# Introduction

This addendum extends the second data analysis in three ways. First, an example is given to show how a different binary classification rule can be devised. Second, a hypothetical “gain” (or loss function depending on your viewpoint) is used to select a cutoff for harvesting abalones. Third, an ROC chart is used to compare two binary classifiers.

# Another Binary Classifier

In the second data analysis, VOLUME was used for making harvesting decisions. An alternative approach is to use WHOLE and LENGTH. First, a regression model relating WHOLE to LENGTH will be developed. Second, a logistic model will be devised to score abalones based on variables resulting from the regression model.

# Simple Linear Regression

Transforming WHOLE and LENGTH into base ten logarithms allows simple linear regression using the function lm(). The code required follows.

> L\_WHOLE <- log10(mydata$WHOLE)

> L\_LENGTH <- log10(mydata$LENGTH)

> test <- data.frame(L\_WHOLE, L\_LENGTH)

> test$type <- combineLevels(mydata$SEX, levs = c("M","F"), "ADULT")

The original levels F I M have been replaced by I ADULT >

> object <- lm(L\_WHOLE ~ L\_LENGTH)

> summary(object)

Call:

lm(formula = L\_WHOLE ~ L\_LENGTH)

Residuals:

Min 1Q Median 3Q Max

-0.22953 -0.03738 -0.00184 0.03636 0.38630

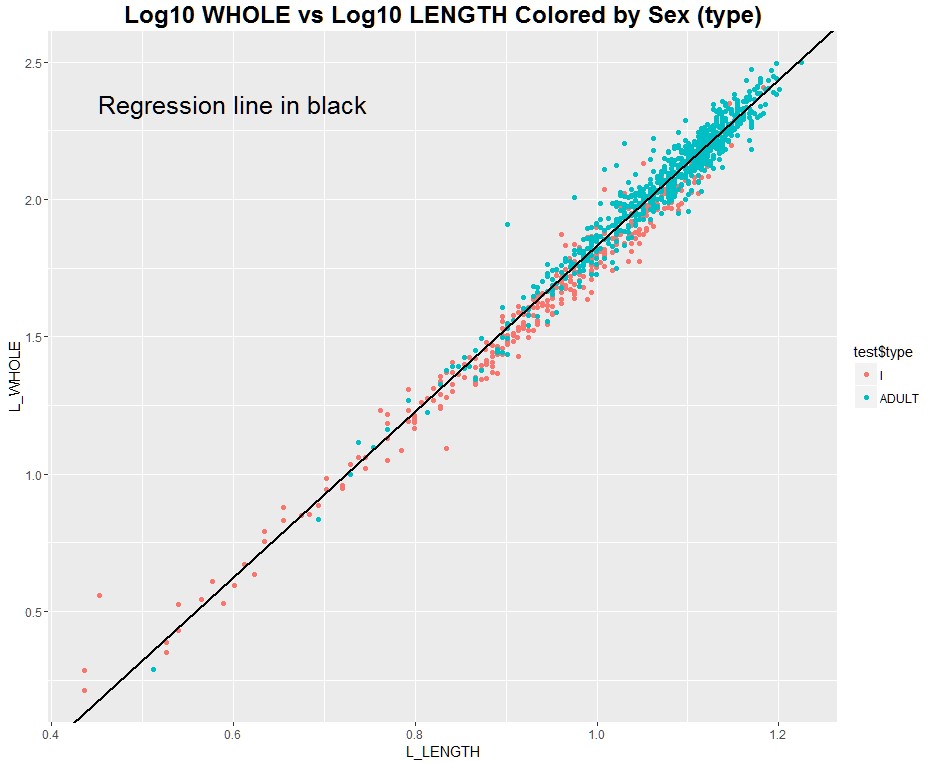
Coefficients:

Estimate Std. Error t value Pr(>|t|)

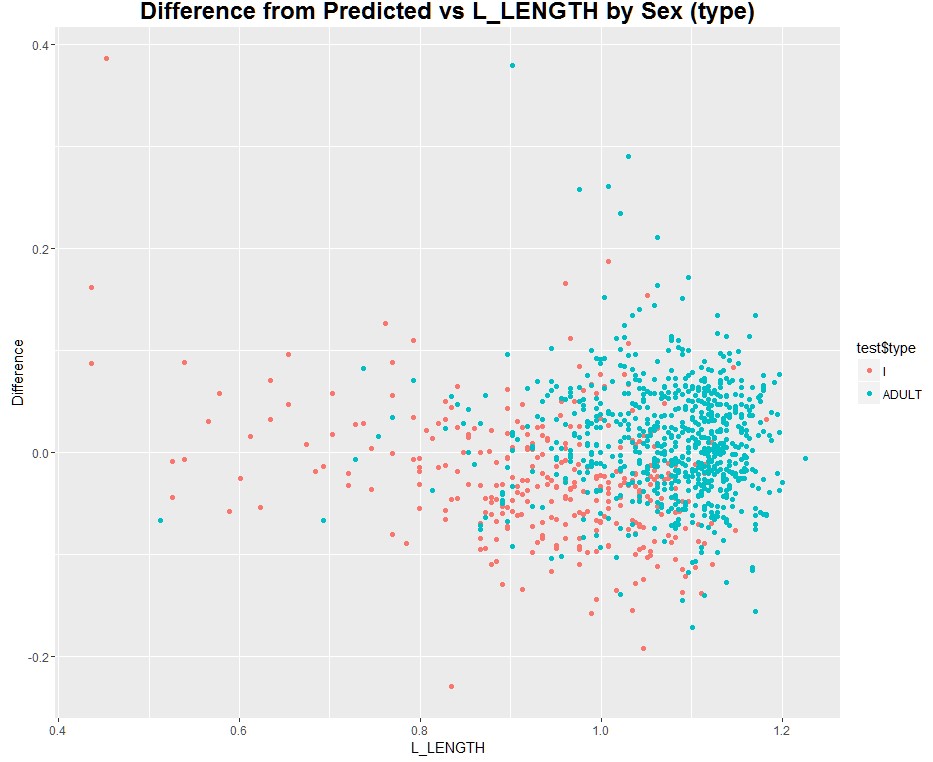
(Intercept) -1.19249 0.01673 -71.28 <2e-16 \*\*\*

L\_LENGTH 3.01732 0.01613 187.10 <2e-16 \*\*\*

The resulting regression model is displayed in the following scatter plot. This model provides a predicted value for L\_WHOLE given a value for L\_LENGTH. A residual equal to the difference between observed and predicted L\_WHOLE values, can be computed for each abalone.



A plot of the residuals versus L\_LENGTH appears below.



EDA suggests older abalones tend to have positive residuals and younger abalones negative residuals. This observation will be the basis for a logistic model.

# Logistic Regression

Using logistic regression, it is possible to predict the probability of harvesting an ADULT abalone as a function of L\_LENGTH and the difference between L\_WHOLE and the predicted value for L\_WHOLE from simple linear regression. It is necessary to use the glm() function to develop this model.

model <- glm(type ~ L\_LENGTH + Difference, family=binomial(link='logit'), data=test)

d <- -12.8924 + 13.5396\*L\_LENGTH +15.5612\*Difference

The variable “Difference” is the residual or difference between L\_WHOLE and a predicted value for L\_WHOLE. The variable “d” is not a predicted probability. It is the logit of the probability of interest. The logit is the natural log of the odds of selecting an ADULT abalone. The variable d can be used in place of VOLUME, and virtually identical coding can be used to determine cutoffs for d, and an ROC curve.

For example, if d = 1.0 the odds are exp(1.0) = 2.718, and the probability an abalone with this score being an ADULT is 0.73. There are different approaches for deciding on a cutoff for d. If a gain (or a loss) function is available, the choice becomes easier.

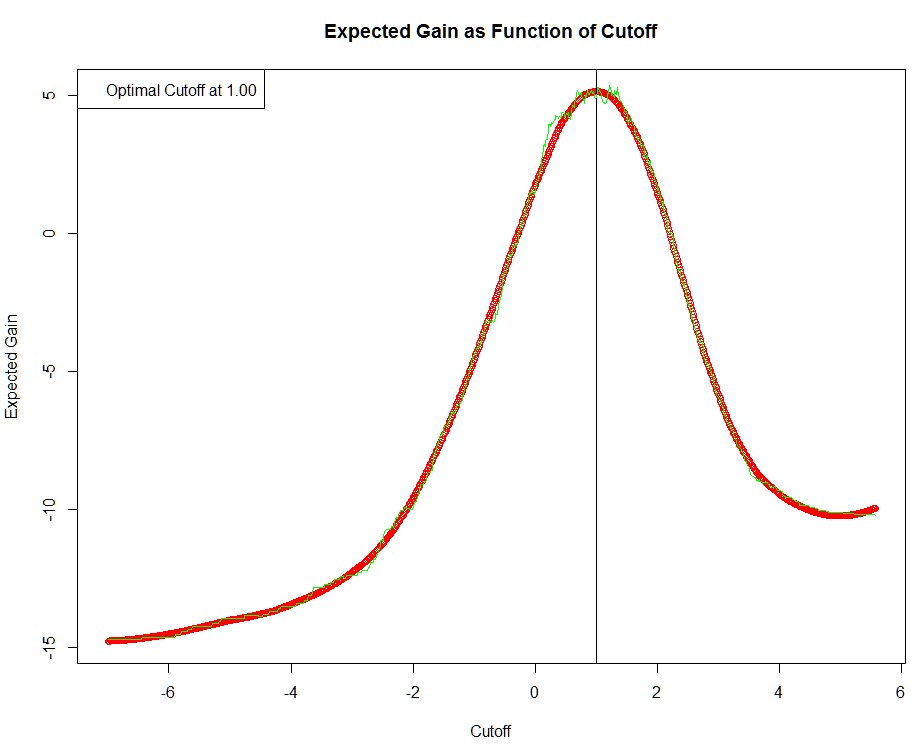
# Gain Function Example

Suppose an ADULT abalone can be sold for $25. Also assume there is a penalty assessed of $100 for each non-ADULT or infant abalone. This could be an imposed fine. Using the variables prop.adults and prop.infants in combination with the prior population proportions for each, Bayes’ Theorem can be used to calculate the expected posterior gain.

profit <- 25 # gain from sale of ADULT abalone penalty <- -100 # loss for harvesting an infant (penalty)

gain <- profit\*(1-prop.adults)\*p.adults+penalty\*(1-prop.infants)\*p.infants

The profit and loss numbers used are purely hypothetical and result in odds of an ADULT harvest equal to 3.31 at a cutoff of 1.003583. This corresponds to an ADULT selection probability equal to 0.77. The probability of harvesting an adult abalone goes up for scores greater than 1.003583. Some infants will still be harvested regardless. The gain function is shown in the following plot. Using this optimal cutoff, the classification frequencies for the available abalones are shown in the table below the plot.

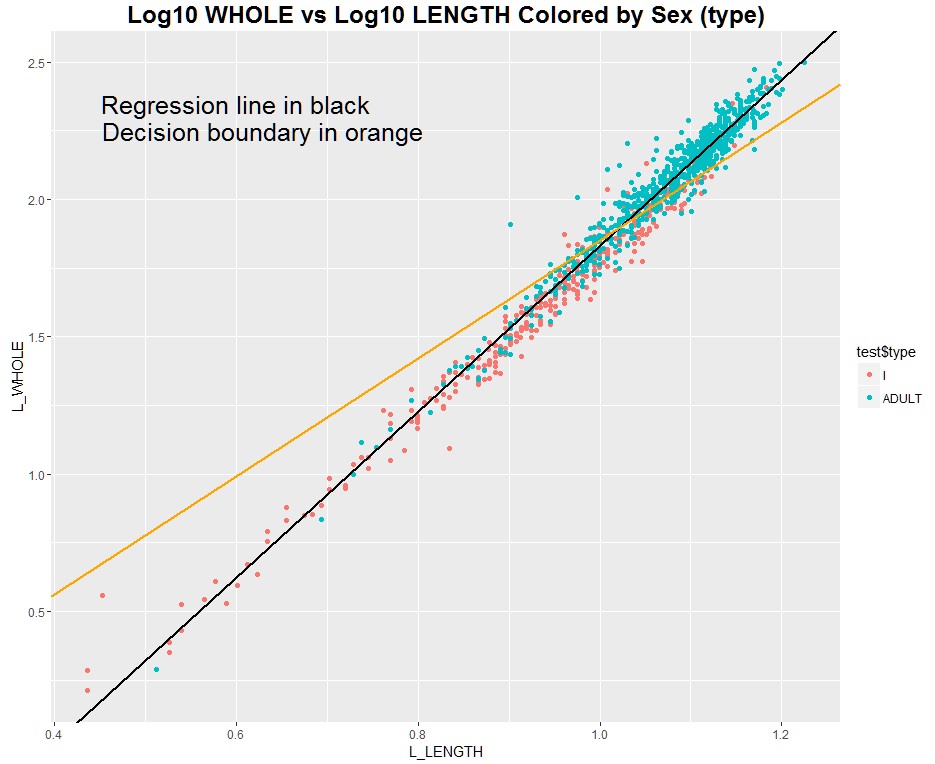


I ADULT Sum

Classified as Infant 269 164 433

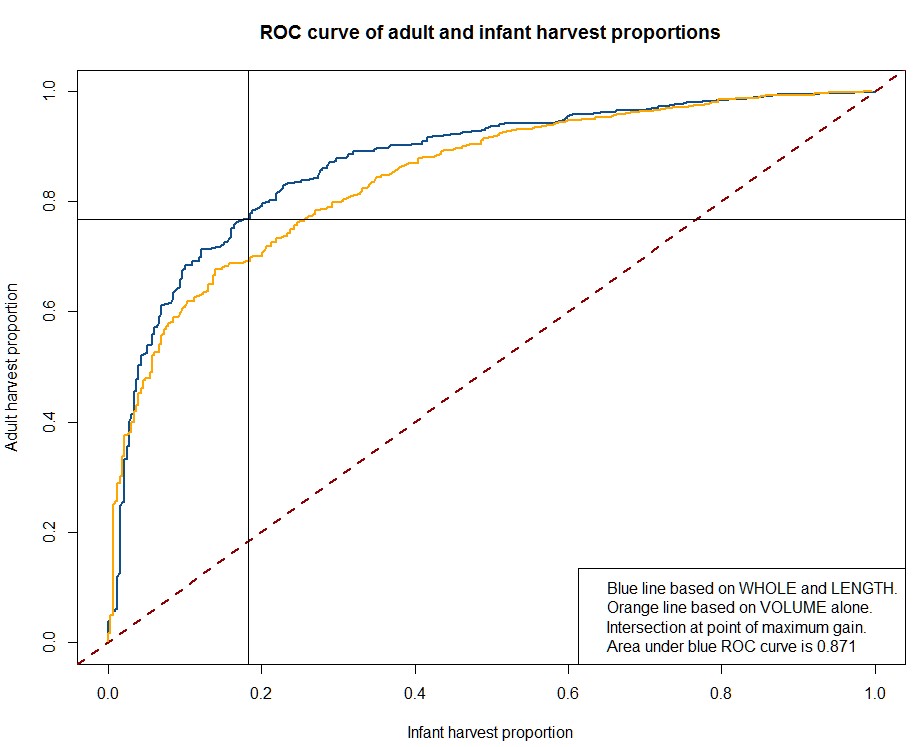
Classified as Adult 60 543 603 Sum 329 707 1036

The decision boundary for classification is shown as an orange line in the plot below.



# ROC Curve Comparison

The foregoing determinations can be displayed in an ROC curve and compared to the performance of VOLUME used as a single variable for harvesting. The area under the blue curve corresponds to the decision rule devised by logistic regression. The area under the blue curve is 0.871. Different choices for the profit and loss numbers would switch the location of the intersection of these rates.



# Concluding Remarks

There are many different binary decision rules. This addendum illustrates how two different binary decision rules might be compared. Changing the composition of the underlying abalone population in terms of age structure, and infant versus adult proportions would affect results. The same is the case with the profit and penalty values. These methods are flexible and could be adapted to differing circumstances. Regardless, any resulting model would have to be evaluated in the field and validated.