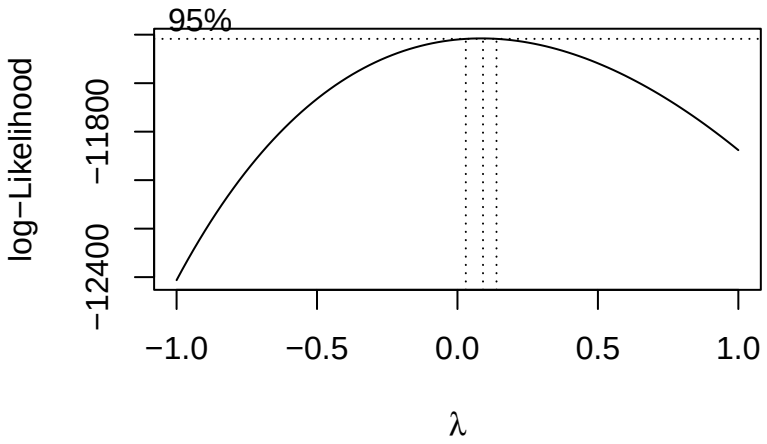


Figure 6: Box-Cox plot described below

We consider the Box Cox maximum likelihood estimate of parameter λ for each simple linear regression model. For each simple linear regression model, the magnitude of the Box Cox maximum likelihood estimate of parameter λ is less than 0.25. We create simple linear regression models with response $ln(rings)$ for each predictor other than *male* and $female$. For each simple linear regression model with $ln(rings)$, we consider a plot of residuals versus predictor values. For example, below is a residual plot for $ln(rings)$ vs. *shell weight*.

Following *Introduction to Linear Regression Analysis* (Sixth Edition) by Douglas C. Montgomery et al., for each plot of residuals versus predictor values, an impression of a horizontal band containing the residuals, centered at $e = 0$, is desirable. A nonlinear pattern in general implies that the assumed relationship between $ln(rings)$ and the predictor is not correct. A curved band that looks parabolic and concave up suggests a transformation $x' = x^2$; a curved band that looks logarithmic suggests a transformation $x' = ln(x)$. The plots of residuals versus predictor for predictors *length*, *diameter*, *male*, and $female$ offer an impression of a horizontal band. The plots of residual versus predictor for predictors *whole weight*, *shucked weight*, *viscera weight*, and *shell weight* have curved bands that look logarithmic. We apply the natural logarithm to the *whole weight*, *shucked weight*, *viscera weight*, and *shell weight* columns of our data set.

We consider the residual plots of each simple linear regression model with $ln(rings)$ and/or $ln(predictor)$. There are 5 models of the form $ln(rings)$ vs. $ln(predictor)$ and 4 models of the form $ln(rings)$ vs. *predictor*. These simple linear regression models offer the impression of a horizontal band centered at $e = 0$. The above simple linear regression assumptions are met for these simple linear regression models. As a reminder, these

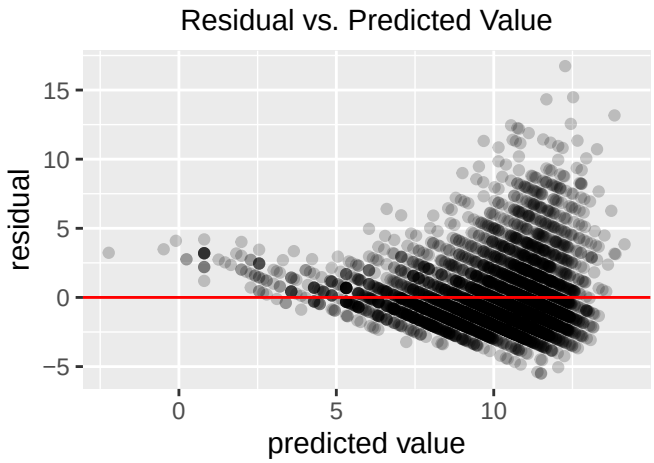


Figure 7: Residual plot for $ln(rings)$ vs. shell weight

simple linear regression assumptions are 1) the relationship between response and predictor is linear, 3) the variance of residuals is constant, and 2) the mean residual is 0.

For example, we present below a residual plot for $ln(rings)$ vs. $ln(shell\ weight)$.

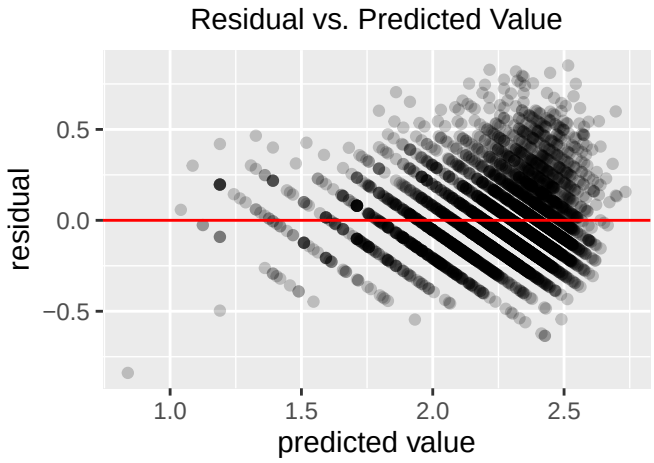


Figure 8: Residual plot for $ln(rings)$ vs. $ln(shellweight)$

For example, we present below the plot of AutoCorrelation Function Value vs. Lag for the simple linear regression model with $ln(rings)$ and $ln(shell\ weight)$. Most AutoCorrelation Function values are insignificant for these simple linear models. The remaining AutoCorrelation Function values for these simple linear models are approximately insignificant. A simple linear regression assumption that residuals are uncorrelated is met approximately for these simple linear regression models.

Additionally, we consider the QQ plots of each simple linear regression model with $ln(rings)$ and/or $ln(predictor)$. We present below the QQ plot for the simple linear regression model with $ln(rings)$ and $ln(shell\ weight)$. The simple linear regression models for $length$, $diameter$, and $height$ seem mildly either light-tailed or right skewed. The remaining simple linear regression models seem mildly either heavy left tailed and light right tailed or positively skewed. A simple linear regression assumption that residuals are normally distributed is approximately met for these simple linear regression models. Simple linear regression is robust to this assumption.

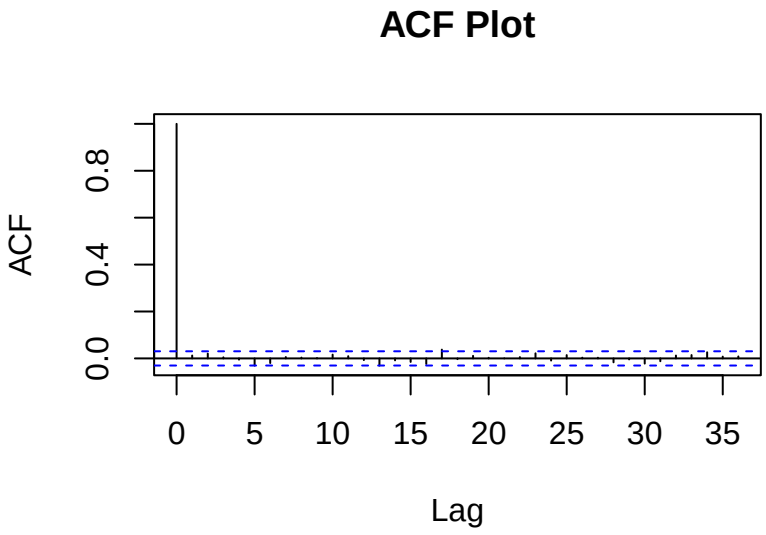


Figure 9: ACF plot for $ln(rings)$ vs. $ln(shellweight)$

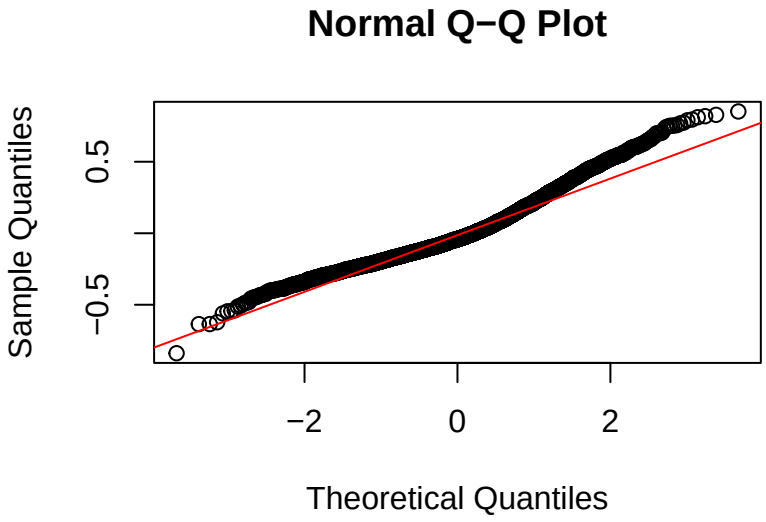


Figure 10: QQ plot for $ln(rings)$ vs. $ln(shellweight)$

Considering R's forward selection algorithm from an intercept only model to each simple linear regression model with $ln(rings)$ and/or $ln(predictor)$ as described above, the simple linear regression model with the smallest Akaike Information Criterion (AIC) is the model involving $shell\ weight$.

Because the transformed models meet the assumptions as described above, and the simple linear regression model $ln(rings)$ vs. $ln(shell\ weight)$ has the lowest AIC among transformed models, we choose as our working model a simple linear regression model with $ln(rings)$ and $ln(shell\ weight)$:

$$ln(rings) = \beta_0 + \beta_1\ ln(shell\ weight) + \epsilon$$

We can consider studying the number of rings / age of abalones with our working model, or a multiple linear regression model informed by our working simple linear regression model.

We analyze the correlation matrix for our data set with additional columns for $ln(variable)$. $ln(shell\ weight)$ is highly correlated with every variable other than $rings$, $male$, and $female$. Because of this, we prefer either a simple linear regression model with predictor $ln(shell\ weight)$, or a multiple linear regression model with predictors $ln(shell\ weight)$, $male$, and $female$.

```
##                 length diameter height whole_weight shucked_weight
## length            1.00     0.99   0.83         0.93           0.90
## diameter          0.99     1.00   0.83         0.93           0.89
## height            0.83     0.83   1.00         0.82           0.77
## whole_weight      0.93     0.93   0.82         1.00           0.97
## shucked_weight    0.90     0.89   0.77         0.97           1.00
## viscera_weight    0.90     0.90   0.80         0.97           0.93
## shell_weight      0.90     0.91   0.82         0.96           0.88
## rings             0.56     0.57   0.56         0.54           0.42
## male              0.24     0.24   0.22         0.25           0.25
## female            0.31     0.32   0.30         0.30           0.26
## ln_rings          0.56     0.57   0.56         0.54           0.42
## ln_shell_weight   0.90     0.91   0.82         0.96           0.88
##                 viscera_weight shell_weight rings  male female ln_rings
## length                    0.90         0.90  0.56  0.24   0.31     0.56
## diameter                  0.90         0.91  0.57  0.24   0.32     0.57
## height                    0.80         0.82  0.56  0.22   0.30     0.56
## whole_weight              0.97         0.96  0.54  0.25   0.30     0.54
## shucked_weight            0.93         0.88  0.42  0.25   0.26     0.42
## viscera_weight            1.00         0.91  0.50  0.24   0.31     0.50
## shell_weight              0.91         1.00  0.63  0.24   0.31     0.63
## rings                     0.50         0.63  1.00  0.18   0.25     1.00
## male                      0.24         0.24  0.18  1.00  -0.51     0.18
## female                    0.31         0.31  0.25 -0.51   1.00     0.25
## ln_rings                  0.50         0.63  1.00  0.18   0.25     1.00
## ln_shell_weight           0.91         1.00  0.63  0.24   0.31     0.63
##                 ln_shell_weight
## length                     0.90
## diameter                   0.91
## height                     0.82
## whole_weight               0.96
## shucked_weight             0.88
## viscera_weight             0.91
## shell_weight               1.00
## rings                      0.63
## male                       0.24
## female                     0.31
## ln_rings                   0.63
```

```
## ln_shell_weight            1.00
```

Based on the output of R's bidirectional-selection function, we determine that a multiple linear regression model with $male$ and $female$ is preferred, and we choose as our working multiple linear regression model for studying the number of rings / age:

$$ln(rings) = \beta_0 + \beta_1\ ln(shell\ weight) + \beta_2\ male + \beta_3\ female + \epsilon$$

Following "Detecting Multicollinearity Using Variance Inflation Factors" (https://online.stat.psu.edu/stat462/node/180/), the variance inflation factor VIF_j corresponding to predictor x_j in a multiple linear regression model quantifies how much the variance of the estimated coefficient β_j corresponding to predictor x_j is inflated due to multicollinearity / correlation between predictor x_j and other predictors. In particular, the variance inflation factor for predictor x_j

$$VIF_j = \frac{1}{1 - {R_j}^2}$$

where ${R_j}^2$ is the coefficient of determination obtained by regressing predictor x_j on the remaining predictors. A VIF of 1 means that there is no correlation among predictor x_j and the remaining predictors and that the variance of the regression coefficient corresponding to predictor x_j is not inflated. The general rule of thumb is that VIF's exceeding 4 warrant further investigation, while VIF's exceeding 10 are signs of serious multicollinearity requiring correction.

The Variance Inflation Factors for our working multiple linear regression model are all between 1 and 2, suggesting that the multicollinearity between $ln(shell\ weight)$, $male$, and $female$ is acceptable.

```
## ln_shell_weight            male          female
##        1.466889        1.793629        1.873197
```

iii. Recommended Model

Summarizing, above we assessed linear regression models by AIC, a residual analysis, a correlation analysis, bidirectional selection, and multicollinearity. Through comparing simple linear regression models by AIC, we chose as intermediary working models $rings$ vs. $shell\ weight$ and $ln(rings)$ vs. $ln(shell\ weight)$. Through bidirectional selection, we added $male$ and $female$. We recommend our working multiple linear regression model

$$ln(rings) = \beta_0 + \beta_1\ ln(shell\ weight) + \beta_2\ male + \beta_3\ female + \epsilon$$

We fit our recommended model to our data set.

```
##
## Call:
## lm(formula = ln_rings ~ ln_shell_weight + male + female, data = data_set)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90611 -0.14461 -0.03735  0.11824  0.84491
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.641156   0.012878 205.094  < 2e-16 ***
## ln_shell_weight 0.266836   0.004923  54.197  < 2e-16 ***
## male            0.070056   0.009171   7.639 2.70e-14 ***
## female          0.079876   0.009736   8.204 3.06e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.2132 on 4173 degrees of freedom
## Multiple R-squared:  0.5552, Adjusted R-squared:  0.5549
## F-statistic:  1736 on 3 and 4173 DF,  p-value: < 2.2e-16
##
## E(y | x) =
##     B_0 +
##     B_ln_shell_weight * ln_shell_weight +
##     B_male * male +
##     B_female * female
## E(y | x) =
##     2.64115640599568 +
##     0.266835929918076 * ln_shell_weight +
##     0.0700560947213758 * male +
##     0.0798757862201906 * female
## Number of observations: 4177
## Estimated variance of errors: 0.0454421885343188
## Prediction R2: 0.554316184461445
## Multiple R:  0.745123193994204   Adjusted R:  0.744908592004766
## Critical value t(alpha/2 = 0.05/2, DFRes = 4173): 1.64521885808865
## Critical value F(alpha = 0.05, DFR = 3, DFRes = 4173): 2.60703733356591
```

When fit to our data set, our recommended model is

$$\hat{\beta}_0 = 2.641$$

$$\hat{\beta}_1 = 0.267$$

$$\hat{\beta}_2 = 0.070$$

$$\hat{\beta}_3 = 0.080$$

$$ln(rings) = 2.641 + (0.267)\ ln(shell\ weight) + (0.070)\ male + (0.080)\ female$$

95-percent confidence intervals for the regression coefficients for $ln(shell\ weight)$, $male$, and $female$

```
##                      2.5 %    97.5 %
## ln_shell_weight 0.2571833 0.2764885
```

```
##           2.5 %    97.5 %
## male 0.05207569 0.0880365
```

```
##             2.5 %     97.5 %
## female 0.06078829 0.09896328
```

The number of rings / age of a blacklip abalone has a linear relationship with approximately the fourth root of shell weight when an indicator of whether the blacklip abalone is male is held constant and an indicator of whether the blacklip abalone is female is held constant.

$$rings = (shell\ weight)^{0.267} exp\left\{2.641 + 0.070\ male + 0.080\ female\right\}$$

When a blacklip abalone is an infant, the slope of the relationship between number of rings and approximately fourth root of shell weight is $exp\{2.641\} = 14.027$. When a blacklip abalone is male, the slope is larger, at $exp\{2.641 + 0.070\} = 15.044$. When a blacklip abalone is female, the slope is largest, at $exp\{2.641 + 0.080\} = 15.196$.

The confidence interval for the expected natural-log number of rings for adult female blacklip abalones with shell weights of the mean shell weight $x_0 = 0.239$ is

```
##        fit      lwr      upr
## 1 2.338923 2.327287 2.350559
```

The prediction interval for the natural-log number of rings for adult female blacklip abalones with shell weights of the mean shell weight $x_0 = 0.239$ is

```
##        fit      lwr      upr
## 1 2.338923 1.920831 2.757015
```

Both intervals include the point estimate for the natural-log expected value of number of rings for the mean shell weight and adult female blacklip abalones. Exponentiating this number yields the point estimate for the expected value of number of rings 10 for the mean shell weight and adult female blacklip abalones.

For an increase in shell weight by proportion p, the number of rings of a blacklip abalone increases by a factor of $(1+p)^{0.267}$. Age is the number of rings plus 1.5. For an increase in shell weight from the mean shell weight of 0.239 g of $p = 0.1$ (10 percent), the number of rings of a blacklip abalone increases by a factor of 1.026 to 0.245.

For a given *shell weight*, a male blacklip abalone has a number of rings that is a factor of $exp(0.070) = 1.073$ larger than an infant abalone. A female blacklip abalone has a number of rings that is a factor of $exp(0.080) = 1.083$ larger than an infant abalone. A female blacklip abalone has a number of rings that is a factor of $exp(0.080 - 0.070) = 1.010$ larger than a male abalone.

iv. Regression Assumptions, Outliers, High-Leverage Observations, and Influential Observations for Our Recommended Model

A plot of residuals versus predicted values, a plot of AutoCorrelation Function Values, and a QQ plot for our recommended model is presented below.

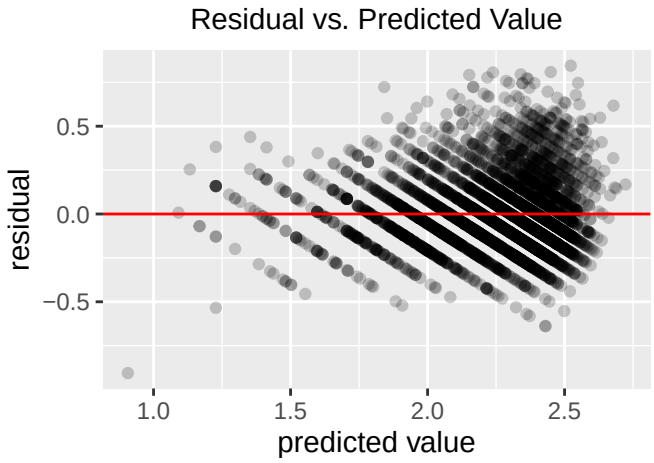


Figure 11: Residual plot for our recommended model

Assumptions for this multiple linear regression model are met.

1. The relationship between $ln(rings)$ and $ln(shell\ weight)$ is linear; the residuals are evenly scattered across $e = 0$.
2. The mean residual is 0; the residuals are evenly scattered across $e = 0$.
3. The variance of the residuals is constant; the residuals are evenly spread for different predicted values.
4. The residuals are uncorrelated; AutoCorrelation Function values are approximately insignificant.
5. The residuals are not normally distributed; the residuals have a moderately right-skewed distribution. However, multiple linear regression is robust to this assumption.
6. The variances of response number of rings for each sex are approximately equal.

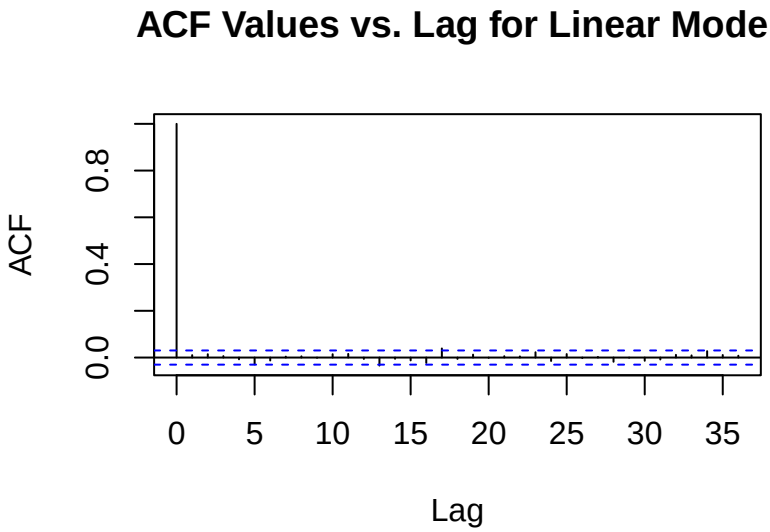


Figure 12: QQ plot for our recommended model

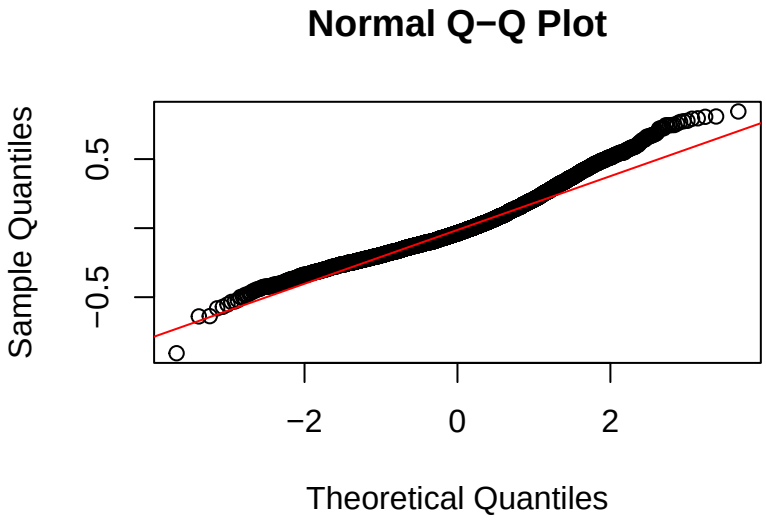
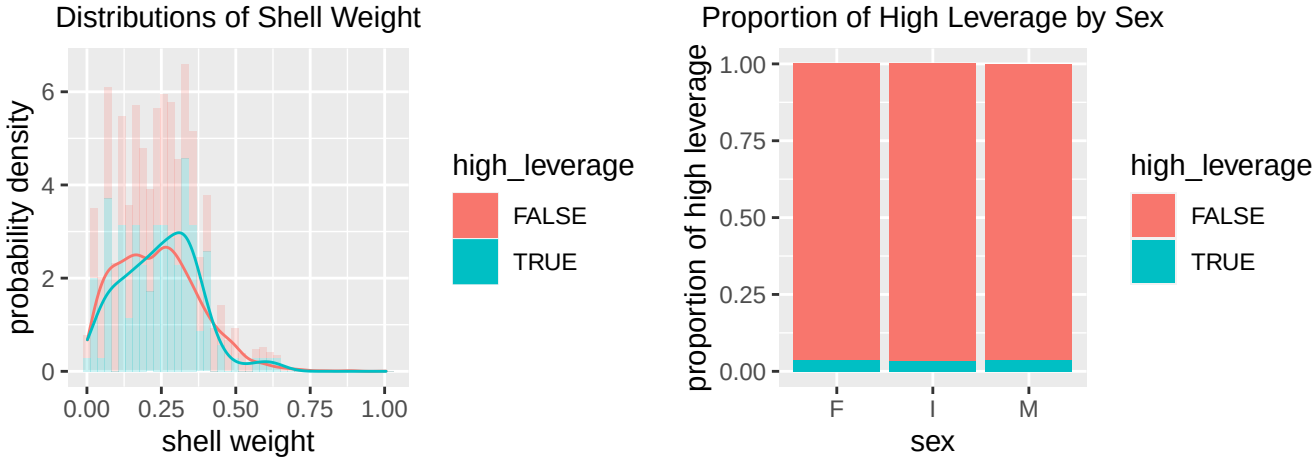


Figure 13: QQ plot for our recommended model

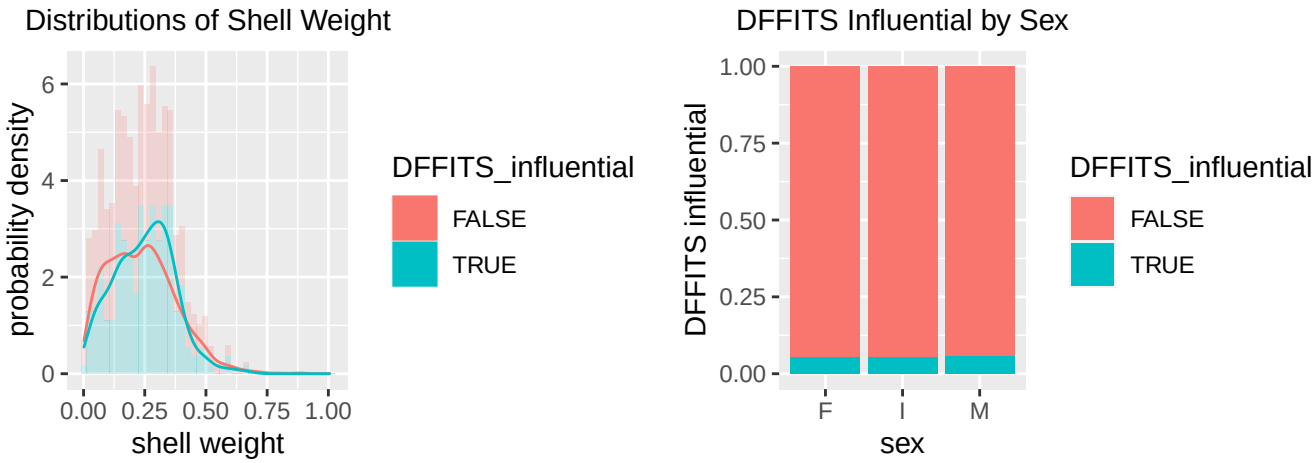
We use the Student's t distribution and the Bonferroni procedure to find a cutoff value for outlier detection using externally studentized residuals. Let $n = 4,177$ be the number of observations and let df_{Res} be the number of residual degrees of freedom of our recommended model. If magnitude of externally studentized residual $|t_i|$ is greater than a critical t value $t_c = t_{\alpha/(2n),df_{Res}-1}$, observation i is deemed an outlier. There are no outliers for our linear model.

A leverage h_{ii} is used to identify how far observation i is from the centroid of the predictor space. If leverage $h_{ii} > \frac{2p}{n}$, then observation i is deemed to have high leverage and is outlying in the predictor space. High leverage observations are data points that are most likely to be influential. There are 148 high-leverage observations. Shell weights corresponding to low and high leverages have similar right-skewed distributions, though the distribution of shell weights corresponding to high leverages seems to be taller and narrower and to have its peak for a higher shell weight. The proportions of low and high leverage observations are similar across sex.

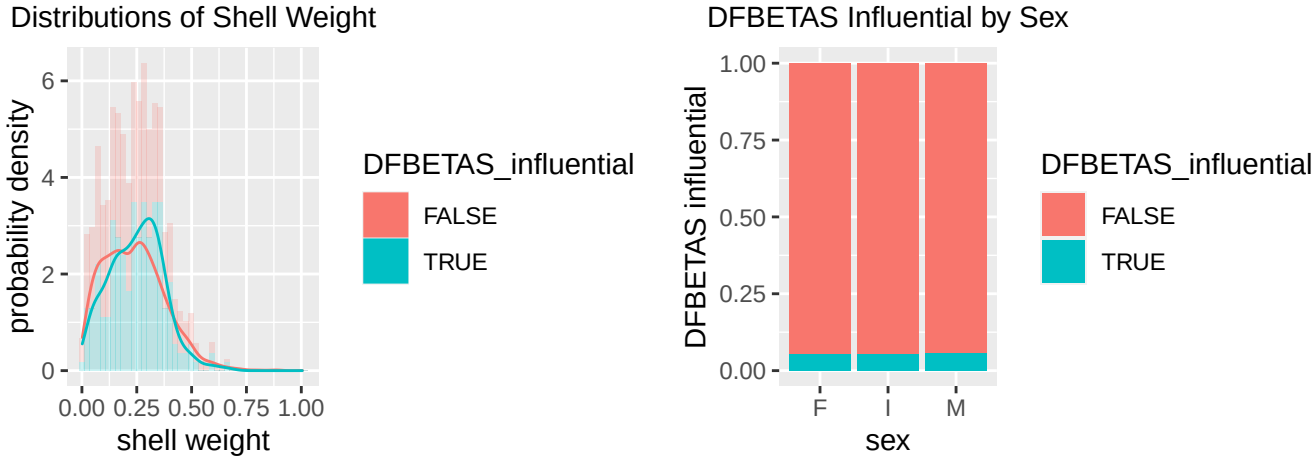


Cook's distance D_i can be interpreted as the squared Euclidean distance that a vector of fitted values moves when observation i is removed from the regression model. A cutoff rule for an influential observation is D_i is greater than a critical value $F_c = F_{0.5,df_R+1,df_{Res}}$. There are no influential observations using Cook's distance.

$DFFITS_i$ measures how the fitted value of observation i changes when it is removed from the regression model. Observation i is influential if $|DFFITS_i| > 2\sqrt{p/n}$. There 230 influential observations using $DFFITS_i$. Shell weights corresponding to non-influential and influential observations have similar right-skewed distributions, though the distribution of shell weights corresponding to influential observations seems to be taller and narrower and to have its peak for a higher shell weight. The proportions of non-influential and influential observations are similar across sex.



$DFBETAS_{j,i}$ measures how much the estimated coefficient $\hat{\beta}_j$ changes when observation i is removed from the multiple linear regression model. The cutoff used is $|DFBETAS_{j,i}| > \frac{2}{\sqrt{n}}$. There are 914 influential observations using $DFBETAS_{j,i}$. Shell weights corresponding to non-influential and influential observations have similar right-skewed distributions, though the distribution of shell weights corresponding to influential observations seems to be taller and narrower and to have its peak for a higher shell weight. The proportions of non-influential and influential observations are similar across sex.



d. Conclusions

i. Addressing Question of Interest

As a reminder, our question of interest is, "How is the number of rings / 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?"

The $p = 10$ columns and the first three rows of our data set are presented below.

```
##      length diameter height whole_weight shucked_weight viscera_weight
## 1422   0.72    0.575  0.170       1.9335          0.913          0.389
## 1017   0.63    0.485  0.155       1.2780          0.637          0.275
## 2177   0.57    0.450  0.170       1.0980          0.414          0.187
##      shell_weight rings male female
## 1422        0.510    13    0      1
## 1017        0.310     8    1      0
## 2177        0.405    20    0      1
```

Our recommended model is

$$ln(rings) = \beta_0 + \beta_1\ ln(shell\ weight) + \beta_2\ male + \beta_3\ female + \epsilon$$

$$ln(rings) = 2.641 + (0.267)\ ln(shell\ weight) + (0.070)\ male + (0.080)\ female$$

Using the summary of our recommended model above, we conduct an ANOVA F Test / Partial F Test involving all predictors. The null and alternate hypothesis for this test are $H_0 : \boldsymbol{\beta} = \mathbf{0}$ and $H_1 : \boldsymbol{\beta} \neq \mathbf{0}$. A test statistic $F = 1736$ is greater than a critical value $F_c = 2.607$. We reject the null hypothesis that all regression coefficients are 0 and prefer our recommended multiple linear regression model to an intercept only model.

Since each test statistic t in the summary of our recommended model is greater than a critical value t_c, for each corresponding predictor, we reject a null hypothesis that the regression coefficient for that predictor is 0, and conclude that each predictor is significant in the context of the multiple linear regression model / all predictors.

However, while the above hypothesis test results seem ideal, since the multiple and adjusted coefficients of determination R^2 for our recommended model are both 0.555, only a moderate proportion of variation in $ln(rings)$ of blacklip abalones can be explained by the predictors $ln(shell\ weight)$, $male$, and $female$. Given an R^2 adequacy criterion of 0.554, our recommended model is R^2 adequate, though barely. Because these coefficients are both less than a common threshold of 0.8, our linear model is not be good for prediction.

In summary, while our recommended multiple linear regression model does act in some degree as a statistically significant model for the true relationship between the various predictors and blacklip abalone age (as evidenced by the hypothesis tests above), unfortunately this model is not good in accurately predicting blacklip abalone age given these predictors.

ii. Insights

Considering potential confusion relating to our recommended multiple linear regression model, our set of boxplots of number of rings by sex suggests that there is overlap between the number of rings of infant and adult blacklip abalones and that maturity is not determined by age alone and/or age is not determined by number of rings alone.

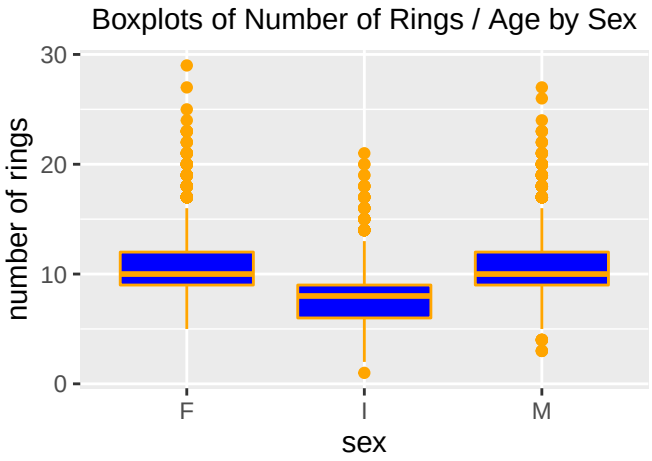


Figure 14: Boxplots comparing numbers of rings for infant, male, and female blacklip abalones

Considering again the value of our recommended multiple linear regression model, using our multiple linear regression model may help marine biologists or environmental scientists to understand better the relationship of the number of rings / age of a blacklip abalone and other physical attributes of the abalone, to determine the number of rings and age of blacklip abalones, and to ensure a balance of ages in populations of blacklip abalones. Nonlinear models may be better.

iii. Challenges

Regarding determining our recommended multiple linear regression model, our greatest challenges were considering transformations for a large number of simple and multiple linear regression models before considering various model selection methods.

Developing, testing, sharing, and using an R package for code research, brainstorming, and modularization was challenging.

Communicating remotely and across continents was challenging.

4. A Second Question of Interest

a. Introduction

i. Question Statement

We conduct data analysis of the above data set towards answering the following question:

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

ii. Value in Exploring Question

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.

b. Data Visualizations

i. Data Wrangling

As this question of interest forms a binary classification problem specifying male and female as the possible responses, the data set is altered to exclude any rows that correspond to infant abalones. Essentially, the rows that contain an 'I' from the sex column are excluded.

Additionally, for logistic regression, the data must be split into a train and test set, in order to to measure potential models' effectiveness on a new data set.

ii - iii. Presentation and Interpretation of Data Visualizations

With the training data created, the first step is to perform exploratory data analysis to understand any potential patterns within the data.

A bar chart comparing proportions of male and female abalones in the training data set is depicted below. There are slightly more male abalones (760) than female abalones (657) in our training data set, but there is sufficient representation for both values of our binary response variable.

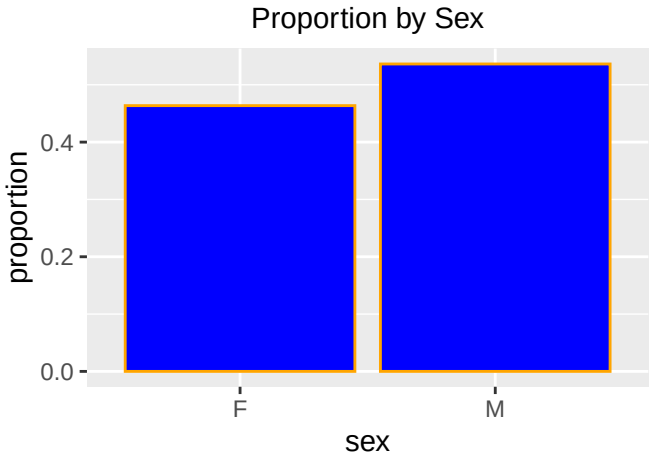


Figure 15: Bar chart comparing proportion between male and female abalones

By comparing boxplots of the physical attributes (e.g. length, shucked weight), any differences between the distributions of these attributes can be assessed between male and female abalones. These boxplots are presented in the figure below.

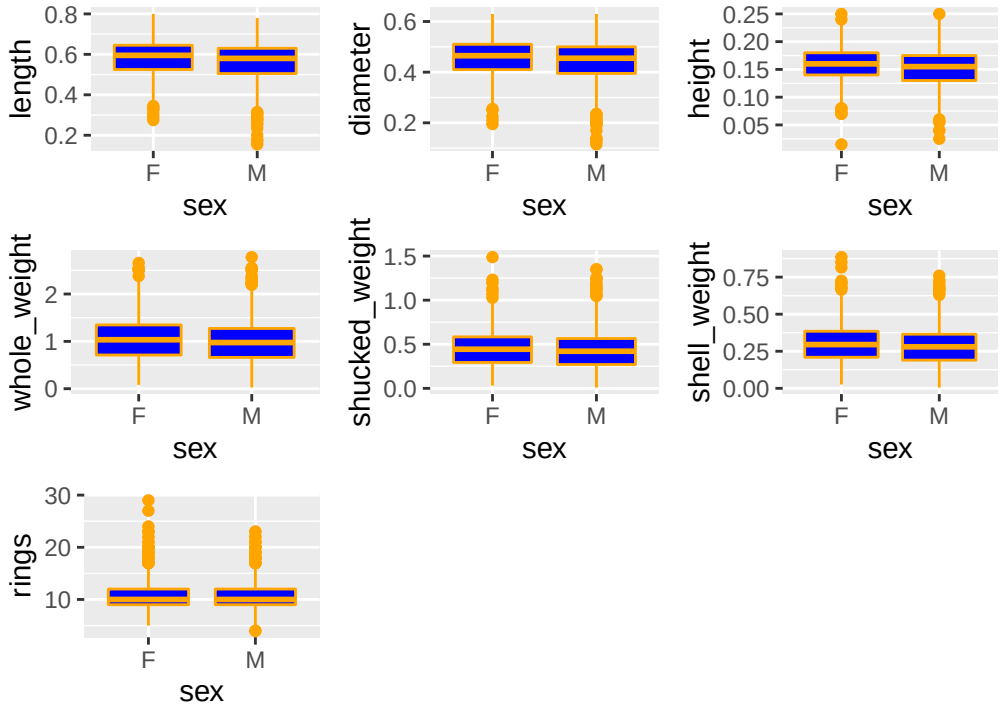


Figure 16: Boxplots comparing the distribution of the seven predictor variables between male and female abalones

It appears that the distributions of the physical attributes do not greatly differ when comparing between male and female abalones. However, it does appear that for several variables, male abalones tend to have slightly lower distribution. Male blacklip abalones have more outliers on the lower half, as seen with length and diameter. Males have a slightly lower minimum, first quartile, median, third quartile, and maximum for the vast majority of these plots. Although there is nothing conclusive in the figure, male blacklip abalones may be smaller than female blacklip abalones. A binary classifier may be able to detect this trend.

Next, depicted below are multivariate plots comparing blacklip abalone physical attributes against one another, with the color of the data point indicating the abalone's sex. This is done to assess whether a potential pattern can be seen that separates the two groups on a scatter plot between the predictors. Multiple plots were observed during exploratory data analysis, and the following six best represent the relationships between variables.

It can be seen that male and female data points appear to be interspersed throughout the graphs. However, the male data points also appear to occupy more of the lower ranges of both axes in each plot. More blue points exist in the bottom-left region, where length ranges from 0 - 0.4 mm.

Thus, we can state that although we can see some very slight differences that could distinguish male and female abalones apart, there is nothing conclusive in our preliminary data visualizations. We are unsure that a logistic regression model will be able to accurately predict an abalone's sex based off of its physical attributes.

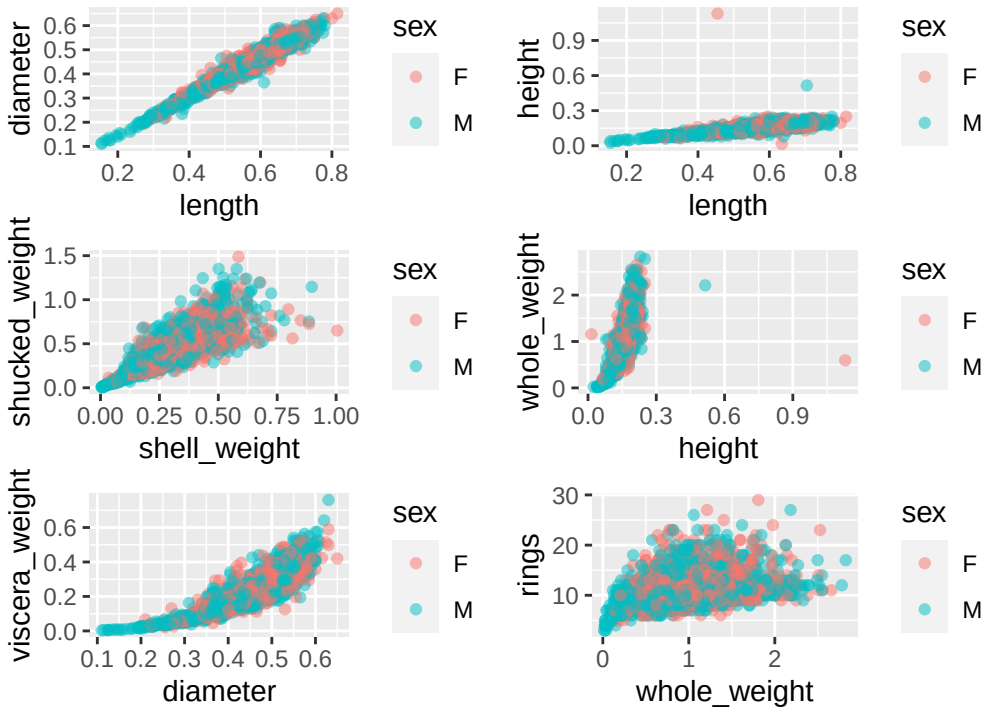


Figure 17: Multivariable scatter plots comparing blacklip abalone physical attributes against one another

c. Model Building

ii. Choice of Initial Logistic Regression Model

Our second question of interest is framed in a way that we want to understand the true relationship between the predictor variables and the sex variable, as well as to predict and extrapolate on new data. Thus, for our initial model, we will keep all variables present as a baseline, but we will try to improve on this model in the section below.

The general form of this initial model, Model 0, can be described as:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1\, length + \hat{\beta}_2\, diameter + \hat{\beta}_3\, height + \hat{\beta}_4\, whole + \hat{\beta}_5\, shucked + \hat{\beta}_6\, viscera + \hat{\beta}_7\, shell + \hat{\beta}_8\, rings$$

Fitting this model to the training data using the "glm" command in R, the initial regression model has an equation of:

$$\begin{aligned} log\frac{\hat{\pi}}{1-\hat{\pi}} = {} & 2.57043 \\ & + 0.73048\ length \\ & - 7.49341\ diameter \\ & - 2.59351\ height \\ & - 0.11939\ whole \\ & + 2.92424\ shucked \\ & - 2.97086\ viscera \\ & + 1.14699\ shell \\ & + 0.00696\ rings \end{aligned}$$

A hypothesis test is performed below to confirm whether this initial model is useful in predicting the log odds of the sex variable:

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = 0$

H_a : At least one of the coefficients in H_0 is not 0.

To perform this test, the ΔG^2 statistic is calculated by getting the difference between the deviance of an intercept-only model and our full model, which is 37.23. Comparing this value to a χ^2 distribution with 8 degrees of freedom, we calculate a p-value of 1.0456×10^{-6}.

Since the p-value is less than $\alpha = 0.05$, we reject H_0. There is sufficient evidence that our initial model is useful in predicting the log odds of abalones' sex.

ii. Improvements to Initial Model

We analyze the correlation matrix between predictors, in order to check for evidence of multicollinearity:

```
##          length diameter height whole shucked viscera shell rings
## length     1.00     0.98   0.82  0.92    0.88    0.89  0.87  0.30
## diameter   0.98     1.00   0.83  0.92    0.87    0.88  0.88  0.34
## height     0.82     0.83   1.00  0.84    0.76    0.80  0.84  0.40
## whole      0.92     0.92   0.84  1.00    0.96    0.95  0.93  0.32
## shucked    0.88     0.87   0.76  0.96    1.00    0.91  0.82  0.16
## viscera    0.89     0.88   0.80  0.95    0.91    1.00  0.86  0.27
## shell      0.87     0.88   0.84  0.93    0.82    0.86  1.00  0.46
## rings      0.30     0.34   0.40  0.32    0.16    0.27  0.46  1.00
```

This matrix shows that there are indeed high correlations between all the predictors in regards to one another, with the exception of *rings*. In fact, *rings* is the only variable that does not have high or very high correlation with the other predictors.

To further receive evidence for multicollinearity in the model, VIFs in the model are calculated below:

```
##         length       diameter         height   whole_weight shucked_weight
##      112.62596      108.46722       16.86897      364.44243      107.35926
## viscera_weight   shell_weight          rings
##       51.80660       70.76763        6.81246
```

The VIF values for all predictors can be considered high (greater than 10), with the exception of rings, so we can conclude that the model has multicollinearity. This may not have been a significant issue if the model was exclusively built for prediction, but question 2 aims to understand the true relationship between these variables and abalone's sex, as well as to be able to extrapolate with new data. For these reasons, this will require reducing multicollinearity as much as possible.

Intuitively, through the description of the data set being used, we know that whole weight, shucked weight, viscera weight, and shell weight are most likely linked because they all represent similar attributes. While shucked weight, viscera weight, and shell weight represent the weights of individual components of an abalone, whole weight represents the entire weight combined. It makes sense that a larger value for one would also mean a larger value for another, hence the very high positive correlation between them.

Similarly, length and diameter are highly correlated (.978) because they are just perpendicular measurements of the shell on the same plane. Height is not considered along with these two for now, as length and diameter are horizontal measurements pertaining to the abalone's shell, while height is a vertical measurement that includes soft tissue as well.

For each of these two groups (whole weight, shucked weight, shell weight and length, diameter), we can choose one variable to represent the group within a model with reduced parameters.

Using combinations of the choices of predictors in both groups, we now have a potential set of eight 4-predictor models. Because high correlation exists between predictors across groups as well (e.g. length and

whole_weight, diameter and viscera weight), we will add 7 two-predictor models, with rings as one predictor and a choice of one out of the remaining 7. It may make sense that a longer abalone may also have a higher weight. Rings is kept in all models, because it has a low to medium correlation with other variables.

Compiling all of these together, we have a set of 15 potential logistic regression models that could help describe the relationship between the the physical attributes of a blacklip abalone and its sex:

Model 1:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_3 height + \hat{\beta}_4 whole + \hat{\beta}_8 rings$$

Model 2:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_3 height + \hat{\beta}_5 shucked + \hat{\beta}_8 rings$$

Model 3:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_3 height + \hat{\beta}_6 viscera + \hat{\beta}_8 rings$$

Model 4:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_3 height + \hat{\beta}_7 shell + \hat{\beta}_8 rings$$

Model 5:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_2 diameter + \hat{\beta}_3 height + \hat{\beta}_4 whole + \hat{\beta}_8 rings$$

Model 6:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_2 diameter + \hat{\beta}_3 height + \hat{\beta}_5 shucked + \hat{\beta}_8 rings$$

Model 7:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_2 diameter + \hat{\beta}_3 height + \hat{\beta}_6 viscera + \hat{\beta}_8 rings$$

Model 8:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_2 diameter + \hat{\beta}_3 height + \hat{\beta}_7 shell + \hat{\beta}_8 rings$$

Model 9:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_8 rings$$

Model 10:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_2 diameter + \hat{\beta}_8 rings$$

Model 11:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_3 height + \hat{\beta}_8 rings$$

Model 12:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_4 whole + \hat{\beta}_8 rings$$

Model 13:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_5 shucked + \hat{\beta}_8 rings$$

Model 14:

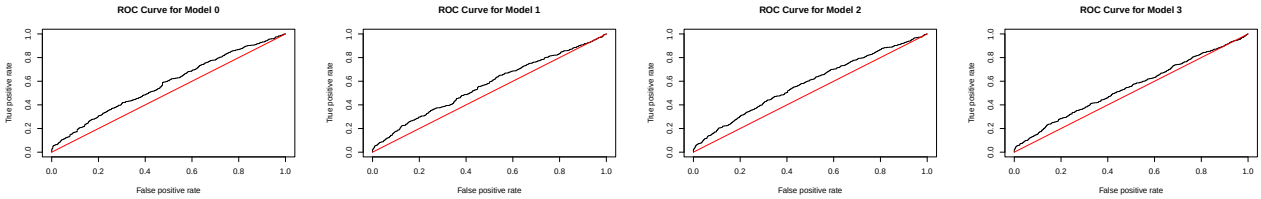
$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_6 viscera + \hat{\beta}_8 rings$$

Model 15:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_7 shell + \hat{\beta}_8 rings$$

iii. Performance of Model

Figure 18 shows the generated ROC curves for each of the models listed above. The results from these plots are not promising, as the ROC curve for all of the models appear to be close to the diagonal. While the curves of Model 2 and Model 6 appear to be further away from the diagonal, it is not by a significant margin. Thus, in general, these plots indicate that our models are not much better than random guessing.

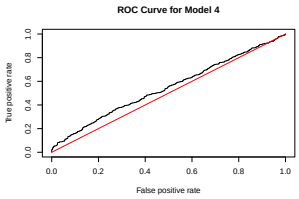


(a) Full Model 0

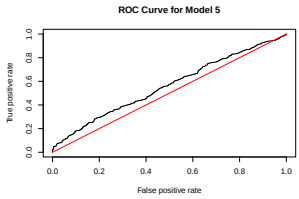
(b) Reduced Model 1

(c) Reduced Model 2

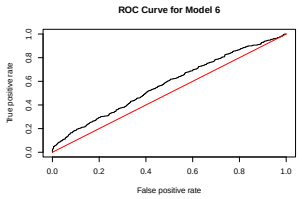
(d) Reduced Model 3



(e) Reduced Model 4

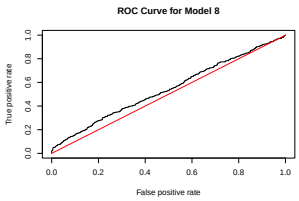


(f) Reduced Model 5

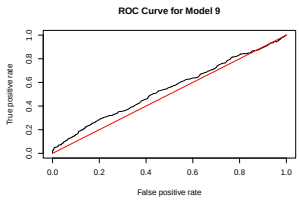


(g) Reduced Model 6

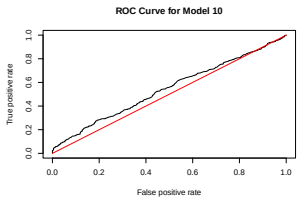
(h) Reduced Model 7



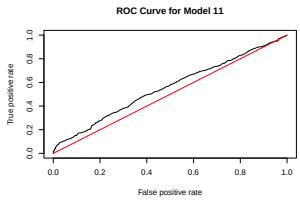
(i) Reduced Model 8



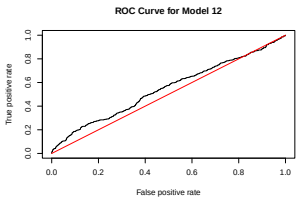
(j) Reduced Model 9



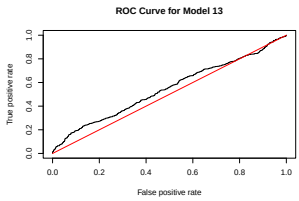
(k) Reduced Model 10



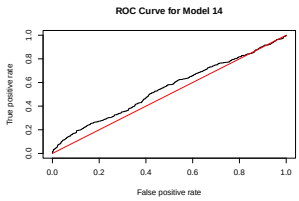
(l) Reduced Model 11



(m) Reduced Model 12



(n) Reduced Model 13



(o) Reduced Model 14

(p) Reduced Model 15

Figure 18: ROC curves of all potential models

The table below ranks the fifteen models by their area under the ROC curve (AUC) values. As seen with the plots in Table 1, model 2 is the best performing model with an AUC value of 0.58. However, the margin between model 2 and the lowest performing model, Model 15, is only approximately 0.04.

Table 1: Potential models ranked by AUC

Model	AUC
2	0.5808754
0	0.5771254

Model	AUC
6	0.5749740
1	0.5656410
5	0.5612851
11	0.5553576
15	0.5528716
14	0.5490835
3	0.5472616
12	0.5470613
4	0.5460276
13	0.5457913
7	0.5438141
9	0.5437921
10	0.5435407
8	0.5429768

In the figure below, we rank the potential models by their error rate. A threshold of 0.5 was chosen for the confusion matrix from which the error rate is calculated, because we are only interested in the lowest error rate rather than trying to limit false positive or false negative rates.

Table 2: Potential models ranked by error rate

Model	Error Rate
2	0.4393512
0	0.4400564
6	0.4407616
1	0.4485190
5	0.4492243
8	0.4569817
9	0.4583921
3	0.4598025
11	0.4612130
10	0.4619182
7	0.4640339
4	0.4654443
15	0.4675599
14	0.4682652
12	0.4724965
13	0.4851904

The model with the lowest error rate is model 2, with an error rate of 0.439. Additionally, the error rates range from 0.439 to 0.485, which again reinforces that the performance between the best and worst models is not drastically different. Again, similar to the AUC results, model 2's error value indicates that while it might be marginally better than the rest, the model does not perform well on the test data set.

iv. Recommended Model

With caution regarding its performance, we would like to hesitantly recommend model 2. Model 2 marginally performs the best out of all models in regards to AUC and error rate, even outperforming the full model.

The general form of the model:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = \hat{\beta}_0 + \hat{\beta}_1 length + \hat{\beta}_3 height + \hat{\beta}_5 shucked + \hat{\beta}_8 rings$$

The model fitted to the training data:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = 2.808184 - 5.281277\ length - 3.840513\ height + 2.080794\ shucked + 0.003041\ rings$$

A hypothesis test is performed below to compare Model 2 to Model 0, the initial model:

$H_0 : \beta_2 = \beta_4 = \beta_6 = \beta_7 = 0$

H_a : At least one of the coefficients in H_0 is not 0.

To perform this test, the ΔG^2 statistic is calculated by getting the difference between the deviance of Model 2 and the full model, which is 8.70. Comparing this value to a χ^2 distribution with 4 degrees of freedom, we calculate a p-value of 0.069.

Since the p-value is not less than $\alpha = 0.05$, we fail to reject H_0. Thus, we go with the 4-predictor model (Model 2) over the full model (Model 0) when predicting the log odds of abalones' sex.

With all that said, the low AUC values and high error rates show that this model is not ideal in predicting the log odds of abalone's sex.

Additionally, below are the VIF values for the *length*, *height*, *shucked weight*, and *rings* predictors:

```
##         length         height shucked_weight          rings
##      27.337074      14.394606      22.263915       5.327977
```

Aside from *rings*, these VIF values are still above 10, which indicates multicollinearity and further reinforces that while this model performs marginally better than the others, it is not ideal. One may consider a two-predictor model to avoid high VIF's and allow for extrapolation; however, this has similarly poor predictive ability. Additionally, a hypothesis test comparing the full model and a two predictor model shows that "dropped" predictors are significant.

d. Conclusions

i. Addressing Question of Interest

Our question of interest propelled us to investigate whether a relationship between any or all physical attributes of an abalone and the abalone's sex could be established, or whether knowing the measurements for the former could help us predict the latter.

The final model that was decided upon was:

$$log\frac{\hat{\pi}}{1-\hat{\pi}} = 2.808184 - 5.281277\ length - 3.840513\ height + 2.080794\ shucked + 0.003041\ rings$$

While it outperformed the other fifteen models by having the highest Area Under ROC Curve value (0.581) and the lowest error rate (0.439), these values along with an ROC Curve that is close to the diagonal indicate that this model is not able to accurately predict log-odds of sex, and unable to accurately represent the relationship between the physical attributes of an abalone and its sex.

The range between the best and worst models for AUC was 0.034, and the range between the best and worst models for error rate was 0.023. This indicates that these models did not differ much regardless of choice of predictors, and were equally ineffective at predicting blacklip abalones' sex.

Thus, we say that question 2 was not answered by the chosen model, and cannot be answered by similar logistic regression models using the provided data. This reinforces the results of the exploratory data analysis conducted beforehand, where there weren't clear difference between male and female abalones across the various plots.

In summary, while the equation above does act in some degree as a statistically significant model for the true relationship between the physical attributes of blacklip abalone and sex (as evidenced by the hypothesis test

in the previous section involving full and partial models), this model is only slightly better than a random logistic regression model in accurately predicting blacklip abalone sex given these predictors.

ii. Insights

Considering our data set and analysis, initially, it was surprising to us that the correlation matrix between predictors had high correlation values for most pairs. However, after understanding the context behind the variables in the data description, it made sense that the variables would be heavily positively correlated. While other weight variables measure components within an abalone, the whole weight measures the entire abalone's weight, soft tissue and shell included. Additionally, it might intuitively make sense that a larger abalone with a heavier shell might be longer along the abalone's shell as well. Thus, the high positive correlations and subsequently examined multicollinearity can be explained.

From a biological / environmental perspective, identifying male and female abalones, inducing reproduction, and ensuring a balance according to sex is valuable to marine biologists, conservationists, and a diverse, robust marine environment and ecosystem. Based on our logistic regression modelling, we encourage those interested in determining sex in relationship to other attributes of blacklip abalones or predicting sex to use established analytical methods; it is hard to identify sex of abalones based on a logistic regression model.

iii. Challenges

The largest challenge for the group with this question was ensuring that all decisions taken during the initial model improvement steps were valid. We were very curious whether components of our analysis from the linear regression section, such as multicollinearity, stepwise regression, and certain hypothesis testing would still be relevant if used when performing logistic regression. This also included making sure that the decision to reduce the number of parameters for potential models based on context from the data description, such as reducing the number of weight variables due to their high correlation and underlying relationship, was also valid.

References

1. Candy Abalone (2022). "Candy Abalone". Accessed on 12/14/22 at https://www.candyabalone.com.
2. Department of Primary Industries in New South Wales (2009). "Blacklip Abalone". Accessed on 12/14/22 at https://www.dpi.nsw.gov.au/___data/assets/pdf_file/0009/375858/BlacklipAbalone.pdf.
3. UCI Machine Learning Repository (2022). "Abalone Data Set". Accessed on 12/14/22 at https://archive.ics.uci.edu/ml/datasets/Abalone.
4. Hao Chen, University of California Davis (2022). "Abalone Age". Accessed on 12/14/22 at https://anson.ucdavis.edu/~haochen/abalone_description.pdf.
5. STAT 462: Applied Regression Analysis, PennState Eberly College of Science (2022). "Detecting Multicollinearity Using Variance Inflation Factors". Accessed on 12/14/22 at https://online.stat.psu.edu/stat462/node/180/.
6. Baptiste Auguie (2022). "Laying out multiple plots on a page". Accessed on 12/14/22 at https://cran.r-project.org/web/packages/egg/vignettes/Ecosystem.html.
7. Yihui Xie (2022). "LaTeX sub-figures". Accessed on 12/14/22 at https://bookdown.org/yihui/rmarkdown-cookbook/latex-subfigure.html.