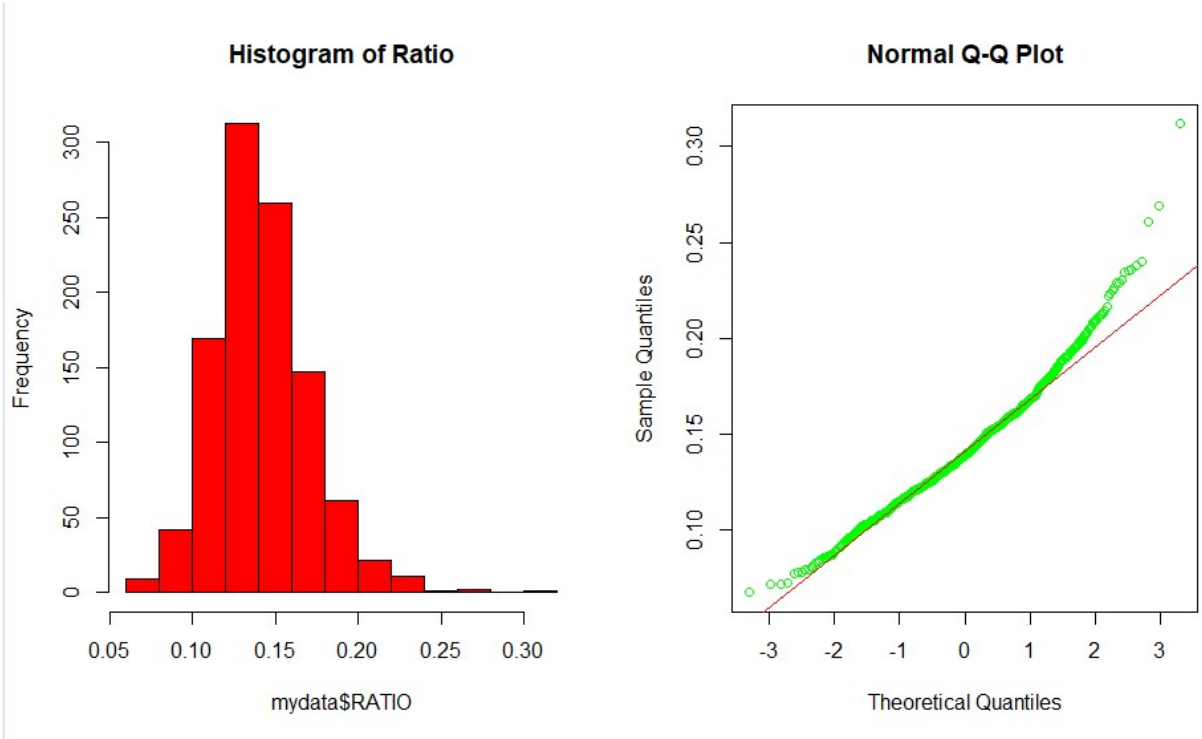
John Barker, Data Analysis with R Assignment 2

#(1)(a) 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$RATIO, col = 'red', main = 'Histogram of Ratio')
> qqnorm( mydata$RATIO, col = 'green', main = 'Normal Q-Q Plot' )
> qqline(mydata$RATIO,col='red')
> par(mfrow = c(1,1))
>
> kurtosis(mydata$RATIO)
[1] 4.676321
> skewness(mydata$RATIO)
[1] 0.7157417
```

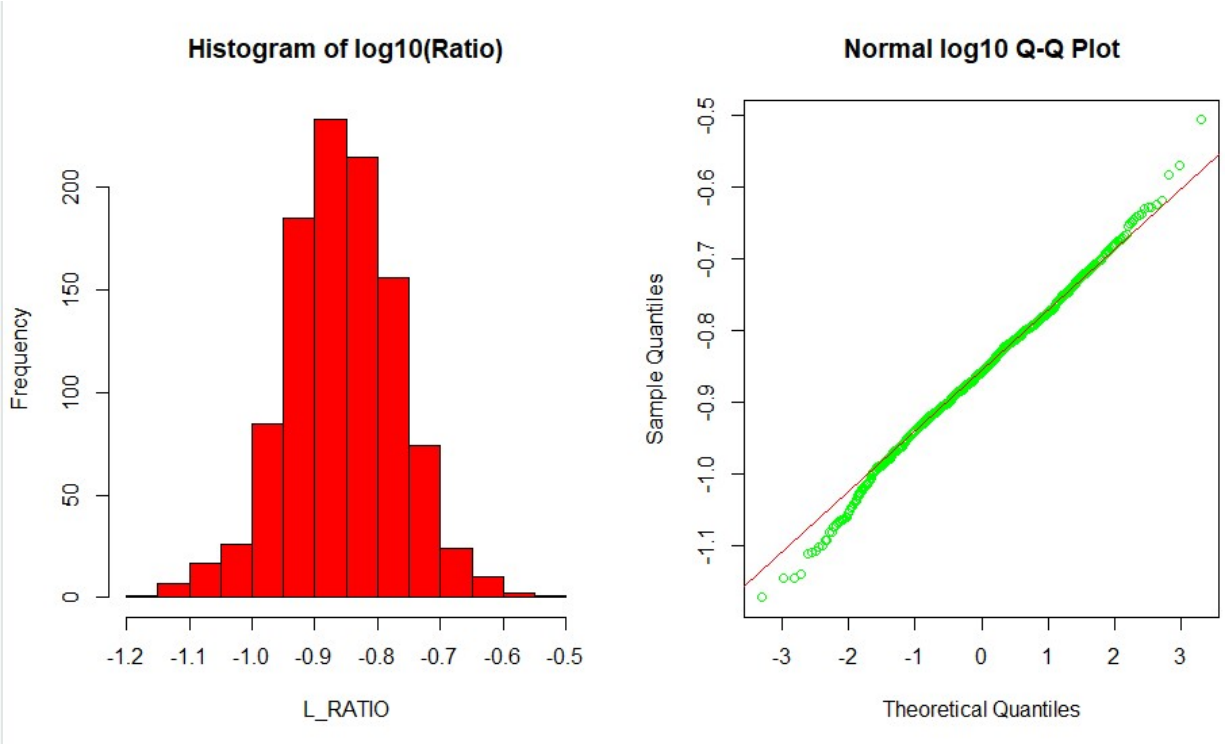


#(1)(b) Tranform RATIO using *log10()* to create L_RATIO (Kabacoff Section 8.5.2, p. 199-200).

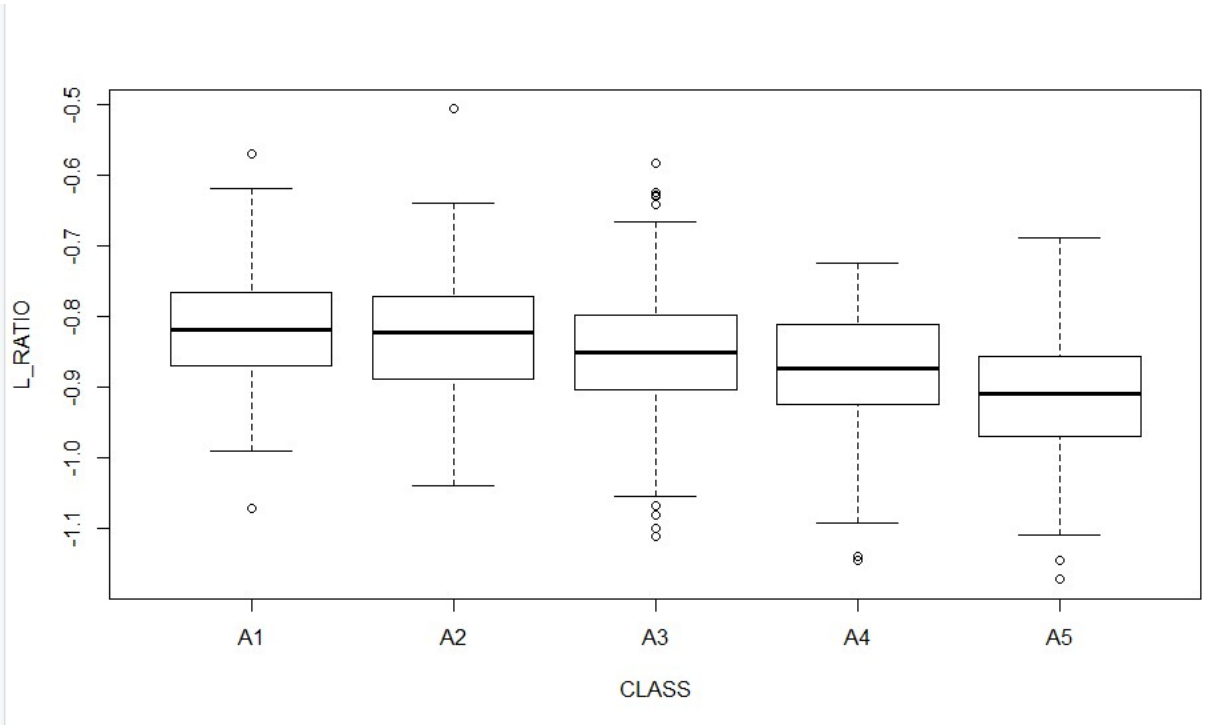
Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis.

Create a boxplot of L_RATIO differentiated by CLASS.

```
> L_RATIO<-log10(mydata$RATIO)
> par(mfrow = c(1,2))
> hist(L_RATIO, col = 'red', main = 'Histogram of log10(Ratio)')
> qqnorm( L_RATIO, col = 'green', main = 'Normal log10 Q-Q Plot' )
> qqline(L_RATIO,col='red')
> par(mfrow = c(1,1))
>
> kurtosis(L_RATIO)
[1] 3.542266
> skewness(L_RATIO)
[1] -0.09405162
```




```
> boxplot(L_RATIO~CLASS, data=mydata, ylab="L_RATIO", xlab = "CLASS")
```



#(1)(c) Test the homogeneity of variance across classes using *bartlett.test()*

 # (Kabacoff Section 9.2.2, p. 222).

```
> bartlett.test(L_RATIO~mydata$CLASS)

        Bartlett test of homogeneity of variances

data:  L_RATIO by mydata$CLASS
Bartlett's K-squared = 3.1891, df = 4, p-value = 0.5267
```

#**Essay Question: 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: (Enter your answer here.)***

The L_RATIO value exhibits better conformance to a normal distribution when comparing the histogram of the RATIO and the histogram of the L_RATIO. The histogram of the RATIO is skewed to the right where the L_RATIO is very close to the perfect symmetry of a normal curve. Looking at the QQ plots RATIO strays from the line just past one and there is more separation from the line than in the QQ plot for L_RATIO. The box plot for L_RATIO displays very little variance with a possible change in class A5. From the output of Bartlett test on L_RATIO we can see the p-value of 0 5267 is not less than the significance value of 0.05 the variance of L_RATIO is not significantly different in the classes.

#(2)(a) Perform an analysis of variance with *aov()* on L_RATIO using CLASS and SEX

as the independent variables (Kabacoff chapter 9, p. 212-229). Assume equal variances.

Perform 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).

```
> anovacs<-aov(L_RATIO~CLASS*SEX, data = mydata)
> summary(anovacs)
             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
> anovancs<-aov(L_RATIO~CLASS+SEX, data = mydata)
> summary(anovancs)
             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

>
```

#**Essay Question: 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: (Enter your answer here.)***

The non-significant interaction term between CLASS and SEX has no relationship with L_RATIO. This is in evidence by the small change in the residuals when CLASS:SEX is removed from the aov(). Of the two factors CLASS has a greater affect on L_RATIO. For every unit increase of CLASS there is an increase of 1.055 and the Pr value has greater significance. There is a significant affect of SEX on L_RATIO it is not has pronounced as CLASS. It is most likely due to infants.

#(2)(b) 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(anovancs)
  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


>
```

#**Additional Essay Question: 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?**

Across CLASSes there is only one set of groups which does not have a significant difference and that is between A1 and A2. There is a 0 in the confidence interval of A1 and A2. All other CLASS group comparisons show a significant group difference. The output gives a significant difference between Infants and males or females. The difference between males and females is insignificant so males and females can be grouped as one category called adults. In the aov test the difference between infants and adults represented all of the affects in the L_RATIO SEX comparison.

#(3)(a1) We combine "M" and "F" into a new level, "ADULT". (While this could be

accomplished using *combineLevels()* from the 'rockchalk' package, we use base R

code because many students do not have access to the rockchalk package.) This

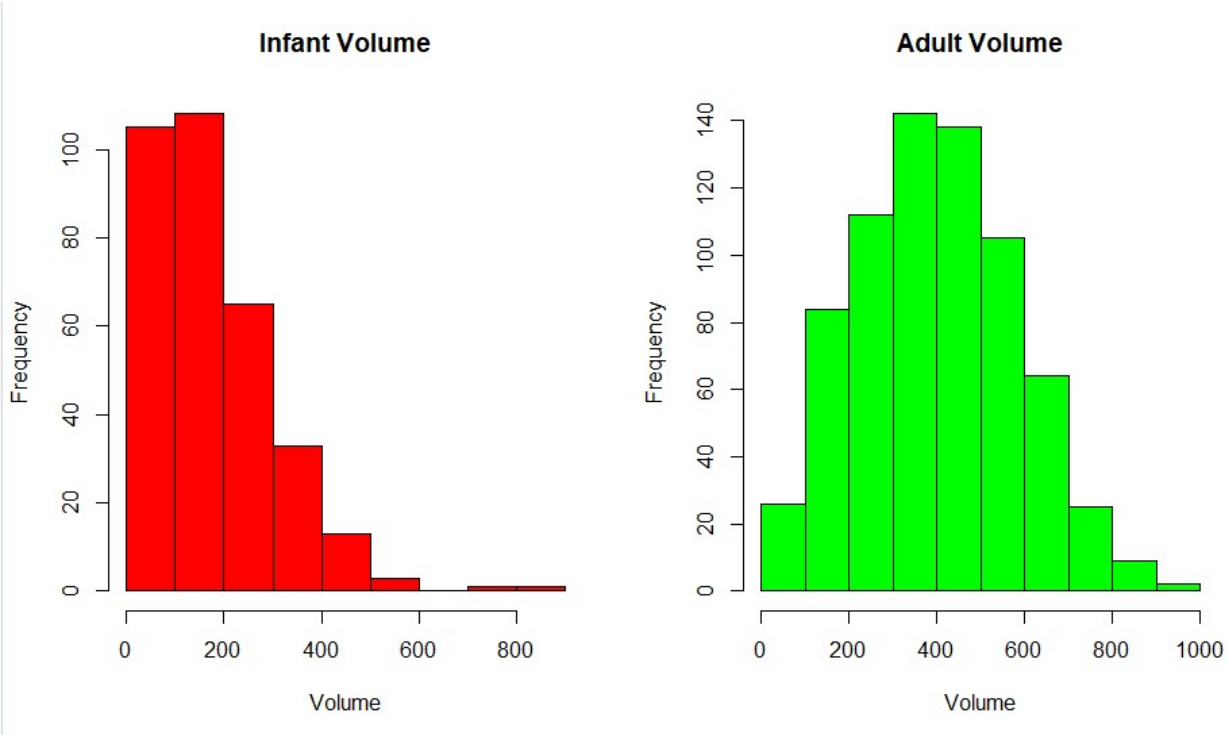
necessitated defining a new variable, TYPE, in mydata which had two levels: "I" and "ADULT".

```
> mydata$TYPE <- character(nrow(mydata))  # initialize the TYPE column as all blanks
> for (i in seq(along = mydata$SEX)) {
+   mydata$TYPE[i] <- 'I'
+   if (mydata$SEX[i] == 'M' || mydata$SEX[i] == 'F') mydata$TYPE[i] <- 'ADULT'
+ }
> mydata$TYPE <- factor(mydata$TYPE)
> cat('\nCheck on definition of TYPE object (should be an integer): ', typeof(mydata$TYPE
Check on definition of TYPE object (should be an integer):  integer> cat('\nmydata$TYPE i
is.factor(mydata$TYPE), '\n')

mydata$TYPE is treated as a factor:  TRUE
> table(mydata$SEX, mydata$TYPE)

    ADULT   I
  F   326   0
  I     0 329
  M   381   0

>
```

#(3)(a2) Present side-by-side histograms of VOLUME. One should display infant

volumes and, the other, adult volumes.

```
par(mfrow = c(1,2))
> hist(mydata$VOLUME[mydata$TYPE == "I"], main = "Infant Volume", col = "red"
, xlab = "Volume")
> hist(mydata$VOLUME[mydata$TYPE == "ADULT"], main = "Adult Volume", col = "g
reen", xlab = "Volume")
> par(mfrow = c(1,1))
```



#**Essay Question: Compare the histograms. How do the distributions differ?

Are there going to be any difficulties separating infants from adults based on VOLUME?**

The adult volume is similar to the normal distribution with the mean somewhere between 400 and 500. While infant volume is heavily skewed right. To separate infants from adults you would have to separate the smaller adults from the rest of the adults to ensure most infants are separated. Somewhere between 200 and 400 would be a good cutoff point.

#(3)(b) Create a scatterplot of SHUCK versus VOLUME and a scatterplot of their

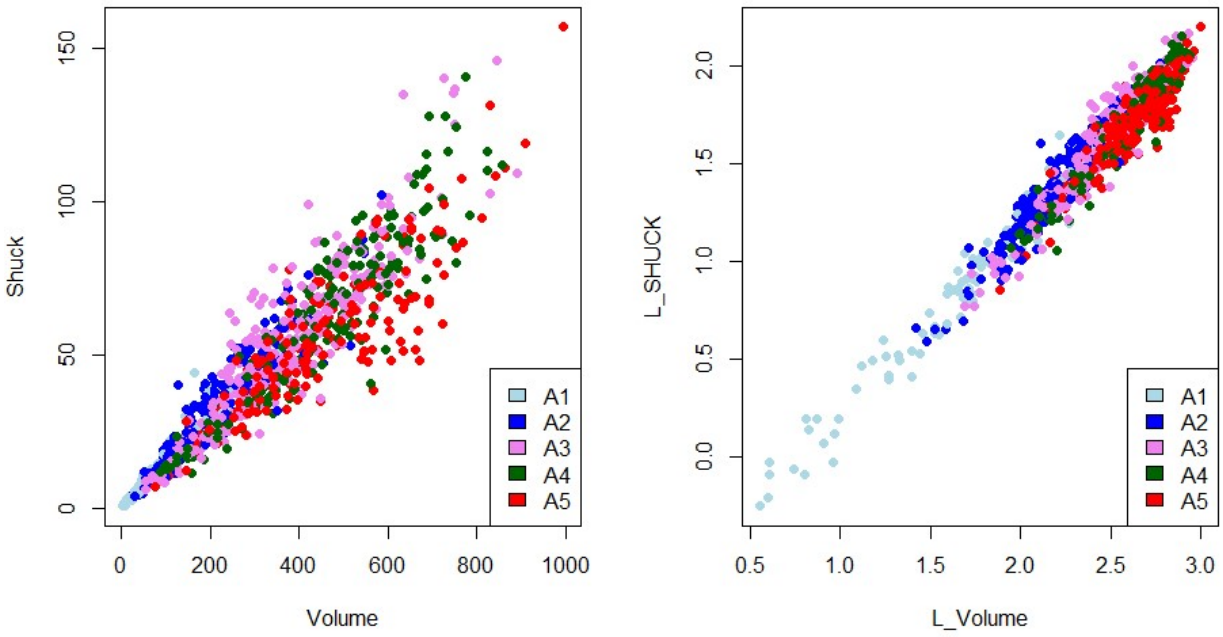
base ten logarithms, labeling the variables as L_SHUCK and L_VOLUME. Please be

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

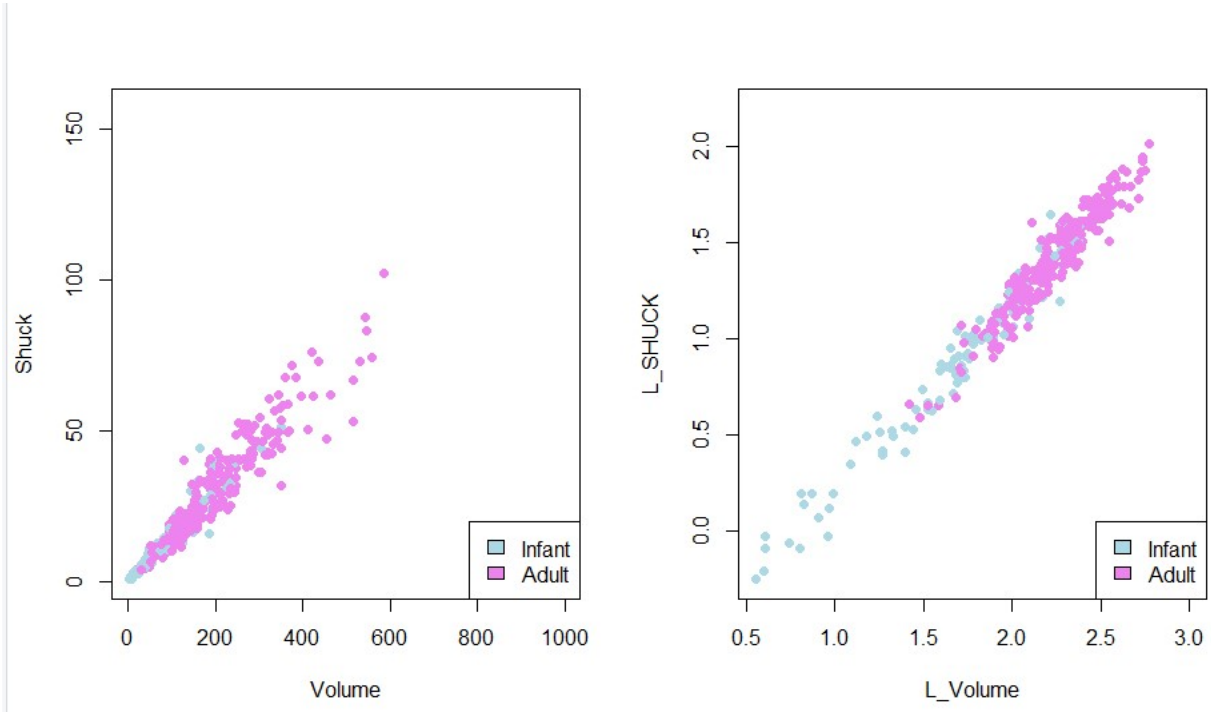
in the plots. Repeat using color to differentiate by TYPE.

```
> L_SHUCK<-log10(mydata$SHUCK)
> L_VOLUME<-log10(mydata$VOLUME)
>
> color = rep(NA, length=length(mydata$CLASS))
> color[which(mydata$CLASS == "A1")] = "lightblue"
> color[which(mydata$CLASS == "A2")] = "blue"
> color[which(mydata$CLASS == "A3")] = "violet"
> color[which(mydata$CLASS == "A4")] = "darkgreen"
> color[which(mydata$CLASS == "A5")] = "red"
>
> par(mfrow = c(1,2))
> plot(mydata$VOLUME,mydata$SHUCK, col = color, pch=16, xlab = "Volume",
+      ylab = "Shuck")
> legend("bottomright", legend=c("A1","A2","A3","A4","A5"),
+        fill=c("lightblue","blue","violet","darkgreen","red"))
> plot(L_VOLUME,L_SHUCK, col = color, pch=16, xlab = "L_Volume",
+      ylab = "L_SHUCK")
> legend("bottomright", legend=c("A1","A2","A3","A4","A5"),
+        fill=c("lightblue","blue","violet","darkgreen","red"))
> par(mfrow = c(1,1))
```




```
> tcolor = rep(NA, length=length(mydata$TYPE))
> tcolor[which(mydata$CLASS == "A1")] = "lightblue"
> tcolor[which(mydata$CLASS == "A2")] = "violet"
>
> par(mfrow = c(1,2))
> plot(mydata$VOLUME,mydata$SHUCK, col = tcolor, pch=16, xlab = "Volume",
+      ylab = "Shuck")
> legend("bottomright", legend=c("Infant","Adult"),
+        fill=c("lightblue","violet"))
```

```
> plot(L_VOLUME,L_SHUCK, col = tcolor, pch=16, xlab = "L_Volume",
+      ylab = "L_SHUCK")
> legend("bottomright", legend=c("Infant","Adult"),
+        fill=c("lightblue","violet"))
> par(mfrow = c(1,1))

>
```



#**Additional Essay Question: 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? Where do the various CLASS

levels appear in the plots? Where do the levels of TYPE appear in the plots?**

#***Answer: (Enter your answer here.)***

As volume increases the variance is greater by CLASS and SEX. The log transformation reduces the increase in variance. As volume gets larger SHUCK gets larger. Infants have a tendency to be smaller and adults have a tendency to be larger. The volume and shuck weight go up as the abalone gets older. There is some overlap at the oldest end with class A4 and A5.

```
#(4)(a1) 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 could have been achieved using

# *combineLevels()*, but only on the abalones in classes A4 and A5. We will do

# this recoding of the TYPE variable using base R functions. We will use this

# recoded TYPE variable, in which the infants in A4 and A5 are reclassified as ADULTS,

# for the remainder of this data analysis assignment.
```

```
> for (i in seq(along = mydata$TYPE)) {
+   if (mydata$CLASS[i] == 'A4' || mydata$CLASS[i] == 'A5') mydata$TYPE[i] <- 'ADULT'
+ }
> mydata$TYPE <- factor(mydata$TYPE)
> cat('\nCheck on redefinition of TYPE object (should be an integer): ', typeof(mydata$TY

Check on redefinition of TYPE object (should be an integer):  integer> cat('\nmydata$TYPE
, is.factor(mydata$TYPE), '\n')

mydata$TYPE is treated as a factor:  TRUE
> cat('\nThree-way contingency table for SEX, CLASS, and TYPE:\n')

Three-way contingency table for SEX, CLASS, and TYPE:
> print(table(mydata$SEX, mydata$CLASS, mydata$TYPE))
, ,  = ADULT


    A1  A2  A3  A4  A5
F    5  41 121  82  77
I    0   0   0  21  19
M   12  62 143  85  79

, ,  = I


    A1  A2  A3  A4  A5
F    0   0   0   0   0
I   91 133  65   0   0
M    0   0   0   0   0


>
```

#(4)(a2) Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and TYPE

(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.

```
> 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) -0.796418   0.021718 -36.672  < 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 ***
TYPEI       -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


>
```

#**Essay Question: Interpret the trend in CLASS levelcoefficient estimates?

(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: (Enter your answer here.)

CLASS A2 is not a predictor of SHUCK but CLASS A3, A4 and A5 are all good predictors of Shuck. With all other value remaining constant with each unit increase of A3 there is a .047310 increase in shuck, for each unit increase in A4 there is a .075782 increase in Shuck and for each unit increase in A5 there is a .117119 increase in Shuck. The pattern is that the older the abalone the greater the affect on Shuck. Which is evidenced by the charts in 3b.

**Additional Essay Question: Is TYPE an important predictor in this regression?

(Hint: This question is not asking if TYPE is statistically significant, but rather

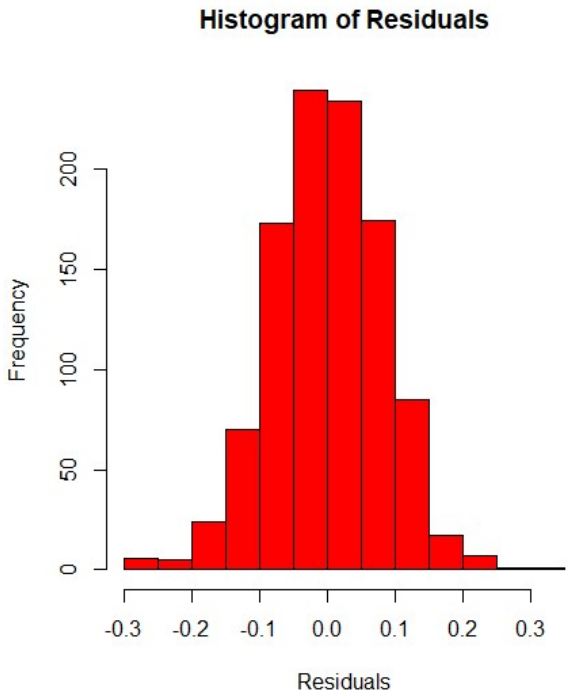
how it compares to the other independent variables in terms of its contribution to

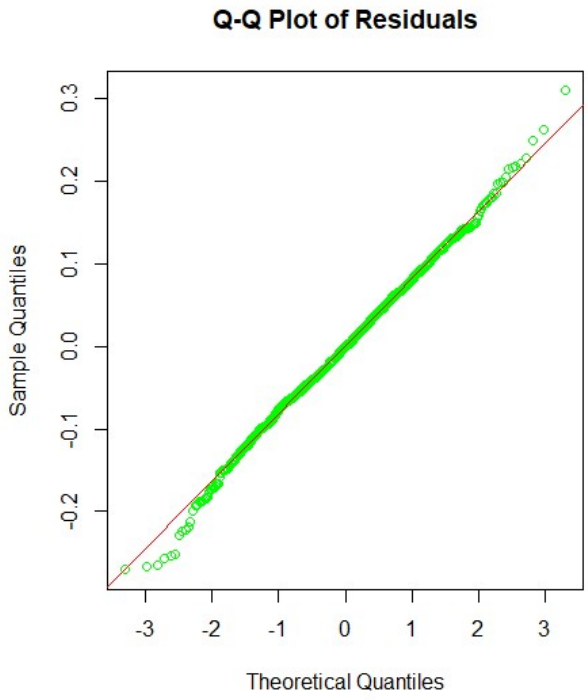
predictions of L_SHUCK for harvesting decisions.) Explain your conclusion.**

Answer: (Enter your answer here.)

Type is a predictor in this regression. As the abalone go from infant to adult their Shuck weight increases. Most of the infants are at the low end of the shuck scale and the adults are in the middle and upper end of the shuck scale. For every unit increase of Type there is a .021093 increase in Shuck.

```
#(5)(a) If "model" is the regression object, use model$residuals and construc
t
# a histogram and QQ plot. Compute the skewness and kurtosis. Be aware that w
ith
# 'rockchalk,' the kurtosis value has 3.0 subtracted from it which differs fr
om the 'moments' package.
#```{r Part_5a}
 > hist(model$residuals)
 > par(mfrow = c(1,2))
 > hist(model$residuals, col = 'red', main = 'Histogram of Residuals', xlab = "Residuals")
 > qqnorm( model$residuals, col = 'green', main = 'Q-Q Plot of Residuals' )
 > qqline(model$residuals,col='red')
 > par(mfrow = c(1,1))
 >
 > kurtosis(model$residuals)
 [1] 3.349772
 > skewness(model$residuals)
 [1] -0.05953853

 >
```





#(5)(b) 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

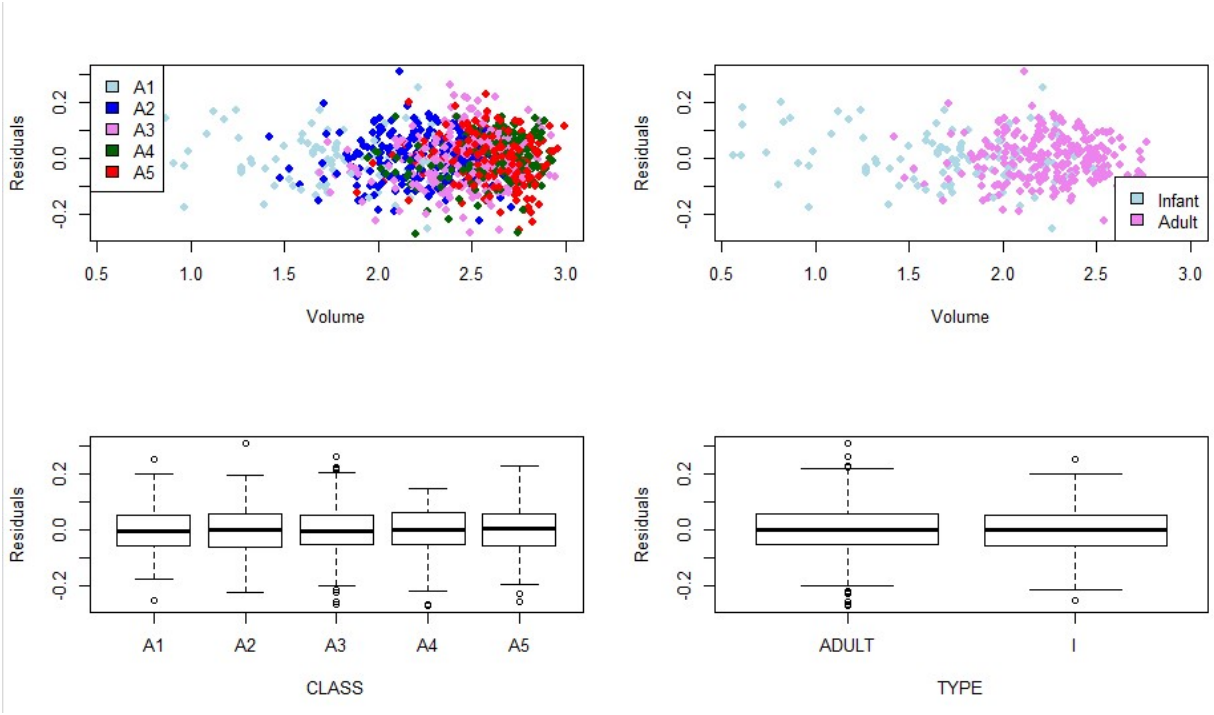
presented on one page using *par(mfrow..)* or *grid.arrange()*. Test the homogeneity of

variance of the residuals across classes using *bartlett.test()* (Kabacoff Section 9.3.2, p. 222).

#```{r Part_5b}

```
> par(mfrow = c(2,2))
> color = rep(NA, length=length(mydata$CLASS))
> color[which(mydata$CLASS == "A1")] = "lightblue"
> color[which(mydata$CLASS == "A2")] = "blue"
> color[which(mydata$CLASS == "A3")] = "violet"
> color[which(mydata$CLASS == "A4")] = "darkgreen"
> color[which(mydata$CLASS == "A5")] = "red"
>
> plot(L_VOLUME, model$residuals, col = color, pch=16, xlab = "Volume",
+      ylab = "Residuals")
> legend("topleft", legend=c("A1","A2","A3","A4","A5"),
+        fill=c("lightblue","blue","violet","darkgreen","red"))
>
> tcolor = rep(NA, length=length(mydata$TYPE))
> tcolor[which(mydata$CLASS == "A1")] = "lightblue"
> tcolor[which(mydata$CLASS == "A2")] = "violet"
>
```

```
> plot(L_VOLUME,model$residuals, col = tcolor, pch=16, xlab = "Volume",
+      ylab = "Residuals")
> legend("bottomright", legend=c("Infant","Adult"),
+        fill=c("lightblue","violet"))
> boxplot(model$residuals~CLASS, data=mydata, ylab="Residuals", xlab = "CLASS")
> boxplot(model$residuals~TYPE, data=mydata, ylab="Residuals", xlab = "TYPE")
> par(mfrow = c(1,1))
>
>
> bartlett.test(model$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


>
```



#**Essay Question: 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, and ultimately VOLUME,

might be useful for harvesting decisions? Discuss.**

#***Answer: (Enter your answer here.)***

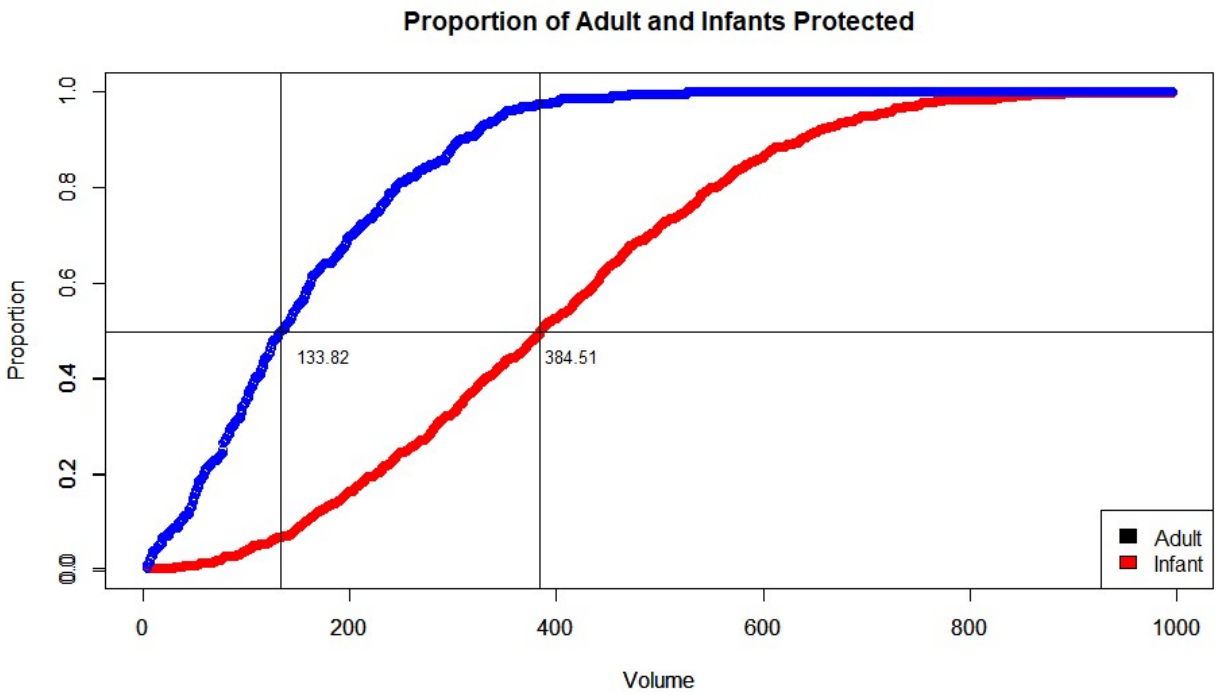
From the charts of residuals versus volume by class and type we see there is a distinct demarcation of infants and adults by volume. The Bartlett test provides results that shows very little variance between classes. The lack of variance is displayed in the boxplots.

#(6)(b) Present a plot showing the infant proportions and the adult proportions versus

volume.value. Compute the 50% "split" volume.value for each and show on the plot.

#```{r Part_6b}

```
> plot(volume.value, prop.adults, col="red", main = "Proportion of Adult and Infants Pro
+      xlab = "Volume", ylab = "Proportion")
> par(new=TRUE)
> plot(volume.value, prop.infants, col="blue",
+      xlab = "Volume", ylab = "Proportion")
> abline(h=0.5)
> abline(v=133.82)
> text(175, 0.45, labels = c("133.82"), cex=0.8)
> abline(v=384.51)
> text(415, 0.45, labels = c("384.51"), cex=0.8)
> legend("bottomright", legend=c("Adult","Infant"),
+        col=c("lightblue","violet"), lty, seq_len(5))

>
```



#**Essay Question: 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: (Enter your answer here.)***

The two 50% splits represent the points at which half the adults are harvested and half the infants are harvested. The adults reach 50% at a volume of 133.82 and the infants reach 50% at a volume of 384.51. Due to the steep initial curve for adult's volume by the time 50% split is reached less than 10% of the infants are harvested. And at 50% harvest for the infants about 95% of the adults are harvested. This is good if the goal of the harvest is to catch adults only.

#(7)(c) Present a plot of the difference

((1 - prop.adults) - (1 - prop.infants)) versus volume.value

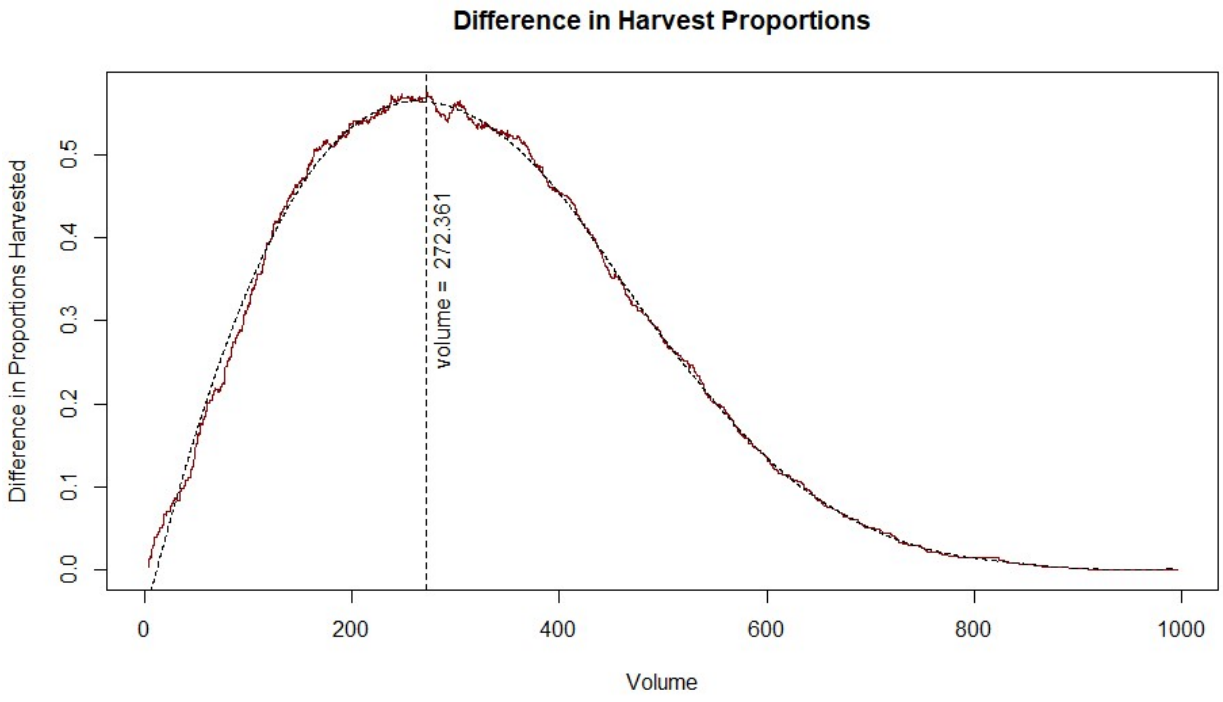
with the variable smooth.difference superimposed. Determine the volume.value

corresponding to the maximum smoothed difference (Hint: use *which.max()*).

Show the estimated peak location corresponding to the cutoff determined.

#```{r Part_7c}

```
> plot(volume.value, difference, type = "l", col="red4", xlab = "Volume",
+      ylab = "Difference in Proportions Harvested",
+      main = "Difference in Harvest Proportions")
> lines(volume.value, smooth.difference, col = "gray2", lwd = 1.5, lty = 2)
> abline(v = volume.value[which.max(difference)], lwd = 1.5, lty = 2)
> text(285, 0.35, labels = paste("volume = ", round(volume.value[which.max(difference)],
x = 1)

>
```



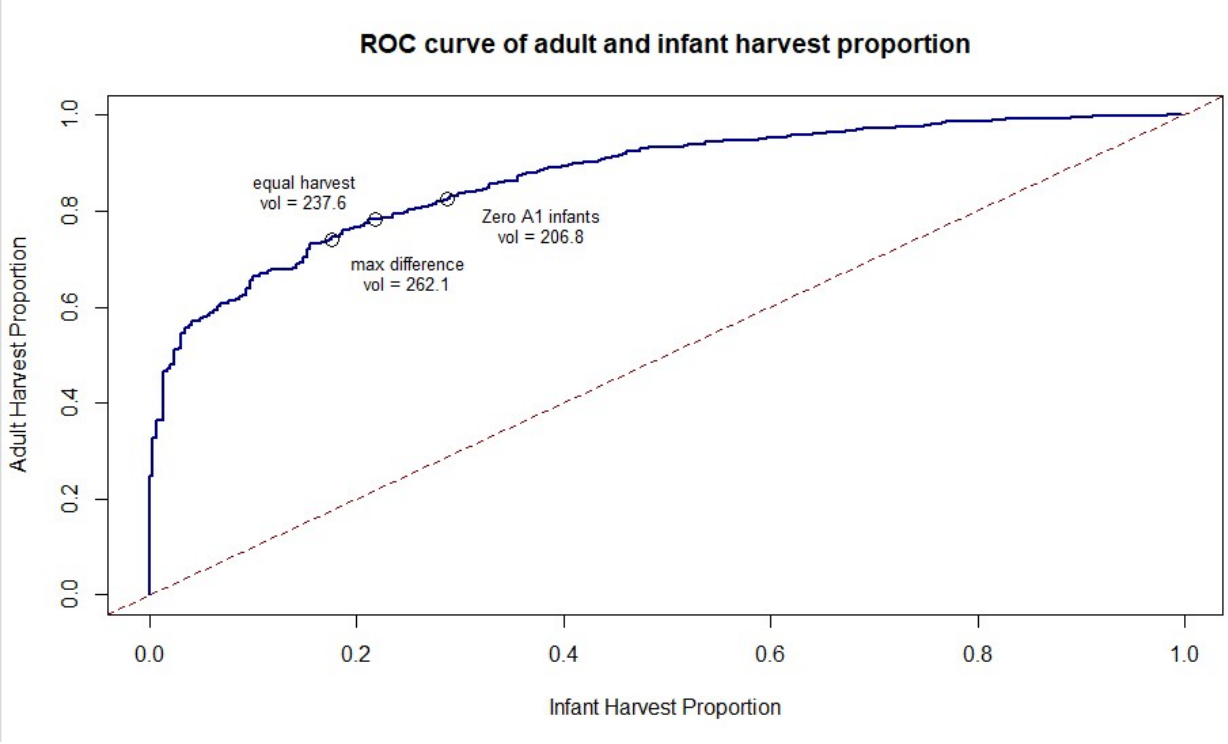
#(9)(a) Construct an ROC curve by plotting (1 - prop.adults) versus (1 - prop.infants).

Each point which appears correspondADULTs to a particular volume.value. Show the location

of the cutoffs determined in (7) and (8) on this plot and label each.

#```{r Part_9}

```
> plot(1-prop.infants, 1-prop.adults, type="l", col="darkblue", lwd=2,
+      main = "ROC curve of adult and infant harvest proportion",
+      xlab = "Infant Harvest Proportion", ylab = "Adult Harvest Proportion")
> abline(0,1, lty=2, col="red4")
>
> points(max.difference.infants, max.difference.adults, cex=1.5)
> text(0.15, 0.82, labels = c("equal harvest\n\n", "vol = 237.6"), cex=0.8)
> points(zero.A1.infants, zero.A1.adults, cex=1.5)
> text(0.25, 0.65, labels = c("max difference\n\n", "vol = 262.1"), cex=0.8)
> points(error.i, error.a, cex=1.5)
> text(0.38, 0.75, labels = c("max difference\n\n", "vol = 206.8"), cex=0.8)
>
```



#(9)(b) 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.

#```{r Part_9b}

```
> auc((1 - prop.infants),(1 - prop.adults))
[1] 0.8666894
```

#(10)(a) 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

#```{r Part_10}

```
> py_1 <- sum(mydata$VOLUME > volume.value[which.max(smooth.difference)])/nrow(mydata)

> py_2 <- sum(mydata$VOLUME> volume.value[volume.value>max(mydata[mydata$CLASS=="A1" &

+                                    mydata$TYPE=="I","VOLUME"])][1])/nrow(mydata)

> py_3 <- sum(mydata$VOLUME>volume.value[which.min(abs(prop.adults-(1-
prop.infants)))])/nrow(mydata)

>

> max.diff <- c(volume.value[which.max(smooth.difference)],max.difference.adults,
max.difference.infants, py_1)

> max.diff <- round(max.diff,digits=3)

> z.a.1 <- c(volume.value[volume.value>max(mydata[mydata$CLASS=="A1" &

+          mydata$TYPE == "I", "VOLUME"])][1],zero.A1.adults, zero.A1.infants, py_2)

> z.a.1 <- round(z.a.1,digits = 3)

> equal_error <- c(volume.value[which.min(abs(prop.adults-(1-prop.infants)))], error.a, error.i, py_3)

> equal_error<-round(equal_error,digits=3)

>

> bind<-rbind(max.diff, z.a.1, equal_error)

> rownames(bind)<-c("Max.Diff", "z.a.1", "equal_error")

> cutoff<-data.frame(bind)

> colnames(cutoff)<-c("Volume", "TPR", "FPR", "PropYield")

> cutoff
```

	Volume	TPR	FPR	PropYield
Max.Diff	262.143	0.742	0.176	0.584
z.a.1	206.786	0.826	0.287	0.676
equal_error	237.639	0.782	0.218	0.625

>

#**Essay Question: 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.**

#***Answer: (Enter your answer here.)***

There were three cutoffs choices for harvesting abalone one is use the maximum difference between adults and infants by volume and use that as the cutoff point. This minimized the amount of infant abalone which were harvested to less than ten percent. There is the method of not harvesting any abalone smaller or equal to the largest A1 class abalone. This reduces the amount of infant abalone harvested while maximizing the adult harvest. The third choice is to harvest at the point where the proportion of adults not harvested equals the proportion of infants harvested this is the equal harvest method. From the ROC chart all three points lie where the adult proportion of harvest is about 80 % while the proportion of infants harvested is about 20 percent. Equal harvest is the best method for minimizing infant harvest and Zero A1 infants is the best method for maximizing adult harvest.

#**Final Essay Question: 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: (Enter your answer here.)***

I would outline the choices for the customer with the specific outcomes the group I am presenting to is looking for. If there is a preservation concern, I would choose the method which harvests the fewest infants. If shear quantity is all the group was after I would suggest the method which creates the largest harvest regardless of type. It is imperative to make your audience aware that it is in their best interest to ensure they can harvest abalone forever.

The research done with the data provided does lack a means to differentiate the infants from the adults at smaller volumes. I would suggest that the company invest in finding out a better means of classifying smaller abalone to get better harvest yields while preserving the yields in the future. Research would be required to determine a better method of classifying small abalone.

I would recommend the equal error method because the harvest would generate .782 percent of the adult population while harvesting .2818 percent of the infants in comparison to the other methods a yield is very close to maximum while infant preservation is very good in this method.

From the past analysis we found that the volume of females was greater than males and it might be possible to use a combination of characteristics to classify adults. A study could be done in a different location. The results might be better in different waters. There may be improved results if the study is done at different times of the year. It may be possible that abalone display different characteristics depending on the time of year.