Data Analysis Assignment 2

Anamitra Bhattacharyya

MSPA Fall 2015
Northwestern University
PREDICT 401-DL Section 58

1. INTRODUCTION

Abalones are an economical and recreational resource in parts of the world. The growth of abalone is affected by a number of factors, and the ability to grow and harvest them is an important part of the industry that has grown up around them. The ability to utilize physical measurements for harvesting of abalone is an important management procedure, which can be used to control harvesting requirements.

This document reports on some exploratory data analysis (EDA) using information derived from an Australian team of investigators using Tasmanian abalone. The original purpose of Australian study was to predict the age of abalone from physical measurements (*e.g.* size, height, volume, weight) to bypass any need for counting rings on abalone shells, which is time-consuming. This report aims to assess whether physical measurements can be used for harvesting abalone, using a 500-abalone sample set, derived from a larger study of 4141 observations. Additionally, this study explores whether a simple decision rule can be devised using physical abalone measurements as a metric for harvesting abalone.

2. RESULTS

A random sample set of 500 of a total 4141 observations of Tasmanian abalone, tracked twelve different metrics or variables, which comprises:

1. SEX = M (male), F (female), I (infant)
2. LENGTH = Longest shell length in mm
3. DIAM = Diameter perpendicular to length in mm
4. HEIGHT = Height perpendicular to length and diameter in mm
5. WHOLE = Whole weight of abalone in grams
6. SHUCK = Shucked weight of meat in grams
7. VISCERA = Viscera weight in grams
8. SHELL = Shell weight after drying in grams
9. RINGS = Age (+1.5 gives the age in years)
10. CLASS = Age classification based on RINGS (A1= youngest to A6=oldest)
11. VOLUME = LENGTH x DIAMETER x HEIGHT (derivative variable)*
12. DENSITY = WHOLE / VOLUME (derivative variable)*

*Note: Variables 11 (VOLUME) and 12 (DENSITY) were derived from Assignment 1

2.1 Pairwise correlation between variables

Rather than use the entire 4141 observational data set, a smaller sample group of 500 randomly selected observations was used to perform the EDA. An initial analysis was performed on the sample data set to explore the pairwise correlations. A graphical summary of the plot matrix plotting each parameter against each other is shown in Figure 1. To more quantitatively evaluate the various pairwise correlations between variables was performed:

1) Linear relationships were observed when plotting,
    a) Length, diameter and height variables against each other (used Pearson Correlation Coefficients)
    b) Whole weight, shuck (meat) and viscera (internal soft organs) and shell weight against each other (Pearson Correlation Coefficient)
2) There were curvilinear distributions away from linearity when plotting any of the dimensional variables (*e.g.* length, diameter, height) against any other weight-dependent parameters (*e.g.* whole weight, shuck, viscera, shell weight). In these instances the Spearman Correlation Coefficient was used.

The correlation coefficients for the linear and curvilinear relationships between variables are shown in Table 1. Correlations are shown only for the pairwise plots in the upper diagonal Figure 1, as the cognate reciprocal plots in the lower diagonal would produce similar results.

Figure 1: Matrix of bivariate plots to explore relationships between length, diameter, height, whole weight, shuck, viscera and shell of abalone

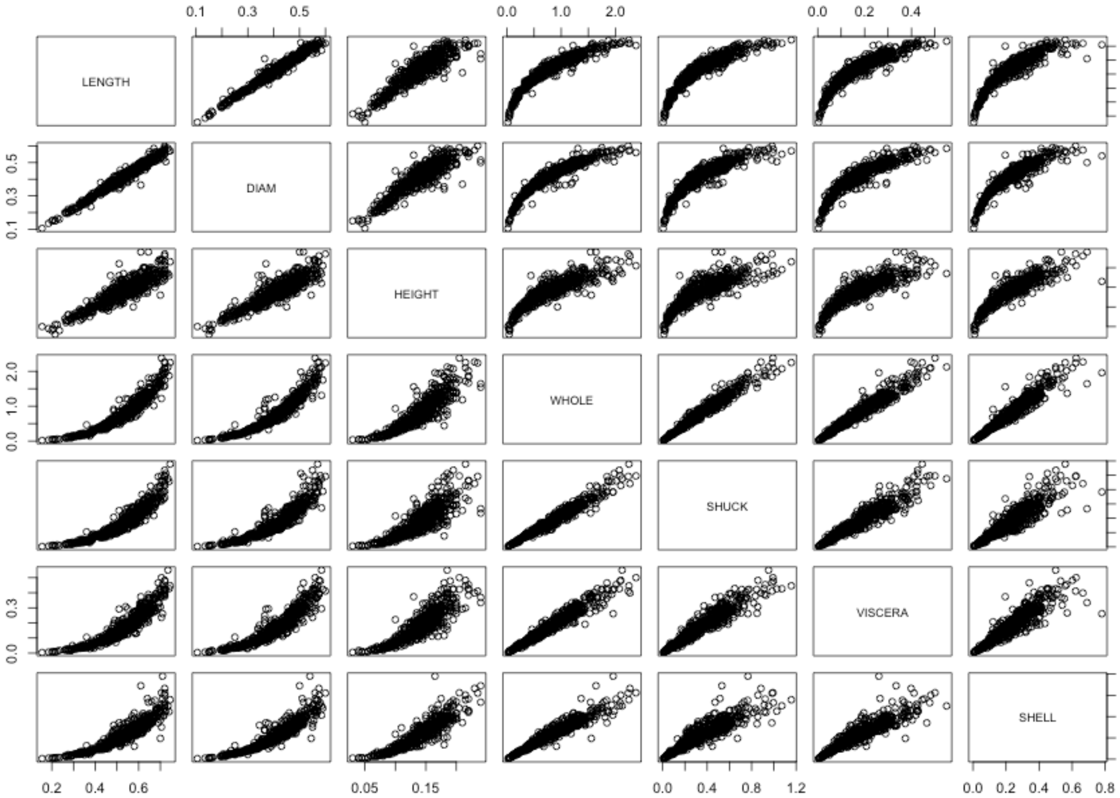


Table 1: Percent pairwise correlations between abalone variables

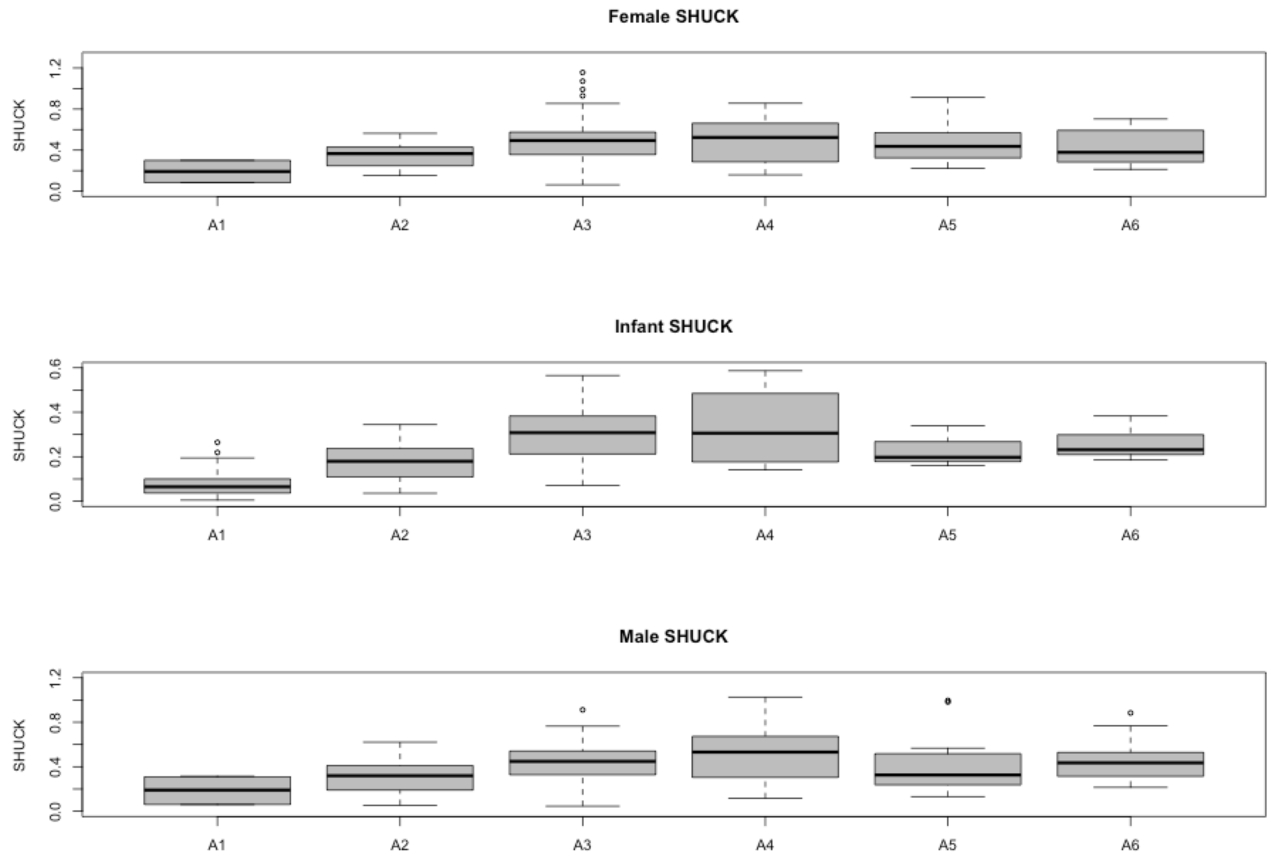
Variable	LENGTH	DIAMETER	HEIGHT	WHOLE	SHUCK	VISCERA	SHELL
LENGTH	NA	**Pearson: 98.39%**	**Pearson: 89.46%**	**Spearman: 97.09%**	**Spearman: 95.97%**	**Spearman: 95.30%**	**Spearman: 93.98%**
DIAMETER		NA	**Pearson: 89.80%**	**Spearman: 96.57%**	**Spearman: 94.92%**	**Spearman: 94.66%**	**Spearman: 94.42%**
HEIGHT			NA	**Spearman: 90.74%**	**Spearman: 87.02%**	**Spearman: 89.74%**	**Spearman: 91.85%**
WHOLE				NA	**Pearson: 97.73%**	**Pearson: 96.83%**	**Pearson: 95.46%**
SHUCK					NA	**Pearson: 94.26%**	**Pearson: 89.50%**
VISCERA						NA	**Pearson: 90.81%**
SHELL							NA

Note: NA=Not applicable

2.2 Relationship between abalone shuck, age-class and sex

As result of factors present during the life of each age-class, each of these age-class populations needed to be treated separately. Consequently, in order to investigate the relationship between shuck weights, age and sex further, the shuck and age-classes were analyzed for females, infants and males separately. The results are shown in the Figure 2. Recall from Assignment 1 that age-classes A1 and A2 are predominantly infant abalone, while A3 and above are male/female. The results indicate for each category of sex (female, infant, male) the mean shuck weight generally increases with age-class (the mean shuck weight in each age-class is shown as a bold horizontal line, see Figure 2) before reaching a plateau, and then decreasing slightly. More specifically there is a lot of variability in shuck weight as a function of age-class, irrespective of sex (female, infant, male). For example, in the A4 age-class the variability in shuck weight, including the box-whiskers, shows an overlap in the shuck weight range between females, infants and males, as well between A4 and the infant age-classes (*e.g.* A1, A2). This indicates great variability in abalone growth (with shuck weight being a measure of growth) as a function of age (age-class is an indicator of abalone age).

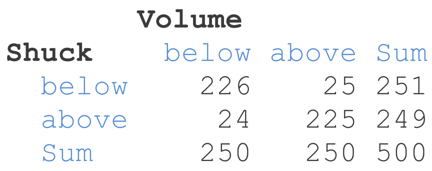
Figure 2: Boxplots showing abalone SHUCK differentiated by CLASS and SEX



Section 2.3 Evaluation of dependencies between abalone shuck weight and volume

In order to investigate whether the variables of shuck weight and volume were independent or not, a Pearson chi-square statistic was determined. The results are shown in Table 2.

Table 2: Test for independence of abalone SHUCK and VOLUME



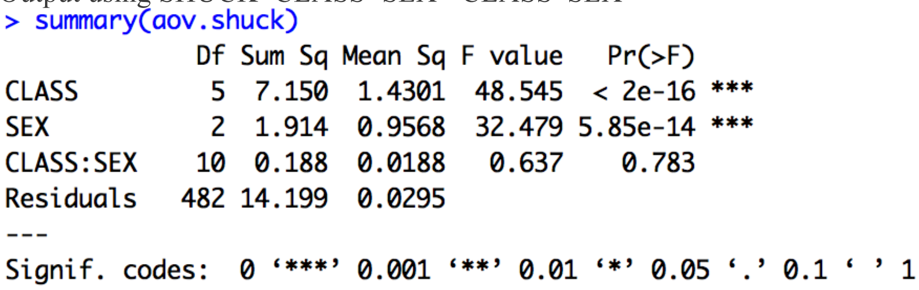
```
              Volume
Shuck     below above Sum
  below     226    25 251
  above      24   225 249
  Sum       250   250 500
```

Chi-square value = 124.5
p-value= 6.547914e-29

This analysis tests the independence of VOLUME and SHUCK using the Chi-squared test. This test evaluates the null hypothesis, which states that these variables are independent. Since the probability or p-value is low (p=6.547914e-29), however, this indicates to reject the null hypothesis and suggests the likelihood that the abalone shuck and volume are dependent, becomes stronger.

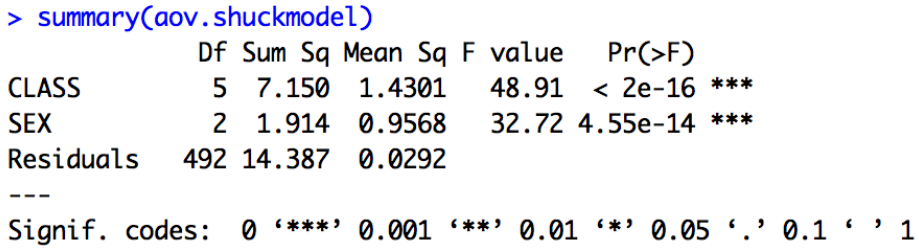
Section 2.4 Analysis of variance of shuck using age-class and sex
An analysis of variance (AOV) was performed on abalone shuck using age-class (CLASS) and sex (SEX) as the grouping variables. AOV results are presented below with (1) and without (2) an interaction term of CLASS*SEX.

1) Results using model with interaction term CLASS*SEX (two-way layout)
Output using SHUCK~CLASS+SEX+ CLASS*SEX



```
> 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
```

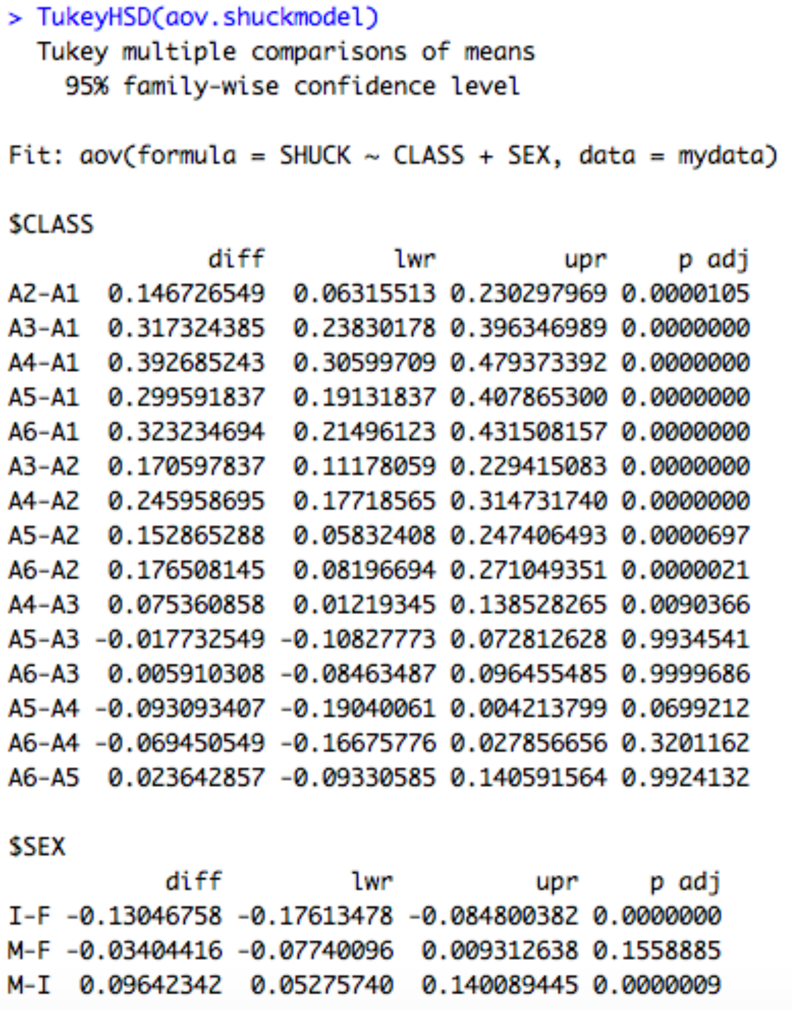
2) Results using model without interaction term CLASS*SEX (one-way layout)
Output using SHUCK~CLASS+SEX



```
> summary(aov.shuckmodel)
             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
```

There is statistical significance in the AOV output table in (1) above for CLASS and SEX but not using the interaction term (CLASS*SEX). The p-values for CLASS and SEX in the AOV output without the interaction term (2) are also significant and aligned with the values from (1). Thus, the interaction between CLASS*SEX, does not appear to be statistically significant. A more efficient way of comparing multiple pairwise comparisons, is for sub-groups in age-class and sex, is by using the Tukey(HSD) method (which decreases the error rate when performing multiple t tests). The output is shown below. For the comparison between age-classes, the pairwise differences in means are all significant, except those between A3-A5/A6, A4-A6 and A5-A6. For the comparison between sexes, this shows statistically significant differences between infants and males/females (adults) but not between male and female adults. So the shuck *versus* sex relationship seems to segregate into infants and adults (males/females), leading to the AOV significance previously. Similarly with shuck *versus* age-classes, A1 and A2 (lower age-classes) show statistical significant differences with all other age-classes producing the AOV significance noted above.

Output from TukeyHSD

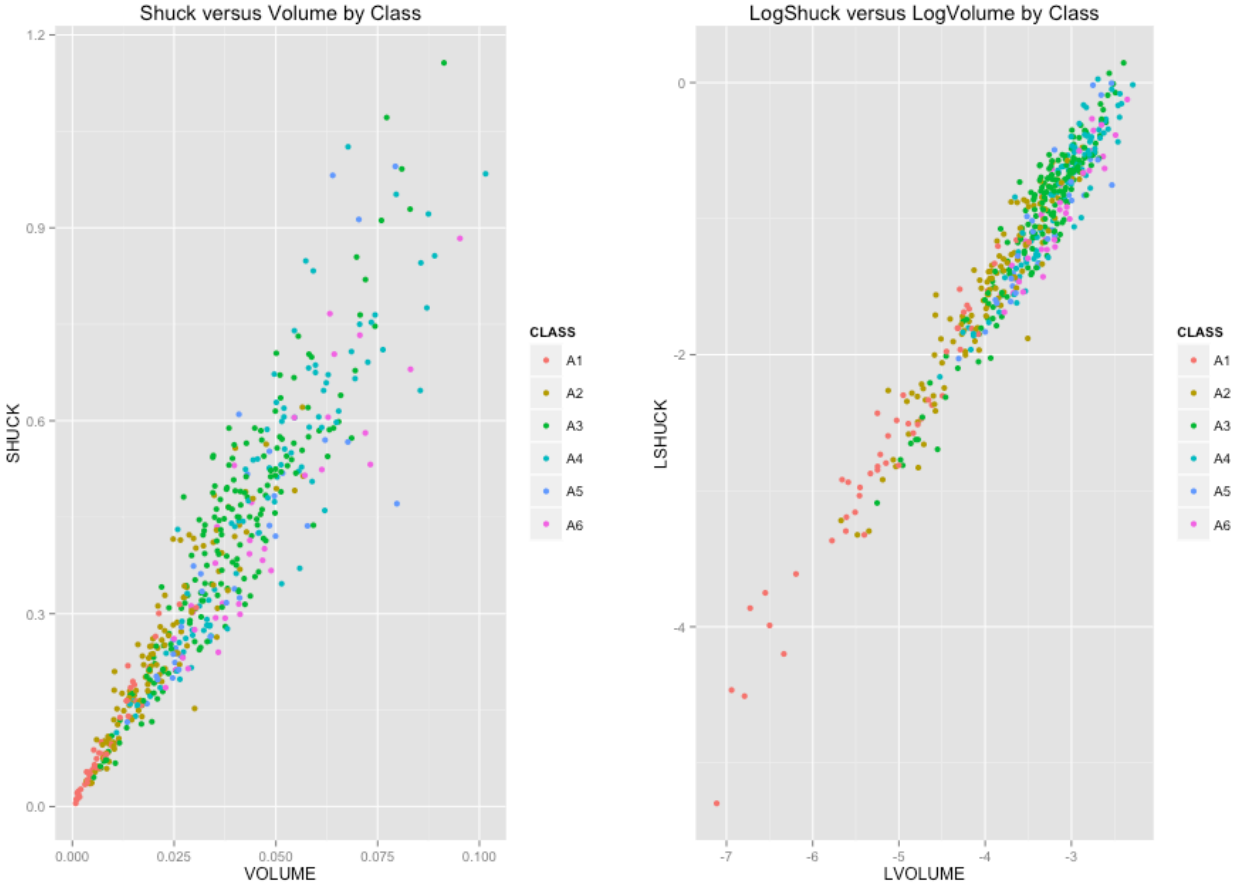


```
> TukeyHSD(aov.shuckmodel)
  Tukey multiple comparisons of means
    95% family-wise confidence level

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
```

Section 2.5 Relationship between abalone shuck and volume

As discussed previously there appears to be a dependency between shuck and volume. In order to investigate the relationship between these variables further a scatter plot of shuck versus volume was constructed differentiated by abalone age-class (Figure 3). Note first that there is proportionality between shuck and volume, such that lower values of shuck and volume are closer to the origin (zero value). As both shuck weight and volume of abalone increases, there is increased scatter (Figure 3, left panel). In order to tighten the scatter of points and create a more ‘discrete’ distribution, the shuck-volume plot was re-plotted in a log-log format (Figure 3, right panel). Interestingly, the infants are the predominant group that clusters together at the lower bounds of each the plots, consistent with them being smaller and more immature than the older age-classes (A3-A6), and show high degree of linearity in both plots.

Figure 3: Scatter plots of abalone shuck *versus* volume (left panel) and log-shuck *versus* log-volume (right panel) differentiated by age-class

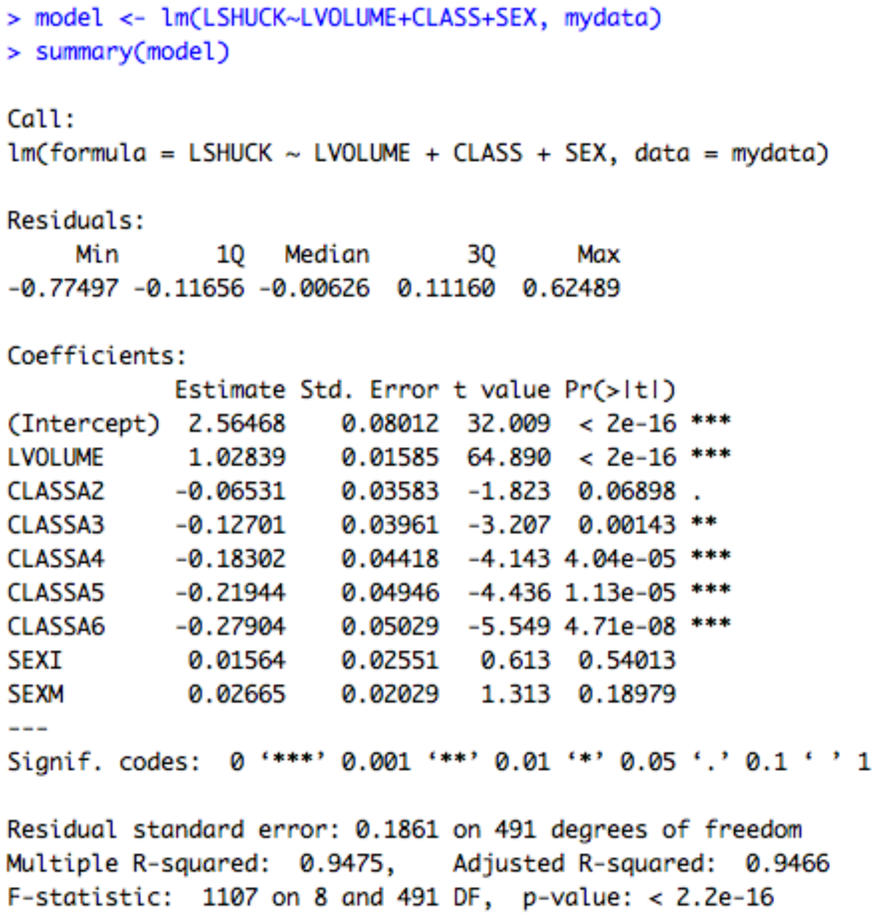


The lower age-classes (A1-A2) cluster to the lower quadrant in both plots, while the higher ages-classes cluster predominantly in the upper quadrants, but there is a high degree of variability in age-classes A3-A6, as confirmed in Figure 2 previously and exemplified in Figure 3 above. The log function helps to decrease the extent of the scatter at higher shuck/volume values, and transforms the relationship between shuck/volume into a more linear function. The implication from this analysis is that there is a broad linear relation between shuck and volume, with more immature abalone at the lower end of the scale, and more higher age classes in the upper end of the scale, but exhibits high variability in the higher age-classes. For example, abalone with age-class A3 is present at the mid and higher end of the linear scale with A4-A6 age classes.

2.6 Investigation of multiple regression model of abalone shuck against volume, class and sex

In order to further explore the extent and quality of the shuck *versus* volume relationship (together with age-class and sex), a linear regression analysis was performed. A summary of the output from the linear regression model is shown below.

Output from linear regression model



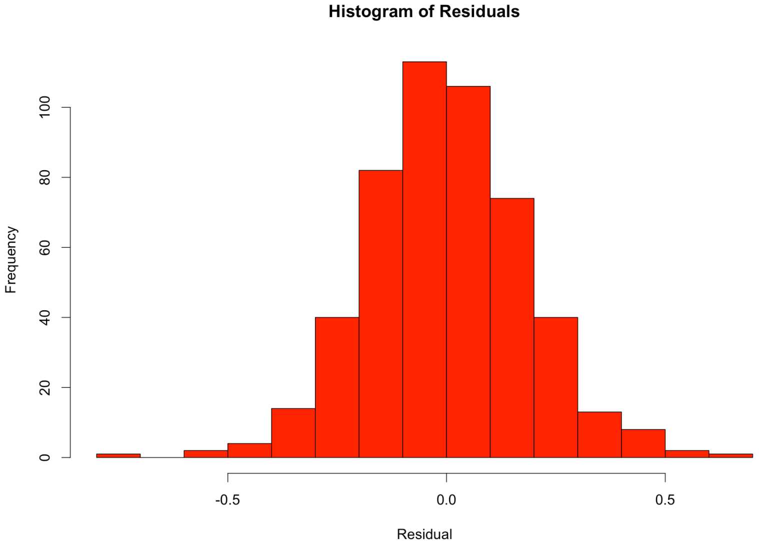
```
> model <- lm(LSHUCK~LVOLUME+CLASS+SEX, mydata)
> summary(model)

Call:
lm(formula = LSHUCK ~ LVOLUME + 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 ***
LVOLUME      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
```

Looking at the coefficients from the output, the baseline includes age-class A1 (and females), which is statistically significant. The intercept is statistically significant as is the relationship between shuck and log volume (LVOLUME). Interestingly, since this is parallel line regression, the slopes of the lines for age-classes A3-A6 display statistically significant slopes but show slight negative values compared to the baseline. The best fit (derived from the p-values) appears to be with the baseline (which includes A1). The coefficient values for sex (infants and males) appear not be significant. This appears somewhat unusual since age-class A1, for instance, would be predicted to comprise infants or juveniles.

2.7 Analysis of residuals from linear regression

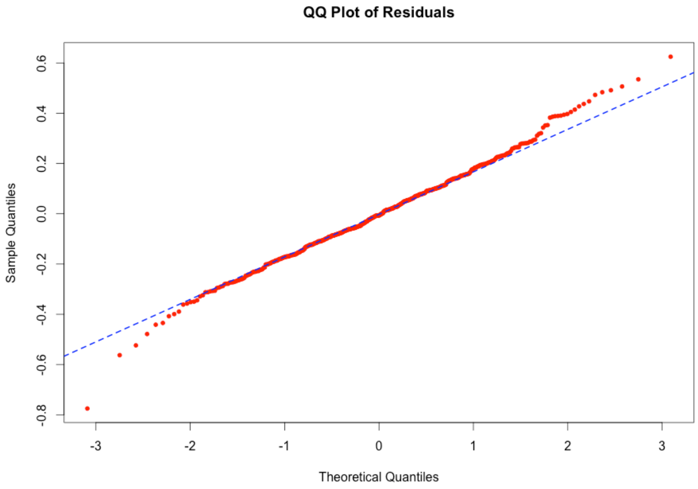
To evaluate the quality of the linear regression, the residuals from the model were plotted as a histogram and are shown in Figure 4. The residuals approximate to a normal, symmetrical distribution.

Figure 4: Histogram of residuals from linear regression model



Furthermore, a quantile-quantile (QQ) plot of the residuals (Figure 5) confirms a normal distribution function, with the points for the most part aligned with the blue-dotted line that defines a model normal distribution. At the extremities of the distribution on the QQ plot there is some departure from normality, as the points deviate from the dotted line of the normal function. The skewness (approximately ≈ 0.048) and kurtosis (≈ 3.735), that are metrics of the quality of the distribution, confirm a normal distribution (Figure 5). For normal distributions, skewness and kurtosis should be close to zero and 3.0, respectively.

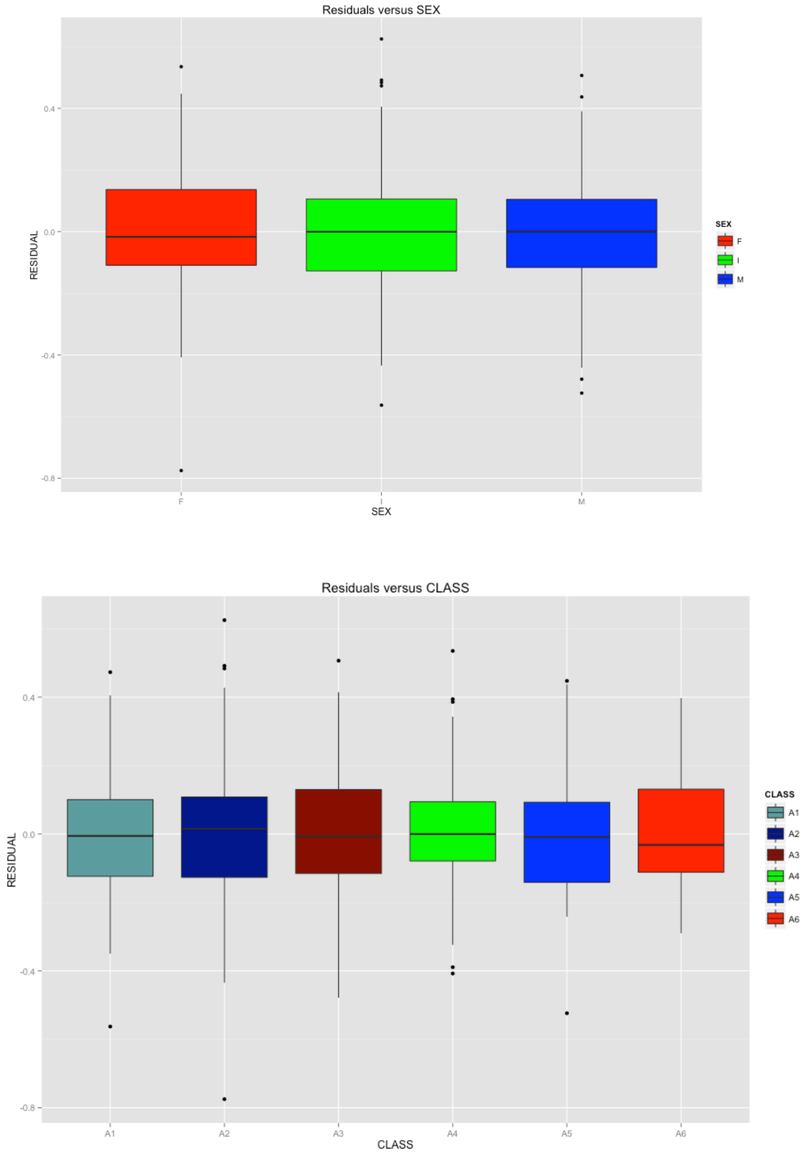
Figure 5: QQ plot of residuals



Skewness = 0.04799289 (for a normal distribution should be close to zero)
Kurtosis = 3.735231 (for a normal distribution should be close to 3)

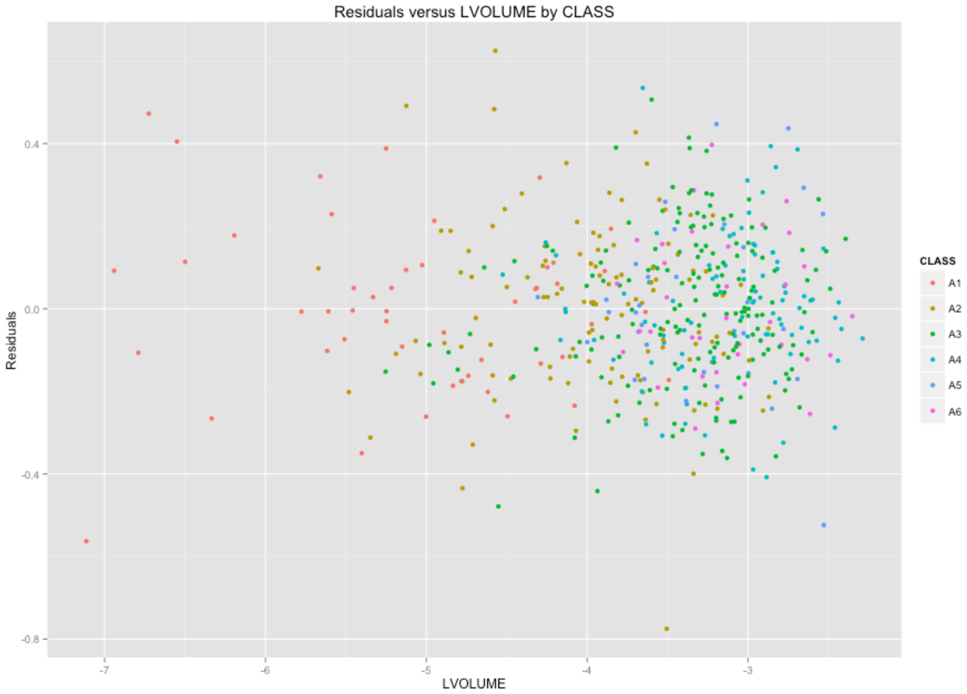
The distribution of residuals differentiated by sex (Figure 6, top panel) and class (Figure 6, bottom panel) are shown below. The boxplots (Figure 6) show that irrespective of sex or class the means of the residuals are relatively similar, despite having some outliers.

Figure 6: Boxplots of residuals differentiated by SEX and CLASS

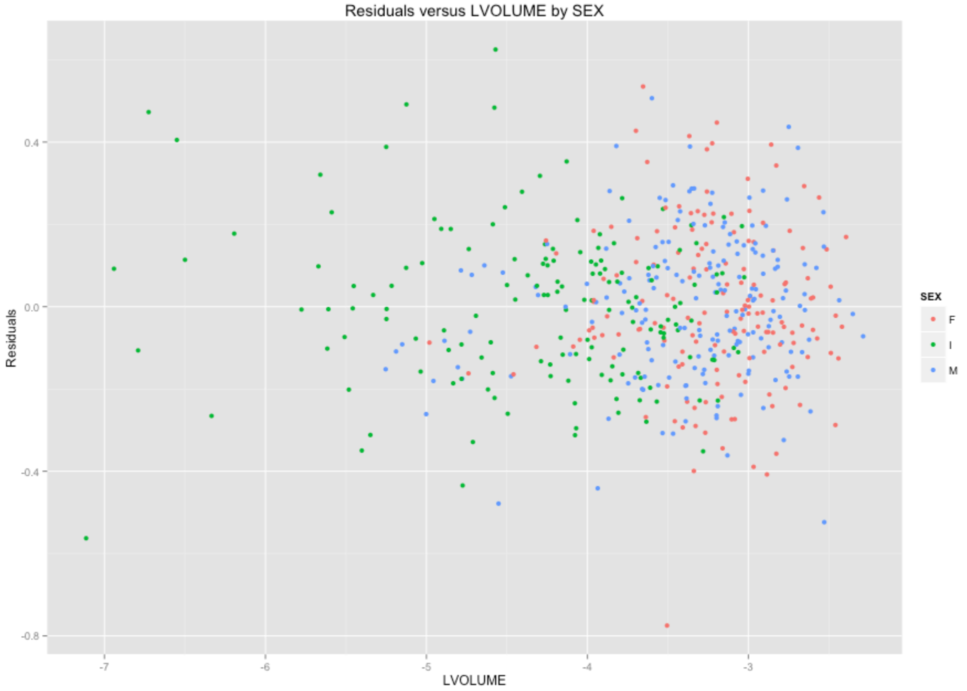


A scatter plot display of the residuals *versus* log volume differentiated by abalone age-class (Figure 7A) and abalone sex (Figure 7B) is also shown. The scatter plots show a relatively broad distribution and scatter of points throughout, and above and below the mid-line. In addition, the distribution of points (see Figures 7A and 7B) do not demonstrate any evidence of other relationships (*e.g.* sinusoidal, linear, *etc.*) present, consistent with a normal distribution with no other extenuating factors influencing the distribution. Overall the regression model fits the data well.

Figure 7: Distribution of residuals *versus* Log volume differentiated by abalone age-class
A. Data points colored by CLASS



B. Data points colored by SEX

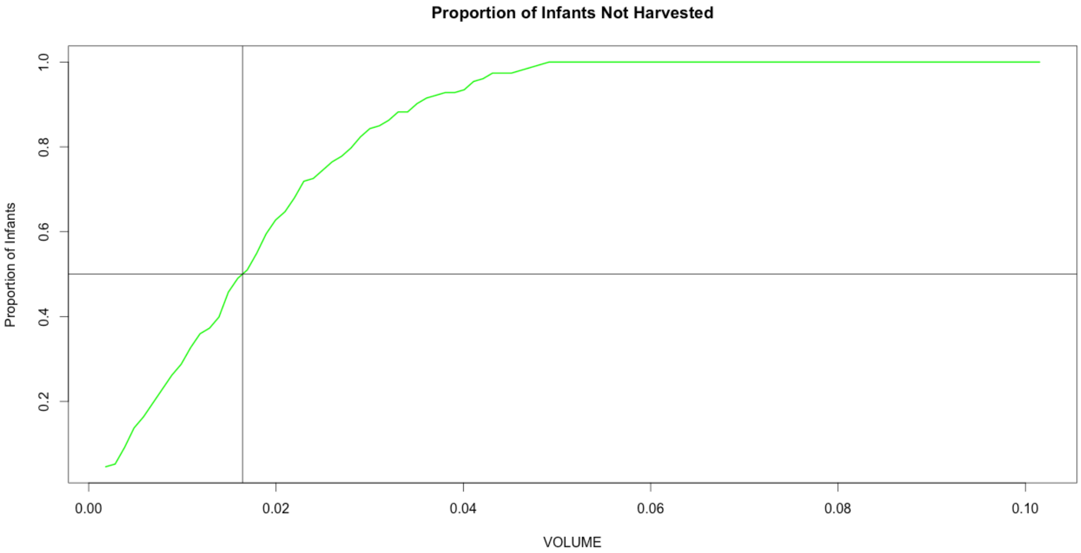


2.8 Decision rule in harvesting abalone by volume threshold

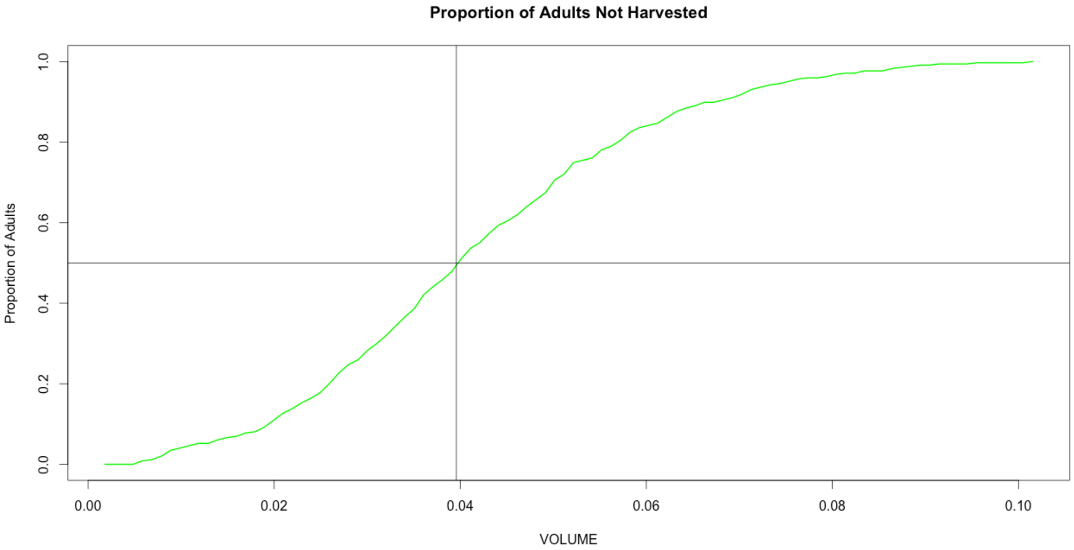
A volume cut-off for selection of infant and adult abalone was devised by plotting proportion of abalone in each category (*e.g.* infants or adults) *versus* volume. These plots are displayed in Figures 7A (infants) and 7B (adults). From the 50% proportion for infants or adults, the commensurate volume cut-off was established (see cross-hairs in Figure 7). For infants and adults, the volume cut-off was 0.0164 mm^3 and 0.0396 mm^3, respectively. The volume cut-off threshold for adults was more than twice that of infants, which is not unreasonable since adults more mature and thus have a bigger volume.

Figure 8: Harvest cut-off threshold based on volumes of infants and adults

A. Infants (cut-off volume = 0.01642322 mm^3)



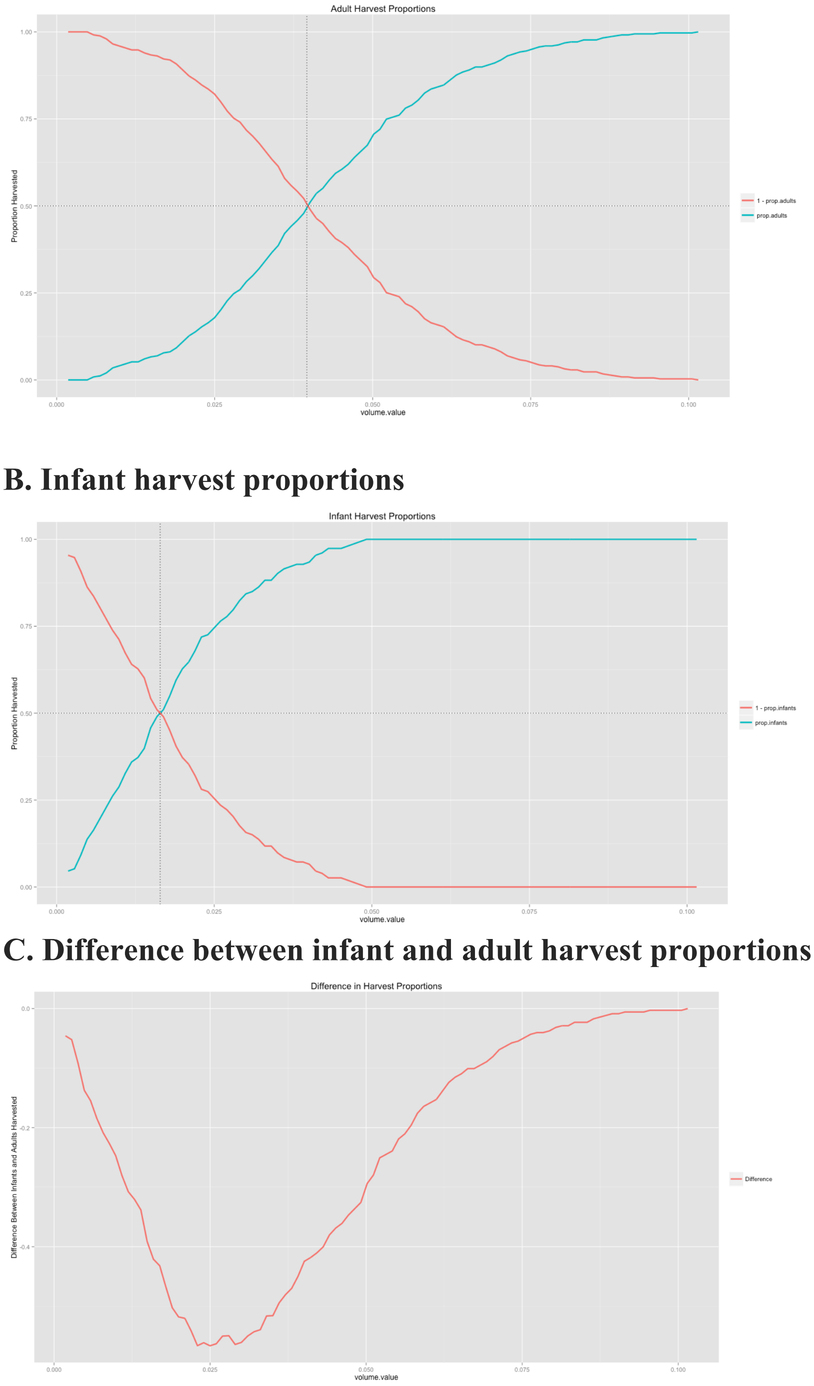
B. Adults (cut-off volume = 0.03958566 mm^3)



2.9 Determination of volumes that maximize differences in proportions of adults and infants

In order to find the set of volumes that maximize differences in proportions of adults and infants, their proportions were calculated and plotted as a function of volume (see Figure 9).

Figure 9: Harvest proportions of adult and infants and their differences



A. Adult harvest proportions

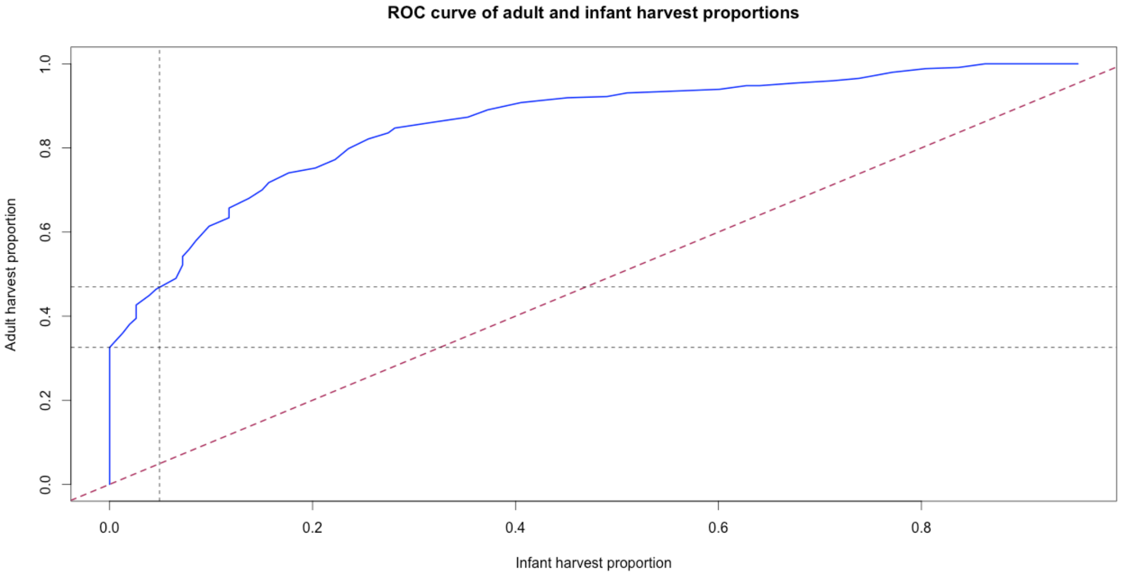
B. Infant harvest proportions

C. Difference between infant and adult harvest proportions

The value of the intercepts or cross over points between the lines in Figures 9A (adults) and 9B (infants) are 0.03958566 mm^3 and 0.01642322 mm^3, respectively, consistent with the threshold cut-offs established earlier (see Section 2.8). Interestingly, from Figure 9C the infant harvest population falls to an inflection point and then adult harvest population rises.

In order to explore trade-offs in the decision rule of what proportions of adults and infants to harvest, a ROC (Receiver Operating Characteristic curve) analysis was performed (Figure 10). The largest cut-off at which no infants are harvested occurs at the volume value of 0.04915 mm^3 (Figure 10, vertical grey dotted line), which correlates to a proportion for adults of 0.47. Although, we are harvesting only approximately half the adults, this seems like a reasonable decision rule, since from earlier we have established that volumes of adults are generally twice that of infants, and since infants are immature their volume is likely to be less anyway. So preserving the infants for further development into adults, in order to generate more harvestable material later, is a tenable argument. The range of potential, tractable adult proportions (and thus their respective volumes) lies in the section of the ROC plot running from 0.3256484 (Figure 10, lower dotted horizontal line) to 0.47 (Figure 10, lower dotted horizontal line).

Figure 10: ROC curve of adult versus infant harvest proportions



2.10 Optimization of harvest volume by screening age-classes

Although a set of volumes and proportions could be used for harvest decision-making purposes, in order to settle on a specific volume an additional evaluation was performed. The additional filter that was applied comprised minimizing harvesting infants in ages-classes A1 and A2, to elicit an optimal harvest volume. Screening of all age-classes to minimize infants in A1/A2 yielded a volume value of 0.035 mm^3. Recall that a volume of 0.049 mm^3 was the level at which no infants were harvested, and this volume at which no A1 and A2 age-classes will be taken, occurs at a value of 0.035 mm^3. Thus, some small proportion of larger volume infants will be taken if the threshold volume is lowered to 0.035 mm^3.

3. Conclusions

Use of physical measurements of abalone for harvesting

This study suggests that specific aspects of abalone physical measurements can be used to devise a simple decision strategy for harvesting abalone (Section 2.8), although there are inherent trade-offs (Sections 2.9 and 2.10). Specifically, maximizing the volume and shuck weight that can be harvested without diminishing the reservoir of infant, immature abalone to preserve future growth stocks. To this end, if the aim is to not to harvest any infant abalone, the criteria noted in Section 2.10 which lowers the threshold to 0.035 mm^3 would potentially result in some larger infants being harvested. The premise to not harvest infants, but rather prefer adults (males and females), is sound both from the maintenance of future abalone stock, but also from an economic perspective as adults have greater shuck and volume compared to infants.

Although this study derived some value from investigating shuck and the volume of abalone, in relation to sex, there may also be some value in exploring other abalone variables (*e.g.* whole weight). This is noted since there is a linear relationship between shuck and whole (and shuck and viscera; see Figure 1 and Table 1). Further whole weight is likely an easier and more rapid method to derive a useful metric. Secondly, although a simple decision rule was devised in this study based a volume criteria to preserve infants and harvest adults, additional criteria might help to provide value to the harvesting heuristics that can be devised.

Difficulties with observational studies

One of the main difficulties with observational studies, such as the one described in this report with abalone, is reliability and quality. Although this was discussed at length in the first report, there are other aspects that need to be addressed. For instance, there may have some aspect(s) of the study that was important for the growth of abalone but that was not monitored by the researchers that performed this study. For example, the position and locational information about where the abalone were sampled from. Specifically, there are examples of the sand or the sub-strata that the abalone grow upon that might influence physical attributes, similarly, if there were location temperature fluctuations in the different cohorts that were sampled. Perhaps some abalone resided closer a hot water outlet than others and that might influence the measured variables. Such uncertainties are always of concern, since they cast doubt not only on the reliability and quality of the study but also on the detail and thoroughness with which it was performed.

What did you learn about abalone with this assignment?

Although I did realize that abalones are edible, I did not appreciate the extent of commercial efforts and interests involved in their husbandry and harvest. I also learnt that abalone have sexes and are not hermaphrodite, have different age-classes, and grow in more tropical parts of the planet. The simple bivariate plots (Figure 1, Table 1) demonstrate the extent of relationships between the physical attributes of abalone. Specifically, while working through the study it was apparent that the volumes and weights of an individual abalone are small and one must harvest a large population to get a ‘tasty morsel’ from them. This suggests that this is a niche market and thus evidently a delicacy in some parts of the world.

APPENDIX 1: R Code Accompanying Report

```
# Assignment 2 401-DL-58
require(moments)
require(ggplot2)
require(rockchalk)

mydata <- read.csv("mydata.csv", sep="")
head(mydata)
tail(mydata)
str(mydata)
summary(mydata)
plot(mydata[,2:8])#review which correlations are linear (Pearson) and which are not (Spearman)

#Q1 Correlations
#Pearson Correlation Coefficients
cor(mydata[,2],mydata[,3], method = "pearson")# LENGTH v DIAM
cor(mydata[,2],mydata[,4], method = "pearson")# LENGTH v HEIGHT
cor(mydata[,3],mydata[,4], method = "pearson")# DIAM v HEIGHT
cor(mydata[,5],mydata[,6], method = "pearson")# WHOLE v SHUCK
cor(mydata[,5],mydata[,7], method = "pearson")# WHOLE v VISCERA
cor(mydata[,5],mydata[,8], method = "pearson")# WHOLE v SHELL
cor(mydata[,6],mydata[,7], method = "pearson")# SHUCK v VISCERA
cor(mydata[,6],mydata[,8], method = "pearson")# SHUCK v SHELL
cor(mydata[,7],mydata[,8], method = "pearson")# VISCERA v SHELL

#Spearman Correlation Coefficients
cor(mydata[,2],mydata[,5], method = "spearman")# LENGTH v WHOLE
cor(mydata[,2],mydata[,6], method = "spearman")# LENGTH v SHUCK
cor(mydata[,2],mydata[,7], method = "spearman")# LENGTH v VISCERA
cor(mydata[,2],mydata[,8], method = "spearman")# LENGTH v SHELL
cor(mydata[,3],mydata[,5], method = "spearman")# DIAM v WHOLE
cor(mydata[,3],mydata[,6], method = "spearman")# DIAM v SHUCK
cor(mydata[,3],mydata[,7], method = "spearman")# DIAM v VISCERA
cor(mydata[,3],mydata[,8], method = "spearman")# DIAM v SHELL
cor(mydata[,4],mydata[,5], method = "spearman")# HEIGHT v WHOLE
cor(mydata[,4],mydata[,6], method = "spearman")# HEIGHT v SHUCK
cor(mydata[,4],mydata[,7], method = "spearman")# HEIGHT v VISCERA
cor(mydata[,4],mydata[,8], method = "spearman")# HEIGHT v SHELL

#Q2 Boxplots: analyze SHUCK differentiated by CLASS and SEX
females <- subset(mydata, subset = (SEX =="F"))
infants <- subset(mydata, subset = (SEX =="I"))
males <- subset(mydata, subset = (SEX =="M"))
#Create 3 x 1 matrix of 18 boxplots
par(mfrow = c(1,1))
boxplot(SHUCK~CLASS, data = females, col = "grey", main = "Female SHUCK", ylab = "SHUCK", ylim = c(0,1.3))
boxplot(SHUCK~CLASS, data = infants, col = "grey", main = "Infant SHUCK", ylab = "SHUCK", ylim = c(0,0.6))
boxplot(SHUCK~CLASS, data = males, col = "grey", main = "Male SHUCK", ylab = "SHUCK", ylim = c(0,1.2))
par(mfrow = c(1,1))

#Q3 Pearson chi square statistic
shuck <- factor(mydata$SHUCK > median(mydata$SHUCK), labels=c("below","above"))
volume <- factor(mydata$VOLUME > median(mydata$VOLUME), labels=c("below","above"))
shuck_volume <- addmargins(table(shuck,volume))

#function which calculates chi-squared value given a 2x2 matrix with marginals
matrix <- function(x){
  e11 <- x[3,1]*x[1,3]/x[3,3]
  e12 <- x[3,2]*x[1,3]/x[3,3]
  e21 <- x[3,1]*x[2,3]/x[3,3]
  e22 <- x[3,2]*x[2,3]/x[3,3]
}
```

```
q <- matrix(shuck_volume)# q=124.5 (chi-squared statistic)
pchisq(q,1,lower.tail=FALSE)# p-value: p=6.547914e-29 (reject null hypothesis)

#Q4 Analysis of variance

aov.shuck <- aov(SHUCK~CLASS+SEX+CLASS*SEX, mydata)#with interaction CLASS*SEX
summary(aov.shuck)

aov.shuckmodel <- aov(SHUCK~CLASS+SEX, mydata)#without interaction CLASS*SEX
summary(aov.shuckmodel)

# Statistically significant F-test results.  Perform TukeyHSD.
TukeyHSD(aov.shuckmodel)

#Q5 ggplots of SHUCK versus VOLUME and their respective log plots
require(ggplot2)
library(gridExtra)

#create LSHUCK and LVOLUME as category variables of SHUCK and VOLUME, respectively and add to mydata
LSHUCK <- log(mydata[,6])#calculate ln(shuck)
LVOLUME <- log(mydata[,11])#calculate ln(volume)
mydata <- data.frame(mydata, LSHUCK, LVOLUME)#added LSHUCK and LVOLUME to mydata for convenience
str(mydata)#check LSHUCK/LVOLUME was added to mydata

grid.arrange(ggplot(data = mydata, aes(x = VOLUME, y = SHUCK)) + geom_point(aes(color = CLASS),size = 2) +
ggtitle("Shuck versus Volume by Class"),
          ggplot(data = mydata, aes(x = LVOLUME, y = LSHUCK)) + geom_point(aes(color = CLASS),size = 2) +
ggtitle("LogShuck versus LogVolume by Class"), nrow = 1)


#Q6 Linear regression
model <- lm(LSHUCK~LVOLUME+CLASS+SEX, mydata)
summary(model)

#Q7 Analysis of residuals

r <- residuals(model)
fitt <- fitted(result)

par(mfrow = c(1,1))
hist(r, col = "red", main = "Histogram of Residuals", xlab = "Residual")
par(mfrow = c(1,1))

qqnorm(r, col = "red", pch = 20, main = "QQ Plot of Residuals")
qqline(r, col = "blue", lty = 2, lwd = 2)

skewness(r)
kurtosis(r)

#ggplot: Create an out regression object
out <- data.frame(mydata$LVOLUME,mydata$CLASS,mydata$SEX)
out <- data.frame(out,r)
colnames(out) <- c("LVOLUME","CLASS","SEX","RESIDUAL")
head(out)
ggplot(out, aes(x = LVOLUME, y = RESIDUAL)) + geom_point(aes(color = CLASS)) + labs(x = "LVOLUME", y =
"Residuals")

ggplot(out, aes(x = SEX, y = RESIDUAL, fill = SEX)) + ggtitle("Residuals versus SEX") +
  geom_boxplot(aes(fill = factor(SEX))) + scale_fill_manual(values = c("red", "green", "blue"))

ggplot(out, aes(x = CLASS, y = RESIDUAL, fill = CLASS)) + ggtitle("Residuals versus CLASS") +
  geom_boxplot(aes(fill = factor(CLASS))) + scale_fill_manual(values = c("cadetblue", "blue4", "darkred", "green",
"blue", "red"))


#ggplot of residuals vs LVOLUME. Color data points by CLASS.
ggplot(out, aes(x = LVOLUME,y = RESIDUAL)) + geom_point(aes(color = CLASS)) +
  labs(x = "LVOLUME", y = "Residuals") +
  theme(legend.direction = "vertical", legend.position = "right") + ggtitle("Residuals versus LVOLUME by CLASS")

#Color data points by SEX
ggplot(out, aes(x = LVOLUME,y = model$residuals)) + geom_point(aes(color = SEX)) +
  labs(x = "LVOLUME", y = "Residuals") +
  theme(legend.direction = "vertical", legend.position = "right") + ggtitle("Residuals versus LVOLUME by SEX")


#Q8
idxi <- mydata[,1]=="I"
idxa <- mydata[,1]!="I"
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 <- length(mydata[idxi,1]) # This value must be changed for adults.
totala <- length(mydata[idxa,1]) # Value changed for adults.

for (k in 1:100) {
  value <- min.v + k*delta
  volume.value[k] <- value
  prop.infants[k] <- sum(mydata$VOLUME[idxi] <= value)/total
  prop.adults[k] <- sum(mydata$VOLUME[idxa] <= value)/totala
}

#Plot infants
n.infants <- sum(prop.infants <= 0.5)
split.infants <- min.v + (n.infants + 0.5)*delta # This estimates the desired volume.
plot(volume.value, prop.infants, col = "green", main = "Proportion of Infants Not Harvested", xlab="VOLUME",
ylab="Proportion of Infants", type = "l", lwd = 2)
abline(h=0.5)
abline(v = split.infants)

#Plot adults
n.adults <- sum(prop.adults <= 0.5)
split.adults <- min.v + (n.adults + 0.5)*delta # This estimates the desired volume.
plot(volume.value, prop.adults, col = "green", main = "Proportion of Adults Not Harvested", xlab="VOLUME",
ylab="Proportion of Adults", type = "l", lwd = 2)
abline(h=0.5)
abline(v = split.adults)

#split value (infants) = 0.01642322 volume @ 50% popn infants
#split value (adults) = 0.03958566 volume @ 50% popn adults

#Q9 Part A

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

total.infants <- length(mydata[idxi,1])  # This value must be changed if adults are being considered.
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
}

difference <- (1-prop.infants)-(1-prop.adults) #

#plot (1) 1-prop.adults vs volume.value
ggplot(mydata, aes(x = volume.value)) + ggtitle("Adult Harvest Proportions") +
  geom_line(aes(y = (prop.adults), color = "prop.adults"), size = 1.1) +
  geom_line(aes(y = (1 - prop.adults), color = "1 - prop.adults"), size = 1.1) +
  geom_hline(yintercept=0.5, linetype="dotted") +
  geom_vline(xintercept=split.adults, linetype="dotted") +
  theme(legend.title=element_blank()) +
  ylab("Proportion Harvested")

#plot (2) 1-prop.infants vs volume.value
ggplot(mydata, aes(x = volume.value)) + ggtitle("Infant Harvest Proportions") +
  geom_line(aes(y = (prop.infants), color = "prop.infants"), size = 1.1) +
  geom_line(aes(y = (1 - prop.infants), color = "1 - prop.infants"), size = 1.1) +
  geom_hline(yintercept=0.5, linetype="dotted") +
  geom_vline(xintercept=split.infants, linetype="dotted") +
  theme(legend.title=element_blank()) +
  ylab("Proportion Harvested")

#plot (2) difference (prop.infants-prop.adults) vs volume.value
ggplot(mydata, aes(x = volume.value)) + ggtitle("Difference in Harvest Proportions") +
  geom_line(aes(y = (difference), color = "Difference"), size = 1.1) +
  theme(legend.title=element_blank()) +
  ylab("Difference Between Infants and Adults Harvested")

#Q9 Part B ROC curve: 1-prop.adults vs 1-prop.infants
plot(1-prop.infants,1-prop.adults, col = "blue", lwd = 2, type = "l",
     main = "ROC curve of adult and infant harvest proportions", ylab = "Adult harvest proportion",
     xlab = "Infant harvest proportion")
abline(a=0, b=1, col = "maroon", lty = 2, lwd = 2)

#Identify the largest infant
max(mydata$VOLUME[mydata$SEX == "I"]) # [1] 0.0485
#smallest volume.value that corresponds to a harvest of zero infants
min(volume.value[(1 - prop.infants) == 0]) # 0.04915275, value can be passed to the ROC curve
which.max(1 - prop.infants == 0) # The 48th element in (1 - prop.infants) is the first zero
#add cross-hairs into plot
abline(h=0.47, lty =2)# value of y-xis intercept with x=0.04915275
abline(h=0.3256484, lty =2)#
abline(v = 0.04915275, lty = 2)
#Proportion of adults harvested at the volume.value threshold (zero infants harvested)
(1 - prop.adults)[48] # 0.3256484

#Q10 Minimizing harvesting of age-classes A1 and A2
cutoff <- 0.035 # this cutoff value works.
index.A1 <- (mydata$CLASS=="A1")
indexi <- index.A1 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A1)# [1] 0
index.A2 <- (mydata$CLASS=="A2")
indexi <- index.A2 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A2)# [1] 0
index.A3 <- (mydata[,10]=="A3")
indexi <- index.A3 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A3)# [1] 0.04545455
```

```
index.A4 <- (mydata[,10]=="A4")
indexi <- index.A4 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A4)# [1] 0.03296703
index.A5 <- (mydata[,10]=="A5")
indexi <- index.A5 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A5)# [1] 0.02857143
index.A6 <- (mydata[,10]=="A6")
indexi <- index.A6 & idxi
sum(mydata[indexi,11] >= cutoff)/sum(index.A6)# [1] 0.05714286
```