STAT 530 – Final Exam – Fall 2022

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.

IMPORTANT: The following problems involve real data sets! The answers you get about the data may not be "perfect" or idealized. Your reports should address the questions of the clients in the best way you can, but it is NOT expected that every question the client wants addressed will have a perfectly clear answer. Your job is to explain and show what you have learned from the data, in no more than three pages (per problem) of clear writing (plus any relevant graphs and plots that you want to include at the end).

Furthermore, it is expected that each statistician (student) will make somewhat different conclusions or illuminate somewhat different aspects of the data set. For many of these questions, there is not only one right answer!

By the way, although the clients' questions are numbered, your report ideally will NOT be in the form of a numbered set of answers. It's better if your report addresses the questions in a more unified and flowing way, and you can arrange the flow of the report however you feel is appropriate. Make your report reflect good statistical work, but just as importantly, do your best to make it interesting to the reader! If the report is so dry and boring that the client's eyes glaze over when he or she tries to read it, then you haven't been a successful statistical communicator.

While you are working on the exam: You can ask me questions, but I may or may not give an answer. If it is a clarification question, I may answer that, but if it is about which approach to take on a problem or how to do a data analysis, I will likely not answer that, since this is a test and I want you to know that (or to look back through your notes and course resources for insight). If you have a question about R code that is not working: If it's basically correct except for a tiny thing causing an error, I may be able to point you in the right direction. But if it is something substantially wrong, I will just advise you to rethink your approach and look back at class examples.

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?

1. 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.

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

NOTE: It was recommended by the ecologists that we use a natural log transformation of the body weight and brain weight variables before doing the analysis (that is why the data you read in is in terms of the natural logarithm of the weight variables). However, when dealing with a numerical predicted weight, to make the prediction more interpretable, you should back-transform (exponentiate) the logged predicted value at the end, so that the weight is back in terms of kg (for body weight) or grams (for brain weight).

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) Is there a way to graphically represent the raw data for the 51 mammals and draw conclusions about the data set from such a graph?

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

(4) Considering the entire data set, we wish to build two models to predict or explain each of the two weight variables (log body weight and log brain weight) based on the six other variables. Construct both such models and evaluate their appropriateness.

(5) Ecologists have been observing another mammal, an Arctic ground squirrel, but they have not been able to weigh it, because every time they get near the darn critter, it bites them. Use your models for predicting each of the two weight variables for this 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. Are your predictions reasonable/sensible?

(6) It is argued by some ecologists that the three “index” variables are subjective and should not be used in the models. Could your models 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 (at most) 3-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-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 code to read the data into R correctly is given on the course web page.

2. You are working as a statistical collaborator for an environmental protection agency examining factors influencing damage from forest fires. Data have been collected on 517 instances of forest fires, some ultimately harmless and others very serious. Each observation has been classified as "None", "Moderate", or "Severe" depending on the area of forest burned. The variables measured include temperature (in degrees Celsius), relative humidity (in percent), wind speed (in km per hour), and rain amount (in mm per square meter). The data file contains the observed values of these 4 variables (plus an indicator of which "burned" category the observation was placed into) for 517 fires.

The questions that the company would like answered include:

(1) Are there notable associations/relationships between some of the variables? (if so, describe them)
(2) Would you say the three categories of fire severity are the same in terms of their average values for the temperature, relative humidity, and wind speed variables in particular? Can you test this formally? Are there interesting differences between the three severity categories that can be displayed using the variables (or combinations of variables) in the data set?
(3) If the company were to consider a new fire with measurements on the four weather conditions, could we determine a rule for classifying the severity of the burning? How accurate could you expect such a rule to be?
(4) Most of the fires in the sample occurred when rainfall was exactly 0. Should a classification rule include rainfall as a variable or not? Discuss.
(5) In particular, we have a hypothetical new fire with weather conditions given as 17 degrees C, 20 percent relative humidity, wind 6 km/hour, and no rainfall. Which severity category would you classify it as coming from, and how confident are you in the classification?
(6) What are any other potentially interesting aspects of the data set?

You will type a (at most) 3-page report detailing your analysis of the data and your conclusions. Keep in mind that the report should be written for two audiences: the environmental agency supervisor, who is not an expert 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-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 "Fire Data" on the course web page. (Original Data collected by P. Cortez and A. Morais). The code to read the data into R correctly is given on the course web page.