Predict 401DL Data Analysis Assignment #2

Data Analysis of Abalone data

Ross, Kari
Predict 401 Section 55
March 13th, 2016

Table of Contents

List of Figures

Introduction

The report provides a summary of the data analysis and modelling performed on sample set from data collected on abalones. The data was obtained for a study, with the study goal to ascertain if the age of abalones can be estimated from physical measurements of the abalones.

This report explores the use of VOLUME, CLASS, and SEX as predictors of SHUCK was explored. Additionally, the independence of CLASS and SEX was explored, as well as multiple cutoff points for abalone harvesting was explored, with the intent to explore methods for managing the abalone harvest.

This report does not draw conclusions the success of the study but presents exploratory the data from the study. This report is a follow onto a previous report, entitled "Predict401DL Data Analysis Assignment #1: Summary of Exploratory Data Analysis of Abalone data".

Results

Part 1: Hypothesis testing of Independence of SHUCK and VOLUME using Person Chi Square Statistic

In order to explore the independence of SHUCK and VOLUME, the chi-square test for independence was used.

The test procedure was as follows:

Hypothesis:

Ho: SHUCK and VOLUME are independent

Ha: SHUCK and VOLUME are not independent

Analysis Plan:

The significance level was assumed to be 0.05

A function was created to calculate the chi-square value and the p-value of the 2x2 contingency table of factors of the SHUCK value and the VOLUME of the sample data set. The contingency table was created by using a factor of values above and below median.

The chi-square test of independence was valid because the method is simple random sampling, the variables were transformed to be categorical, and the expected frequency was greater than 5 for the contingency table.

Results:

The results are show below in Figure 1:

```
chi(shuck_volume))
chi-squared: 323.2132

chi(shuck_volume))

p-value:
2.89077e-72
```

Figure 1: Chi-Square Test of Independence Results

Interpret Results:

Since the P-value (2.89077e-27) is less than the significance level we must reject the null hypothesis. Thus we can conclude that there is a relationship between median values of SHUCK and VOLUME.

Part 2 : Analysis of Variance on SHUCK using CLASS and SEX, with and without interaction

An two-way analysis of variance was performed on the data set, to determine if CLASS and SEX are explanatory variables for SHUCK.

The analysis was performed with several models

1. Analysis was performed using a model for the interaction of CLASS and SEX.
2. Analysis was performed using a model where CLASS and SEX interaction was not considered
3. Lastly, analysis was performed using a model the TUKEYHSD
    a. TUKEYHSD was used to provide a test of the pairwise interactions between the groups, as the previous steps will determine if the interactions are significant, but doesn't differential which variables differ from each other

The results are shown below in Figures 2,3 and 4:

```
> summary(SHUCK_class_sex_interaction)
             Df Sum Sq Mean Sq F value  Pr(>F)
CLASS         5   7.15   1.430   48.54 < 2e-16 ***
SEX           2   1.91   0.957   32.48 5.9e-14 ***
CLASS:SEX    10   0.19   0.019    0.64    0.78
Residuals   482  14.20   0.029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 2: Impact on SHUCK with CLASS and SEX interaction

```
> summary(SHUCK_class_sex_no_interaction)
             Df Sum Sq Mean Sq F value  Pr(>F)
CLASS         5   7.15   1.430    48.9 < 2e-16 ***
SEX           2   1.91   0.957    32.7 4.6e-14 ***
Residuals   492  14.39   0.029
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3: Impact on SHUCK with CLASS and SEX not interacting

```
> TukeyHSD(SHUCK_class_sex)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = SHUCK ~ CLASS + SEX, data = my_data)

$CLASS
             diff        lwr       upr     p adj
A2-A1  0.14672655  0.0631551 0.2302980 0.000011
A3-A1  0.31732439  0.2383018 0.3963470 0.000000
A4-A1  0.39268524  0.3059971 0.4793734 0.000000
A5-A1  0.29959184  0.1913184 0.4078653 0.000000
A6-A1  0.32323469  0.2149612 0.4315082 0.000000
A3-A2  0.17059784  0.1117806 0.2294151 0.000000
A4-A2  0.24595869  0.1771856 0.3147317 0.000000
A5-A2  0.15286529  0.0583241 0.2474065 0.000070
A6-A2  0.17650815  0.0819669 0.2710494 0.000002
A4-A3  0.07536086  0.0121935 0.1385283 0.009037
A5-A3 -0.01773255 -0.1082777 0.0728126 0.993454
A6-A3  0.00591031 -0.0846349 0.0964555 0.999969
A5-A4 -0.09309341 -0.1904006 0.0042138 0.069921
A6-A4 -0.06945055 -0.1667578 0.0278567 0.320116
A6-A5  0.02364286 -0.0933058 0.1405916 0.992413

$SEX
          diff        lwr         upr    p adj
I-F -0.1304676 -0.1761348 -0.08480038 0.000000
M-F -0.0340442 -0.0774010  0.00931264 0.155889
M-I  0.0964234  0.0527574  0.14008945 0.000001
```

Figure 4: Tukey HSD pairwise group comparison

By examination of Figure 2, the p-values show that the interaction between CLASS and SEX is not significant and therefore can be dropped in further analysis. In other words, male and female abalones can be combined into a single category, ADULT, for further investigation.

The Tukey pairwise group comparison shows that there is less interaction with lower classes than the upper classes, as noted by A5-A3, A5-A4, A6-A3, and A5-A6.

Part 3 : SHUCK versus VOLUME Scatterplot Exploration

Shuck was plotted versus volume, both in linear and log scale. The results are shown below in Figure 5.

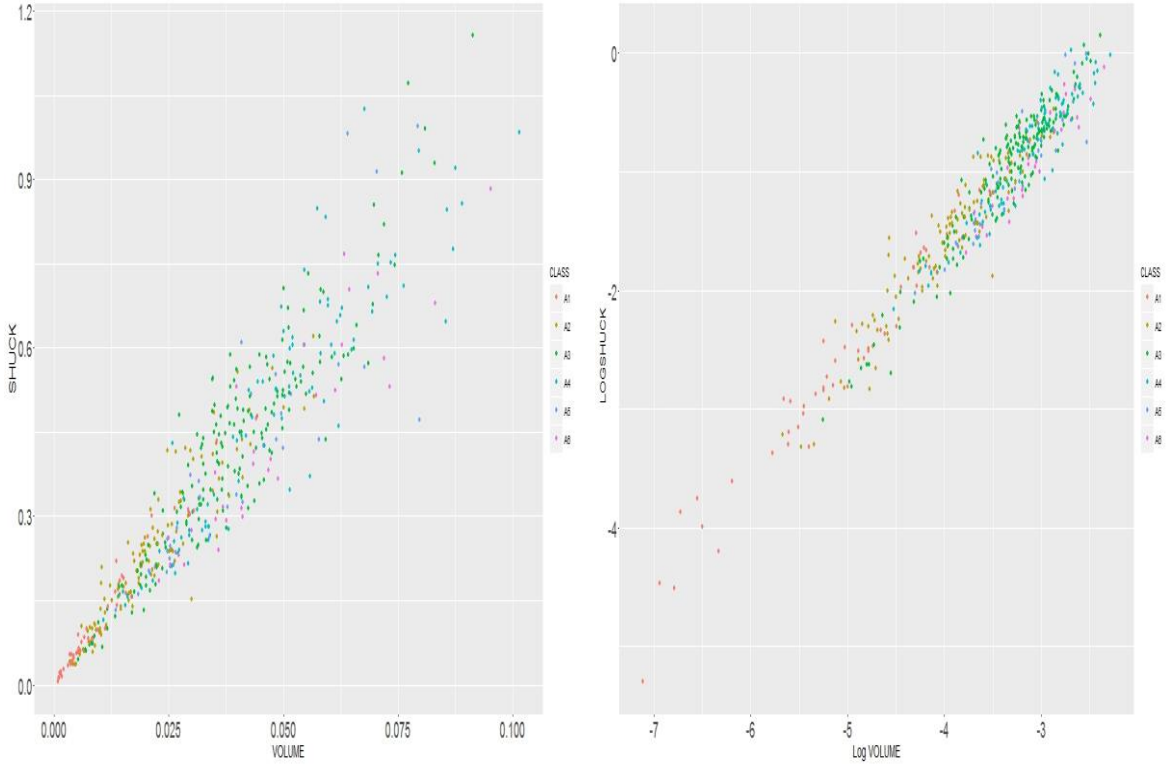

Figure 5: SHUCK versus VOLUME plots

Exploration of the SHUCK versus Volume plots show that as Volume increases, the class increases. For instance, lower class categories, A1, have a lower volume and lower shuck value. Higher classes tend to have higher Volume and Shuck values.

Part 4 : Regression of SHUCK on VOLUME, SEX, and CLASS

In order to ascertain if it is possible to predict SHUCK based on VOLUME, SEX, and CLASS, a regression analysis was performed, with SHUCK as the dependent variable, with VOLUME, SEX, and CLASS as explanatory variables.

A simple linear regression was performed on the log transformed SHUCK and VOLUME. The summary of the regression results are shown in Figure 6.

```
> summary(log_shuck_LM)

Call:
lm(formula = L_SHUCK ~ L_VOLUME + CLASS + SEX, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.77497 -0.11656 -0.00626  0.11160  0.62489

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.56468    0.08012  32.009  < 2e-16 ***
L_VOLUME     1.02839    0.01585  64.890  < 2e-16 ***
CLASSA2     -0.06531    0.03583  -1.823  0.06898 .
CLASSA3     -0.12701    0.03961  -3.207  0.00143 **
CLASSA4     -0.18302    0.04418  -4.143 4.04e-05 ***
CLASSA5     -0.21944    0.04946  -4.436 1.13e-05 ***
CLASSA6     -0.27904    0.05029  -5.549 4.71e-08 ***
SEXI         0.01564    0.02551   0.613  0.54013
SEXM         0.02665    0.02029   1.313  0.18979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1861 on 491 degrees of freedom
Multiple R-squared:  0.9475,    Adjusted R-squared:  0.9466
F-statistic:  1107 on 8 and 491 DF,  p-value: < 2.2e-16
```

Figure 6: Linear Regression for SHUCK predicted by Volume, Class, and Sex

The analysis was performed using log values of both SHUCK and VOLUME in order to better fit a model for prediction.

Assuming a significance level of 0.05,From the Pr(>|t|) column, it is shown that

CLASSA3 has a significance level of less than 0.01

CLASSA4, CLASSA5, CLASSA6 has a significance level of less than 0.001

CLASSSA2 has a significance level of 0.05.

This shows that there is significance relationship between the variables CLASS and VOLUME.

Additionally, the coefficient estimate shows a negative relationship for CLASS.

However, the model shows that SEXI and SEXM have a significance level of 0.54 (for INFANT) and 0.18979. This shows that SEX is not an important predictor in this regression model.

The Multiple R-squared statistic indicates that model accounts for 94.75% of the variance in the SHUCK.

Part 5 : Analysis of Residuals of Regression Model

In order to determine the difference between the observed data of the dependent variable SHUCK and the fitted values from the Regression model of SHUCK, a residual plot and analysis was performed.

Figure 7 shows the Histogram of the plot of residuals.

Figure 8 shows the Normal Q-Q plot of the Residuals

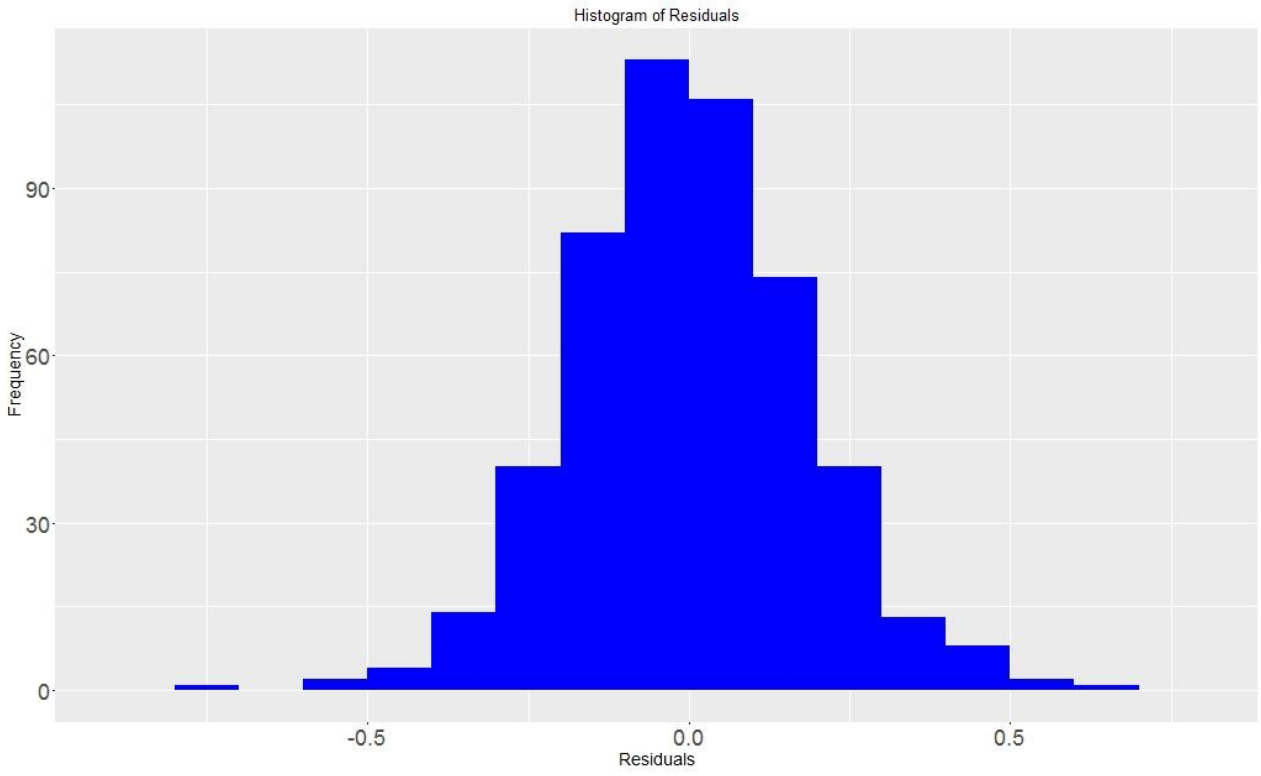

Figure 7: Histogram of Residuals

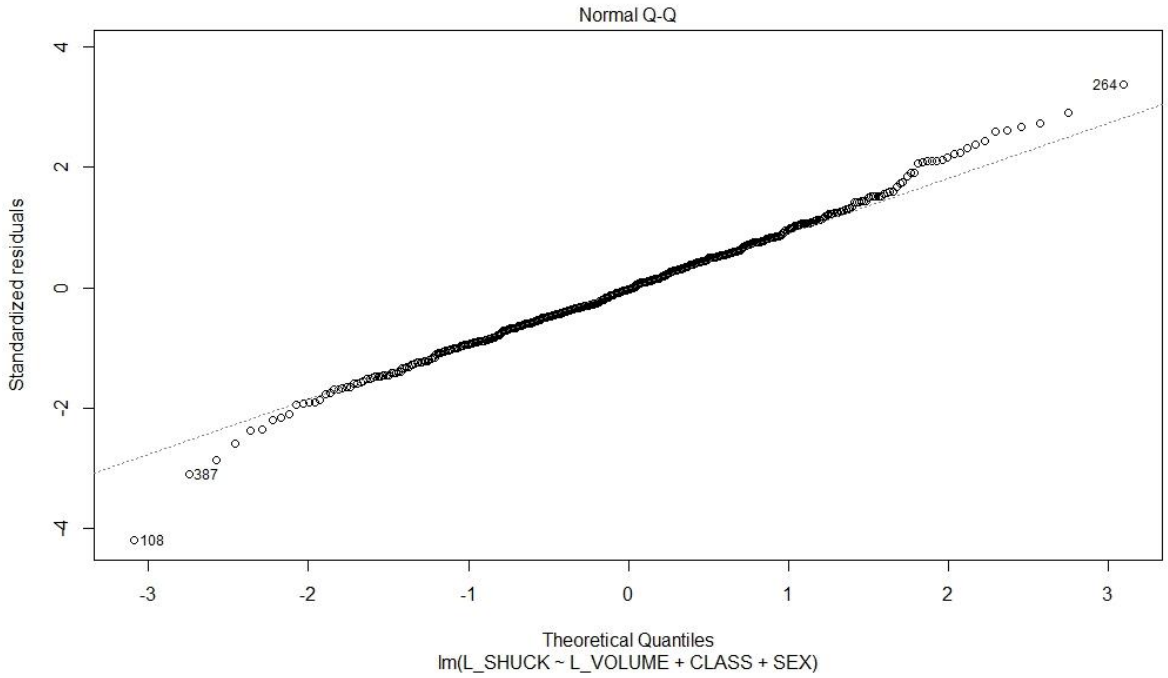

Figure 8: Q-Q Plot

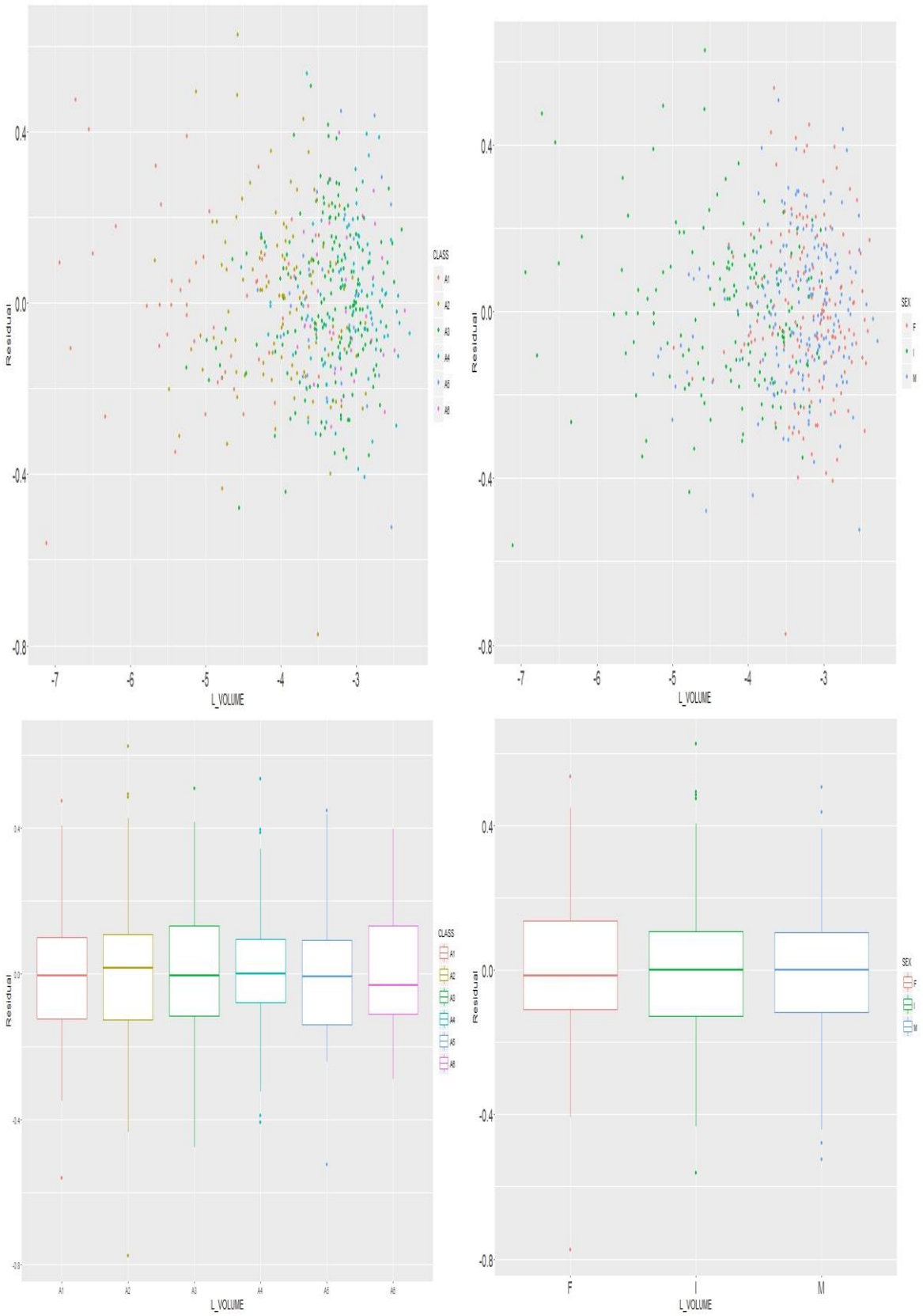

Figure 9: Scatter & Box Plots of Residuals

Examination of Figure 8 , Q-Q plots show that normality assumption is violated after +/- 2 standard deviations. The tails drift away from the normal line, indicating that there is not a good fit of the model throughout all data.

Examination of Figure 9 Scatter Plots show that there is some relationship between the Volume and both SEX and individual class, as the pattern in the residuals is randomly dispersed.

Part 6 : Cutoff for 50% harvest points in Infants and Adults

In order to manage the tradeoff faced in managing the abalone harvest, plots were created where the proportions of Infants and Adults were plotted at a 50% mark.

Figure 10 shows the data for the proportions of infants and proportions of adults

```
> prop.infants
  [1] 0.0457516 0.0522876 0.0915033 0.1372549 0.1633987 0.1960784
  [7] 0.2287582 0.2614379 0.2875817 0.3267974 0.3594771 0.3725490
 [13] 0.3986928 0.4575163 0.4901961 0.5098039 0.5490196 0.5947712
 [19] 0.6274510 0.6470588 0.6797386 0.7189542 0.7254902 0.7450980
 [25] 0.7647059 0.7777778 0.7973856 0.8235294 0.8431373 0.8496732
 [31] 0.8627451 0.8823529 0.8823529 0.9019608 0.9150327 0.9215686
 [37] 0.9281046 0.9281046 0.9346405 0.9542484 0.9607843 0.9738562
 [43] 0.9738562 0.9738562 0.9803922 0.9869281 0.9934641 1.0000000
 [49] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [55] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [61] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [67] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [73] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [79] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [85] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [91] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [97] 1.0000000 1.0000000 1.0000000 1.0000000
```

```
> prop.adults
  [1] 0.00000000 0.00000000 0.00000000 0.00000000 0.00864553 0.01152738
  [7] 0.02017291 0.03458213 0.04034582 0.04610951 0.05187320 0.05187320
 [13] 0.06051873 0.06628242 0.06916427 0.07780980 0.08069164 0.09221902
 [19] 0.10951009 0.12680115 0.13832853 0.15273775 0.16426513 0.17867435
 [25] 0.20172911 0.22766571 0.24783862 0.25936599 0.28242075 0.29971182
 [31] 0.31988473 0.34293948 0.36599424 0.38616715 0.42074928 0.44092219
 [37] 0.45821326 0.47838617 0.51008646 0.53602305 0.55043228 0.57348703
 [43] 0.59365994 0.60518732 0.61959654 0.63976945 0.65706052 0.67435159
 [49] 0.70605187 0.72046110 0.74927954 0.75504323 0.76080692 0.78097983
 [55] 0.78962536 0.80403458 0.82420749 0.83573487 0.84149856 0.84726225
 [61] 0.86167147 0.87608069 0.88472622 0.89048991 0.89913545 0.89913545
 [67] 0.90489914 0.91066282 0.91930836 0.93083573 0.93659942 0.94236311
 [73] 0.94524496 0.95100865 0.95677233 0.95965418 0.95965418 0.96253602
 [79] 0.96829971 0.97118156 0.97118156 0.97694524 0.97694524 0.97694524
 [85] 0.98270893 0.98559078 0.98847262 0.99135447 0.99135447 0.99423631
 [91] 0.99423631 0.99423631 0.99423631 0.99711816 0.99711816 0.99711816
 [97] 0.99711816 0.99711816 0.99711816 1.00000000
>
```

Figure 10: Proportions of Adults and Infants

Figure 11 shows the results of protecting half the abalone harvest

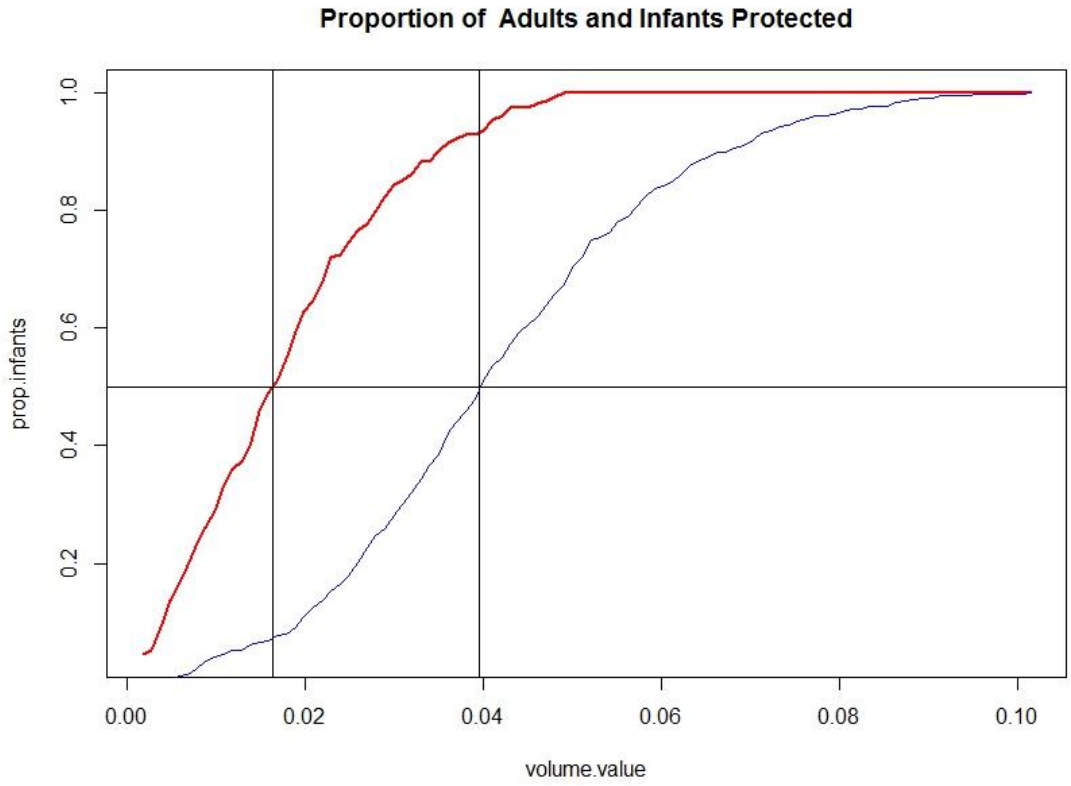

Figure 11: Proportion of Adults and Infants protected

Part 7 : Cutoff for 50% harvest points in Infants and Adults

The maximum value for the volume which corresponds to the observed difference in the maximum difference of harvest percentages of adults and infants was determined. The graphical result is show in Figure 12, below.

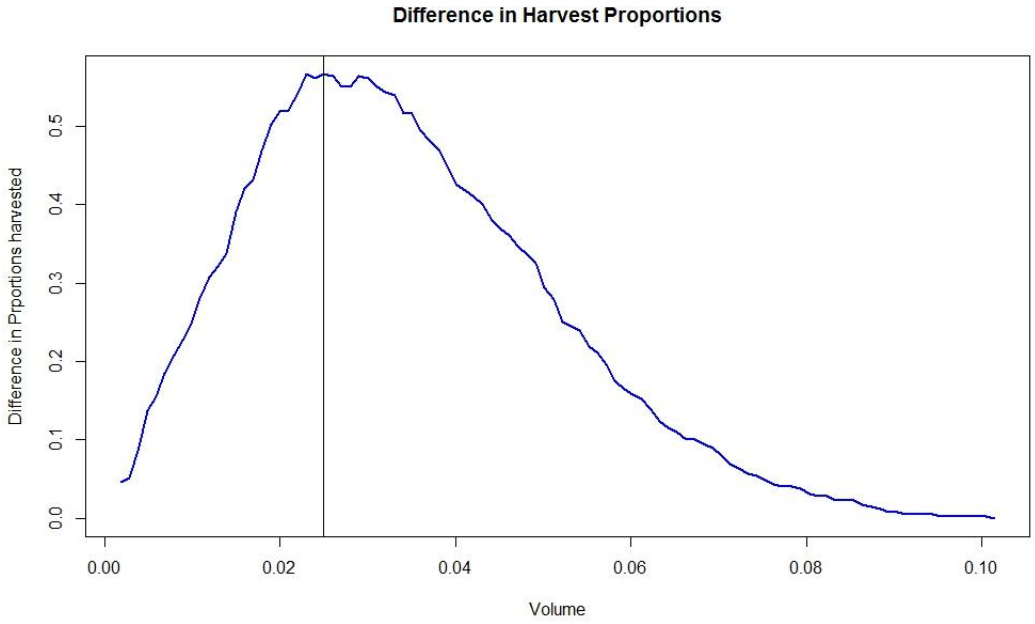

Figure 12: Difference in Harvest Populations

The maximum value difference in population was determined to be 0.566, which corresponds to a Volume Value of 0.02498

Part 8 : ROC curve

A ROC curve was created to illustrate the tradeoff's involved in the decision rules. The ROC is shown below in Figure 13.

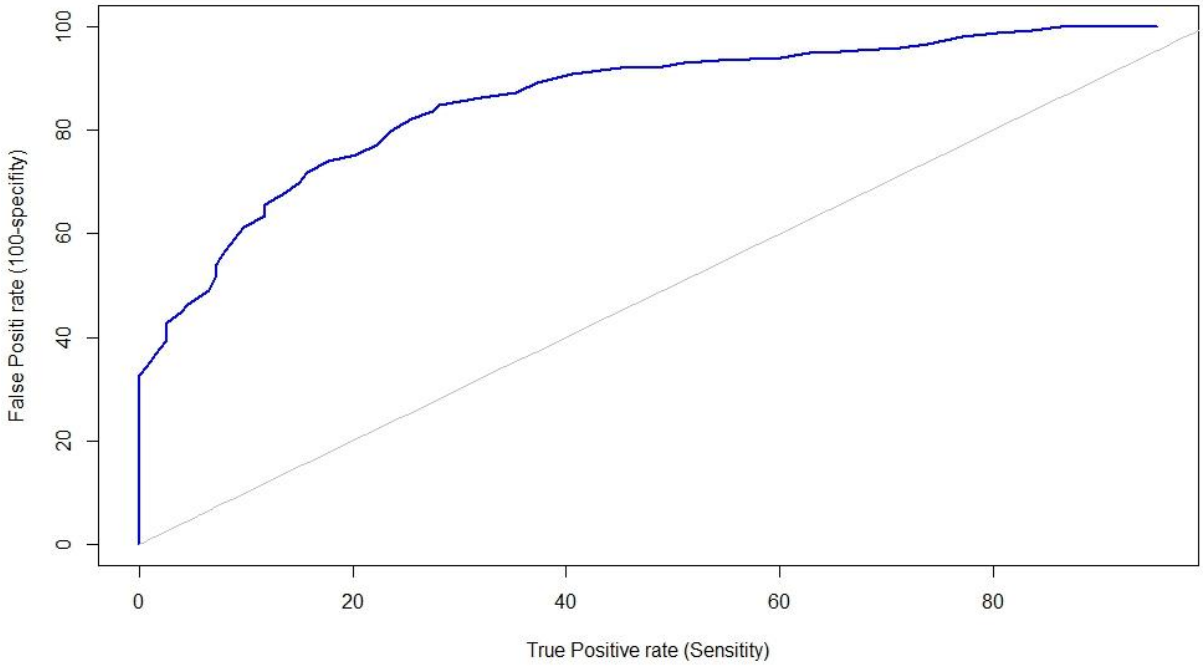

Figure 13: ROC Curve

The ROC curve shows that the cutoff rate around 25 - 30 will reduce the number of false positives (the infant is being treated as an adult), which corresponds to the value show in Figure 12, the maximum difference between volumes of infants and adults, although the ROC curves shows a slightly higher value of volume than the difference in population proportions

Part 9 : Reduction of harvesting of infants in Class A1 and A2

Code was utilized to perform cutoff analysis of Volumes for A1 and A2, to reduce the number of infants in included in the harvesting. The data showed that for A1, the volume is 0.02 and for A2, the volume is 0.034. This corresponds to the values found in Part 7 and Part 8.

Part 10 : Summary of Cutoffs determined in Study

	cutoff volume	A1	A2
1	0.036	0.000000	0.000000
2	0.035	0.000000	0.000000
3	0.034	0.000000	0.008772
4	0.033	0.000000	0.008772
5	0.032	0.000000	0.008772
6	0.031	0.000000	0.008772
7	0.030	0.000000	0.008772
8	0.029	0.000000	0.017544
9	0.028	0.000000	0.026316
10	0.027	0.000000	0.035088
11	0.026	0.000000	0.035088
12	0.025	0.000000	0.052632
13	0.024	0.000000	0.061404
14	0.023	0.000000	0.070175
15	0.022	0.000000	0.096491
16	0.021	0.000000	0.131579
17	0.020	0.020408	0.140351
18	0.019	0.020408	0.175439
19	0.018	0.020408	0.201754
20	**0.017**	**0.020408**	**0.236842**

Figure 14: Summary of Cutoff Points for CLASS A1 & A2

Conclusions

Based on the results of the abalone data analysis and modelling, there is some opportunity for using physical measurements as method for abalone harvesting controls and population management, although it mostly limited to reducing harvesting of infants in CLASS A1 and CLASS A2, with use of VOLUME measurements.

It was shown that there is little interaction between CLASS and SEX, and therefore could be dropped . It was also shown SHUCK and VOLUME are related.

Linear regression, ANOVA, and ROC all showed that a subset of CLASS does show a correlation with physical measurements. For instance, class A4,A5,and A6 showed a correlation with Volume. This correlation was less significant for CLASS A2 and A3 and not significant for Class A1.

Investigations on reducing the harvesting of A1 and A2 by using VOLUME as an indicator is underway and shows some promise of being able to be reduced. The analysis of residuals showed that it is possible to separate out Infants in CLASS A1 and CLASS A2 from adults the more readily than with CLASS A3,A4,and A5.

There are additional issues with using the data in this report for broad statements on abalone harvesting by use of physical characteristics. One of the major issues is that there is no control for confounding and nuisance variables. An additional study would need to be performed in order to verify the findings of this study.

Appendices

Appendix 1: Source of Original Data and Sample Selection

The data for this report was obtained by taking a sampling of a larger data set. The original data for this report was downloaded from the Predict401 Section XX course site and entitled "abalone.csv". The sample was chosen randomly using R, with known seed, obtained by the command of set.seed(123) and the sample function.

For further details on the code used to create the report, see *Appendix 3: R source code , (page 16).*

Appendix 2: Resources for Supplemental Information on Abalones

Additional information on abalones may be obtained from the following sources:

http://www.dpi.nsw.gov.au/__data/assets/pdf_file/009/375858/BlacklipAbalone.pdf
http://www.fishtech.com/facts.html
http://www.marinebio.net/marinesscience/06future/abitro.htm

Appendix 3: R source code

```
#############################################################################
# Kari Ross
# PREDICT 401, Data Analysis Assignment 2
# 2016 March
#############################################################################

#############################################################################
# Packages and Libraries Used
# install.packages("moments")
# install.packages("ggplot2")
#############################################################################
require(moments)
require(ggplot2)
closeAllConnections()

#############################################################################
# FUNCTIONS USED
#############################################################################
```

```
create_file_keep_1_old_file <- function( file_to_write) {
   if ( file.exists( file_to_write ) == FALSE) {
      file.create( file_to_write )
   } else {
      file.copy( file_to_write, paste(file_to_write, "_old.csv", sep="" ) , overwrite = TRUE)
      file.remove( file_to_write )
      file.create( file_to_write  )
      }
}

chi <- function(input_table, result) {
   options( digits = 6 )
   expected_freq <- matrix(0 , nrow = (nrow( input_table ) - 1), ncol = (ncol( input_table) - 1) )
   rows <- nrow(input_table)
   cols  <- ncol(input_table )
   chi_sq_value <- 0
   for (i in 1:(rows-1) ) {
      for ( j in 1:(cols-1) )  {
         expected_freq[i,j] <- input_table[i,cols] * input_table[rows,j] / input_table[rows,cols]
         chi_sq_value <- chi_sq_value + ( input_table[i,j] - expected_freq[i,j] ) ^ 2 / expected_freq[i,j]
      } # end j loop
   } # end i loop

   p_stat   <-pchisq(chi_sq_value, df=1, lower.tail=F)


   if (result == "p_value") {
      cat(sprintf("\nchi(shuck_volume)\n"))
      cat(sprintf("p-value:\n"))
      cat(p_stat)
      return(p_stat)
   }
   else {
      cat(sprintf("\nchi(shuck_volume)\n"))
      cat(sprintf("chi-squared: %.4f\n", chi_sq_value))
      return(chi_sq_value)
   }
} # end function chi()

#############################################################################
# Cleanup RStudio Environment before Analysis starts
# Set Paths
# Set Seed
 # setwd("C:/Users/Kari/source_code/PREDICT/401_stats/DataAnalysis/DA2/src")
#############################################################################
# remove variables, keep functions
rm(list=setdiff(ls(), lsf.str()))

## Set Debug Flags & printing to screen flags
PRINT_TO_FILE      <- FALSE
PRINT_TO_SCREEN <- TRUE

# setup paths and seed
input_file     <- "project1_mydata.csv"
base_path    <- "C:/Users/Kari/source_code/PREDICT/401_stats/DataAnalysis/"
data_path     <- paste( base_path ,     "output/" ,          sep = "" )
output_path <- paste( base_path ,     "DA2/output/",       sep = "" )
output_csv   <- paste ( output_path , "project2_mydata.csv" , sep = "" )
output_log    <- paste ( output_path , "project2_log.txt" ,         sep = "" )

create_file_keep_1_old_file( output_log )
create_file_keep_1_old_file( output_csv )

my_data   <- read.csv(paste(data_path, input_file,  sep= ""), header=TRUE)

#############################################################################
# STEP 1
#############################################################################
# Chi-Square Test of Independence (Chapture 16, page 676)
```

```
shuck   <- factor( my_data$SHUCK   > median( my_data$SHUCK   ), labels = c("below","above") )
volume <- factor( my_data$VOLUME > median( my_data$VOLUME ), labels = c("below","above") )
shuck_volume <- addmargins(table(shuck, volume) )

# calculate expected frequecies
test_statistic         <- chi(shuck_volume, "stat")
test_statistic_p_val <- chi(shuck_volume, result = "p_value")
chi_function           <- chisq.test(shuck_volume[1:2, 1:2], correct = F)

#############################################################################
# STEP 2 - 2way ANOVA
#############################################################################
# Two-way factoria ANOVA
SHUCK_class_sex_interaction <- aov( SHUCK ~ CLASS * SEX, data = my_data)

# ANOVA without CLASS*SEX interaction
SHUCK_class_sex_no_interaction <- aov(SHUCK ~ CLASS + SEX, data = my_data)

TukeyHSD(SHUCK_class_sex_no_interaction)
summary(SHUCK_class_sex_interaction)
summary(SHUCK_class_sex_no_interaction)

#############################################################################
# STEP 3 -  Create Scatterplots
#############################################################################
# add natural logs to my_data (for use later)

# note: function "1og"  is natural log, and "log10" is the log base 10
my_data <- transform( my_data,
                      L_VOLUME = log(VOLUME))
my_data <- transform( my_data,
                      L_SHUCK = log(SHUCK))

# Labels for Graphs
#plot_a_label <- paste("Diff Bit Count: ",text_label,sep="")

# Create gaphs
g_3a  <- ggplot(data= my_data, aes(x=VOLUME, y=SHUCK, colour = CLASS) ) + geom_point()
g_3a <- g_3a + labs(x = "VOLUME", y = "SHUCK")
g_3a <- g_3a + theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 16))
g_3a <- g_3a + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )


g_3b  <- ggplot(data= my_data, aes(x=L_VOLUME, y=L_SHUCK, colour = CLASS) ) + geom_point()
g_3b <- g_3b + labs(x = "Log VOLUME", y = " LOGSHUCK")
g_3b <- g_3b +theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 14))
g_3b <- g_3b + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

#jpeg(file= paste(output_path,"part3.jpg",sep =""), width= 1200, height = 800, quality=100)
#    multiplot(g_3a, g_3b, cols=2)
#dev.off()

#############################################################################
# STEP 4 -  Perform Linear Regression
#############################################################################
# outcome variable (what is being predicted) is called the response, also called dependent
# Input Variable( what we are using to predict) is called the predictor.
log_shuck_LM <- lm(L_SHUCK ~ L_VOLUME+CLASS+SEX, data=my_data)
summary(log_shuck_LM)

#############################################################################
# STEP 5 - Analysis of the residuals of Regression
#############################################################################
log_shuck_LM$residuals
g_5a <- ggplot(log_shuck_LM, aes(x = .resid)) + geom_histogram(binwidth=.1, fill = "blue")
g_5a <- g_5a +ggtitle("Histogram of Residuals")
g_5a <- g_5a + labs(x="Residuals", y="Frequency")
```

```
g_5a <- g_5a +theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 14))
g_5a <- g_5a + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

g_5b <- ggplot(log_shuck_LM, aes=(x = .resid)) + stat_qq() + geom_abline()
g_5b <- g_5b +theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 14))
g_5b <- g_5b + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

g_5c <- ggplot(log_shuck_LM, aes(L_VOLUME, y=log_shuck_LM$residuals)) +geom_point(aes(color=CLASS))
g_5c <- g_5c + labs(x="L_VOLUME", y="Residual")
g_5c <- g_5c +theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 14))
g_5c <- g_5c + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

g_5d <- ggplot(log_shuck_LM, aes(L_VOLUME, y=log_shuck_LM$residuals)) +geom_point(aes(color=SEX))
g_5d <- g_5d + labs(x="L_VOLUME", y="Residual")
g_5d <- g_5d +theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 14))
g_5d <- g_5d + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

g_5e <- ggplot(log_shuck_LM, aes(CLASS, y=log_shuck_LM$residuals)) +geom_boxplot(aes(color=CLASS))
g_5e <- g_5e + labs(x="L_VOLUME", y="Residual")
g_5e <- g_5e +theme(axis.title.x=element_text(colour="black", size=14),
axis.title.y=element_text(colour="black", size = 14))
g_53 <- g_5e + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

 g_5f <- ggplot(log_shuck_LM, aes(SEX, y=log_shuck_LM$residuals)) +geom_boxplot(aes(color=SEX))
 g_5f <- g_5f + labs(x="L_VOLUME", y="Residual")
g_5f <- g_5f +theme(axis.title.x=element_text(colour="black", size=14), axis.title.y=element_text(colour="black",
size = 14))
g_5f <- g_5f + theme(axis.text.x = element_text(size =rel(2)), axis.text.y = element_text(size = rel(2) ) )

#############################################################################
# STEP 6 - Decision Rule, based on Volume
#############################################################################

idxi  <- my_data[,2]== "I"
idxf  <- my_data[,2]== "F"
idxm <- my_data[,2] == "M"

idxa  <-  idxf | idxm

max.v <- max(my_data$VOLUME)
min.v <- min(my_data$VOLUME)
delta  <- (max.v - min.v) / 100

## Set minimum values
prop.infants   <- numeric(0)
prop.adults    <- numeric(0)
volume.value <- numeric(0)

total_infants   <- length(my_data[idxi,1])  # this value must be  changed for adults
total_adults     <- length(my_data[idxa,1])
for (k in 1:100) {
 value                 <- min.v + k*delta
  volume.value[k] <- value
  prop.infants[k] <- sum(my_data$VOLUME[idxi] <= value ) / total_infants
  prop.adults[k]  <- sum(my_data$VOLUME[idxa] <= value ) / total_adults

}

## prop.infants shows the impact of increasing the volume cutoff for harvesting
## The following code shows how to split the population at at 50% harvest of infants
n.infants <- sum(prop.infants <= 0.5)
n.adults  <- sum(prop.adults  <= 0.5)
split.infants <- min.v + (n.infants + 0.5)*delta # this estimates the desired volume
split.adults <- min.v + (n.adults + 0.5)*delta
```

```
plot(volume.value, prop.infants, col="red", main = "Proportion of  Adults and Infants Protected", type = "l",
lwd=2)
lines(volume.value, prop.adults, col="blue")
abline(h=0.5)
abline(v=split.infants)
abline(v=split.adults)
legend(2000,9.5, c("Infants", "Adults"), lty=c(1,1), col=c("red", "blue") )

#############################################################################
# STEP 7 - Difference in Harvest Proportions
#############################################################################
adult_harvested  <- 1 - prop.adults
infant_harvested <- 1 - prop.infants
delta_harvested <- numeric(0)
for (k in 1:100) {
   delta[k] <- adult_harvested[k] - infant_harvested[k]
}

line_position <- volume.value[which.max(delta)]
plot(volume.value, delta, type = "l", lwd=2, col="blue", xlab = "Volume", ylab="Difference in Prportions
harvested", main="Difference in Harvest Proportions")

abline(v= line_position)
text(1,0,"volume =")


#############################################################################
# STEP 8 - ROC Curve
#############################################################################
plot(infant_harvested*100, adult_harvested*100,  type = "l", lwd=2, col="blue", xlab="True Positive rate
(Sensitity)", ylab= "False Positi rate (100-specifity)")
lines(adult_harvested*100,adult_harvested*100, col="grey")

#############################################################################
# STEP 9 - Cutoff points for A1 and A2
#############################################################################
cutoff         <- 0.036
A1.sum      <- numeric(0)
A2.sum      <- numeric(0)
cutoff_values      <- numeric(0)


for (k in 1:20) {
   cutoff_values[k] <- cutoff
   index.A1   <- (my_data$CLASS == "A1")
   indexi       <- index.A1 & idxi
   A1.sum[k] <- sum(my_data[indexi,12] >=cutoff) / sum(index.A1)
   index.A2   <- (my_data$CLASS == "A2")
   indexi       <- index.A2 & idxi
   A2.sum[k] <- sum(my_data[indexi,12] >=cutoff) / sum(index.A2)
   cutoff <- cutoff - 0.001
}
cells <- c(cutoff_values, A1.sum, A2.sum)
cnames <- c("cutoff volume", "A1", "A2")
rnames <- c(1:20)
summary_cutoff <- matrix(cells, ncol=3, byrow=FALSE, dimnames=list(rnames,cnames))

            write.csv(x=summary_cutoff, file ="cutoff_point.csv")

#############################################################################
# Create Files
#############################################################################
if (PRINT_TO_SCREEN) {
    print(g_3a)
    print(g_3b)
   print(g_5a)
   print(g_5b)
   print(g_5c)
```

```
    print(g_5d)
    print(g_5e)
    print(g_5f)
}


if (PRINT_TO_FILE) {
    jpeg(file= paste(output_path,"part3.jpg",sep =""), width= 1200, height = 800, quality=100)
        multiplot(g_3a, g_3b, cols=2)
    dev.off()

    jpeg(file= paste(output_path,"part5a.jpg",sep =""), width= 1200, height = 800, quality=100)
        multiplot(g_5a, g_5b, cols=1)
    dev.off()

}
```

Appendix 4: Software Version and Computer System

```
> R.version()
               _
platform       x86_64-w64-mingw32
arch           x86_64
os             mingw32
system         x86_64, mingw32
status
major          3
minor          2.3
year           2015
month          12
day            10
svn rev        69752
language       R
version.string R version 3.2.3 (2015-12-10)

nickname       Wooden Christmas-Tree

> Sys.info()
  sysname       release       version      nodename       machine
"Windows"    ">= 8 x64"   "build 9200"     "KARI_PC"      "x86-64"
```