Predict 401DL Data Analysis Project Assignment #2

Project Assignment #2 (100 points due the end of session 10)

Overview

The first assignment dealt with physical measurement and age prediction. In this assignment a different questions will be considered. To what extent can physical measurements be used as a basis for harvesting abalone? Can a useful decision rule be constructed? The first part of this assignment continues the EDA already started. Analysis of variance, linear regression and other forms of analysis will be performed. Conclusions will be sought regarding decision rules for abalone harvesting. Different decision rules can be presented and discussed in the report. Snipets of code are supplied for each of the steps in the assignment. This report should comply with the report template and be a separate report in its own right. Results from the first assignment may be referenced as needed.

Report Template
Assignment #
(Enter your name)

Introduction:

The introduction should describe the purpose of the assignment. It should be clear to me that you understand the rationale behind the steps in the assignment. State the objective of the report.

Results:

Display the results for your assignment and comment. Your discussion should be intertwined with (or linked to) the results of your analysis with the discussion on or near the page containing analytical results. Reports should be written to inform a business reader, not overwhelm the reader with distracting detail. Present appropriate graphics and tables. **All graphics and tables should be identified by number and title for reference purposes.** Do not show unnecessary R output. For example, you should never include a printout of the data set or extensive tables of summary statistics in the body of the report. Use the appendix for this information. Intermediate steps in the analysis could also be eliminated. While graphical output may take up space, such output can be reduced in size provided some important feature is not lost or obscured.

Conclusions:

Summarize assignment results. Responses to the questions placed at the end of the assignment should appear here. State your conclusions succinctly and clearly.

Submission:

The report should be submitted in pdf format. **R code should not appear in the body of the report.** Attach your R code as an appendix. This should be all the code used for your report.

Predict 401DL Data Analysis Project Assignment #2

1) Construct a plot matrix using plot(mydata[,2:8]). Determine for which pairs of variables a Pearson Correlation Coefficient or a Spearman Correlation Coefficient is appropriate. Compute and present these coefficients in a table.
   a) It is possible to use cor() for this part. Cor() can be used to calculate correlations using two dataframes. Both Pearson and Spearman coefficients are options. Check help(cor).
   b) For example, cor(x,y,method=c("spearman")) would calculate the Spearman Correlation Coefficient for the variables present in data frames x and y.

2) SHUCK represents harvestable meat. Form a matrix of boxplots showing SHUCK differentiated by CLASS and SEX. There will be 18 boxplots. Each age CLASS is a separate population due to factors present during the life of the population. What does this reveal about the variability in abalone growth and aging?
   a) A convenient way to do this is with indices and the par() function with mfrow = c(3,1).
   b) For example, idxi <- mydata[,1] == "I" creates a logical vector of TRUE and FALSE values which can be used for selecting observations i.e. mydata$SHUCK[idxi].

3) Write a function in R to calculate the Pearson chi square statistic on 2x2 contingency tables which have the marginal totals. Test for independence using this function on SHUCK and VOLUME. Show the chi square value and p-value in a table. Discuss the results.
   a) An example of code to include in the function is shown below. The function would start with a table that has marginal totals. These statements calculate the expected value for each cell in the table. Using these statements the Pearson Chi Square Statistic may be calculated within the function and the value returned.

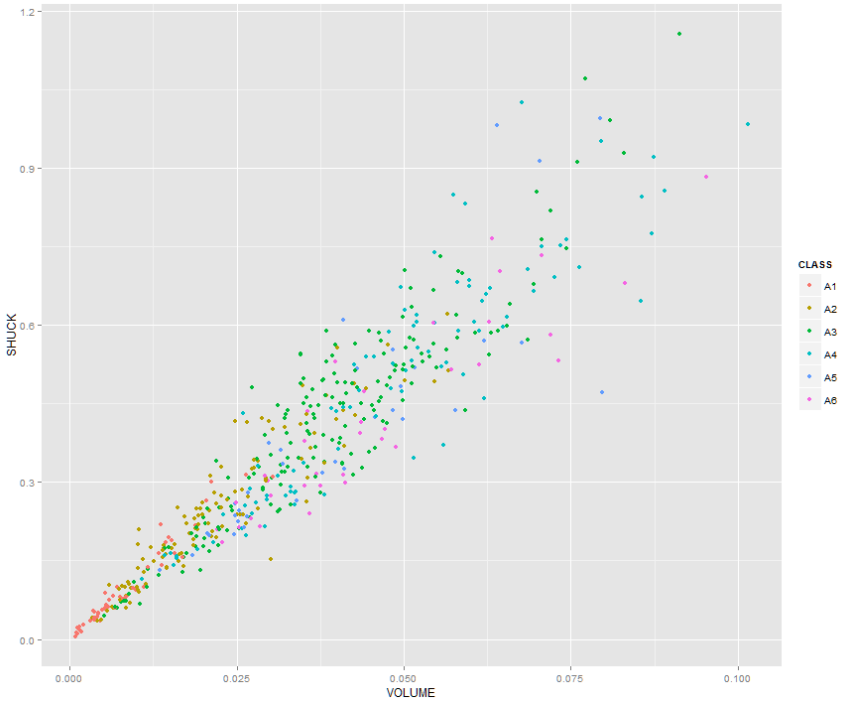
```
# To be used with 2x2 contingency tables that have margins added.
e11 <- x[3,1]*x[1,3]/x[3,3]
e12 <- x[3,2]*x[1,3]/x[3,3]
e21 <- x[3,1]*x[2,3]/x[3,3]
e22 <- x[3,2]*x[2,3]/x[3,3]
```

   b) To dichotomize SHUCK and VOLUME use statements similar to this:
   shuck <- factor(mydata$SHUCK > median(mydata$SHUCK, labels=c("below","above")
   c) To generate a table use shuck_volume <- addmargins(table(shuck,volume)) This would generate a table which could be submitted to the user-supplied chi-square function.
   d) pchisq(q, 1, lower.tail = FALSE) will give the p-value given a quantile q.

4) Perform an analysis of variance with aov() on SHUCK using CLASS and SEX as the grouping variables. Assume equal variances. First use the model with an interaction term CLASS*SEX and then without CLASS*SEX. Use summary() to obtain the analysis of variance table. Follow up with the TukeyHSD() function for the model without CLASS*SEX. Interpret the results. (TukeyHSD() will adjust for unequal sample sizes.).

Predict 401DL Data Analysis Project Assignment #2

5) Use ggplot2 to form a scatterplot of SHUCK versus VOLUME and a scatterplot of their logarithms labeling the variables as L_SHUCK and the latter as L_VOLUME. Use color to differentiate CLASS in the plots. Compare the two scatterplots. Where do the various CLASS levels appear in the plots? What are the implications of the observed pattern?
   a) ggplot2 must be installed from CRAN. Use library(ggplot2) prior to executing code.
   b) Here is an example of what should be produced using ggplot.



6) Regress L_SHUCK as the dependent variable on L_VOLUME, CLASS and SEX. Follow the steps shown in Section 16.1 of Lander but use the following multiple regression model: L_SHUCK~L_VOLUME+CLASS+SEX. Use summary() on the object and comment.
   a) summary() applied to the regression object produces significance test results.

7) Perform an analysis of the residuals. If "out" is the regression object, use out$residuals and construct a histogram and QQ plot. Compute the skewness and kurtosis. The residuals should approximate a normal distribution. Describe the distribution of residuals. Use ggplot. Plot the residuals versus L_VOLUME coloring the data points by CLASS and a second time coloring the data points by SEX. Also use ggplot to present boxplots of the residuals differentiated by SEX and CLASS. How well does the regression model fit the data?
   a) The package "moments" will need to be installed on the computer.
   b) Here is the type of code needed: ggplot(out, aes(x = L_VOLUME,y = out$residuals)) + geom_point(aes(color = CLASS)) + labs(x = "L_VOLUME", y = "Residual")

8) The next portion is a study of the potential use of VOLUME as a means of selecting abalone for harvest. There is a tradeoff faced in managing the abalone harvest. The infant population must be protected since that represents the future harvest. On the other hand, the harvest should be designed to be efficient with a sufficient yield to justify the effort. This part will

use a very simple decision rule. If the VOLUME for an abalone is below a specified volume, that individual will not be harvested. If above, it will be harvested.

a) This part of the assignment will calculate the proportion of infant abalone which fall beneath a specified volume or "cutoff". A series of volumes covering the range from the minimum volume to the maximum volume will be used in a "for loop". This calculation will show how the harvest proportion of infants changes as the volume used for assessment changes. Example code for doing this is supplied below.

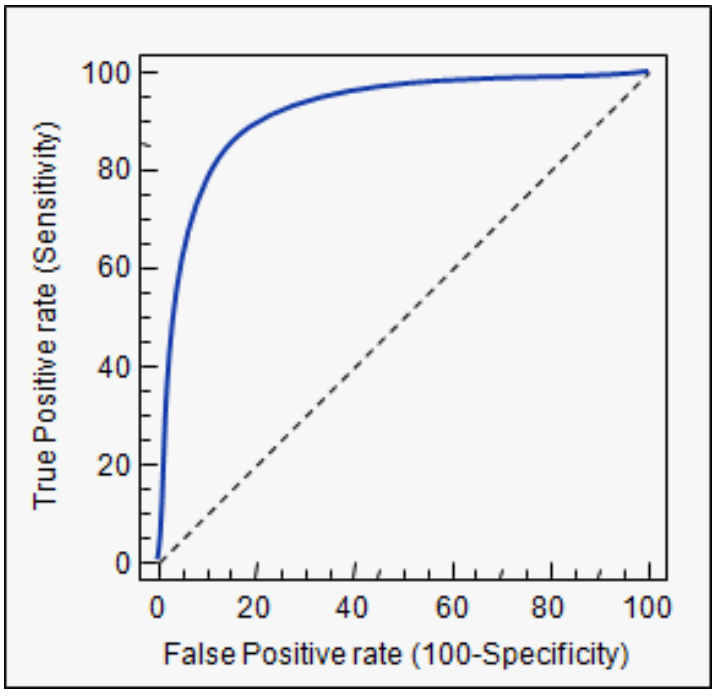
```
idxi <- mydata[,1]=="I"
idxf <- mydata[,1]=="F"
idxm <- mydata[,1]=="M"

max.v <- max(mydata$VOLUME)
min.v <- min(mydata$VOLUME)
delta <- (max.v - min.v)/100
prop.infants <- numeric(0)
volume.value <- numeric(0)
total <- length(mydata[idxi,1])  # This value must be 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
}
If an abalone has a volume below a specified cutoff, it is not harvested.  The vector
# prop.infants shows the impact of increasing the volume cutoff for harvesting.  An
# increasing proportion of the infant population does not qualify for harvesting.

# This shows how to determine volume levels which "split" the population at
# any specified harvest proportion.  For example using a 50% harvest of infants, the
corresponding volume is:

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",
     type = "l", lwd = 2)
abline(h=0.5)
abline(v = split.infants)
```

8 continued) Modify the code. This time instead of counting infants, count adults. Present a plot showing the adult proportions versus volume. Compute the 50% "split" volume value for the adults and show on the plot similarly to the plot for infants. Compare the "split" values for infants and adults and comment on the difference.

Predict 401DL Data Analysis Project Assignment #2

It is essential that the males and females be combined into a single count as "adults" for computing the proportion for "adults". Part #9) will require plotting of infants versus adults. For this plotting to be accomplished, a "for loop", similar to the one above, must be used to compute the adult harvest proportions. It must use the same value for the constants min.v and delta. It must also use the statement "for (k in 1:100)". Otherwise, the resulting adult proportions cannot be directly compared to the infant proportions.

9) This part will address the determination of a volume or set of volumes which maximize the difference in percentages of adults and infants. To calculate this result, the proportions from #8) must be used. These proportions must be converted from "not harvested" proportions to "harvested" proportions by subtracting (1-prop.infants) from (1-prop.adults). (The reason the proportion for infants drops sooner than adults, is that infants are maturing and becoming adults with larger volumes.) From the plot generated, select a range of values which have the potential to maximize the difference in harvest proportions.
   a) Present three plots: 1) a plot of (1-prop.adults) versus volume.value, and 3) a plot of (1-prop.infants) versus volume.value, and 3) a plot of the difference (prop.infants-prop.adults) versus volume.value. Use volume.value from #8.
   b) Now construct an ROC curve by plotting (1-prop.adults) versus (1-prop.infants). Each point which appears corresponds to a particular volume.value. This curve is used often to illustrate the tradeoffs involved with decision rules. Find the largest cutoff for which no infant is harvested. Report this value and the 1-prop.aduts value. Comment. Does this seem to be a reasonable choice for a decision rule? Here is an example of an OC curve.



For abalone harvesting, if an infant is harvested according to the volume decision rule, that is a false positive. The infant is being treated as an adult. The true positive rate is the proportion of adults harvested.

Predict 401DL Data Analysis Project Assignment #2

10) In #9) a set of volumes were identified which could be used for decision making during abalone harvest. To settle on one volume, which may not be included in the volumes calculated in #9), an additional evaluation will be performed. Harvesting of infants in classes A1 and A2 must be minimized. With these data it is possible to find volume cutoffs for which this infant harvest is zero. The minimum volume that does this is to be determined.
    a) Using the code supplied, find the largest volume which produces a zero harvest of infant abalone in classes A1 and A2. This can be accomplished by substituting values into the code supplied. The level of precision required can be achieved with values such as 0.036. More than that is unnecessary. Report your result and discuss how this compares to the volumes identified from the plot of differences in harvesting proportions shown in #9).
    b) Here is an example of the code with the result for 0.036. Smaller values will work.

```
> cutoff <- 0.036 # Example volume cutoff value. Hint, a smaller volume cutoff 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
```

In your report conclusions, discuss what you have learned about the use of physical measurements as a basis for abalone harvesting. Consider feasibility and the tradeoffs involved. How much reliance would you place in the volume cutoff determined in step #10)? What else might be done to verify this conclusion? If you have specific harvesting recommendations or strategy, discuss them. Also discuss what you see as difficulties in analyzing data from an observational study involving different classes or cohorts of subjects. What cautions come to mind? What did you learn about abalone with this assignment?