STAT 530 – Final Exam – Fall 2016

Note: For this final exam, you are not allowed to receive help from anyone except me on the exams. For example, you may not talk to other students about the exam problems, and you may not look at other students’ exams. Violations of this policy may result in a 0 on the exam, an F for the course, and/or punishment by the USC Office of Academic Integrity.

1. You are working as a statistician doing oversight for a food-safety and nutrition board, as part of a governmental investigation. The agency has selected 43 types of cereal produced by three companies (General Mills, Kellogg’s, Quaker). Eight numerical nutritional characteristics have been measured for each cereal. These 8 variables are: Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, Potassium. In addition, the data file gives the name of each cereal and the company that produced the cereal.

The questions that the nutrition board would like answered include:

(1) Are there notable associations/relationships between some of the variables? (if so, describe them)

(2) Is there a way to graphically represent the raw data for the 43 cereals and draw conclusions about the data set from such a graph?

(3) What are the basic underlying groups that the cereals form? Can you plot the data in a small number of dimensions, showing the group separation of the cereals?

(4) Would you say the three companies are the same in terms of their average values for the eight variables? Can you test this formally? Are there interesting differences between the three companies’ cereals that can be displayed using the variables (or combinations of variables) in the data set?

(5) If the researchers were to investigate a new cereal observation that had known measurements for the eight variables, could we determine a rule for predicting it to belong to one of the three companies? How accurate could you expect such a rule to be?

(6) In particular, we have a new cereal with measurements given as 150 calories, protein content of 3, fat content of 2, sodium content of 130, fiber content of 3.5, carbohydrates of 12, sugar content 10, and potassium content 180. Which company would you predict that this came from, and how confident are you in the classification?

(7) What are any other potentially interesting aspects of the data set?

You will type a roughly 3-page or 4-page report detailing your analysis of the data and your conclusions. Your report should be understandable and meaningful to the general public, to the leaders of the cereal industry, and to statisticians who will be reviewing your report. You may include graphs that illustrate and/or support your findings (such graphs don’t count against the 3-4 page limit). Do NOT include computer code within the main body of your report. You may include such code in an appendix if you wish. The data for this problem are given at the link “Cereal Data” on the course web page. (Original Data collected by Chad Dacus). The following code should read the data into R correctly:

```r
cer <- read.table("http://www.stat.sc.edu/~hitchcock/cereals.txt", header=T); attach(cer)
cer.data <- cbind(Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugar, Potassium)
cereal.names <- as.character(Brand); cereal.company <- as.character(Company)
```
2. You are working as a statistical consultant for an ecological study about the weights of mammals. Data were gathered on 51 mammals. The variables measured include weight measurements (body weight and brain weight); measurements of characteristics of the mammal (total daily sleep, maximum lifespan, and gestation time), and three indices noted by ecological experts. The units of measurement are given below. The data file contains the observed values of the 8 numeric variables (plus a labeling column with the names of the species) for 51 mammals.

<table>
<thead>
<tr>
<th>species of animal</th>
<th>body weight in kg</th>
<th>brain weight in g</th>
<th>total sleep (hrs/day)</th>
<th>maximum life span (years)</th>
<th>gestation time (days)</th>
<th>predation index (1-5)</th>
<th>sleep exposure index (1-5)</th>
<th>overall danger index (1-5)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1 = minimum (least likely to be preyed upon)</td>
<td>1 = least exposed (e.g. animal sleeps in a well-protected den)</td>
<td>(based on the above two indices and other information)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5 = maximum (most likely to be preyed upon)</td>
<td>5 = most exposed</td>
<td>5 = most danger (from other animals)</td>
</tr>
</tbody>
</table>

NOTE: It is recommended by the ecologists that you use a natural log transformation of the body weight and brain weight variables before doing the analysis (see the code below).

The questions that the ecologists would like answered include:

(1) Are there notable associations/relationships between some of the variables? (if so, describe them)
(2) Can you use the dissimilarities among the mammals to plot the mammals on a 2-dimensional map? If so, how would you characterize those two dimensions, and what do you learn about these data from such a map?
(3) Considering the entire data set, we wish to build a model to predict or explain the two weight variables (log body weight and log brain weight) based on the six other variables. Construct such a model and evaluate its appropriateness.
(4) Use your model for prediction of the two weight variables for another mammal, which has total sleep 16.5 hours per day, maximum life span 11.0 years, gestation time 25 days, predation index 5, sleep exposure index 2, and overall danger index 3.
(5) BREAKING NEWS: It is later discovered that the mammal whose weight variables you predicted above is actually an Arctic ground squirrel. Knowing this, comment on whether your prediction seems sensible or reasonable.
(6) It is argued by some ecologists that the three “index” variables are subjective and should not be used in the model. Determine whether the model could predict the weight variables just as well using only total daily sleep, maximum lifespan, and gestation time as it could using all six variables.
(7) What are any other potentially interesting aspects of the data set?

You will type a roughly 3-page or 4-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the ecological researchers, who are not experts in statistics; and your own supervisor at the statistical consulting company, who will be judging
you and deciding on your possible promotion based on the statistical competency of the report. Your report should be understandable and meaningful to both audiences.

You may include graphs that illustrate and/or support your findings (such graphs don’t count against the 3-4 page limit). Do NOT include computer code within the main body of your report. You may include such code in an appendix if you wish.

The data for this problem are given at the link “Mammal Data” on the course web page. (Original Data collected by Allison, T. and Cicchetti, D.). The following code should read the data into R correctly:

```r
mammals530 <- read.csv("http://people.stat.sc.edu/Hitchcock/mammaldata.txt",header=T);
attach(mammals530)
mammal.species <- as.character(species)
mammal.numeric <- mammals530[,,-1]
lnbodyweight <- log(bodyweight)
lnbrainweight <- log(brainweight)
```

**Grading Scale:**

Each problem will be worth 30 points, for a total of 60 points. For each problem, your report will be graded based on Writing, Analysis, and Context. For example:

**Writing** (out of 10 points): How organized, clearly written, comprehensible, and grammatically correct is the report? Would the client reading this report be confident that it was written by an educated, well-trained statistical scientist?

**Analysis** (out of 10 points): Were the graphs and data analyses appropriate for the problem? Were the analyses carried out correctly? Were your statistical conclusions about the data set sensible and clearly justified by numerical or graphical evidence?

**Context** (out of 10 points): Were the questions answered in terms of the variables of the data set? Although you are not an expert in the field as your client is, have you attempted to frame your conclusions and interpretations in a subject-matter context rather than treating the data as simply a meaningless set of numbers?