MSDS 401DL Data Analysis Project Assignments

Data Analysis Overview:

Two data analysis projects using abalone data are required in this course. The first project entails exploratory data analysis. The second project involves statistical inference using analysis of variance and linear regression. Binary decision rules will be evaluated and a Receiver Operating Characteristic (ROC) curve developed. These projects will require application of course concepts and use of R Studio and R Markdown resulting in submission of a .Rmd file and .html document.

Topics covered during Sessions 1-5 pertain to the first analysis project. Assignment submissions are due the end of Session 5. The second project uses topics covered during Sessions 6-9. Assignment submissions are due at the end of the course.

Overall Background:

We are bombarded daily with statements or claims arising from surveys and studies that use data to generate statistics. In a world where data are becoming more abundant every day, as an educated consumer, it is essential to think critically about the information we receive and the decisions that will be made based on that information. Part of this involves considering the source of the data, how it was collected, how it was analyzed and whatever limitations there may be to the conclusions reached and claims being made.

As analysts we must be prepared to apply sound statistical and critical thinking to a wide variety of situations. Being able to do so is the mark of an accomplished analyst. This project assignment is one example.

Project Background:

Abalones are an economic and recreational resource that is threatened by a variety of factors which include: pollution, disease, loss of habitat, predation, commercial harvesting, sport fishing and illegal harvesting. Environmental variation and the availability of nutrients affect the growth and maturation rate of abalones. Over the last 20+ years it is estimated the commercial catch of abalone worldwide has declined in the neighborhood of 40%. Abalones are easily over harvested because of slow growth rates and variable reproductive success. Being able to quickly determine the age composition of a regional abalone population would be an important capability. The information so derived could be used to manage harvesting requirements.

Supplemental information may be obtained from the following sources:

http://www.fishtech.com/facts.html
http://www.marinebio.net/marinescience/06future/abintro.htm

Background information concerning the assignment data:

The assignment data are derived from an observational study of abalones. The intent of the investigators was to predict the age of abalone from physical measurements thus avoiding the necessity of counting growth rings for aging. Ideally, a growth ring is produced each year of age. Currently, age is determined by drilling the shell and counting the number of shell rings using a microscope. This is a difficult and time consuming process. Ring clarity can be an issue. At the completion of the breeding season sexing abalone can be difficult. Similar difficulties are experienced when trying to determine the sex of immature abalone.

The study was not successful. The investigators concluded additional information would be required such as weather patterns and location which affect food availability.

Assignment 1 is an exploratory data analysis to determine plausible reasons why the original study was unsuccessful in predicting abalone age based on physical characteristics.

Assignment 2 will involve development of a regression model; and, also address development of binary decision rules and a Receiver Operating Characteristic (ROC) curve.

Data set: abalones.csv

Data Description: This data file is derived from study of abalones in Tasmania. There are 1036 observations and eight variables. The CLASS variable has been added for this assignment. Note: **When data sets are made available for public use, the original owners may obscure variable names or scale the data differently from original measurements.** There are different reasons for this. This is the case with these data and will be ignored for this assignment. Basic facts remain.

1. SEX=M(male), F (female), I (infant)
2. LENGTH= Longest shell length in cm
3. DIAM = Diameter perpendicular to length in cm
4. HEIGHT = Height perpendicular to length and diameter in cm
5. WHOLE = Whole weight of abalone in grams
6. SHUCK = Shucked weight of meat in grams
7. RINGS = Age (+1.5 gives the age in years)
8. CLASS = Age classification based on RINGS (A1= youngest,., A6=oldest)

Additional Features:

Additional features that are relevant to answer the questions can be created. In this case, there are 2 additional features that seem relevant. (i) Calculate a new variable VOLUME by multiplying LENGTH, DIAM and HEIGHT together. VOLUME is related to the overall size of an abalone. (ii) Calculate a new variable called RATIO by dividing SHUCK by VOLUME. RATIO is related to the proportion of meat in an abalone.

Data Analysis Project Assignment 1 (50 points due Session 5 (check the syllabus))

Exploratory data analysis (EDA is a process of detective work which may lead to important insights. EDA by its nature tends to be visual. When starting to analyze data, a few good plots may save you hours of pouring over tables and summary statistics. This assignment will use important EDA methods to display aspects of these data such as: 1) the center or location of distributions, 2) the variation in different variables, 3) the shape of various distributions, 4) the presence of outliers, and 5) differences in data characteristics between abalone classifications. Real data are usually not perfect and that is the case here. This work may suggest hypotheses that need confirmatory testing, or it may identify difficulties with the data that need to be addressed in subsequent analyses or future studies of abalones. Assignment 2 will continue with the analysis and build upon what is found in Assignment 1.

Before starting, be sure to review the Data Analysis Video #1 and the self-check page posted on the course site. This latter page shows what the displays should look like.

Data Analysis Project 2 Using R (75 points due Session 9 (check the syllabus))

This assignment will involve development of a regression model; and, also address development of binary decision rules and a Receiver Operating Characteristic (ROC) curve for harvesting abalones. Use your mydata file from the first assignment for this assignment. Results from the first assignment may be referenced as needed, but need not be included or reproduced.

Before starting, be sure to review the Data Analysis Video #2 and the self-check page posted on the course site. This latter page shows what the displays should look like.