Data Analysis Assignment #2

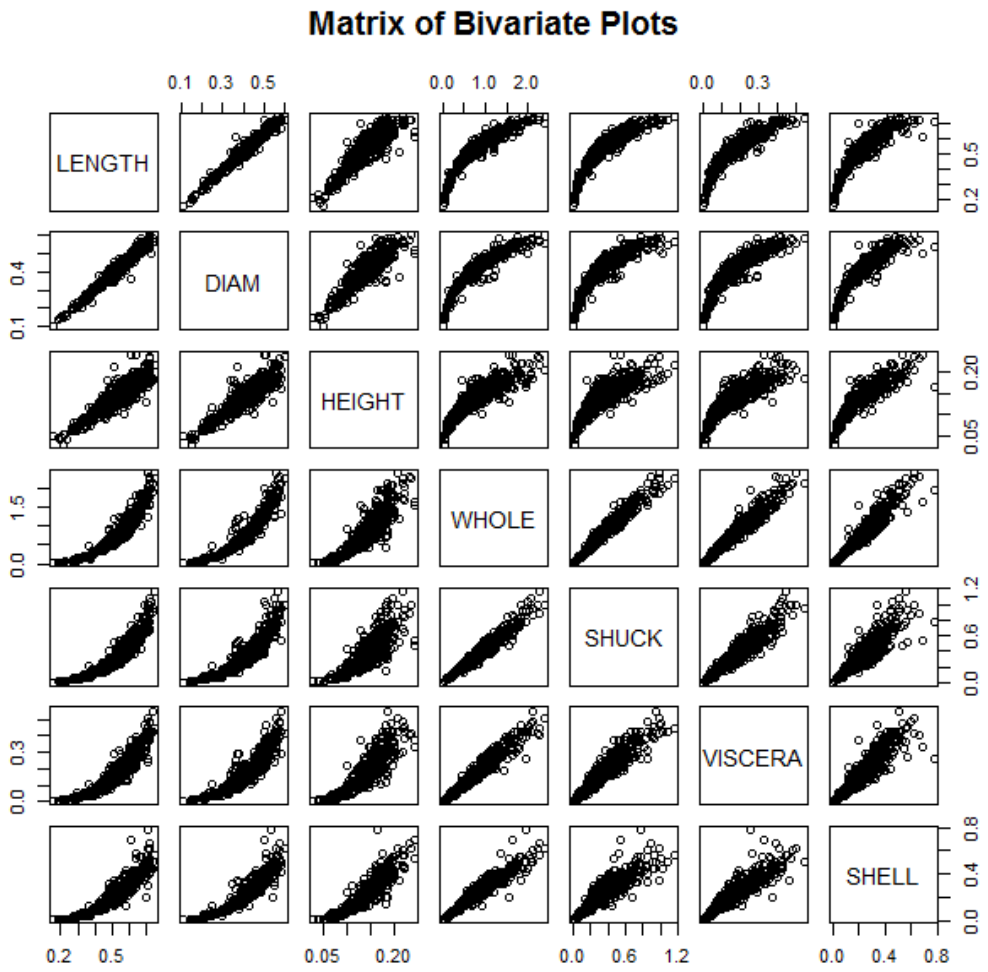
Kevin Wong
Northwestern University MSPA
Predict 401 Sec 58
Fall 2015

Introduction:

Abalone are often overharvested. The goal of this report is to come up with a decision rule to manage the harvesting of abalone. To accomplish our goal, we will use the same dataset of 500 observations from assignment #1 and perform additional exploratory data analysis. After the exploratory data analysis, we will dive deeper in our analysis using more advanced statistical methods like analysis of variance and linear regression to help us come up with suitable decision rules for harvesting abalone.

Results:

Figure 1



The bivariate plots are worth another look as represented in figure 1. Figure 1 shows scatterplots between all physical measurement variables. For the most part, this plot reveals a fairly strong positive linear relationship between all physical measurements. However, this begs the question of how strong are the correlations between each variable. We examine these coefficients in table 1 below.

Table 1 Correlation Coefficient Table

Variable	Variable	Coefficient	Method
DIAM	LENGTH	0.984	Pearson
SHUCK	WHOLE	0.977	Pearson
LENGTH	WHOLE	0.971	Spearman
VISCERA	WHOLE	0.968	Pearson
DIAM	WHOLE	0.966	Spearman
LENGTH	SHUCK	0.960	Spearman
SHELL	WHOLE	0.955	Pearson
LENGTH	VISCERA	0.953	Spearman
DIAM	SHUCK	0.949	Spearman
DIAM	VISCERA	0.947	Spearman
DIAM	SHELL	0.944	Spearman
VISCERA	SHUCK	0.943	Pearson
LENGTH	SHELL	0.940	Spearman
HEIGHT	SHELL	0.918	Spearman
SHELL	VISCERA	0.908	Pearson
HEIGHT	WHOLE	0.907	Spearman
HEIGHT	DIAM	0.898	Pearson
HEIGHT	VISCERA	0.897	Spearman
SHELL	SHUCK	0.895	Pearson
HEIGHT	LENGTH	0.895	Pearson
HEIGHT	SHUCK	0.870	Spearman

Table 1 shows a complete pairing of each physical measurement variable and their correlation coefficient sorted from the highest to the lowest coefficient value. This confirms our initial examination of Figure 1 that all physical measurements have strong linear relationships.

Figure 2

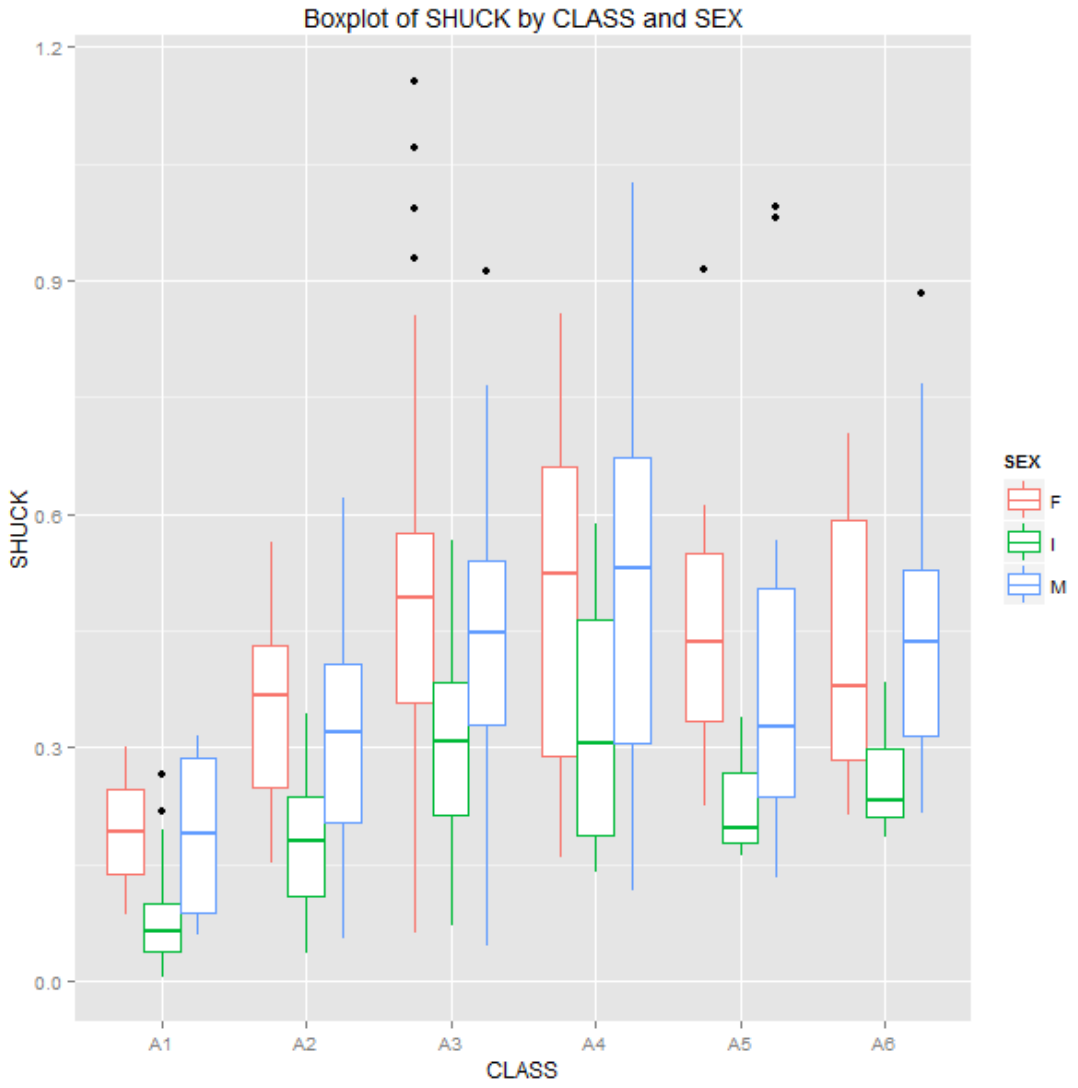


Figure 2 shows boxplots of shuck differentiated by age class and sex. This figure reveals there is an obvious distinction of shuck from age classes A1 to A2. Starting from A3 onward, there is high variability in shuck which indicates shuck as a poor indicator in determining age class from A3 to A6. There are also outliers frequently present throughout the higher age classes which possibly confirms our suspicion that there is much variability of shuck between age class and sex.

Table 2 Shuck and Volume Contingency Table

	below	above	Sum
below	226	25	251
above	24	225	249
Sum	250	250	500

Table 3 Calculated Chi Square Statistic and P-Value

chi.squared	p.value
323.213171410743	2.89077097070329e-72

Table 2 shows a 2x2 contingency table of our dataset for abalone below and above the median shuck value and median volume value. Margins have been added to the table for the purpose of computing the chi square statistic. Table 3 is the resulting chi square test of independence with a value of 323.213 and a significant p-value of 2.89e-72. These calculated values suggest that the variables are related.

Next, we perform an analysis of variances on shuck using class and sex with class*sex interaction term. The results are shown in the appendix in appendix 1. Class and sex had a significant effect on shuck, but class*sex interaction term had no significant effect due to a p-value of 0.783. Another analysis of variance was performed, but this time without the class*sex interaction term. The result was that class and sex still had significant effect on shuck. Through the use of Tukey's Honestly Significant Difference, it's determined that there are significant differences between lower age classes (A1 and A2) and higher age classes (A3 to A6). There were also significant differences between infant and adult (male and female). The results of our chi square test of independence and analysis of variance leads us to take a closer look at the relationship between shuck and volume.

Figure 3

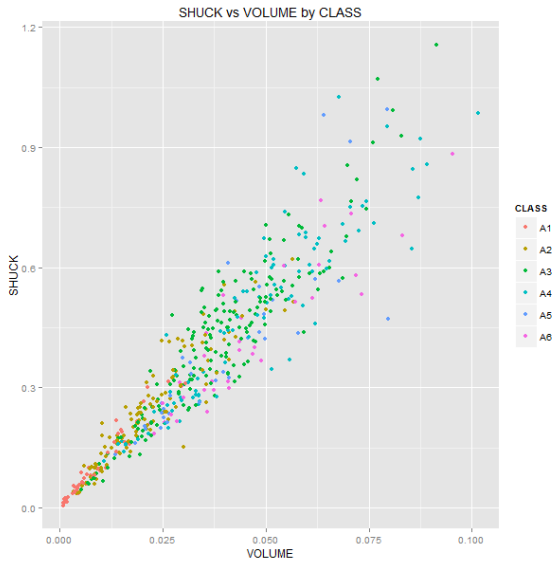


Figure 4

Figure 3 shows a scatterplot of shuck versus volume and figure 4 shows a scatterplot of their logarithms differentiated by age class. In figure 3, age classes A1 and A2 fall in the lower left region while other age classes are scattered throughout. This indicates that there is a linear relationship between shuck and volume, however it is difficult to differentiate between age classes A3 to A6 due to variability. In figure 4, majority of age classes are clustered in the upper right region while age classes A1 and A2 are spread across the plot. There is a lot more variability in figure 3 compared to figure 4. It seems that using the log of the variables may create a better model.

We produced a regression model where log of shuck is the dependent variable on log of volume, age class and sex. The summary output is shown in appendix 2. Female abalone and age class A1 are used as the baseline. The coefficients produced by the summary indicate negative coefficients. This trend seems to indicate that if you move to a higher age class, the coefficient becomes increasingly negative. This means as age class increases, the log of shuck decreases. This relationship is odd since our prior analyses have indicated that higher age classes have high variability.

Figure 5

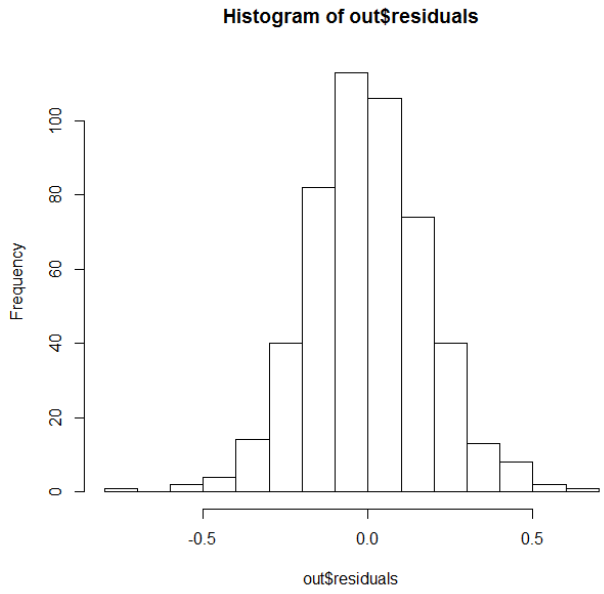


Figure 5 is a histogram of the residuals from our regression model. This histogram reveals the distribution of the residuals is approximately normal. Skewness and kurtosis values are calculated to be 0.048 and 3.735 respectively. This is expected as there appears to be some outliers.

Figure 6

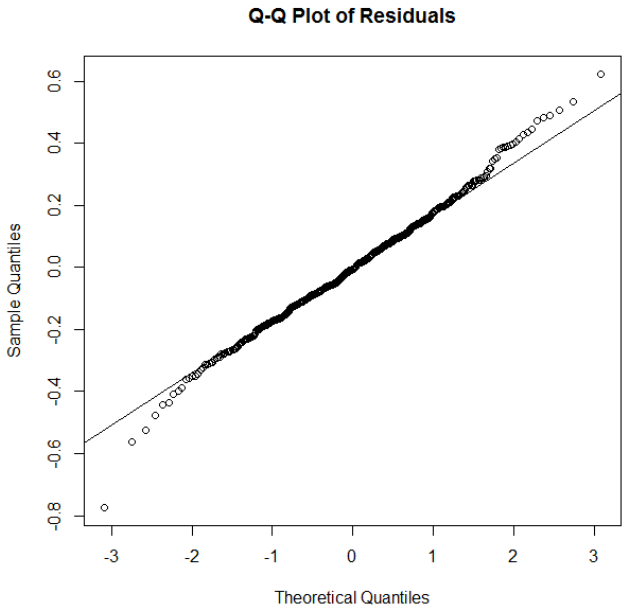


Figure 6 shows a Q-Q plot of the residuals where a majority of the points fall on the normal line with some slight skewing at both extremes. This again confirms that the residuals are approximately normal as indicated in figure 5.

Figure 7

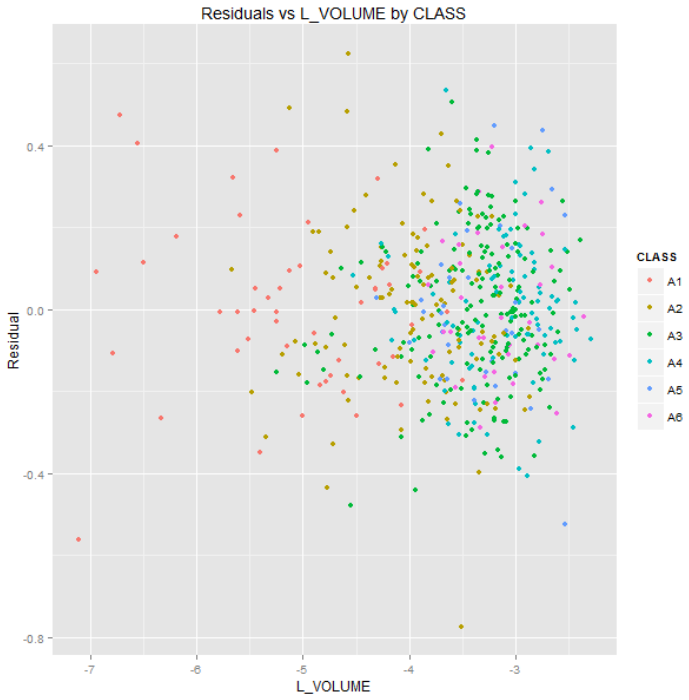


Figure 8

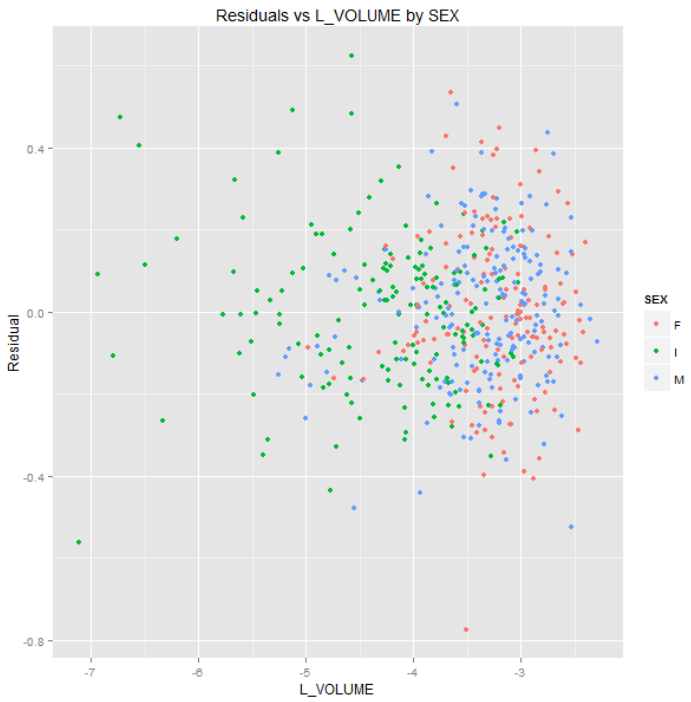


Figure 7 and 8 are scatterplots of the residuals versus the log of volume. Figure 7 is differentiated by age class while figure 8 is differentiated by sex. In figure 7, there is a noticeable distinction of lower age classes (A1 and A2) dispersed across more negative log volume values while higher age classes are dispersed across less negative values on the right hand side of the plot. In figure 8, we can see that the residuals of infant abalone are scattered across the left hand side of the plot, while female and male abalone are clustered to the right.

Figure 9

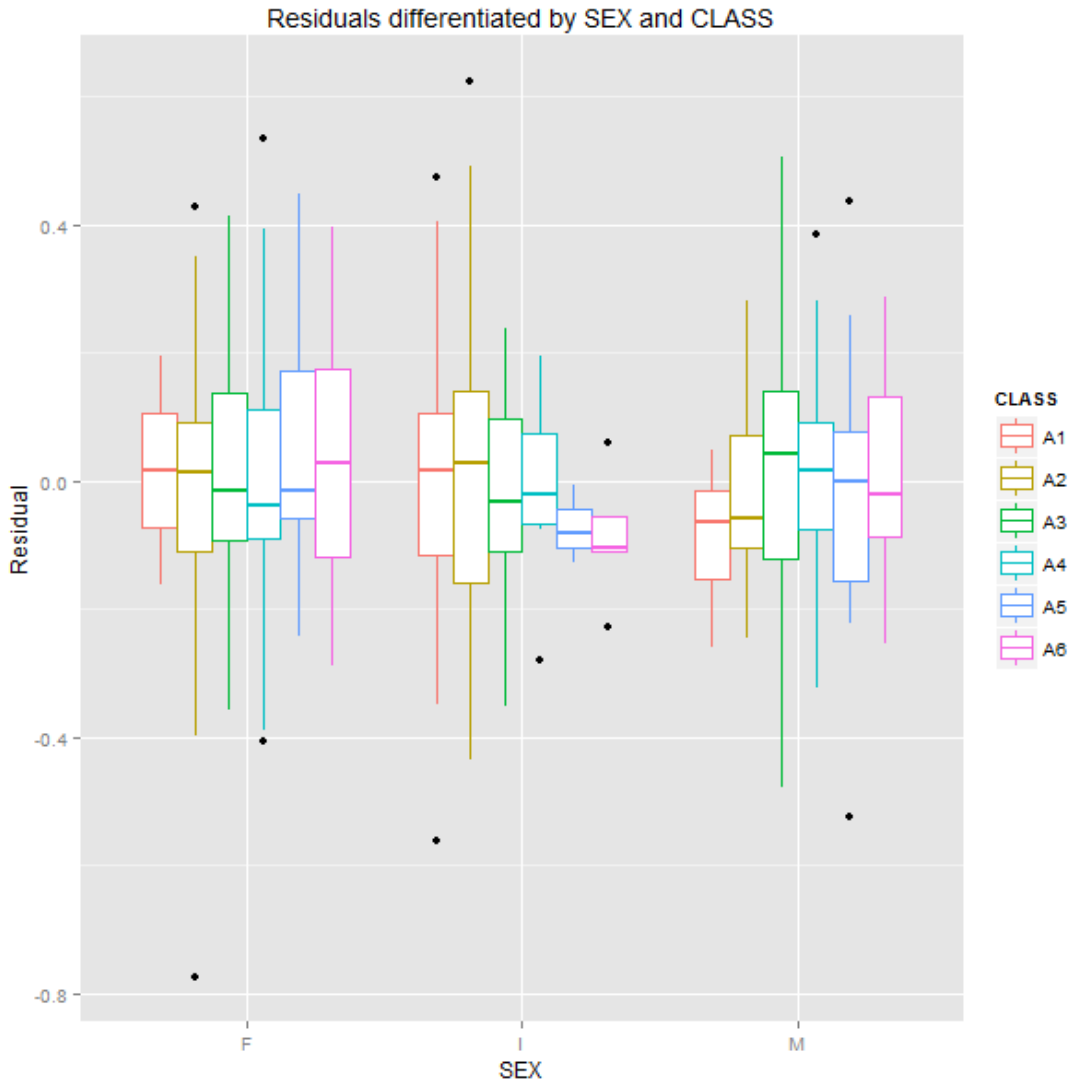
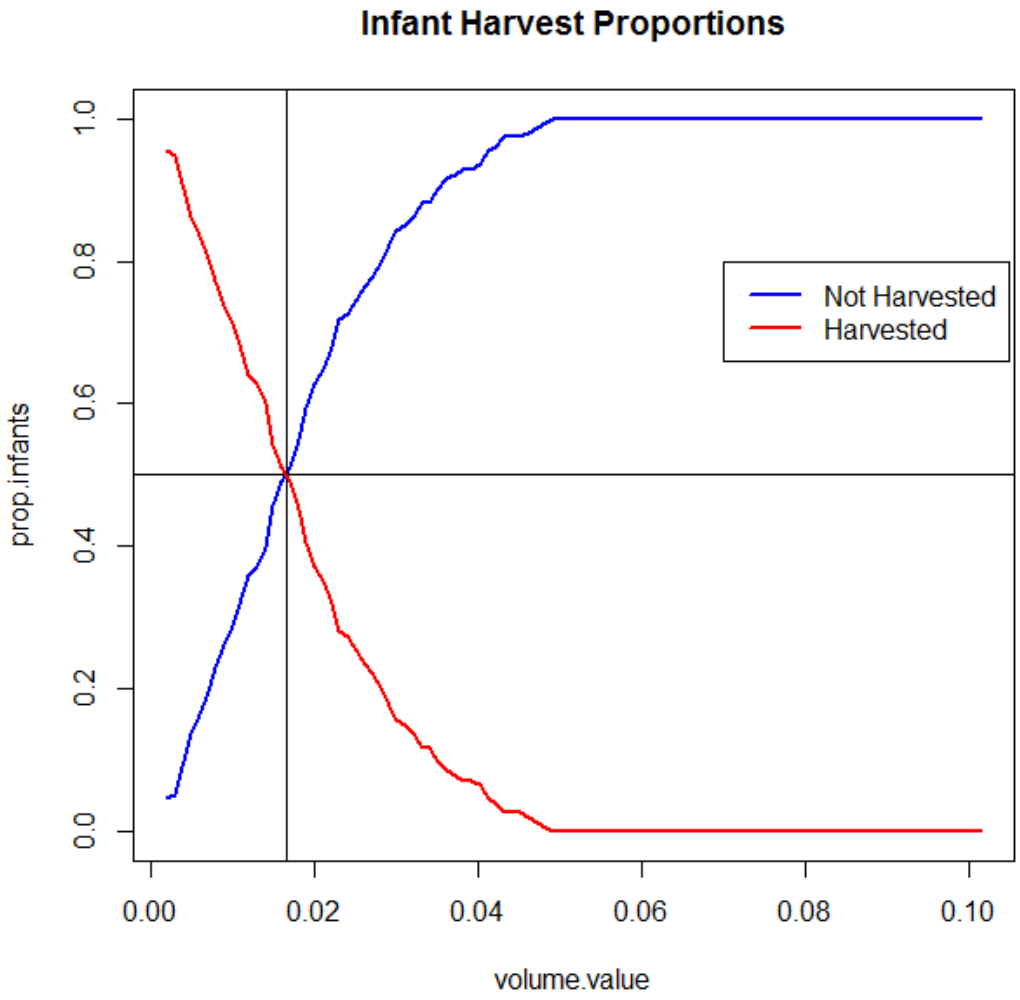


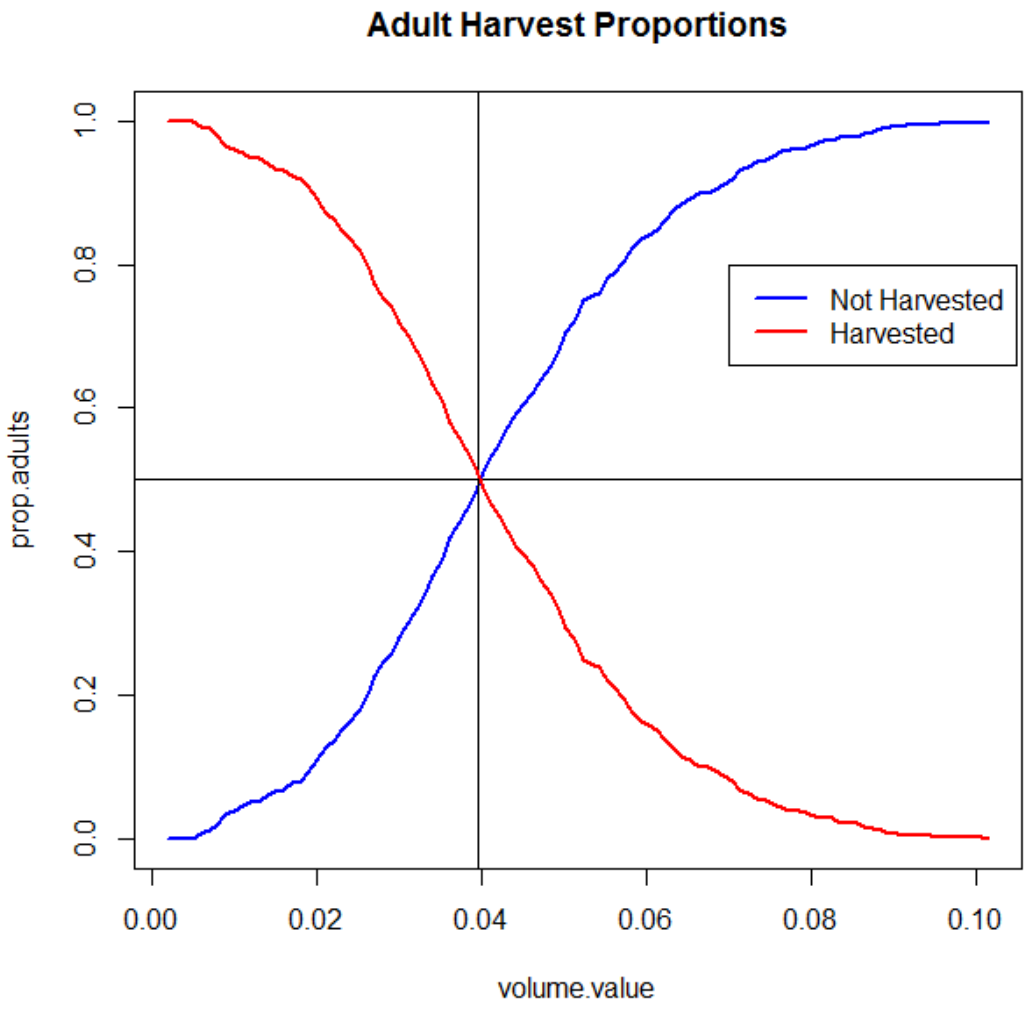
Figure 9 shows boxplots of the residuals differentiated by sex and age class. The regression model seems to fit the data as a majority of points fall around the 0.0 residual value across class and sex. There are some outliers of residual values, however this is indicative of the variability between physical measurements and aging.

Figure 10



So how do we determine a decision rule for harvesting abalone? Ultimately, we want to protect infants, but maximize our harvest potential of adults. We use a simple approach here which is to use a cutoff value for volume. Figure 10 shows the harvest proportions of infants versus volume. As volume changes, the harvest proportion of infants changes.

Figure 11



Similarly, figure 11 shows the harvest proportions of adults versus the volume. In both figures 10 and 11, we compute a 50% split of the volume value for infants and adults. At a 50% split, volume value for infants is 0.016 and volume value for adults is 0.039. This indicates that the cutoff of the volume value of adults is higher than that of infants, which is expected because infants generally have smaller volumes compared to adults. These 50% split values could possibly serve as a decision rule to harvest abalone between 0.016 and 0.039.

Figure 12

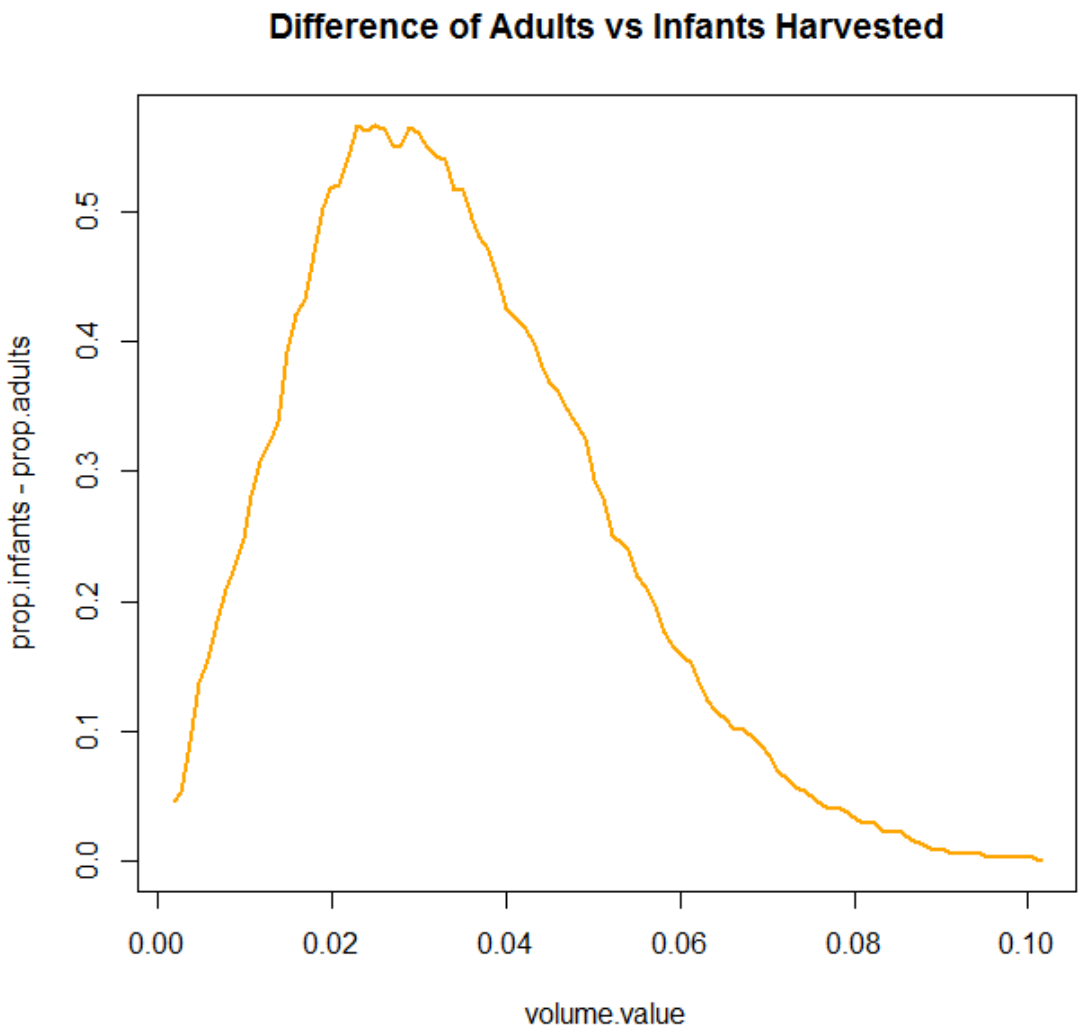


Figure 12 shows the difference between harvested infants and harvested adults versus volume value. It appears that the largest difference between adults and infants harvested occur between the volume values of approximately 0.02 and 0.04. This range of values is similar to the range we calculated for the 50% split cutoff.

Figure 13

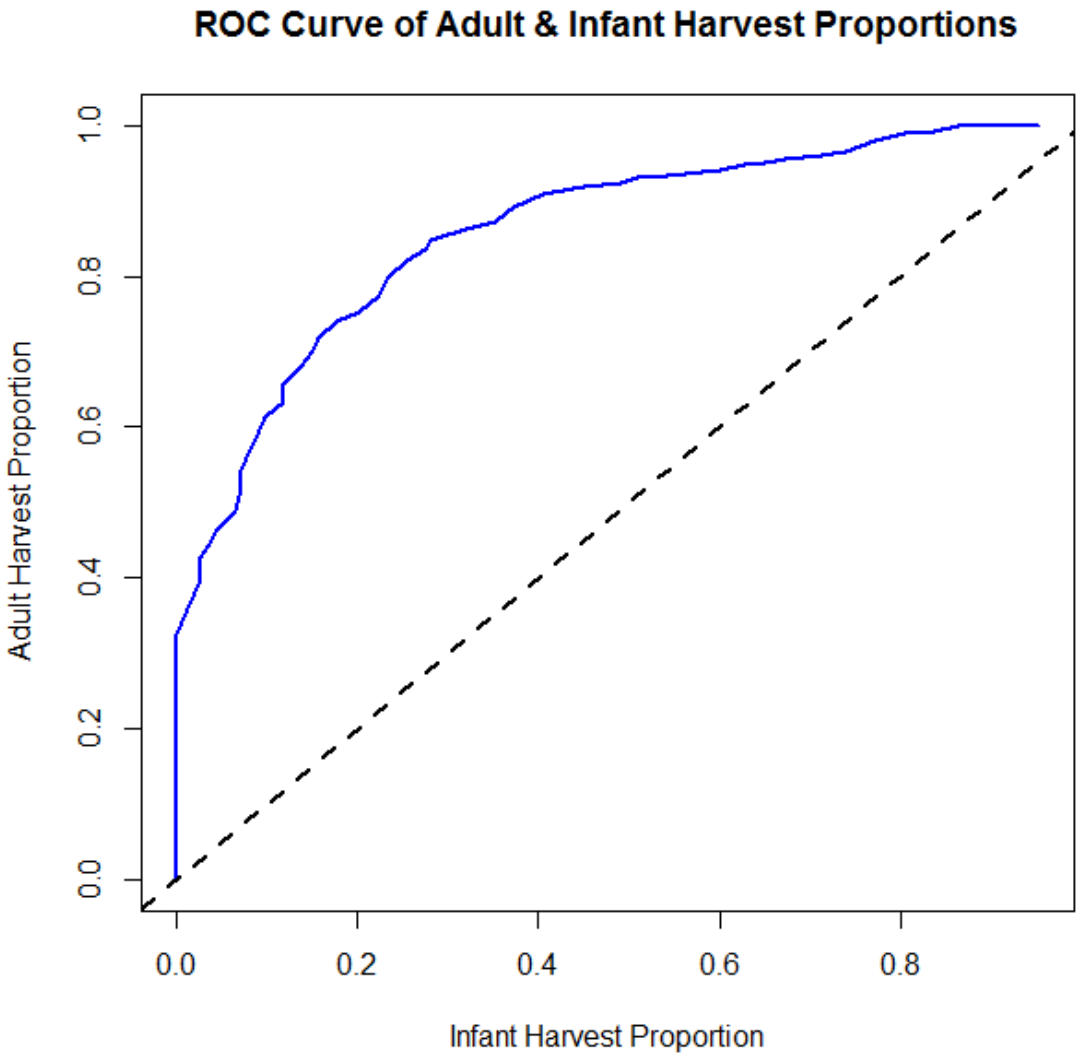


Figure 13 shows a ROC of the tradeoff of our volume values between infant and adult harvest proportions. Based on our volume decision rule, if infants are harvested, we have a false positive. The true positive corresponds to the adult harvest proportion. The smallest volume value that corresponds to no infants harvested is 0.049. The largest cutoff volume value where no infant is harvested for the adult harvest proportions is 0.326. So this range of volume values can be used as a possible decision rule which is between 0.049 and 0.326. If we recall from assignment #1 and previous plots in this assignment, physical measurements are a poor predictor for older age classes A3 to A6. So by using this potential decision rule, we may still harvest infants which will result in false positives.

Alternatively, physical measurements have some prediction power for age classes A1 and A2. It may be more sensible to determine a single volume value as a cutoff that protects

infants in age classes A1 and A2 from being harvested instead. Based on our calculations, the determined cutoff volume value is 0.0347. At 0.0347, age classes A3 and above will be maximized while zero infants will be harvested from age classes A1 and A2.

Conclusion:

Based on our analysis, using physical measurements as basis for abalone harvesting largely depends on the requirement of being able to maximize yield and minimize the harvest of infant abalone. If our model can satisfy that requirement, then a decision rule based on physical measurements can be viable. However, physical measurements have largely been imprecise in classifying abalone age class except in the lower age classes.

It is possible to rely on the volume cutoff value 0.0347 as a simple method of creating a decision rule to manage abalone harvesting. Because physical measurements seem to be somewhat reliable in classifying lower age classes from our analysis, this volume cutoff value may be a sensible decision rule.

In order to verify our conclusion, we can consider resampling from the original abalone dataset and running the same analysis and comparing results. Otherwise, it may be wise to bring the analyses to the original investigators to discuss other variables or factors that may shed light on a more optimal solution for harvesting abalone.

When analyzing data from observational studies involving different classes, we must consider whether the data are representative of the population and whether these different classes accurately describe the data. We must also consider how the data were collected. It's extremely important to keep in mind that measurements in observational studies do not indicate causality.

In conclusion, I learned it is very difficult to classify the age of abalone using physical measurements. Abalone can vary tremendously in physical measurements especially among adults. It's important to pose questions to the original investigators regarding other factors that could possibly affect the variability in the abalone dataset.

Appendix:

Appendix 1

```
> aov.shuck <- aov(SHUCK ~ CLASS+SEX+CLASS*SEX, mydata)
> summary(aov.shuck)
             Df Sum Sq Mean Sq F value   Pr(>F)
CLASS         5  7.150  1.4301  48.545  < 2e-16 ***
SEX           2  1.914  0.9568  32.479 5.85e-14 ***
CLASS:SEX    10  0.188  0.0188   0.637    0.783
Residuals   482 14.199  0.0295
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> aov.shuck2 <- aov(SHUCK ~ CLASS+SEX, mydata)
> summary(aov.shuck2)
             Df Sum Sq Mean Sq F value   Pr(>F)
CLASS         5  7.150  1.4301   48.91  < 2e-16 ***
SEX           2  1.914  0.9568   32.72 4.55e-14 ***
Residuals   492 14.387  0.0292
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> TukeyHSD(aov.shuck2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

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

$CLASS
              diff         lwr         upr     p adj
A2-A1  0.146726549  0.06315513 0.230297969 0.0000105
A3-A1  0.317324385  0.23830178 0.396346989 0.0000000
A4-A1  0.392685243  0.30599709 0.479373392 0.0000000
A5-A1  0.299591837  0.19131837 0.407865300 0.0000000
A6-A1  0.323234694  0.21496123 0.431508157 0.0000000
A3-A2  0.170597837  0.11178059 0.229415083 0.0000000
A4-A2  0.245958695  0.17718565 0.314731740 0.0000000
A5-A2  0.152865288  0.05832408 0.247406493 0.0000697
A6-A2  0.176508145  0.08196694 0.271049351 0.0000021
A4-A3  0.075360858  0.01219345 0.138528265 0.0090366
A5-A3 -0.017732549 -0.10827773 0.072812628 0.9934541
A6-A3  0.005910308 -0.08463487 0.096455485 0.9999686
A5-A4 -0.093093407 -0.19040061 0.004213799 0.0699212
A6-A4 -0.069450549 -0.16675776 0.027856656 0.3201162
A6-A5  0.023642857 -0.09330585 0.140591564 0.9924132

$SEX
           diff         lwr          upr     p adj
I-F -0.13046758 -0.17613478 -0.084800382 0.0000000
M-F -0.03404416 -0.07740096  0.009312638 0.1558885
M-I  0.09642342  0.05275740  0.140089445 0.0000009
```

Appendix 2

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

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
```

R Code:

```
# Data Analysis Assignment 2

# Question 1
# read in and examine data
mydata <- read.csv("./Data Analysis 2/mydata.csv", sep="")
str(mydata)
summary(mydata)

plot(mydata[,2:8], main = "Matrix of Bivariate Plots")

# create table of correlation coefficient values
allCorrPearson <- cor(mydata[,2:8], method = "pearson")
allCorrSpearman <- cor(mydata[,2:8], method = "spearman")

library(gridExtra)
frame() # start new plot frame
grid.table(allCorr, rows=NULL) # for pretty print
frame() # start new plot frame
grid.table(allCorrSpearman, rows=NULL) # for pretty print
```

```
# Question 2
library(ggplot2)
ggplot(mydata, aes(x=SEX, y=SHUCK)) + geom_boxplot(aes(color=CLASS)) +
ggtitle("Boxplot of SHUCK by CLASS and SEX")
ggplot(mydata, aes(x=CLASS, y=SHUCK)) + geom_boxplot(aes(color=SEX)) +
ggtitle("Boxplot of SHUCK by CLASS and SEX") # or is this better plot?

# Question 3
# define pearson chi square statistic function
chisquared <- function(x) {
   e11 <- x[3,1]*x[1,3]/x[3,3] # expected value of x[1,1]
   e12 <- x[3,2]*x[1,3]/x[3,3] # expected value of x[1,2]
   e21 <- x[3,1]*x[2,3]/x[3,3] # expected value of x[2,1]
   e22 <- x[3,2]*x[2,3]/x[3,3] # expected value of x[2,2]

   chisqStat <- (x[1,1]-e11)^2/e11 + (x[1,2]-e12)^2/e12 + (x[2,1]-e21)^2/e21 + (x[2,2]-
e22)^2/e22

   return(list("chi-squared" = chisqStat, "p-value" = pchisq(chisqStat, 1, lower.tail = F)))
}

# dichotomize SHUCK and VOLUME
shuck <- factor(mydata$SHUCK > median(mydata$SHUCK), labels = c("below",
"above"))
volume <- factor(mydata$VOLUME > median(mydata$VOLUME), labels = c("below",
"above"))

# generate table
shuck_volume <- addmargins(table(shuck, volume))
shuck_volume
frame()
grid.table(shuck_volume) # pretty print of table

# apply user-defined function on table
frame()
grid.table(data.frame(chisquared(shuck_volume)),rows=NULL)
# compare p-value with built-in chi square test in R (should match)
chisq.test(shuck_volume[1:2, 1:2], correct = F)
pchisq(323.213, 1, lower.tail = FALSE)

# Question 4
aov.shuck <- aov(SHUCK ~ CLASS+SEX+CLASS*SEX, mydata)
summary(aov.shuck)
aov.shuck2 <- aov(SHUCK ~ CLASS+SEX, mydata)
```

```
summary(aov.shuck2)
TukeyHSD(aov.shuck2)

# Question 5
library(ggplot2)
ggplot(mydata, aes(x=VOLUME, y=SHUCK)) + geom_point(aes(color=CLASS)) +
ggtitle("SHUCK vs VOLUME by CLASS")
L_SHUCK <- log(mydata$SHUCK)
L_VOLUME <- log(mydata$VOLUME)
ggplot(mydata, aes(x=L_VOLUME, y=L_SHUCK)) + geom_point(aes(color=CLASS))
+ ggtitle("L_SHUCK vs L_VOLUME by CLASS")

# Question 6
out <- lm(L_SHUCK~L_VOLUME+CLASS+SEX, mydata)
summary(out)

# Question 7
library(moments)
hist(out$residuals) # NAME THESE GRAPHS
qqnorm(out$residuals, main="Q-Q Plot of Residuals")
qqline(out$residuals)
skewness(out$residuals)
kurtosis(out$residuals)

ggplot(out, aes(x=L_VOLUME, y=out$residuals)) + geom_point(aes(color=CLASS)) +
labs(x="L_VOLUME",y="Residual") + ggtitle("Residuals vs L_VOLUME by CLASS")
ggplot(out, aes(x=L_VOLUME, y=out$residuals)) + geom_point(aes(color=SEX)) +
labs(x="L_VOLUME",y="Residual") + ggtitle("Residuals vs L_VOLUME by SEX")
ggplot(out, aes(x=SEX, y=out$residuals)) + geom_boxplot(aes(color=CLASS)) +
labs(x="SEX",y="Residual") + ggtitle("Residuals differentiated by SEX and CLASS")
ggplot(out, aes(x=L_VOLUME, y=out$residuals)) + geom_boxplot(aes(color=SEX)) +
labs(x="L_VOLUME",y="Residual") + ggtitle("Residuals vs L_VOLUME by CLASS")

# Question 8
idxi <- mydata[,1] == "I"
idxf <- mydata[,1] == "F"
idxm <- mydata[,1] == "M"

max.v <- max(mydata$VOLUME)
min.v <- min(mydata$VOLUME)
delta <- (max.v - min.v)/100
prop.infants <- numeric(0)
prop.adults <- numeric(0)
volume.value <- numeric(0)
```

```
total.infants <- length(mydata[idxi,1])
total.adults <- length(mydata[idxf,1]) + length(mydata[idxm,1])

for (k in 1:100) {
   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[idxf] <= value) +
sum(mydata$VOLUME[idxm] <= value))/total.adults
}

# infants
n.infants <- sum(prop.infants <= 0.5)
split.infants <- min.v + (n.infants + 0.5)*delta
plot(volume.value, prop.infants, col = "blue", main = "Infant Harvest Proportions", type
= "l", lwd = 2, ylim = c(0,1))
abline(h=0.5)
abline(v=split.infants)
lines(volume.value,1-prop.infants,col="red",lwd=2)
legend(0.07,0.8,c("Not
Harvested","Harvested"),lty=c(1,1),lwd=c(2,2),col=c("blue","red"))

# adults
n.adults <- sum(prop.adults <= 0.5)
split.adults <- min.v + (n.adults + 0.5)*delta
plot(volume.value, prop.adults, col = "blue", main = "Adult Harvest Proportions", type =
"l", lwd = 2)
abline(h=0.5)
abline(v=split.adults)
lines(volume.value,1-prop.adults,col="red",lwd=2)
legend(0.07,0.8,c("Not
Harvested","Harvested"),lty=c(1,1),lwd=c(2,2),col=c("blue","red"))

# Question 9
plot(volume.value,1-prop.adults,col="red",main="Proportion of Adults
Harvested",type="l",lwd=2)
plot(volume.value,1-prop.infants,col="blue",main="Proportion of Infants
Harvested",type="l",lwd=2)
plot(volume.value,prop.infants-prop.adults,col="orange",main="Difference of Adults vs
Infants Harvested",type="l",lwd=2)

# roc curve of infant and adult harvest proportions
plot(1-prop.infants, 1-prop.adults, col = "blue", main="ROC Curve of Adult & Infant
Harvest Proportions", type="l",lwd=2, xlab="Infant Harvest Proportion", ylab="Adult
Harvest Proportion")
```

```
abline(a=0,b=1,lty=2,lwd=2)

# find harvest threshold volume that protects infants and gives largest harvest of adults
# identify largest infant
max(mydata$VOLUME[mydata$SEX == "I"])
# find smallest volume value that corresponds to 0 harvested infants
min(volume.value[(1-prop.infants) == 0])
max(1-prop.adults[(1-prop.infants) == 0])

# Question 10
cutoff <- 0.0347

index.A1 <- (mydata$CLASS == "A1")
indexi <- index.A1 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A1)

index.A2 <- (mydata$CLASS == "A2")
indexi <- index.A2 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A2)

index.A3 <- (mydata$CLASS == "A3")
indexi <- index.A3 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A3)

index.A4 <- (mydata$CLASS == "A4")
indexi <- index.A4 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A4)

index.A5 <- (mydata$CLASS == "A5")
indexi <- index.A5 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A5)

index.A6 <- (mydata$CLASS == "A6")
indexi <- index.A6 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A6)
```