Data Analysis Assignment #2

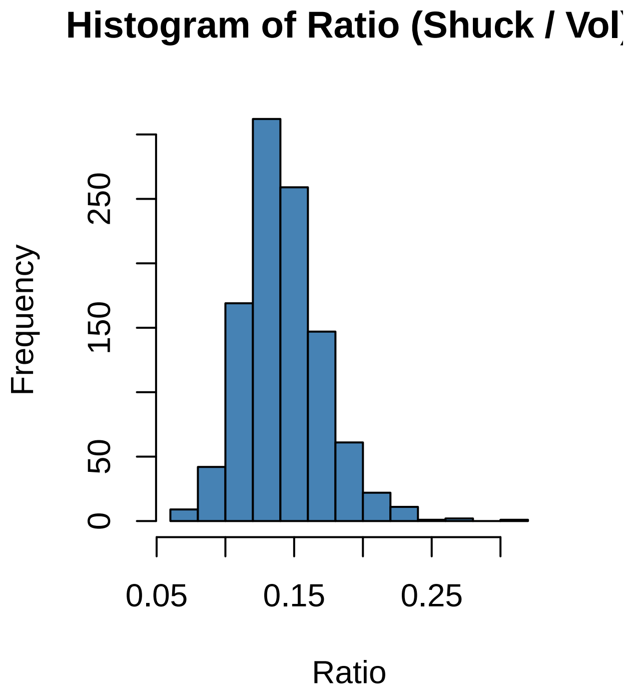
Granitz, Stephan

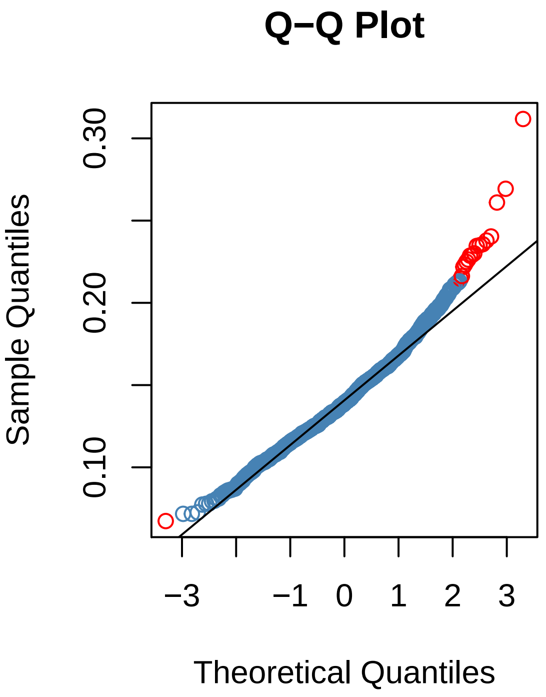
```
## 'data.frame':    1036 obs. of  8 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 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 ...
```

---

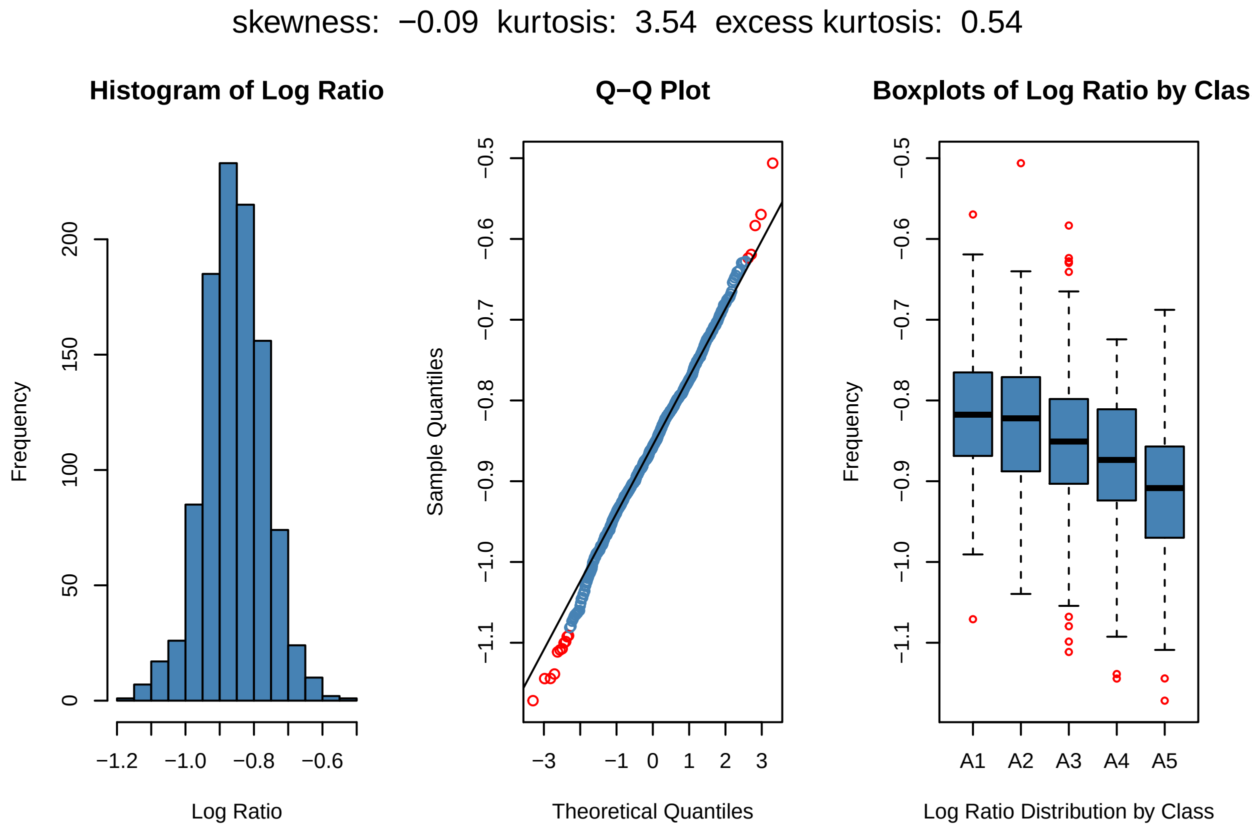
(1)(a) (1 point) Form a histogram and QQ plot using RATIO. Calculate skewness and kurtosis.

skewness: 0.71 kurtosis: 4.67 excess kurtosis: 1.67





(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 six boxplots of L_RATIO differentiated by CLASS.



(1)(c) (1 point) Test the homogeneity of variance across classes.

```
## [[1]]
## [1] "ratio:"
##
## [[2]]
##
##  Bartlett test of homogeneity of variances
##
## data:  RATIO by CLASS
## Bartlett's K-squared = 21.49, df = 4, p-value = 0.0002531
##
##
## [[3]]
## [1] "log ratio:"
##
## [[4]]
##
##  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 exhibits better conformance to a normal distribution with homogeneous

variances across age classes. This is shown with less skew in the histogram, more evenly distributed outliers in the QQ plot and boxplots, less skewness in the QQ plot, and failing to reject the null hypothesis with the Bartlett test of homogeneity of variances.

---

(2)(a) (2 points) Perform an analysis of variance on L_RATIO using CLASS and SEX as the independent variables. Assume equal variances. Peform two analyses. First, fit a model with the interaction term CLASS:SEX. Then, fit a model without CLASS:SEX. Obtain the analysis of variance tables.

```
##               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
```

```
##               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: Adding the interaction term had very little effect and was not significant. While the main effects of CLASS and SEX are statistically significant in a model of L_RATIO, the interaction between the two variables is not significant.

(2)(b) (2 points) For the model without CLASS:SEX (i.e. an interaction term), obtain multiple comparisons and interpret the results at the 95% confidence level.

```
##   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  0.0167236691 0.9412689
## M-I  0.017959386  0.003340824  0.0325779478 0.0111881
```

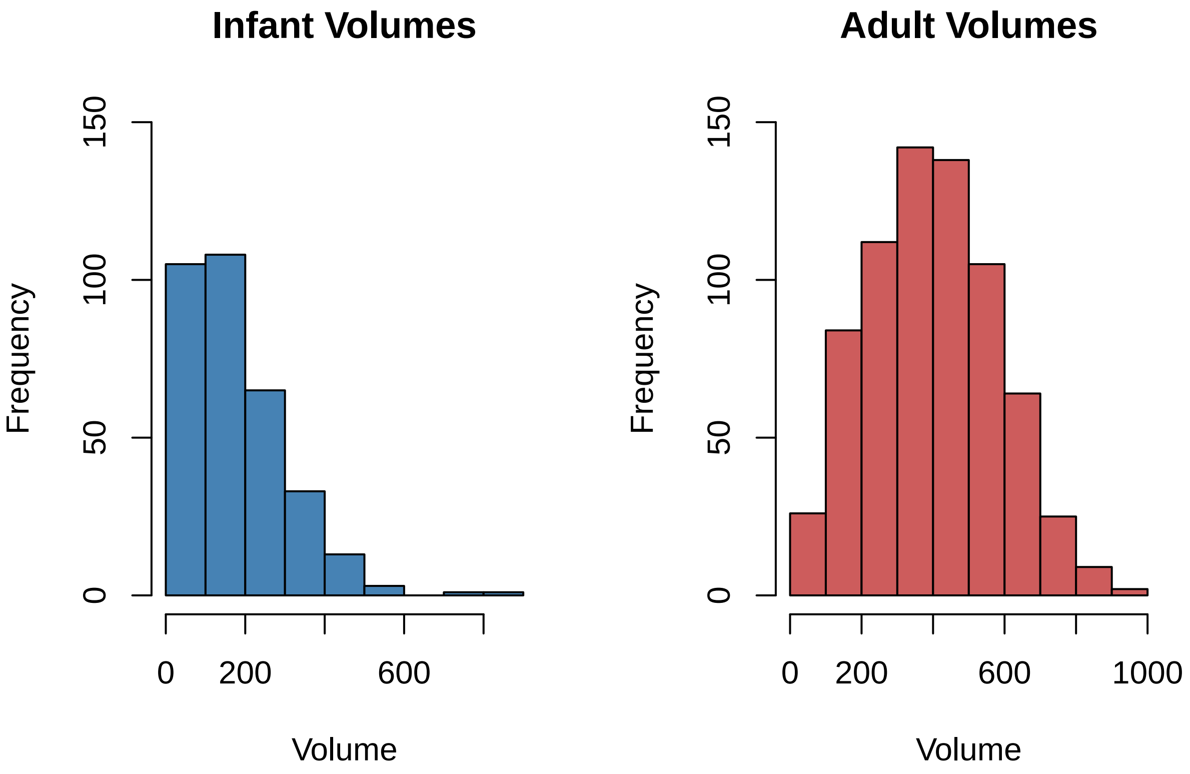
Question (2 points) : Interpret the trend across classes. Do these results suggest male and female abalones can be combined into a single category labeled as 'adults?' If not, why not?

Answer: Rejecting the null hypothesis that Infants and Male or Females are the same, and failing to reject the same for Males and Females suggests that Males and Females can be combined into an Adult group.

---

(3)(a) (2 points) Combine "M" and "F" into a new level, "ADULT". Form two histograms of VOLUME. One should display infant volumes, and the other: adult volumes.

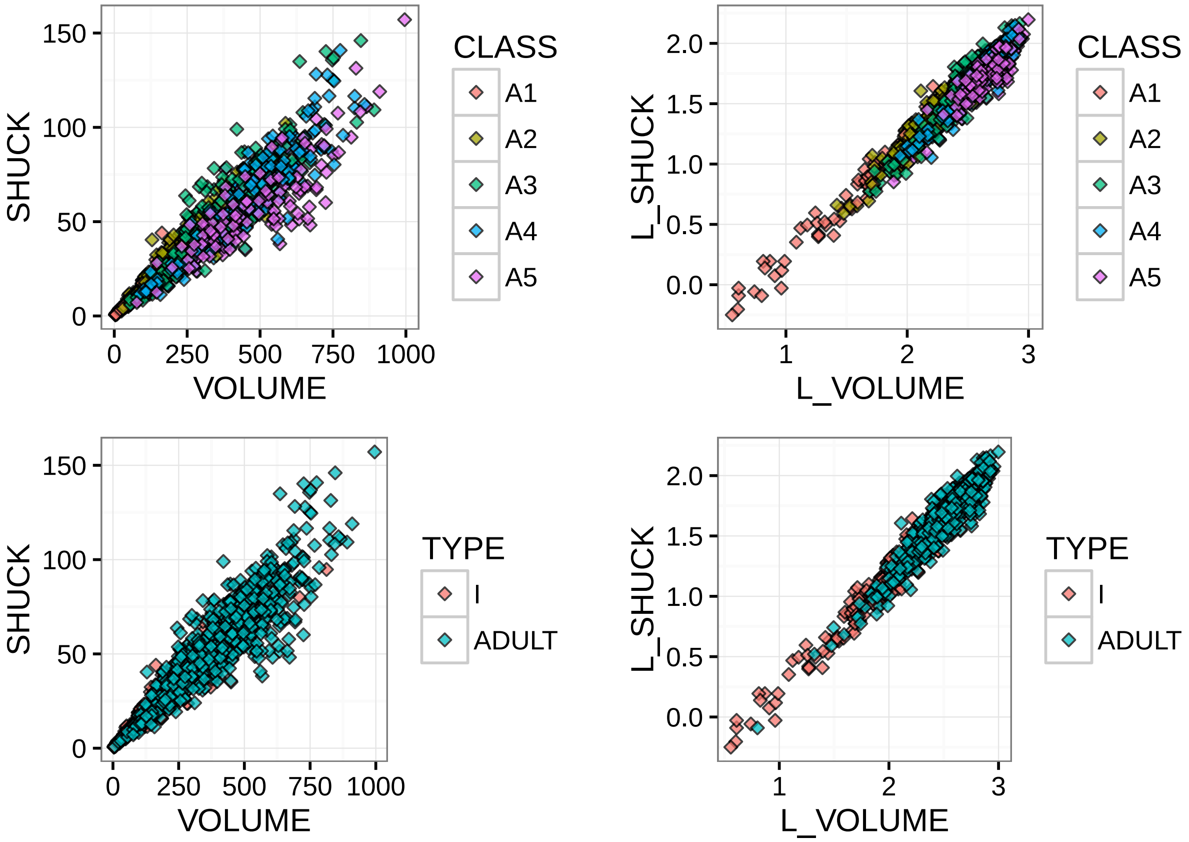
```
## The original levels F I M
## have been replaced by I ADULT
```



Question (2 points): Compare the histograms. How do the distributions differ? What are the implications in terms of identifying and, ultimately, separating infants from adults based on VOLUME?

Answer: Infants are right skewed, more heavily distributed among lower volumes with possible high volume outliers whie the Adult distribution is more normal. The majority of the Adult distribution is >300 whereas the Infants are mostly <300. This suggests that Volume is a part of the solution for splitting abalones.

(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.



Question (3 points): Compare the two scatterplots. What effect(s) does log-transformation appear to have on any relationship between SHUCK weight and VOLUME? Where do the various CLASS levels appear in the plots? Where do the levels of TYPE appear in the plots?

Answer: The measured VOLUME and SHUCK plots have a lot of overlap, making it difficult to distinguish clear lines between CLASS or TYPE. The log-transformed measures have a clear cut between a large group of A1 and INFANT abalones from the rest. The Infant Type has a large group at (<1.75, <0.75) and Adult type is almost completely above and to the right of these points.

---

(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. 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.

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

```
##
## 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 estimated coefficients suggest a stronger decrease in L__SHUCK for the higher the CLASS. This combined with the previous charts suggests that L__SHUCK increases more significantly at the lower CLASSes and then either flattens out or even decreases on average.

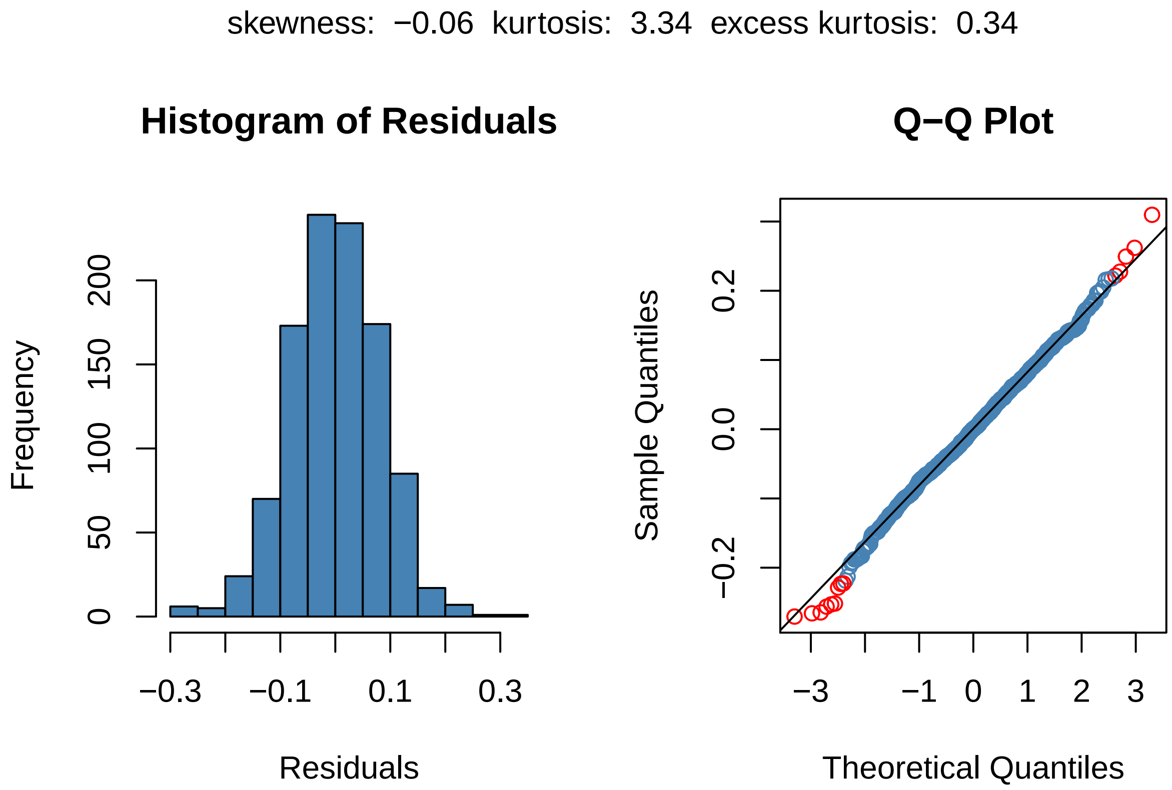
Question (2 points): Is TYPE an important predictor in this regression? (Hint: This question is not asking if TYPE is statistically significant, but rather how it compares to the other independent variables in terms of its contribution to predictions of L__SHUCK.) Explain your conclusion.

Answer: TYPE is less important than most of the CLASSes and much less important than L__VOLUME. This suggests that TYPE may not help in predicting L__SHUCK and possibly the reverse is also true.

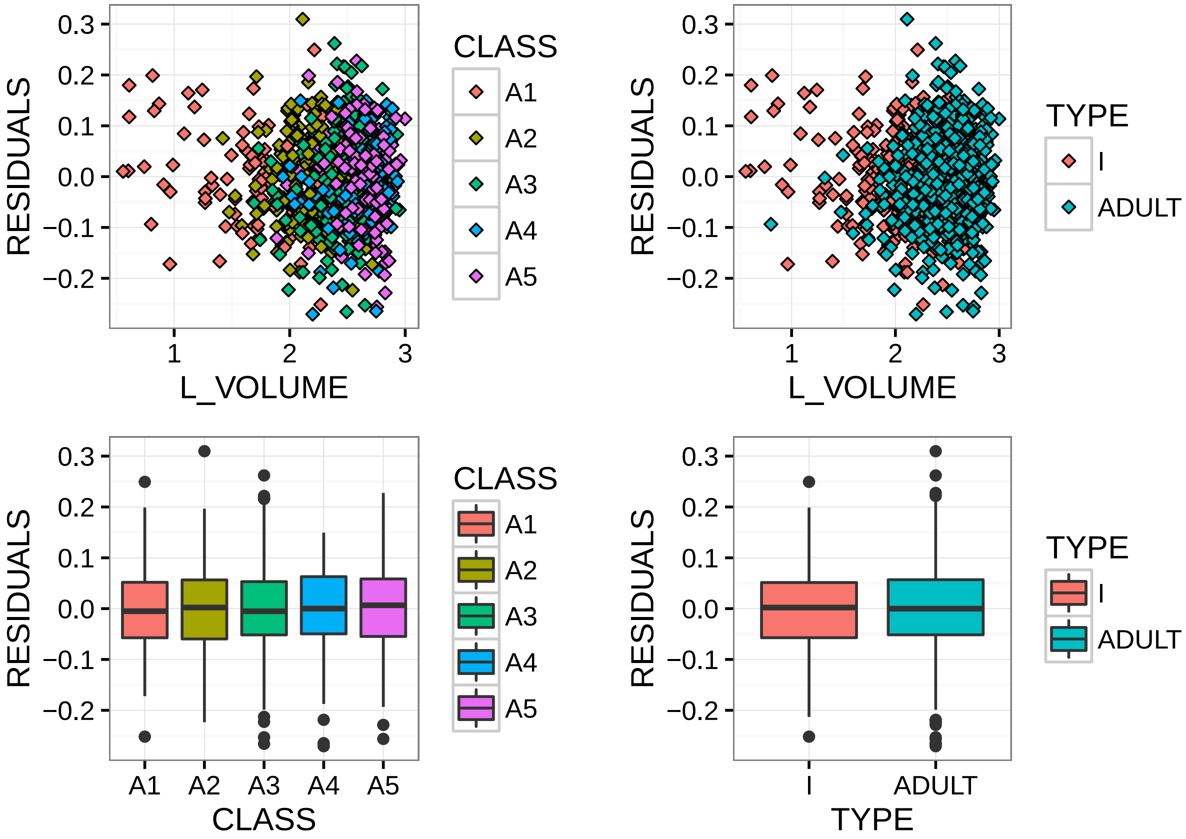
The next two analysis steps involve an analysis of the residuals resulting from the regression model in (4)(b) (see Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2).

---

(5)(a) (3 points) If "model" is the regression object construct a histogram and QQ plot. Compute the skewness and kurtosis.



(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. Test the homogeneity of variance of the residuals across classes.



```
##
##  Bartlett test of homogeneity of variances
##
## data:  RESIDUALS by 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: The RESIDUALS seem to be fairly evenly distributed and close to zero on both sides. There doesn't seem to be too defined of a pattern or shape but the large cluster to the right and more widely distributed scatter on the left when plotted against VOLUME may suggest their could be improvements made to 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.

---

(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.

```
delta <- diff(range(mydata$VOLUME)) / 1000
prop_infants <- numeric(1000)
prop_adults  <- numeric(1000)
vol_value <- numeric(1000)

inf_ind <- mydata$TYPE == "I"
adu_ind <- mydata$TYPE == "ADULT"

tot_infants <- sum(inf_ind)
tot_adults  <- sum(adu_ind)
min_vol <- min(mydata$VOLUME)

for (k in 1:1000) {
    value <- min_vol + k * delta
    vol_value[k] <- value
    prop_infants[k] <- sum(mydata$VOLUME[inf_ind] <= value) / tot_infants
    prop_adults[k]  <- sum(mydata$VOLUME[adu_ind] <= value) / tot_adults
}

num_infants <- sum(prop_infants <= 0.5)
split_infants <- min_vol + (num_infants + 0.5) * delta

num_adults <- sum(prop_adults <= 0.5)
split_adults <- min_vol + (num_adults + 0.5) * delta

head(vol_value)
```

```
## [1] 4.603851 5.595913 6.587974 7.580036 8.572097 9.564159
```

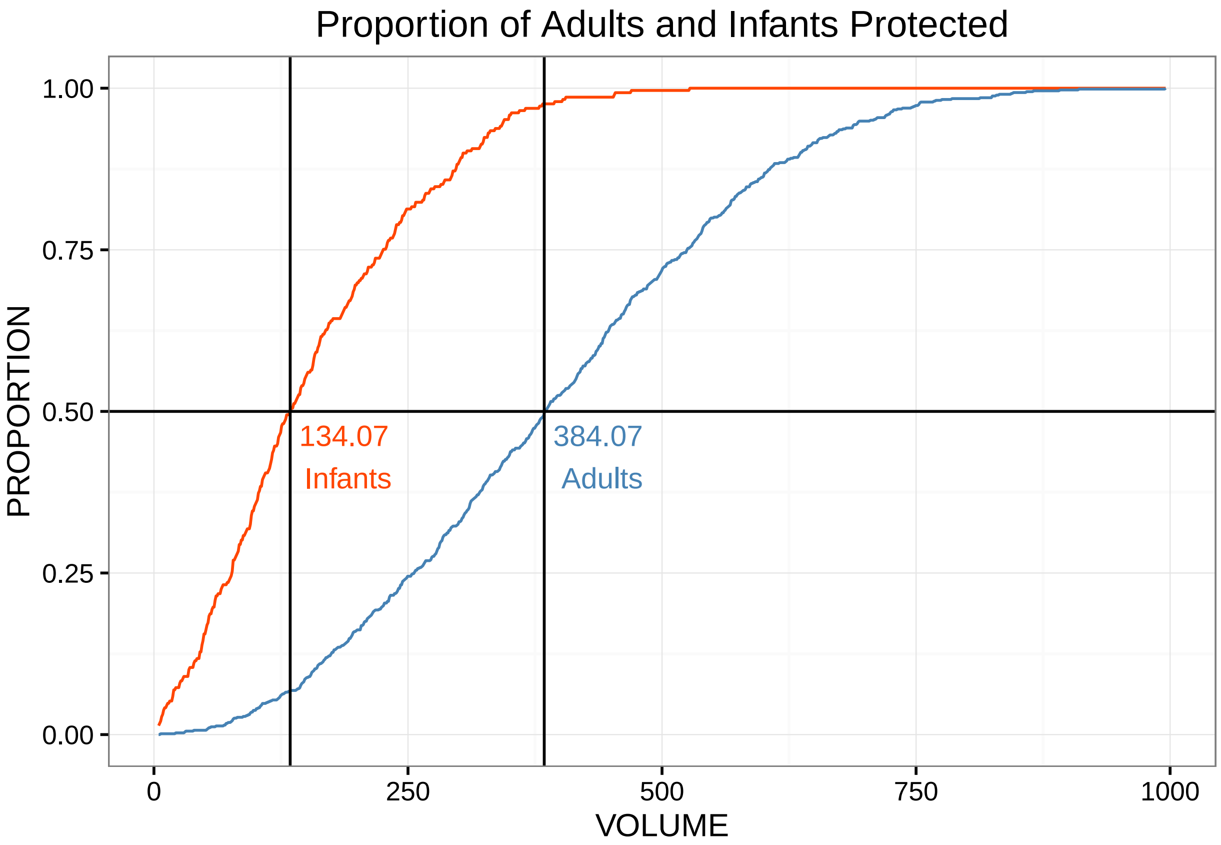
```
head(prop_infants)
```

```
## [1] 0.01384083 0.01730104 0.02076125 0.02768166 0.03114187 0.03806228
```

```
head(prop_adults)
```

```
## [1] 0.000000000 0.000000000 0.001338688 0.001338688 0.001338688 0.001338688
```

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



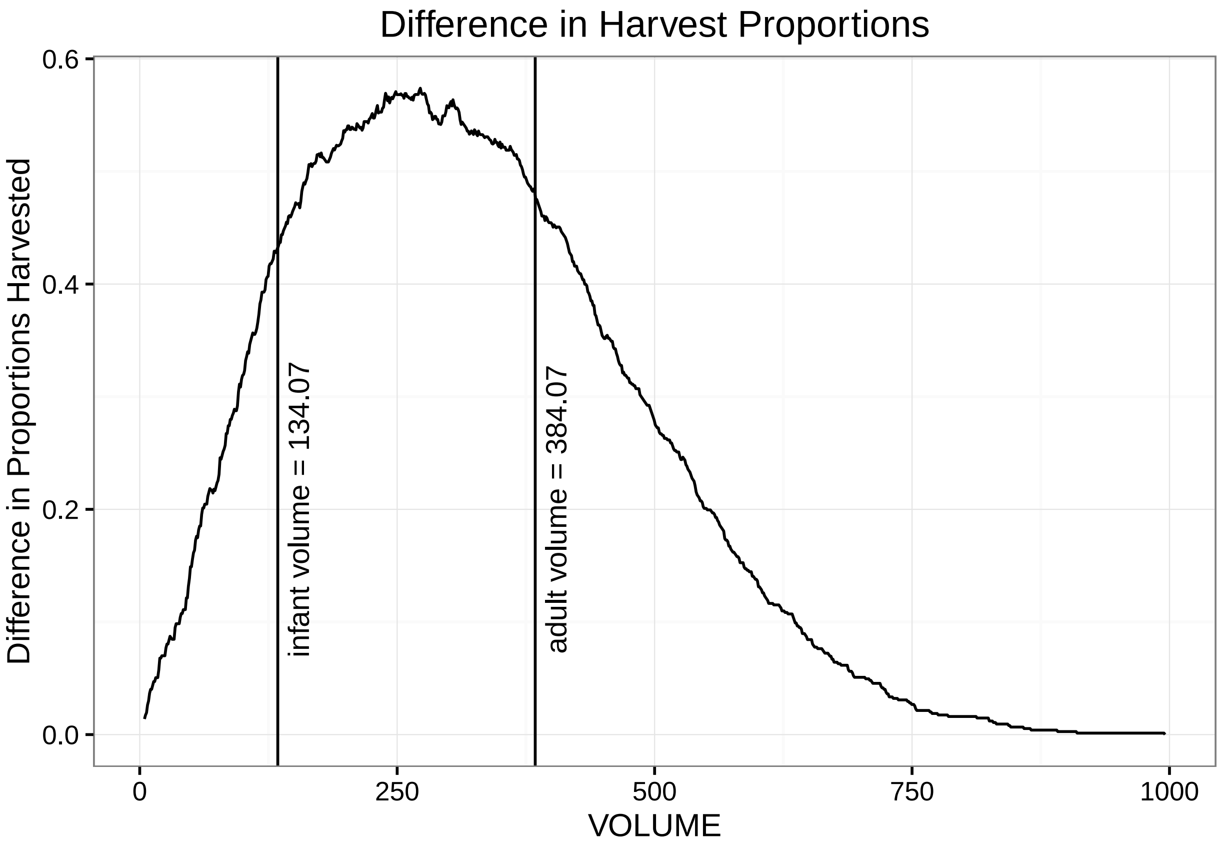
Question (2 points): The two 50% "split" values serve a descriptive purpose illustrating the difference between the populations. What do these values suggest regarding possible cutoffs for harvesting?

Answer: It appears there is a good cutoff between the two splits which would reduce both false positives and true negatives.

This part will address the determination of a volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. These proportions must be converted from "not harvested" to "harvested" proportions. The reason the proportion for infants drops sooner than adults is that infants are maturing and becoming adults with larger volumes.

---

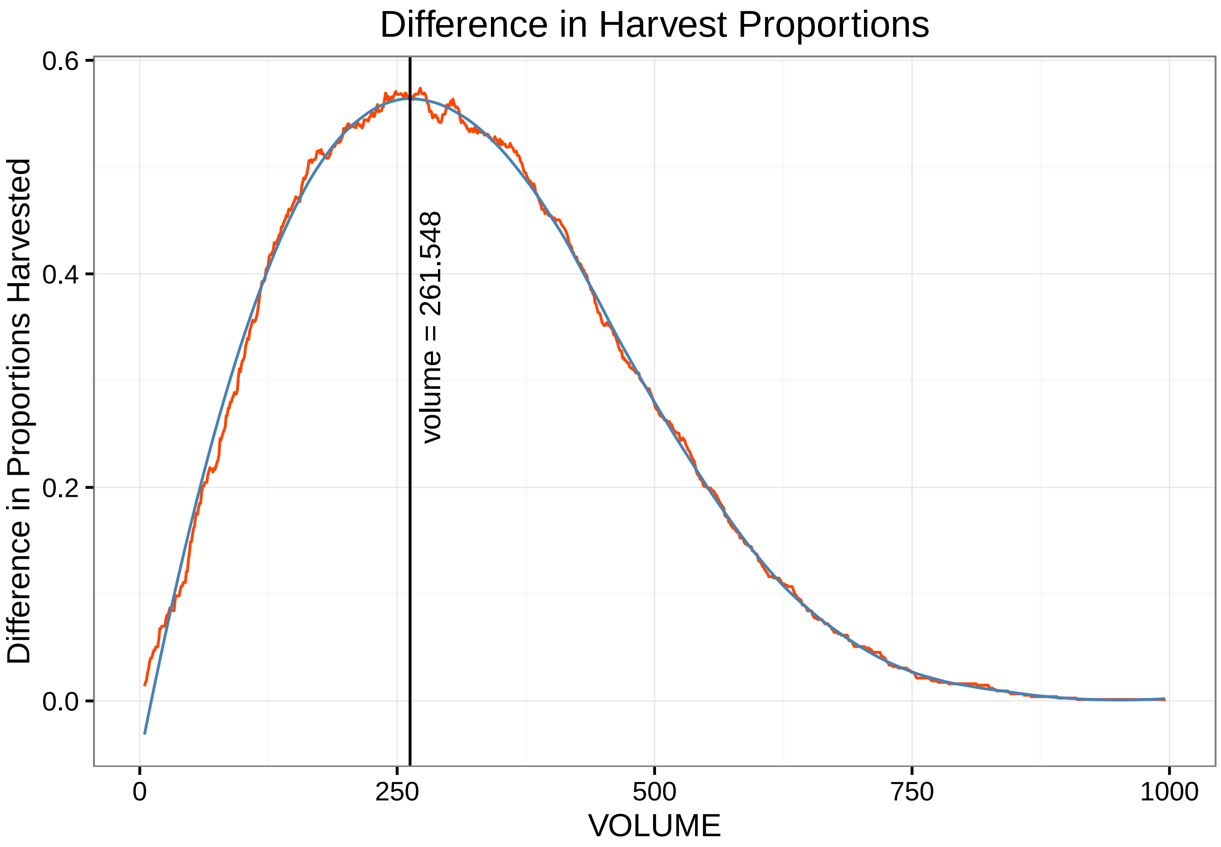
(7)(a) (1 point) Evaluate a plot of the difference versus volume. Compare to the 50% "split" points determined in (6)(a). There is considerable variability present in the peak area of this plot. The observed "peak" difference may not be the best representation of the data. One solution is to smooth the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



(7)(b) (1 point) Individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference.

```
y_loess_a <- loess(1 - prop_adults ~ vol_value,
                   span = 0.25, family = c("symmetric"))
y_loess_i <- loess(1 - prop_infants ~ vol_value,
                   span = 0.25, family = c("symmetric"))
smooth_diff <- 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. Show the estimated peak location corresponding to the cutoff determined.



(7)(d) (1 point) What separate harvest proportions for infants and adults would result if this cutoff is used?

```
## [1] "True positive rate: 0.741633199464525"
```

```
## [1] "False positive rate: 0.176470588235294"
```

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.

```
## [1] "cutoff: 206.9843918625"
```

```
## [1] "True positive rate: 0.825970548862115"
```

```
## [1] "False positive rate: 0.28719723183391"
```

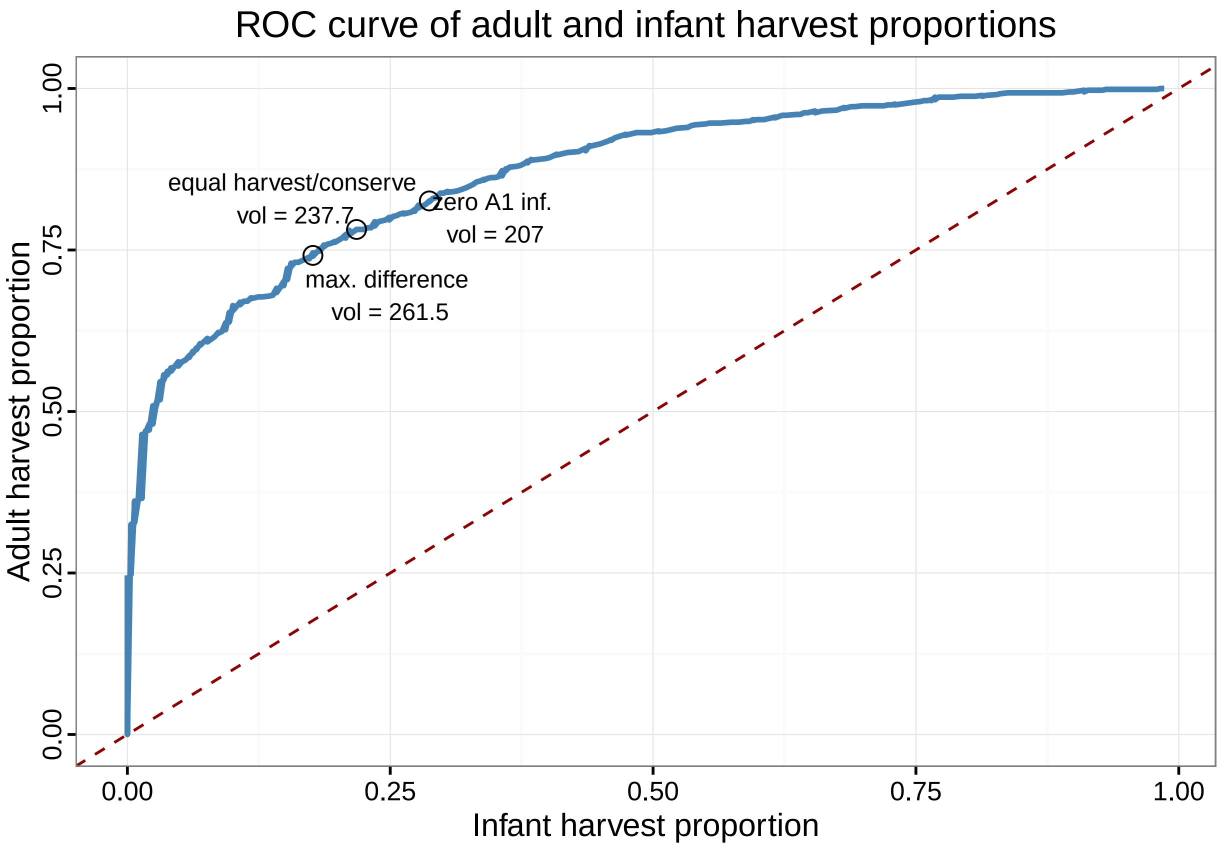
(8)(b) (2 points) Another cutoff can be determined for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: 'missed' adults and wrongly-harvested infants. This leaves for discussion which is a greater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

```
## [1] "cutoff: 237.73829751"
```

```
## [1] "True positive rate: 0.781793842034806"
## [1] "False positive rate: 0.217993079584775"
```

---

(9)(a) (7 points) Construct an ROC curve by plotting (1 - prop.adults) versus (1 - prop.infants). Show the location of the cutoffs determined in (7) and (8) on this plot and label each.



(9)(b) (1 point) Numerically integrate the area under the ROC curve and report your result.

```
## [1] "Area under ROC curve: 0.856331902002477"
```

---

(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

```
## # A tibble: 3 × 5
##          strategy   volume       tpr       fpr prop_yield
##             <chr>    <dbl>     <dbl>     <dbl>      <dbl>
## 1  max difference 261.5478 0.7416332 0.1764706  0.5839768
## 2 zero A1 infants 206.9844 0.8259705 0.2871972  0.6756757
## 3     equal error 237.7383 0.7817938 0.2179931  0.6245174
```

Question: (3 points) 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:* *The 'max difference' cutoff is the most conservative, reducing the false positive rate but also having the lowest proportional yield. The 'zero A1 infants' cutoff is the most aggressive with the highest true positive rate and proportional yield, however there is a nearly 0.3 false positive rate. The 'equal error' is in between these two.

Question (5 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: If I was presenting these results I would (1) not suggest a specific strategy but would outline the risks and rewards of a couple of strategies describing the effects on abalone populations with false positives and on profitability with true negatives and the proportion of each in the strategies. I would caution that (2) every analysis is limited by the quality of the data collection and that there are many challenges in propoerly measuring some of the key inputs for determining the age of abalones. The study had a number of outliers suggesting the abalones could have been mislabeled, already known to be a difficult process. Considering the risks of overharvesting, (3) if we must move forward on the current analysis I would push towards a more conservative cutoff to avoid harvesting too many young infants which are required to continue the sustainable growth of the abalones. Going forward (4) I would suggest testing new measures of the abalones to try and find a better way to reduce both false positives and true negatives. These measurements should include environmental, geographical, and dietary factors.