Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Introduction

It is imperative to the survival and prosperity of the abalone species and the harvesting community to clearly delineate a measure of acceptable abalone harvest levels. In a previous exploratory data analysis report, Data Analysis Assignment 1, results demonstrated the methods being used to develop a binary rule based system to determine a cutoff value for age. Thus, in this report, methods utilizing regressions analysis, ROC, kurtosis, log transformations, variance homogeneity tests, and graphical analysis are carried out to create the binary decision rules using only infant and adult cutoff values based on a new classification system for harvesting.

Results

Data from Data Analysis Assignment 1 was re-examined before implementing any new methods, to gain a baseline understanding of the data. The rockchalk package in R, which subtracts three from the kurtosis to ontain the excess kurtosis making the kurtosis of a normal distribution zero, was used to obtain a kurtosis of 1.667. This value is slightly outside the acceptable range for normal distribution and represents that the dataset has more weight in the tails of the distribution, thus is leptokurtic with data more clustered around the mean value (**Figure 1a**) producing a distribution with a higher peak. This means that small changes in **RATIO** will occur less, but within the fat tails large fluctuations are more likely to be observed. This type of distribution is more susceptible to extreme observations should be closely examined when making decisions of great impact. The skewness of 0.715 for **RATIO** indicates that the distribution is slightly skewed to the right. Therefore, the return distribution is not consistent with normal distribution.

Figure 1a: Histogram of RATIO Frequency Distribution

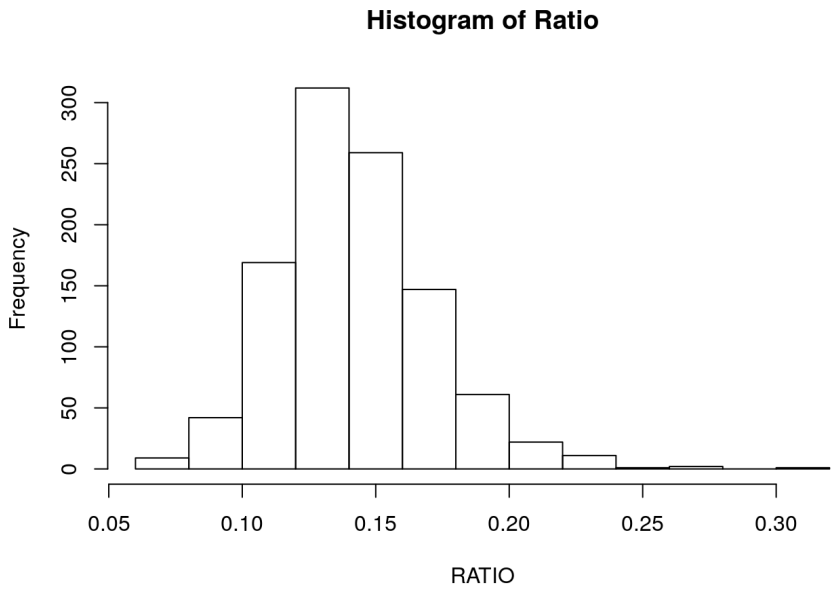
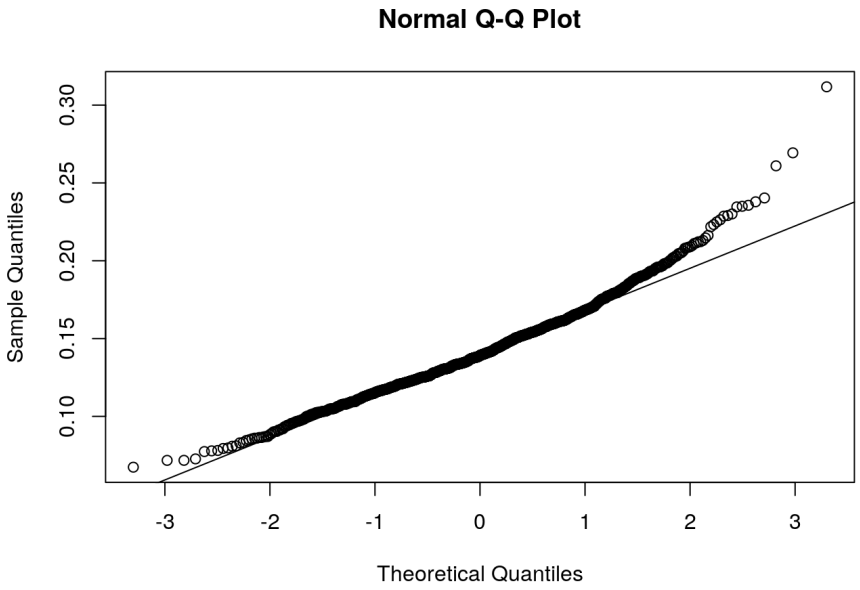
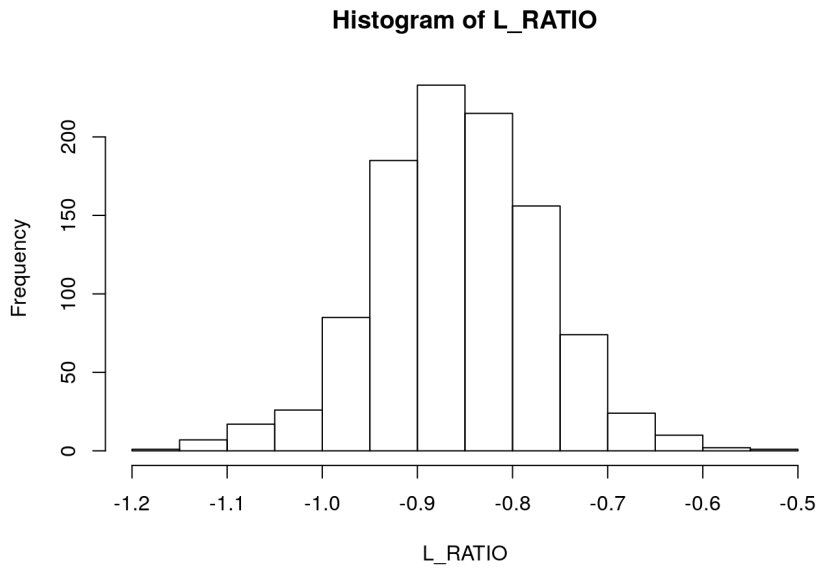


Figure 1b: Normal Q-Q Plot of RATIO



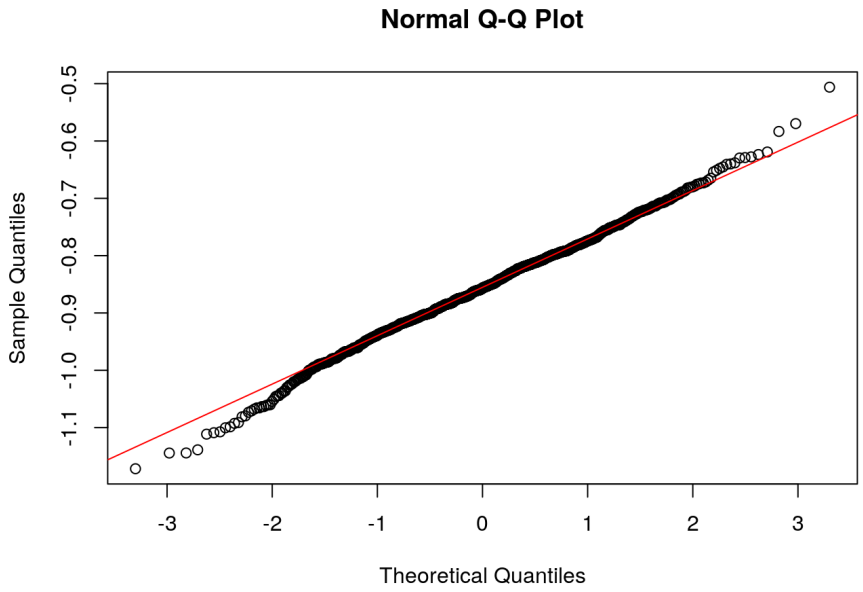
Log transformations on **RATIO** were carried out, **L_RATIO**, to rescale and normalize the data. Then the above analysis was repeated. The kurtosis dropped to 0.5354, which is better than before, but the distribution is still slightly outside acceptable ranges to have a normal distribution. The tails have thinned out and reducing the possibility of uncertainty being introduced via outlier events, while reducing the occurrence of extreme observations (**Figure 1c**). Since the kurtosis is positive it is still considered leptokurtic and is now slightly skewed to the left (**Figure 1c**) with a value of -0.0939, but it is acceptable since it is close zero, and thus is close to having a normal distribution. A Q-Q plot (**Figure 1d**) was constructed to gain a different perspective of the distribution **L_RATIO.** The normal line (red) shows that the distribution still deviates from the norm roughly outside the range of [-1.6, 2.2], but fits a normal distribution more than before the logit transformation on **RATIO** was carried out.

Figure 1c: Histogram of Frequency Distributions for L_RATIO



Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Figure 1d: Q-Q Plot



Since the survivability of the abalone species and industry depend on well examined data based decisions a Bartlett's test heteroscedasticity of variances across classes, with a null hypothesis that the variance across the six classes is the same, was also conducted and boxplot of **CLASS** verses **L_RATIO** was constructed to look at the outliers. The Bartlett's test produced a p-value of 0.6884, which is greater than significance level of 0.05, thus we fail to reject the null hypothesis. Therefore, there is not significant difference variance among the six classes.

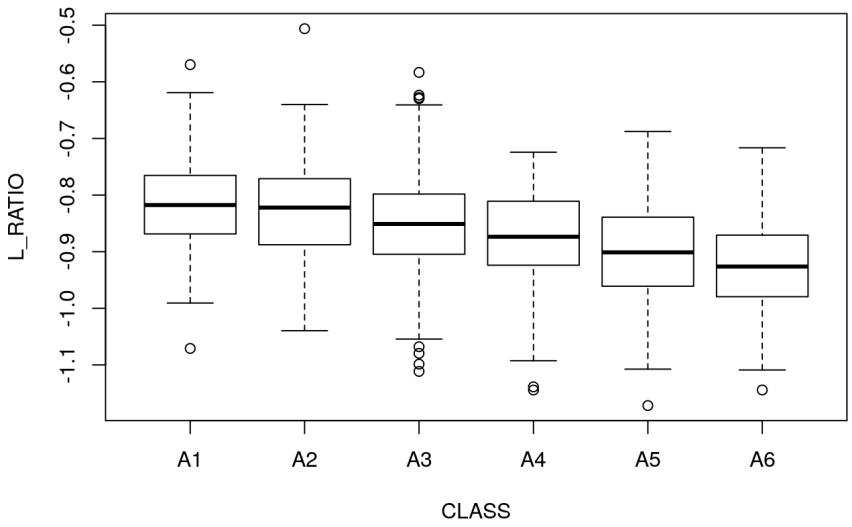
Table 1: Bartlett Test for L_RATIO Across All Classes

Bartlett test of homogeneity of variances
Bartlett's K-squared = 3.0749, df = 5, p-value = 0.6884

Boxplots, differentiated by **CLASS** (**Figure 1e**), reveal outliers in all six of the classes. This indicates all classes contain **L_RATIO** values that do not fit within the correct age range for that **CLASS**. The distribution for classes A1, A2, A3, A5, and A6 are reasonably symmetric with class A4 having a distribution slightly skewed left. Further, classes A1 and A2 have centers, or means, that are very close in value, with A2 having slightly more variability than A1. Class A3 contains the most outliers with slightly less variance in distribution when compared to other classes.

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Figure 1e: Boxplot of L_RATIO Differentiated by CLASS



An analysis of variance on **L_RATIO** using **CLASS** and **SEX** as independent variables, with the assumption of equal variances, was performed for two different models; Model_1 contains an interaction term **CLASS:SEX** and Model_2 does not have an interaction term. When comparing the two models show that both **CLASS** and **SEX** have a significant effect on **L_RATIO**, while the interaction term **CLASS:SEX** is not significant.

Table 2a: Model_1 AOV() on L_RATIO using CLASS and SEX, with Interaction Term CLASS:SEX

	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		

Table 2b: Model_2 AOV() on L_RATIO using CLASS and SEX, with no 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		

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

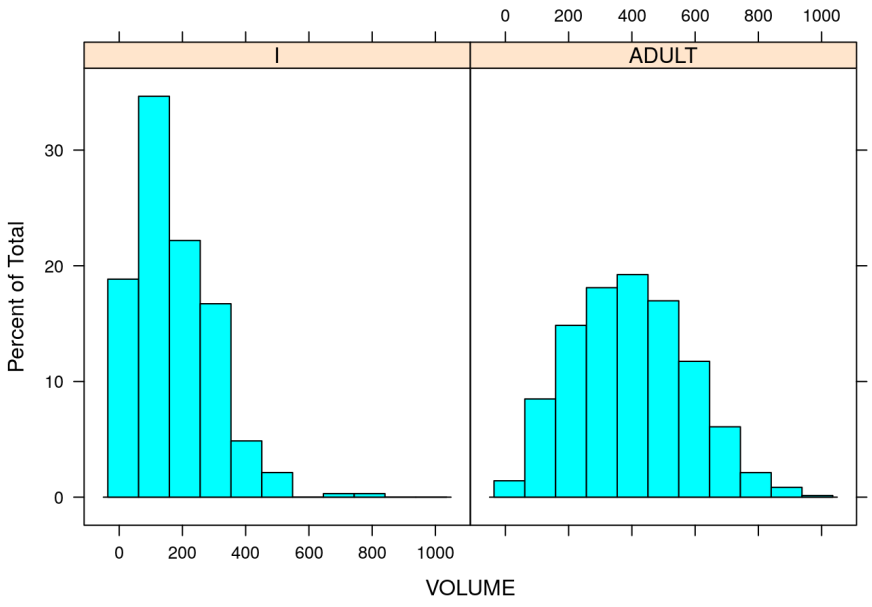
Model_2 was used to obtain multiple comparison across **CLASS** and **SEX** via the TukeyHSD() function in R at the 95% confidence level. When looking at the results for multiple comparison across **CLASS** there is no significant difference between classes A2-A1, A5-A4, and A6-A5 since their respective p-values are greater than the significance level of 0.05 (**Table 3b** in Appendix). This means that A2-A1 could be combined into one category, with A5-A4 and A6-A5 presented with the same possibility. Note that A3-A2 have a p-value of 0.0224, which is less than the significant value of 0.05, but if it makes sense these classes could be combined as well. Further, analysis is need to make that decision. Then looking at the three factors of abalone sex, male (M), female (F), and infant (I) only the comparison M-F showed no significant difference with a p-value of 0.9415, which is greater than the significant level of 0.05 (**Table 3a**). Thus, results suggest that male and female abalones could be combined into one category 'Adult.'

Table 3a: Multiple Comparisons with the TukeyHSD() Function using Model_2

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

Given the result in Table 3a, "M" and "F" were combined into a new level labeled "ADULT" producing a new variable **TYPE**. To examine the distributions regarding separation of infants, "I," from adults based on VOLUME side by side histograms were used (**Figure 2**). Clearly, the infant distribution skewed to the right, while the distribution of adults is slightly skewed to the right. Further, we can see that the distribution for infants is leptokurtic with a large peak.

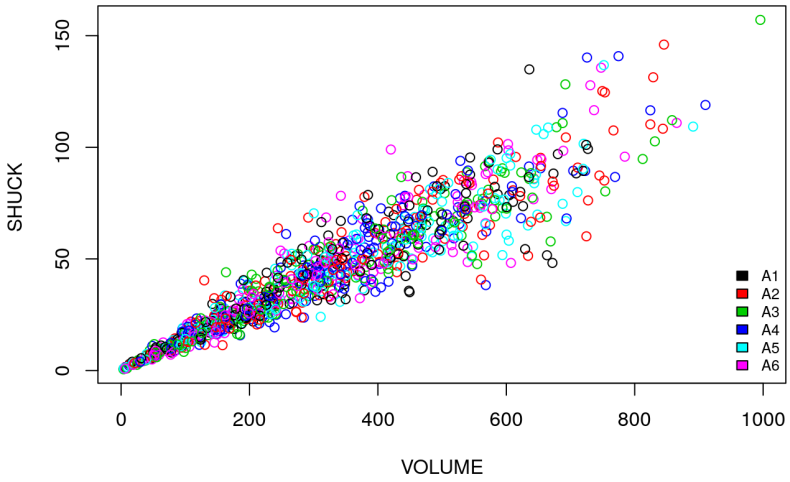
Figure 2: Histograms Infant (left) and Adult (right) Volumes



Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

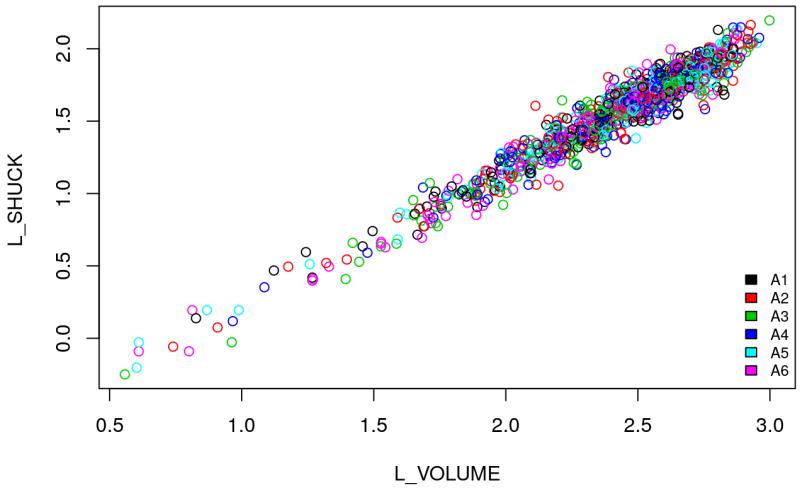
Looking at the scatterplot of **SHUCK** versus **VOLUME** differentiated by **CLASS** levels (**Figure 3a)**, harvesting volumes mostly resides below a volume value range of 500 to 600 and is tightly compact and skewed right. This makes it difficult to see the differences in the volumes per class and reveals that the mean is greater than its median. It does show that above this range there are fewer infants being harvested, but the other classes harvesting values also becomes sparse due to the rarity of such a large shuck weight and volume. Thus, scatterplot of their base ten logarithm transformations **L_SHUCK** versus **L_VOLUME**, differentiated by **CLASS** was constructed to compare these two methods. The variables **L_SHUCK** and **L_VOLUME** present the data as orders of magnitude (**Figure 3b)**. This transformation allows to drill down to the main shuck weight verses volume for abalones. There are still infants present in the volume range, but are mostly harvested in the **L_VOLUME** range 1.5 to 3.0, or 31.622 to 1000. This is quite different than the results presented in Figure 3a, and skews the distribution to the left resulting in a mean that is smaller than the median.

Figure 3a: Scatterplot SHUCK versus VOLUME, Differentiated by CLASS



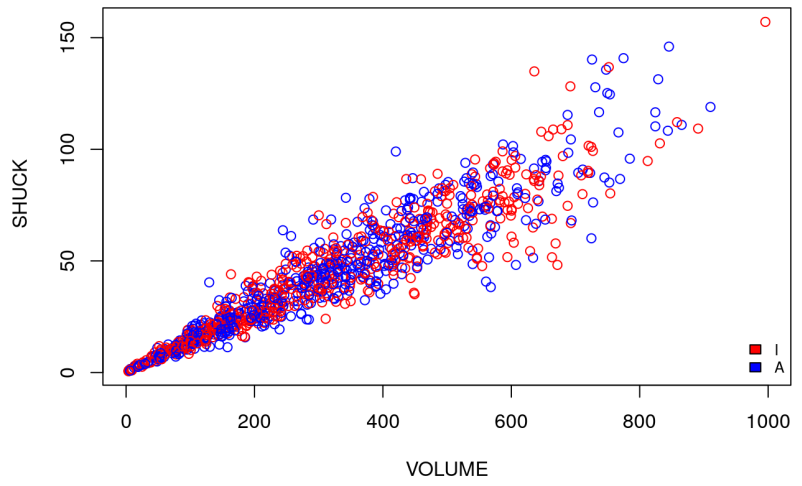
Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Figure 3b: Scatterplot L_SHUCK versus L_VOLUME, Differentiated by CLASS



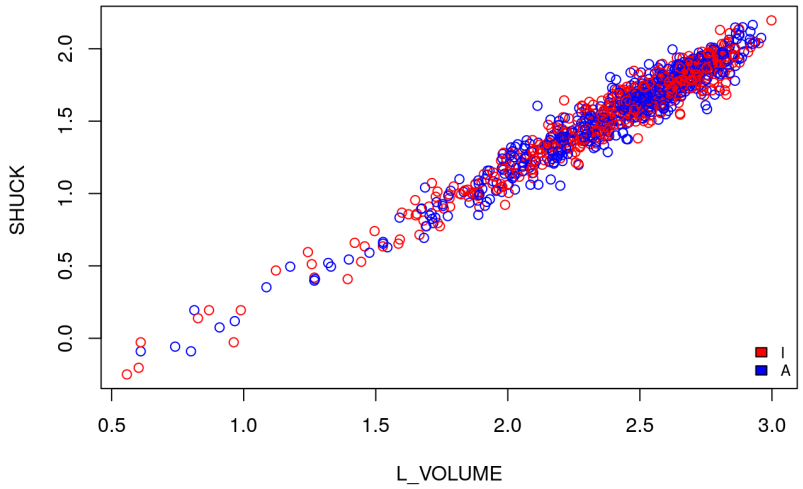
To see the different shuck weight verses volume at harvest for adults and infant abalones a scatterplot of **SHUCK** versus **VOLUME** differentiated by **TYPE** levels (**Figure 3c)** was constructed. In **Figure 3c** harvesting volumes still mostly resides below a volume value range of 500 to 600 and is tightly compact and skewed right as it was when differentiated by **CLASS**. A scatterplot of their base ten logarithms, **L_SHUCK** verses **L_VOLUME**, differentiated by **TYPE**, also demonstrates that there are still infants present in the volume range, but are mostly harvested in the **L_VOLUME** range 1.5 to 3.0, or 31.622 to 1000. These results demonstrating that the base ten logarithm transformations for the analysis are appropriate.

Figure 3c: Scatterplot SHUCK versus VOLUME, Differentiated by TYPE



Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Figure 3d: Scatterplot L_SHUCK versus L_VOLUME, Differentiated by TYPE



A2 is the only class that is not significantly different at the 0.05 significance level (**Table 4a)** since it has a p-value of 0.1129. Note, a trend exists that as the classes increase, starting at A3 with a p-value of 0.0001 and ending at A6 with a p-value of 2.28e-16, the p-values get exponentially smaller (**Table 4a)**. This means that as the class increases there is a significant effect on **L_SHUCK**. Further, notice that **L_VOLUME** has a p-value of 2e-16 meaning that it also has a significant effect on **L_SHUCK**. The last independent variable to examine in this regression is **TYPE**. The p-value for **TYPE** is 0.0002, which is much smaller than the 0.05 significance level. Thus, **TYPE** also has a significant effect on **L_SHUCK**. Therefore, an abalones log(volume), class increase log(shuck) will increase due to growth that can occur in each class and there is a delineation established between infants and adult abalones, which also demonstrates that as the abalone moves up in class, from A3 to A4, and increases in volume the abalone will be reclassified from infant to adult.

Table 4a: Regression L_SHUCK (dependent variable) on L_VOLUME, CLASS, TYPE

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

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

After class A3, infants that find themselves in A4 and A5 classes become eligible for harvesting. Thus, infants found in classes A4 and A5 were reclassified as **ADULTS**. Then Model_2 was regressed again with this new classification. In **Table 5a**, we see the same trend of p-values as in **Table 4a**. The p-value for class A2, 0.1011, is still well above the significance level of 0.05 and the p-value for TYPEADULT, 0.0041, is still below the significance level leading to the same interpretation as in **Table 4a**, but there is a slight decrease in their respective p-values.

Table 5a: Model_2 Regressed with New Classification, A4 & A5 'ADULTS'

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.817335	0.019036	-42.937	< 2e-16
L_VOLUME	0.999170	0.010260	97.381	< 2e-16
CLASSA2	-0.018025	0.010985	-1.641	0.101135
CLASSA3	-0.047614	0.012419	-3.834	0.000134
CLASSA4	-0.075936	0.013974	-5.434	6.87e-08
CLASSA5	-0.104175	0.015392	-6.768	2.19e-11
CLASSA6	-0.127145	0.015237	-8.344	2.28e-16
TYPEADULT	0.021418	0.007451	2.874	0.004131

Residuals of the regression performed on Model_2 are value that are produced over after subtracting predicted values from observed values. Residuals can give an idea of how much the model with actual data deviates from the ideal predicted behavior. These residual values can be analyzed both contextually and visually to help make decisions to improve the model. Contextually the kurtosis and skewness are examined. The kurtosis for the new model is 0.3514, demonstrating a close to normal distribution, and is very slightly skewed left with a value of -0.0616. These results tell us that the Model_2 does not contain or contains very little unobservable errors unaccounted for in Model_2. This conclusion can be graphically validated by observing the histogram (**Figure 4a**) and Q-Q plot (**Figure 4b**) for the residuals of the regression for Model_2.

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Figure 4a: Model_2 Histogram of Residuals

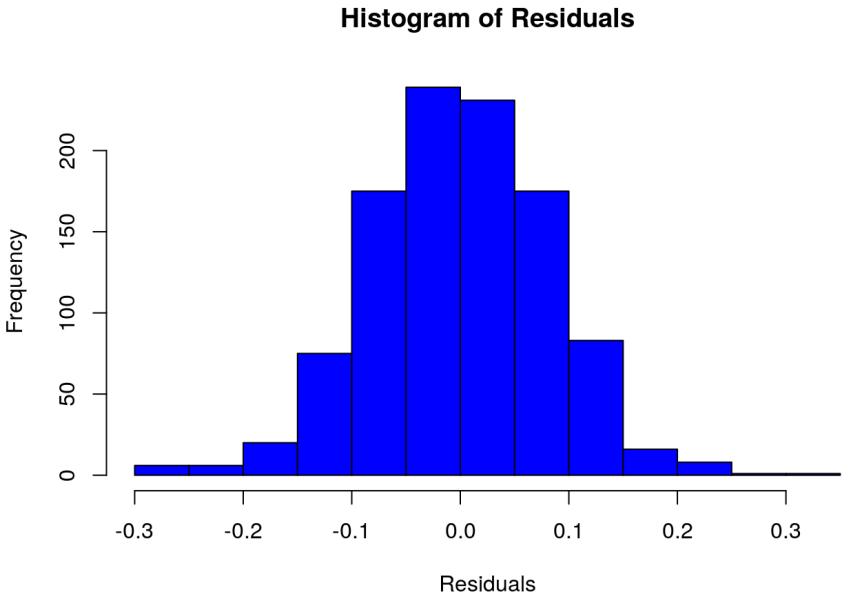
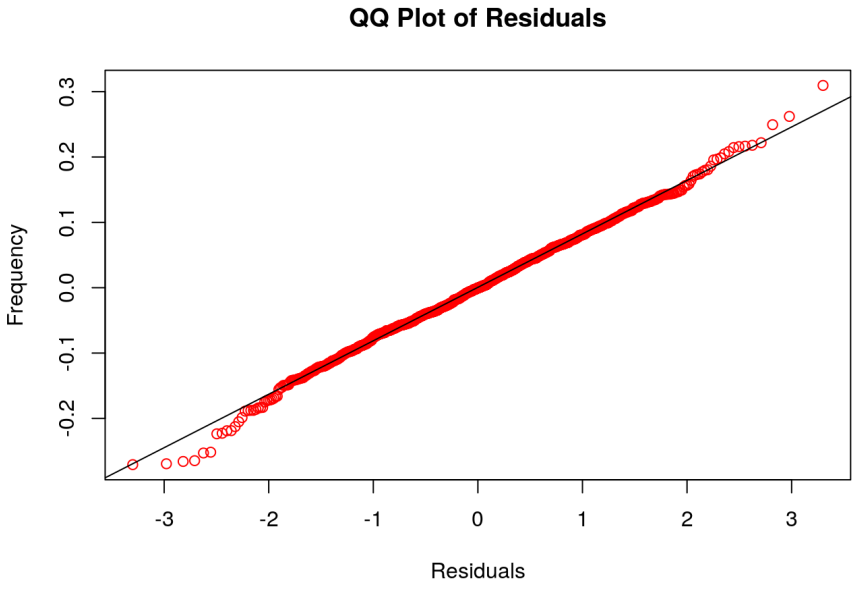


Figure 4b: Model_2 Q-Q Plot of Residuals



Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

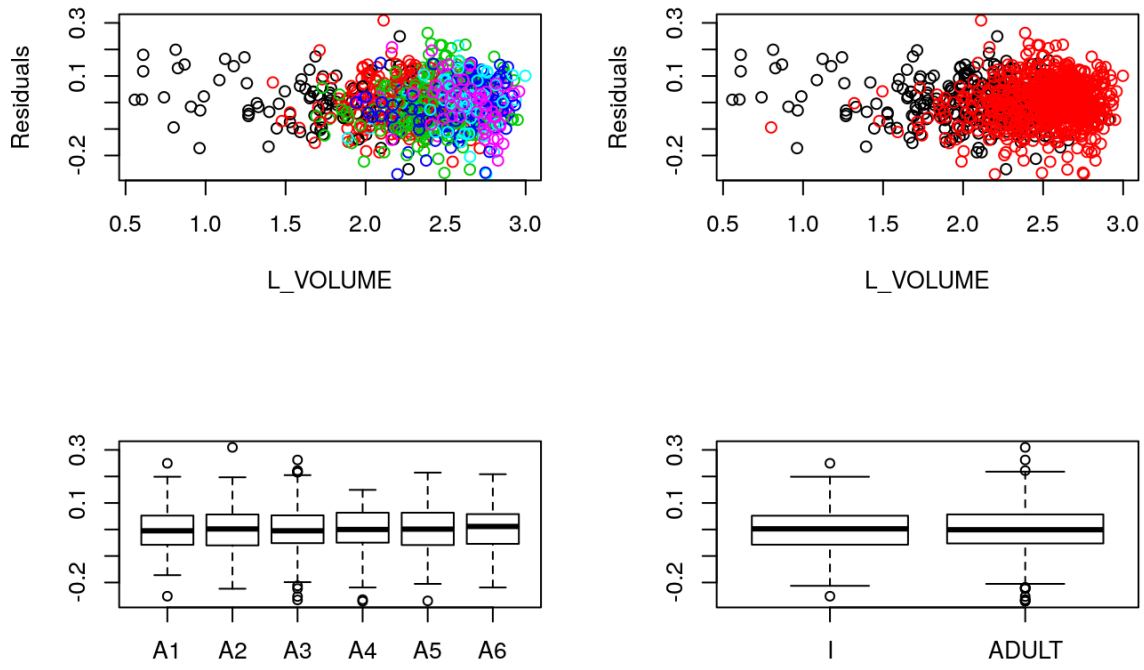
In **Figure 4c**, the distributions differentiated by **CLASS** and by **TYPE**, are still skewed left, but are more tightly compact than previously. Further, the boxplot for residuals differentiated by class show that the means are more inline with each other, with fewer outliers present. The boxplot differentiated by type also reflect this same trend, but with more outliers in the residuals for **L_VOLUME**. This shows that the base ten logarithim transformation on **L_VOLUME** fits the data better than with out the transformation.

A Bartlett test of heteroscedasticity was conducted to evaluate the variance of residuals across the classes (**Table 6)**. The resulting p-value is greater than the significance value of 0.05. Therefore, there is no significant difference in the variances between classes.

Table 6: Bartlett Test of Heteroscedasticity of Variance of Residuals Across CLASS

Bartlett test of homogeneity of variances
Bartlett's K-squared = 3.5356 df = 5, p-value = 0.618

Figure 4c: Scatterplots (top) and Boxplots (bottom) of L_VOLUME Differentiated by CLASS (left), Differentiated by TYPE (right)



Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

In **Figure 5**, a plot of the infant proportions (red) and adult proportions (blue) versus volume is presented. The points marked on each line represent a 50% split value for **VOLUME** and it is within this interval where the potential cutoff points could be located.

Figure 5: Proportion of Adults, Infants Protected, with 50% Split Value for VOLUME

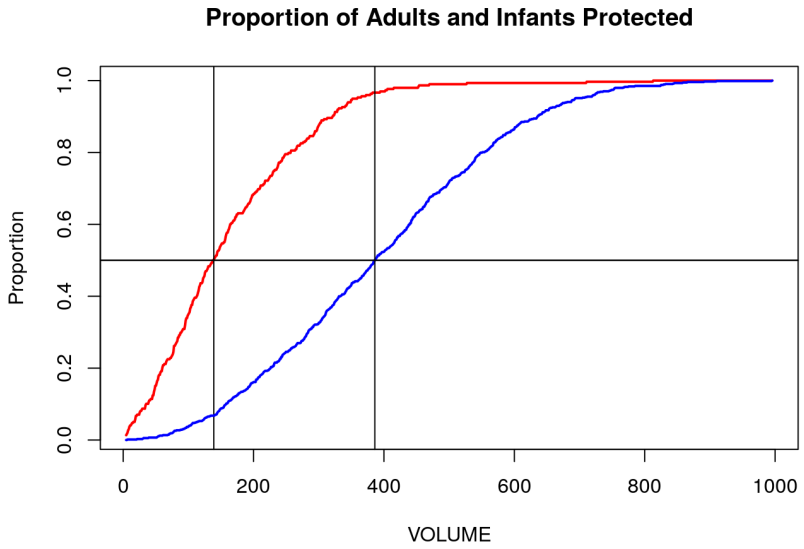
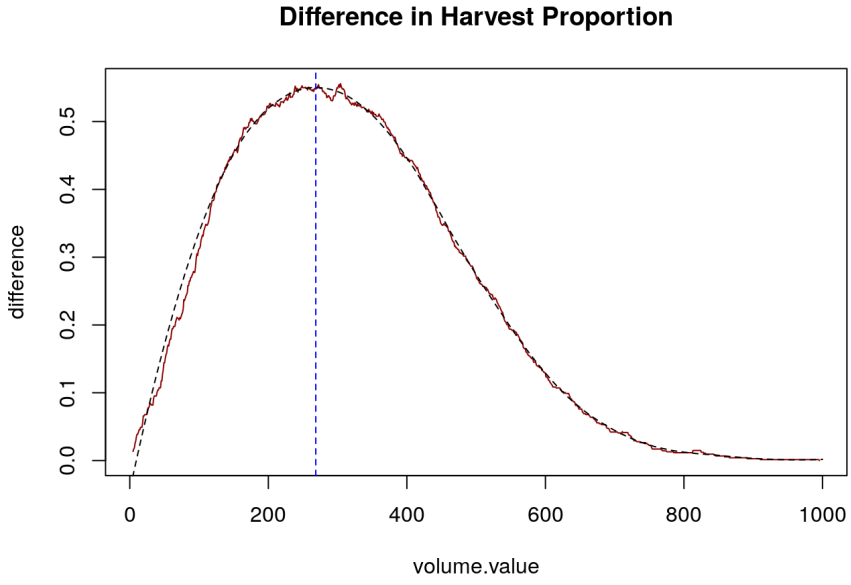


Figure 6 presents the difference in true positive rates and false positive rates versus **VOLUME** values. At the peak there is a jagged line fit to the curve which indicates variability is present at the peak area and does not represent the data well. Thus, to improve the representation of the data, the data was smoothed and plotted (black dashed line). When compared to the 50% split, the resulting distribution starts to take a more normalized shape, but produces a max cutoff volume.value of 268.492.

Figure 6: Difference ((1-prop.adults)-(1-prop.infants)) versus Volume.value

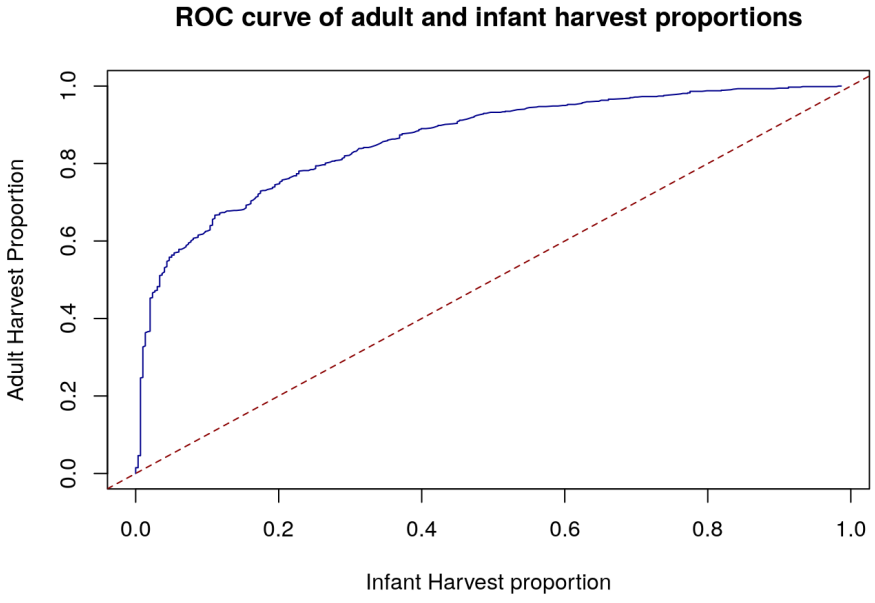


Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

If the cutoff value of 206.984 were used the harvest portion for infants would give the result of 0.1812. Using the same cutoff volume value for the adult harvest proportion ("true positive rate") the value obtained is 0.7304. This seems like a reasonable value, but a better defined cutoff value needs to be examined. To minimize the harvesting of infants in **CLASS** A1, a volume cutoff value of 241.7065 is needed to reduce the harvesting of the infants in **CLASS** A1 to zero resulting in the harvest population proportions of infants, with VOLUME greater than 206.984, of 0.3020. The smallest volume value greater than the largest **VOLUME** among **CLASS** A1, for the adult population, is 350.833 with an adult population portion of 0.8317. The equated volume value is 253.6113 with a 0.2675 harvest proportion for infants and 0.7779 harvest proportion for adult abalones.

The ROC curve purpose is to visualize and quantify the tradeoff made between the two measures, true positive rate (TPR) and false positive rate (FPR). The TPR is the adult harvest proportion on the y-axis and the FPR is the infant harvest proportion on the x-axis at various cutoff values ranging between 0 and 1. The cutoff values cause the "bend" in the line representing a decrease in the rate of change quickly the infant harvest proportion increases past zero and greater than 0.4 of the adult harvest proportion. The value of for the area under the curve (AUC) of 0.848 was obtained. This value shows that the model is performing well when compared to the linear, dashed red line, which represents an average model with an AUC of 0.5.

Figure 7:



In **Table 7**, the final resulting three potential cutoff values identified, along with the respective true positive rates, false positive rates, and total proportional yield where the proportion harvested considers both adult and infant abalones are presented.

Table 7: True Positive Rate, False Positive Rate, Harvest Proportion of the Total Population (all adults and infants considered)

	Volume	TPR	FPR	PropYield
max.difference	268.492	0.730	0.181	0.572
zero.A1.infants	206.984	0.831	0.340	0.676
equal.error	253.611	0.778	0.267	0.593

Conclusion

Data coming in from the real world rarely displays a normal distribution, thus using a base 10 logarithm transform on the variables **SHUCK**, **VOLUME**, and **RATIO** contributed to a more normalized data distribution allowing for a regression analysis to completed in order to improve appropriate cutoff values for harvest levels. Note, actual field data was transformed the magnitude for comparison these results will not correctly reflect reality, but is a good approximation. Further, with the outliers present in the classes, more data needs to be collected to try to minimize the impact on the results. Thus, for implementing a selected cutoff I feel that hedging on a higher cutoff will protect more infants and in return allow more infants to grow to a maturity where harvesting will be a return on investment while allowing the near adult abalones to thrive and grow larger for next harvest. This cutoff will need to be adjusted as new population data arrives each season and this analysis should continue to evolve along with the updates.

References

Anon, (2017). [online] Available at: http://investopia [Accessed 28 Aug. 2017].

Davies, T. (2016). *The Book of R: A First Course in Programming and Statistics*. San Francisco, CA: William Pollock.

Galili, T., Smith, D., Galili, T., Lamstein, A., Goldfeld, K., Mind, T., Sun, S., Software, E., Touzin, G., Mount, J., Random, T., Carpenter, B., Smith, D., Kasvikis, E., box, T., Oberg, R., Walia, A., Schlegel, A., Blogger, G., Ghosh, B., Magnusson, K., Blog, D. and Guides, E. (2017). *R-bloggers | R news and tutorials contributed by (750) R bloggers*. [online] R-bloggers. Available at: http://R-bloggers.com [Accessed 28 Aug. 2017].

Kabacoff, R. (2015). *R in Action: Data Analysis and Graphics with R*. 2nd ed. Shelter Island, NY: Manning Publishing Co.

Wilcox, R. (2009). *Basic Statistics: Understanding Conventional Methods and Modern Insights.* New York, NY: Oxford University Press.

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Appendix

Table 3b: Multiple Comparisons with the TukeyHSD() Function using Model_2

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mydata$L_RATIO ~ CLASS + SEX, data = mydata)
##
## $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
##
## $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
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Table 4b: Regression L_SHUCK (dependent variable) on L_VOLUME, CLASS, TYPE

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

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

Table 5b: : Model_2 Regressed with New Classification, A4 & A5 'ADULTS'

```
##
## Call:
## lm(formula = L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)
##
## Residuals:
##       Min         1Q    Median         3Q       Max
## -0.270689 -0.054513 -0.000241  0.055806  0.309518
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.817335   0.019036 -42.937  < 2e-16 ***
## L_VOLUME     0.999170   0.010260  97.381  < 2e-16 ***
## CLASSA2     -0.018025   0.010985  -1.641 0.101135
## CLASSA3     -0.047614   0.012419  -3.834 0.000134 ***
## CLASSA4     -0.075936   0.013974  -5.434 6.87e-08 ***
## CLASSA5     -0.104175   0.015392  -6.768 2.19e-11 ***
## CLASSA6     -0.127145   0.015237  -8.344 2.28e-16 ***
## TYPEADULT    0.021418   0.007451   2.874 0.004131 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.08286 on 1028 degrees of freedom
## Multiple R-squared:  0.9506, Adjusted R-squared:  0.9503
## F-statistic:  2825 on 7 and 1028 DF,  p-value: < 2.2e-16
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

R-Code

```
#libraries
library(stats)
library(rockchalk)
library(ggplot2)

#reads in data from csv file, stores in dataframe "mydata", had to use sep="
" since not comma seperated
mydata <- read.csv(file.path("/home/Dominator/Documents/PREDICT_401/DataAnaly
sis_02/","mydata.csv"), sep = " ")

#sanity check
str(mydata)

## 'data.frame':    1036 obs. of  10 variables:
##  $ SEX   : Factor w/ 3 levels "F","I","M": 2 2 2 2 2 2 2 2 2 2 ...
##  $ LENGTH: num  5.57 3.67 10.08 4.09 6.93 ...
##  $ DIAM  : num  4.09 2.62 7.35 3.15 4.83 ...
##  $ HEIGHT: num  1.26 0.84 2.205 0.945 1.785 ...
##  $ WHOLE : num  11.5 3.5 79.38 4.69 21.19 ...
##  $ SHUCK : num  4.31 1.19 44 2.25 9.88 ...
##  $ RINGS : int  6 4 6 3 6 6 5 6 5 6 ...
##  $ CLASS : Factor w/ 6 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 ...


#histogram using RATIO
hist(mydata$RATIO, main ="Histogram of Ratio", xlab = "RATIO", ylab = "Freque
ncy")

#Q-Q plot using RATIO
qqnorm(mydata$RATIO)
qqline(mydata$RATIO)

#skewness
rockchalk::skewness(mydata$RATIO)

## [1] 0.7147056

#kurtosis
rockchalk::kurtosis(mydata$RATIO)

#Transforms Ratio usling log10() to create L_RATIO
mydata$L_RATIO <- log10(mydata$RATIO)

#Histogram using L_RATIO
hist(mydata$L_RATIO, main = "Histogram of L_RATIO", xlab = "L_RATIO", ylab =
"Frequency")
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

```r
#QQ plots using L_RATIO
qqnorm(mydata$L_RATIO, main = "Normal Q-Q Plot")
qqline(mydata$L_RATIO, col = 2)

#skewness
rockchalk::skewness(mydata$L_RATIO)
```

```
## [1] -0.09391549
```

```r
#kurtosis
rockchalk::kurtosis(mydata$L_RATIO)
```

```
## [1] 0.535431
```

```r
#box plots by class
#Boxplot of model residuals differentiated by CLASS and TYPE

boxplot(split(mydata$L_RATIO, mydata$CLASS), xlab = "CLASS", ylab ="L_RATIO")

# bartlett.test() tests null hypothesis of homogeneity of variance of a numer
ic variable # across two (2) or more groups or levels of a factor.

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

#analysis of variance on L_RATIO, CLASS and SEX independent vars
#CLASS:SEX is the interaction term
varAnalysInteraction <-  aov(mydata$L_RATIO ~ CLASS * SEX, data = mydata)


#analysis of variance on L_RATIO, CLASS and SEX independent vars
#No interaction term
varAnalysNonInteract <- aov(mydata$L_RATIO ~ CLASS + SEX, data = mydata)

#summary with CLASS:SEX interaction term
summary(varAnalysInteraction)

#summary without CLASS:SEX interaction term
summary(varAnalysNonInteract)

#multple comparision using TukeyHSD
multiComparision <- TukeyHSD(varAnalysNonInteract, conf.level = 0.95)

#print the TukeyHSD comparision
multiComparision

# the rockchalk package was previously loaded
mydata$TYPE <- combineLevels(mydata$SEX, levs = c("M","F"), "ADULT")
```

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

```r
#setup side by side histograms
par(mfrow = c(1,2))
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

```
#histogram infant voumes
require(lattice)

## Loading required package: lattice

histogram(~VOLUME|TYPE, data = mydata)

library(lattice)
# define the base ten logarithm vectors
mydata$L_SHUCK <- log10(mydata$SHUCK)
mydata$L_VOLUME <- log10(mydata$VOLUME)

colours <- c(1,2,3,4,5,6)

#scatterplots color differentiated by CLASS
## SHUCK vs VOLUME
plot( mydata$VOLUME, mydata$SHUCK,  col = colours, xlab = "VOLUME", ylab ="SH
UCK ")
legend("bottomright", c("A1","A2","A3","A4","A5", "A6"), cex=.75, bty="n", fi
ll= colours, col = colours)

## L_SHUCK vs L_VOLUME
plot(mydata$L_VOLUME, mydata$L_SHUCK,  col = colours,xlab = "L_VOLUME", ylab
="L_SHUCK ")
legend("bottomright", c("A1","A2","A3","A4","A5", "A6"), cex=.75, bty="n", fi
ll= colours,  col = colours)

#scatterplots color differentiated by TYPE

## SHUCK vs VOLUME
plot(mydata$VOLUME, mydata$SHUCK,xlab = "VOLUME", ylab =" SHUCK", col = c("re
d", "blue"))
legend("bottomright", c("I","A"), cex=.75, bty="n", fill=  c("red", "blue"),
col = c("red", "blue"))

## L_SHUCK vs L_VOLUME
plot(mydata$L_VOLUME, mydata$L_SHUCK, xlab = "L_VOLUME", ylab ="SHUCK ", col
= c("red", "blue"))
legend("bottomright", c("I","A"), cex=.75, bty="n", fill = c("red", "blue"),
col = c("red", "blue"))

#OLS: Dependent variable = L_SHUCK, Independent variables: L_VOLUME, CLASS, T
YPE
linear_model_4a <- lm(L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)

#print OLS results
summary(linear_model_4a)

# note the code below for reclassifying infants
index <- (mydata$CLASS == "A5")|(mydata$CLASS == "A4")
mydata$TYPE[index] <- combineLevels(mydata$TYPE[index],
                         levs = c("I", "ADULT"), "ADULT")
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

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

#OLS: Dependent variable = L_SHUCK, Independent variables: L_VOLUME, CLASS, T
YPE
linear_model_4b <- lm(L_SHUCK ~ L_VOLUME + CLASS + TYPE, data = mydata)

#print OLS results
summary(linear_model_4b)

# Histogram, base R example
hist(linear_model_4b$residuals,xlab= "Residuals", ylab= "Frequency", col  = "
blue",main = "Histogram of Residuals" )

# QQ plot, base R example
qqnorm(linear_model_4b$residuals, main ="QQ Plot of Residuals", xlab = "Resid
uals", ylab = "Frequency", col = "red")
qqline(linear_model_4b$residuals)

# skewness() and kurtosis() functions are defined by both the "moments"
# and "rockchalk" packages. You can specify the package you want the
# function of by adding "package_name::" before the function.
#"rockchalk" that the kurtosis value has 3.0 subtracted from which it differs
from the "moments" package
moments::kurtosis(linear_model_4b$residuals)

## [1] 3.357953

# OR,
rockchalk::kurtosis(linear_model_4b$residuals)

## [1] 0.3514734

#set up plot locations
par(mfrow = c(2,2))

# Scatterplot of model residuals as a function of L_VOLUME, CLASS, color diff
erientiated by CLASS
plot( mydata$L_VOLUME,linear_model_4b$residuals, xlab = "L_VOLUME", ylab = "R
esiduals",col = c(mydata$CLASS))

# Scatterplot of model residuals as a function of L_VOLUME, CLASS, color diff
erentiated by TYPE
plot(mydata$L_VOLUME,linear_model_4b$residuals,  xlab = "L_VOLUME", ylab = "R
esiduals", col = c(mydata$TYPE))

#Boxplot of model residuals differentiated by CLASS and TYPE

boxplot(split(linear_model_4b$residuals, mydata$CLASS))
boxplot(split(linear_model_4b$residuals, mydata$TYPE))

# Barlett test of homogeneity of variances
bartlett.test(linear_model_4b$residuals ~ CLASS, data = mydata)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  linear_model_4b$residuals by CLASS
## Bartlett's K-squared = 3.5356, df = 5, p-value = 0.618

#clear par(mfrow)
par(mfrow = c(1,1))

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

max.v <- max(mydata$VOLUME)
min.v <- min(mydata$VOLUME)

delta <- (max.v - min.v)/1000

prop.infants <- numeric(0)
prop.adults <- numeric(0)

volume.value <- numeric(0)

total.infants <- length(mydata$TYPE[idxi])
total.adults <- length(mydata$TYPE[idxa])

for (k in 1:1000) {
 value <- min.v + k*delta
  volume.value[k] <- value
  prop.infants[k] <- sum(mydata$VOLUME[idxi] <= value)/total.infants
  prop.adults[k] <- sum(mydata$VOLUME[idxa] <= value)/total.adults
}

#These proportions show the impact of increasing the volume cutoff for harves
ting. The following code shows how to "split" the population at a 50% harvest
level.
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

head(volume.value)

## [1] 4.603851 5.595913 6.587974 7.580036 8.572097 9.564159

head(prop.adults)

## [1] 0.000000000 0.000000000 0.001355014 0.001355014 0.001355014 0.00135501
4

head(prop.infants)

## [1] 0.01342282 0.01677852 0.02013423 0.02684564 0.03020134 0.03691275
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

```
# ?plot(), ?abline() to review documentation pages

plot(volume.value, prop.infants, col = "red", main = "Proportion of Adults an
d Infants Protected", xlab = "VOLUME", ylab ="Proportion",
     type = "l", lwd = 2)
abline(h=0.5)
abline(v = split.infants)


lines(volume.value, prop.adults, col = "blue",
     type = "l", lwd = 2 )
abline(h=0.5)
abline(v = split.adults)

split.infants

## [1] 139.0282

split.adults

## [1] 386.0515

#proportion difference between adults and infants
difference <- ((1-prop.adults) - (1-prop.infants))

#max difference

max(difference)

## [1] 0.5559284

head(difference)

# loess, local polynomial regression fitting
y.loess.a <- loess(1-prop.adults ~ volume.value, span = 0.25, family = c("sym
metric"))

y.loess.i <- loess(1-prop.infants ~ volume.value, span = 0.25, family = c("sy
mmetric"))

smooth.difference <- predict(y.loess.a) - predict(y.loess.i)

#determine volume.value corresponding to maximum of var smooth.difference
which.max(smooth.difference)

## [1] 267

# ?plot(), ?abline(), ?text() to review documentation pages
plot(volume.value, difference, col = "darkred", main = "Difference in Harvest
Proportion",
     type = "l", lwd = 1)

#adds a smoothed spline to plot
#ss <- smooth.spline(volume.value, difference, df = 10 )
```

```
lines(smooth.difference, type = "l", lty = 2)

#Add the Max difference
abline(v = volume.value[which.max(smooth.difference)], col = "blue", lty = 2)

volume.value[which.max(smooth.difference)]

## [1] 268.4922

# The relevant harvest proportions may be found similarly, by passing the ele
ment number of the largest value in smooth.difference as a bracketed index fo
r (1-prop.infants) and (1-prop.adults).

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

## [1] 0.1812081

# [1] 0.1764706


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

## [1] 0.7303523

# Although the relevant volume.value - 207 - is given to you, we can demonstr
ate # 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.

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

## [1] 206.9844

# [1] 206.9844

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

## [1] 350.8333

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

sum(mydata[mydata$TYPE == "I", "VOLUME"] > 206.9844) / sum(mydata$TYPE == "I"
)

## [1] 0.3020134

# [1] 0.2871972

#adults
```

```
sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > 206.9844) / sum(mydata$TYPE == "ADULT")
```

```
## [1] 0.8265583
```

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

```
## [1] 241.7065
```

```
# [1] 237.7383

# The infant and adult harvest proportions can be determined in much the same way
volume.value[which.min(abs(prop.infants - (1-prop.adults)))]
```

```
## [1] 241.7065
```

```
# we calculated proportions for item (8)(a).
sum(mydata[mydata$TYPE == "I", "VOLUME"] > 241.7065) / sum(mydata$TYPE == "I")
```

```
## [1] 0.2248322
```

```
# [1] 0.2871972

#adults
sum(mydata[mydata$TYPE == "ADULT", "VOLUME"] > 241.7065) / sum(mydata$TYPE == "ADULT")
```

```
## [1] 0.7737127
```

```
plot((1-prop.infants),(1-prop.adults),
     xlab = "Infant Harvest proportion",
     ylab = "Adult Harvest Proportion",
     main = "ROC curve of adult and infant harvest proportions",
     col = "darkblue",
     type = "l")
abline(0,1, col = "darkred", lty = 2)

#points()
#text()

require(MESS)
```

```
## Loading required package: MESS
```

```
## Loading required package: geepack
```

```
## Loading required package: geeM
```

```
## Loading required package: Matrix
```

```
auc((1-prop.infants),(1-prop.adults), type = 'spline')
```

```
## [1] 0.8479908
```

Brandon O'Briant
PREDICT 401
Data Analysis Assignment 2

```
auc <- -1*sum(diff((1-prop.infants))*(head((1-prop.adults),-1)+tail((1-prop.a
dults),-1)))/2
auc
```

```
## [1] 0.8481043
```

```
sum(mydata$VOLUME >= volume.value[which.max(smooth.difference)])/ (total.adul
ts + total.infants)
```

```
## [1] 0.5723938
```

```
# [1] 0.5839768
```

```
sum(mydata$VOLUME >= 206.984)/ (total.adults + total.infants)
```

```
## [1] 0.6756757
```

```
sum(mydata$VOLUME >= 253.611)/ (total.adults + total.infants)
```

```
## [1] 0.5936293
```

```
tableOfCutoffs <- matrix(c(268.492, 0.730, 0.181, 0.572,206.984, 0.831, 0.340
, 0.676, 253.611, 0.778, 0.267, 0.593), ncol = 4, byrow = TRUE)
rownames(tableOfCutoffs) <- c("max.difference", "zero.A1.infants", "equal.err
or")
colnames(tableOfCutoffs) <- c("Volume", "TPR", "FPR", "PropYield")
tableOfCutoffs <- as.table(tableOfCutoffs)
tableOfCutoffs
```

```
##                  Volume     TPR     FPR PropYield
## max.difference  268.492   0.730   0.181     0.572
## zero.A1.infants 206.984   0.831   0.340     0.676
## equal.error     253.611   0.778   0.267     0.593
```