Data Analysis Project #2 due at the end of Session 10 (75 points)

Overview

The primary objective of this assignment is to devise and evaluate binary decision rules for harvesting abalones. This will be based upon an investigation of the variables RATIO and VOLUME. **Use your saved mydata file from the first assignment for this assignment.** Snippets of code are supplied in the assignment. This report should comply with the report template. Results from the first assignment may be referenced as needed, but need not be included or reproduced.

Before starting, review the Data Analysis Report Example, the Report Grading document, the Data Analysis Video #2 and the self-check page all of which are posted in the data analysis module. A thorough review of these materials will help prepare you for the assignment. It will be necessary to install and load the rockchalk package on your machine. It is permissible to use either base R or ggplot2 for graphical displays. If you choose to use the latter, you will need to install and load ggplot2 and gridExtra on your machine.

Project Assignment 2 (75 points due the end of Session 10)

(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.). Discuss. Do these data come from a normal distribution?

```
# ?hist(), ?qqnorm(), ?qqline() to review documentation pages
library(rockchalk)
```

(1)(b) Transform RATIO using log10() to create L_RATIO (see Kabacoff Section 8.5.2, p. 199-200.). Form a histogram and QQ plots using L_RATIO. Calculate the skewness and kurtosis. Display boxplots of L_RATIO differentiated by CLASS.

```
mydata$L_RATIO <- log10(mydata$RATIO)
```

(1)(c) Test the homogeneity of variance across classes using the bartlett.test() (see Kabacoff Section 9.3.2, p. 222.). Comment on your findings. Based on the bartlett.test(), is it reasonable to assume a normal distribution for L_RATIO with homogeneous variances across classes?

```
# 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() to review documentation page
```

(2)(a) Perform an analysis of variance with aov() on L_RATIO using CLASS and SEX as the independent variables (review Kabacoff chapter 9, p. 212-229). 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. Compare the two analyses (see Kabacoff chapter 9, p. 227).

```
# aov(), example
data(mtcars)

cars_anova <- aov(mpg ~ factor(cyl)*factor(am), data = mtcars) # w/ interaction term
summary(cars_anova)

##                        Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(cyl)             2  824.8   412.4  44.852 3.73e-09 ***
## factor(am)              1   36.8    36.8   3.999   0.0561 .  
## factor(cyl):factor(am)  2   25.4    12.7   1.383   0.2686    
## Residuals              26  239.1     9.2                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

cars_anova2 <- aov(mpg ~ factor(cyl) + factor(am), data = mtcars) # w/o interaction term
summary(cars_anova2)

##             Df Sum Sq Mean Sq F value   Pr(>F)    
## factor(cyl)  2  824.8   412.4  43.657 2.48e-09 ***
## factor(am)   1   36.8    36.8   3.892   0.0585 .  
## Residuals   28  264.5     9.4                     
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(2)(b) For the model without CLASS:SEX, obtain multiple comparisons with the TukeyHSD() function. Interpret the results (TukeyHSD() will adjust for unequal sample sizes). Comment on the results. Interpret the trend across classes. Do these results suggest male and female abalones can be combined into a single category labeled as 'adults?' If not, why not?

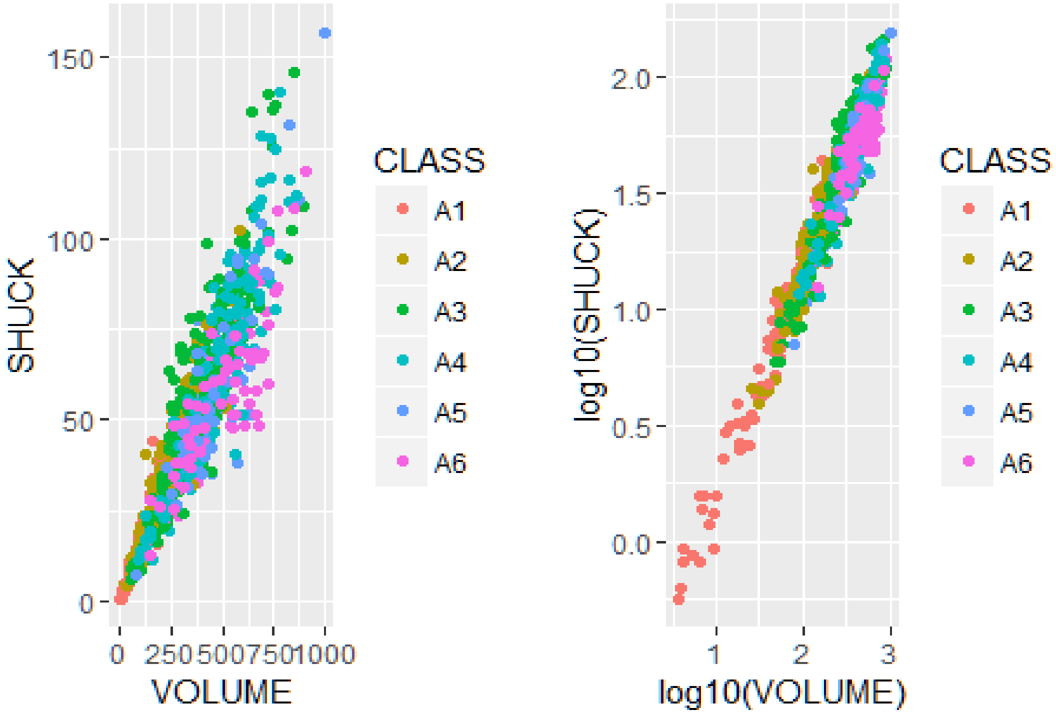
```
TukeyHSD(cars_anova2) # Tukey's Honest Significant Difference test

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
## 
## Fit: aov(formula = mpg ~ factor(cyl) + factor(am), data = mtcars)
## 
## $`factor(cyl)`
##           diff        lwr       upr     p adj
## 6-4  -6.920779 -10.597684 -3.243875 0.0002034
## 8-4 -11.563636 -14.627723 -8.499549 0.0000000
## 8-6  -4.642857  -8.163225 -1.122489 0.0079071
## 
## $`factor(am)`
##         diff       lwr      upr     p adj
## 1-0 1.860708 -0.405361 4.126776 0.1036939
```

(3)(a) Use combineLevels() from the rockchalk package to combine "M" and "F" into a new level "ADULT". This will necessitate defining a new variable, TYPE, in mydata which will have two levels" "I" and "ADULT". Use par() to form two histograms using VOLUME. One would display infant volumes, and the other ADULT volumes. Compare the histograms and discuss the implications regarding separation of infants from adults based on VOLUME.

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

(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. Compare the two scatterplots. What are the harvesting implications? Where do the various CLASS levels appear in the plots? An example of the type of plot is shown below.

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



(4)(a) Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and TYPE (see Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2, Black section 14.2). 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 = mydata)
summary(linear_model)
```

(4)(b) What implications are suggested by the coefficient estimates for CLASS levels (hint: this question is not asking if the estimates are statistically significant. It is asking for an interpretation of any pattern in these coefficients, and how this may relate to earlier displays)?

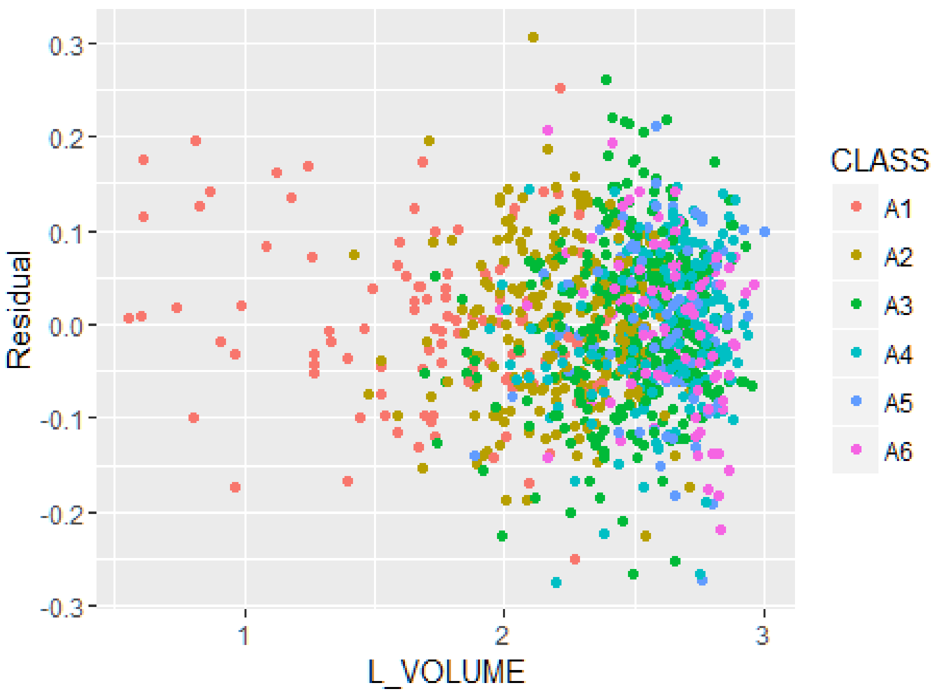
(4)(c) Is TYPE an important predictor in this regression (hint: this question is not asking if TYPE is statistically significant, but rather how it compares to the other independent variables in terms of its contribution when predictions of L_SHUCK might be made)?

(5)(a) Perform an analysis of the residuals resulting from the regression model in (3) (see Kabacoff Section 8.2.4, p. 178-186, the Data Analysis Video #2.) If "linear_model" is the regression object, use linear_model$residuals and construct a histogram and QQ plot. Compute the skewness and kurtosis (be aware with rockchalk the kurtosis value has 3.0 subtracted from which it differs from the moments package). What is revealed by these displays and calculations?

```r
# Histogram, base R example
hist(linear_model$residuals, ...)

# QQ plot, base R example
qqnorm(linear_model$residuals, ...)
qqline(linear_model$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.

moments::kurtosis(linear_model$residuals) # OR,
rockchalk::kurtosis(linear_model$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 (keep in mind the y-axis and x-axis may be disproportionate which will amplify the variability in the residuals). Present boxplots of the residuals differentiated by CLASS and TYPE (these four plots can be conveniently presented on one page using *par(mfrow..)* or *grid.arrange()*. Test the homogeneity of variance of the residuals across classes using the bartlett.test() (see Kabacoff Section 9.3.2, p. 222). How well does the regression model fit the data? Here is the type of code needed:

```r
# Scatterplot of model residuals as a function of L_VOLUME, CLASS, ggplot2
ggplot(linear_model, aes(x = L_VOLUME,y = linear_model$residuals)) +
  geom_point(aes(color = CLASS))  + labs(x = "L_VOLUME", y = "Residual")

# You will need to modify the above code to color by TYPE.

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



There is a tradeoff faced in managing the abalone harvest. The infant population must be protected since that represents future harvests. On the other hand, the harvest should be designed to be efficient with a yield to justify the effort. This assignhment will use VOLUME to form binary decision rules. If VOLUME is below a "cutoff" (i.e. specified volume), that individual will not be harvested. If above, it will be harvested.

This assignment will require plotting of infants versus adults. For this plotting to be accomplished, "for loops" will be used to compute the harvest proportions. These loops must use the same value for the constants min.v and delta; and, use the same statement "for(k in 1:1000)". Otherwise, the resulting infant and adult proportions cannot be directly compared and plotted as requested. Note the example code supplied below.

(6)(a) 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. Example code for doing this is supplied below.

```
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
# harvesting. 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
```

(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. This is for descriptive purposes to illustrate the difference between populations. The two split points suggest an interval within which potential cutpoints may be located.

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

This part will address the determination of a volume.value corresponding to the observed maximum difference in harvest percentages of adults and infants. To calculate this result, the proportions from (6) must be used. These proportions must be converted from "not harvested" proportions to "harvested" proportions by using (1-prop.infants) for infants, and (1-prop.adults) for adults. The reason the proportion for infants drops sooner than adults, is that infants are maturing and becoming adults with larger volumes.

(7)(a) 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(), ?abline(), ?text() to review documentation pages
```

(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). The procedure is to individually smooth (1-prop.adults) and (1-prop.infants) before determining an estimate of the maximum difference. Determine the volume.value corresponding to the maximum of the variable smooth.difference (hint: use which.max(smooth.difference)).

```
# 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(), ?abline(), ?lines(), ?text() to review documentation pages

# The peak can be found by identifying the relevant - i.e. largest - value
# in the smooth.difference vector and passing the element number, as opposed
# to the value, as a brackteted index for volume.value; volume.value[ ... ]

# ?which(), ?which.max() to review documentation pages
```

(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.")

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

(1-prop.infants)[which.max(smooth.difference)] # [1] 0.2036474

# ?which(), ?which.max() to review documentation pages
```

There are alternative ways to determine cutoffs for harvesting. Two such cutoffs are described below:

(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.

```
# 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.

volume.value[volume.value > max(mydata[mydata$CLASS == "A1" &
  mydata$TYPE == "I", "VOLUME"])][1] # [1] 206.9844

# 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] 0.3404255
```

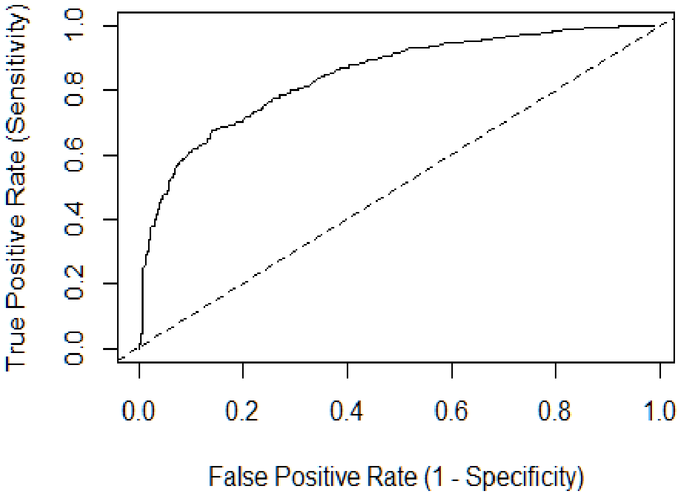
(8)(b) Another cutoff can be determined for which the proportion of adults not harvested equals the proportion of infants harvested. This cutoff would equate these rates; effectively, our two errors: 'missed' adults and wrongly-harvested infants. This leaves for dicussion which is a greater loss: a larger proportion of adults not harvested or infants harvested? This cutoff is 253.6. Calculate the separate harvest proportions for infants and adults using this cutoff and report.

```
# ?abs(), ?which(), ?which.min() to review documentation pages

# Although the relevant volume.value - 253.6 - 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 absolute difference between prop.adults and  (1 - prop.infants).

volume.value[which.min(abs(prop.adults - (1-prop.infants)))] # [1] 253.6113

# The infant and adult harvest proportions can be determined in much the same way
# we calculated proportions for item (8)(a).
```

(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. This is most easily done with the auc() function from the "flux" package.

Areas-under-curve, or AUCs, greater than 0.8 are taken to indicate good discrimination potential. Do you agree with this general rule? The expected ROC curve is given below and on the self-check page (see Kabacoff Section 17.6, p. 405-408).

Receiver Operating Characteristic (ROC) graphics are useful for organizing binary classifiers and visualizing their performance. ROC graphs are used in medical decision making, in machine learning and data mining research to compare different classification strategies.



(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). Based on the ROC curve, it is evident a wide range of possible "cutoffs" exist. Compare and discuss the different cutoffs. Is it possible to make a final selection at this time? What additional information would be helpful?

```
# 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:

sum(mydata$VOLUME >= volume.value[which.max(smooth.difference)])/
  (total.adults + total.infants) # [1] 0.5501931
```

The required table may be created in R, but this is not required. The table on the self-check page was output directly to the console as a data frame with the rows (and columns) named.

```
# ?cbind(), ?data.frame(), ?rownames(), ?colnames() to review documentation pages
```

Conclusions (Address each part)

Assume you are expected to make a presentation to the investigators for the purposes of determining a cutoff to implement based on your analysis. How would you do so?

- What qualifications or considerations would you present regarding your analysis?
- Would you make a specific recommendation or outline various choices and tradeoffs?
- Would you pose questions to the investigators? If so, what?
- What suggestions would you have for implementation of a selected cutoff?
- Would you indicate the need for additional data, investigations or consideration of alternative approaches?