Herman, Scott
Predict 401 Section 58
TITLE: Abalone Exploratory Data Analysis 2

INTRODUCTION

The purpose of this assignment is to further understand the relationship between abalone physical characteristics to more accurately identify and predict abalone age. Our findings will enable abalone harvesters to devise more useful decisions which will promote their harvests. This analysis builds upon our findings from an earlier analysis that observed an abalone data set that contains twelve different variables defined by various abalone physical characteristics and consists of 4,141 observations. A random sample of 500 observations was utilized throughout this analysis.

RESULTS

The first step in this analysis will explore the relationship between abalone Shuck and abalone volume to determine whether or not these variables are independent from one another. To test for independence, we will utilize Pearson's Chi-squared Test. In reviewing these results, which are shown in Figure 1 below, we can see that our observed p value of .07 is greater than our critical p value which allows us to reject the assumption that abalone shuck size and abalone volume are independently related. This means that it does appear that there is a relationship between these two variables, and we will need to explore this further to understand exactly how shuck and volume are related.

Figure 1:

```
        Pearson's Chi-squared test

data:  shuck_volume[1:2, 1:2]
X-squared = 323.21, df = 1, p-value < 2.2e-16
```

```
> pchisq(chisq,df=1,lower.tail=FALSE)
[1] 0.06964401
```

Since it does appear that a relationship exists between shuck size and volume, we will create a new variable which will give us the ratio between shuck size and volume for a given abalone. This newly defined variable will be used to measure the relationship between abalone ratio and class to determine if there is substantial variation between the two variables. This is shown in Figure 2 below. We will do the same for ratio and sex, displayed in Figure 3. The boxplots, as well as the analysis of variance calculation both indicate that there appears to be a much greater variance between abalone ratio and class, than there is between ratio and sex or between sex and class.

Figure 2:

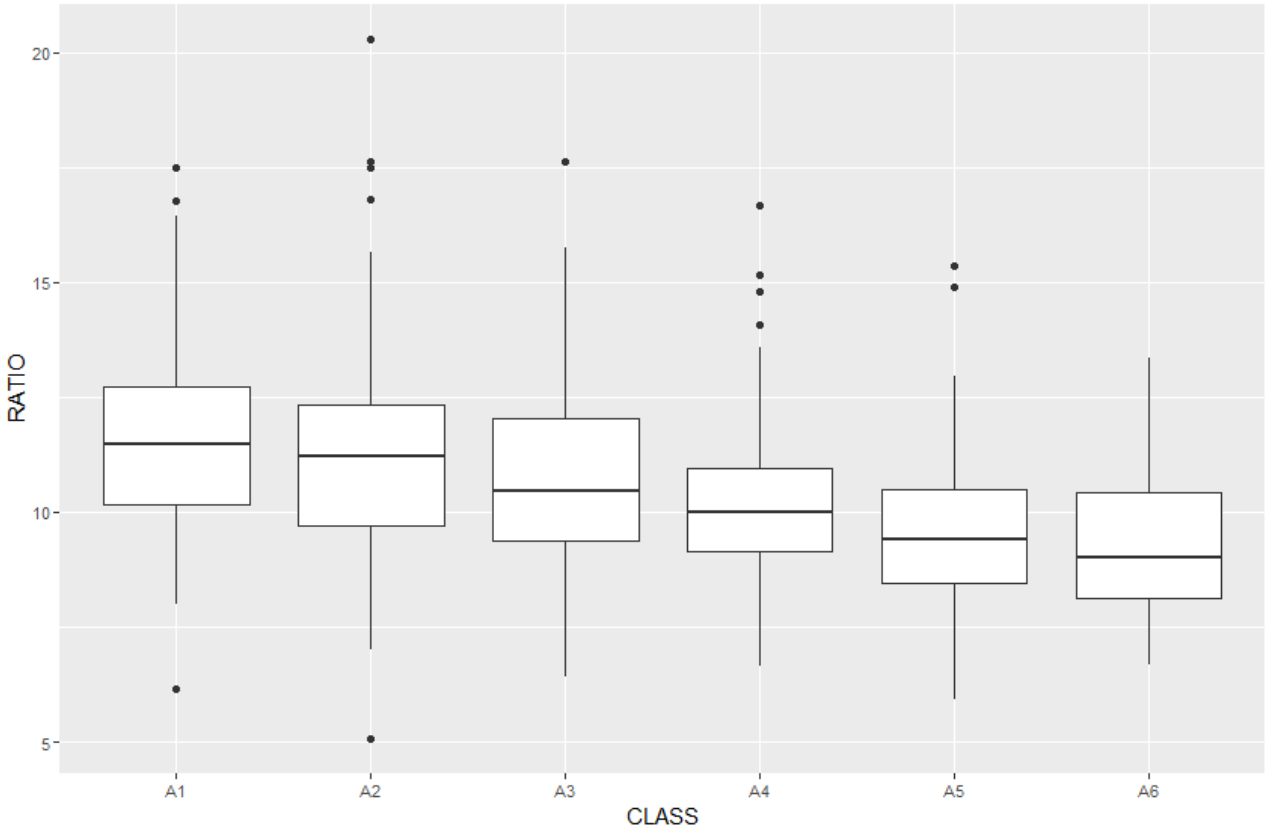


Figure 3:

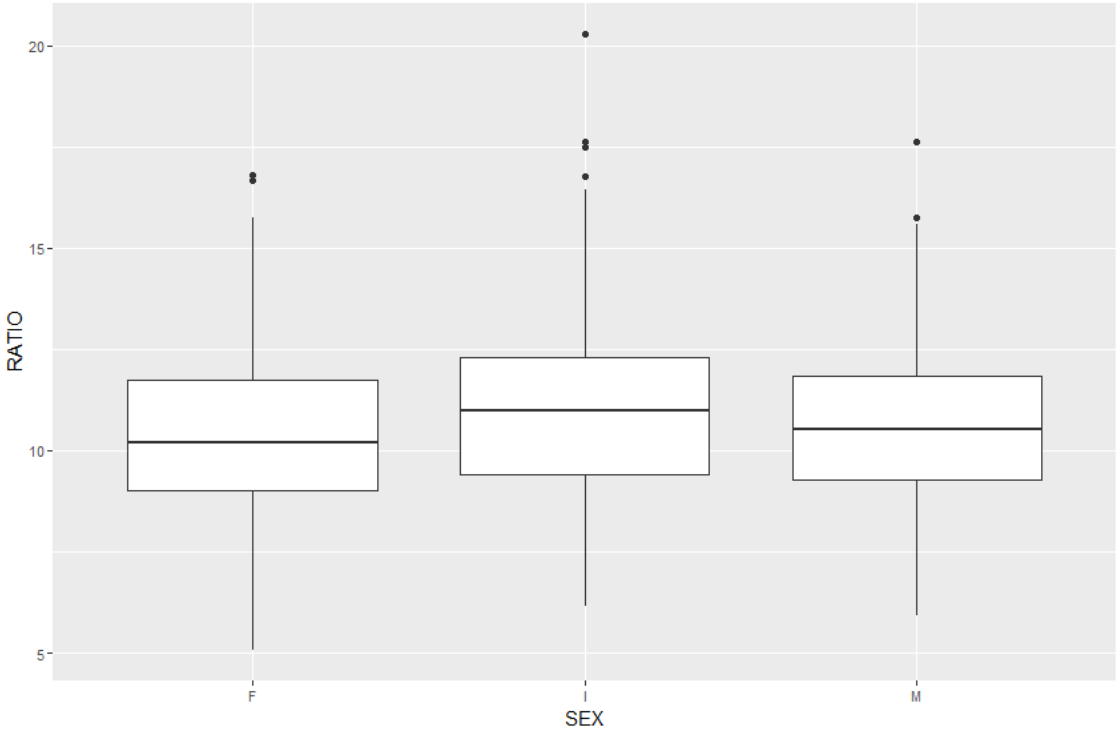


Figure 4 shows the analysis of variance between abalone ratio and class, ratio and sex, and sex and class. It appear that class represents the only statistically significant variance as indicated by our low p value. Figure 5 confirms these results when analyzing ratio to class as compared to ratio and sex. Again, the low p value here indicates significant variance between classes.

Figure 4:

```
  RATIO      SEX     CLASS
1     F 10.42679 11.658530
2     I 11.02888 11.255956
3     M 10.60720 10.736015
4     F 10.42679 10.190765
5     I 11.02888  9.860751
6     M 10.60720  9.237746
> aov.1<-aov(RATIO~CLASS+SEX+CLASS*SEX,mydata)
> summary(aov.1)
             Df Sum Sq Mean Sq F value   Pr(>F)
CLASS         5  203.4   40.67   9.784 6.43e-09 ***
SEX           2    4.8    2.42   0.581    0.560
CLASS:SEX    10   26.1    2.61   0.627    0.791
Residuals   482 2003.7    4.16
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5:

```
> summary(aov.2)
             Df Sum Sq Mean Sq F value   Pr(>F)
CLASS         5  203.4   40.67   9.858 5.37e-09 ***
SEX           2    4.8    2.42   0.586    0.557
Residuals   492 2029.8    4.13
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 6 shows the results from our Tukey Test and displays the variance in means between ratio and class, as well as ratio and sex.

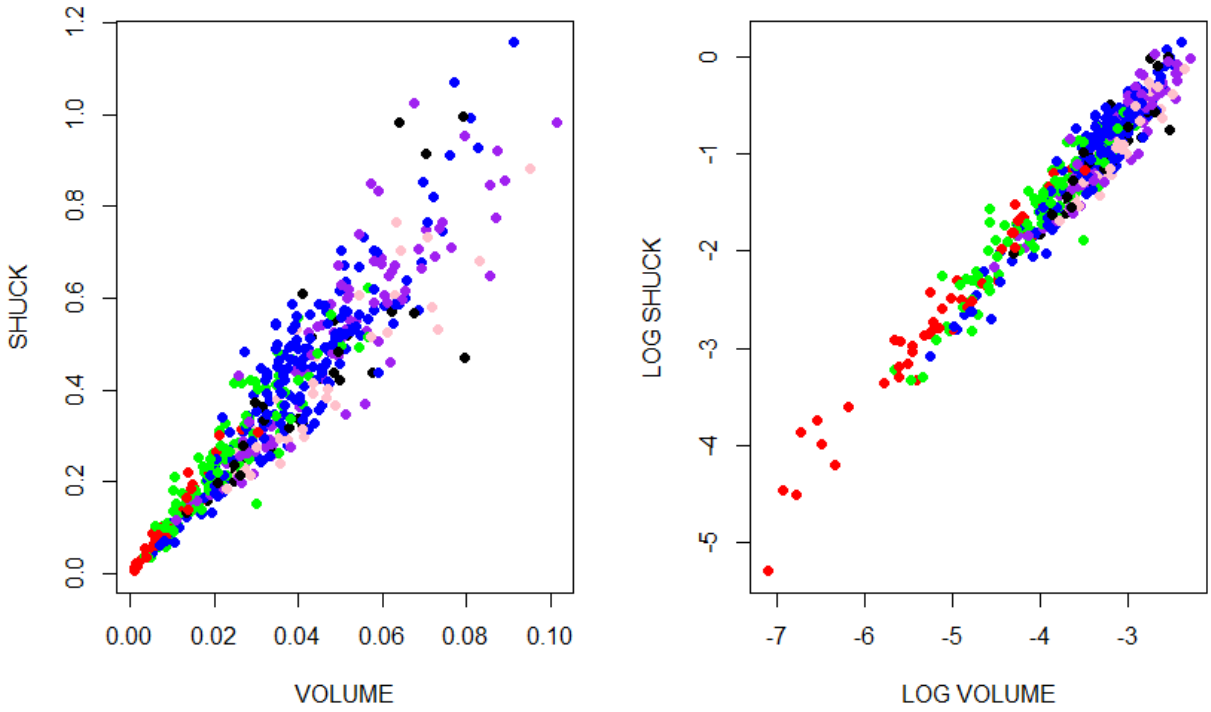
Figure 6:

```
> TukeyHSD(aov.2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = RATIO ~ CLASS + SEX, data = mydata)

$CLASS
            diff       lwr          upr     p adj
A2-A1 -0.4025742 -1.395244  0.59009557 0.8554223
A3-A1 -0.9225151 -1.861154  0.01612334 0.0572400
A4-A1 -1.4677656 -2.497456 -0.43807507 0.0007488
A5-A1 -1.7977797 -3.083863 -0.51169659 0.0010296
A6-A1 -2.4207844 -3.706867 -1.13470129 0.0000017
A3-A2 -0.5199409 -1.218578  0.17869624 0.2739224
A4-A2 -1.0651915 -1.882085 -0.24829831 0.0029108
A5-A2 -1.3952055 -2.518175 -0.27223557 0.0055278
A6-A2 -2.0182102 -3.141180 -0.89524028 0.0000058
A4-A3 -0.5452505 -1.295559  0.20505827 0.2999601
A5-A3 -0.8752646 -1.950769  0.20024012 0.1846703
A6-A3 -1.4982693 -2.573774 -0.42276459 0.0010875
A5-A4 -0.3300140 -1.485839  0.82581072 0.9643895
A6-A4 -0.9530188 -2.108844  0.20280601 0.1729769
A6-A5 -0.6230047 -2.012133  0.76612378 0.7942576

$SEX
          diff        lwr       upr     p adj
I-F 0.01219174 -0.5302479 0.5546313 0.9984620
M-F 0.19861165 -0.3163848 0.7136080 0.6363763
M-I 0.18641991 -0.3322495 0.7050893 0.6752336
```

The relationship of shuck in terms of volume is given by Figure 7a below and shows a positive relationship. In the far left corner, the data seem to be much more closely related, and in the far right corner we can see that the points are much further spread out which indicates a higher level of variance. As volume increases, there is more variability. We cannot apply a linear regression here because we do not have a constant variance. Though, it does appear that there is an underlying proportional linear relationship between these two variables. In an attempt to re-express this relationship we will apply a transformation to create a new scale of measurement that will better allow to run a linear regression test. This will be done to handle problems with heteroscedasticity.

We will take the logarithm of Shuck and the log of Volume to see if we can improve limit the variability between these two variables. The plot of Log Shuck in terms of Log Volume appear in Figure 7b and display a linear relationship with equal variability. This will allow us to use these newly created variable and plug them into a multiple regression model to interpret how these factors interact with our other variables in terms of predicting abalone age.

Figure 7a and Figure 7b:



In applying these variables into our fitted model, we will select the log of shuck as the baseline and compare this against log volume, class, and sex. The summary results of our fitted model are presented in Figure 8 and indicate that log shuck and log volume display both a strongly correlated relationship and one that is statistically significant given the low p-value. We do have some statistical significance in terms of the slightly negative linear relationship between Log Shuck and a few of the different Class levels as well. Specifically, Classes 4, 5 and 6. Though, there is a slightly positive relationship between Log Shuck and Sex, this relationship is insignificant given the larger p-values.

Figure 8:

```
Call:
lm(formula = L_SHUCK ~ L_VOLUME + CLASS + SEX, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77497 -0.11656 -0.00626  0.11160  0.62489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.56468    0.08012  32.009  < 2e-16 ***
L_VOLUME     1.02839    0.01585  64.890  < 2e-16 ***
CLASSA2     -0.06531    0.03583  -1.823  0.06898 .
CLASSA3     -0.12701    0.03961  -3.207  0.00143 **
CLASSA4     -0.18302    0.04418  -4.143 4.04e-05 ***
CLASSA5     -0.21944    0.04946  -4.436 1.13e-05 ***
CLASSA6     -0.27904    0.05029  -5.549 4.71e-08 ***
SEXI         0.01564    0.02551   0.613  0.54013
SEXM         0.02665    0.02029   1.313  0.18979
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1861 on 491 degrees of freedom
Multiple R-squared:  0.9475,    Adjusted R-squared:  0.9466
F-statistic:  1107 on 8 and 491 DF,  p-value: < 2.2e-16
```

We will now plot the residuals of our fitted model to understand the level of normality with this distribution. Ideally, we would want to see a normal distribution of residuals. The histogram below in Figure 9 displays the frequency of residuals and it does appear that these are normally distributed. Figure 10 utilizes a QQ plot to see how our model stands up against skewness and outliers. It appears that there is some departure from normality in the lower and upper corners of this plot, however, this mostly shows that our model does indicate a positive linear relationship that is normally distributed. Skewness is close to zero, but kurtosis is not close to 3 like we would want it to be.

The scatterplot in Figure 11 shows the residuals of log volume based upon class mostly show equal variance among these variables, and also appear to be normally distributed.

Figure 9:

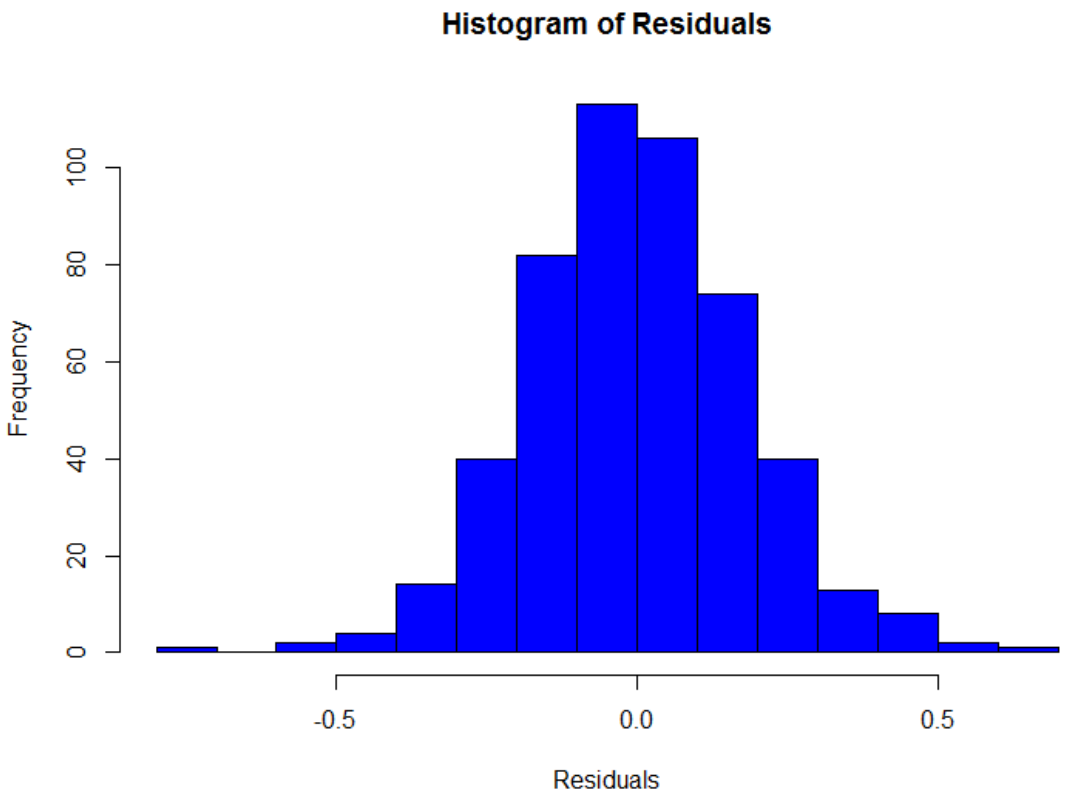
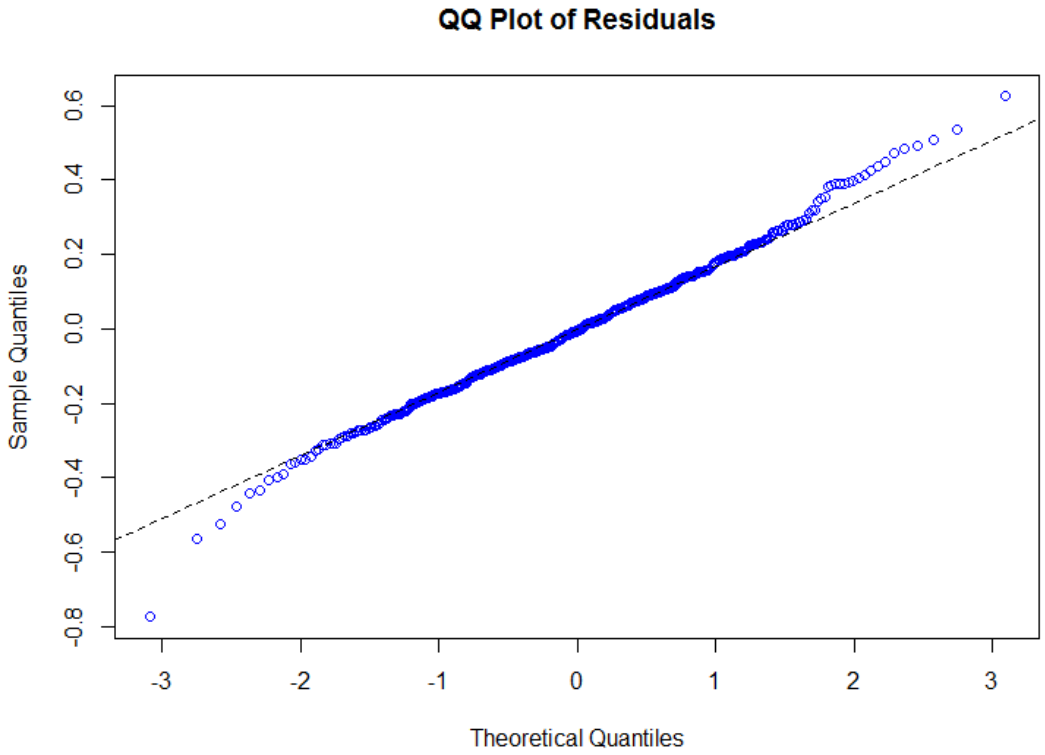


Figure 10:




```
> skewness(r)
[1] 0.04784899
> kurtosis(r)
[1] 0.7203053
```

Figure 11:

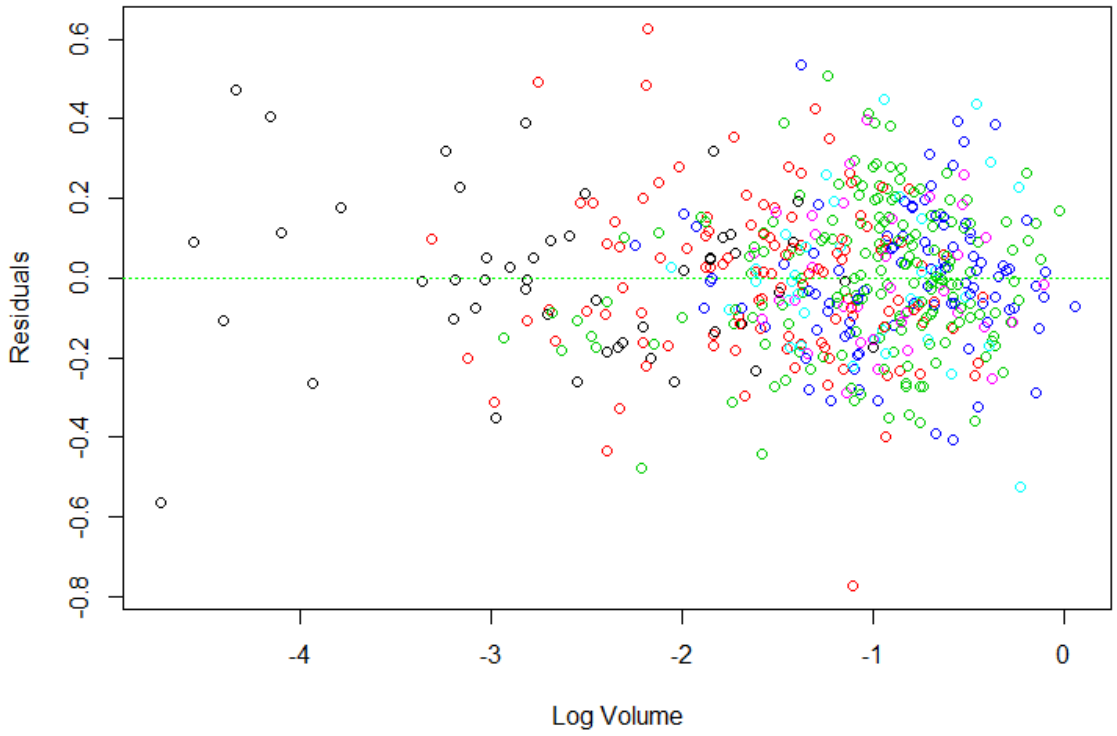
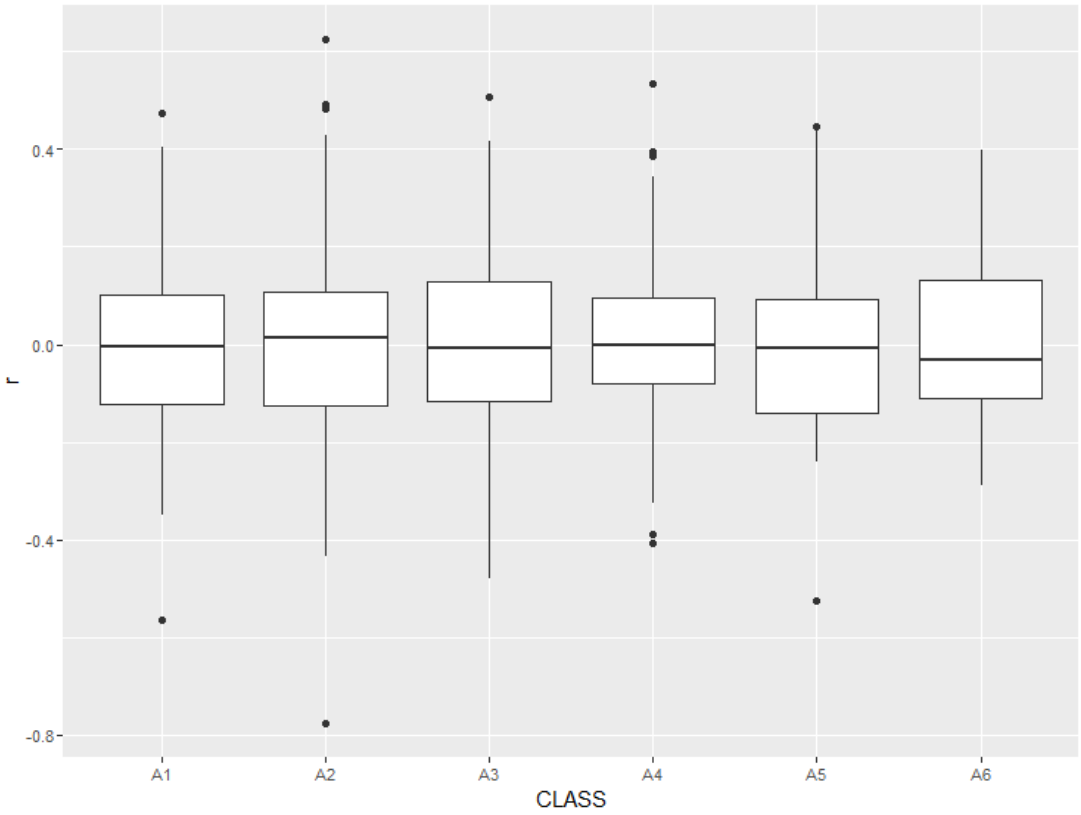


Figure 12 presents a box plot of the residuals based upon the various classes of abalone. These data points do appear to be normally distributed as well.

Figure 12:



Next we will explore the harvesting management of the abalone population. The harvesting of abalones presents us with the challenge of harvesting enough abalones to warrant the involved effort as well as minimizing the harvesting of infant abalone to protect the future of the abalone population. To help overcome this challenge, we will identify a cutoff value based upon abalone volume which will enable us to maximize adult abalone harvests while minimizing infant abalone harvested. We'll utilize three different calculations to identify the best cutoff value for the problem at hand.

First we'll calculate three vectors, one corresponding to infant proportion protected or not harvested at each volume and one with the adult proportion protected or not harvested at each volume. Then we'll calculate the volumes that split the harvest to 50% of each of the populations. Figure 14 below displays the adult and infant proportions based upon volume, along with the 50% split volumes for each. These points indicated in the plot show that 50% of infants are protected at a volume of 0.0164, and 50% of adults are protected at a volume of 0.0396.

Figure 13:

```
> prop.infants
  [1] 0.04575163 0.05228758 0.09150327 0.13725490 0.16339869
  [6] 0.19607843 0.22875817 0.26143791 0.28758170 0.32679739
 [11] 0.35947712 0.37254902 0.39869281 0.45751634 0.49019608
 [16] 0.50980392 0.54901961 0.59477124 0.62745098 0.64705882
 [21] 0.67973856 0.71895425 0.72549020 0.74509804 0.76470588
 [26] 0.77777778 0.79738562 0.82352941 0.84313725 0.84967320
 [31] 0.86274510 0.88235294 0.88235294 0.90196078 0.91503268
 [36] 0.92156863 0.92810458 0.92810458 0.93464052 0.95424837
 [41] 0.96078431 0.97385621 0.97385621 0.97385621 0.98039216
 [46] 0.98692810 0.99346405 1.00000000 1.00000000 1.00000000
 [51] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [56] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [61] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [66] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [71] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [76] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [81] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [86] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [91] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
 [96] 1.00000000 1.00000000 1.00000000 1.00000000 1.00000000
```

```
prop.adults
 [1] 0.000000000 0.000000000 0.000000000 0.000000000 0.008645533
 [6] 0.011527378 0.020172911 0.034582133 0.040345821 0.046109510
[11] 0.051873199 0.051873199 0.060518732 0.066282421 0.069164265
[16] 0.077809798 0.080691643 0.092219020 0.109510086 0.126801153
[21] 0.138328530 0.152737752 0.164265130 0.178674352 0.201729107
[26] 0.227665706 0.247838617 0.259365994 0.282420749 0.299711816
[31] 0.319884726 0.342939481 0.365994236 0.386167147 0.420749280
[36] 0.440922190 0.458213256 0.478386167 0.510086455 0.536023055
[41] 0.550432277 0.573487032 0.593659942 0.605187320 0.619596542
[46] 0.639769452 0.657060519 0.674351585 0.706051873 0.720461095
[51] 0.749279539 0.755043228 0.760806916 0.780979827 0.789625360
[56] 0.804034582 0.824207493 0.835734870 0.841498559 0.847262248
[61] 0.861671470 0.876080692 0.884726225 0.890489914 0.899135447
[66] 0.899135447 0.904899135 0.910662824 0.919308357 0.930835735
[71] 0.936599424 0.942363112 0.945244957 0.951008646 0.956772334
[76] 0.959654179 0.959654179 0.962536023 0.968299712 0.971181556
[81] 0.971181556 0.976945245 0.976945245 0.976945245 0.982708934
[86] 0.985590778 0.988472622 0.991354467 0.991354467 0.994236311
[91] 0.994236311 0.994236311 0.994236311 0.997118156 0.997118156
[96] 0.997118156 0.997118156 0.997118156 0.997118156 1.000000000
```

Figure 14:

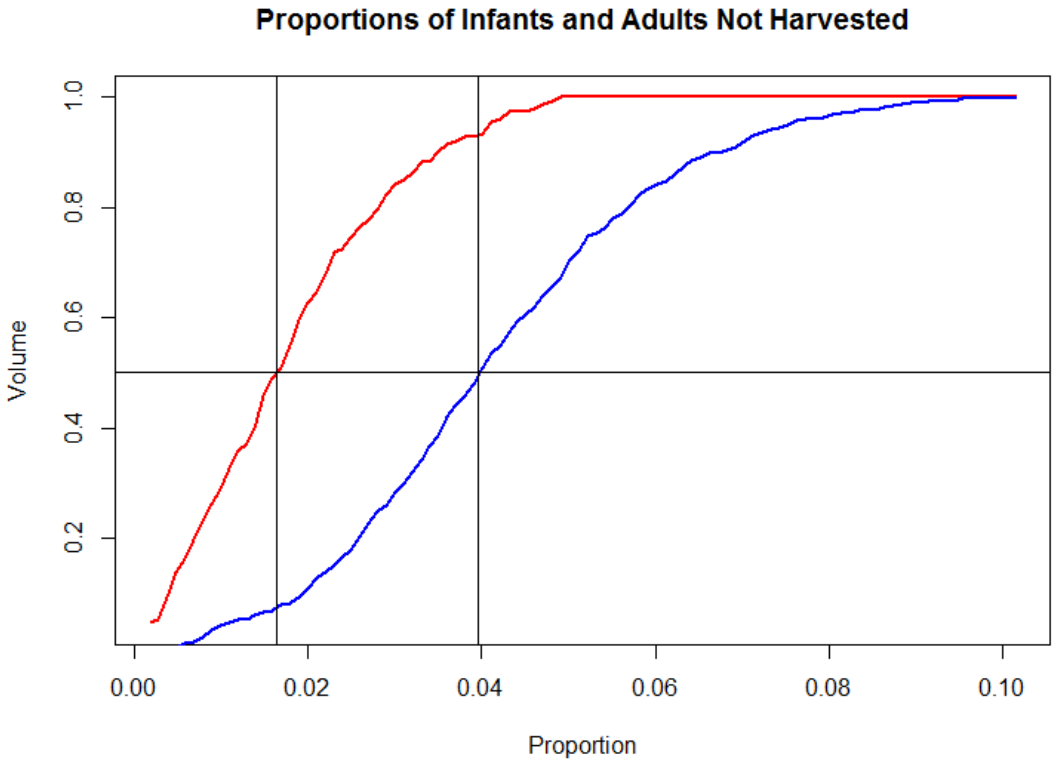
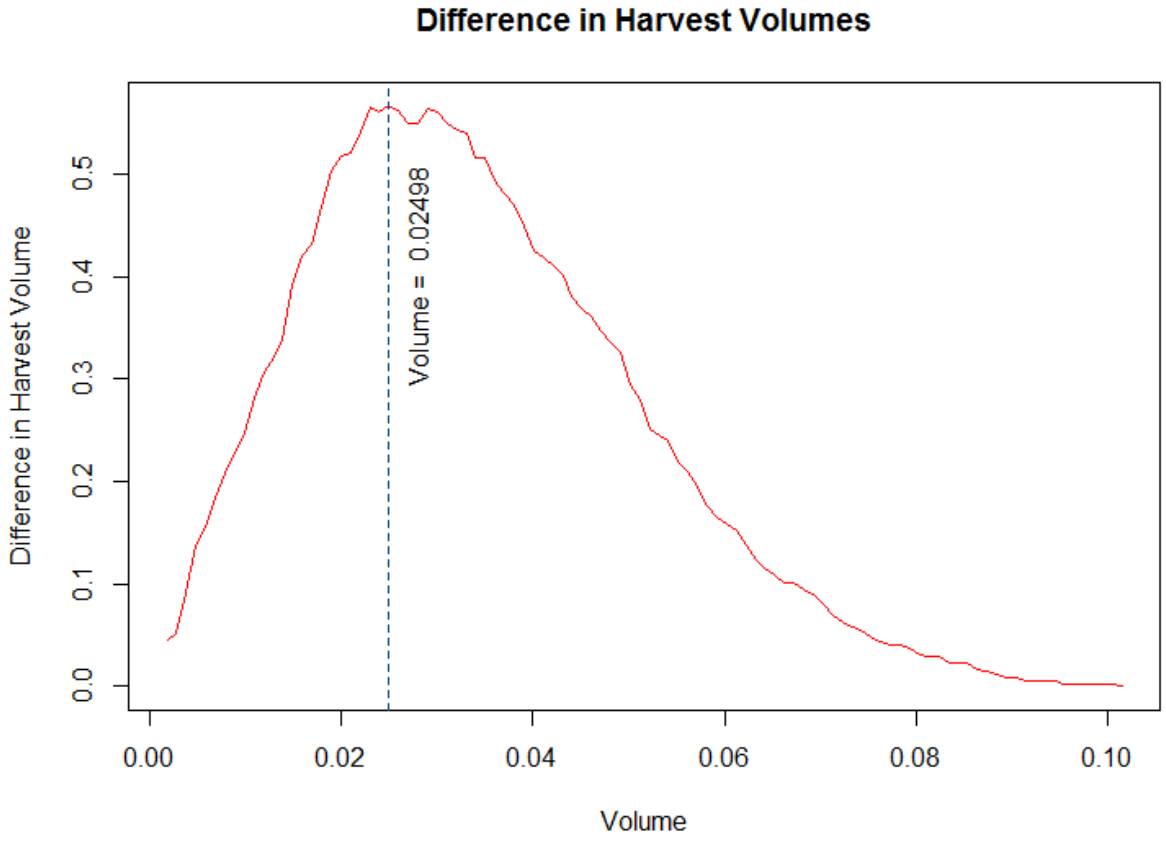


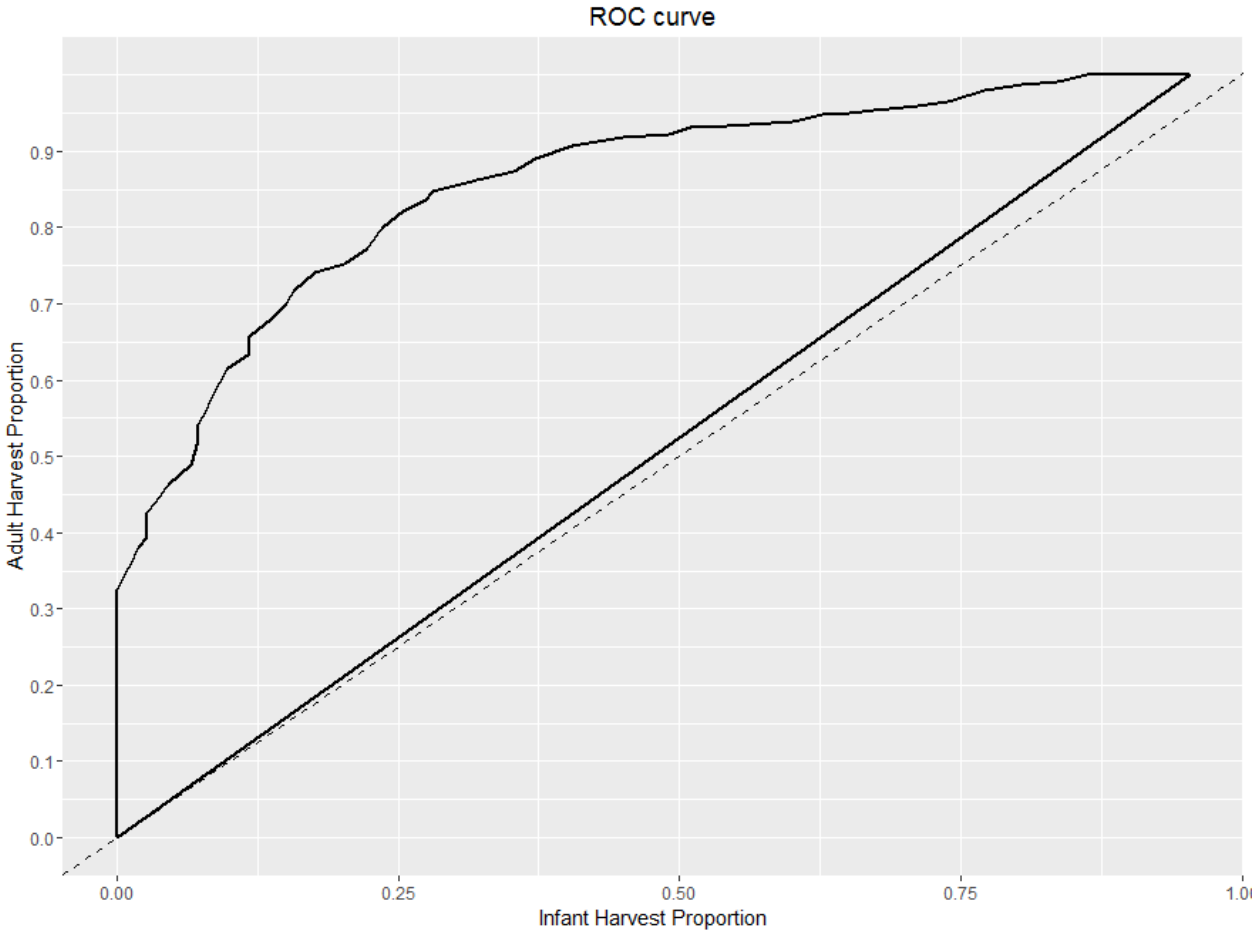
Figure 15 displays the difference in harvest proportions between adult and infant abalone. The calculated volume indicating the maximum difference is 0.02498. The harvest proportion for infants at this volume is 0.2549 and the adult proportion is 0.8213. Although this volume maximizes the difference in the harvest proportions we will consider some additional calculations for cutoffs to see if we can further minimize the proportion of infants harvested.

Figure 15:



Our secondary method for determining a cutoff considers the difference between harvesting an infant, a false positive, and harvesting an adult, a true positive. Figure 16 considers the volume values that gives us zero false positives, or no infants harvested. Utilizing the vectors we created in Figure 13, we can calculate the minimum volume for when no infants are harvested to be 0.04915. This corresponds to an adult harvest volume of 0.3256.

Figure 16:



The final method we'll utilize for calculating a potential cutoff volume will further explore class as a classifying characteristic of abalone age. Since we want to minimize the proportion of infants harvested, we will look at cutoff volumes greater than 0.034, 0.035, and 0.036. Our aim here is to find the minimum value for when there are no abalone in class A1 and A2 that are harvested. Figure 16 indicates that the lowest volume with no infants harvested is 0.035.

Figure 17:

```
> adults
       0.034     0.035     0.036
A1 0.0000000 0.0000000 0.0000000
A2 0.3962264 0.3584906 0.2830189
A3 0.7092199 0.6808511 0.6453901
A4 0.7108434 0.6987952 0.6987952
A5 0.5312500 0.5312500 0.5312500
A6 0.8000000 0.8000000 0.6666667
> infants
        0.034     0.035     0.036
A1 0.00000000 0.0000000 0.0000000
A2 0.01639344 0.0000000 0.0000000
A3 0.28571429 0.2571429 0.2285714
A4 0.50000000 0.3750000 0.3750000
A5 0.33333333 0.3333333 0.3333333
A6 0.40000000 0.4000000 0.4000000
```

Now that we've come up with three different cutoff levels for volume, we can compare them against one another to determine which cutoff will provide the best measure for maximizing the number of adult abalone harvested while limiting the number of infants harvested.

Our first cutoff value for volume of 0.02498325 will manage to harvest over 25% of infants and 82% of adults. Although this approach does reduce the proportion of infants harvested, we may want to look at an alternative cutoff level that will further minimize the proportion of infants harvested. The final cutoff value of 0.04915275 will result in an infant harvest proportion of zero and an adult proportion of 33%. This may be a bit conservative and may not yield a large enough population of adults for harvest.

Our secondary cutoff value of 0.035 will result in an infant abalone harvest of less than 10%, while bringing in 62% of adults harvested. This appears to be the best approach when considered against the other two.

Figure 18:

```
> volume.value[which.max(difference)]   # volume
[1] 0.02498325
> 1-prop.infants[which.max(difference)] # infant proportion
[1] 0.254902
> 1-prop.adults[which.max(difference)]  # adult proportion
[1] 0.8213256
> max(1-prop.adults[(1 - prop.infants) == 0]) # adult proportion
[1] 0.3256484
> min(volume.value[(1 - prop.infants) == 0]) # volume
[1] 0.04915275
> sum(mydata[idxa, 11] >= 0.035) / nrow(mydata[idxa, ]) # adult proportion
[1] 0.6167147
> sum(mydata[idxi, 11] >= 0.035) / nrow(mydata[idxi, ]) # infant proportion
[1] 0.09803922
```

CONCLUSION

The results of our analysis demonstrate that physical characteristics can be utilized as a predictor of abalone age. Our cutoff values allowed us to determine how well volume and class can be utilized to maximize the number of adult abalone harvested while minimizing the proportion of infants harvested. If the harvesters use an abalone volume of 0.035 as the cutoff value for harvesting, they will be able to harvest less than 10% of infant abalone while managing to harvest 62% of the adult population. If this approach is used, it will help enable the growth of future abalone population.

APPENDIX:

```
setwd("C:/Users/hermsc01/Desktop/NW/401/R_Directory")

require(moments)

require(ggplot2)

require(rockchalk)

mydata<-read.csv("mydata.csv",sep="")

str(mydata)

head(mydata)

summary(mydata)


#1

shuck<-factor(mydata$SHUCK>median(mydata$SHUCK),labels=c("below","above"))

volume<-factor(mydata$VOLUME>median(mydata$VOLUME),labels=c("below","above"))

shuck_volume<-addmargins(table(shuck,volume))
```

```
shuck_volume
chisq.test(shuck_volume[1:2,1:2],correct=F)
pchisq(chisq,df=1,lower.tail=FALSE)



#2
mydata$RATIO<-mydata[,6]/mydata[,11]
head(mydata$RATIO)
data.frame(mydata$RATIO)
CLASS=mydata$CLASS
RATIO=mydata$Ratio
SEX=mydata$SEX

par(mfrow=c(1,2))
ggplot(mydata,aes(x=CLASS,y=RATIO))+
geom_boxplot()+
theme(axis.text=element_text(size=8))
ggplot(mydata,aes(x=SEX,y=RATIO))+
  geom_boxplot()+
  theme(axis.text=element_text(size=8))
par(mfrow = c(1, 1))



my<-aggregate(RATIO~SEX,data=mydata,mean)
mx<-aggregate(RATIO~CLASS,data=mydata,mean)
mx<-mx[,2]
overview<-cbind(my,mx)
colnames(overview)<-c("RATIO","SEX","CLASS")
overview
aov.1<-aov(RATIO~CLASS+SEX+CLASS*SEX,mydata)
```

```
summary(aov.1)
aov.2<-aov(RATIO~CLASS+SEX,mydata)
summary(aov.2)
TukeyHSD(aov.2)

#3
L_SHUCK<-log(mydata$SHUCK,base=exp(1))
head(L_SHUCK)
L_VOLUME<-log(mydata$VOLUME,base=exp(1))
head(L_VOLUME)



par(mfrow=c(1,2))
plot(mydata$VOLUME, mydata$SHUCK, main = "",
    xlab = "VOLUME", ylab = "SHUCK",
    col = c("red", "green", "blue", "purple", "black", "pink")[mydata$CLASS],
    pch = 16)

plot(L_VOLUME,L_SHUCK, main = "",
    xlab = "LOG VOLUME", ylab = "LOG SHUCK",
    col = c("red", "green", "blue", "purple", "black", "pink")[mydata$CLASS],
    pch = 16)
par(mfrow = c(1, 1))

#4
model=lm(formula=L_SHUCK~L_VOLUME + CLASS + SEX, data=mydata)
summary(model)

#5
r <- residuals(model)
```

```
fitt <- fitted(model)
par(mfrow = c(1,1))
hist(r, col = "blue", main = "Histogram of Residuals", xlab = "Residuals")
qqnorm(r, col = "blue", pch = 1, main = "QQ Plot of Residuals")
qqline(r, col = "black", lty = 2, lwd = 1)
skewness(r)
kurtosis(r)
plot(fitt,r, main = " ", xlab = "Log Volume",
    ylab = "Residuals", col = CLASS)
abline(h = 0, lty = 3, col = "green")
abline(h = 87.65, lty = 2, col = "blue")
abline(h = -87.65, lty = 2, col = "blue")

ggplot(mydata,aes(x=CLASS,y=r),xlab="CLASS", ylab="Residuals",col=CLASS)+
 geom_boxplot()+
 theme(axis.text=element_text(size=8))



#6
idxi <- mydata[,1]=="I"
idxf <- mydata[,1]=="F"
idxm <- mydata[,1]=="M"
idxa <- mydata[, 1] == "M" | mydata[, 1] == "F"
max.v <- max(mydata$VOLUME)
min.v <- min(mydata$VOLUME)
delta <- (max.v - min.v)/100
prop.infants <- numeric(0)
volume.value <- numeric(0)
total <- length(mydata[idxi,1])
for (k in 1:100)
```

```
{
  value <- min.v + k*delta
  volume.value[k] <- value
  prop.infants[k] <- sum(mydata$VOLUME[idxi] <= value)/total
}
prop.infants

prop.adults <- numeric(0)
volume.value <- numeric(0)
total <- length(mydata[idxa,1])
for (k in 1:100)
{
  value <- min.v + k*delta
  volume.value[k] <- value
  prop.adults[k] <- sum(mydata$VOLUME[idxa] <= value)/total
}
prop.adults

n.infants <- sum(prop.infants <= 0.5)
split.infants <- min.v + (n.infants + 0.5)*delta
plot(volume.value, prop.infants, col = "green", main = "Proportion of Infants Not Harvested",
    type = "l", lwd = 2)
abline(h=0.5)
abline(v = split.infants)

n.adults <- sum(prop.adults <= 0.5)
split.adults <- min.v + (n.adults + 0.5)*delta

plot(volume.value, prop.infants, col = "red",
    main = "Proportions of Infants and Adults Not Harvested",xlab="Proportion",ylab="Volume",
```

```
    type = "l", lwd = 2)
lines(volume.value, prop.adults, col = "blue", type = "l", lwd = 2)
abline(h=0.5)
abline(v=split.adults)
abline(v=split.infants)

#how to find intercepts?

#7
difference <- (1-prop.adults)-(1-prop.infants)
max.diff <- max(difference)
index <- difference ==max.diff
volume.value[index]
plot(volume.value,difference, col = "red",
    main = "Difference in Harvest Volumes",
    xlab = "Volume", ylab = "Difference in Harvest Volume",
    type = "l", lwd = 1)
abline(v = volume.value[which.max(difference)], lty = 2, col = "dodgerblue4")
text(volume.value[which.max(difference)] + 0.003, 0.4,
    paste("Volume = ",
        round(volume.value[which.max(difference)], 5)), srt = 90)
#8
mydata <- cbind(mydata,prop.adults,prop.infants)
ggplot(mydata,aes(1-prop.infants,1-prop.adults))+geom_path(size = 1)+
  labs(title= "ROC curve",y = "Adult Harvest Proportion", x = "Infant Harvest
Proportion")+geom_abline(linetype=2)+ scale_y_continuous(breaks = round(seq(min(volume.value[(1-
prop.infants)==0]), max(prop.adults), by = .1),1))



#9
classes <- levels(mydata$CLASS)
```

```
cutoffs <- c(0.034, 0.035, 0.036)
adults <- matrix(NA, nrow = length(classes), ncol = length(cutoffs),
                 dimnames = list(classes, cutoffs))
infants <- matrix(NA, nrow = length(classes), ncol = length(cutoffs),
                  dimnames = list(classes, cutoffs))
adults.temp <- numeric(3)
infants.temp <- numeric(3)

for(i in classes) {
  for(j in 1:length(cutoffs)) {
    adults.temp[j] <- sum(mydata[mydata$CLASS == i & (mydata$SEX == "F" | mydata$SEX == "M"),
                                 11] >= cutoffs[j]) / nrow(mydata[mydata$CLASS == i & (mydata$SEX == "F" |
mydata$SEX == "M"), ])
    infants.temp[j] <- sum(mydata[mydata$CLASS == i & mydata$SEX == "I", 11] >= cutoffs[j]) /
      nrow(mydata[mydata$CLASS == i & mydata$SEX == "I", ])
  }
  adults[i, ] <- adults.temp
  infants[i, ] <- infants.temp
}

adults
infants

#10
#7
volume.value[which.max(difference)]     # volume
1-prop.infants[which.max(difference)]   # infant proportion
1-prop.adults[which.max(difference)]    # adult proportion

#8
```

```
max(1-prop.adults[(1 - prop.infants) == 0]) # adult proportion
min(volume.value[(1 - prop.infants) == 0]) # volume

#9
sum(mydata[idxa, 11] >= 0.035) / nrow(mydata[idxa, ]) # adult proportion
sum(mydata[idxi, 11] >= 0.035) / nrow(mydata[idxi, ]) # infant proportion
```