Data Analysis Assignment 2

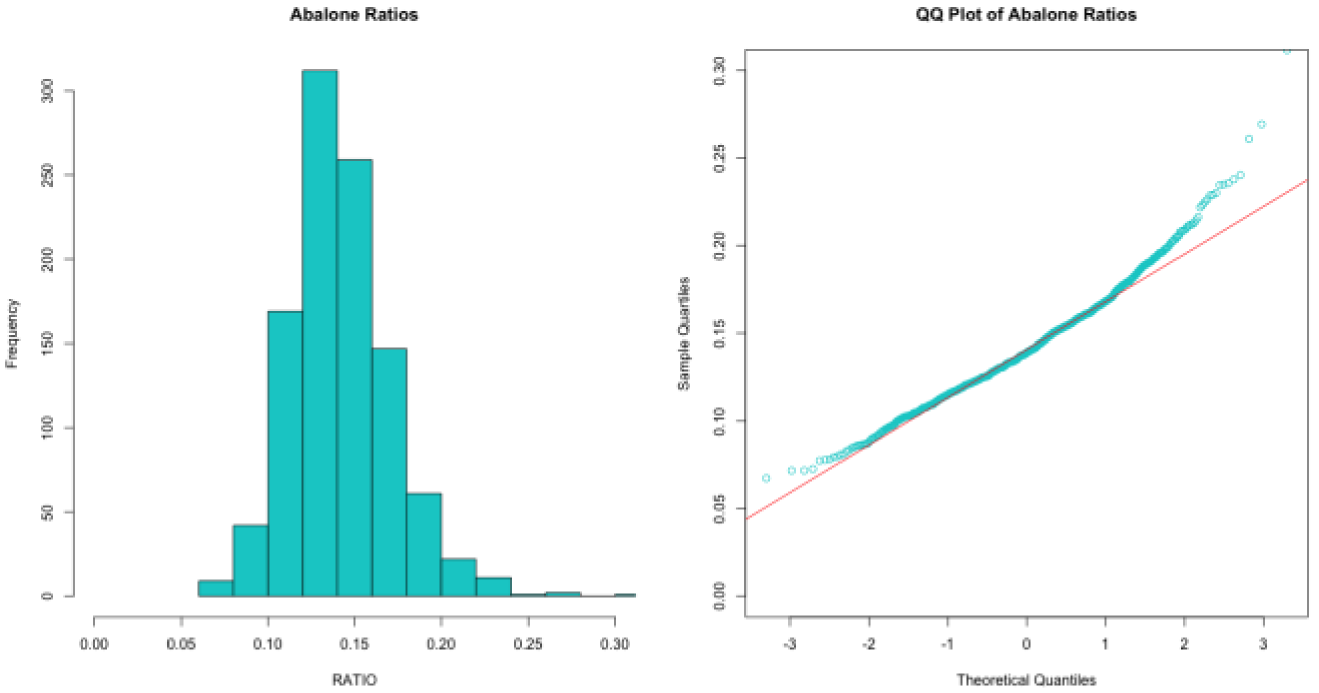
Introduction

The purpose of this assignment is to use analysis of variance, data transformation and simple linear regression techniques to create and test binary decision rules for the harvesting of abalones. Abalone are an important commercial resource. Establishing good binary harvesting rules can help preserve a healthy population while maximizing yield from the harvest.

Results

Our first task is to evaluate the shape of the Abalone data. The histogram for RATIO reveals a tail to the right, and the slight concavity of QQ plot also indicates a right skew.

Figure 1
Histogram and QQ Plot of Abalone RATIO data



Using both the Rockchalk and the Moments packages, I generated Table 1 to compare the skewness and kurtosis values for RATIO from the 2 packages. All remaining skewness and kurtosis calculations in this document are done using Rockchalk. Regardless of the package used, the ultimate interpretation of the returned values is the same. The skewness values of greater than 0, indicate we are dealing with an asymmetric distribution, which has a tail to the right. The kurtosis values of greater than 3.0 for Moment or greater than 0.0 for Rockchalk indicate the distribution is leptokurtic. The Abalone

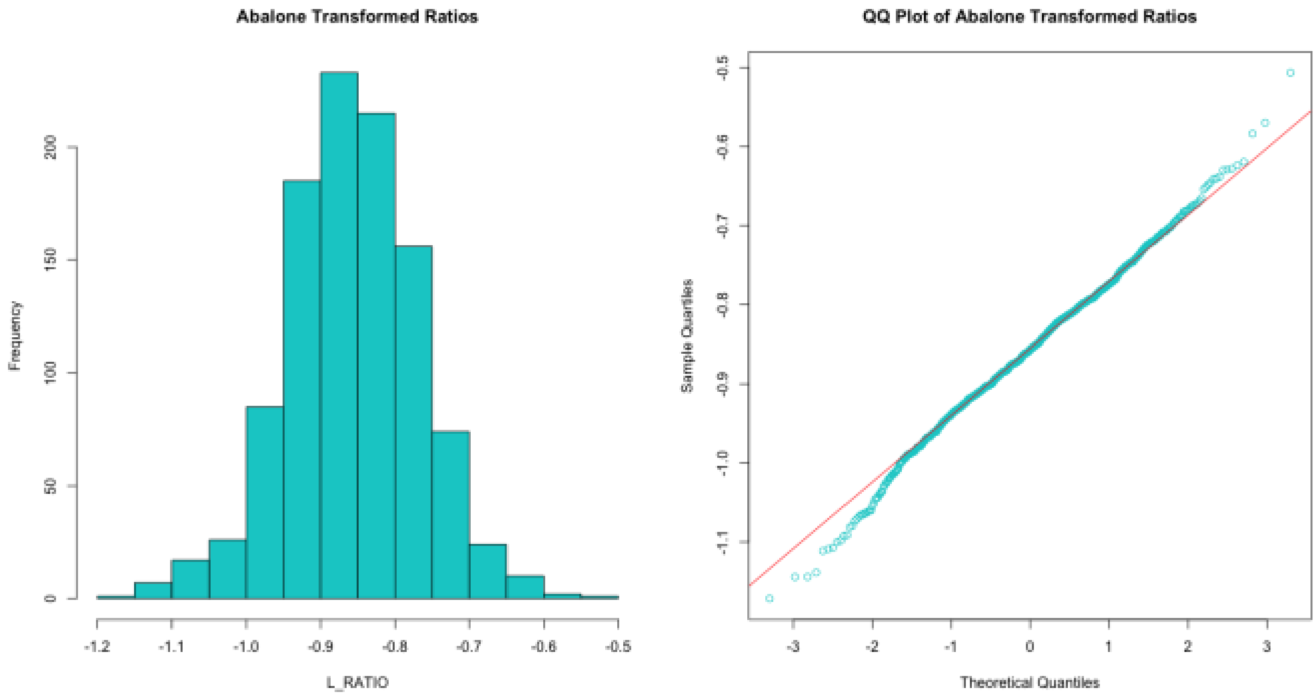
RATIO data does not form a Normal distribution given the differences in symmetry and shape from Normal.

Table 1
Comparision of skewness and kurtosis calcuated with the Moment and Rockchalk* packages

	Rockchalk	Moment	Difference
Skewness	0.7147056	0.7157417	0.0010361
Kurtosis	1.667298	4.676321	3.009023

In order to normalize the RATIO data, the RATIO data was transformed by computing the $\log_{10}$(RATIO) value for each datum and the results were saved to the data frame as L_RATIO. Visualizations of the results of this transformation on the data are presented in Figure 2.

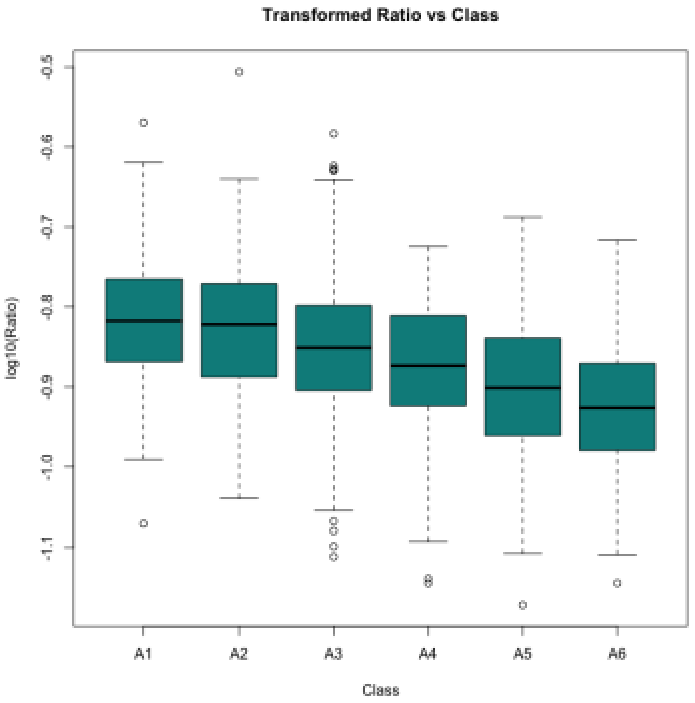
Figure 2
Histogram and QQ Plot of Abalone RATIO data tranformed by taking log10(RATIO)



The Rockchalk value for the skewness of the transformed Ratio data is -0.09391549, the kurtosis is 0.535431. This transformation brings the RATIO data closer to the Normal curve values of skew = 0 and kurtosis = 0, but doesn't completely normalize the data. L_RATIO is still somewhat leptokurtic, but much closer to symmetric.

The boxplots of L_RATIO by when differentiated by CLASS (Figure 3) show no obvious violations of normality. They do show some variation in Log_{10}(RATIO) between each of the classes.

Figure 3
Boxplots of the log10(RATIO) data by CLASS



Testing for the homogeneity of the variance across the classes gives the following result:

Bartlett test of homogeneity of variances:

data: L_RATIO by CLASS
Bartlett's K-squared = 3.0749, df = 5, p-value = 0.6884

By definition, the p-value is the smallest value of alpha for which the null hypothesis can be rejected. In this case, the null hypothesis is that the variance in each of the groups is the same. So, Bartlett's test is saying that .6884 is the smallest alpha for which we can reject the hypothesis that the variance between groups is the same. This is a large p-value, and I would *fail* to reject the null hypothesis with this value. So, putting it all together, I fail to reject the variances are equal (homogeneous), which meets the test for normality. Looking at the boxplots, histogram and QQ Plot for the log10(RATIO) data, it seems reasonable to assume a normal enough distribution with variances across the classes.

Table 2
Analysis of variance of CLASS and SEX with CLASS:SEX interaction term

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CLASS	5	1.076	0.21512	31.313	< 2e-16 ***
SEX	2	0.096	0.04782	6.960	0.000995 ***
CLASS:SEX	10	0.029	0.00290	0.421	0.936789
Residuals	1018	6.994	0.00687		

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 3
Analysis of variance of CLASS and SEX without an interaction term

```
                      Df    Sum Sq    Mean Sq    F value          Pr(>F)
 CLASS                 5     1.076    0.21512     31.490      < 2e-16 ***
 SEX                   2     0.096    0.04782      6.999     0.000957 ***
 Residuals          1028     7.023    0.00683
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In Table 2 the CLASS:SEX values show that there is no significant interaction between SEX and CLASS for the L_RATIO value. Put another way, the relationship between L_RATIO and CLASS does not depend on SEX.

Sum of Squares, Mean Square, and Degrees of Freedom for CLASS and SEX are the same between the two tables, however the F statistics differ slightly. Also, the p-value for SEX in Table 3 is .00038 smaller than that in Table 2, which is small enough to be of no consequence. The small p-values in both tables for CLASS and SEX indicate that for an alpha of .05, or even .01 the null hypothesis would be rejected for the assumption of equal variance.

Focusing on the results in Table 3, a multiple comparison test was conducted, with the following results:

Table 4
TukeyHSD analysis of variation between CLASS catagories

```
 Tukey multiple comparisons of means
   95% family-wise confidence level

 Fit: aov(formula = L_RATIO ~ CLASS + SEX, data = dfAbalone)

      $CLASS
                    diff           lwr             upr        p adj
          A2-A1  -0.01248831  -0.03990346   0.014926837   0.7848170
          A3-A1  -0.03451323  -0.06067382  -0.008352646   0.0024066
          A4-A1  -0.05863763  -0.08713038  -0.030144884   0.0000001
          A5-A1  -0.08685165  -0.12129814  -0.052405154   0.0000000
          A6-A1  -0.11174297  -0.14532240  -0.078163549   0.0000000
          A3-A2  -0.02202492  -0.04214244  -0.001907396   0.0224189
          A4-A2  -0.04614932  -0.06921824  -0.023080398   0.0000002
```

```
A5-A2   -0.07436334   -0.10447811   -0.044248565   0.0000000
A6-A2   -0.09925466   -0.12837366   -0.070135660   0.0000000
A4-A3   -0.02412440   -0.04568735   -0.002561445   0.0180550
A5-A3   -0.05233842   -0.08131574   -0.023361091   0.0000045
A6-A3   -0.07722974   -0.10517079   -0.049288694   0.0000000
A5-A4   -0.02821402   -0.05931298    0.002884949   0.1005227
A6-A4   -0.05310534   -0.08324107   -0.022969608   0.0000085
A6-A5   -0.02489132   -0.06070873    0.010926085   0.3520976
```

The class-to-class comparisons show that A2-A1, A5-A4 and A6-A5 have no significant differences at the 95% level; the other pairs do show significant differences. Generally, comparisons between adjacent classes, like A1-A2 have higher p-values (closer variances) than comparisons between non-adjacent classes, like A1 – A3; with widely separated classes like A5 – A2 having p-values at or near 0, indicating rejection of the null hypothesis of equal variances.

Intuitively, these results make sense. CLASS is an ordinal-level variable, and you would anticipate greater differences between classes that are further apart. The boxplots in Figure 3 show that the CLASS 1 box is visually a subset of the CLASS 2 box; CLASS 1 is completely overlapped by CLASS 2. The corresponding p-value, .78 says that one would *fail* to reject the null hypothesis that CLASS 1 and CLASS 2 have the same $\log_{10}$(L_RATIO) mean. The mean for L_RATIO of CLASS == A1 is -0.8154 and that for CLASS == A2 is -0.8279; the means differ by only .0125, which is smaller than the critical value for 95% confidence.

Table 5
TukeyHSD analysis of variation between SEX catagories

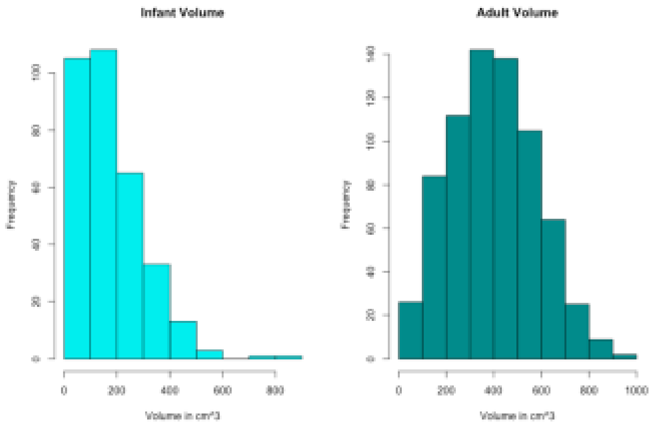
```
 Tukey multiple comparisons of means
   95% family-wise confidence level

 Fit: aov(formula = L_RATIO ~ CLASS + SEX, data = dfAbalone)


$SEX
              diff          lwr           upr         p adj
I-F   -0.016277335  -0.031437534  -0.001117136   0.0318479
M-F    0.002062021  -0.012574216   0.016698257   0.9415134
M-I    0.018339356   0.003739124   0.032939587   0.0091596
```

Comparing SEX, we can see that there are significant differences between Infants and Males, and between Infants and Females. There are not significant differences between Males and Females. This confirms the validity of creating a single “Adult” class which is a combination of Males and Females.

Figure 4
Histograms of Infant VOLUME and Adult VOLUME measures



Looking at the distribution of VOLUME following an Infant/Adult split, we see Infants' VOLUME skews strongly right. It makes sense that younger Abalone would be in general, smaller than adults, and that the volume of these individuals would cluster at the lower end of the scale. Looking at skew and kurtosis we see:

Table 6
Skew and kurtosis tests of Infant and Adult VOLUME

	Infant	Adult
Skewness	1.165	.229
Kurtosis	2.285	-.441

In making this separation, the Infant population has a VOLUME distribution which clearly violates the assumption of normality. The ADULT distribution is closer to Normal, but it is not fully normal either.

Unlike RATIO, VOLUME is a measure where using the $\log_{10}$ transformation is not helpful in trying to get to a more Normal distribution. As shown in Figure 5 and Table 7, that transformation is not helpful.

Figure 5
Histograms of Infant and Adult VOLUMES transformed using log10

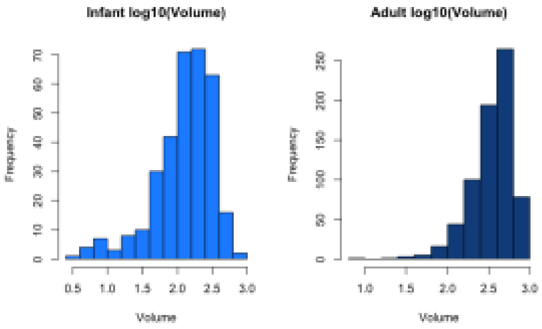
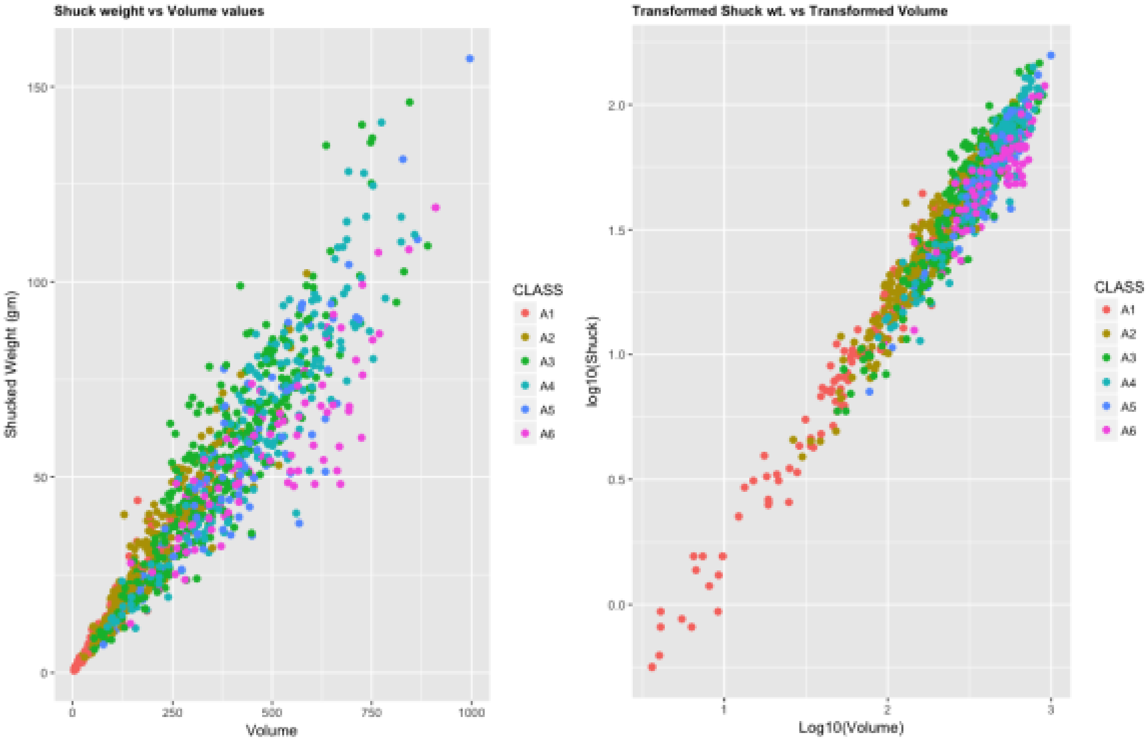


Table 7
Skew and kurtosis tests of $\log_{10}$(VOLUME)

	Infant	Adult
Skewness	-1.276953	-1.500244
Kurtosis	2.005294	4.320206

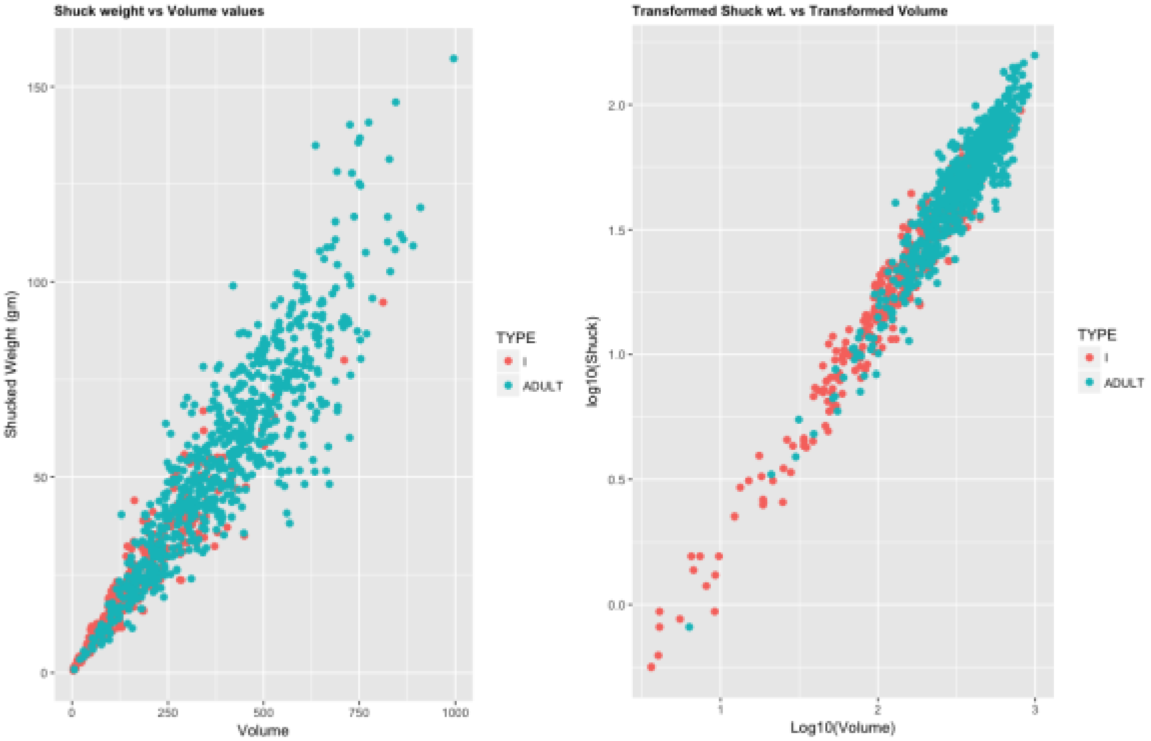
Figure 6
Scatterplots of SHUCK weight vs VOLUME and SHUCK weight vs log10(VOLUME) by CLASS



In the Volume versus Shuck chart, the A1 points are clustered toward the lower left, near the origin. The other classes are dispersed throughout the chart; with overlap between classes in all areas of the chart. The over-all shape of the point cloud flairs out in a way that is indicative of heteroscedasticity.

In the chart using the transformed, $\log_{10}$(VOLUME) data, there is more of a visual separation of points by class. The majority of A1 and some A2 points are to the left of the vertical “2” line, while the older A3 – A6 specimens are, for the most part, are to the right of the “2” line. If one were to select “2.5” on this chart as an arbitrary dividing line for harvesting, harvesting individuals to the right of the line, it appears that all the A1 and A2s would be preserves, as well as some members of all classes. The overall shape of this point cloud is basically linear, indicating homoscedasticity. I find the density of points in the upper right quadrant of the chart interesting, but I can't relate that pattern to anything we covered in class.

Figure 7
Scatterplots of SHUCK weight vs VOLUME and $\log_{10}$(SHUCK) vs $\log_{10}$ (VOLUME)


Shuck weight vs Volume values

Transformed Shuck wt. vs Transformed Volume

In Figure 7, the chart of the original data shows the majority of Infants clustered to the lower left of the chart, but there are a few specimens scattered across the entire band.

In the chart using transformed $\log_{10}$(VOLUME) data, we can see more of a separation between the Infants and the Adults, although the separation is not exact. A hypothetical harvest cut-off line (harvesting to the right) through the 2.5 vertical would preserve the majority (but not all) of Infants and a selection of Adults. If the theoretical harvest line was moved to the 2 vertical, a number of Infants could be harvested as well as the majority of Adults. Since a goal is to maximize the grams of Abalone meat, harvesting from Adult specimens, setting limits that are to the right of the log10(VOLUME)==2.5 line, should be highly effective. That VOLUME cutoff is associated with the higher SHUCK values, and fewest Infants collected.

Table 8
Summary of lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = dfAbalone)

```
Call:
lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = dfAbalone)

Residuals:
 Min          1Q          Median       3Q          Max
-0.274844    -0.054213    -0.001639    0.055975    0.306985

Coefficients:
                Estimate      Std. Error     t value     Pr(>|t|)
 (Intercept)    -0.812384      0.019103      -42.528     < 2e-16 ***
 L_VOLUME       0.995930       0.010315       96.554     < 2e-16 ***
 CLASSA2        0.017359       0.010942       -1.587     0.112927
 CLASSA3        0.047442       0.012266       -3.868     0.000117 ***
 CLASSA4        0.073368       0.013588       -5.399     8.30e-08 ***
 CLASSA5        0.101482       0.015019       -6.757     2.36e-11 ***
 CLASSA6        0.127006       0.015060       -8.433     < 2e-16 ***
 TYPEADULT      0.025179       0.006818        3.693     0.000233 ***
---

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08265 on 1028 degrees of freedom
Multiple R-squared:  0.9508,  Adjusted R-squared:  0.9505
F-statistic:  2841 on 7 and 1028 DF,  p-value: < 2.2e-16
```

CLASS A2 is the only one not having a significant result. In Figure 6, we can see that the distribution of A2 points for L_SHUCK versus L_VOLUME spreads across nearly all the other classes. It is poorly differentiated by shuck and volume values. Table 8 shows the TukeyHSD results for L_RATIO ~ CLASS. Ratio is defined as shuck/volume. Table 8 also shows that class A1 fails to reject the null hypothesis of equal variances for A1-A2, and depending on your critical value, possibly A2-A3.

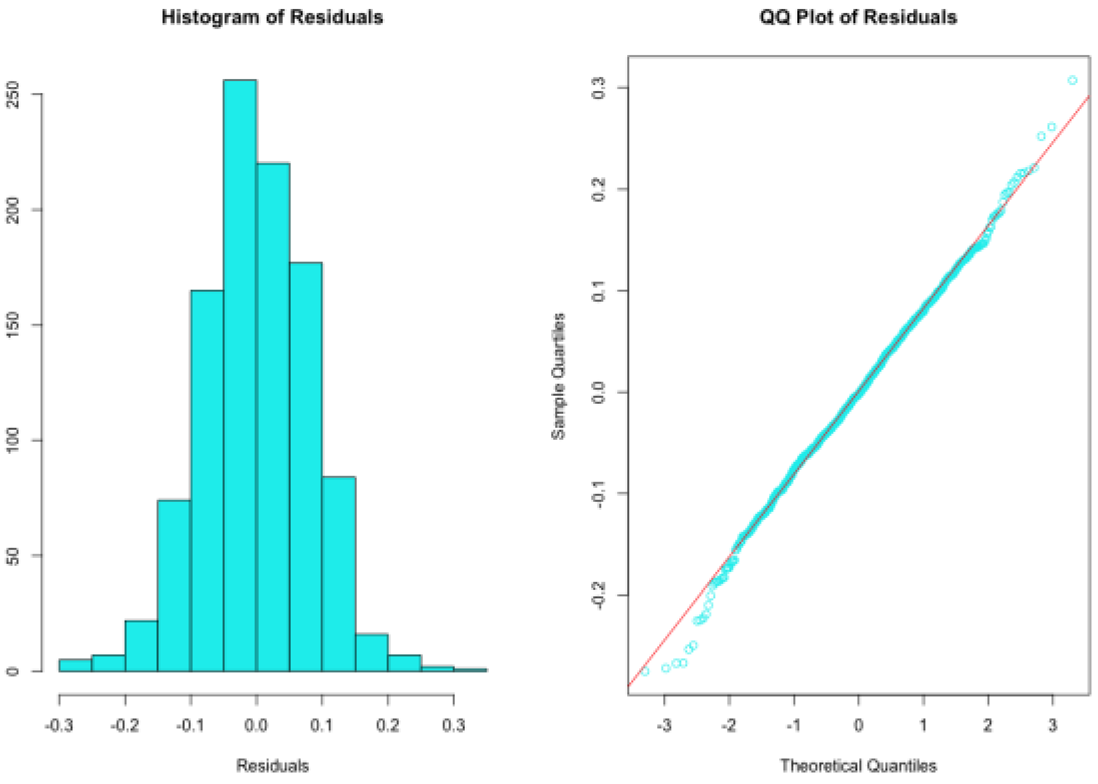
Per Kabacoff[1], *"when there is more than one predictor variable, the regression coefficient indicates the increase in the dependent variable for a unit change in a predictor variable, holding all other predictor variables constant"*. In Table 8, this would mean the .127 regression coefficient for CLASSA6 can be interpreted as log$_{10}$ (SHUCK) increases 12.7% for every 1% increase in CLASSA6, when controlling for the other predictors. Similarly, the other 4 CLASS variables listed in Table 8 show increases, indicating that each, when other predictor variables are controlled, cause an increase in log$_{10}$ (SHUCK).

[1] Kabaoff, R., "*R In Action, Data Analysis and Graphics with R*", 2015, Manning Publications Co., pg 190

In the case of TYPEADULT, the regression coefficient indicates that $\log_{10}$ (SHUCK) increases 2.51% for each 1% of TYPEADULT. While not as strong a predictor as some of the CLASS variables, like CLASS A5 or A6, it is a predictor.

The Multiple R-Squared value of .9508 indicated that 95.08% of the variation in L_SHUCK is accounted for by the independent variables L_VOLUME, CLASS and TYPE.

Figure 8
Histogram and QQ Plot of the Residuals for linear model associated with Table 8



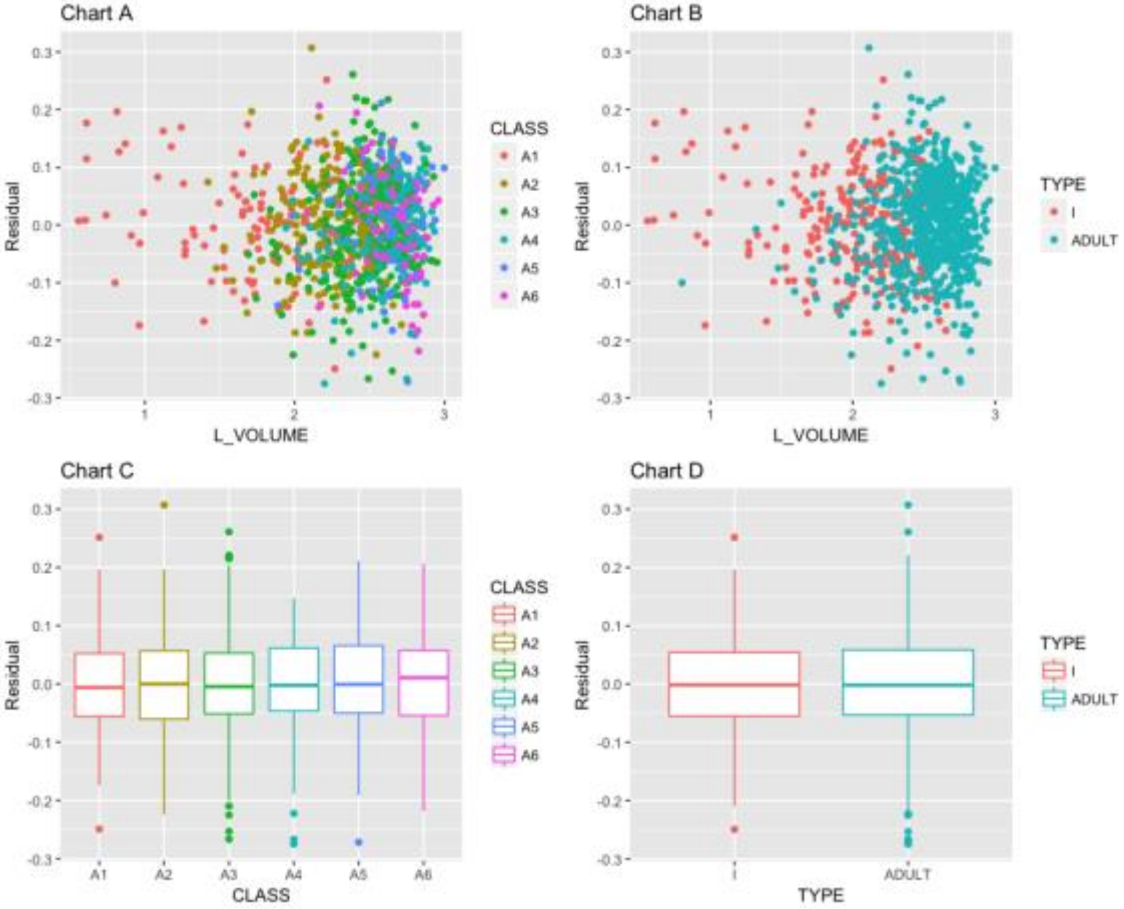
The calculation for skewness of the residuals, using rockchalk is:
[1] -0.06942973

The negative, non-zero value indicates an asymmetrical distribution, with a tail to the left.

The kurtosis calculation:
[1] 0.3615913

The value is greater than 0, and indicative of a distribution that is more peaked than a Normal distribution, which is to say, leptokurtic.

Figure 9
Residuals vs different factors, examined by boxplots and scatterplots

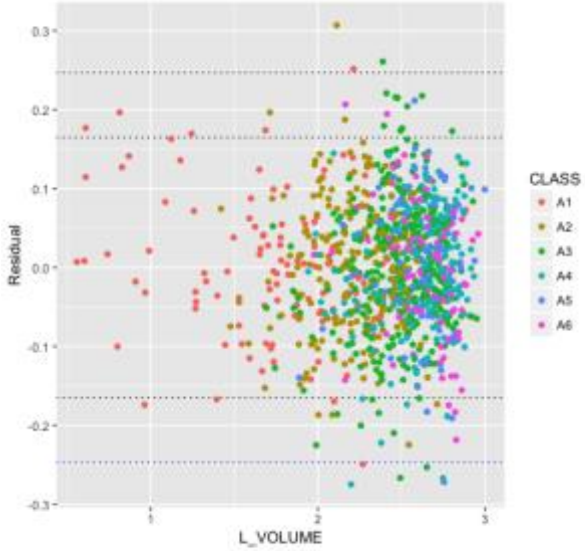


```
        Bartlett test of homogeneity of variances
data:  linear_model$residuals by CLASS
Bartlett's K-squared = 3.6657, df = 5, p-value = 0.5985
```

To have a good fit, you want the residuals to have a normal distribution, with no pattern when compared to the independent variables. Additionally, you would like to have the points lie within 2 standard deviations of the mean of 0. All 4 charts in Figure 9 are centered roughly around y = 0 line, which is necessary to have a good fit. However, Charts A and B show a much higher density of points toward the right side of the chart. That density on one side of the chart, while remaining symmetric about the y=0 line is not something seen in our text’s sample charts, I am unclear how to interpret the distribution. The overall distribution sort of has the appearance of the “flair” shape that is indicative of heteroscedasticity, which means that the assumption of homoscedasticity is questionable for this data.

Looking at the standard deviations for the plot, show by the dotted lines in Figure 10, it is clear there are a lot of points outside the $\pm 2\sigma$ zone, and even some beyond $\pm 3\sigma$. However, the preponderance of points lie within the desired range.

Figure 10



Given the p-value of the Bartlett test, I would *fail* to reject the null hypothesis the variances are equal. The mean of the residuals by CLASS and by TYPE are hovering near 0, which is indicative of a normal distribution. Based on these charts and the Bartlett test, I say the regression model fits the data well.

Sample of results from the calculations of cutoff to verify correctness of proportion calculations, item (6)(a) from the assignment brief, are shown in Table 10 in the Appendix.

Figure 11
Proportion of Infants and Adults protected, showing a split at the 50% proportion level

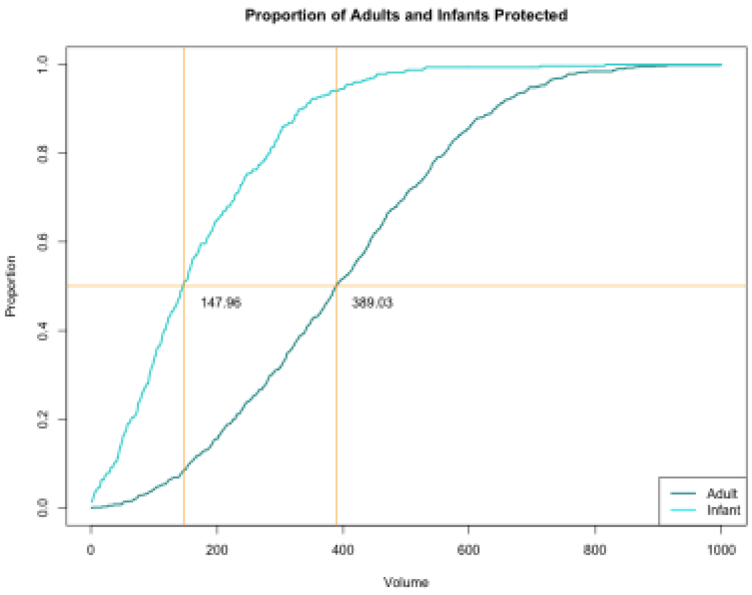
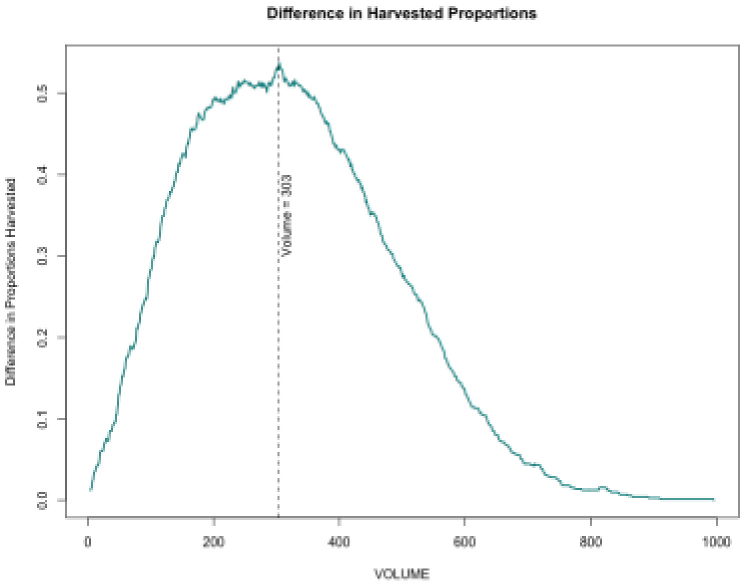
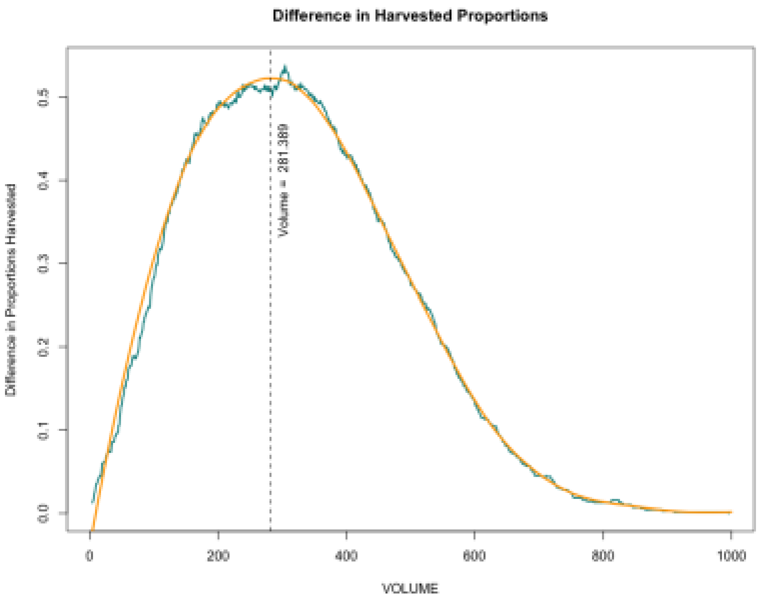


Figure 12
Plot of the difference in porportions harvested vs volume.value



Plotting the difference in proportions reveals an observed maximum volume at 303. This differs from the 50% split values of volumes of 147.96 for Infants and 389.03 for Adults shown in Figure 11. With the addition of curve smoothing, the volume maximum changes to 281.389 as seen in Figure 13.

Figure 13
Plot of the difference in porportions harvested vs volume.value with smoothing line



Adult Harvest proportion (TPR) for using the maximum difference as the cut-off is:
[1] 0.7114569

Infant Harvest proportion (FPR) for using the maximum difference as the cut-off is:
[1] 0.2036474

Adult Harvest proportion (TPR) when minimizing harvesting of A1 individuals
[1] 0.8302687

Infant Harvest proportion (FPR) when minimizing harvesting of A1 individuals
[1] 0.3404255

Adult Harvest proportion (TPR) when harvesting equal proportions of Adults and Infants
[1] 0.7553041

Infant Harvest proportion (FPR) when harvesting equal proportions of Adults and Infants
[1] 0.2431611

Figure 14
Reciever Operating Characteristic graph of harvest, showing potential harvest cutoff volume values

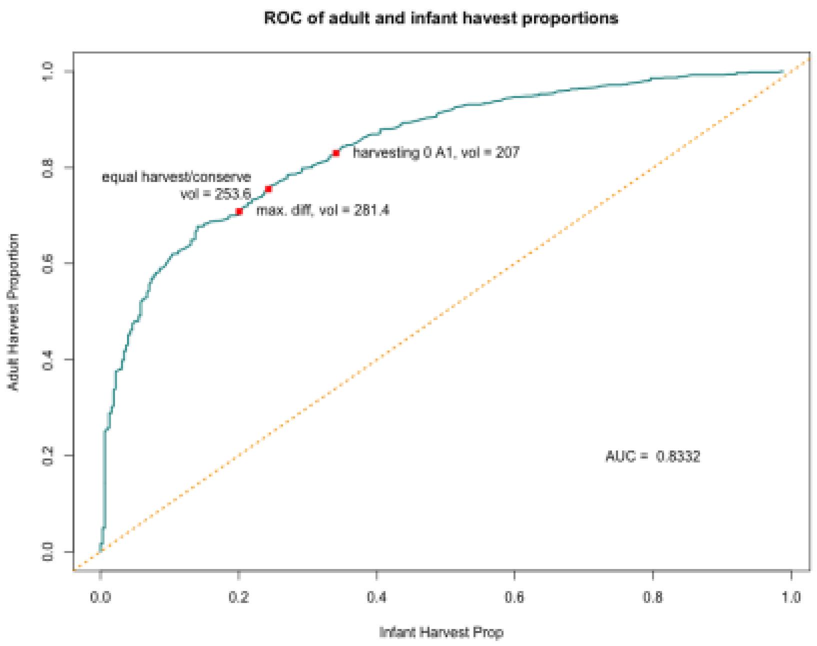


Figure 14 shows the Reciever Operating Characteristic graph for Adult harvest proportion versus Infant harvest proportion. Potential harvest cutoff volume values are shown, with their "strategy" label. The total area under the curve is .8332. This chart does have good discremination potential; the 3 separate proposed cutoffs are clearly seen and are well separated along the line of the chart.

Table 9
Comparison of potenital harvest cutoff values

Harvest	Volume	TPR	FPR	PropYield
zero harvest	206.984	0.83	0.34	0.676
equal harvest	253.611	0.755	0.243	0.594
max difference	281.389	0.711	0.204	0.55

I see competing factors in each choice of harvest strategy. In reviewing Table 9, it might appear that the "Zero Harvest A1" should be the preferred strategy. PropYield is defined to be the sum of all the volumes above the threshold, divided by the number of individuals and the "zero harvest" strategy maximizes PropYield. However, looking at the ROC graph, one can see that strategy allows for taking the largest total number of individuals, including the largest number of infants. It also has the highest False-Positive rate among the 3 strategies. The maximum difference strategy actually seems to best restrict the number of Infants taken, but has the lowest PropYield. Equal harvest seems to be the compromise option.

To make a choice between the competing strategies, I would need more information on what is needed to maintain a healthy, sustainable fishery. Without knowing more exact parameters for how many Infants and Adults are needed to sustain a breeding eco-system, I cannot tell how many to harvest. It is also unclear if there are additional factors outside the scope of this data which have a meaningful bearing on the decision. Do El Niño or La Niña weather patterns effect breeding, infant survivability, growth? Does location play a role in growth, maturity, and viability of a sustainable population? This would need to be determined before a metric could be set to act as an authoritative cutoff for harvesting.

Conclusions

In order to present material like this, I find putting together a PowerPoint or other narrative style presentation to be an effective communication device. I'd start by setting context with an overview of the data collection methods used, high-level view of data set using the table of summary statistics we generated in EDA1 and perhaps the Volume versus Class boxplot from EDA1 since we are going to be using volume in our cutoff discussions.

I would structure the presentation as a step-by-step walk thru of the salient points of the work done, methods used, and results discovered. I would include all the charts and tables from this paper *except* Figure 5 and Table 7. The 2 excluded items are exploratory work on my part that represent a dead end, and are not appropriate to a presentation.

I would note the following consideration: volume is not normally distributed for these Abalone. One of the steps we took in this analysis was the separation of the Abalone population into Infants and Adults. The Adult Volume distribution is close enough to Normal that the tools assuming normalcy *should* be valid; the Infant data distribution is decided not normal. I believe the deviation from normal will not render the results invalid, but since we are using volume in setting/modeling/testing harvest thresholds is should be mentioned.

I also found the clustering of residuals on the right side of Figure 9, to be worth noting. I'm not sure how to interpret that behavior. It does not look like a "funnel" shape to me, so I am ruling it out as an indicator of heteroscedasticity. Since it is a deviation from a classic normal residual plot, I'd want to call it out in any presentation. It is probably an indicator of something worth delving into. Also, the fact some points are outside $\pm 2\sigma$ and even $\pm 3\sigma$ zones as seen in Figure 10, should be recognized. We know there are outliers in the data, and seeing some outliers in the residual scatterplots, while interesting, is probably typical, but still worth acknowledging.

I am not an expert in Fishery Management, and would therefore not make specific recommendations. I feel there are other factors which impact management of the resource and I would not want an assumption I made unknowingly to color any rules put in place. I would outline the tradeoffs between the proportions of Adults and Infants collected as shown by the ROC curve in Figure 14. The experts in resource management can weigh in on what is the best strategy for a maintaining a sustainable fishery.

I do have questions about the consistency of Abalone growth and breeding patterns. If there are large variations in the amount individual grow, numbers of offspring produced, and numbers that survive to maturity, based on weather patterns, or locale, those factors could have long-term impacts on the success of using this data to set harvest cutoffs. For example, if these data were collected during typical weather years and next year is a drought year, what is the impact of that on the growth and breeding of the Abalone? Do fewer individuals survive to maturity? Are fewer infants spawned? This could impact management of the fishery and alter where the cutoff should be set.

I suggest that the threshold be implemented along with a multi-year monitoring study. I suggest starting with a conservative cutoff, perhaps lower than what you want to target as the long-term, permanent threshold. Monitor the health of the population and over time, adjust the threshold upward if the data support the position more abalone can be taken without damage to the resource. It is easy to adjust the harvest to take more individuals, but much harder to fix damage from over-collection.

I recommend collecting several seasons worth of data before adjusting the threshold upward. It is important to ensure a single year does not skew the results of the success of the threshold. In the event the new data suggest there is damage to the population from the threshold collection rules, have a plan prepared for dealing with the contingency before setting the threshold in place.

As we see more extreme weather, I think an investigation into the relationship between the abalone population and temperature and rainfall is likely worthwhile. Does temperature impact the number of larvae produced? Does it impact survival? Does drought impact production or survival of the young?

It might also be worth considering a flexible targeting strategy for threshold. Rather than having a blanket threshold, set a target threshold at the start of each year based on the state of the population at that point. Additional studies would need to be done to see if these annual constraints were as effective, more effective, or less effective than a blanket threshold at preserving the health of the population.

Appendix

R code to produce results and graphics in this paper

```
# Code for Data Analysis Project 2
# NWU Predict 401, Sec 55, Sp2017
# Written by - Tamara Williams

#Clear Workspace
rm(list=ls())
# Clear Console:
cat("\014")

setwd('~/NorthwesternU_MSPA/Statistics_Predict_401/Data Analysis Assignments/Assignment2')
library(stats)

dfAbalone <- read.csv("myAbaloneData.csv", header = T, sep = " ", stringsAsFactors = T)


# (1)(a) Form a histogram and QQ plots using RATIO. Calculate the skewness and kurtosis
# (be aware with rockchalk the kurtosis value has 3.0 subtracted from it which differs from the
# moments package.).
library(rockchalk)
library(moment)

par(mfrow = c(1,2))
hist(dfAbalone$RATIO, main = 'Abalone Ratios', xlab = 'RATIO', ylab = 'Frequency',
    xlim = c(0.0, .30),  col = 'cyan3')
qqnorm(dfAbalone$RATIO, main = 'QQ Plot of Abalone Ratios', xlab = 'Theoretical Quantiles',
      ylab = 'Sample Quartiles', col = 'cyan3', ylim = c(0.0, .30))
qqline(dfAbalone$RATIO,datax = FALSE, distribution = qnorm,  col = 'red')
par(mfrow = c(1, 1)) ## reset 'mfrow' to default value

rc_skew <- rockchalk::skewness(dfAbalone$RATIO, na.rm = TRUE, unbiased = TRUE)
m_skew <- moments::skewness(dfAbalone$RATIO)
rc_skew
m_skew
rc_kurtosis<-rockchalk::kurtosis(dfAbalone$RATIO, na.rm = TRUE, unbiased = TRUE)
m_kurtosis <- moments::kurtosis(dfAbalone$RATIO)
rc_kurtosis
m_kurtosis

detach('package:moment', unload = TRUE)
```

```
# (1)(b) Transform RATIO using log10() to create L_RATIO
dfAbalone$L_RATIO <- log10(dfAbalone$RATIO)

# Form a histogram and QQ plots using L_RATIO.
par(mfrow = c(1, 2))
hist(dfAbalone$L_RATIO, main = 'Abalone Transformed Ratios', xlab = 'L_RATIO',
    ylab = 'Frequency', col = 'cyan3')
qqnorm(dfAbalone$L_RATIO, main = 'QQ Plot of Abalone Transformed Ratios',
      xlab = 'Theoretical Quantiles', ylab = 'Sample Quartiles', col = 'cyan3')
qqline(dfAbalone$L_RATIO,datax = FALSE, distribution = qnorm,  col = 'red')
par(mfrow = c(1, 1)) ## reset 'mfrow' to default value

# checking skew and kurtosis.  Skew = 0 for Normal dist. and kurtosis = 0 (in Rockchalk)
# for a Normal dist.
rcl_skew <- rockchalk::skewness(dfAbalone$L_RATIO, na.rm = TRUE, unbiased = TRUE)
rcl_skew
rcl_kurtois<-rockchalk::kurtosis(dfAbalone$L_RATIO, na.rm = TRUE, unbiased = TRUE)
rcl_kurtois

# Display boxplots of L_RATIO differentiated by CLASS.
boxplot(dfAbalone$L_RATIO~dfAbalone$CLASS, xlab = 'Class', ylab = 'log10(Ratio)',
       col = "cyan4", main = 'Transformed Ratio vs Class')

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

# bartlett.test() tests null hypothesis of homogeneity of variance of a numeric
#variable across two (2) or more groups or levels of a factor.
# bartlett.test(numeric ~ factor, data = ...)
# OR, bartlett.test(x = numeric, g = factor, data = ...)

bartlett.test(L_RATIO~CLASS, data=dfAbalone)

# (2)(a) Perform an analysis of variance with aov() on L_RATIO using CLASS and SEX as
# the independent variables
# Assume equal variances. Perform two analyses. First, use the model *with* an interaction
# term CLASS:SEX, and then a model *without* the interaction term CLASS:SEX.
# Use summary() to obtain the analysis of variance table.

lr_anova_terms <- aov(L_RATIO ~CLASS*SEX, data = dfAbalone)
summary(lr_anova_terms)
lr_anova_noterms <- aov(L_RATIO ~CLASS+SEX, data = dfAbalone)
```

```
summary(lr_anova_noterms)
#export it to wordfile, to assisting in report writing
capture.output(summary(lr_anova_terms),file="2a_lrAnovaTerms.doc")
capture.output(summary(lr_anova_noterms),file="2a_lrAnovaNoTerms.doc")

#2)(b) For the model without CLASS:SEX, obtain multiple comparisons with the TukeyHSD() function.

TukeyHSD(lr_anova_noterms)
capture.output(TukeyHSD(lr_anova_noterms),file="Tukey_lrAnovaNoTerms.doc")

# (3)(a) Use combineLevels() from the rockchalk package to combine "M" and "F" into a
# new level "ADULT".
# Use par() to form two histograms using VOLUME.
# One would display infant volumes, and the other ADULT volumes.

dfAbalone$TYPE <- combineLevels(dfAbalone$SEX, levs = c("M","F"), "ADULT")

par(mfrow = c(1, 2))
dfAdult <- subset(dfAbalone,dfAbalone$TYPE =='ADULT')
dfInfant <- subset(dfAbalone,dfAbalone$TYPE =='I')
hist(dfInfant$VOLUME, main = 'Infant Volume',  xlab = 'Volume in cm^3', ylab = 'Frequency',
    col = 'cyan2')
hist(dfAdult$VOLUME, main = 'Adult Volume', xlab = 'Volume in cm^3', ylab = 'Frequency',
    col = 'cyan4')
par(mfrow = c(1, 1))

# (3)(b) Form a scatterplot of SHUCK versus VOLUME and a scatterplot of their base ten logarithms,
# labeling the variables as L_SHUCK and the latter as L_VOLUME.
# 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.

# define the base ten logarithm vectors
dfAbalone$L_SHUCK <- log10(dfAbalone$SHUCK)
dfAbalone$L_VOLUME <- log10(dfAbalone$VOLUME)
require(ggplot2)
require(gridExtra)

plot1 <- ggplot(dfAbalone, aes(x=dfAbalone$VOLUME, y=dfAbalone$SHUCK, color=CLASS)) +
  geom_point()+geom_point(size=2)+ggtitle('Shuck weight vs Volume values')+
  labs(x= "Volume", y="Shucked Weight (gm)")+
  theme(plot.title = element_text(size = 10, face = "bold"))
```

```
plot2 <- ggplot(dfAbalone, aes(x=dfAbalone$L_VOLUME, y=dfAbalone$L_SHUCK, color=CLASS)) +
  geom_point()+geom_point(size=2)+ggtitle('Transformed Shuck wt. vs Transformed Volume')+
  labs(x= "Log10(Volume)", y="log10(Shuck)")+
  theme(plot.title = element_text(size = 10, face = "bold"))
grid.arrange(plot1, plot2, nrow=1, ncol=2)

plot3 <- ggplot(dfAbalone, aes(x=dfAbalone$VOLUME, y=dfAbalone$SHUCK, color=TYPE)) +
  geom_point()+geom_point(size=2)+ggtitle('Shuck weight vs Volume values')+
  labs(x= "Volume", y="Shucked Weight (gm)")+
  theme(plot.title = element_text(size = 10, face = "bold"))

plot4 <- ggplot(dfAbalone, aes(x=dfAbalone$L_VOLUME, y=dfAbalone$L_SHUCK, color=TYPE)) +
  geom_point()+geom_point(size=2)+ggtitle('Transformed Shuck wt. vs Transformed Volume')+
  labs(x= "Log10(Volume)", y="log10(Shuck)")+
  theme(plot.title = element_text(size = 10, face = "bold"))
grid.arrange(plot3, plot4, nrow=1, ncol=2)

# (4)(a) Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and TYPE
# Use the multiple regression model: L_SHUCK~L_VOLUME+CLASS+TYPE.
# Apply summary() to the model object to produce results.

linear_model <- lm(L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = dfAbalone)
summary(linear_model)
capture.output(summary(linear_model),file="4a_linearModel.doc")

# (5)(a) Perform an analysis of the residuals resulting from the regression model in (3)
# If "linear_model" is the regression object, use linear_model$residuals and construct a
# histogram and QQ plot. Compute the skewness and kurtosis.

par(mfrow = c(1, 2))
hist(linear_model$residuals, main = 'Histogram of Residuals',  xlab = 'Residuals', ylab = ' ',
    col = 'cyan2')
qqnorm(linear_model$residuals,main = 'QQ Plot of Residuals', xlab = 'Theoretical Quantiles',
      ylab = 'Sample Quartiles', col = 'cyan2')
qqline(linear_model$residuals, datax = FALSE, distribution = qnorm,  col = 'red')
par(mfrow = c(1, 1))

rockchalk::skewness(linear_model$residuals)
rockchalk::kurtosis(linear_model$residuals)

# (5)(b) Plot the residuals versus L_VOLUME coloring the data points by CLASS,
# and a second time coloring the data points by TYPE
# Present boxplots of the residuals differentiated by CLASS and TYPE
# Test the homogeneity of variance of the residuals across classes using the bartlett.test()

#Scatterplot of model residuals as a function of L_VOLUME, CLASS, ggplot2
plot5 <- ggplot(linear_model, aes(x = L_VOLUME,y = linear_model$residuals)) +
   ggtitle('Chart A') +
   geom_point(aes(color = CLASS)) + labs(x = "L_VOLUME", y = "Residual")
plot6 <- ggplot(linear_model, aes(x = L_VOLUME,y = linear_model$residuals)) +
   ggtitle('Chart B') +
   geom_point(aes(color = TYPE)) + labs(x = "L_VOLUME", y = "Residual")
plot7 <- ggplot(linear_model, aes(x = CLASS,y = linear_model$residuals)) +
   ggtitle('Chart C') +
   geom_boxplot(aes(color = CLASS)) + labs(x = "CLASS", y = "Residual")
plot8 <- ggplot(linear_model, aes(x = TYPE,y = linear_model$residuals)) +
   ggtitle('Chart D') +
   geom_boxplot(aes(color = TYPE))+ labs(x = "TYPE", y = "Residual")
grid.arrange(plot5, plot6,plot7, plot8, nrow=2, ncol=2)

# Barlett test of homogeneity of variances
bartlett.test(linear_model$residuals ~ CLASS, data = dfAbalone)

# (6)(a) Calculate the proportion of infant abalones 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.
# -- Code from instructor's RMD file --
idxi <- dfAbalone$TYPE=="I"
idxa <- dfAbalone$TYPE=="ADULT"
max.v <- max(dfAbalone$VOLUME)
min.v <- min(dfAbalone$VOLUME)
delta <- (max.v - min.v)/1000
prop.infants <- numeric(0)
prop.adults <- numeric(0)
volume.value <- numeric(0)
total.infants <- length(dfAbalone$TYPE[idxi])
total.adults <- length(dfAbalone$TYPE[idxa])
for (k in 1:1000) {
  value <- min.v + k*delta
  volume.value[k] <- value
  prop.infants[k] <- sum(dfAbalone$VOLUME[idxi] <= value)/total.infants
  prop.adults[k] <- sum(dfAbalone$VOLUME[idxa] <= value)/total.adults
}
```

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

# check the outcome of the split
head (prop.adults, 20)
head (prop.infants,20)
head (volume.value, 20)

# (6)(b) 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.
# The two split points suggest an interval within which potential cutpoints may be located.
par(mfrow = c(1, 1))
plot(prop.adults, main = 'Proportion of Adults and Infants Protected', xlab = 'Volume',
    ylab = 'Proportion', col = 'cyan4', type = 'l', lwd = 2)
lines(prop.infants, col = 'cyan3', lwd = 2)
legend('bottomright',legend = c('Adult', 'Infant'), cex = 1, bg = "transparent",
      col = c('cyan4', 'cyan2'), lty = 1, lwd = 2)
abline(v= split.adults, col='orange')
abline(v= split.infants, col='orange')
abline(h= .50, col='orange')
text(split.adults, .46,  round(split.adults, 2),  pos = 4, offset = 1, col = "black")
text(split.infants, .46,  round(split.infants, 2),  pos = 4, offset = 1, col = "black")

# (7)(a) Evaluate a plot of the difference ((1-prop.adults)-(1-prop.infants)) versus volume.value.
# Compare to the 50% split points determined in (6)(b). 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.

difference <- (1-prop.adults) - (1-prop.infants)
plot(volume.value, difference, main = "Difference in Harvested Proportions", col = 'cyan4', type = "l",
lwd = 2, ylab = 'Difference in Proportions Harvested',
    xlab = 'VOLUME')

peak_x <- which.max(difference)
peak_x
abline(v=peak_x, lty = 2)
text(305, .3, paste("Volume =",peak_x), pos = 4, srt = 90)
```

```
# (7)(b) 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).

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


# (7)(c) Present a plot of the difference ((1-prop.adults)-(1-prop.infants)) versus
# volume.value with the variable smooth.difference superimposed.  Show the estimated peak
# location corresponding to the cutoff determined.
plot(volume.value, difference, main = "Difference in Harvested Proportions", col = 'cyan4', type = "l",
lwd = 2,
     ylab = 'Difference in Proportions Harvested', xlab = 'VOLUME')
lines(smooth.difference, col = 'orange', lwd = 2.5, lty=1)
max_val <- which.max(smooth.difference)
max_smooth <- volume.value[max_val]
abline(v=max_smooth, lty = 2, col = 'black')
note <- paste('Volume = ', round(max_smooth, 4))
text(max_smooth+20, .4, note, col = 'black', srt = 90)

# (7)(d) 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.")
TP_maxdiff <- (1-prop.adults)[which.max(smooth.difference)]
FP_maxdiff <- (1-prop.infants)[which.max(smooth.difference)]
TP_maxdiff
FP_maxdiff


# (8)(a) Harvesting of infants in CLASS "A1" must be minimized. The volume.value cutoff that
# produces a zero harvest of infants from CLASS "A1" is 207. Any smaller cutoff would result in
# harvesting infants from CLASS "A1." Calculate the separate harvest proportions for infants
# and adults if this cutoff is used. Report your results.

TP_0A1 <- (1-prop.adults)[207]
FP_0A1 <- (1-prop.infants)[207]
TP_0A1
FP_0A1


# Although the relevant volume.value - 207 - is given to you, we can demonstrate
# how it was arrived at. Specifically, we want to return the volume.value corresponding,
# element-wise, to the smallest volume.value greater than the largest VOLUME among
# CLASS "A1" infants.
v1 <- volume.value[volume.value >
    max(dfAbalone[dfAbalone$CLASS == "A1" & dfAbalone$TYPE == "I", "VOLUME"])][1] # [1] 206.9844
v1


# Now, to determine the proportions harvested, we can look to the proportions # of infants
# and adults with VOLUMEs greater than this threshold.

tot_p_1<-
sum(dfAbalone["VOLUME"]>v1)/(sum(dfAbalone$TYPE=="ADULT")+sum(dfAbalone$TYPE=="I"))


# (8)(b) 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.

TP_hc <- (1-prop.adults[253.6])
FP_hc <- (1-prop.infants)[253.6]
TP_hc
FP_hc


v2 <- volume.value[which.min(abs(prop.adults - (1-prop.infants)))]     # [1] 253.6113
v2
tot_p_2<-
sum(dfAbalone["VOLUME"]>v2)/(sum(dfAbalone$TYPE=="ADULT")+sum(dfAbalone$TYPE=="I"))


# (9) 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 locations of the cutoffs determined in (7) and (8) on this plot.
# Numerically integrate the area under the ROC curve and report your result.

plot(1-prop.infants, 1-prop.adults, type = 'l', lwd = 2, col='cyan4',
   main = "ROC of adult and infant havest proportions",
   xlab = 'Infant Harvest Prop', ylab='Adult Harvest Proportion')
abline(0, 1, col = 'orange', lty = 3, lwd = 2)

points((1-prop.infants)[207],(1-prop.adults)[207], col = 'red', pch=15)
text((1-prop.infants)[207],(1-prop.adults)[207], "harvesting 0 A1, vol = 207",
    col = 'black', pos = 4, offset = 1)

points((1-prop.infants)[253.6],(1-prop.adults)[253.6], col = 'red', pch=15)
text((1-prop.infants)[253.6],(1-prop.adults)[253.6], "equal harvest/conserve\n vol = 253.6",
    col = 'black', pos = 2, offset = 1)

points((1-prop.infants)[281.4],(1-prop.adults)[281.4], col = 'red', pch=15)
text((1-prop.infants)[281.4],(1-prop.adults)[281.4], "max. diff, vol = 281.4",
    col = 'black', pos = 4, offset = 1)
require(flux)
area <- round(flux::auc((1-prop.infants), (1-prop.adults)), 4)
note <- paste('AUC = ', area)
text (.8, .2, note)

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

# To calculate the total harvest proportions, we need to consider all individuals,
# regardless of SEX or TYPE, and what proportion are greater|less than a given
# cutoff. An example calculation, for the "maximum difference" approach is given here:
v3 <- volume.value[which.max(smooth.difference)]
tot_p_3<-
sum(dfAbalone$VOLUME>=volume.value[which.max(smooth.difference)])/(total.adults+total.infants)
# [1] 0.5501931
v3
tot_p_3

harvest<- c("zero harvest", "equal harvest", "max difference")
volume <- round(c(v1,v2,v3), 3)
FPR <- round(c(FP_0A1,FP_hc,FP_maxdiff), 3)
TPR <-round(c(TP_0A1,TP_hc,TP_maxdiff), 3)
PropYield <- round(c(tot_p_1, tot_p_2, tot_p_3),3)
prop_table <- cbind (harvest, volume, TPR, FPR, PropYield)
prop_table
capture.output(prop_table,file="10_Table.doc")

## ***** Auxilary Material, not part of assignment PDF
# Figure 5
par(mfrow = c(1, 2))
dfAdult$L_VOL <- log10(dfAdult$VOLUME)
dfInfant$L_VOL <- log10(dfInfant$VOLUME)
```

```
hist(dfInfant$L_VOL, main = 'Infant log10(Volume)',  xlab = 'Volume', ylab = 'Frequency',
      col = 'dodgerblue')
hist(dfAdult$L_VOL, main = 'Adult log10(Volume)', xlab = 'Volume', ylab = 'Frequency',
      col = 'dodgerblue4')
par(mfrow = c(1, 1))

# Data that got typed into Table 7
rockchalk::skewness(dfInfant$L_VOL)
rockchalk::kurtosis(dfInfant$L_VOL)
rockchalk::skewness(dfAdult$L_VOL)
rockchalk::kurtosis(dfAdult$L_VOL)


## Skew and kurtosis for Volume
rockchalk::skewness(dfAdult$VOLUME)
rockchalk::kurtosis(dfAdult$VOLUME)
rockchalk::skewness(dfInfant$VOLUME)
rockchalk::kurtosis(dfInfant$VOLUME)

## Figure 10, Chart 9A with 2sd lines
r_sd <- sd(linear_model$residuals)
ggplot(linear_model, aes(x = L_VOLUME,y = linear_model$residuals)) +
  geom_hline(yintercept = 2*r_sd, lty = 3) + geom_hline(yintercept = -2*r_sd, lty = 3)+
  geom_hline(yintercept = 3*r_sd, lty = 3) + geom_hline(yintercept = -3*r_sd, lty = 3, col='blue')+
  geom_point(aes(color = CLASS)) + labs(x = "L_VOLUME", y = "Residual")

# getting the class means of L_RATIO for Tukey discussion
C1 <-dfAbalone[dfAbalone$CLASS == 'A1',]
summary(C1)
C2 <-dfAbalone[dfAbalone$CLASS == 'A2',]
summary(C2)
```

Table 10
Sample of results from the calculations of cutoff to verify correctness of proportion calculations

```
> head (prop.adults, 20)
 [1] 0.000000000 0.000000000 0.001414427 0.001414427 0.001414427 0.001414427
 [7] 0.001414427 0.001414427 0.001414427 0.001414427 0.001414427 0.001414427
[13] 0.001414427 0.001414427 0.001414427 0.001414427 0.001414427 0.002828854
[19] 0.002828854 0.002828854


> head (prop.infants,20)
 [1] 0.01215805 0.01519757 0.01823708 0.02431611 0.02735562 0.03343465 0.03647416
 [8] 0.03647416 0.03951368 0.04255319 0.04255319 0.04559271 0.04559271 0.04559271
[15] 0.05167173 0.06079027 0.06079027 0.06382979 0.06382979 0.06382979


> head (volume.value, 20)
 [1]  4.603851  5.595913  6.587974  7.580036  8.572097  9.564159 10.556220
 [8] 11.548282 12.540343 13.532405 14.524466 15.516528 16.508589 17.500651
[15] 18.492712 19.484774 20.476835 21.468897 22.460958 23.453019
```