Increase Efficiency of Age Determination: Predicting Age of Abalone Based on Physical Traits

Presented for STA 206

Kate Jones (katjones@ucdavis.edu), Gianni Spiga (glspiga@ucdavis.edu)

2022-12-05

Abstract

Data was collected on stunted blacklip abalone to determine if there is a more efficient method to determine the age of an abalone rather than counting its rings. The abalone were collected from 5 regions in Bass Strait: Kent Islands, Waterhouse Island, Hogan Islands, Babel Island, and Barren Island. Various statistical methods were implemented to determine the physical traits of abalone that result in the best prediction of their age. It was found early on that many of the variables in the data set were highly correlated, and tackling this challenge will be discussed in the Methods section. Our model selection process determined that infancy, diameter of the shell, height of the shell, shucked weight, and shell weight were the most significant predictors of age.

Contents

Introduction

Abalone are classified as marine gastropod molluscs, and are also known as marine snails. They have a worldwide distribution, but are mostly found off the coasts of New Zealand, South Africa, Australia, Western North America, and Japan. In this study, samples of Abalone are taken from Islands near Tasmania. The data includes physical characteristics of 1477 stunted blacklip abalone. The traits recorded are as follows:

- Sex: Male (M), Female (F), and Infant (I)
- Length: longest shell measurement (mm)
- Diameter: perpendicular to length (mm)
- Height: with meat in shell (mm)
- Whole weight: whole abalone (g)
- Shucked weight: weight of meat (g)
- Viscera weight: gut weight after bleeding (g)
- Shell weight: after being dried (g)

Sex is a categorical variable while the other variables are all quantitative, hence we will make sure to declare Sex as a factor when building our model. The purpose of this study is to make determining the age of abalone more efficient for researchers. The current method for determining the age of an abalone is cutting the shell through the cone, staining it, and counting the number of rings through a microscope. This process is very tedious and time consuming. Determining each of the physical parameters above would also be an inefficient process, so the goal is to produce a model that requires the fewest number of predictors while also remaining significant. Abalone are researched for various reasons: risk of extinction, abalone farming, and their ecological importance in controlling algal density and supporting diversity of kelp species. Due to illegal harvesting, disease, and low reproduction rates, abalone populations are on the decline. There are various researchers around the world, including Dr. Kristin Aquilino of the UC Davis Bodega Bay Marine Laboratory, working on restoring the world's abalone population. Abalone are not only responsible for supporting kelp forests, but are also a significant contributor to the economy of many countries, including South Africa (NOAA).

Methods and Results

Exploratory Data Analysis

In order to gain a better understanding of the data, we first created histograms for the quantitative variables and a pie chart for our one qualitative variable (Appendix 1, Figures 1-8). We can see that length is left-skewed with a mean of 0.524mm and a median of 0.545mm. Diameter has a similarly shaped distribution with a mean of 0.408mm and a median of 0.425mm. The height variable has a very narrow distribution with a slight left skew with a mean of 0.1395 and a median of 0.14. Each of the weight variables has a fairly similar distribution that is mildly right skewed. Whole weight has a larger spread with a mean of 0.829g and a median of 0.7995g. Shucked weight has a mean of 0.359g and a median of 0.336g. Viscera weight has a mean of 0.181g and a median of 0.171g. Lastly, shell weight has a mean of 0.239g and a median of 0.234. The response variable, rings, is also plotted. We can observe a distribution that is right skewed with a mean of 9 and a median of 9.933 (Appendix 1, Figure 9).

Due to many of the variables being related to weight, we had a hunch that there would be a high level of multicollinearity in the data. Hence, we examined the correlation values between each of the variables, as well as variance inflation factors (VIF). As can be seen in the correlation plot (Appendix 1, Figure 10), many of the variables have a correlation coefficient value greater than 0.9. Length and diameter have a correlation coefficient of 0.99, hinting that we will later need to disregard one of these variables in our final model. Further, these two variables being highly correlated is fairly intuitive since abalone are fairly round and consistent in shape. The VIF values are as follows: Sex- 1.543, Length - 40.946, Diameter - 42.380, Height - 3.581, Whole.weight - 109.769, Shucked.weight - 28.551, Viscera.weight - 17.445, Shell.weight - 21.263. Any VIF value greater than 10 suggests high multicollinearity, and many of the VIF values far exceed this value. Discovering this high level of multicollinearity will help guide our model-building process.

Checking Model Assumptions

Our first step of building a predictive model is ensuring that we meet the assumptions to build a proper model. We check that we do not violate homoscedasticity via a Residuals vs. Fitted Values plot. We see no obvious sign of a pattern, so we can conclude that we have no heteroscedasticity (Appendix 1, Figure 11). We also must check the normality of our residuals with a Normal Quantile-Quantile plot (Appendix 1, Figure 12). Upon plotting, we immediately observe that our residuals are right skewed, as hinted by the histogram of our response variable, Rings. From here, a Box-Cox transformation was performed on an initial full model to find the optimal transformation for our response variable. Suggested from a 95% confidence interval (Appendix 1, Figure 14) that a lambda value equal to zero would be optimal, we performed a log transformation on our response variable and then reassessed our assumptions. Our new log-linear model provides much stronger evidence for approximate normality in our residuals, allowing us to carry on with our analysis. The slight caveat however, we do have two outliers that skew that approximation of normality, observation 237 and 2052.

These two observations have extremely large residuals, relative to the rest of the data. Along with that, a plot of our residuals vs. leverage plot shows us that the Cook's distance (Appendix 1, Figure 13) for observation 2052 is nearly 500 times larger than the rest of the points. Investigating these two points, we first recognize observation 237, which is a very young and tiny abalone, much younger and smaller than the majority of infant abalones in our data. Observation 2052 however, is an extremely tall and heavy female abalone, relative to other females in the data. Due to these influential observations, we will remove them from the data, as to not skew our predictability.

Model Selection

Now that we have made clear that our assumptions for the building of a linear model are met, as well as removing influential points, we then switch our focus to the quality of the predictors in the model. A general F-test tells us that our model is highly significant at any reasonable alpha. A t-test for each predictor coefficient shows us that the dummy variable of male sex is not significant, despite the observation that sex being an infant is a highly significant piece of information for the model. Due to this, we create a new dummy variable; one which only will provide us information about whether the abalone is an infant or an adult.

The next issue faced was the challenge of multicollinearity, as seen in the exploratory data analysis, which is persistent throughout the variables in our data. Due to this, we first try to remove shucked weight from our prediction, given that it's correlated with whole weight at a Pearson correlation coefficient of 0.97. Using this logic, we also removed viscera weight, shell weight, and length, whose correlation coefficient with diameter was 0.99. Combining this with our new infant variable in hopes of a simpler model, we find that our adjusted R squared value shrinks by about 0.09, leaving us to infer there is room for improvement.

Changing direction, we switched to a ridge regression model, which would allow us to observe which variables should be shrunk due to the penalization of an L2 norm. We first find an optimal lambda value for our ridge model and observe our coefficients. We observe that the smallest coefficient is whole weight, with infant, length, and viscera weight not far behind. Despite this, we decided to approach a new model suggested by our ridge regression with the deletion of the whole weight first. This new, significant model shows us that all but viscera weight are significant, so we update once more to a model that only includes Infant status, length, diameter, height, shucked weight, and shell weight.

Treating this as a new full model, we then perform forward stepwise selection on sub-models based on AIC and BIC criterion. Our stepwise regression based on AIC suggests the AIC lowers with all variables except length. BIC, in further support, increases when length is left in the model. This positive result allows us to further justify removing length from the model as we did earlier, recalling that the correlation between length and diameter was near one.

We perform our stepwise regression one more time, with both AIC and BIC criterion, on our length-less model. In both iterations, we indeed find that the model with height, shell weight, shucked weight, diameter, and infant status is the best model to predict the log of rings based on the mentioned criterion. This model also provides us with an adjusted R squared of 0.5956, which is near the highest we have seen, while also being simplified and reducing multicollinearity. We have one final component to add and expand our model, interactions.

We test all interactions with our remaining variables as listed in the previous paragraph. Individual t-tests tell us interactions such as shucked weight and height as well as shell weight and infant are non-significant. We reduce the model to find by the same process that the interaction between diameter and and our infant variable is non-significant. We then perform a stepwise selection just as before, using the AIC and BIC criterion. Finally we are left with a model containing height, shell weight, shucked weight, diameter, infant, and interactions between the following: height and diameter, shucked weight and diameter, and shucked weight and infant. With our new model reporting an adjusted R squared of 0.6273, we see an obvious improvement from our previous 0.5956. Here, we decide that these variables are the best set of predictors to be our final model.

Model Interpretation

Before interpreting the model, we first need to keep in mind that the response variable underwent a log transformation, hence care needs to be taken in understanding how changes in the predictor variables affect the value we want to predict, which is the number of rings. The final model equation is as follows:

$$\hat{Y}_{Rings} = 0.919 + 11.076_{Height} - 2.256_{Shucked.weight} + 1.676_{Shell.weight} + 2.548_{Diameter} - 0.201_{Infant} \\ -21.144_{Height:Diameter} + 2.810_{Shucked.weight:Diameter} + 0.447_{Shucked.weight:Infant}$$

The model indicates that diameter, height, and shell weight are all positively related to the log number of rings. Additionally, the interaction terms of shucked weight/diameter and shucked weight/infant are positively correlated to the response variable. Since y = log(x) is a monotone increasing function, this means that each of these variables is also positively related to the number of rings. Infancy and shucked weight are both negatively related to the number of rings. The height/diameter term is also negatively correlated to the response variable. Intuitively, the relationship between infancy and the response variables makes sense. Being an infant means that you are younger and hence will have fewer rings than adult abalone. The shucked weight variable is more open to interpretation. One could hypothesize that younger abalone tend to eat more since they are still growing and developing. Hence, if they consume more, their shucked weight should be greater.

We have three interaction terms in the model. The first suggests that height has a different effect on the number of rings based on the diameter of the abalone. A similar interpretation can be made for shucked weight and varying values of diameter. The influence of shucked weight is further explained by if the abalone is an infant or not. If the abalone is an infant, the infant variable evaluates to 1 and shucked weight has an increased effect on the predicted value for the number of rings.

Conclusion and Discussion

Throughout the model building process, we determine that a linear model including some interaction terms provides the best fit for our data. We also determined that the data has high levels of multicollinearity and accounted for this obstacle via Ridge regression. With this, we determined that many of the variables did not need to be included in our model. Overall, we believe our final model for the age of stunted blacklip abalone is the best we can build given the parameters. However, we would like to identify a few places where we could have improvements. For instance, we are only given abalone data from one specific region. We would be curious to test this model with populations from abalone regions all around the world and compare. Similar in theme, we would also like to have more data available on different species of abalone outside of the stunted abalone. Regarding the abalone in this data, we would also be curious to see how other factors account for age. Perhaps older abalone have a different diet than those who are younger. Older abalone might be collected in different regions of coastline than younger ones. Lastly, how has the abalone population changed over time? Can we find data that is available from past abalone populations to compare in microevolutions over time? Answering these questions, we believe, is key to improving our predictions of collect abalone age. For now, we have created a linear model that only requires a few pieces of physical data, making it easier for researchers and scientists to determine the number of rings on an abalone shell.

Appendices

Appendix 1

Figure 1: Pie Chart of Abalone Sex

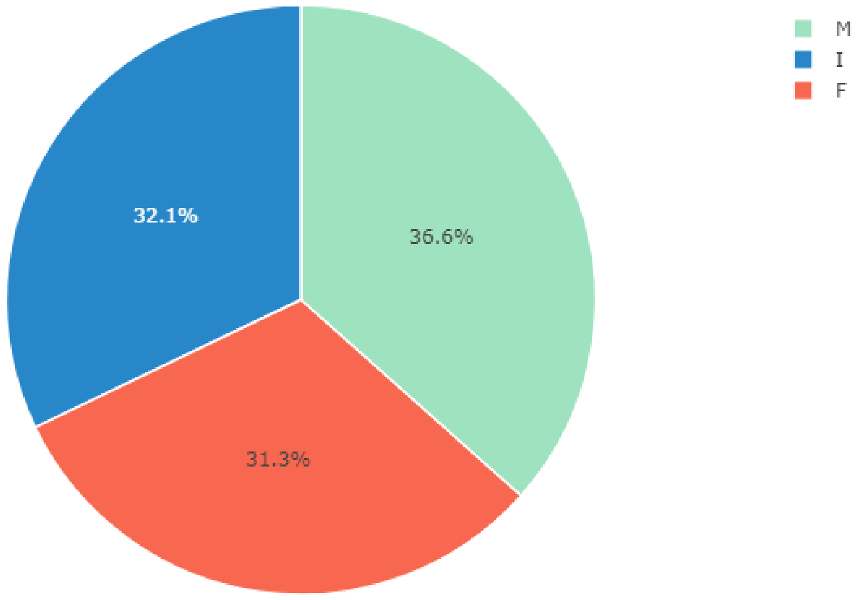


Figure 2

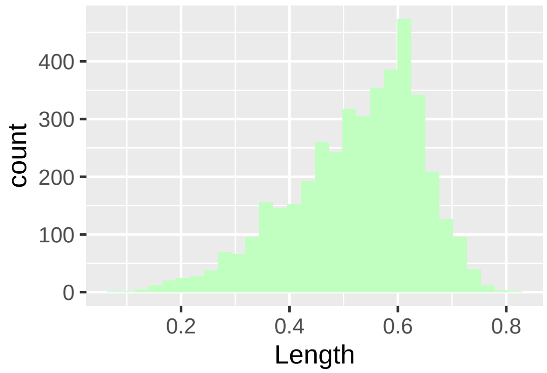


Figure 3

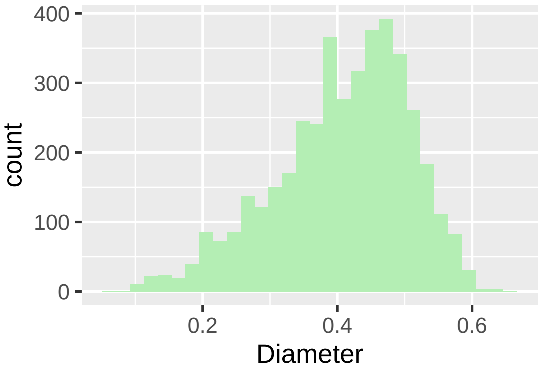


Figure 4

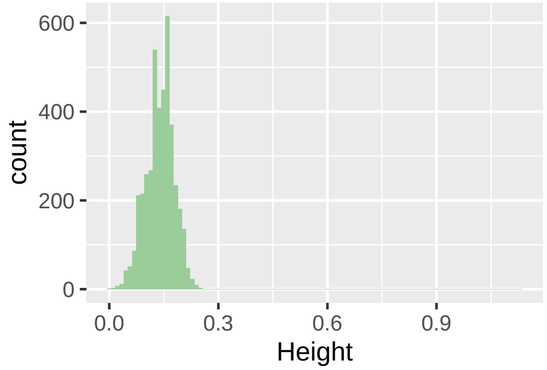


Figure 5

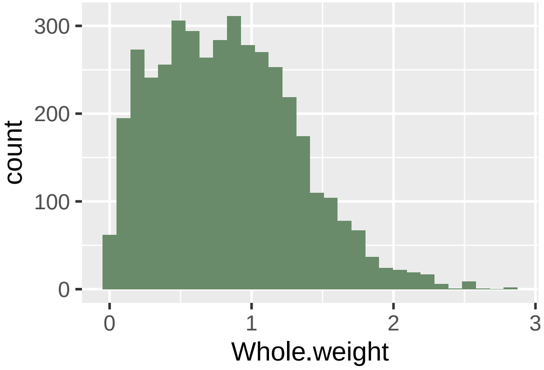


Figure 6

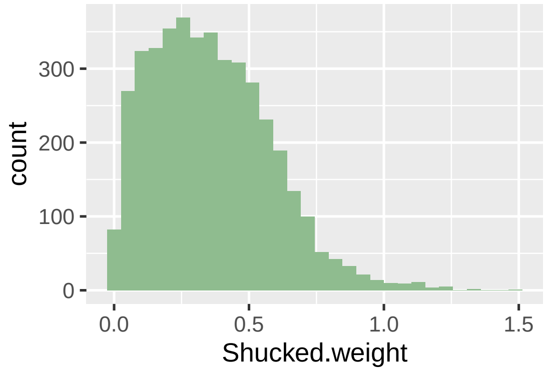


Figure 7

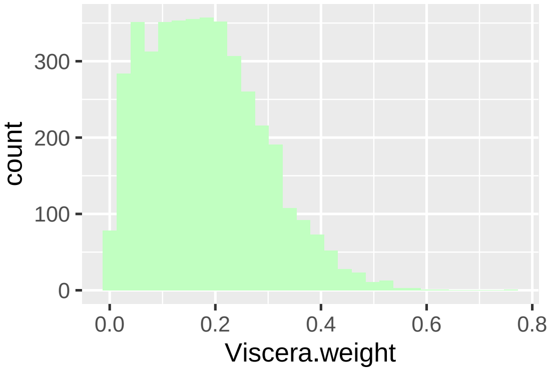


Figure 8

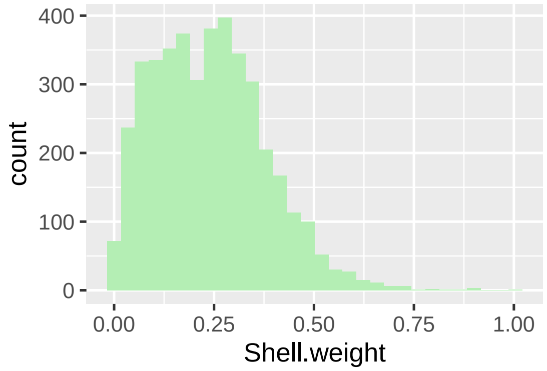


Figure 9

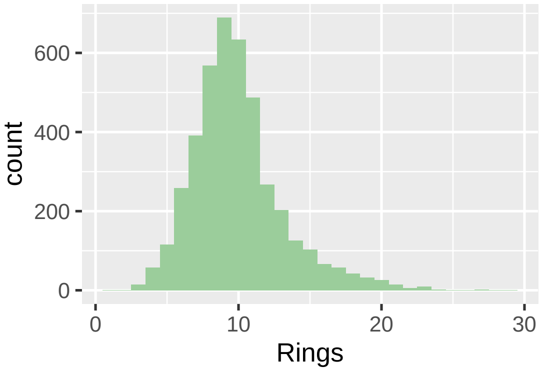


Figure 10

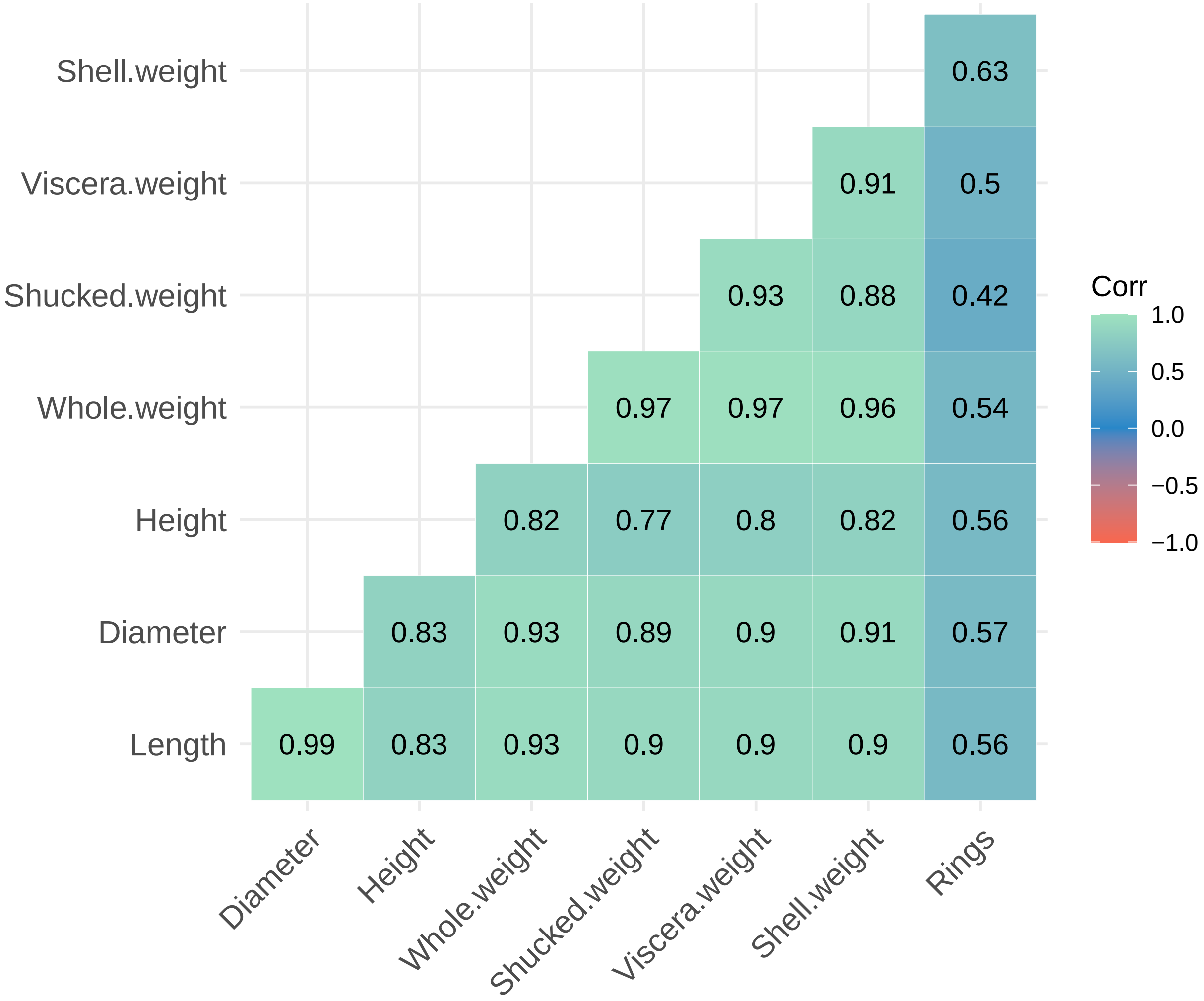
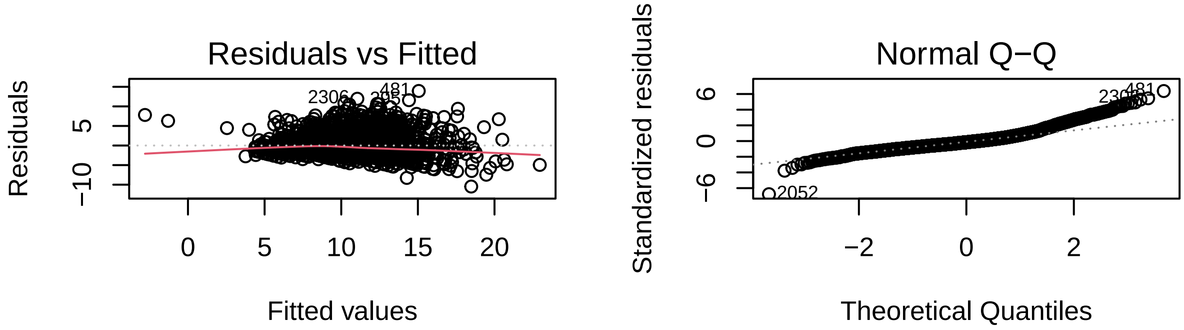
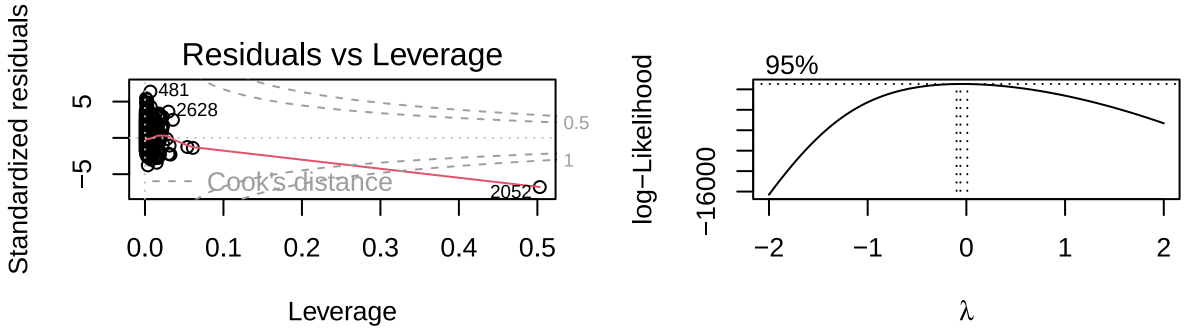
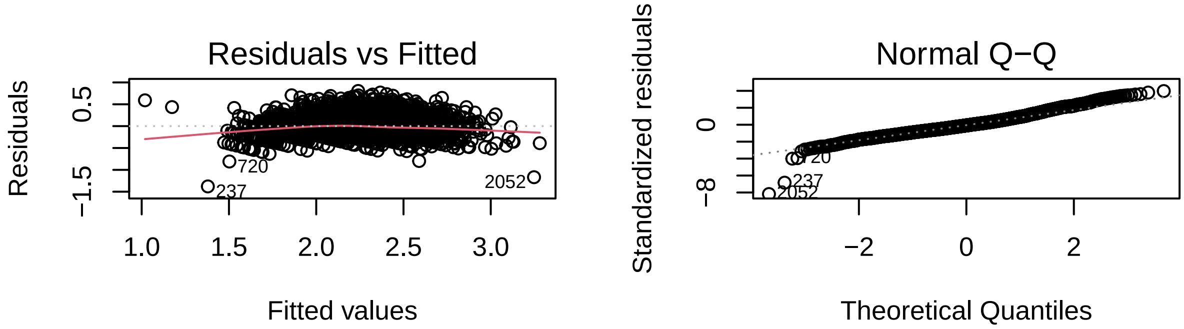

Shell.weight
Viscera.weight
Shucked.weight
Whole.weight
Height
Diameter
Length
0.63
0.91
0.5
0.93
0.88
0.42
0.97
0.97
0.96
0.54
0.82
0.77
0.8
0.82
0.56
0.83
0.93
0.89
0.9
0.91
0.57
0.99
0.83
0.93
0.9
0.9
0.9
0.56
Diameter
Height
Whole.weight
Shucked.weight
Viscera.weight
Shell.weight
Rings
Corr
1.0
0.5
0.0
−0.5
−1.0

Figure 11, 12, 13

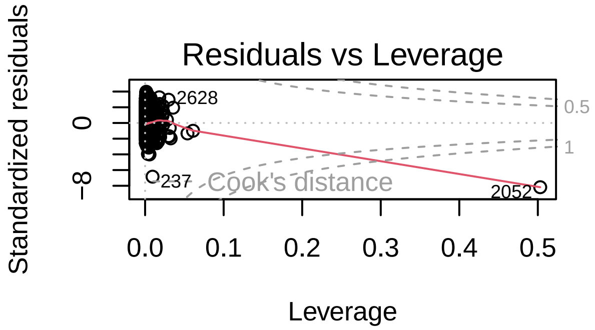


Figures 11, 12, 13, 14





Figures 15, 16, 17



Appendix 2

```r
library(ggplot2)
library(GGally)
library(plotly)
library(ggcorrplot)
library(car)
library(gridExtra)
library(MASS)
library(glmnet)

aba <- read.table("abalone.txt", sep = ",")
# Fixing Column names
names(aba) <-
  c(
    "Sex",
    "Length",
    "Diameter",
    "Height",
    "Whole.weight",
    "Shucked.weight",
    "Viscera.weight",
    "Shell.weight",
    "Rings"
  )

# Change Sex into factor
aba$Sex <- as.factor(aba$Sex)

# Initial Model
fullMod <- lm(Rings ~ ., data = aba)

# Now we use full log model
logFullMod <- lm(log(Rings) ~ ., data = aba)

vif(fullMod)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Sex              1.543449  2        1.114610
## Length          40.945763  1        6.398888
## Diameter        42.379841  1        6.509980
## Height           3.581369  1        1.892450
## Whole.weight   109.768710  1       10.477056
## Shucked.weight  28.550546  1        5.343271
## Viscera.weight  17.445012  1        4.176723
## Shell.weight    21.263272  1        4.611212
```

```
# Now we use full log model
logFullMod <- lm(log(Rings) ~ ., data = aba)
summary(logFullMod)
```

```
## 
## Call:
## lm(formula = log(Rings) ~ ., data = aba)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -1.37909 -0.13172 -0.01587  0.11120  0.80427 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     1.341185   0.026914  49.832  < 2e-16 ***
## SexI           -0.092485   0.009452  -9.785  < 2e-16 ***
## SexM            0.008926   0.007694   1.160  0.24605    
## Length          0.533049   0.166998   3.192  0.00142 ** 
## Diameter        1.423575   0.205598   6.924 5.06e-12 ***
## Height          1.206625   0.141805   8.509  < 2e-16 ***
## Whole.weight    0.608252   0.066961   9.084  < 2e-16 ***
## Shucked.weight -1.657046   0.075449 -21.963  < 2e-16 ***
## Viscera.weight -0.835499   0.119425  -6.996 3.05e-12 ***
## Shell.weight    0.606814   0.103823   5.845 5.46e-09 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 0.2025 on 4167 degrees of freedom
## Multiple R-squared:  0.5991, Adjusted R-squared:  0.5982 
## F-statistic: 691.8 on 9 and 4167 DF,  p-value: < 2.2e-16
```

```
# We will drop observations 237 and 2052 since
# it is harming our assumption of normality in residuals
aba <- aba[-c(237, 2052), ]

# Male sex is insignificant, so lets make new column where
# abalone is infant or not
aba$Infant <- as.factor(ifelse(aba$Sex == 'I', 'Y', 'N'))

# Clean model by removing Length, and Viscera weight
```

```r
redMod <-
  lm(log(Rings) ~ Infant + Diameter + Height + Whole.weight, data = aba)
summary(redMod)
```

```
##
## Call:
## lm(formula = log(Rings) ~ Infant + Diameter + Height + Whole.weight,
##     data = aba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.28577 -0.15073 -0.03279  0.11363  0.86321
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.306416   0.027303   47.85   <2e-16 ***
## InfantY      -0.104956   0.009094  -11.54   <2e-16 ***
## Diameter      1.598008   0.104530   15.29   <2e-16 ***
## Height        3.632715   0.215262   16.88   <2e-16 ***
## Whole.weight -0.222528   0.019728  -11.28   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.224 on 4170 degrees of freedom
## Multiple R-squared:  0.5031, Adjusted R-squared:  0.5026
## F-statistic:  1056 on 4 and 4170 DF,  p-value: < 2.2e-16
```

```r
### RIDGE REGRESSION
fullMod <- lm(Rings ~ ., data = aba)
logMod <- lm(log(Rings) ~ ., data = aba)

guess <-
  lm(
    log(Rings) ~ Infant + Diameter + Height + Whole.weight + Shucked.weight +
      Shell.weight,
    data = aba
  )

response <- aba$Rings
predictors <-
  data.matrix(aba[, c(
    "Infant",
    "Length",
    "Diameter",
    "Height",
    "Whole.weight",
    "Shucked.weight",
    "Viscera.weight",
    "Shell.weight"
  )])

ridgeModel <- glmnet(predictors, response, alpha = 0)

cv_model <- cv.glmnet(predictors, response, alpha = 0)
best_lambda <- cv_model$lambda.min

best_model <-
  glmnet(predictors, response, alpha = 0, lambda = best_lambda)
coef(best_model)
```

```
## 9 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)     4.8050468
## Infant         -0.8948481
## Length          1.5250822
## Diameter        5.1909877
## Height         21.9506558
## Whole.weight    0.7861055
## Shucked.weight -7.6029887
## Viscera.weight -2.5434894
## Shell.weight   12.0512875
```

```r
# Model suggested by ridge regression in mine.RMD, took out whole weight
ridgeSug <-
  lm(
    log(Rings) ~ Infant + Length +  Diameter + Height +
      Shucked.weight + Viscera.weight + Shell.weight,
    data = aba
  )
#summary(ridgeSug)

# Lets drop viscera weight
ridgeSug <-
  lm(log(Rings) ~ Infant  + Length +  Diameter + Height +
       Shucked.weight + Shell.weight,
     data = aba)
summary(ridgeSug)
```

```
##
## Call:
## lm(formula = log(Rings) ~ Infant + Length + Diameter + Height +
##     Shucked.weight + Shell.weight, data = aba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79163 -0.13330 -0.01758  0.11109  0.83488
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.333896   0.026197  50.918  < 2e-16 ***
## InfantY        -0.095571   0.008199 -11.656  < 2e-16 ***
## Length          0.379711   0.165732   2.291    0.022 *
## Diameter        1.302831   0.205709   6.333 2.65e-10 ***
## Height          2.327363   0.199101  11.689  < 2e-16 ***
## Shucked.weight -1.114422   0.034851 -31.977  < 2e-16 ***
## Shell.weight    1.209153   0.061214  19.753  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2019 on 4168 degrees of freedom
## Multiple R-squared:  0.5966, Adjusted R-squared:  0.5961
## F-statistic:  1028 on 6 and 4168 DF,  p-value: < 2.2e-16
```

```
# Stepwise Regression
nullModel <- lm(log(Rings) ~ 1, data = aba)
fullModI <- lm(log(Rings) ~ . - Sex, data = aba)
```

```
# Length is the last to be added via step AIC, as well
# as it is highly correlated with variables (diameter = 0.99), we will drop
# BIC also suggests we drop Length
postStepModel <-
  lm(log(Rings) ~ Infant + Diameter + Height + Shucked.weight + Shell.weight,
    data = aba)
summary(postStepModel)
```

```
##
## Call:
## lm(formula = log(Rings) ~ Infant + Diameter + Height + Shucked.weight +
##     Shell.weight, data = aba)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79746 -0.13338 -0.01796  0.11211  0.83301
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.35305    0.02484   54.47   <2e-16 ***
## InfantY        -0.09452    0.00819  -11.54   <2e-16 ***
## Diameter        1.72100    0.09493   18.13   <2e-16 ***
## Height          2.35568    0.19882   11.85   <2e-16 ***
## Shucked.weight -1.09671    0.03400  -32.26   <2e-16 ***
## Shell.weight    1.20329    0.06119   19.66   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.202 on 4169 degrees of freedom
## Multiple R-squared:  0.5961, Adjusted R-squared:  0.5956
## F-statistic:  1231 on 5 and 4169 DF,  p-value: < 2.2e-16
```

```
# Reducing the abalone dataset
aba_red <-
  aba[, c("Height",
          "Shucked.weight",
          "Shell.weight",
          "Diameter",
          "Infant",
          "Rings")]
head(aba_red, 10)
```

```
##    Height Shucked.weight Shell.weight Diameter Infant Rings
## 1   0.095         0.2245        0.150    0.365      N    15
## 2   0.090         0.0995        0.070    0.265      N     7
## 3   0.135         0.2565        0.210    0.420      N     9
## 4   0.125         0.2155        0.155    0.365      N    10
## 5   0.080         0.0895        0.055    0.255      Y     7
## 6   0.095         0.1410        0.120    0.300      Y     8
## 7   0.150         0.2370        0.330    0.415      N    20
## 8   0.125         0.2940        0.260    0.425      N    16
## 9   0.125         0.2165        0.165    0.370      N     9
## 10  0.150         0.3145        0.320    0.440      N    19
```

```
# Now we build a model with all our variabels and interactions
intModel <- lm(log(Rings) ~ . ^ 2, data = aba_red)
summary(intModel)
```

```
##
```

```
## Call:
## lm(formula = log(Rings) ~ .^2, data = aba_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.00767 -0.12436 -0.01731  0.10257  0.88396
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   0.84931    0.09231   9.201  < 2e-16 ***
## Height                        8.83831    1.21936   7.248 5.01e-13 ***
## Shucked.weight               -4.27703    0.26876 -15.914  < 2e-16 ***
## Shell.weight                  5.35773    0.45977  11.653  < 2e-16 ***
## Diameter                      3.35677    0.34607   9.700  < 2e-16 ***
## InfantY                      -0.06098    0.06419  -0.950  0.34221
## Height:Shucked.weight         0.31981    1.26642   0.253  0.80064
## Height:Shell.weight           6.33208    2.26123   2.800  0.00513 **
## Height:Diameter             -21.40292    3.49424  -6.125 9.90e-10 ***
## Height:InfantY                0.54434    0.57404   0.948  0.34305
## Shucked.weight:Shell.weight  -0.34069    0.40164  -0.848  0.39635
## Shucked.weight:Diameter       7.06436    0.70384  10.037  < 2e-16 ***
## Shucked.weight:InfantY        0.93495    0.13326   7.016 2.65e-12 ***
## Shell.weight:Diameter        -9.26856    1.27925  -7.245 5.12e-13 ***
## Shell.weight:InfantY         -0.38206    0.23462  -1.628  0.10351
## Diameter:InfantY             -0.69799    0.26794  -2.605  0.00922 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1914 on 4159 degrees of freedom
## Multiple R-squared:  0.6385, Adjusted R-squared:  0.6372
## F-statistic: 489.7 on 15 and 4159 DF,  p-value: < 2.2e-16
```

```
# Drop all non-signicant variables: height:shucked weight, height:infant,
# shucked.weight:shellweight, shell.weight:infant

intModel.red <-
  lm(
    log(Rings) ~ Height + Shucked.weight + Shell.weight + Diameter + Infant +
      Height:Diameter + Shucked.weight:Diameter + Shucked.weight:Infant +
      Diameter:Infant,
    data = aba_red
  )
summary(intModel.red)
```

```
##
```

```
## Call:
## lm(formula = log(Rings) ~ Height + Shucked.weight + Shell.weight +
##     Diameter + Infant + Height:Diameter + Shucked.weight:Diameter +
##     Shucked.weight:Infant + Diameter:Infant, data = aba_red)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8755 -0.1266 -0.0181  0.1036  0.8295
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.86915    0.06420  13.538  < 2e-16 ***
## Height                   11.45374    0.69816  16.405  < 2e-16 ***
## Shucked.weight           -2.33382    0.16988 -13.738  < 2e-16 ***
## Shell.weight              1.66960    0.06752  24.729  < 2e-16 ***
## Diameter                  2.68164    0.17153  15.634  < 2e-16 ***
## InfantY                  -0.14457    0.05008  -2.887  0.00391 **
## Height:Diameter         -21.97905    1.55021 -14.178  < 2e-16 ***
## Shucked.weight:Diameter   2.93929    0.32156   9.141  < 2e-16 ***
## Shucked.weight:InfantY    0.56110    0.11135   5.039 4.88e-07 ***
## Diameter:InfantY         -0.22929    0.19302  -1.188  0.23494
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1939 on 4165 degrees of freedom
## Multiple R-squared:  0.6282, Adjusted R-squared:  0.6274
## F-statistic: 781.8 on 9 and 4165 DF,  p-value: < 2.2e-16
```

```r
# Removing non significant variable Diameter:Infant
intModel.red <-
  lm(
    log(Rings) ~ Height + Shucked.weight + Shell.weight + Diameter + Infant +
      Height:Infant + Height:Diameter + Shucked.weight:Diameter +
      Shucked.weight:Infant,
    data = aba_red
  )
summary(intModel.red)
```

```
##
## Call:
## lm(formula = log(Rings) ~ Height + Shucked.weight + Shell.weight +
##     Diameter + Infant + Height:Infant + Height:Diameter + Shucked.weight:Diameter +
##     Shucked.weight:Infant, data = aba_red)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88208 -0.12675 -0.02021  0.10518  0.83143
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.81681    0.06545  12.479  < 2e-16 ***
## Height                   12.27594    0.80677  15.216  < 2e-16 ***
## Shucked.weight           -2.45135    0.17762 -13.801  < 2e-16 ***
## Shell.weight              1.66836    0.06733  24.780  < 2e-16 ***
## Diameter                  2.74128    0.15355  17.853  < 2e-16 ***
## InfantY                  -0.11610    0.03982  -2.915  0.00357 **
## Height:InfantY           -1.05688    0.45335  -2.331  0.01979 *
## Height:Diameter         -23.37651    1.68056 -13.910  < 2e-16 ***
## Shucked.weight:Diameter   3.17709    0.34097   9.318  < 2e-16 ***
## Shucked.weight:InfantY    0.62970    0.09647   6.528 7.48e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1938 on 4165 degrees of freedom
## Multiple R-squared:  0.6285, Adjusted R-squared:  0.6277
## F-statistic:   783 on 9 and 4165 DF,  p-value: < 2.2e-16
```

```r
# BIC suggests we drop interaction between height:infant
intModel.final <-
  lm(
    log(Rings) ~ Height + Shucked.weight + Shell.weight + Diameter + Infant+
      Height:Diameter + Shucked.weight:Diameter + Shucked.weight:Infant,
    data = aba_red
  )
summary(intModel.final)
```

```
##
## Call:
## lm(formula = log(Rings) ~ Height + Shucked.weight + Shell.weight +
##     Diameter + Infant + Height:Diameter + Shucked.weight:Diameter +
##     Shucked.weight:Infant, data = aba_red)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87340 -0.12638 -0.01934  0.10330  0.83261
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.91888    0.04868  18.875  < 2e-16 ***
## Height                  11.07601    0.62158  17.819  < 2e-16 ***
## Shucked.weight          -2.25556    0.15659 -14.404  < 2e-16 ***
## Shell.weight             1.67646    0.06727  24.920  < 2e-16 ***
## Diameter                 2.54760    0.12920  19.718  < 2e-16 ***
## InfantY                 -0.20083    0.01629 -12.329  < 2e-16 ***
## Height:Diameter        -21.14424    1.38183 -15.302  < 2e-16 ***
## Shucked.weight:Diameter   2.80996    0.30258   9.287  < 2e-16 ***
## Shucked.weight:InfantY    0.44693    0.05624   7.947 2.44e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1939 on 4166 degrees of freedom
## Multiple R-squared:  0.628,  Adjusted R-squared:  0.6273
## F-statistic: 879.3 on 8 and 4166 DF,  p-value: < 2.2e-16
```