MATH 8850

Homework Assignment 3

PROBLEM 1: On the course web page, you will find a data set consisting of data collected on a drug rehabilitation program. The encoding of this data is as follows: Column 1, patient ID; Column 2, age (in years) of patient at enrollment; Column 3, patient depression score; Column 4, Drug use history (1=Never, 2=Previous, 3=Recent); Column 5, the number of times that the patient has undergone prior drug treatments; Column 6, race (0=White, 1=Otherwise); Column 7, treatment duration (0=Short, 1=Long); Column 8, location of treatment (0=Site 1, 1=Site 2); Column 9, whether the patient remained drug free for 12 months (1=Yes, 0=No). Your goal in this problem is to analyze these data and to develop a regression model that can be used to predict whether a new patient is going to be successful at remaining drug free. Your analysis (which should be typed) should include the following:

a A brief description of the data and the goals of the analysis, possibly discuss your intuition about how things are related, etc. We are not just number crunchers, we have to learn to discuss data and the objectives on a analysis.

b Describe in detail the modeling approach that you take, be brief but be comprehensive.

c Issues to consider, link functions, higher order terms, multicollinearity, influential observations, lack of fit, interaction effects, combing qualitative variables.

d Note, every assertion you make needs to be backed up in some way; e.g., plots, hypothesis tests, etc.

e Using what you have learned I want you to identify several candidate models. Then, recalling the goal of the study, I want you to think of a way to assess the predictive power of each of these models, with the model with the best predictive power being selected as your final model.

f Finally, discuss any short comings that the data/model/analysis have that you cannot account for with the given information; e.g., predictor variables that could have been collected, the effect of influential observations, etc.

PROBLEM 2: On the course web page, you will find a data set consisting of data collected during a study examining nesting horseshoe crabs. Each female horseshoe crab in the study had a male crab attached to her in her nest. The study investigated factors that affect whether the female crab had any other males, called satellites, residing near her. Explanatory variables that are thought to affect this included the female crabs color (Column 2), spine condition (Column 3), carapace width (Column 4), and weight

(Column 5). The response outcome for each female crab is her number of satellites (Column 6). Your goal in this problem is to analyze these data and to develop a regression model that can be used to predict the number of satellites a female crab will have. Your analysis (which should be typed) should include all of the elements discussed in Problem 1.

PROBLEM 3: In this problem you will expand on the marathon data analysis that we examined in class. In particular, you are going to compile a data set consisting of the 10 Big City Marathons during 2014; i.e., Houston, L.A., Boston, Portland, Twin Cities, Chicago, Marine Corps, New York City, Philadelphia, Toronto. The data set should consist of the participants ages, gender, and finishing time. Once you have compiled this data I want you to add to it two additional variables (which you will have to search for); mean temperature the day of the race and the elevation of the city in which the race was run. Your goal in this problem is to analyze these data and to develop a regression model that can be used to predict the finishing time of a participant. Your analysis (which should be typed) should include all of the elements discussed in Problem 1.