Data Analysis Assignment #2

McCoy, Jason

Data Analysis #2

```
# Perform the following steps to start the assignment.

# 1) Load/attach the following packages via library():  flux, ggplot2, gridExtra, moments, roc
kchalk, car.
# NOTE:  packages must be installed via install.packages() before they can be loaded.

library(flux)
library(ggplot2)
library(gridExtra)
library(moments)
library(rockchalk)
library(car)
library(cowplot)

# 2) Use the "mydata.csv" file from Assignment #1 or use the file posted on the course site.
Reading
# the files into R will require sep = "" or sep = " " to format data properly.  Use str() to c
heck file
# structure.

mydata = read.csv("mydata.csv", sep = ",")
# mydata <- read.csv(file.path("c:...", "mydata.csv"), sep = ",")
#mydata <- read.csv(file.path("c:/Rabalone/", "mydata.csv"), sep = ",")
str(mydata)
```

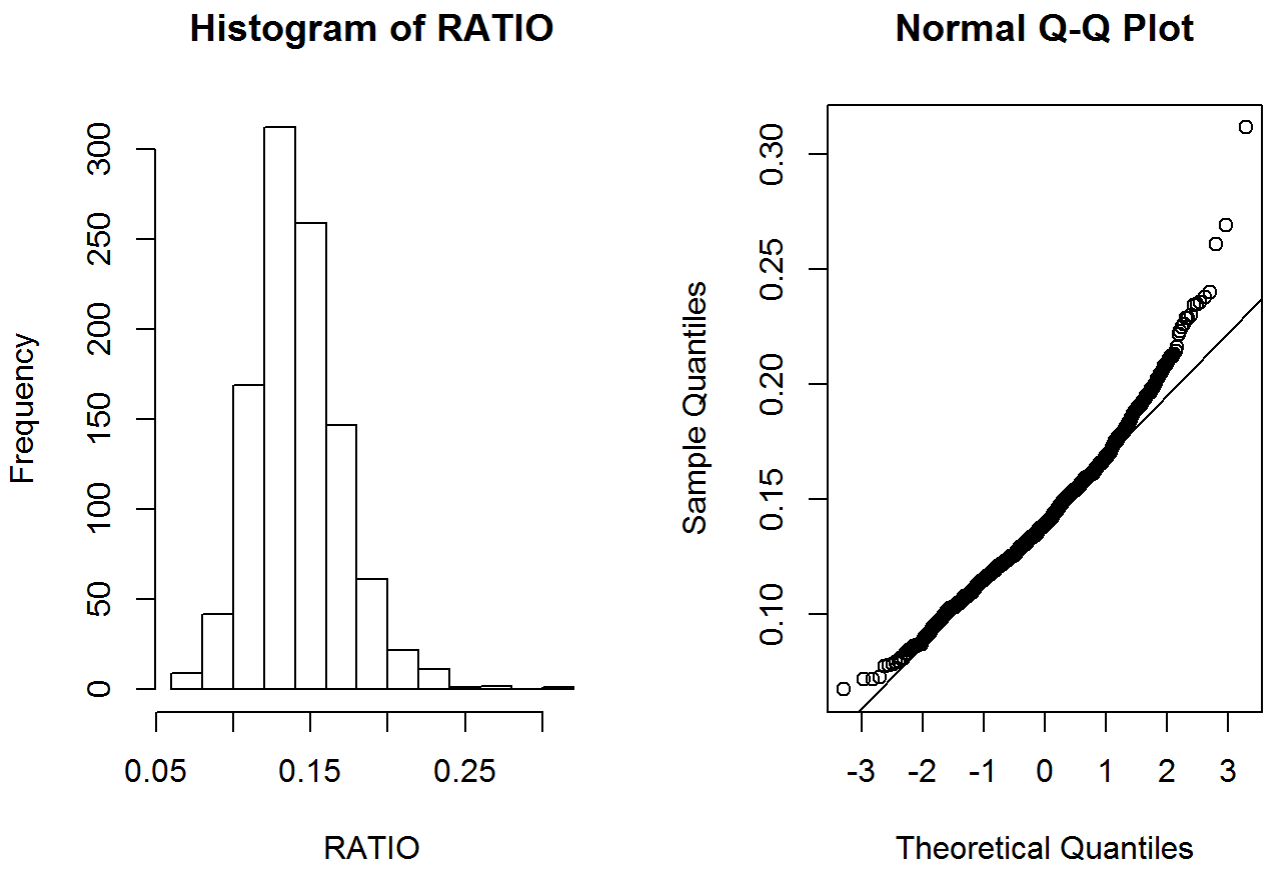
```
## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : Factor w/ 5 levels "A1","A2","A3",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ VOLUME: num  28.7 8.1 163.4 12.2 59.7 ...
##  $ RATIO : num  0.15 0.147 0.269 0.185 0.165 ...
```

---

(1)(a) (1 point) Form a histogram and QQ plot using RATIO. Calculate skewness and kurtosis using ‘rockchalk.’ Be aware that with ‘rockchalk’, the kurtosis value has 3.0 subtracted from it which differs from the ‘moments’ package.

```
par(mfrow=c(1,2))
```

```
hist(mydata$RATIO, main="Histogram of RATIO", xlab="RATIO")
qqnorm(mydata$RATIO)
qqline(mydata$RATIO)
```




```
skewness(mydata$RATIO, na.rm = TRUE, unbiased = TRUE)
```

```
## [1] 0.7147056
```

```
#Positively skewed
kurtosis(mydata$RATIO, na.rm = TRUE, unbiased = TRUE)
```

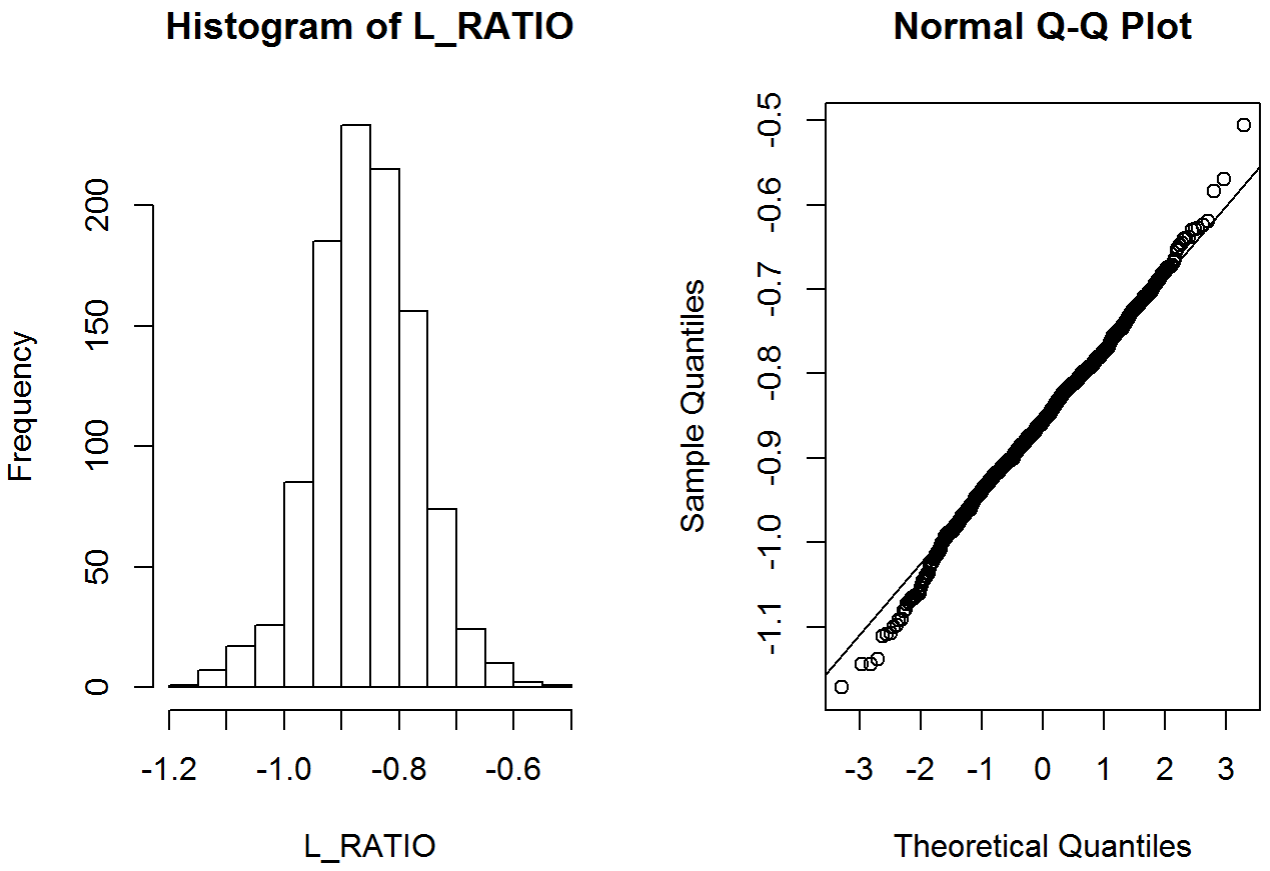
```
## [1] 1.667298
```

```
#Leptokurtic
```

(1)(b) (2 points) Tranform RATIO using log10() to create L_RATIO (see Kabacoff Section 8.5.2, p. 199-200). Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis. Create a display of five boxplots of L_RATIO differentiated by CLASS.

```
mydata$L_RATIO = log10(mydata$RATIO)
par(mfrow=c(1,2))
hist(mydata$L_RATIO, main="Histogram of L_RATIO", xlab="L_RATIO")
qqnorm(mydata$L_RATIO)
```

```
qqline(mydata$L_RATIO)
```




```
skewness(mydata$L_RATIO, na.rm = TRUE, unbiased = TRUE)
```

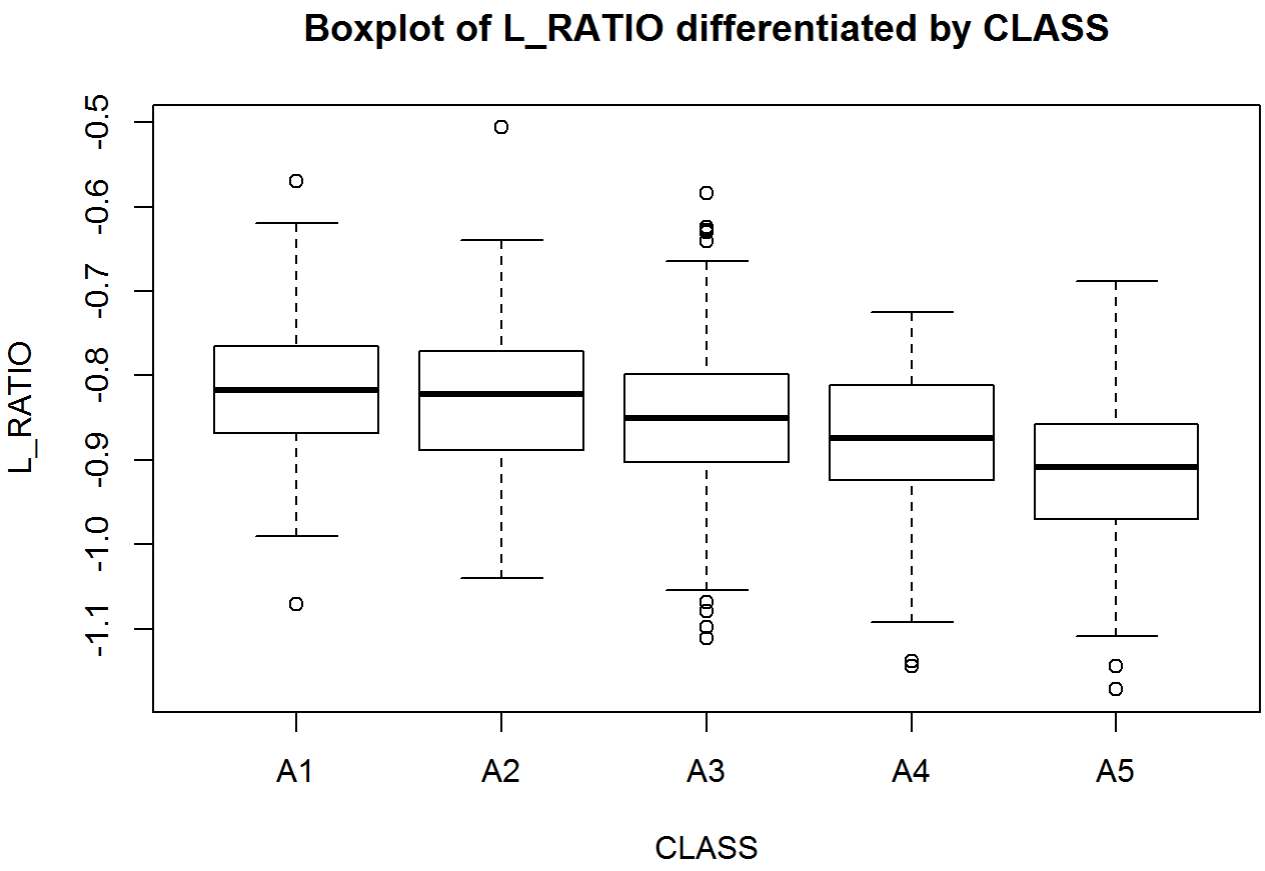
```
## [1] -0.09391548
```

```
#Slight negative skewness (Close to no skewness)
kurtosis(mydata$L_RATIO, na.rm = TRUE, unbiased = TRUE)
```

```
## [1] 0.5354309
```

```
# Close to mesokurtic. Approximately normal.

#Boxplots
par(mfrow=c(1,1))
boxplot(mydata$L_RATIO ~ mydata$CLASS, xlab="CLASS", ylab="L_RATIO", main="Boxplot of L_RATIO
differentiated by CLASS")
```



(1)(c) (1 point) Test the homogeneity of variance across classes using the bartlett.test() (see Kabacoff Section 9.2.2, p. 222).

```
bartlett.test(L_RATIO ~ CLASS, data = mydata)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  L_RATIO by CLASS
## Bartlett's K-squared = 3.1891, df = 4, p-value = 0.5267
```

Question (2 points): Based on steps 1.a, 1.b and 1.c, which variable RATIO or L_RATIO exhibits better conformance to a normal distribution with homogeneous variances across age classes? Why?

Answer: L_RATIO is closer to a normal distribution as compared to RATIO. This is because fro L_RATIO, skewness and kurtosis are both closer to 0. Moreover, Bartlett test for homogeneity of variance shows that P-value > 0.05 which means that the hypothesis of homogeneity of variances can not be rejected. Hence, L_RATIO exhibits better conformance to a normal distribution with homogeneous variances across age classes.

---

(2)(a) (2 points) Perform an analysis of variance with aov() on L_RATIO using CLASS and SEX as the independent variables (see Kabacoff chapter 9, p. 212-229). Assume equal variances. Peform two analyses. First, fit a model with the interaction term CLASS:SEX. Then, fit a model without CLASS:SEX. Use summary() to obtain the analysis of variance tables (Kabacoff chapter 9, p. 227).

```
anova1 = aov(L_RATIO ~ CLASS*SEX, data=mydata)
summary(anova1)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## CLASS          4  1.055 0.26384  38.370 < 2e-16 ***
## SEX            2  0.091 0.04569   6.644 0.00136 **
## CLASS:SEX      8  0.027 0.00334   0.485 0.86709
## Residuals   1021  7.021 0.00688
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova2 = aov(L_RATIO ~ CLASS+SEX, data=mydata)
summary(anova2)
```

```
##               Df Sum Sq Mean Sq F value  Pr(>F)
## CLASS          4  1.055 0.26384  38.524 < 2e-16 ***
## SEX            2  0.091 0.04569   6.671 0.00132 **
## Residuals   1029  7.047 0.00685
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question (2 points): Compare the two analyses. What does the non-significant interaction term suggest about the relationship between L_RATIO and the factors CLASS and SEX?

Answer: The non-significant interaction term suggests that the joint effect of CLASS and SEX on L_RATIO is not significantly different (or higher) than the sum of the individual effects of CLASS and SEX. The individual effects of both CLASS and SEX are significant as p-value is less than 0.05, however since interaction term has p-value greater than 0.05, it is not significant and need not be considered.

(2)(b) (2 points) For the model without CLASS:SEX (i.e. an interaction term), obtain multiple comparisons with the TukeyHSD() function. Interpret the results at the 95% confidence level (TukeyHSD() will adjust for unequal sample sizes).

```
TukeyHSD(anova2)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = L_RATIO ~ CLASS + SEX, data = mydata)
##
## $CLASS
##               diff         lwr          upr     p adj
## A2-A1 -0.01248831 -0.03876038  0.013783756 0.6919456
## A3-A1 -0.03426008 -0.05933928 -0.009180867 0.0018630
## A4-A1 -0.05863763 -0.08594237 -0.031332896 0.0000001
## A5-A1 -0.09997200 -0.12764430 -0.072299703 0.0000000
## A3-A2 -0.02177176 -0.04106269 -0.002480831 0.0178413
## A4-A2 -0.04614932 -0.06825638 -0.024042262 0.0000002
## A5-A2 -0.08748369 -0.11004316 -0.064924223 0.0000000
## A4-A3 -0.02437756 -0.04505283 -0.003702280 0.0114638
## A5-A3 -0.06571193 -0.08687025 -0.044553605 0.0000000
```

```
## A5-A4 -0.04133437 -0.06508845 -0.017580286 0.0000223
##
## $SEX
##              diff          lwr           upr     p adj
## I-F -0.015890329 -0.031069561 -0.0007110968 0.0376673
## M-F  0.002069057 -0.012585555  0.0167236690 0.9412689
## M-I  0.017959386  0.003340824  0.0325779478 0.0111881
```

```
#lwr and upr gives the upper and the lower bounds of the 95% confidence interval. If p-adj < 0
.05, the difference is significant and if p adj > 0.05, then the difference is not significant
. For example: The difference in L_RATIO between classes A2 and A1 is not significant while th
e difference in L_RATIO between classes A3 and A1 is statistically significant. A1 and A2 can
be combined, likewise M and F can also be combined as differences between them is not statisti
cally different.
```

Question (2 points): First, interpret the trend in coefficients across age classes. What is this indicating about L_RATIO? Second, do these results suggest male and female abalones can be combined into a single category labeled as 'adults'? If not, why not? ***Answer: The trend is that L_RATIO declines with age. This can be inferred from the negative "diff" in the R output. In DA-1, when RATIO was plotted by sex and class, we observed that RATIO also declined with age for all the three sex categories. RATIO=SHUCK/VOLUME. As age increases, RATIO declines because increase in volume is more than that in SHUCK weight. As a result L_RATIO also declines with age. Yes, male and females abalones can be combined into a single category labeled as 'adults'. This is because the difference between male and female is not statistically significant (p adj > 0.05), however different of both males and females with infants has p-adj of less than 0.05 and hence statistically significant. The difference in L_RATIO between males and females is not significant and hence these two can be combined together as 'adults'. In DA-1, when mean RATIO was plotted by sex and Class, we saw that infants differ on average from 'adults' across the classes.***

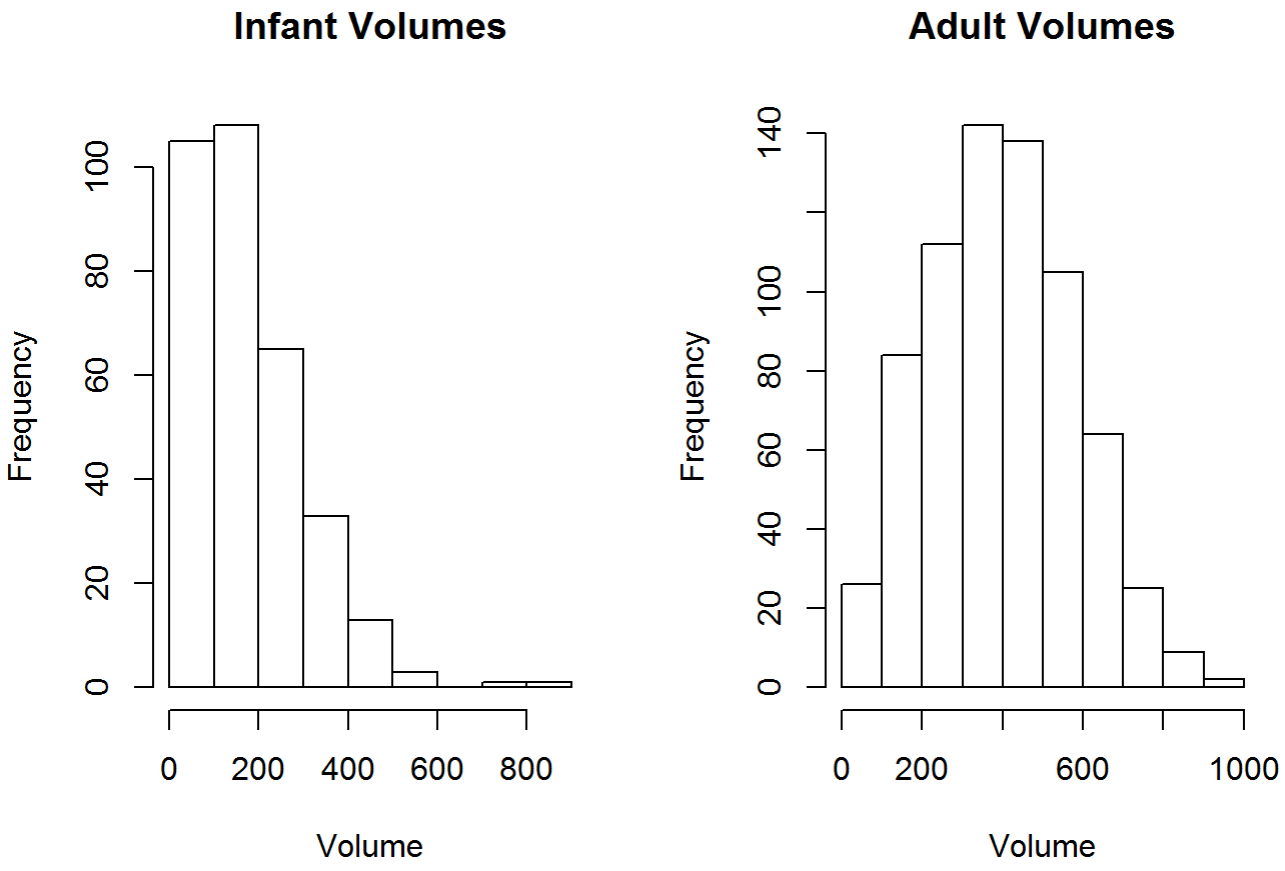
---

(3)(a) (2 points) Use combineLevels() from the 'rockchalk' package to combine "M" and "F" into a new level, "ADULT". This will necessitate defining a new variable, TYPE, in mydata which will have two levels: "I" and "ADULT". Use par() to form two histograms of VOLUME. One should display infant volumes, and the other: adult volumes.

```
mydata$TYPE = combineLevels(mydata$SEX, levs = c("F","M"), "ADULT")
```

```
## The original levels F I M
## have been replaced by I ADULT
```

```
par(mfrow=c(1,2))
hist(mydata$VOLUME[mydata$TYPE=="I"], main="Infant Volumes", xlab="Volume", ylab="Frequency")
hist(mydata$VOLUME[mydata$TYPE=="ADULT"], main="Adult Volumes", xlab="Volume", ylab="Frequency
")
```

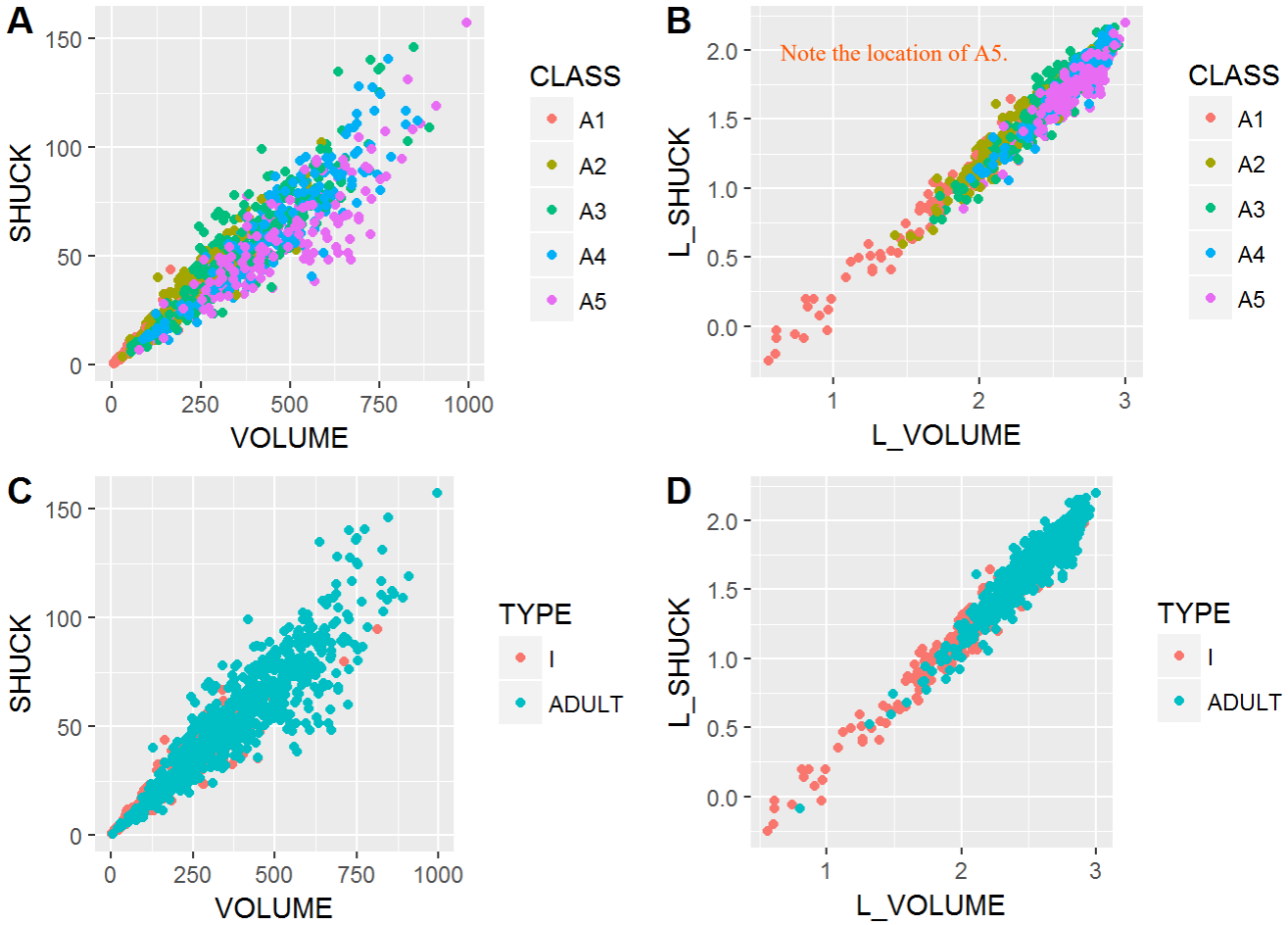


Question (2 points): Compare the histograms. How do the distributions differ? Are there going to be difficulties separating infants from adults based on VOLUME?

Answer: The distribution for infacnts seems to be right skewed while that for adult is more or less normally distributed. However, the most important thing is that majority of the infants have valume less than 400 with many having very low volumes. There are very few adults having as low volumes as infants and most of them seem to be having higher volumes. Thus, infants can be separated from adults based on VOLUME. Having said that, there are some overlaps of volume for infants and adults which may create difficulties.

(3)(b) (3 points) Create a scatterplot of SHUCK versus VOLUME and a scatterplot of their base ten logarithms, labeling the variables as L_SHUCK and L_VOLUME. Please be aware the variables, L_SHUCK and L_VOLUME, present the data as orders of magnitude (i.e. VOLUME = 100 = 10^2 becomes L_VOLUME = 2). Use color to differentiate CLASS in the plots. Repeat using color to differentiate only by TYPE.

```
mydata$L_SHUCK = log10(mydata$SHUCK)
mydata$L_VOLUME = log10(mydata$VOLUME)
theme_set(theme_gray())
g1=ggplot(mydata, aes(x = VOLUME, y = SHUCK, colour = CLASS)) + geom_point()
g2=ggplot(mydata, aes(x = L_VOLUME, y = L_SHUCK, colour = CLASS)) + geom_point()
g3=ggplot(mydata, aes(x = VOLUME, y = SHUCK, colour = TYPE)) + geom_point()
g4=ggplot(mydata, aes(x = L_VOLUME, y = L_SHUCK, colour = TYPE)) + geom_point()
plot_grid(g1, g2, g3, g4,labels = "AUTO")
```




```
#cowplot package used to create ggplots side by side. It is not absolutely required. We can ha
ve the plots one below the other too.
```

Question (3 points): Compare the two scatterplots. What effect(s) does log-transformation appear to have on the variability present in the plot? What are the implications for linear regression analysis? Additionally, where do the various CLASS levels appear in the plots? Where do the levels of TYPE appear in the plots?

stabilized the variability

Answer: (a)The left plots have higher variability (especially for larger volumes) than the plots on the right which are log-transformations. Log-transformation appears to have reduced the variability in the plot. The variability for larger volumes is much lesser when the log-transformation is applied. (b) This means that log-transformations would be better for linear regression analysis as the scatteplots show more of a "linear" relationships as compared to the case without log-transformations. A better linear regression line can be fit when log-transformations of variables are taken. (c) Higher CLASS levels, i.e. A3, A4, A5 are associated with larger volumes and mostly appear towards the top-right in the graphs.However for the same volume, a higher CLASS appears to be below the lower CLASS in the graph. Similarly, adults are associated with larger volumes and mostly appear towards the top-right in the graphs as compared to infants. We discussed a similar conclusion in DA-1. Abalone growth slows after A3 class.

---

(4)(a) (3 points) Since abalone growth slows after class A3, infants in classes A4 and A5 are considered mature and candidates for harvest. Reclassify the infants in classes A4 and A5 as ADULTS. This reclassification can be achieved using combineLevels(), but only on the abalones in classes A4 and A5. You will use this recoded TYPE variable, in which the infants in A4 and A5 were reclassified as ADULTS, for the remainder of this data analysis assignment.

Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and TYPE (see Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2 and Black Section 14.2). Use the multiple regression model: L_SHUCK ~ L_VOLUME + CLASS + TYPE. Apply summary() to the model object to produce results.

```
index <- (mydata$CLASS == "A5") | (mydata$CLASS == "A4")
mydata$TYPE[index] <- combineLevels(mydata$TYPE[index],
         levs = c("I", "ADULT"), "ADULT")
```

```
## The original levels I ADULT
## have been replaced by ADULT
```

```
# Or, alternatively:
# mydata$TYPE[with(mydata, CLASS=='A4' | CLASS=='A5')] <- 'ADULT'

model=lm(L_SHUCK~L_VOLUME+CLASS+TYPE, data=mydata)
summary(model)
```

```
##
## Call:
## lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.270634 -0.054287  0.000159  0.055986  0.309718
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.817512   0.019040 -42.936  < 2e-16 ***
## L_VOLUME     0.999303   0.010262  97.377  < 2e-16 ***
## CLASSA2     -0.018005   0.011005  -1.636 0.102124
## CLASSA3     -0.047310   0.012474  -3.793 0.000158 ***
## CLASSA4     -0.075782   0.014056  -5.391 8.67e-08 ***
## CLASSA5     -0.117119   0.014131  -8.288 3.56e-16 ***
## TYPEADULT    0.021093   0.007688   2.744 0.006180 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08297 on 1029 degrees of freedom
## Multiple R-squared:  0.9504, Adjusted R-squared:  0.9501
## F-statistic:  3287 on 6 and 1029 DF,  p-value: < 2.2e-16
```

Question (2 points): Interpret the trend in coefficient estimates for CLASS levels (Hint: this question is not asking if the estimates are statistically significant. It is asking for an interpretation of the pattern in these coefficients, and how this pattern relates to the earlier displays).

Answer: The coefficient estimates of CLASSA2, CLASSA3, CLASSA4 and CLASSA5 are all negative and they become more and more negative as one moves from A2 through A5. This indicates that ceterus paribus, i.e. keeping TYPE and L_VOLUME constant, L_SHUCK decreases as one moves moves lower class A1 to higher CLASS A5. This is as expected, because in the earlier display of L_SHUCK vs L_VOLUME by CLASS, for any gives L_VOLUME, lower CLASS was generally above the higher CLASS on the graph. Note that CLASSA2 is not statistically significant.

Question (2 points): Is TYPE an important predictor in this regression? (Hint: This question is not asking if TYPE is statistically significant, but rather how it compares to the other independent variables in terms of its contribution to predictions of L_SHUCK.) Explain your conclusion.

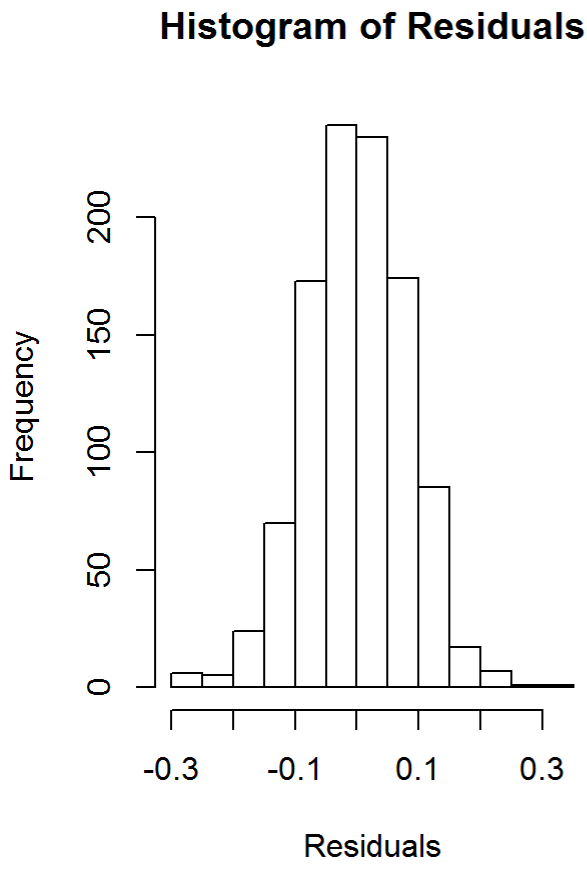
Answer: TYPE is statistically significant (p-value < 0.01). However, going by the magnitude of the coefficient, it seems like CLASS may be a more important contributor of L_SHUCK. Ceterus paribus, on an average, L_SHUCK of adult will be 0.021093 more than that of infant. On the other hand, ceterus paribus, moving from CLASS A1 to say CLASS A4, result in a decrease in L_SHUCK of 0.075782 which is of higher magnitude. So, TYPE is statistcally significant but may not be as important a contributor in the prediction of L_SHUCK as some other variables.

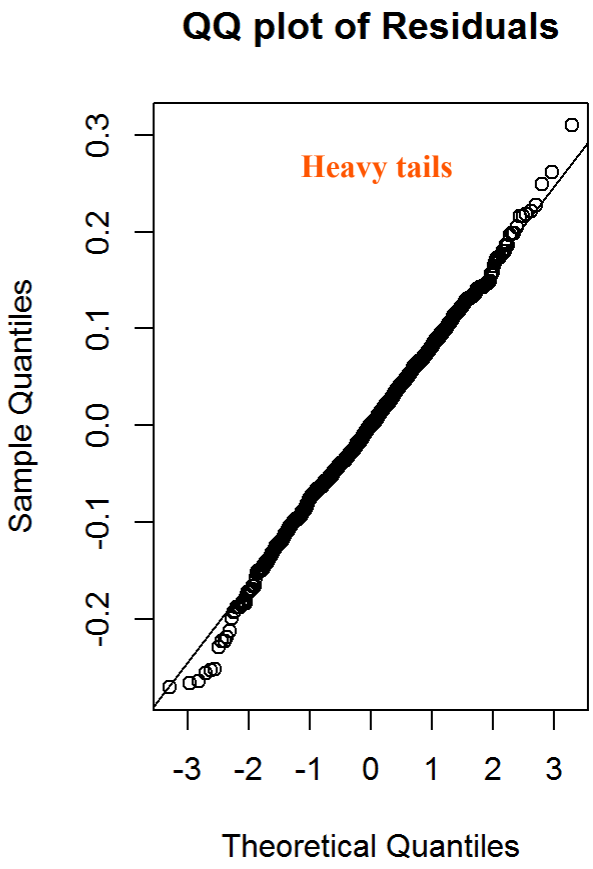
The next two analysis steps involve an analysis of the residuals resulting from the regression model in (4)(a) (see Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2).

---

(5)(a) (3 points) If “model” is the regression object, use model$residuals and construct a histogram and QQ plot. Compute the skewness and kurtosis. Be aware that with ‘rockchalk,’ the kurtosis value has 3.0 subtracted from it which differs from the ‘moments’ package.

```
par(mfrow=c(1,2))
hist(model$residuals, xlab="Residuals",ylab="Frequency", main="Histogram of Residuals")
qqnorm(model$residuals, main="QQ plot of Residuals")
qqline(model$residuals)
```






```
skewness(model$residuals, na.rm = TRUE, unbiased = TRUE)
```

```
## [1] -0.05945234
```

```
#Slight negative skewness (Close to no skewness)
kurtosis(model$residuals, na.rm = TRUE, unbiased = TRUE)
```

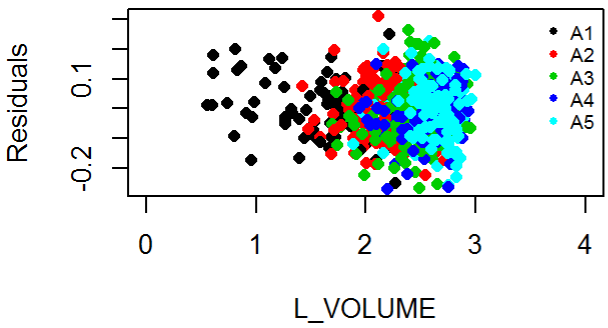
```
## [1] 0.3433082
```

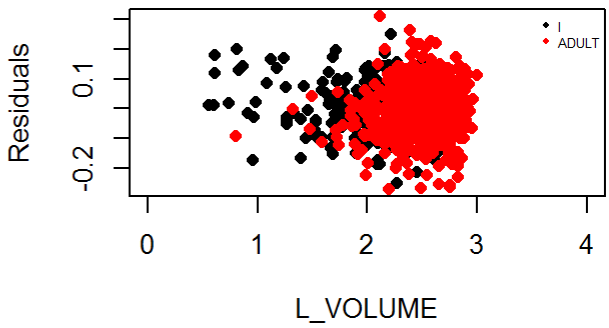
```
# Close to zero (mesokurtic). Approximately normal.
```

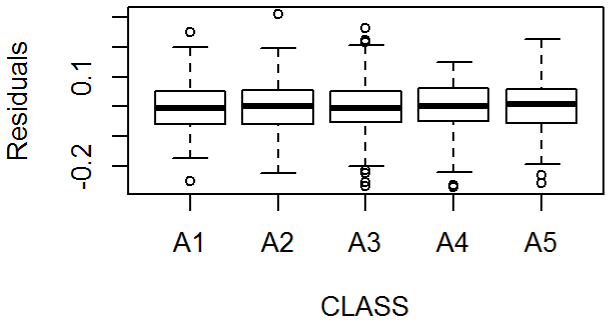
(5)(b) (3 points) Plot the residuals versus L_VOLUME coloring the data points by CLASS, and a second time coloring the data points by TYPE (Keep in mind the y-axis and x-axis may be disproportionate which will amplify the variability in the residuals). Present boxplots of the residuals differentiated by CLASS and TYPE (These four plots can be conveniently presented on one page using par(mfrow..) or grid.arrange(). Test the homogeneity of variance of the residuals across classes using the bartlett.test() (see Kabacoff Section 9.3.2, p. 222).

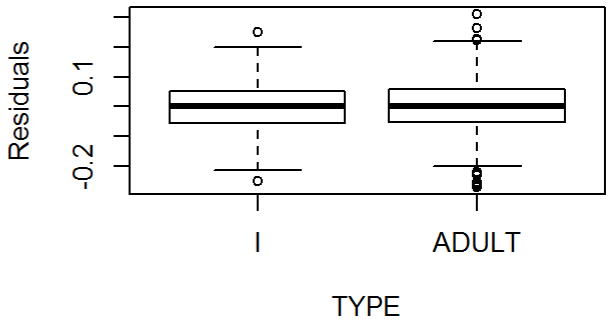
```
par(mfrow=c(2,2))
plot(mydata$L_VOLUME, model$residuals, pch=16,col=mydata$CLASS, ylab="Residuals", xlab="L_VOLU
ME", xlim=c(0,4))
legend('topright', legend = levels(mydata$CLASS), col = 1:5, cex = 0.7, pch = 16, bty="n")
plot(mydata$L_VOLUME, model$residuals, pch=16,col=mydata$TYPE, ylab="Residuals", xlab="L_VOLUM
```

```
E", xlim=c(0,4))
legend('topright', legend = levels(mydata$TYPE), col = 1:2, cex = 0.5, pch = 16, bty="n")
boxplot(model$residuals ~ mydata$CLASS, xlab="CLASS", ylab="Residuals")
boxplot(model$residuals ~ mydata$TYPE, xlab="TYPE", ylab="Residuals")
```










```
bartlett.test(model$residuals ~ mydata$CLASS)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  model$residuals by mydata$CLASS
## Bartlett's K-squared = 3.6882, df = 4, p-value = 0.4498
```

Question (3 points): What is revealed by the displays and calculations in (5)(a) and (5)(b)? Does the model 'fit'? Does this analysis indicate that L_VOLUME might be useful for harvesting decisions? Discuss.

Answer: 5(a) shows that the residuals are approximately normally distributed. This is because both skewness and kurtosis are close to 0 and even the histogram and QQ plots helps conclude that residuals are normally distributed. in 5(b), we try to see if residuals are correlated with our independent variables. The boxplots of residuals vs CLASS look similar for all the class and the residuals do not seem to be dependent on CLASS. Likewise for TYPE, residuals do not seem to be dependent on TYPE. The plot of Residuals versus L_VOLUME also do not form any pattern. This shows that residuals are not correlated with L_VOLUME too. Bartlett test shows homogenity of variances of the residuals (as p value > 0.05). Thus, the assumptions of linear regression model are satisfied and we can say that the model 'fits'. Yes, L_VOLUME might be useful for harvesting decisions

because the assumptions of the linear regression model are satisfied and L_VOLUME was statistically significant in the model.

There is a tradeoff faced in managing abalone harvest. The infant population must be protected since it represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. This assignment will use VOLUME to form binary decision rules to guide harvesting. If VOLUME is below a “cutoff” (i.e. specified volume), that individual will not be harvested. If above, it will be harvested. Different rules are possible.

The next steps in the assignment will require plotting of infants versus adults. For this plotting to be accomplished, similar “for loops” must be used to compute the harvest proportions. These loops must use the same value for the constants min.v and delta; and, use the same statement “for(k in 1:1000).” Otherwise, the resulting infant and adult proportions cannot be directly compared and plotted as requested. Note the example code supplied below.

---

(6)(a) (2 points) Calculate the proportion of infant and adult abalones which fall beneath a specified volume or “cutoff.” A series of volumes covering the range from minimum to maximum abalone volume will be used in a “for loop” to determine how the harvest proportions change as the “cutoff” changes. Example code for doing this is provided.

```
idxi <- mydata$TYPE == "I"
idxa <- mydata$TYPE == "ADULT"

max.v <- max(mydata$VOLUME)
min.v <- min(mydata$VOLUME)
delta <- (max.v - min.v)/10000
prop.infants <- numeric(10000)
prop.adults <- numeric(10000)
volume.value <- numeric(10000)

total.infants <- sum(idxi)
total.adults <- sum(idxa)

for (k in 1:10000) {
    value <- min.v + k*delta
    volume.value[k] <- value
    prop.infants[k] <- sum(mydata$VOLUME[idxi] <= value)/total.infants
    prop.adults[k] <-  sum(mydata$VOLUME[idxa] <= value)/total.adults
}

# prop.infants shows the impact of increasing the volume cutoff for
# harvesting. The following code shows how to "split" the population at
# a 50% harvest of infants.

n.infants <- sum(prop.infants <= 0.5)
split.infants <- min.v + (n.infants + 0.5)*delta  # This estimates the desired volume.
split.infants
```

```
## [1] 133.8199
```

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

```
split.adults
```

```
## [1] 384.5138
```

```
length(volume.value)
```

```
## [1] 10000
```

```
length(prop.infants)
```

```
## [1] 10000
```

```
length(prop.adults)
```

```
## [1] 10000
```

```
head(volume.value)
```

```
## [1] 3.710996 3.810202 3.909408 4.008615 4.107821 4.207027
```

```
head(prop.infants)
```

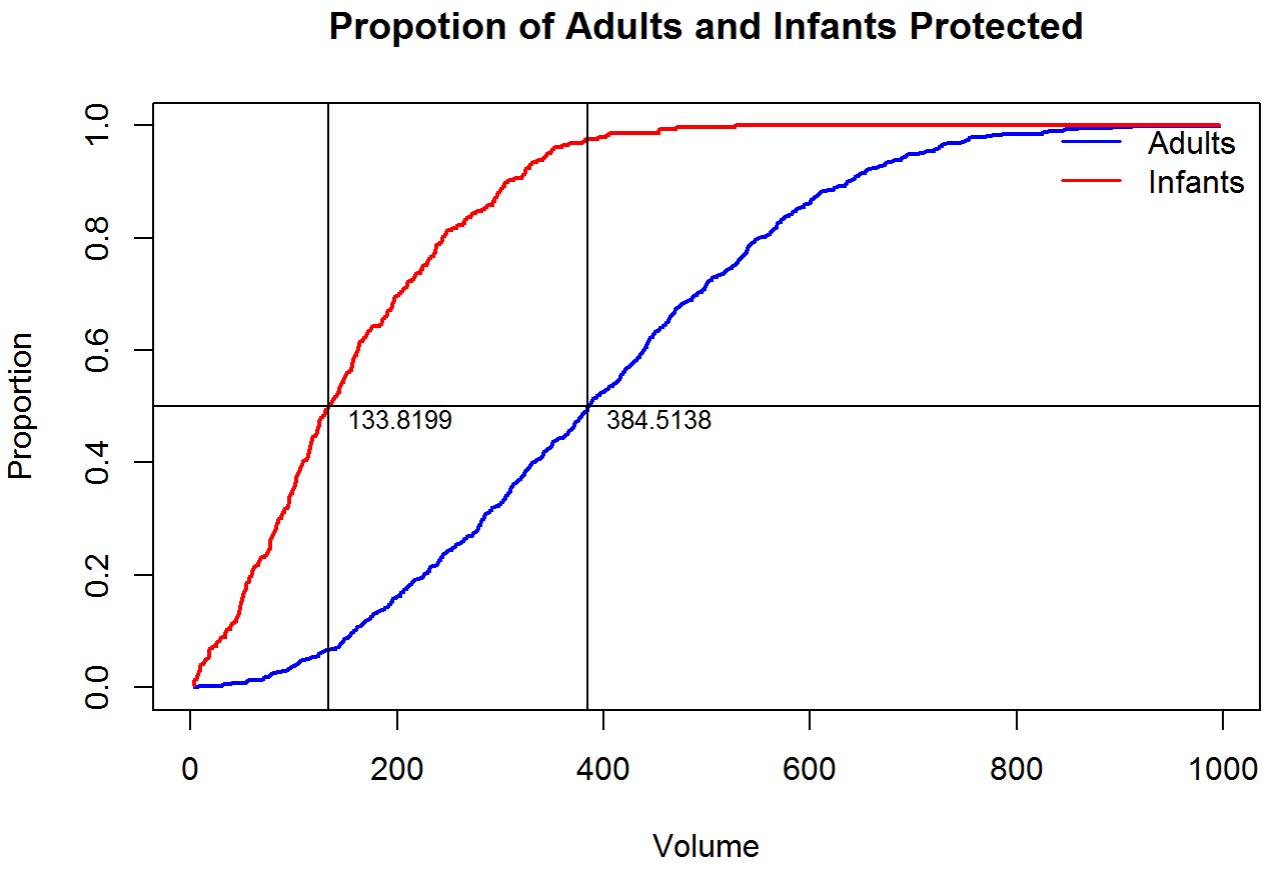
```
## [1] 0.003460208 0.003460208 0.003460208 0.006920415 0.013840830 0.013840830
```

```
head(prop.adults)
```

```
## [1] 0 0 0 0 0 0
```

(6)(b) (2 points) Present a plot showing the infant proportions and the adult proportions versus volume. Compute the 50% "split" volume.value for each and show on the plot.

```
par(mfrow=c(1,1))
plot(volume.value, prop.adults, col = "blue", lwd = 2, type = "l",main = "Propotion of Adults
and Infants Protected",
    xlab = "Volume", ylab = "Proportion")
lines(volume.value, prop.infants, col = "red", lwd = 2, type = "l")
abline(h = 0.5)
abline(v = c(split.adults, split.infants))
text(split.infants, 0.475, round(split.infants, 4), pos = 4, cex = 0.8)
text(split.adults, 0.475, round(split.adults, 4), pos = 4, cex = 0.8)
legend("topright", c("Adults", "Infants"), col = c("blue", "red"),
    lwd = 1.5, bty="n")
```



Question (2 points): The two 50% "split" values serve a descriptive purpose illustrating the difference between the populations. What do these values suggest regarding possible cutoffs for harvesting?

Answer: For the infant population, 50% infants have volume below 133.8199 and would not be harvested while the 50% cutoff for adults is 384.5138. This suggests that for adults, for 50% proportion to be not harvested, the volume cutoff is higher than that of infants. So, for volumes between 133.8199 and 384.5138, if it is an infant, they would be harvested, but if it is an adult, they will not e harvested. This shows the difference between the two populations.

-1 point Simply put, the two 50% split points indicate reasonable bounds for potential cutoffs.

This part will address the determination of a volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. To calculate this result, the proportions from (6) must be used. These proportions must be converted from "not harvested" to "harvested" proportions by using (1 - prop.infants) for infants, and (1 - prop.adults) for adults. The reason the proportion for infants drops sooner than adults is that infants are maturing and becoming adults with larger volumes.

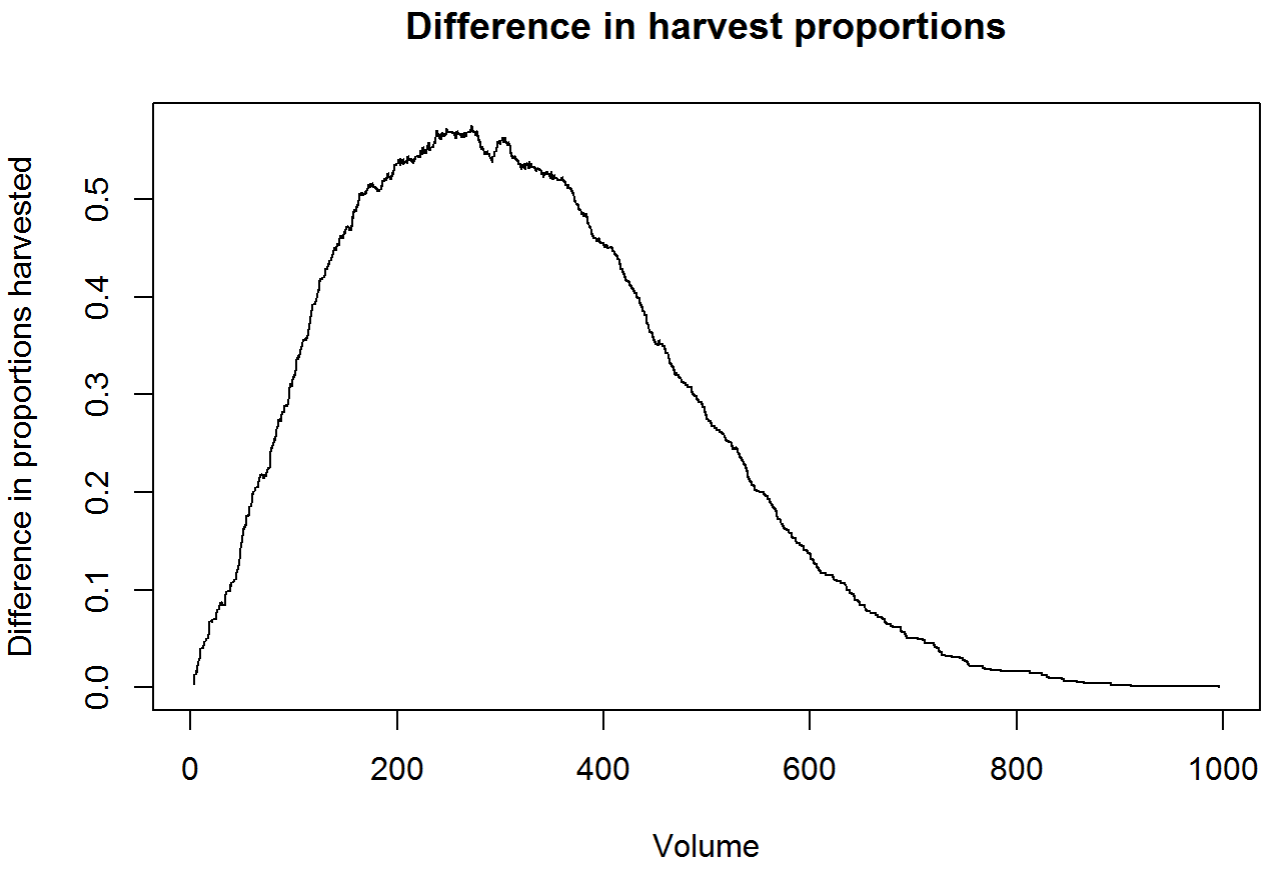
(7)(a) (1 point) Evaluate a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value. Compare to the 50% "split" points determined in (6)(a). There is considerable variability present in the peak area of this plot. The observed "peak" difference may not be the best representation of the data. One solution is to smooth the data to determine a more representative estimate of the maximum difference.

```
prop.infants.h = 1-prop.infants
prop.adults.h = 1-prop.adults
```

```
difference = prop.adults.h - prop.infants.h
head(difference)
```

```
## [1] 0.003460208 0.003460208 0.003460208 0.006920415 0.013840830 0.013840830
```

```
plot(volume.value, difference, main="Difference in harvest proportions", xlab="Volume", ylab="
Difference in proportions harvested", type="l")
```



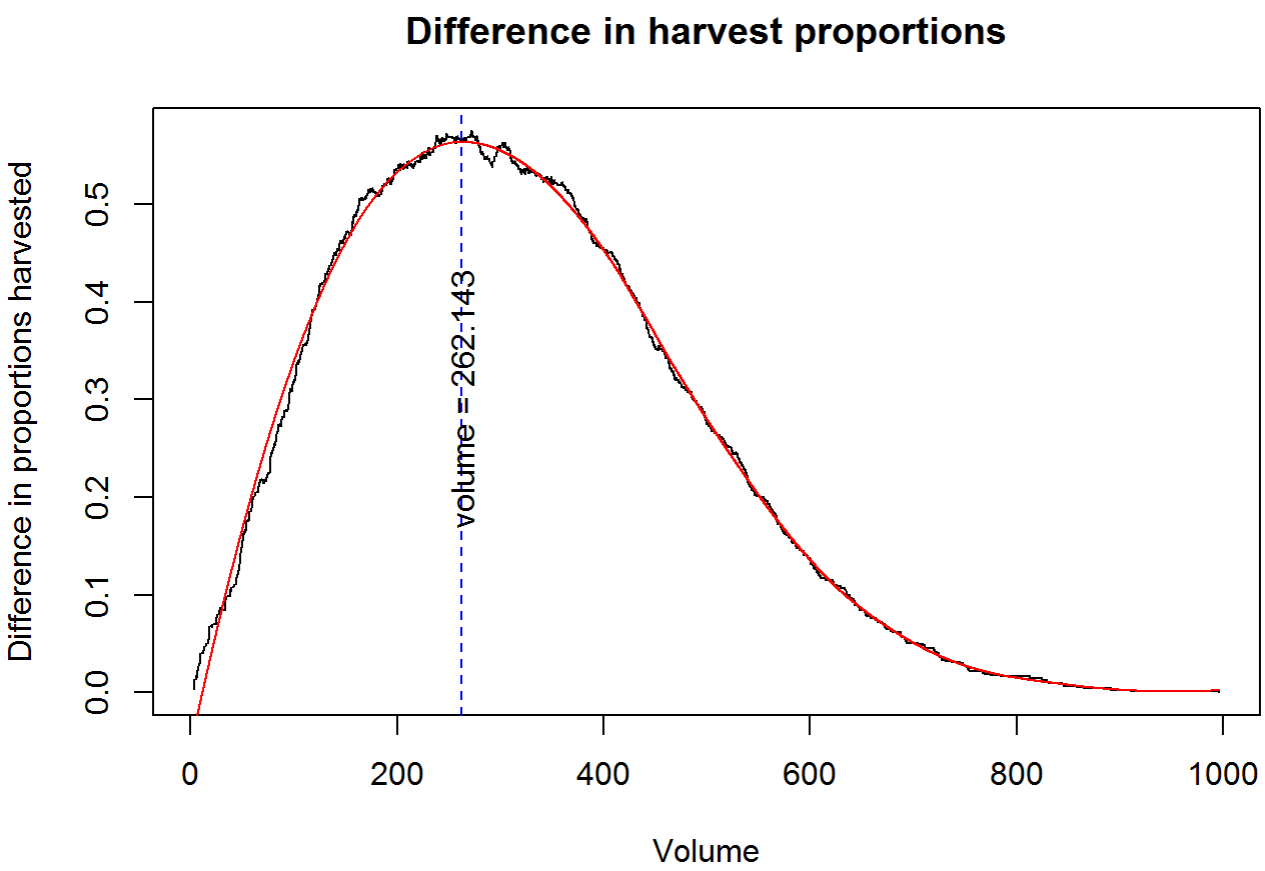
(7)(b) (1 point) Since curve smoothing is not studied in this course, code is supplied below. Execute the following code to determine a smoothed version of the plot in (a). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

```
y.loess.a <- loess(1 - prop.adults ~ volume.value, span = 0.25,
    family = c("symmetric"))
y.loess.i <- loess(1 - prop.infants ~ volume.value, span = 0.25,
    family = c("symmetric"))
smooth.difference <- predict(y.loess.a) - predict(y.loess.i)
```

(7)(c) (3 points) Present a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value with the variable smooth.difference superimposed. Determine the volume.value corresponding to the maximum of the variable smooth.difference (Hint: use which.max()).Show the estimated peak location corresponding to the cutoff determined.

```
plot(volume.value, difference, main="Difference in harvest proportions", xlab="Volume", ylab="
Difference in proportions harvested", type="l")
```

```
lines(volume.value, smooth.difference, col = "red", lwd = 1, type = "l")
abline(v = volume.value[which.max(smooth.difference)], lty = 2, col = "blue")
text(volume.value[which.max(smooth.difference)] + 0.0003, 0.3,
    paste("volume =",round(volume.value[which.max(smooth.difference)], 3)), srt = 90)
```



(7)(d) (1 point) What separate harvest proportions for infants and adults would result if this cutoff is used? (NOTE: the adult harvest proportion is the “true positive rate” and the infant harvest proportion is the “false positive rate.”)

Code for calculating the adult harvest proportion is provided.

```
(1 - prop.infants)[which.max(smooth.difference)] # [1] 0.1764706
```

```
## [1] 0.1764706
```

```
(1 - prop.adults)[which.max(smooth.difference)] #[1] 0.7416332
```

```
## [1] 0.7416332
```

There are alternative ways to determine cutoffs. Two such cutoffs are described below.

---

(8)(a) (2 points) Harvesting of infants in CLASS “A1” must be minimized. The smallest volume.value cutoff that produces a

zero harvest of infants from CLASS “A1” may be used as a baseline for comparison with larger cutoffs. Any smaller cutoff would result in harvesting infants from CLASS “A1.”

Compute this cutoff, and the proportions of infants and adults with VOLUME exceeding this cutoff. Code for determining this cutoff is provided.

```
volume.value[volume.value > max(mydata[mydata$CLASS == "A1" &
  mydata$TYPE == "I", "VOLUME"])][1] # [1] 206.786
```

```
## [1] 206.786
```

```
sum(mydata$VOLUME[idxi] > 206.786)/total.infants #[1] 0.2871972
```

```
## [1] 0.2871972
```

```
sum(mydata$VOLUME[idxa] > 206.786)/total.adults #[1] 0.8259705
```

```
## [1] 0.8259705
```

(8)(b) (2 points) Another cutoff can be determined for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: ‘missed’ adults and wrongly-harvested infants. This leaves for discussion which is a greater loss: a larger proportion of adults not harvested or infants harvested? This cutoff is 237.6391. Calculate the separate harvest proportions for infants and adults using this cutoff. Code for determining this cutoff is provided.

```
volume.value[which.min(abs(prop.adults - (1-prop.infants)))] # [1] 237.6391
```

```
## [1] 237.6391
```

```
sum(mydata$VOLUME[idxi] > 237.6391)/total.infants #[1] 0.2179931
```

```
## [1] 0.2179931
```

```
sum(mydata$VOLUME[idxa] > 237.6391)/total.adults #[1] 0.7817938
```

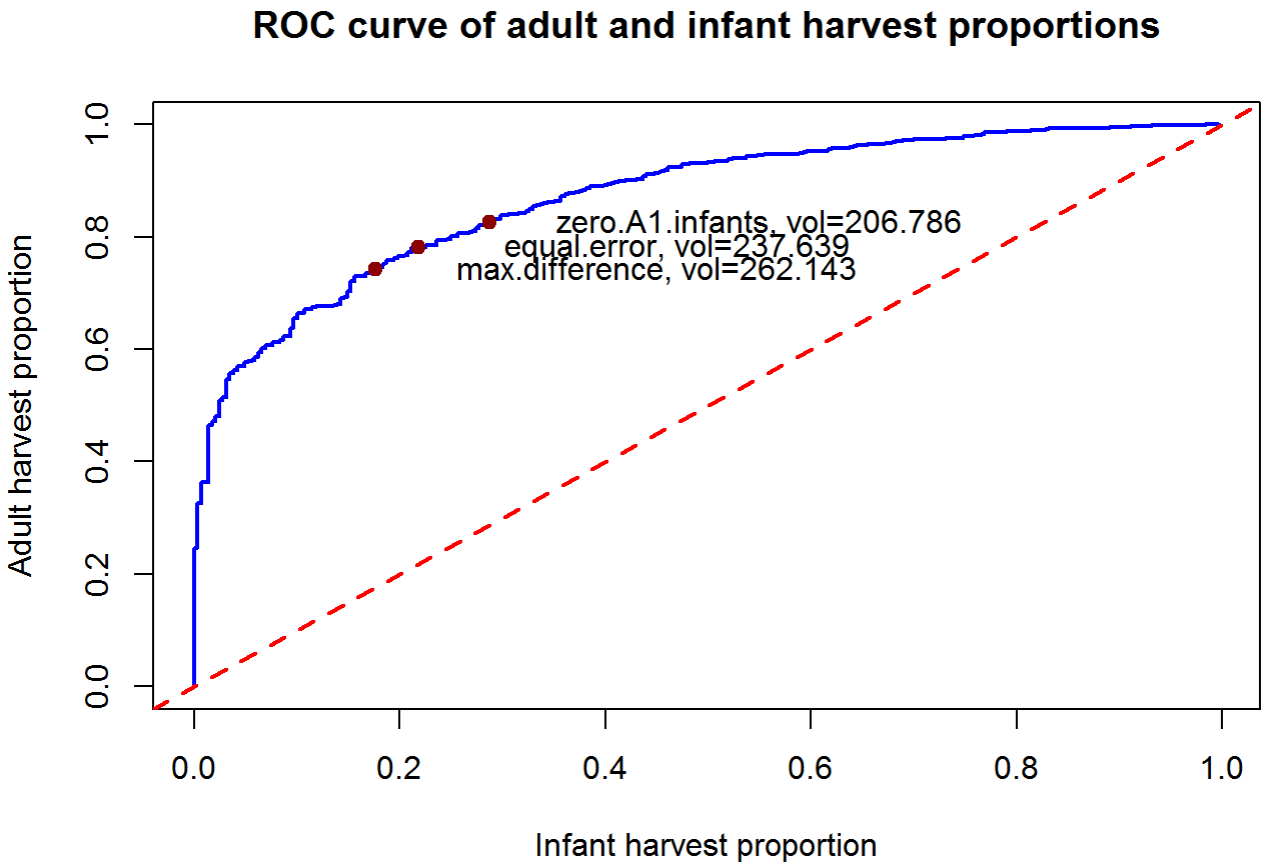
```
## [1] 0.7817938
```

---

(9)(a) (6 points) Construct an ROC curve by plotting (1 - prop.adults) versus (1 - prop.infants). Each point which appears corresponds to a particular volume.value. Show the location of the cutoffs determined in (7) and (8) on this plot and label each.

```
plot(1-prop.infants,1-prop.adults, col = "blue", lwd = 2, type = "l",
    main = "ROC curve of adult and infant harvest proportions", ylab = "Adult harvest proporti
on",
```

```
    xlab = "Infant harvest proportion")
abline(a=0, b=1, col = "red", lty = 2, lwd = 2)
points(0.176, 0.742, type="p", col="darkred",pch=19)
text(0.45, 0.742, "max.difference, vol=262.143")
points(0.287, 0.826, type="p", col="darkred",pch=19)
text(0.55, 0.826, "zero.A1.infants, vol=206.786")
points(0.218, 0.782, type="p", col="darkred",pch=19)
text(0.47, 0.782, "equal.error, vol=237.639")
```



(9)(b) (1 point) Numerically integrate the area under the ROC curve and report your result. This is most easily done with the auc() function from the “flux” package. Areas-under-curve, or AUCs, greater than 0.8 are taken to indicate good discrimination potential.

```
round(auc(1-prop.infants,1-prop.adults),3) #[1] 0.867
```

```
## [1] 0.867
```

---

(10)(a) (3 points) Prepare a table showing each cutoff along with the following: 1) true positive rate (1-prop.adults), 2) false positive rate (1-prop.infants), 3) harvest proportion of the total population

```
table=data.frame(c(262.143, 206.786,237.639), c(0.742, 0.826,0.782), c(0.176, 0.287, 0.218), r
ow.names = c("max.difference","zero.A1.infants","equal.error"))
colnames(table)=c("VOLUME","TPR","FPR")
```

```
table$PropYield = round((table$TPR * total.adults + table$FPR * total.infants)/(total.adults+t
otal.infants),3)
table
```

```
##                  VOLUME   TPR   FPR PropYield
## max.difference  262.143 0.742 0.176     0.584
## zero.A1.infants 206.786 0.826 0.287     0.676
## equal.error     237.639 0.782 0.218     0.625
```

Question: (1 point) Based on the ROC curve, it is evident a wide range of possible "cutoffs" exist. Compare and discuss the three cutoffs determined in this assignment. How might this display be used with the investigators?

Answer: There are 3 cutoffs. zero.A1.infants has the highest harvest proportion of the total population. max.difference has the least harvest proprotion of the total population. However, it has the least false positive rate, i.e. proportion of infants harvesting. equal.error seems to be a balance wherein the proportion of infants harvesting is somewhere in between as well as the harvest proportion of the total population is also somewhere in the middle. There is a tradeoff faced in managing abalone harvest. The infant population must be protected since it represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. So, equal.error offers a middle ground.

There are an infinite number of possible cutoffs. These three illustrate the tradeoffs involved in making a choice, and should be useful for stimulating discussion with the investigators.

Question (8 points): Assume you are expected to make a presentation of your analysis to the investigators How would you do so? Consider the following in your answer: 1) Would you make a specific recommendation or outline various choices and tradeoffs? 2) What qualifications or limitations would you present regarding your analysis? 3) If it is necessary to proceed based on the current analysis, what suggestions would you have for implementation of a cutoff? 4) What suggestions would you have for planning future abalone studies of this type?

Answer: 1) Yes, I would like to outline the tradeoff. There is a tradeoff faced in managing abalone harvest. The infant population must be protected since it represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. Hence, I won't recommend simply minimizing the Harvesting of infants in CLASS "A1". I would rather recommend cutoff for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: 'missed' adults and wrongly-harvested infants. 2) There is a tradeoff and there is nothing like a best cutoff. Even equal.error method leaves for discussion which is a greater loss: a larger proportion of adults not harvested or infants harvested. This is a limitation of the analysis. 3) Yes, I would say the best thing to do should be implemented. And given our choices,I would rather recommend cutoff for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: 'missed' adults and wrongly-harvested infants. 4) In future, a more detailed stufy should be done looking at not just volume, but further dimensions and also looking at other variables to improve the accuracy or the predictive power of the linear models.

-2 points I have a slightly different perspective.

I would keep it visual and simple at the start minimizing analytical details. The origin of the data needs to be explained since this determines limitations pertaining to the results. The results may not pertain to other locations. Get the investigators thinking about feasible choices and the tradeoffs involved. I would not present a recommendation unless it became necessary. If the investigators are uncertain, considering the concerns about over harvesting, my choice would be to do the least damage to the overall infant population. The maximum difference is a starting point. Different locations and more extensive data on the environment are needed. This analysis is just a first step.

72 points