Data Analysis Assignment #2

Thakur, Prabhat

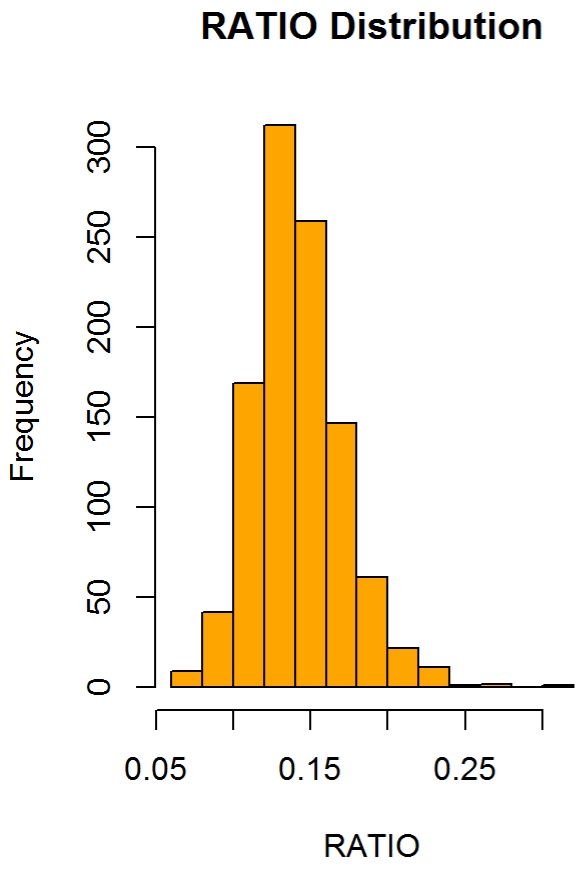
Submit both the .Rmd and .html files for grading. You may remove the instructions and example problem above, but do not remove the YAML metadata block or the first, “setup” code chunk. Address the steps that appear below and answer all the questions. (75 points possible)

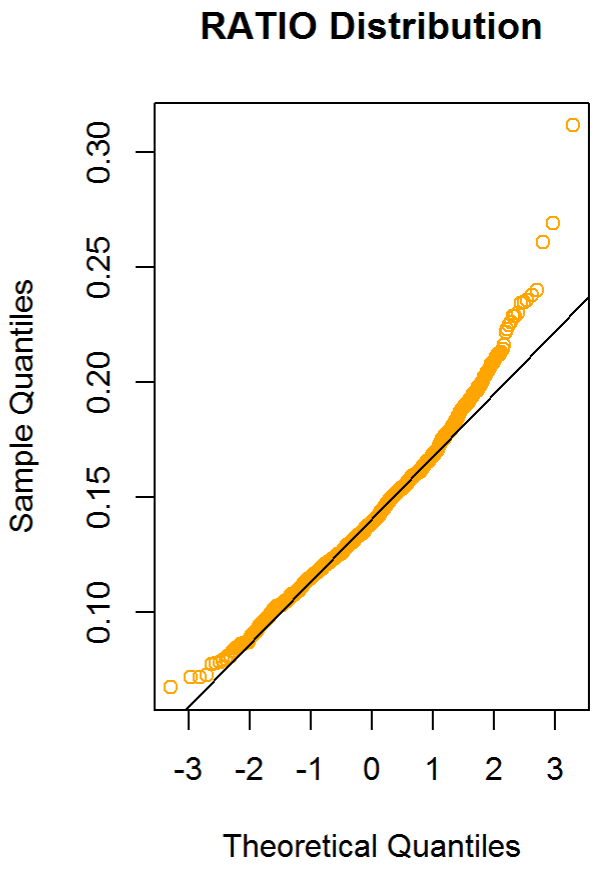
Data Analysis #2

```
## '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.

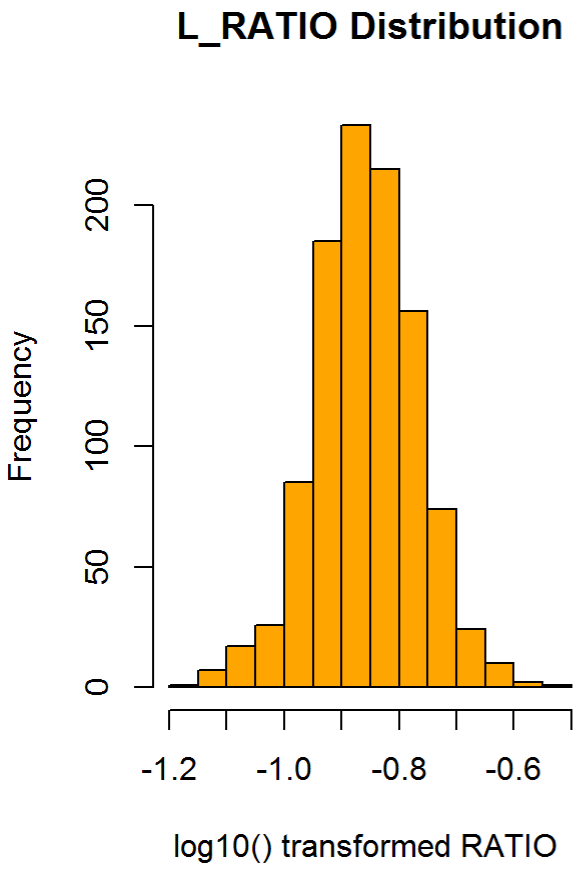


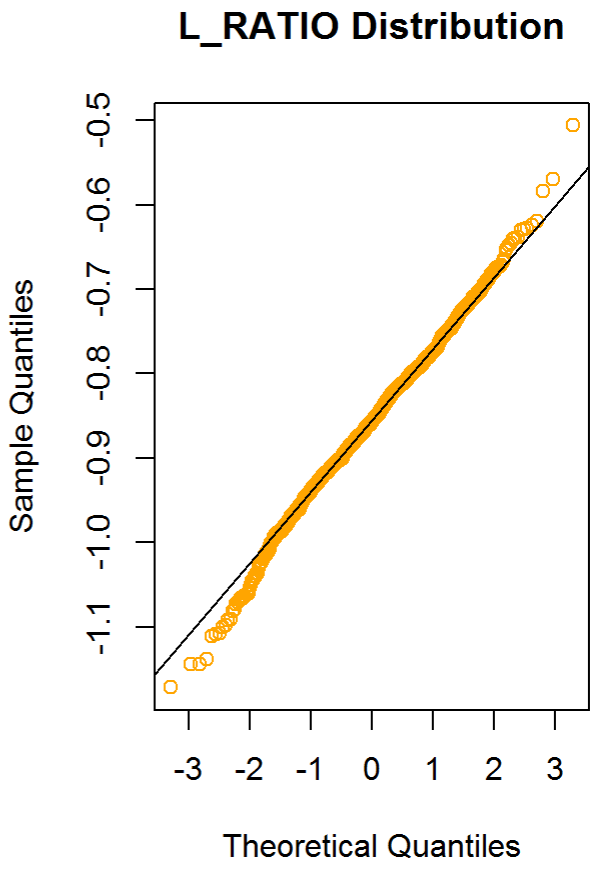



```
## RATIO skewness & kurtosis using 'rockchalk' package and 'moment' package.
##
##
```

```
##          'rockchalk' package 'moment' package
## skewness           0.7157417        0.7157417
## kurtosis           1.6763211        4.6763211
```

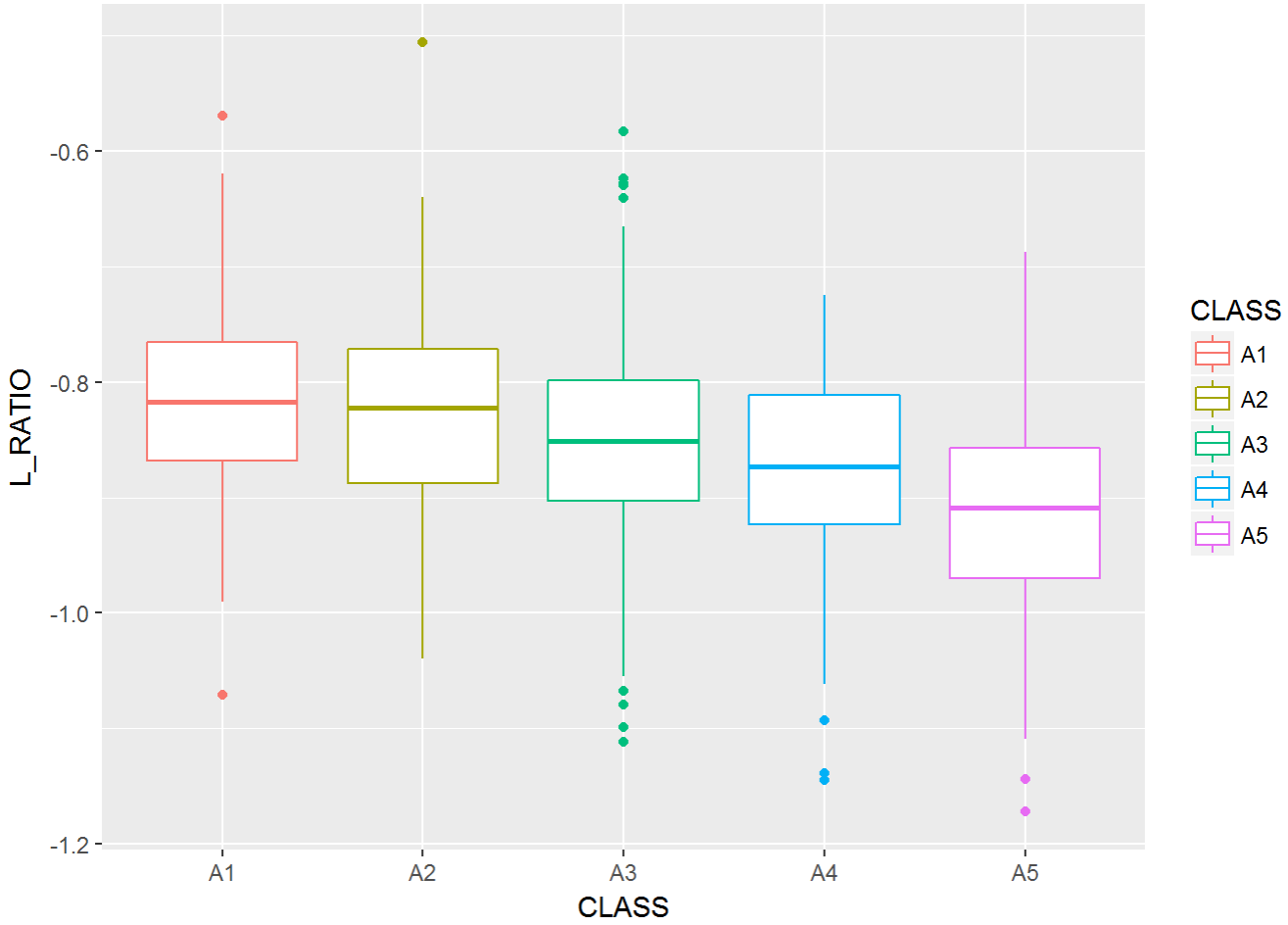
(1)(b) (2 points) Tranform RATIO using log10() to create L_RATIO (see Kabacoff Section 8.5.2, p. 199-200). Form a histogram and QQ plot using L_RATIO. Calculate the skewness and kurtosis. Create a display of five boxplots of L_RATIO differentiated by CLASS.






```
## L_RATIO skewness & kurtosis using 'rockchalk' package and 'moment' package.
##
##
```

```
##          'rockchalk' package 'moment' package
## skewness         -0.09405162      -0.09405162
## kurtosis          0.54226600       3.54226600
```



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

```
##
## Testing the null hypothesis that RATIO and L_RATIO variances are all equal across Classes u
sing the bartlett.test()
```

```
## Degrees of freedom:  df = 4
```

```
## Significance level:  alpha = 0.05
```

```
##
## Test homogeneity of variance in RATIO across classes.
```

```
## ------------------------------------------------------------------
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  RATIO by CLASS
## Bartlett's K-squared = 21.49, df = 4, p-value = 0.0002531
```

```
## Critical value:  Chi-square =  9.487729
```

```
## Test statistic:  T =  21.49048
```

```
## Critical region: Reject H0 if Test statistic T >  9.487729
```

```
##
## RATIO variances are not equal across Classes. Because the test statistic is larger than the
 critical value, we reject the null hypotheses at the 0.05 significance level and conclude tha
t at least one Class variance is different from the others. Reject the null hypothesis.
```

```
##
##
## Test homogeneity of variance in L_RATIO across classes.
```

```
## ------------------------------------------------------------------
```

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

```
## Critical value:  Chi-square =  9.487729
```

```
## Test statistic:  T =  3.189142
```

```
## Critical region: Reject H0 if Test statistic T >  9.487729
```

```
##
## L_RATIO variances are all equal across Classes. Because the test statistic is less than the
 critical value, we pass the null hypotheses at the 0.05 significance level and conclude that
variances in each of the classes are the same. No reason to reject null hypothesis.
```

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: From comparing skewness and kurtosis statistic of RATIO and L_RATIO, its clear that L_RATIO exhibits better conformance to a normal distribution. L_RATIO distribution is very close to the normal distribution with skewness -0.09405162 and 'rockchalk' excess kurtos is 0.542266. When compared to normal distribution values 0 and 0 respectively. Histograms and QQ plots of RATIO or L_RATIO also confirms the above conclusion. From Bartlett's test for homogeneity of variances in RATIO and L_RATIO across Classes, we can conclude that RATIO variances are not equal across Classes whereas L_RATIO variances are. To perform one-way analysis of variance, assumption of equal variances is made. The log10() transformed value of RATIO (L_RATIO) confirms this assumption. A one-way analysis of variance can be performed on L_RATIO.

---

(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).

```
## Model with the interaction term CLASS:SEX
```

```
##               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
```

```
##
##
## Model without the interaction term CLASS:SEX
```

The analysis of variance is comparing group means. There is no regression slope involved.

```
##               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
```

The ANOVA indicates CLASS and SEX are statistically significant factors that account for variability in L_RATIO. Absence of an interaction means the effect of CLASS or SEX does not change when the level of the other factor is changed.

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: This is an example of The ANCOVA F test. Standard ANCOVA designs assumes homogeneity of regression slopes. It's assumed that the regression slope for predicting L_RATIO from CLASS is same in each of the 3 SEX groups. A significant interaction would imply that the relationship between CLASS and L_RATIO depends on the Type of SEX variable. From the test we can see that CLASS and SEX has a significant effect on L_RATIO, but CLASSSEX* interaction term has no significant effect due to a p-value of 0.86709, the interaction is non-significant. This supports the assumption of equality of regression slopes for predicting L_RATIO from CLASS is same in each of the 3 SEX groups. In another words, L_RATIO in different CLASS doesn't very by SEX groups. In the second analysis of variance without the CLASS*SEX interaction term. The result is that CLASS and SEX still has significant effect on L_RATIO

(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).

```
##   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
```

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: Tukey's Honestly Significant Difference (HSD) Test determines the critical difference necessary between the means of any two treatment levels for the means to be significantly different. For 95% confidence level, we can see that in age classes, all pairs except A1 and A2, are significantly different from each other (p < 0.05). L_RATIO mean for A1 and A2 Classes aren't significantly different from each other. Similarly for Sex group, we can see significant differences between infant- male and infant-female however L_RATIO mean for Male(M) and Female (F) SEX aren't significantly different. For M-F pair, P value = 0.94 suggests that there is not much difference in the mean values of M and F abalones and both can be combined into a single category for further analysis.

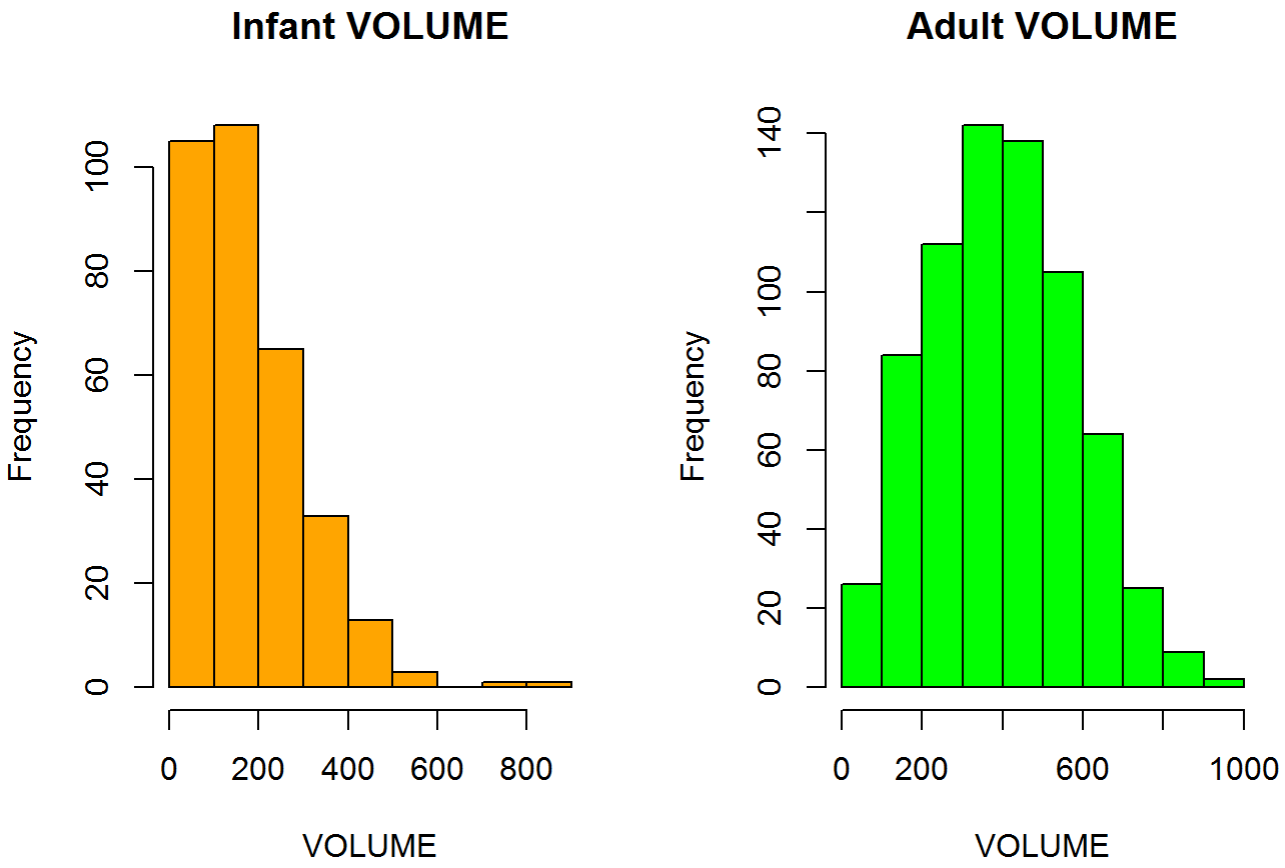
-1 point The trend in the coefficients for age class indicates average L_RATIO is declining with age.

(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.

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

```
##  SEX         LENGTH           DIAM            HEIGHT
##  F:326   Min.   : 2.73   Min.   : 1.995   Min.   :0.525
##  I:329   1st Qu.: 9.45   1st Qu.: 7.350   1st Qu.:2.415
##  M:381   Median :11.45   Median : 8.925   Median :2.940
##          Mean   :11.08   Mean   : 8.622   Mean   :2.947
##          3rd Qu.:13.02   3rd Qu.:10.185   3rd Qu.:3.570
##          Max.   :16.80   Max.   :13.230   Max.   :4.935
##      WHOLE             SHUCK             RINGS        CLASS
##  Min.   :  1.625   Min.   :  0.5625   Min.   : 3.000   A1:108
##  1st Qu.: 56.484   1st Qu.: 23.3006   1st Qu.: 8.000   A2:236
##  Median :101.344   Median : 42.5700   Median : 9.000   A3:329
##  Mean   :105.832   Mean   : 45.4396   Mean   : 9.993   A4:188
##  3rd Qu.:150.319   3rd Qu.: 64.2897   3rd Qu.:11.000   A5:175
##  Max.   :315.750   Max.   :157.0800   Max.   :25.000
##      VOLUME            RATIO            L_RATIO           TYPE
```

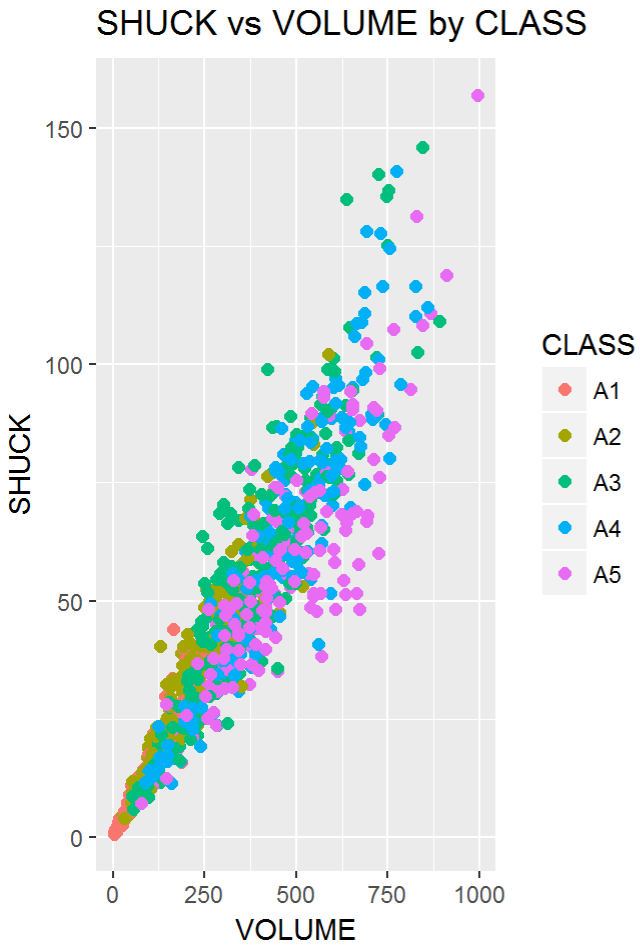
```
##  Min.   :  3.612   Min.   :0.06734   Min.   :-1.1717   I    :329
##  1st Qu.:163.545   1st Qu.:0.12241   1st Qu.:-0.9122   ADULT:707
##  Median :307.363   Median :0.13914   Median :-0.8565
##  Mean   :326.804   Mean   :0.14205   Mean   :-0.8566
##  3rd Qu.:463.264   3rd Qu.:0.15911   3rd Qu.:-0.7983
##  Max.   :995.673   Max.   :0.31176   Max.   :-0.5062
```

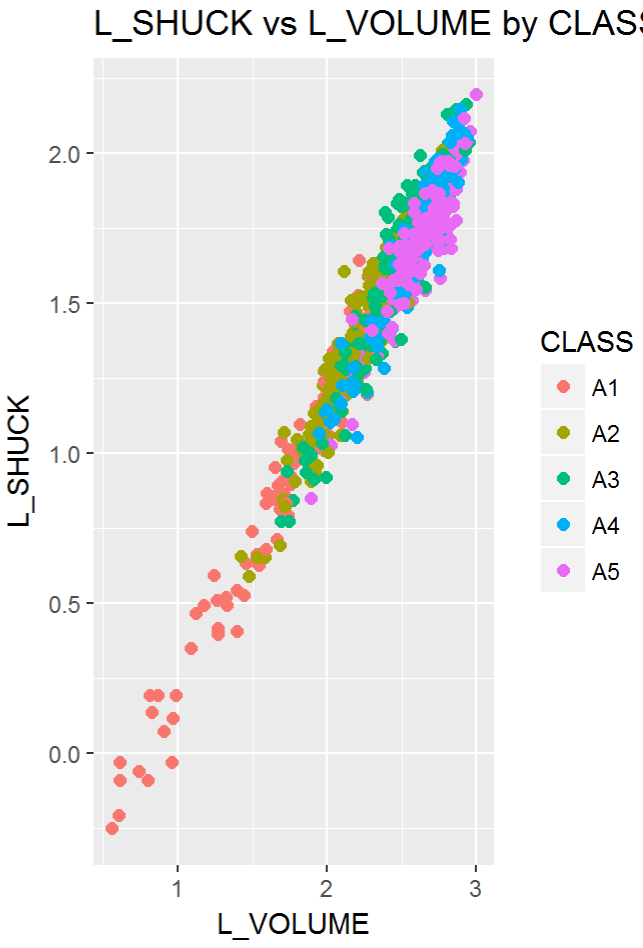


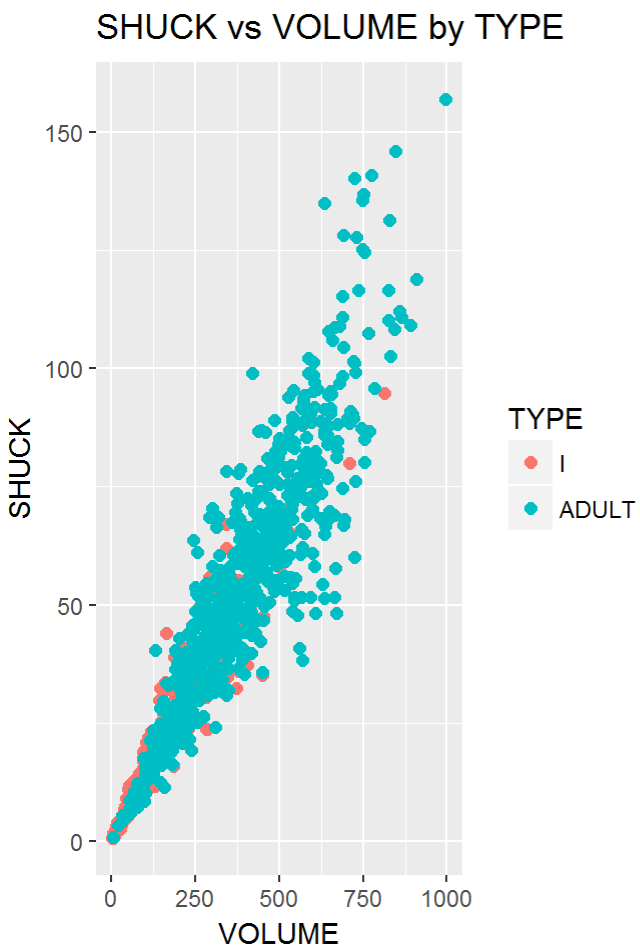
Question (2 points): Compare the histograms. How do the distributions differ? Are there going to be difficulties separating infants from adults based on VOLUME?

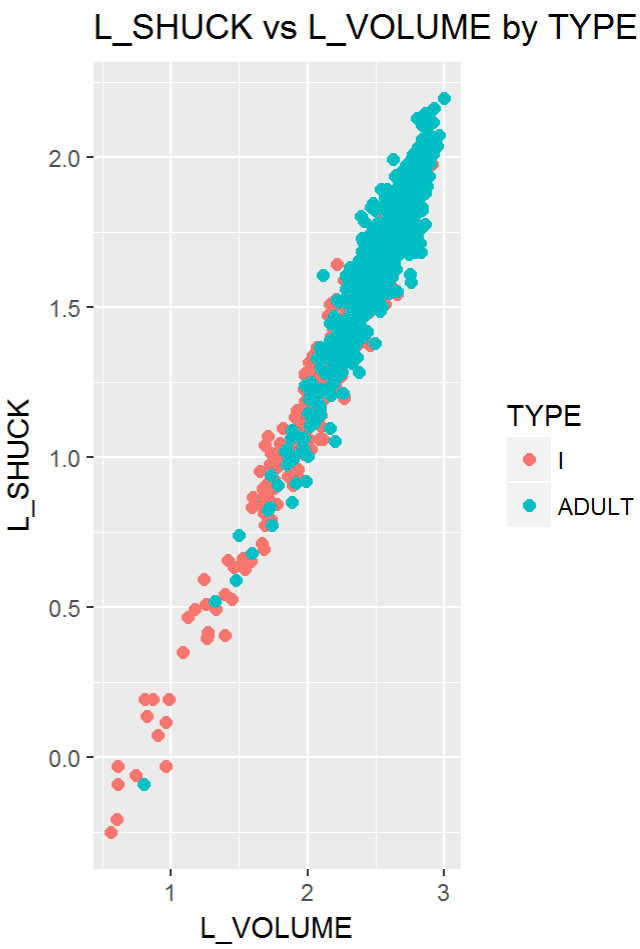
Answer: From the infant volume histogram, we can observe that Infant volume is right skewed and have multiple outliers which suggests that it is not normally distributed. On the other hand from adult volume histogram, the data is slightly right skewed and have only 1 outlier and it appears that adult volume is very much similar to normal distribution. We can see that, dispersion of Infant and Adult volume is overlapping considerably up to 400 unit. About 75% infant's volume ranges from 3.6 to 247 units overlaps with about 25% of adult volume which ranges from 6 to 258 unit. Also there are about 25% infants spread in the volume range where 75% of adults are. It is apparent from the analysis that separating infants from adults based on VOLUME will not produce correct results and will result in errors and wrong TYPE classification for some 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 (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.


SHUCK vs VOLUME by CLASS
SHUCK
VOLUME
0
50
100
150
0
250
500
750
1000
CLASS
A1
A2
A3
A4
A5


L_SHUCK vs L_VOLUME by CLASS
L_SHUCK
L_VOLUME
0.0
0.5
1.0
1.5
2.0
1
2
3
CLASS
A1
A2
A3
A4
A5


SHUCK vs VOLUME by TYPE
SHUCK
VOLUME
0
50
100
150
0
250
500
750
1000
TYPE
I
ADULT


L_SHUCK vs L_VOLUME by TYPE
L_SHUCK
L_VOLUME
0.0
0.5
1.0
1.5
2.0
1
2
3
TYPE
I
ADULT

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?

Answer: SHUCK and VOLUME has lot of variability in both CLASS as well as TYPE levels on the right upper side. For both, the variability is more in 4th quartile (between 3Q and max value), which can also be confirmed by looking at the difference in their 3Q and max value. For SHUCK, 3Q value is 64.29 and max value is 157.1 and for VOLUME is it 463.3 and 995.7. IF we look further, it appears that, A1 and A2 classes are most homogenous and A4 and A5 classes has most variability. This variability can be linked to right skewness and outliers which may result in weak linear relationship. On the other hand, it can be observed from both scatterplots of log transformed values L_SHUCK vs L_VOLUME that, the variability has reduced significantly compared to original values. The log transformed values are showing strong and positive linear association and appears to be highly correlated. Another observation from both scatterplots of original values is that, all CLASSES and TYPES are overlapping but in log transformed plots, CLASS A1 is somewhat separated from other CLASSES A2, A3, A4 and A5 also in TYPE differentiated plot we can see that most of the Infants are in the lower left side of the plot and Adults are in the upper right side. We can see that log transformed values L_SHUCK vs L_VOLUME are normally distributed. In infect it has transformed both values to satisfy Normality, Linearity and Homoscedasticity assumption for developing multiple linear regression model. It seems Log transformation of variables can be helpful in further analysis of abalones data. From the plots differentiated by CLASS we can see that, SHUCK weight is not increasing in same ratio as volume in the CLASS A3, A4 and A5, and this characteristic is increasing as abalone class is increasing. We can see from the plot that more abalones from CLASS A5 are in the lower side of the plot. From the plots differentiated by TYPE we can see that, SHUCK weight is increasing more in Adults when compared to VOLUME. We can see that more Adult abalones are in the upper side of the plot.

---

(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.

```
## 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: From the summary output of regression model, we can see that age class A1 and Infant abalones are used as the baseline and added in intercept coefficient. The Volume is emerging as dominate predictor in this model, reconfirming out finding in earlier analysis. The regression coefficients estimates for CLASS levels indicates negative coefficients and are increasing in negative direction. This trend seems to indicate that if you move to a higher age CLASS, the coefficient becomes increasingly negative which suggests inverse relationship between age CLASS and log SHUCK variable. In another words as age class increases, the log of shuck decreases when log of VOLUME parameter is kept constant. The regression coefficients for CLASS A3, A4, and A5 are significantly different from zero (p < 0.001) and change in these predictor variable has impact on abalone log SHUCK weight. However compared to Volume predictor, CLASS are very weak predictor of log SHUCK variable. Similar observation was made from the earlier scatter plots. Growth in SHUCK weight gets slower compared to Volume for CLASSES in A3, A4 and A5.

SHUCK relative to VOLUME decreases on average as a function of CLASS.

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.

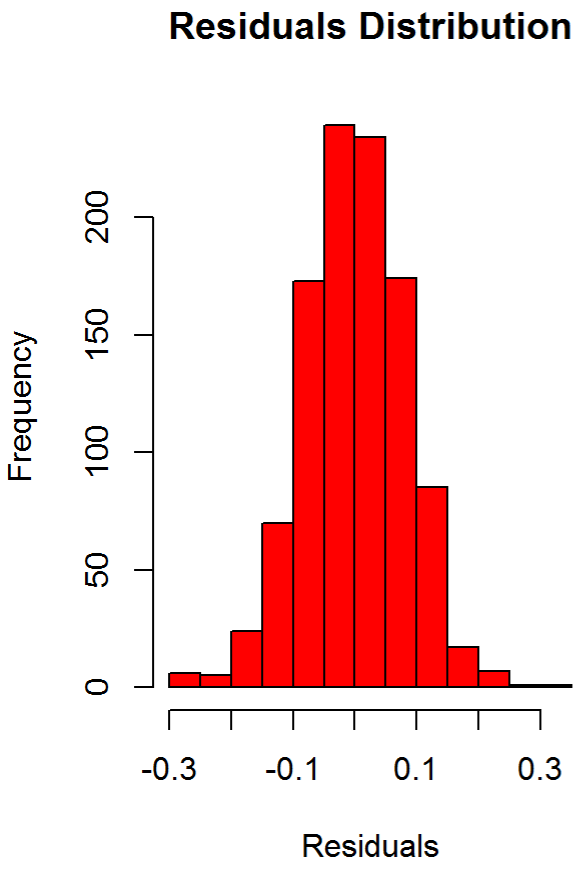
Answer: The summary output estimate coefficient for TYPE suggests that TYPE has very less contribution in predicting abalones log SHUCK weight compared to log Volume and CLASS variables. For example, the regression coefficient for TYPE Adult is 0.02, suggesting that an increase of 1 percent in Adult is associated with a 0.02% percent increase in the log SHUCK weight, when CLASS variable is kept constant. On the other hand, the coefficient CLASS variables are -0.047,-0.0757 and -0.117 for A3, A4 and A5 respectively. Clearly, increase of 1 percent in CLASS variable will have more effect in log SHUCK weight. The regression coefficients for CLASS A3, A4, and A5 are significantly different from zero (p < 0.001) 99.9% whereas for TYPE Adult, p-value is significant different at <0.01 99%. It appears that TYPE is not a significant predictor of log SHUCK variable.

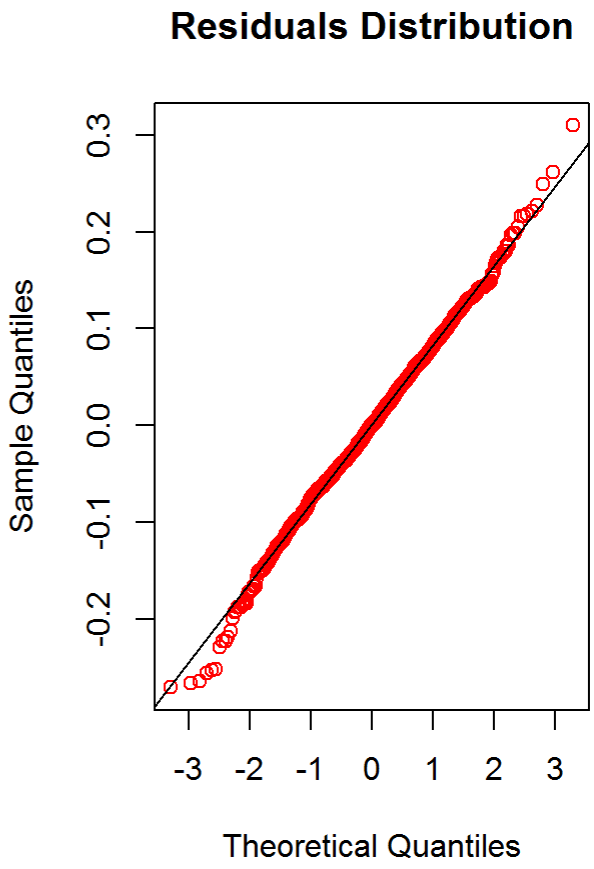
It is small in magnitude compared to the intercept term. Excluding it changes the adjusted R-squared from 0.9501 to 0.9498. Relatively speaking, it is not an important variable in this regression model in terms of explaining the variation in L_SHUCK.

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.

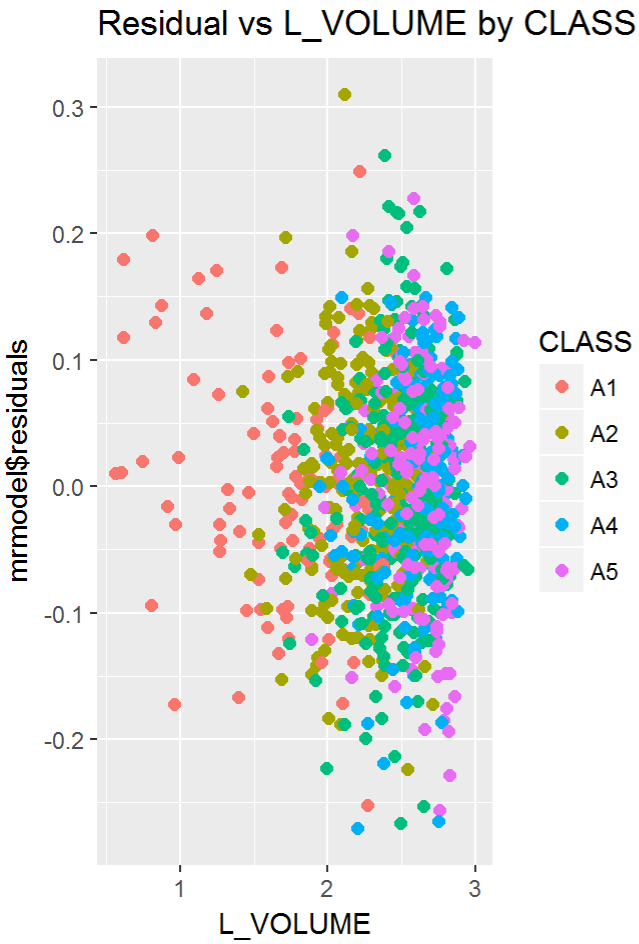


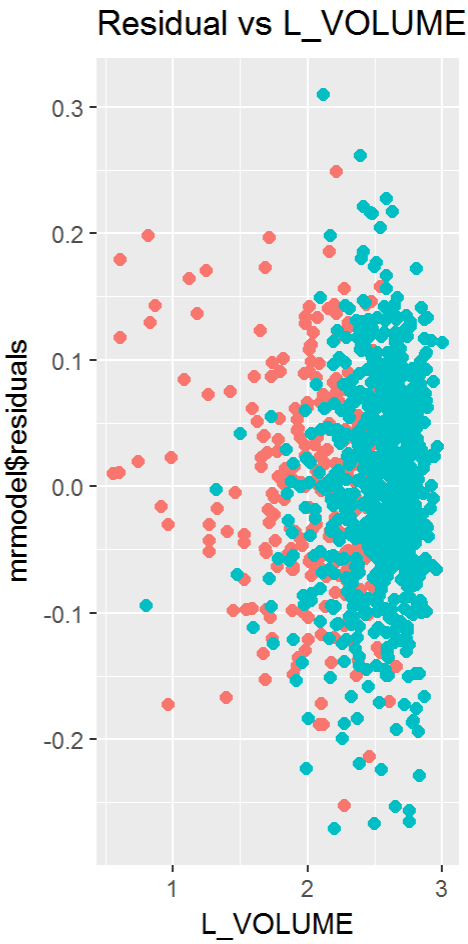



```
## Regression model Residuals skewness & kurtosis using 'rockchalk' package and 'moment' packa
ge.
##
##
```

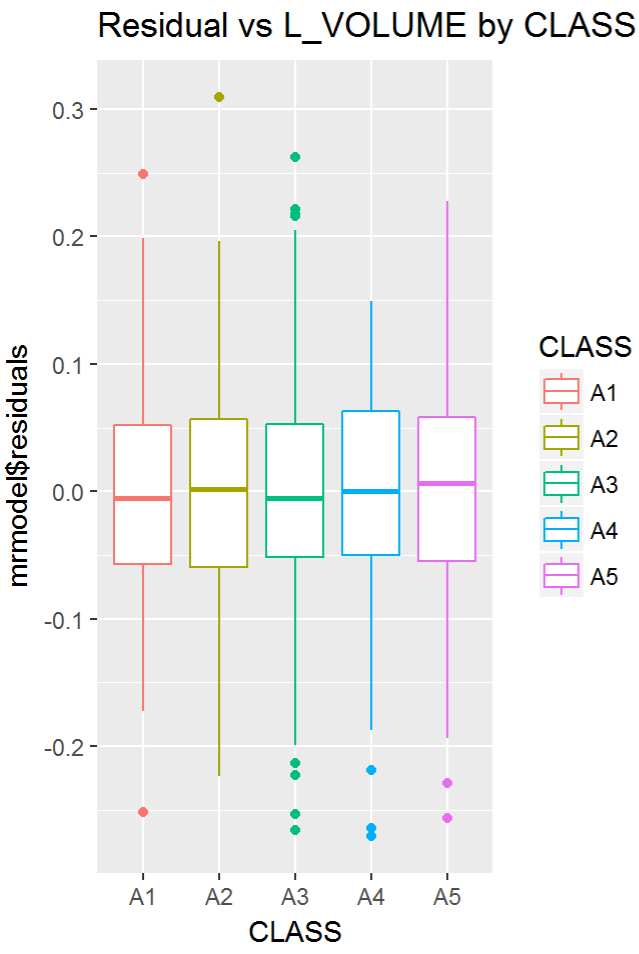
```
##          'rockchalk' package 'moment' package
## skewness         -0.05953853      -0.05953853
## kurtosis          0.34977180       3.34977180
```

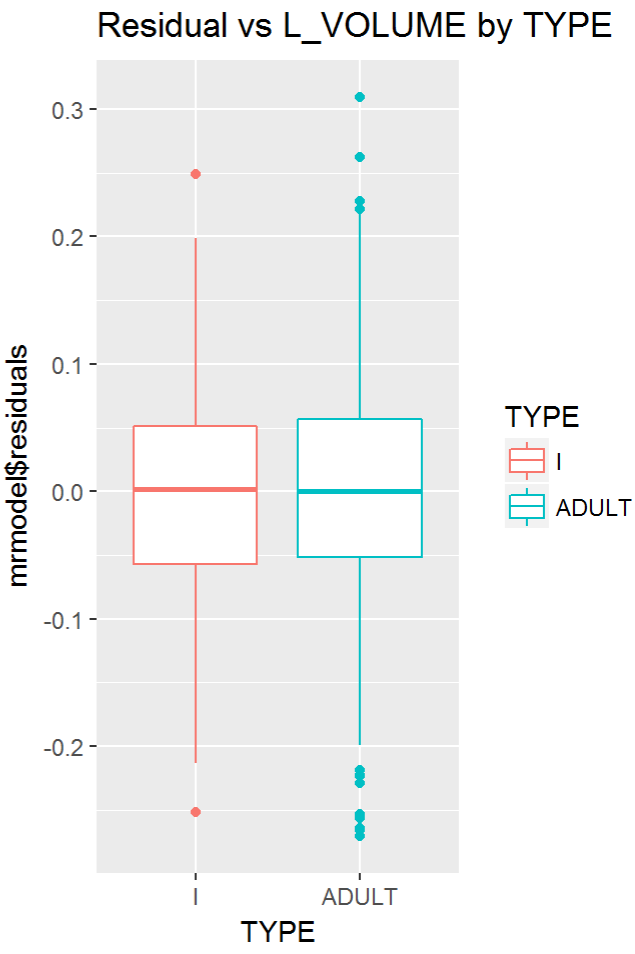
(5)(b) (3 points) Plot the residuals versus L_VOLUME coloring the data points by CLASS, and a second time coloring the data points by TYPE (Keep in mind the y-axis and x-axis may be disproportionate which will amplify the variability in the residuals). Present boxplots of the residuals differentiated 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).





The proportions are way off for these plots. The range for the residuals is much smaller than suggested visually.





```
##
## Test the homogeneity of variance in Residuals across classes using the bartlett.test()
```

```
## Degrees of freedom:  df = 4
```

```
## Significance level:  alpha = 0.05
```

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

```
## Critical value:  Chi-square =  9.487729
```

```
## Test statistic:  T =  3.688189
```

```
## Critical region: Reject H0 if Test statistic T >  9.487729
```

```
##
## Residuals variances are all equal across Classes. Because the test statistic is less than t
he critical value, we pass the null hypotheses at the 0.05 significance level and conclude tha
t variances in each of the classes are the same.No reason to reject null hypothesis.
```

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: Histogram and QQ Plots of the residuals from our regression model, reveals that distribution of the residuals is approximately normal. Skewness and rockchalk excess Kurtosis values are -0.0595 and 0.3498 respectively which is expected as there appears to be some outliers. The Q-Q plot of the residuals where a majority of the points fall on the normal line with some slight skewing at both extremes. This again confirms that the residuals are approximately normal as indicated in residuals histogram. In the scatterplots, there is a noticeable distinction of lower age classes (A1 and A2) dispersed across lower log volume values while higher age classes are dispersed across higher values on the right hand side of the plot. We can see that the residuals of infant abalone are scattered across the left hand side of the plot, while Adult abalone are clustered to the right. From boxplot we can see that the residual mean is almost same across CLASSES and TYPE and majority of points fall around the 0.0 residual value that suggests that CLASS and TYPE has very minute influence on predicting log SHUCK weight and the model seems to fit the data.There are some outliers of residual values, however this is expected as there are some variability between physical measurements and abalones age. Normal distribution of Residuals along with Residual standard error and Multiple R squared value of the model also suggests that the model is a good fit for the further analysis. From this analysis also we can conclude that log VOLUME variable is the most dominate predictor in this model and can be useful for the abalone harvesting decisions.

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.

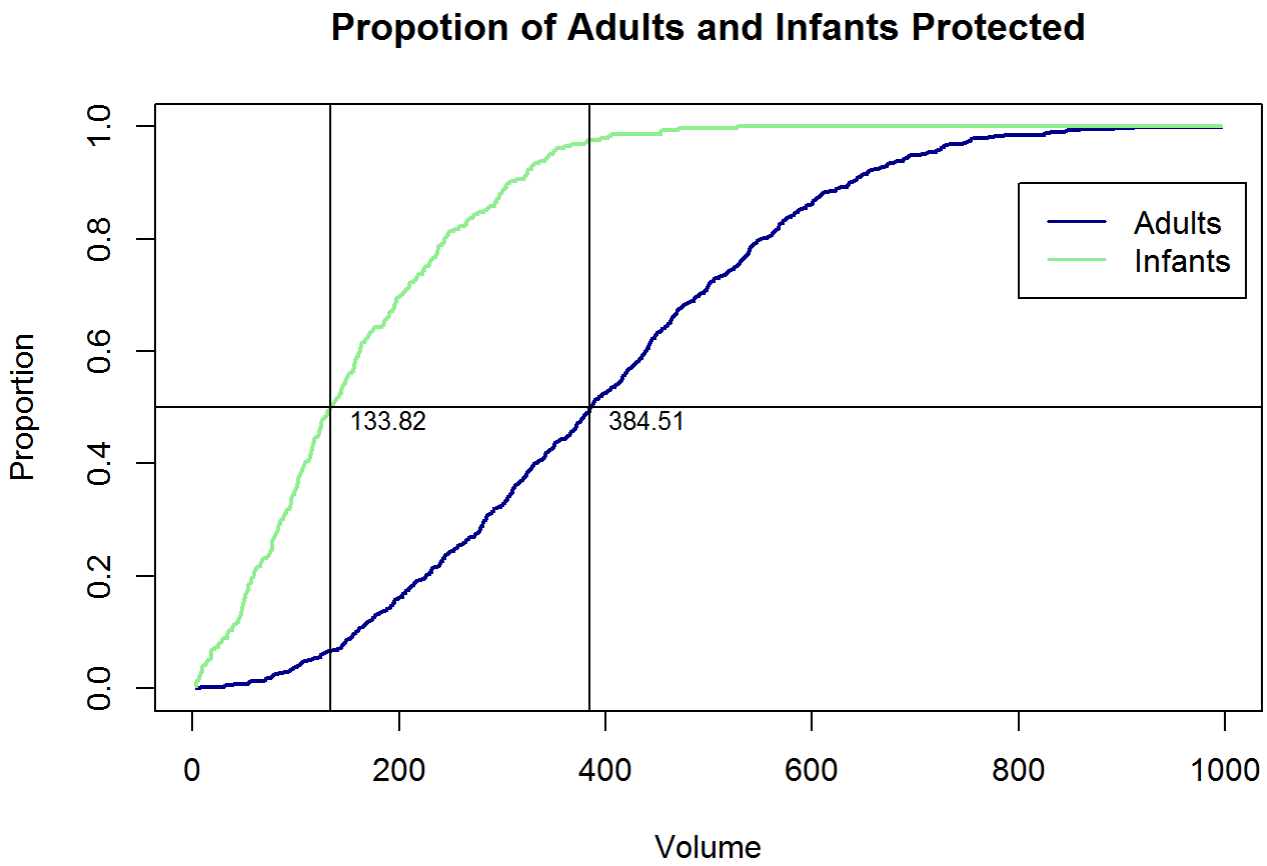
---

(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.

```
## Infant abalones Volume in which <= 50% Infant abalones fall:  133.8199
```

```
##
## Adult abalones Volume in which <= 50% Adult fall:  384.5138
```

(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.



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?

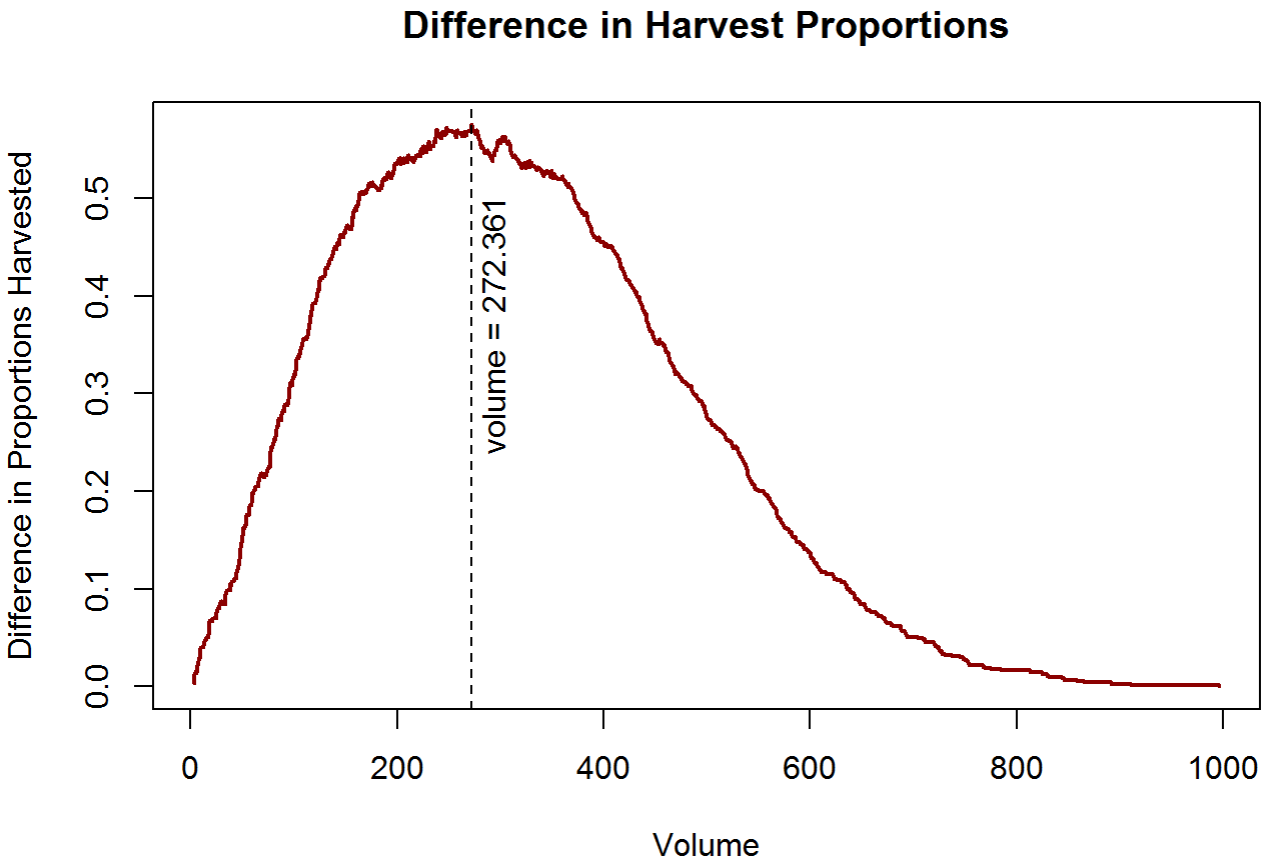
Answer:After the new classification of TYPE into Infant and Adult, Infants are around 38% of Adult abalones population. At 50% split, if we decide to protect 50% of infant abalones, it means that all the abalones with volume > 133.82 will be harvested and in that scenario approximately 93% of adult abalones will get harvested and the total yield will be around 83%. On the other hand if we decide to protect 50% of Adult abalones, which means harvest all abalones with volume > 384.51, only 2 to 3% of the infant abalones will be harvested and yield will be around 37%. Infant abalones population should be considered when deciding on cutoff volume for harvesting. It seems, a cutoff volume between 133.82 and 384.51 would result in a produce a better yield and still protect infant abalones.

-1 point 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.

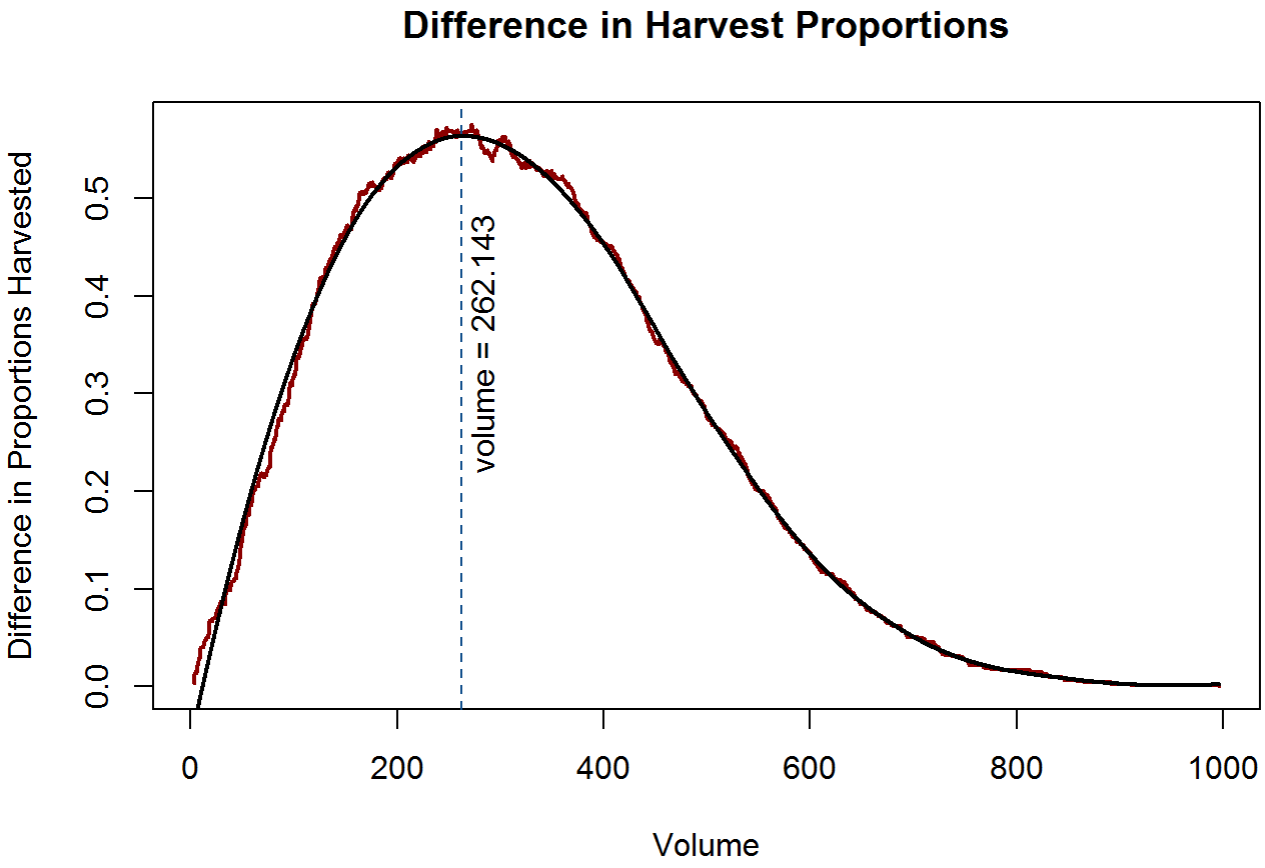
(7)(a) (1 point) Evaluate a plot of the difference ((1 - prop.adults) - (1 - prop.infants)) versus volume.value. Compare to the 50% “split” points determined in (6)(a). There is considerable variability present in the peak area of this plot. The observed “peak” difference may not be the best representation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

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



(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.

(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.



(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.

```
## Max difference cutoff volume:  262.143
```

```
##
## ----------------------------------------
```

```
##
## Adult harvest proportion :  0.742
```

```
##
## Infant harvest proportion:  0.176
```

```
##
## Total proportion :  0.584
```

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.

```
## Zaro A1 Infant cutoff volume:  206.786
```

```
##
## ----------------------------------------
```

```
##
## Adult harvest proportion :  0.826
```

```
##
## Infant harvest proportion:  0.287
```

```
##
## Total proportion :  0.676
```

(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.

```
## Equal harvest cutoff volume:  237.639
```

```
##
## ----------------------------------------
```

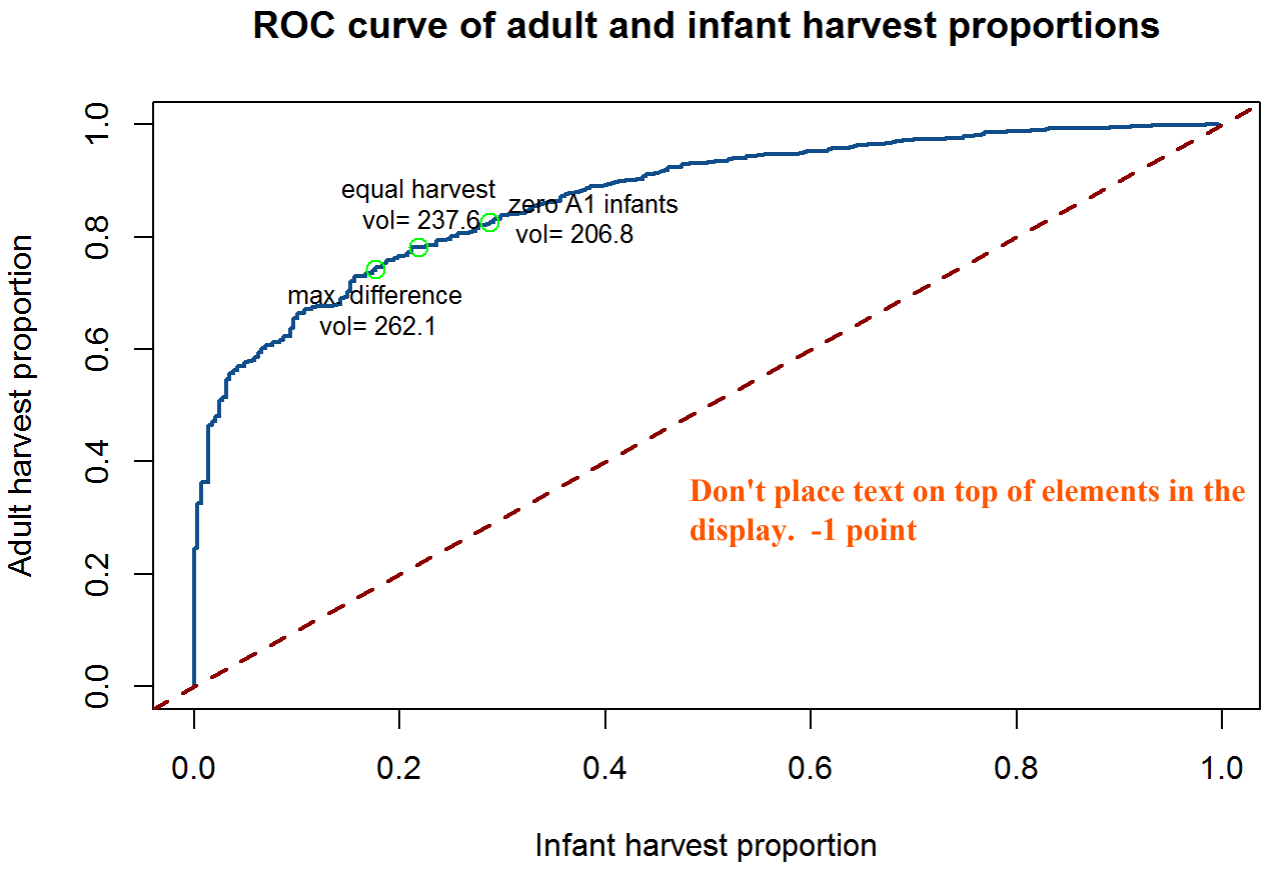
```
##
## Adult harvest proportion :  0.782
```

```
##
## Infant harvest proportion:  0.218
```

```
##
```

```
## Total proportion :  0.625
```

(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.



(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.

```
## Area under the ROC curve:  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

```
##      Cutoffs options  Volume   TPR   FPR PropYield
## 1     max.difference 262.143 0.742 0.176     0.584
## 2 zero.A1.difference 206.786 0.826 0.287     0.676
## 3        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: The main objective is to protect infants, but maximize harvest potential of adults. A ROC plot shows the tradeoff of cutoff volume for harvesting between infant and adult harvest proportions. Based on our cutoff volume decision rule, if infants are harvested, we have a false positive. The true positive corresponds to the adult harvest proportion. Different cutoffs gives an idea of what % of each type abalones (Infant and Adult) will be harvested at different volume cutoffs. Also it give an idea on the total yield. Max difference cutoff protects most Infant abalones compared to other two cutoff but it gives the least total yield. A cutoff volume between Zero A1 Infant abalones cutoff volume and Equal error cutoff volume appears to be a sensible choice here. We may still harvest infants which will result in false positives.

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: Our exploratory data analysis on abalones dataset was based on physical measurement of abalones. The objective was to analyze different physical measurement to find out if they can be used as predictor of abalones age which can help in deciding age cutoff for harvesting abalones to get maximum harvest yield and minimize the harvest of infant abalone. I would prefer giving a range of cutoff volumes and tradeoffs in terms of harvest % of Infant and adult along with total yield when changing the cutoff value in that range so that a well informed decision can be made.

Test should be run on developed model using sample data taken from the abalone dataset and the result should be compared with the original values. If there are lot of variance between test result and original values, it would be a good decision to discuss the analysis with investigators to understand if there are any other factors that may explain the variation. By including those factors in the study we can arrive on an optimal solution for abalone age prediction and harvesting cutoff decision.

To arrive on a harvesting decision rule based on physical measurement is possible if our regression model fits well in predicting abalones age based on the collected physical measurement samples. From our analysis we have found that physical measurements have largely been imprecise in classifying abalone age class except in the lower age classes. Because physical measurements seem to be somewhat reliable in classifying lower age classes, it is logical to give a range of volume from 206.786 to 237.639 for maximizing harvest yield and minimize the harvest of infant abalone.

Based on the exploratory data analysis performed, there are issues with using physical measurements as a way to predict the age of abalone. Physical measurements are useful when predicting age in younger abalone specifically from class A1 to A2, but become less useful for predicting older, adult abalone specifically class A3 to A5. The investigators seemed to have misclassified some infant abalone as adult abalone possibly caused by an issue with ring clarity. When analyzing data from observational studies involving different classes, we must consider whether the data are representative of the population and whether these different classes accurately describe the data. We must also consider how the data were collected. It's extremely important to keep in mind that measurements in observational studies do not indicate causality. It is very difficult to classify the age of abalone using physical measurements. Abalone can vary tremendously in physical measurements especially among adults. It's important to pose questions to the original investigators regarding other factors that could possibly affect the variability in the abalone dataset. Some questions which could be asked: were the samples drawn randomly selected from the population? How were variables determined? Were there any biases that could plague data gathering and analysis? Are there other variables we may not have looked at? ***

-2 points As far as the presentation is concerned, I would keep it visual and simple at the start minimizing analytical details. The origin of the data needs to be explained since this ndetermines limitations pertaining to the results. The results may not pertain to other locations. I would get the investigators thinking about feasible choices and the tradeoffs involved. They are the experts and need to use their judgment at this time. 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. Tracking the health of specific abalone populations subject to different harvesting rules would be a possibility. This analysis is just a first step. More work is needed.

70 points